A quantitative framework for valuation multiples computation in mergers and acquisitions: integrating bidirectional encoder representations from transformers, principal component analysis, and predictive modeling

by Giovanni Maria Parlangeli

ABSTRACT - Artificial intelligence technologies are increasingly infiltrating the financial industry to optimize processes previously subject to human intervention and interference. The mergers and acquisitions (M&A) sector is a fertile ground for harnessing the advantages these tools can offer, in terms of productivity and financial accuracy. This study starts from the traditional process of estimating a company's valuation multiples and proposes a rigorously quantitative approach to generalize it. The study focuses on publicly traded U.S. companies under SIC code 35 to develop a machine learning model that, based on the financial and operational characteristics of these companies, can predict the most appropriate valuation multiple for a target firm in the same industry. By testing three different algorithms (Random Forest, Gradient Boosting, and Support Vector Machine) with three different test set sizes, the study presents a model that can describe the variability in the sample data with an R-squared of 0.314. Furthermore, the study demonstrates that the models including business descriptions encoded into vectors via BERT (Bidirectional Encoder Representations from Transformers) outperform the models employing the traditional SIC code classification method in a statistically significant manner.

INDEX

1. Introduction to the topic and its relevance

1.1 Mergers and acquisitions

The term Mergers and acquisitions (M&A) describes a set of business transactions aimed at consolidating the operations and assets of two or more companies. We can outline the difference between these two specific operations as follows [8].

A merger takes place when companies combine to form a single entity, pooling their assets, resources, and operations to create a broader organization. The parties involved usually agree to unify their ownership, management, and control, and aim at achieving strategic goals. Such goals include increasing market share, entering new markets, achieving economies of scale, and leveraging competitive advantages.

In an acquisition, we identify a company willing to take over another, gaining control over its operations, assets, and resources. The acquiring company absorbs the target firm by purchasing its shares, assets, or a significant ownership stake. Usually, the acquirer integrates the target company's management and ownership with its existing entity. Acquisitions as well are motivated by strategic objectives, such as expanding market presence, acquiring new technologies, accessing new markets, hiring skilled employees, or reducing competition.

In 2023, the M&A industry proved again to be a significant force in the global economy, with over 3,059 announced and completed deals valued at least $25 million each, for a total transaction value of approximately $3.1 trillion. However, this represented a 16% decline from the previous year's results. Besides valuation discrepancies, increased regulatory scrutiny, and political pressures, this recent downtrend can be addressed to the peculiar macroeconomic landscape, characterized by high interest rates and economic uncertainty [11].

1.2 The M&A process

According to past research and literature, six main stages can be identified throughout a typical M&A transaction: initiation, target identification, due diligence, deal structuring and negotiation, post-merger integration, and performance measurement [8].

1. The initiation phase stems from a strategic decision by the acquiring or target firm and can be influenced by external actors like investment banks or investors. As mentioned in the previous paragraph, the typical drivers behind this decision are strengthening market power, creating economies of scale, diversifying risks, and taking countermeasures against current threats. Past research also proves that CEO characteristics and industry trends can play significant roles in igniting and shaping the first stage of the M&A process [8].

2. In the target identification phase, the goal is to find potential target firms through extensive market research. At this early stage, the focus is on gathering basic information and conducting preliminary analyses that can help build a list of suitable candidates, in terms of market position, financial status, and potential synergies to be created. Usually, advisors and their networks play a major role in spotting firms that fit the client's needs.

3. The Due Diligence phase starts after the most suitable candidate for the transaction has been identified. In this stage, the interested parties audit the firms' financial results and legal compliance to check that the facts and information reported correspond to reality before entering into proper agreements.

4. During the Deal Structuring and Negotiation stage, the parties define and agree upon the financial details of the deal, including the acquired stake, the payment method, and the terms and conditions of the deal. A key aspect of this phase is the business valuation process, where the enterprise value of the target company is determined. This is the sub-phase that this research tries to revolutionize, defining a

generalizable framework that does not rely upon human consideration and intervention. Currently, the most popular valuation methods include the discounted cash flow (DCF) analysis, which estimates the present value of the firm's future cash flows, and the multiples approach, which compares the target to similar companies using financial ratios like price-to-earnings (P/E) and enterprise-value-to-sales (EV/Sales). Such ratios are computed for publicly traded companies deemed "comparable" to the target and then applied to the latter's appropriate financial metrics. The methods used to assess whether a firm is comparable to the target do not follow a rigorous framework and can lead to potentially significant valuation discrepancies according to the operating choices of the advisors involved.

5. The Post-Merger Integration stage allows to practically merge the operations, cultures, and systems of the interested entities.

6. The final Performance Measurement has the objective to assess whether and up to which extent the M&A deal has achieved the strategic objectives outlined in the initial stages of the transaction, usually keeping track of key performance indicators.

1.3 Artificial Intelligence: basic concepts and applications in M&A

The technologies that fall under the scope of Artificial Intelligence are reshaping the way we conceive information, productivity, and human work. By reproducing the cognitive capabilities of a human being, while being able to handle much more intensive computations, these tools provide fast and reliable solutions across a diverse set of business processes, allowing to streamline repetitive tasks and keeping output quality high. In the M&A industry, the opportunity to let the human capital spend more time on assignments carrying greater added value for the organization can lead to a notable competitive advantage. The field of Artificial Intelligence encloses a wide range of technologies. In this paper, we will focus on and leverage a specific branch of them, which can be identified as Machine Learning. This term defines a set of computational methodologies that are able to learn from data and can subsequently carry out prediction tasks.

According to the “learning architecture” of each model, Machine learning algorithms can be categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning models work with datasets including both features - the input variables - and labels - the variables we aim to predict - for each data point. The term "supervised" recalls the fact that for each prediction that our model makes we can compare the estimated label with the actual one and adjust consequently the models' parameters to obtain more accurate predictions. At this point, we can also make a distinction between the two main different types of tasks carried out by such models: classification and regression. Via classification, your intent is to categorize the data according to discrete labels, like grouping emails as spam or not. Via regression, you try to estimate continuous target variables, like predicting the price of a car on the market from its technical features.

