Annals of Operations Research (2025) 353:667–692
https://doi.org/10.1007/s10479-025-06574-z

ORIGINAL RESEARCH

Business Failure Prediction From Textual and Tabular Data With Sentence-Level Interpretations

Henri Arno[1] · **Klaas Mulier**[2] · **Joke Baeck**[2] · **Thomas Demeester**[1]

Received: 1 March 2024 / Accepted: 7 March 2025 / Published online: 18 April 2025
© The Author(s) 2025

Abstract
Business failure prediction models are crucial in high-stakes domains like banking, insurance, and investing. In this paper, we propose an interpretable model that combines numerical and sentence-level textual features through a well-known attention mechanism. Our model demonstrates competitive performance across various metrics, and the attention weights help identify sentences intuitively linked to business failure, offering a form of interpretability. Furthermore, our findings highlight the strength of traditional financial ratios for business failure prediction while textual data—particularly when represented as keywords—is mainly useful to correctly classify corporate disclosures where the possibility of failure is explicitly mentioned.

Keywords Decision support systems · Business failure prediction · Natural language processing · Text analytics

1 Introduction

Business failure prediction models can be used to evaluate the financial health of companies, serving as a valuable tool for data-driven decision support (Borchert et al., 2023). Academic interest in the development of such models dates back to at least the 1960s (Beaver, 1966; Altman, 1968), and by now, business failure prediction models have applications in a wide array of economic sectors. For instance, they are of great importance for loan officers in their corporate lending decisions, for insurance brokers in the pricing of their contracts, for investment analysts to determine the risk of investing in a certain stock or bond, for credit bureaus in advising their clients on the creditworthiness of other businesses, or for banking regulators in their assessment of credit risks in the banking system. Having a highly performing business failure prediction model is thus of great economic importance.

Research gap. Over the years, corporate disclosures have become increasingly more complex and lengthy, mainly driven by stricter disclosure requirements (Dyer et al., 2017). Moreover, corporate disclosures do not only contain numerical accounting data but also substantial amounts of text. At the same time, with artificial intelligence (AI), machine learning

✉ Henri Arno
Henri.Arno@UGent.be

1 Ghent University - imec, Technologiepark 126, 9052 Ghent, Belgium

2 Ghent University, Universiteitstraat 4, 9000 Ghent, Belgium

(ML) and large language models (LLMs) developing fast, the possibilities of incorporating multi-modal data to predict business failure look promising. However, superior prediction capabilities alone do not suffice to ensure that business failure prediction models using ML techniques will find their way to practice. User trust is critical for the way that people make sense of and use the output of AI systems (Glikson & Woolley, 2020). Recent insights show that explainability and interpretability are strong determinants of such user trust (Shin, 2021; Bauer et al., 2023). Therefore, we aim to fill this gap by proposing a new multi-modal neural architecture that is interpretable by design and we compare the performance of our model with various classical and state-of-the-art classifiers.[1]

Research focus. The empirical setting in which we train and evaluate our models combines three widely used data sources containing information on listed companies operating in the United States: Edgar, Compustat, and the LoPucki Bankruptcy Research Database. The first data source provides us with the text from 10K's that companies need to file to the Securities and Exchange Commission each year. From these, we take the management discussion and analysis section (MD&A), which contains textual information provided by top-management about the state of the company in a given fiscal year. The text in the MD&A's is relatively long, with 6,810 words on average. The second data source provides us with numerical accounting data from which we derive a number of financial ratios representing the financial state of the company in a given fiscal year. The third data source provides us with information on whether and when a company filed for bankruptcy (either under chapter 7 or 11 of the U.S. bankruptcy code). Our models predict if a company will file for bankruptcy during the next year, based on the information contained in a 10K filing. In total, the combined dataset contains textual and numerical data relating to 84,652 observations, i.e. 10K filings, of which 662 were filed before bankruptcy.

Results. The model we propose combines numerical features and sentence-level textual features through an established attention mechanism, prevalent in transformer models (Vaswani et al., 2017), and attains competitive results on a variety of performance metrics (including the `ROC-AUC`). Notably, the sentences that received a particularly high attention weight from disclosures that were assigned a high likelihood of business failure by our prediction model contain information users may find highly relevant. In these sentences, for example, top-management refers to declining demand for the companies' products, the closing of stores, the breaching of loan covenants, debt renegotiations with creditors, etc. As a result, the model not only performs well in predicting business failure but also provides interpretability as an additional functionality by highlighting the specific parts of the text most indicative of such potential failure.

Our study further examines the impact of the different data modalities in corporate disclosures on the performance of business failure prediction models. Our results underscore the importance of numerical data, in particular traditional financial ratios (liquidity, solvency and profitability ratios) as these features are the main drivers behind the best-performing models (in terms of `ROC-AUC`). In contrast, the textual data—when represented as keywords—enables the models to correctly classify a small subset of disclosures where the management explicitly mentions in the MD&A that the company considers to file for bankruptcy. For these particular instances, the prediction task based on keywords such as *'chapter'* or *'reorganisation'* becomes rather trivial. Additionally, we evaluated classifiers based on state-of-the-art pretrained document embeddings, but they perform poorly on the task. This suggests that these embeddings capture little information relevant to predict business failure and that tailored textual features are better suited for this purpose.

[1] The code to replicate the experiments is available at https://github.com/henriarnoUG/ECL

Contributions. We contribute to the literature in two ways. First, we contribute to the growing body of research on eXplainable Artificial Intelligence in Operations Research (XAIOR) (see, e.g. De Bock et al., 2024; Sobrie et al., 2024; Coussement and Benoit, 2021) by proposing an interpretable neural architecture for business failure prediction based on numerical and sentence-level textual features. According to Rudin (2019), predictive models that are interpretable by design might be preferred over post-hoc methods (such as LIME or SHAP) that aim to explain the predictions of black-box models, especially for high-stakes applications. As our proposed model is interpretable by design, we strengthen this part of the literature.

Second, we contribute to the strand of Operations Research (OR) literature focused on business failure prediction (see, e.g. Mai et al., 2019; De Bock, 2017; Du Jardin, 2021; Borchert et al., 2023). We thoroughly evaluate a multitude of uni-modal and multi-modal business failure prediction models and highlight the distinct role of numerical and textual data in corporate disclosures in the context of this challenging task.

In summary, our study presents a new attention-based model for business failure prediction that (1) is competitive on a variety of performance metrics (including `ROC-AUC`) and (2) offers a form of interpretability since it can highlight the sentences that are indicative of corporate distress. In addition, our findings highlight the importance of traditional financial ratios in predicting business failure and show that keyword-based textual features are mainly useful to detect a subset of corporate disclosures close to failure, where this is explicitly mentioned in the text.

The remainder of the paper is structured as follows. In Sect. 2, we review the literature on business failure prediction and more generally, the literature on predictive modelling techniques related to our methodological contribution. Section 3 details the data used in our study. In Sect. 4, we elaborate on the design and training of our proposed model, and in Sects. 5 and 6, we cover the results and draw conclusions.

2 Literature review

2.1 Business failure prediction

2.1.1 Advancements in predictive modelling techniques

Business failure prediction was pioneered by Beaver (1966), who demonstrated that financial ratios can effectively be used to discriminate between failing and non-failing companies through univariate analysis. The purpose of this foundational work was to empirically verify the usefulness of accounting data, *"...for any purpose, and not merely for solvency determination"*. Building on this, Altman (1968) introduced the Z-score, a multivariate predictive model combining five financial ratios, which became a widely adopted tool for failure prediction in practice. Later, Ohlson (1980) proposed the O-score, a probabilistic model that also relied on financial ratios, a third key contribution in the field.

Following this work, the field has evolved in two directions. First, researchers have tried to improve predictive performance by relying on more flexible modelling techniques. Examples include the application of feed-forward neural networks (Odom and Sharda, 1990), ensembles (Cortés et al., 2007) or convolutional neural networks (Hosaka, 2019) to predict business failure. Second, studies have explored the predictive value of variables beyond financial ratios, such as corporate governance variables (Liang et al., 2016), industry and country

characteristics (Doumpos et al., 2017), and more recently, textual features extracted from corporate disclosures, which will be discussed further below.

2.1.2 The role of ensembles in failure prediction

The main idea behind ensembles is to combine multiple learning algorithms into a single model that outperforms each individual component. Numerous ensemble methods have been proposed for a variety of tasks and we refer to Sagi and Rokach (2018) for an overview. These techniques have been extensively studied in the context of business failure prediction due to their superior performance compared to single classifiers.

Cortés et al. (2007) were among the first to use ensembles for failure prediction, and in their experiments, AdaBoost (adaptive boosting) (Freund and Schapire, 1996) showed significant improvements over the decision tree baselines. Similarly, Alfaro et al. (2008) benchmarked AdaBoost against feed-forward neural networks and also report improved performance. Kim and Kang (2010) compared AdaBoost to bagging ensembles (bootstrapped aggregation) (Breiman, 1996) and found that bagging yields the best results. More recently, García et al. (2019) studied the predictive performance of various traditional ensemble methods—including AdaBoost, bagging, random forest, and rotation forest—and conclude that the optimal ensemble method depends on the properties of the dataset.

While these contributions highlight the strength of traditional ensembles for business failure prediction, Du Jardin (2021) argues that the performance improvements over single classifiers are often limited. Therefore, he proposed an ensemble technique tailored to the task, based on self-organising neural networks, capable of modelling both general and specific financial patterns leading to failure, and showing improved results over various baselines. In addition, De Bock (2017) introduced *spline rule ensemble*—another non-traditional ensembles method for business failure prediction—that effectively balances model strength and interpretability.

2.1.3 The predictive value of textual disclosures

One of the first studies examining the relationship between business failure and the textual information in corporate disclosures was performed by Holder-Webb and Cohen (2007). They find evidence that managers increase the quality of the management discussion and analysis section (MD&A) of 10K filings in the early stages of financial distress, primarily for economic reasons (e.g. to reduce the cost of capital). While this study established a link between textual disclosures and financial distress, subsequent work has directly incorporated features from these texts into business failure prediction models. Initially, these features were extracted with dictionary-based techniques, which produced promising results and established the predictive value of disclosure texts for failure prediction (Cecchini et al., 2010; Mayew et al., 2015). Recently, deep learning techniques have been used to represent the text as well, leading to improved predictive performance, especially when combined with financial ratios (Mai et al., 2019; Arno et al., 2022).