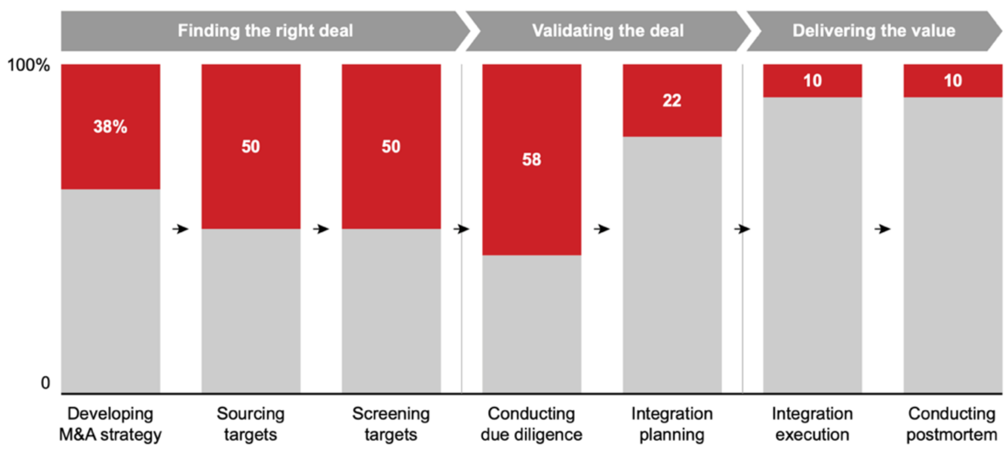
In unsupervised learning, there are no target variables to compare the model's predictions and adjust its parameters accordingly. These models are fed with unlabeled datasets, consisting of the input feature space only, and try to recognize without "supervision" similarities and patterns to describe the data.

Reinforcement learning follows a different setting. In this setup we typically have an agent that learns to make decisions via a trial and error approach, performing actions in an environment that has to be explored to understand which behaviors maximize his cumulative reward. For this reason, reinforcement learning performs much better in solving optimally sequential decision-making problems, rather than simple classification or regression tasks.

Since the model I developed relies heavily on the transformation of verbal business descriptions into numerical vectors that synthesize their content, I think it's important to provide a background on Generative AI technologies, which allowed me to practically carry out this "translation". Generative AI models focus on creating new output resembling existing patterns in the input data: the information within human-generated content like text, images, or music, is synthesized into vectors and used to understand how to reproduce autonomously such content, including improvements or other changes.

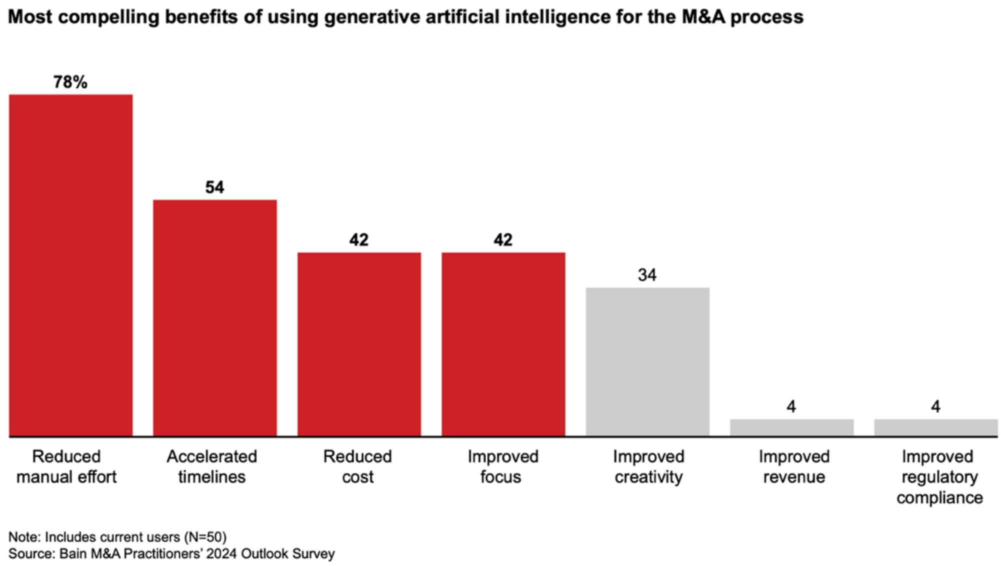
In the M&A field, generative AI gained lots of attention over the last few years, given its potential to significantly improve output quality and reduce time spent on repetitive workloads. A recent survey from Bain & Company of over 300 M&A practitioners [2] indicates that while current adoption of AI in M&A processes is limited to 16%; there is a strong expectation of growth with 80% of respondents planning to integrate AI within the next three years, in particular, in sectors such as technology, healthcare and finance. In accordance with the information collected in this survey, generative AI is generally implemented to produce new ideas and review data during the due diligence phase. This technology can make a difference even in more complex tasks like identifying potential acquisition targets. Early adopters have reported benefits such as: increased productivity, faster timelines, cost reductions and enhanced employee focus with 85% of users stating that AI capabilities met or exceeded their expectations.

Percentage of M&A practitioners using generative artificial intelligence at each step



Note: Includes current users (N=50)
Source: Bain M&A Practitioners' 2024 Outlook Survey

Currently, there are plenty of relevant problems to address in order to ensure a correct implementation of these technologies: data inaccuracy, privacy breaches and cybersecurity risks. It is evident that these innovative tools bring a series of opportunities and limitations to the firms exploiting them. Therefore, it's becoming crucial for companies to integrate them strategically in areas requiring high manual effort, repetitive tasks, or significant creative input. According to Bain & Company's report, the best steps that employers can take to ensure a safe use are investing in proprietary databases, providing appropriate employee training, checking data accuracy and performing human review.