2.2 Multi-modal and explainable predictive modelling

2.2.1 Classification from textual and tabular data

Our objective is the build an interpretable business failure prediction model capable of processing the numerical and textual data from corporate disclosures. Recently, Sleeman et al. (2022) introduced a taxonomy for multi-modal classification systems, highlighting three ways to combine the data modalities: before, during, or after the primary learning stage, referred to as early, cross-modality, and late fusion, respectively. Most classification models using numerical and textual data rely on early or late fusion. For instance, Panda et al. (2020) developed a model for emotion detection that combines electrical brain signals (numerical data) and customer reviews (textual data) through early fusion. Manually extracted features from both modalities were concatenated into a single representation to train the classifier. In contrast, Xu et al. (2019) used late fusion for diseases classification based on multi-modal electronic health records. They trained separate models for the textual and numerical data and combine the outputs in their final prediction. Drawing from this line of work, we will include both early and late fusion classifiers as baselines to evaluate the performance of our proposed model.

2.2.2 Explainable artificial intelligence in operations research

Numerous prediction tasks, including business failure prediction, have seen tremendous performance improvements with the advent of modern deep learning. Nonetheless, deep neural networks typically lack the ability to explain their inference processes or to provide interpretable insights into their predictions (Xu et al., 2019). This is especially problematic for predictive models deployed in high-stakes applications (Rudin, 2019). Consequently, eXplainable Artificial Intelligence (XAI) has emerged as a prominent research area within the AI community.

XAI is also gaining considerable attention in operations research. De Bock et al. (2024) recently introduced a normative framework for eXplainable Artificial Intelligence for Operations Research (XAIOR), emphasizing the need for solutions that are simultaneously *performant, attributable* and *responsible* to enhance decision-making. A notable example of XAIOR is the *spline rule ensemble* method, which enables flexible modelling while ensuring interpretability through human-understandable rules (De Bock, 2017; De Bock & De Caigny, 2021). Additionally, researchers have used model-agnostic methods like LIME and SHAP to explain their model predictions (Sobrie et al., 2024; Chen et al., 2024). Our work contributes to this emerging field by proposing a business failure prediction model that is interpretable by design.

3 Data and prediction task

In Sect. 3.1 we cover the dataset used in our experiments. Section 3.2 expands on the business failure prediction task, including our labelling strategy.

3.1 The Edgar-Compustat-LoPucki dataset

We aim to predict business failure from the numerical and textual data contained in *corporate disclosures*. To this end, we focus on 10K filings, which are extensive documents prepared annually by large—mostly listed—companies operating in the United States. These filings are submitted to the Securities and Exchange Commission (SEC) and contain detailed information about the performance of the company in the past fiscal year. It is a primary source of information for investors, analysts and other stakeholders. A 10K is organised in 15 different sections (denoted *items*), each focusing on a particular topic such as a description of the business (item 1), pending legal proceedings (item 3), the management discussion and analysis (item 7) and the consolidated financial statements (item 8). Most of this information is textual while the included financial statements (e.g., the balance sheet) are mainly numerical.

In recent work, Arno et al. (2023) presented the Edgar-Compustat-LoPucki (ECL) dataset, which combines three established data sources.[2] The dataset contains the textual and numerical data from 170,139 10K records, filed between 1993 and 2023. The text in these 10K filings is relatively long, averaging 29,247 words per filing, with the management discussion and analysis (item 7) being the longest individual item, averaging 6,810 words per filing. Due to a size-related selection criterion in the LoPucki Bankruptcy Research Database, only a subset of the 10K filings in the ECL dataset have associated bankruptcy labels, which is further discussed in Sect. 3.2 below.

3.2 The business failure prediction task

The studies discussed in Sect. 2.1 have operationalized *business failure* in various ways. For instance, Beaver (1966) classified a company as failing if it experienced bankruptcy, a bond default, an overdrawn bank account or nonpayment of a preferred stock dividend. Similarly, Cortés et al. (2007) also adopted a broad definition of failure by classifying companies in the failing group not only in the case of bankruptcy or temporary receivership, but also when a company was acquired or dissolved. In this paper, we follow the majority of previous studies and use a more narrow definition of business failure. Specifically, the failing companies in our sample are those that file for bankruptcy under chapter 7 (liquidation) or chapter 11 (reorganisation) of the United States bankruptcy code.

We only have bankruptcy information for a subset of the 10K filings in the dataset. Only companies with a total asset value exceeding 100,000,000 (measured in 1980 dollars) are considered eligible for inclusion in the LoPucki Bankruptcy Research Database. For eligible companies, we know if they filed for bankruptcy or not, and we know the associated bankruptcy filing date (if applicable). Based on this information, binary bankruptcy labels were assigned to the 10K filings of eligible companies following the strategy depicted in Fig. 1.

A 10K filing covers a fiscal year (T_{FY}) that ends on the fiscal year-end (t_{FY}), and is submitted to the SEC on the filing date (t_{FD}). While the accounting information in a 10K strictly relates to the period T_{FY}, the text could also contain information from the entire pre-filing period T_{PF}. If the company filed for bankruptcy in the year following the filing date (during $T_{Prediction}$), the binary bankruptcy label is `true`, and `false` otherwise. This strategy resulted in a labelled subset of 84,652 10K filings in the dataset, that can be used to predict if a company filed for bankruptcy in the next year (during $T_{Prediction}$), given the numerical and textual data in its 10K filing.

[2] See (Arno et al., 2023) for details on the construction of the dataset and descriptive statistics.

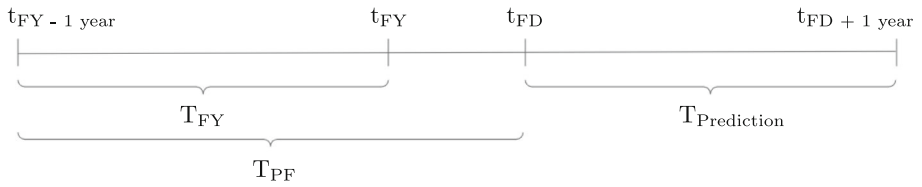


Fig. 1 A timeline illustrating the fiscal year coverage (T_{FY}) and the filing date (t_{FD}) of the 10K filings in the dataset. Based on this information, we assign binary labels to the 10K filings for the business failure prediction task

Since bankruptcy of large listed companies does not occur often, only a small number of 10K filings in the labelled dataset were filed in the year before bankruptcy (the *'positive'* examples), which leads to a strong class imbalance. In fact, only 662 out of the 84,652 labelled 10K filings were filed before bankruptcy. This means that we approximately have 1 positive example for every 127 negative examples in the dataset.

We performed a temporal train, validation and test split that closely reflects how business failure prediction models are constructed in practice. All 10K records filed prior to 2012 are used to train the models, the 10K records filed between 2012 and 2015 are assigned to the validation set and the remaining 10K records, filed after 2015, comprise the test set. The resulting dataset consists of 54,039 instances in the training set, 12,324 in the validation set, and 18,289 in the test set, with 481, 59 and 122 positive examples in each set, respectively.

In line with the work of Mai et al. (2019), we include a selection of financial ratios (reported in item 8 of a 10K filing) as numerical features in our experiments (see Appendix A for an overview). For the text, we follow the majority of previous studies (Cecchini et al., 2010; Mayew et al., 2015; Mai et al., 2019; Arno et al., 2023) and restrict our focus to the management discussion and analysis section (item 7). This section contains a description of the most important events of the past fiscal year as perceived by top-management (such as market trends, strategic decisions and operational matters) and is therefore well suited for business failure prediction.

4 Model design and experimental setup

In Sect. 4.1, we describe the architecture of our proposed model, while the uni-modal and multi-modal baselines are covered in Sects. 4.2 and 4.3, respectively. The experimental setup and evaluation metrics are detailed in Sect. 4.4. For clarity, all models are assigned a name that will be used throughout the work. A summary of these models and their defining characteristics is provided in Table 1.

4.1 Proposed model architecture

We propose a new interpretable model for business failure prediction, denoted `NT-att[num,sent-emb]`, that combines numerical features and sentence-level textual features based on a well-known attention mechanism. To ensure reproducibility, we give a detailed model description in the following paragraphs. A visual overview of the proposed architecture is given in Fig. 2.

Table 1 An overview of the models presented in this work. Numerical data is represented as normalised financial ratios (`num`), while textual data is represented as keyword-based `tfidf` features (`tfidf`), document-level embeddings (`emb`) or sentence-level embeddings (`sent-emb`). Prefixes `N` and `T` refer to uni-modal models on numerical and textual inputs, respectively, while `NT` denotes multi-modal models using both numerical and textual inputs

Model	Description
`NT-att[num,sent-emb]`	Attention-based model on the numerical features and sentence embeddings
`N-Z'[num]`	Altman Z'-score obtained from the numerical features
`N-LR[num]`	Logistic regression classifier on the numerical features
`N-MLP[num]`	Multi-layer perceptron classifier on the numerical features
`N-XGB[num]`	XGBoost classifier on the numerical features
`T-LR[tfidf]`	Logistic regression classifier on all `tfidf` features
`T-XGB[tfidf]`	XGBoost classifier on selected `tfidf` features
`T-LR[emb]`	Logistic regression classifier on the document-level embeddings
`NT-XGB[num+tfidf]`	XGBoost classifier on the combined numerical and selected `tfidf` features
`NT-stack[XGB(num),XGB(tfidf)]`	Stacked combination of the `N-XGB(num)` and `T-XGB(tfidf)` models

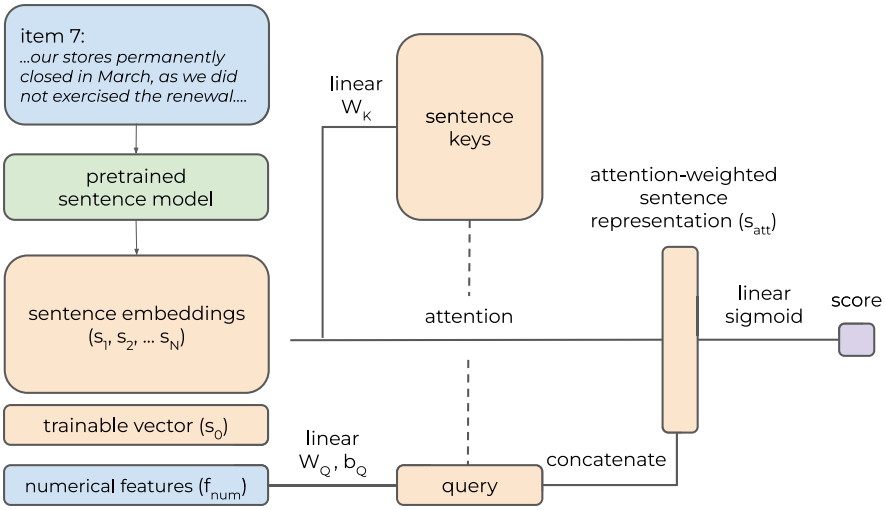


Fig. 2 The architecture of the proposed model `NT-att[num,sent-emb]`, which predicts business failure by combining numerical features and sentence-level textual features. The textual representation $\mathbf{s}_{\text{att}}$, is an attention-weighted sum of the individual sentence embeddings $\mathbf{s}_i$ $(i = 1, \ldots, N)$, reflecting their compatibility with the numerical feature vector $\mathbf{f}_{\text{num}}$ or the prediction target (the business failure label)