1.4 M&A Research Institute Holdings

An eminent example of the successful implementation of AI in the M&A process is the case of M&A Research Institute Holdings, a listed M&A advisory firm founded in 2018 by Shunsaku Sagami. Starting from the founder's experience with artificial intelligence and his intuition about the revolutionary impact it could have in such industry, the company built a proprietary AI platform to ease hundreds of M&A deals focusing on small and medium-sized firms, which risked shutting down because they lacked a succession plan.

The Japanese government stated that by 2025 there will be 1.25 million national businesses at risk by reason of no continuity plans, leading to 6.5 million jobs lost and $240 billion worth of GDP erased [10]. Sagami's disruptive innovation aims to provide a future to these firms by leveraging the potential of strategic M&A combined with the efficiency of machine learning algorithms.

His company's platform integrates a diverse set of data, spanning from financial information and management team profiles to competitive landscapes, and identifies optimal buyer-seller matches for its clients, reshaping even the cultural perception of selling out a company. In these few years of activity, the institute has significantly improved the speed and efficiency of M&A transactions, reaching an average deal closing time of approximately 6.2 months, with some deals closing in as few as 49 days [1]. This efficient approach has allowed M&A Research Institute to grow to more than 160 employees and about 500 deals in the works as of May 2023, doubling its sales as well to ¥3.9 billion, compared to the previous year [10].

This firm's fee structure is also breaking new ground since it relies completely on performance fees with no retainer or interim charges. This approach creates a competitive advantage and makes the service more attractive: this generates a high volume of transactions and consequently allows to refine regularly the advisor's proprietary AI models. M&A Research Institute Holdings exemplifies how AI can transform M&A deals by making them faster, cheaper and more efficient. The success of this business model serves as an inspiration for my thesis, displaying how AI can revolutionize the traditional M&A landscape.

2. Research question

The research question of this thesis starts from the corporate valuation process for firms. In this context, valuation multiples play a paramount role as metrics for assessing the worth of private companies. The so called "multiples approach" entails taking ratios between the enterprise value (or equity value) and other financial metrics (like earnings, sales, or EBITDA) for a set of similar public companies, and then multiplying either the average or the median of such ratios by the appropriate financial metric of the target private company, thus obtaining its estimated enterprise value (or equity value). A central assumption of this method is that companies with comparable characteristics should have similar valuations.

Through this research I want to question the appropriateness of the traditional approach for computing such multiples, since it lacks a standardized, universally accepted framework. Consequently, this process can vary significantly from case to case, with differences in the method of selecting comparable companies, or in the type of statistic used to extrapolate the final multiple from the firms' sample set.

The following pages propose an innovative and rigorous framework for computing valuation multiples via a machine learning model trained both on financial data and operational features of firms in a specific US industry. Analyzing the patterns bounding such features to the valuation multiples for public companies, this framework will predict proper multiples for the target private companies whose features we might input.

In contrast to the traditional models based only on financial ratios, my framework implements a popular generative AI model to encode business descriptions into numerical vectors. These are consequently added to the initial feature space. The procedure is an attempt to prove that incorporating detailed business descriptions should generate models with greater accuracy and predictive power compared to others relying on traditional business classification codes like the Standard Industrial Classification (SIC) codes.

It is important to note that, although past literature includes several papers on predicting which companies might be optimal acquisition targets by leveraging machine learning models and large language models, there are no examples of an approach similar to the one proposed in this research for a rigorous and universally applicable computation of valuation multiples.

3. Methodology for collecting data and developing model

3.1 Financial data selection

The dataset used for this study is sourced from Orbis version 347 update 347001. To make sure our results are robust and avoid industry-based biases, the study focused on companies operating within the same macro-industry, in this case identified by the Standard Industrial Classification (SIC) code 35 (industrial and commercial machinery and computer equipment). The additional selection criteria for the companies to be included in this study were an active company status, a United States domicile, and public listing status. In order to preserve data relevance, all companies lacking financial data in the last five years or categorized as public authorities/governments were excluded from consideration. This initial filtering process yielded a dataset of 395 companies, with data spanning fiscal years 2018 through 2022 to compute compound annual growth rates and to avoid eventual missing data from incomplete 2023 statements.

The feature selection process in this paper was inspired by the study titled "A Methodological Study on Selection of Comparable Companies for Mergers and Acquisitions of Listed Companies Based on KNN Algorithm" by Zihan Gao (2023). This research used a robust and comprehensive set of features, including total asset turnover, capital intensity, return on investment, and growth rates, to determine optimal comparable companies via a KNN algorithm. Given the similarity of the research purpose, the business metrics proposed in that paper were taken into careful consideration and some of them are included here.

The final dataset contains 18 features, besides the ones coming from the encoded business descriptions we will analyze later. Here is the list of the selected features along with their variable names:

1. Primary SIC number (sic_prim)
2. Absolute number of employees in 2022 (n_empl_22)
3. EV/SALES multiple as of 2022 (ev_sales_22)
4. Absolute number of publications (n_publications)
5. CAGR of revenues from 2018 to 2022 (cagr_revs)*
6. CAGR of total assets from 2018 to 2022 (cagr_ta)*
7. Average total assets in the last 2 years (2y_avg_ta)
8. EBIT margin as of 2022 (ebit_m_22)
9. Net income margin as of 2022 (ni_m_22)
10. EBITDA margin as of 2022 (ebitda_m_22)
11. Capital expenditures over revenues as of 2022 (capex_revs_22)
12. CAGR of capital expenditures from 2018 to 2022 (cagr_capex)*
13. Average capital expenditure in the last 2 years (2y_avg_capex)
14. Revenues over current assets as of 2022 (ca_turnover_22)
15. Total assets over total revenues as of 2022 (cap_intensity_22)
16. Return on assets as of 2022 (roa_22)
17. Total assets over total liabilities as of 2022 (ta_tl_22)
18. English description of the firm's operations (td)

In the following pages, I will develop a model that takes as input financial and qualitative features (from the business descriptions) to predict the EV/Sales multiple. I decided to focus on this particular multiple because of several considerations that prove the suitability and robustness of this variable as optimal target for the machine learning model I will outline.