4.1.1 Sentence embeddings

By only focusing on the management discussion and analysis (MD&A) of the 10K filings, we already drastically reduce the amount of text, filtering out parts that are irrelevant for business failure prediction. Nonetheless, we assume that the most useful information for the task at hand is likely concentrated in only a few highly informative sentences. As a first step, we therefore identify individual sentences in the text, and separately encode them into dense D-dimensional vector representations with a pretrained sentence embedding model.[3]

Since each MD&A has a different number of sentences, we initially encode the first 300 sentences of each text, assuming that the most relevant content for business failure prediction can be found at the beginning. This strategy allows us to fully process 52.7% of the MD&A's in the dataset, as they contain fewer than 300 sentences. Alternatively, we encode the first and last 250 sentences, based on the assumption that the conclusions at the end might also contain useful information for the task. With this strategy, that encodes up to 500 sentences, 84.3% of the MD&A's are fully processed. If an MD&A contains fewer than 300 or 500 sentences, we pad the sentence embeddings accordingly.

4.1.2 Attention-based cross-modality fusion

Given the sentence representations $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$ for all N sentences of a given text, the idea is to let the model discover which of those are most compatible with the numerical feature vector $\mathbf{f}_{\text{num}}$ (and therefore potentially interesting from an interpretability perspective), or which ones are simply more strongly related to the prediction target than others and could therefore bring complementary value to the numerical features alone. We first construct a so-called *attention-weighted* representation $\mathbf{s}_{\text{att}}$ for the entire text as a weighted sum of all sentence embedding vectors as follows:

$$\mathbf{s}_{\text{att}} = \sum_{i=1}^{N} \alpha_i \, \mathbf{s}_i, \quad \text{with} \quad \alpha_i = \frac{e^{\frac{1}{\sqrt{d}}(W_Q \mathbf{f}_{\text{num}} + \mathbf{b}_Q)\cdot(W_K \mathbf{s}_i)}}{\sum_{j=1}^{N} e^{\frac{1}{\sqrt{d}}(W_Q \mathbf{f}_{\text{num}} + \mathbf{b}_Q)\cdot(W_K \mathbf{s}_j)}}$$

in which the attention weight α_i of sentence embedding $\mathbf{s}_i$ is obtained by measuring how compatible it is with the numerical feature vector $\mathbf{f}_{\text{num}}$ (compared to the other sentences), through the standard scaled dot product attention. In particular, $\mathbf{f}_{\text{num}} \in \mathbb{R}^F$ is mapped into d-dimensional space through an affine transformation with trainable weights $W_Q \in \mathbb{R}^{d \times F}$ and bias vector $\mathbf{b}_Q \in \mathbb{R}^d$, where it functions as the attention 'query' to detect compatible sentences. Compatibility is measured by the scaled dot product—as is common practice in transformer models (Vaswani et al., 2017)—with the attention 'keys' obtained through a linear projection of the sentence embeddings $\mathbf{s}_i \in \mathbb{R}^D$ into d-dimensional space, with weights $W_K \in \mathbb{R}^{d \times D}$.

Note that the bias term $\mathbf{b}_Q$ allows the model to assign high attention weights to particular sentences, even if there is no substantial contribution from the term involving the numerical features. This expresses our intuition that some sentences may contain clear information on business failure or financial health, even when not reflected in the available financial ratios.

Furthermore, we extended the list of sentence embeddings with an additional normalised but trainable vector $\mathbf{s}_0 \in \mathbb{R}^D$, which lends the model an interesting property in terms of interpretation. If its corresponding attention weight α_0 is higher than the weights of all of

[3] We used the `all-MiniLM-L6-v2` model from HuggingFace with an embedding dimension of $D = 384$. For more details, see https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

the actual sentence embeddings, this can be interpreted as the model explicitly ignoring the sentence embeddings, only taking into account the numerical features in making a prediction. If, however, one or a few sentences contain clear information related to the financial ratios and / or the outcome, their corresponding attention weights will likely become much larger than the others (including the attention weight for $\mathbf{s}_0$). Note that the sentences that lead to high attention weights are considered by the model to be informative for the predicted *outcome*, either hinting towards business failure, or instead, a very healthy financial situation. This is illustrated in Tables 6 and 9, and will be discussed further in Sect. 5.2.1.

We subsequently concatenate $\mathbf{s}_{\text{att}}$, the attention-weighted sum of sentence representations, with the representation $(W_Q \mathbf{f}_{\text{num}} + \mathbf{b}_Q)$ of the numerical features in d-dimensional space (i.e., the query vector in the attention step), and finally add a logistic regression layer (linear projection to a scalar, followed by a sigmoid activation) to predict business failure. The training details of the model are given in Appendix B.

4.2 Uni-modal baselines

In this section, we cover the baseline models that predict business failure from either numerical or textual inputs alone. These models helps us understand the predictive value of each data modality in the corporate disclosures and their performance serves as a lower bound for the performance of the multi-modal models.

4.2.1 Numerical baselines

The Z-score model (Altman, 1968)—introduced in Sect. 2.1—remains a frequently used business failure prediction model today. However, the original model requires both accounting and market data as one of the included financial ratios is based on the market value of equity. In a revised version of the model, the Z'-score, the market value was replaced by the book value of equity (Altman, 2013). This version of the Z-score model can be computed from the accounting data in 10K filings, which is why `N-Z'[num]` is the first baseline in our experiments. For a more detailed formulation of this baseline model, we refer to Appendix C.

From the numerical features (see Appendix A for an overview), we have trained three additional baseline classifiers. An L2-regularised logistic regression classifier (`N-LR[num]`), a multi-layer perceptron classifier (`N-MLP[num]`) and as ensemble method an XGBoost classifier (`N-XGB[num]`). The training details for these models are given in Appendix B.

4.2.2 Textual baselines

For the textual baselines, we used two techniques to extract meaningful features from the text. First, we used *term frequency-inverse document frequency* (`tfidf`) features. This approach starts by constructing a vocabulary that consists of all occurring unigrams and/or bigrams in the training set. A text is then represented as a vocabulary-sized vector where each position corresponds to a term in the vocabulary. The value at each position quantifies how often the term occurs in the text while also accounting for the rarity of this term over all training documents (Manning et al., 2008). In order to manage the size of the vocabulary, we performed several standard text cleaning operations including removal of case, stopwords, punctuation, numerals and inflected word forms. With these keyword-based features, we have trained two baseline models. An L1-regularised logistic regression classifier (`T-LR[tfidf]`) and an XGBoost classifier (`T-XGB[tfidf]`). The L1 regularisation serves as an automatic feature

selection method, retaining approximately 1,200 features after hyperparameter optimisation. Based on these findings, we selected 1,500 `tfidf` features using a chi-squared test, serving as input for the XGBoost classifier.

Second, we represented the texts as dense vectors obtained from a pretrained and commercially available text embedding model from OpenAI.[4] This transformer-based neural sequence encoder is capable of encoding documents up to 8,191 tokens into a single vector, capturing the semantic information of the texts (longer documents were truncated to fit within the token limit). Using these embeddings, we have trained an L2-regularised logistic regression classifier (`T-LR[emb]`) as our third textual baseline. For training details we refer to Appendix B.

4.3 Multi-modal baselines

In this section, we cover the baselines that predict business failure from both numerical and textual inputs. Following the taxonomy of Sleeman et al. (2022), we distinguish between early and late fusion models, depending on whether the numerical and textual data is combined before or after the primary learning stage. Both models are based on XGBoost classifiers, given their strong performance on numerical and textual inputs individually (as discussed in Sect. 5.1).

For the early fusion model, we trained an XGBoost classifier (`NT-XGB[num+tfidf]`) on the concatenated numerical features and selected `tfidf` features (based on a chi-squared test). For the late fusion model, we follow a stacking approach (Wolpert, 1992). The numerical and textual data are processed separately by their respective best-performing uni-modal classifiers (`N-XGB[num]` and `T-XGB[tfidf]`). The normalised predictions from these classifiers on the validation set are then used to train a stacking classifier:

$$P(\textit{Business failure}) = \sigma(\beta_0 + \beta_1 S_{numerical} + \beta_2 S_{textual})$$

where *P(Business failure)* is the likelihood of business failure assigned by the stacking classifier, $\sigma(.)$ denotes the sigmoid transformation, $S_{numerical}$ and $S_{textual}$ are the normalised scores of the uni-modal classifiers, and the β_i's are the parameters of the model (denoted as `NT-stack[XGB(num),XGB(tfidf)]`). Training details are given in Appendix B.

4.4 Performance evaluation

The models discussed above output continuous scores, where a higher score indicates that business failure is more likely. We could transform these scores into class labels (failure vs. no failure) by using a 'cut-off' set to optimise some performance metric. Nonetheless, the optimal cut-off will depend on the application of the model. Therefore, we decided to evaluate the models in terms of their ranking performance through a variety of cut-off independent performance metrics presented below.

4.4.1 Area under the receiver operating curve

The area under the receiver operating curve (`ROC-AUC`) is a metric commonly used to evaluate the overall performance of binary classification models. It summarises the receiver

[4] We used the `text-embedding-ada-002` model from OpenAI. For more details, see https://platform.openai.com/docs/models/embeddings.

operating curve, plotting the true positive rate (or recall) against the false positive rate at each possible cut-off. In our context, this metric can be interpreted as the probability that a classifier assigns a higher score to a randomly sampled 10K filed in the year before failure (a positive) compared to a randomly sampled 10K that was not filed before failure (a negative) (Fernández et al., 2018). Despite this intuitive interpretation, the `ROC-AUC` can be overly optimistic with heavily imbalanced data (Davis and Goadrich, 2006), which is why we report additional metrics as well.

4.4.2 Average precision

The average precision (`AP`) is a ranking metric that captures the performance of a classifier on the minority class. It summarises the precision-recall curve, plotting the precision against the recall at each possible cut-off. The `AP` is calculated as the weighted mean of the precision at each cut-off, where the increase in recall (from the previous cut-off) is used as weight.

4.4.3 Recall@100

The `recall@100` reflects the ability of a model to detect 10K filings of failing companies with a fixed 'budget' (specifically when only 100 filings are considered). This performance metric is the fraction of true positives in the 100 highest-ranked instances as predicted by a model. In our context, it measures the proportion of 10K records filed before failure that appear in the 100 10K filings that were assigned the highest likelihood of failure by the model. With a test set containing 122 positives, a perfect model would achieve a `recall@100` of 81.97%, which is the highest achievable value for this metric given our dataset.