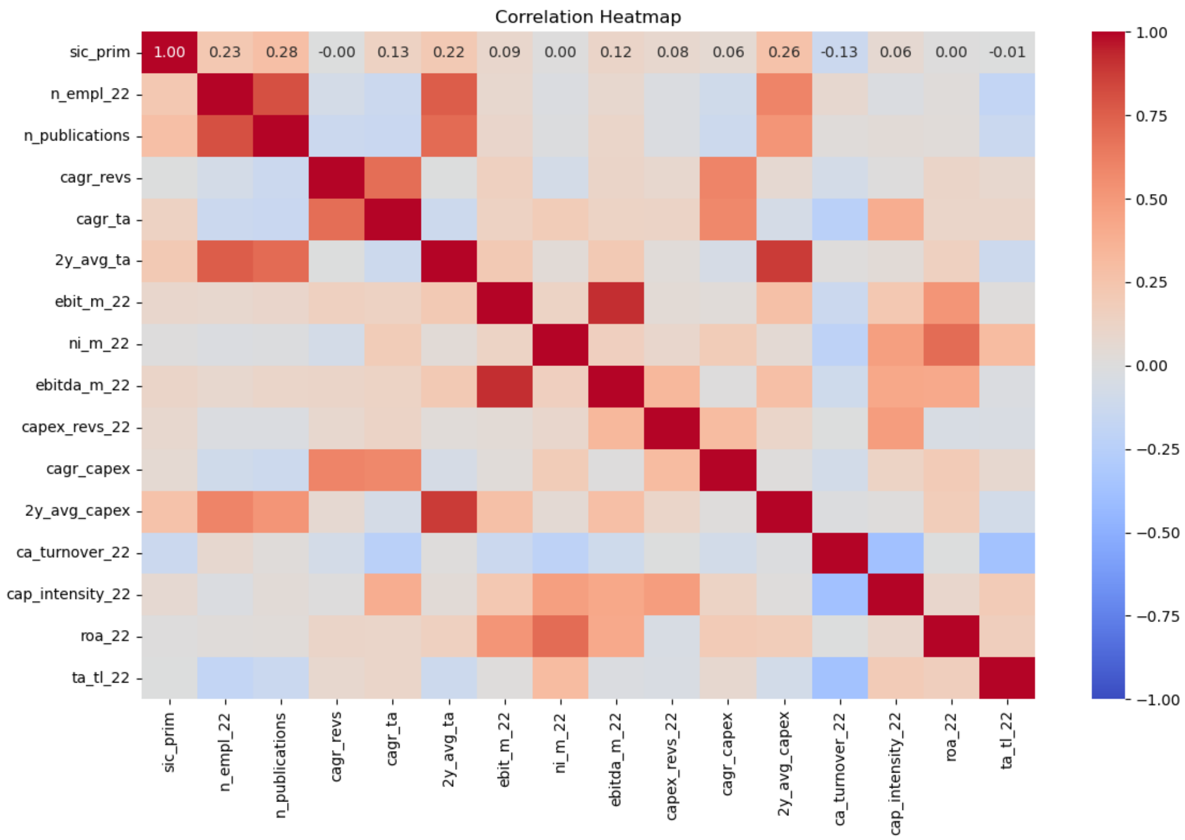
First of all, if we compare EV/Sales to other common multiples, like the Price to Earnings ratio (P/E) and the Enterprise Value to EBITDA multiple (EV/EBITDA), EV/Sales stands out as it cannot take on negative values. This peculiarity is pivotal for our research because relying on potentially negative multiples could lead us to exclude loss making companies or introduce biases due to nonsensical negative values.

In addition to this, the EV/Sales multiple is based on revenue, a financial metric that is less susceptible to accounting manipulations like non-cash charges, and it is generally reported by all companies, regardless of their industry. This broad applicability allows to develop more accurate models, which should also generalize well across diverse datasets.

In the last step of this data selection process, I wanted to make sure the whole framework would have been consistent and reliable; thus, all missing values in the original dataset were not approximated via feature engineering methods, and the companies reporting them were excluded from the analysis. Consequently, our final dataset comprised 152 firms.

3.2 Features Selection Process

Before feeding our model with the data outlined above, it’s important to make sure that the features selected are relevant and independent among them. Potentially, in my specific case, I am going to measure the independence of the features through the linear Pearson correlation. High levels of correlation generally bring the model to multicollinearity problems. I can empirically assess whether this problem arises by taking a look at the p-values of the single features and at the F-test on the overall model. For this purpose, we start by developing a correlation matrix.



From this correlation matrix, we can clearly identify two pairs of variables standing out for their strong correlation coefficient: EBIT Margin (ebit_m_22) and EBITDA Margin (ebitda_m_22); 2-Year Average Capital Expenditure (2y_avg_capex) and 2-Year Average Total Assets (2y_avg_ta).

Between EBIT Margin (ebit_m_22) and EBITDA Margin (ebitda_m_22) I decided to retain only EBITDA Margin (ebitda_m_22) because it is less susceptible to non cash expenses and hence provides a reliable and more comparable value across companies with different magnitude of capital investments.

Comparing 2-Year Average Capital Expenditure (2y_avg_capex) and 2-Year Average Total Assets (2y_avg_ta) I decided to keep in the model only 2-Year Average Total Assets (2y_avg_ta) because it provides a broader and more stable representation of the firm's resources.

As an overall check before proceeding with model implementation, after I had removed the redundant variables, I conducted an F-test to make sure that the dataset as a whole was significant to describe our target. The F-test yielded an F-statistic of 15.86 and a p-value of approximately 4.95e-35, thus rejecting the hypothesis of statistical non-significance of the financial input data.