4.4.4 Cumulative accuracy profile ratio

Finally, we report the cumulative accuracy profile ratio (`CAP ratio`). This ratio summarises the cumulative accuracy profile curve, plotting the recall at various data proportions based on the scores assigned by a model. Essentially, this curve shows the `recall@k` for varying values of `k`. The metric is then computed as the ratio of (1) the area between the `CAP` curve of the evaluated model and the `CAP` curve of a random model and (2) the area between the `CAP` curve of a perfect model and the `CAP` curve of a random model. Similarly to the `ROC-AUC`, the `CAP ratio` captures the overall ranking performance of a classifier.

5 Results and discussion

In Sect. 5.1 we discuss the results of our experiments. Afterwards, in Sect. 5.2 we give an in-depth analysis of these results, with a particular focus on the interpretabilty of the proposed model and the predictive value of the numerical accounting data and the text in corporate disclosures for business failure prediction.

5.1 Business failure prediction results

The results in Table 2 show that among the numerical baselines, the XGBoost classifier (`N-XGB[num]`) performs best on `ROC-AUC` and `CAP ratio`, while the MLP classifier (`N-MLP[num]`) achieves the highest `AP`. Both models perform equally well in terms of

Table 2 The results of the numerical baselines on the test set. To account for model specific variability and data sampling, the mean results (± std.) are shown over 5 training runs. The best results for each metric are shown in bold (within the table) or marked with a star (*) (across all experiments). If the difference between the best models is statistically insignificant (p-value > 0.05, based on an independent samples t-test), both are shown in bold or marked with a star (*)

Model	N-Z'[num]	N-LR[num]	N-MLP[num]	N-XGB[num]
ROC-AUC	74.93%	91.48%	92.61% (±0.16%)	**93.73% (±0.02%)**
AP	9.95%	11.50%	**17.52% (±0.48%)**	15.71% (±0.09%)
recall@100	17.21%	14.75%	**19.84% (±1.47%)**	**19.18% (±0.45%)**
CAP ratio	49.85%	82.97%	85.23% (±0.31%)	**87.46% (±0.05%)**

Table 3 The results of the textual baselines on the test set. See Table 2 for details on the notation

Model	T-LR[tfidf]	T-XGB[tfidf]	T-LR[emb]
ROC-AUC	87.94%	**90.14% (±0.50%)**	81.79%
AP	23.66%	**26.89% (±1.33%)***	3.89%
recall@100	**28.36%***	**28.69% (±0.73%)***	5.41%
CAP ratio	75.88%	**80.29% (±1.00%)**	63.58%

Table 4 The results of the multi-modal baselines on the test set. See Table 2 for details on the notation

Model	NT-XGB[num+tfidf]	NT-stack[XGB(num),XGB(tfidf)]
ROC-AUC	**94.94% (±0.17%)***	**94.96% (±0.08%)***
AP	**26.09% (±1.29%)***	**27.24% (±1.51%)***
recall@100	**27.38% (±0.84%)***	**28.89% (±1.03%)***
CAP ratio	**89.88% (±0.33%)***	**89.93% (±0.16%)***

recall@100. For the textual baselines, we can see from Table 3 that the XGBoost classifier trained on selected tfidf features (T-XGB[tfidf]) is the best model on every metric, although the logistic regression classifier (T-LR[tfidf]) matches its performance in terms of recall@100. If we compare all uni-modal baselines (Tables 2 and 3), the numerical model N-XGB[num] attains the highest ROC-AUC and CAP ratio while the textual model T-XGB[tfidf] is superior for recall@100 and AP.

Table 4 presents the results for the multi-modal baselines. These results show that the early (NT-XGB[num+tfidf]) and late (NT-stack[XGB(num),XGB(tfidf)]) fusion models perform similarly across all metrics, which indicates that combining the numerical and textual features before or after the primary learning stage leads to comparable results. Furthermore, both models achieve the best results over all experiments, although the textual baseline T-XGB[tfidf] performs equally well on AP and recall@100.

Finally, when evaluating our proposed model (Table 5), we notice that the strategy of encoding the first 300 sentences (NT-att[num,sent-emb-300]) outperforms the alternative where we encode the first and last 250 sentences (NT-att[num,sent-emb-500]) (although both achieve the same AP and recall@100). Across all experiments, our proposed model is competitive in terms of ROC-AUC and CAP ratio, although the perfor-

Table 5 The results of the proposed model on the test set (across the different strategies of encoding 300 sentences or 500 sentences, respectively). See Table 2 for details on the notation

Model	NT-att[num,sent-emb-300]	NT-att[num,sent-emb-500]
ROC-AUC	**94.53% (±0.05%)**	94.17% (±0.15%)
AP	**16.15% (±1.02%)**	**16.91% (±0.66%)**
recall@100	**17.21% (±2.09%)**	**18.36% (±1.49%)**
CAP ratio	**89.06% (±0.09%)**	88.33% (±0.30%)

mance difference with the best-performing model (NT-stack[XGB(num),XGB(tfidf)]) is statistically significant.

5.2 In-depth analysis

In this section we provide an in-depth analysis of our results. We focus on the interpretability of our proposed model in Sect. 5.2.1 and we discuss the predictive value of the numerical and the textual data in corporate disclosures for business failure prediction in Sect. 5.2.2.