3.3 Verbal Firm Description Encoding via BERT and PCA

As anticipated earlier, in this paper I aim to leverage the verbal descriptions of the firms' operations to capture their qualitative and non-financial characteristics, in an attempt to build a model able to identify patterns and similarities among firms not only in the financial sphere, but also from an operating perspective. For this purpose, I employed the Bidirectional Encoder Representations from Transformers (BERT) model, a natural language processing framework that allowed me to encode the textual descriptions into high-dimensional vectors.

First of all, I want to provide a detailed conceptual background on BERT (Bidirectional Encoder Representations from Transformers) and PCA (Principal Component Analysis), so that it's easier to understand the steps following this section of the research.

3.3.1 Bidirectional Encoder Representations from Transformers (BERT)

BERT models (Bidirectional Encoder Representations from Transformers) can be defined as a subset of generative AI models; they were made popular thanks to their notable ability to understand text and were consequently applied to more demanding natural language processing workloads. They were first developed by Google AI and leverage a particular deep learning architecture called "transformer", which was initially introduced by Ashish Vaswani in his groundbreaking paper "Attention is all you need" [14]. In the case of this specific model, a set of multi-layered self-attention mechanisms allows the model to focus on different parts of the input text simultaneously. Also, BERT employs an innovative bidirectional approach, which allows to consider both preceding and succeeding terms when processing each word in a sequence: this has enhanced significantly the model's ability to capture contextual and semantic meaning [4]. Besides the exceptional technical progress, the key to BERT's astonishing results are extensive training text corpora to catch linguistic patterns and contextual distinctions effectively.

Looking at the specific natural language understanding tasks, BERT models are mostly used in the fields of language inference, sentiment analysis and question-answering tasks. In this study, employing BERT allowed to analyze and extract meaningful qualitative features, translated by BERT into numerical vectors, from the companies' business descriptions. This passage aims to uncover essential aspects of a company's operations and industry positioning, in an attempt to broaden and deepen the range of information provided by traditional industry classification methods, like SIC codes.

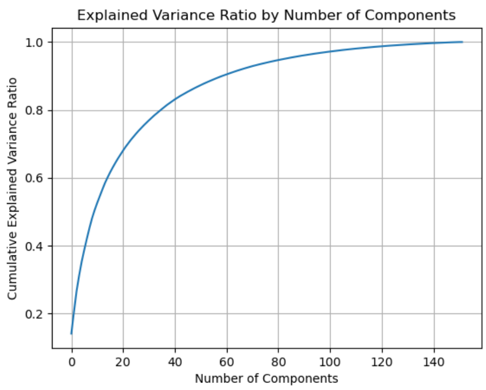
3.3.2 Principal Component Analysis (PCA)

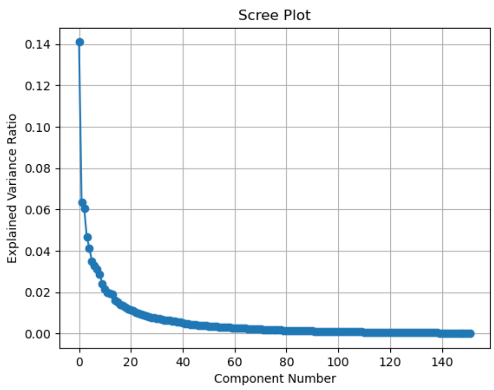
Principal Component Analysis (PCA) is a statistical procedure that allows to transpose high-dimensional data into a lower-dimensional space. In plain words, PCA is a popular technique employed to condense the original data structure without losing the information that describes most of the variability of such data [6]. All this is performed by establishing a new coordinate system for the data. Within this new space, the "principal components", which are basically the original information vectors shrunk in lower-dimension, determine the directions of maximum variance. These components can be ordered among them, and as we go from the first to the last component, the respective explanatory power decreases. For these reasons, PCA is convenient for tasks like data classification, visualization, and storage, and proves to be a crucial aid for carrying out tasks involving machine learning models, in particular with a vast feature space.

3.3.3 Text Preprocessing, Encoding and Features Selection via PCA

We cannot feed the BERT model without first performing textual preprocessing on the firms' descriptions: this involves removing punctuation and stopwords to focus on the key terms sharing insights about the companies' activities. Only at this point BERT comes into place. After loading the processed 152 texts in our GenAI model, I got as output the same number of numerical embeddings, each capturing the business peculiarities of the respective firm in a high-dimensional vector space.

Consequently, we applied PCA to the 152 vectors containing the descriptions' embeddings. To get a clearer picture of what I'm doing, it's important to specify that BERT gave us back a dataset of 152 rows (number of firms included in our analysis) and 768 columns (number of numerical features needed to describe each firm in the latent space that BERT generated). It is clear that 768 features per firm just to describe the respective activities would make our model infeasible. This is where PCA comes into place as a powerful tool to reshape this original high-dimensional space into a lower-dimensional one and distill the essential "columns" from it.





The plots above report the cumulative explained variance ratio and the individual explained variance ratio: on the left side we can see how the share of total variance explained changes by increasing the number of principal components – a very helpful tool to determine the optimal number of components to retain. In this type of process, the goal is to strike a balance between dimensionality reduction and information preservation. On the right, instead, we can see how much of the total variability each ordered principal component is able to explain.

Personally, I considered the practical and computational implications on the model, and I finally chose to continue the research with 15 principal components, able to explain the 61.6% of the total variability in the original dataset, which integrated 768 features in total. In this way, I am sure to include the majority of informative details about the firms without overloading the models.

3.4 Model Development

Before proceeding with the models' implementation, in this section, I will provide a brief overview of the three types of machine learning algorithms that will be trained, tested, and compared: Random Forest, Gradient Boosting, and Support Vector Machine. I decided to focus on such architectures because, according to their technical capabilities, given the dataset selected and the supervised regression task, they should provide sufficient flexibility and robustness.

3.4.1 Random Forest

Random Forest models are an example of ensemble learning methods. When we talk about ensemble learning we refer to all those models that describe their target variable by training many individual models and merging them together. In the case of Random Forests, the individual models to be combined are multiple decision trees, an example of non-parametric, supervised algorithms used for both classification and regression. Each tree is trained on a subset of the entire dataset and with only a subset of all the features. Random Forest is able to mitigate overfitting problems by aggregating the predictions of all these individual trees and can provide a robust estimate of each feature's importance in describing the target, which proves to be a very insightful property in the case of a research question like this one. Random Forest's drawbacks include struggling with highly correlated features and poor descriptive performance with noisy or imbalanced data.

3.4.2 Gradient Boosting

Gradient Boosting is also an ensemble learning technique. Differently from Random Forests, Gradient Boosting models build sequential trees trained to correct the errors made by the previous ones. This iterative process proves to be extremely accurate in capturing subtle relationships within complex datasets and, for this reason, might provide very interesting results for this study. A peculiarity of Gradient Boosting models is their robustness to outliers and missing values, thanks to their sequentially improving architecture. On the other hand, these models are sensitive to overfitting, especially when the number of sequential trees to be built is not fine-tuned, and might be computationally expensive to train with large datasets.

3.4.3 Support Vector Machine (SVM)

Support Vector Machines, usually referred to as SVM, follow a completely different procedure: they work by mapping the input data into a higher-dimensional feature space, where they try to construct an optimal hyperplane to separate the different classes of data. For this reason, SVM models work well with complex classification problems, but can also be employed to define decision boundaries for regression tasks. The main drawbacks are: worse performance when applied to large datasets, since training times scale quadratically to the number of samples, and poor descriptive ability with noisy data or coincidental classes.

3.4.4 Models Implementation

Now that I have outlined the main characteristics of the models employed, I will dive deep into the procedural phases I followed in the study. Each algorithm (Random

Forest, Gradient Boosting and SVM) is trained using two different datasets: the first one contains the encoded company descriptions, without the primary SIC code for each firm; the second contains only the primary SIC codes without the business descriptions. At this point, the input data are normalized and the dataset is split into training and test sets according to three different test set sizes: 20%, 25%, and 30%.

Since this research aims to draft a universal approach for a case-specific process, typically built on data manipulated according to various financial methods, it is not trivial to identify the most appropriate metrics to evaluate its performance. I did not identify any alternative method in the literature that would suit the specific purpose of this study, hence I decided to employ some of the traditional evaluation metrics for regression problems: MAE (Mean Absolute Error), MSE (Mean Squared Error), and R-squared.

The models' performance according to these metrics varied across different test rounds. In order to provide a solid evaluation, I trained each dataset with each model 50 times, using a different random seed every time. Random seeds are numbers used to initialize the machine learning models and can affect how the data is split into training and test sets, the models' initial parameters, and the order of data processing; all these factors impact the models' performance. By averaging the results over 50 runs, I tried to reduce the impact of this variability and obtain a robust assessment of the models' capabilities. At this point, I think it's also important to specify that each time a model was trained, the code included a snippet to optimize the hyperparameters for that specific iteration.

Once all the models completed their test phase, I computed the mean and standard deviation for each distribution of model evaluation metrics and analyzed the significance of the differences between the models including only SIC codes and the ones with encoded business descriptions.

4. Results

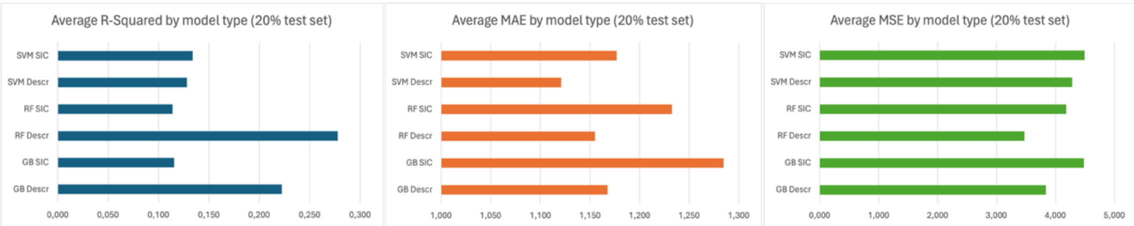
In this section, I present the results of our model performance evaluation. First of all, to ensure a clear understanding of the results presented, I will go over the meaning of the performance scores employed, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2).

I start by introducing the metric I will mainly use to compare description-based and SIC-based models: R-squared. This statistical measure shows how much of the variability in the dependent variable can be described by the model. An R-squared value of 0.70 means that the model explains 70% of the variance in the dependent variable and, following the same logic, an R-squared equal to 0 suggests the model cannot describe any of the variance. Incurring in a negative R-squared warns that the model has an even worse predictive power than the sample average of the observed variable. We can say that, in general, a higher R-square indicates better model approximation of the data, but it's crucial to understand that this does not confirm the validity of the model as-is, because this metric only reflects the fitting power within the sample space. Another relevant consideration in the context of this research is that the R-squared can be a misleading metric if there are outliers in the input data or if the model follows non-linear relationships.

Mean Absolute Error (MAE), the second metric I used, is the average of the errors between estimated and actual test values and does not take into account the direction of such errors. MAE keeps the same measure unit as the target variable, making it very easy to interpret the average magnitude of the errors.
The Mean Squared Error (MSE) measure also computes the average error, but squares the differences between actual and predicted values before averaging them, hence giving more weight to larger errors. This feature allows to identify easily abnormal deviations, but can also take on disproportionate values because of outliers.

In the following charts, the Gradient Boosting model will be referred to as "GB", the Random Forest model as "RF", and the Support Vector Machine model as "SVM". For each of them, the “Descr” label indicates the model was trained including the verbal description for each firm, while the “SIC” label indicates that the dataset employed included just the primary SIC codes.

4.1 Test set size: 20%



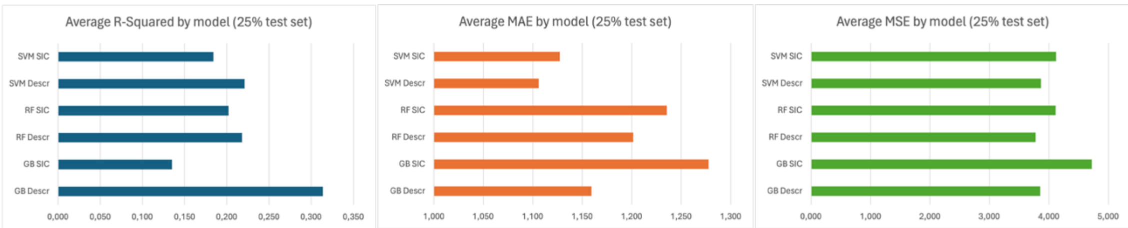
Looking at the models trained with 80% of the whole dataset, the Gradient Boosting model including descriptions (GB Descr) reported an average MAE of 1.168 and an average MSE of 3.834, topping an R-squared of 0.222. The same model, trained with primary SIC codes instead of descriptions (GB SIC), showed a worse performance with a MAE of 1.285, a MSE of 4.480, and an R-squared of 0.116. Although these differences seem consistent in absolute terms, the t-test to prove their statistical significance returned a p-value of 0.0796, thus rejecting the thesis that the description-based model performed significantly better than the SIC-based one, at least at a 5% significance level, which is the standard threshold we use in this type of analysis.

The best-performing algorithm with this test set size, in terms of fitting power via R-squared, was the Random Forest trained with descriptions (RF Descr), reporting an average MAE of 1.155, MSE of 3.471, and R-squared equal to 0.278. Also with this algorithm, the SIC-based model showed worse results, with a higher MAE of 1.233, a MSE of 4.180, and R-squared of 0.114. In this case, the difference in fitting performance between the models is statistically significant at a 5% level, since the t-test for the R-squared values returned a p-value of 0.0323.

On the Support Vector Machine side, the "Descr" model achieved an average MAE of 1.121, MSE of 4.281, and an R-squared equal to 0.128. If we look at the model trained with SIC codes, it reported a MAE of 1.177, MSE of 4.492, and an R-squared of 0.134.

From these results, we can hypothesize that the differences in performance are due to randomness and cannot be deemed to be a direct consequence of using encoded descriptions instead of SIC codes. Supporting this thesis, the t-test for the R-squares returned a p-value of 0.9418, indicating no statistically significant difference at all between the models' fitting power.

4.2 Test set size: 25%

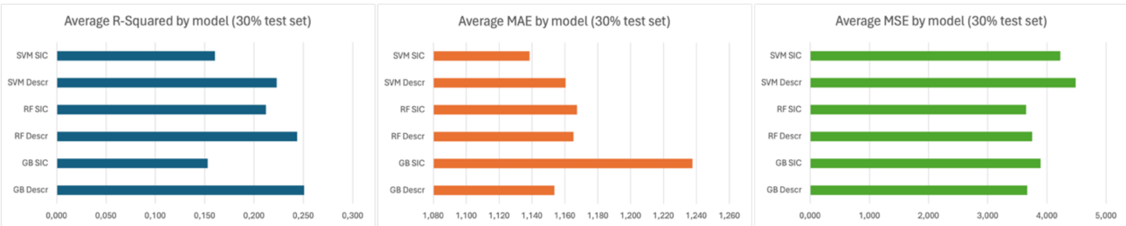


Looking at the models trained with 75% of the dataset, the Gradient Boosting model with descriptions (GB Descr) reported an average MAE of 1.159 and an average MSE of 3.853, topping an R-squared of 0.314. This model, trained with primary SIC codes instead of descriptions (GB SIC), showed a worse performance with MAE of 1.278, MSE of 4.722, and an R-squared of 0.135. The consistency of the difference in fitting power is proved by the t-test conducted, which returned a p-value of 1.37e-5: this rejects the hypothesis that the difference in performance can be addressed to randomness even at a 0.0014% significance level, proving with confidence that encoded business descriptions improve drastically the ability to describe the EV/Sales multiple compared to traditional methods and beyond financial metrics.

For this test set size category, the Random Forest algorithm returned the worst results overall: the model leveraging business descriptions (RF Descr) reported an average MAE of 1.202, MSE of 3.775, and R-squared equal to 0.218. The SIC-based model showed slightly worse results, with MAE of 1.235, MSE of 4.114, and R-squared of 0.202. Given how close such results were, the t-test was crucial to assess that the difference in fitting performance was not statistically significant but likely addressable to randomness, as confirmed by the p-value equal to 0.7149.

On the Support Vector Machine side, the "Descr" model reported an average MAE of 1.106, MSE of 3.869, and an R-squared equal to 0.221. The model trained with SIC codes had a MAE of 1.127, MSE of 4.121, and an R-squared of 0.184. These results were even better than the gradient boosting ones in terms of MAE, but still hinted that the differences in performance could be a simple consequence of randomness; the t-test for the R-squares confirmed this with a p-value of 0.3758, indicating no statistically significant difference between the models' fitting power.

4.3 Test set size: 30%



Among the models trained with 70% of the dataset, Gradient Boosting with descriptions (GB Descr) returned an average MAE of 1.154, an average MSE of 3.665, and an R-squared value of 0.251. The same model with primary SIC codes instead of descriptions (GB SIC) reported MAE of 1.238, MSE of 3.893, and a drastically lower R-squared of 0.153. Since the t-test for the significance of the difference between R-squares returned a p-value of 0.0334, I can state that, once again, embedded firms' descriptions manage to describe the data significantly better than SIC codes via Gradient Boosting.

The second overall best algorithm with a 30% test set size was the Random Forest trained with descriptions (RF Descr), reporting an average MAE of 1.165, MSE of 3.750, and R-squared equal to 0.244. Looking at the SIC-based model's results, we notice MAE of 1.167, MSE of 3.645, and an R-squared equal to 0.212. As we can easily imagine from these figures, the difference in performance might just be a consequence of randomness; in fact, the t-test conducted on the difference between R-squares returned a p-value of 0.5221.

In the context of Support Vector Machines, the description-based model reached an average MAE of 1.160, MSE of 4.486, and an R-squared of 0.223. The model employing SIC codes reported a MAE of 1.138, MSE of 4.224, and an R-squared of 0.160. It's interesting to notice how, in this particular case, substituting descriptions with SIC codes leads to a lower average R-squared on one side, but also to lower average MAE and MSE. However, such differences in performance are just due to casualty, as proved by the t-test for the R-squares returning a p-value of 0.1844.

4.4 Results Summary

Across all three test set sizes, the Gradient Boosting models trained with business descriptions consistently outperformed those trained with SIC codes. Although this conclusion was statistically significant at a 5% level for the models with 25% and 30% test set sizes, also in the case of 20% test size, the model showed considerably lower average Mean Absolute Error (MAE), lower Mean Squared Error (MSE) values, and a higher R-square metric when embedded descriptions were included. From the results yielded by this algorithm, I can conclude with confidence that the inclusion of detailed business descriptions enhances the model's ability to explain the variability in the target variable.

Regarding the other architectures employed, Random Forest models reported less consistent improvements in performance, with a peak explanatory power and significant difference between description and SIC based models in the case of a 20% test set size. The results for Support Vector Machine proved that this algorithm might not be a good choice to describe the variability of a target like EV/Sales with a financially based dataset, and its performance does not get any better by providing a richer qualitative feature set.

As a final overall consideration, beyond the choice between traditional SIC codes and embedded descriptions, we can see how the fitting performance diverges across different algorithms and test set sizes: this highlights how important is to select appropriate models and hyperparameters based on the specific research context and dataset employed.

5. Conclusion

This thesis aims to outline a quantitative, rigorous, and standardized method for conducting one of the fundamental and most common practices in the process of company valuation. The process of selecting comparable companies for the target, which precedes the computation of valuation multiples, still relies heavily on subjective methods and is overly exposed to human variables. The objective of this thesis is to suggest an approach that makes this process more objective. The computation of the "optimal" multiple using the proposed machine learning models presupposes the identification of companies that are most "similar" to the target in a multidimensional space described by financial and, in this case, operational variables, thanks to a detailed encoding of how the company positions itself within its industry through business descriptions.

Although this thesis demonstrates that including business descriptions significantly enhances the descriptive power of the model, it is not trivial to determine an appropriate method to evaluate this descriptive capability in the context of such target variables, given the highly heterogeneous financial and accounting processes subject to human discretion that generate them. However, I tried to outline a basic version for a standardized procedure in business valuation because I believe that, If properly developed, shared, and applied in the industry, it could significantly reduce market asymmetries.

To conclude, I acknowledge that several limitations of this model could be removed by carrying out a more extensive feature research and selection process, testing a wider range of algorithms, and employing a larger dataset, maybe trying even to extend the study to multiple industries simultaneously to capture cross-sectoral firms' characteristics. As an interesting proposal to evolve this framework, I would suggest including sentiment data for each company at valuation time, with the aim to incorporate in the output price also market opinions, following a strictly quantitative and generalizable approach.

6. Bibliography

1. **M&A Research Institute Inc.** "M&A Research Institute Inc." (2024).
2. **Bain & Company's Global M&A and Divestitures.** Global M&A Report 2024. (2024).
3. **Bouchlaghem, Younes, Yassine Akhiat, and Souad Amjad.** "Feature Selection: A Review and Comparative Study." E3S Web of Conferences 351, (2022): 1046.
4. **Devlin, Jacob, Chang Ming-Wei, Kenton Lee, and Kristina Toutanova.** "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." arXiv.Org (2019).
5. **Gao, Zihan.** A Methodological Study on Selection of Comparable Companies for Mergers and Acquisitions of Listed Companies Based on KNN Algorithm. (2023).
6. **Garzon, Max, Ching-Chi Yang, Deepak Venugopal, Nirman Kumar, Kalidas Jana, and Lih-Yuan Deng.** Dimensionality Reduction in Data Science. 1st ed. Cham: Springer International Publishing AG. (2022).
7. **Hastie, Trevor, Robert Tibshirani, and Jerome Friedman.** Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer. (2009).
8. **Kaunisto, Leevi.** "Estimating Private Company Financial Figures in Early-Phase M&A: A Machine Learning Approach." Industrial Engineering and Management. (2023).

9. **Lajoux, Alexandra Reed.** The Art of M & A : A Merger, Acquisition, and Buyout Guide / by Alexandra Reed Lajoux ; with Capital Expert Services, LLC. New York: McGraw-Hill. (2019).
10. **Lee, Yoojung.** "This 32-Year-Old Became a Billionaire by using AI to Broker Mergers, Financial Review." (2023).
11. **McKinsey & Company.** Top M&A Trends in 2024: Blueprint for Success in the Next Wave of Deals. (2024).
12. **Murphy, Kevin P.** Machine Learning : A Probabilistic Perspective / Kevin P. Murphy. Cambridge, Massachusetts: The MIT Press. (2012).
13. **Nobre, João and Rui Ferreira Neves.** "Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to Trade in the Financial Markets." Expert Systems with Applications 125, (2019): 181-194.
14. **Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.** "Attention is all You Need." arXiv.Org (2017).
15. **Zhong, Xiao and David Enke.** "Forecasting Daily Stock Market Return using Dimensionality Reduction." Expert Systems with Applications 67, (2017): 126-139.