5.2.1 Attention-based model interpretability

In Sect. 4.1 we have covered the design of our proposed model and argued that due to its architecture, it offers a form of interpretability through the attention weights. Sentences that are assigned a high attention weight are either informative to predict the outcome or align well with the numerical features. To empirically validate this, we present in Table 6, the sentences with the highest attention weights for the top ranked instances from the test set (the 10K filings with the highest likelihood of business failure according to the model NT-att[num,sent-emb-300]).

Table 6 reveals that the sentences with the highest attention weights cover a variety of topics that can be intuitively linked to business failure such as liquidity or solvency issues, declining demand, breaching of loan covenants and the closing of stores. Interestingly, the model also tends to ignore the text for a number of 10K filings. This is the case for 24 out of the 100 highest ranked instances by our proposed model.[5]

In order to fully understand the interpretability of the attention weights assigned by the model, we have included the selected sentences for several randomly sampled 10K filings from the test set in Appendix D as well. In contrast to the top ranked instances highlighted in Table 6, the sentences within these randomly sampled 10K filings tend to have much lower attention weights. This suggest that the model did not identify sentences that clearly signal business failure (or clear signs of financial health), but instead it distributed its attention more evenly across the text. Although certain selected sentences do hint at either a healthy or a poor financial situation, no clear topics stand out.

Taking a closer look at the results from Tables 2- 5, another interesting finding emerges. The competitive performance of our proposed model in terms of ROC-AUC and CAP ratio contrasts with its poor results in terms of AP and recall@100. Furthermore, all models trained with numerical features (num) outperform their counterparts trained without these features in terms of ROC-AUC and CAP ratio. Additionally, the models trained with

[5] From these 24 instances, 7 were filed before failure (positives) and 17 were not (negatives).

Table 6 For the 10 highest ranked 10K filings in the test set by the `NT-att[num,sent-emb-300]` model, we give the rank, the business failure label, the predicted score ranging from 0 (perfect financial health predicted) up to 1 (certain failure predicted), and the 3 sentences with the highest attention weights. When the model only attends to the trainable vector, no sentences are selected

Rank	Label	Score	Selected sentence (attention weight)
1	`True`	1.0	no sentences selected
2	`True`	0.99	no sentences selected
3	`True`	0.99	There can be no assurance that our efforts will result in any agreement or what the terms of any agreement will be. (16.13%)
			If we do not obtain a waiver or other suitable relief from the lenders under the Credit Agreement or the Term Loan Agreement before the expiration of the 30-day grace period, an event of default under each of the Credit Agreement and Term Loan Agreement would occur. (7.85%)
			If the company does not comply with the covenants in the Revolving Credit Facility, the lenders may, subject to customary cure rights, require immediate payment of all amounts outstanding under the Revolving Credit Facility and may terminate any outstanding unfunded commitments. (5.87%)
4	`True`	0.99	no sentences selected
5	`False`	0.99	no sentences selected
6	`False`	0.99	There can be no assurances that the securitization lenders will agree to any extension of the Securitization Forbearance Agreement or that if such forbearance agreement is terminated early or expires, that the securitization lenders will not pursue any and all remedies available to them. (11.51%)
			There can be no assurance that these efforts will result in any agreement. (10.92%)
			There can be no assurance that any restructuring will be possible on acceptable terms, if at all. (6.27%)
7	`False`	0.99	By March 2020, all of our operating stores were temporarily closed (including our one new store that opened in March 2020). (8.40%)
			Our store in [city] permanently closed in March 2019, as we did not exercise the renewal option, and has been excluded from fiscal 2019 store counts and comparable store sales. (6.10%)
			Nearly all of our store workforce, with the exception of a small team of essential personnel, were furloughed in mid-March 2020. (3.90%)
8	`False`	0.98	We cannot assure you that we will be successful in our recapitalization efforts. (10.31%)
			To date, we have not yet reached an agreement with [legal entity], and at this juncture, it is too early to state with certainty whether the parties will ever enter into such an agreement. (8.87%)
			We do not currently have sufficient funds legally available to be able to satisfy the conditions for terminating the Voting Rights Triggering Event. (4.81%)
9	`True`	0.98	On December 15th, the Loan Parties entered into a First Amendment to Standstill Agreement and Second Amendment to Credit Agreement with certain Standstill Lenders, pursuant to which the maximum duration of the "Standstill Period" was extended from December to February. (4.08%)

Table 6 continued

Rank	Label	Score	Selected sentence (attention weight)
			As described in greater detail below, the extension of the Standstill Period that prohibits the Administrative Agent and the Lenders from exercising their default-related rights and remedies with respect to specified events of default under the Term Loan Agreement will expire on the earliest of the delivery of a notice of termination of the Standstill Period by the Standstill Lenders upon the occurrence of (i) a default under the Term Loan Agreement, or (ii) a breach of, or non-compliance with certain provisions of the Second Amended Standstill Agreement. (3.70%)
			The Company's failure to comply with the Negative Covenants and Excluded Milestones during the Standstill Period would permit the Required Lenders to terminate the Standstill Agreement and constitute an immediate Event of Default under the Term Loan Agreement. (3.52%)
10	`False`	0.98	There can be no assurance that our subscriber base will not continue to decline and that the pace of such decline will not accelerate. (11.94%)
			Given the devastation and loss of power, substantially all customers in those areas were unable to receive our service as of September. (10.75%)
			There can be no assurance that these additional services or other offers will positively affect our results of operations or our [product] subscribers. (3.24%)

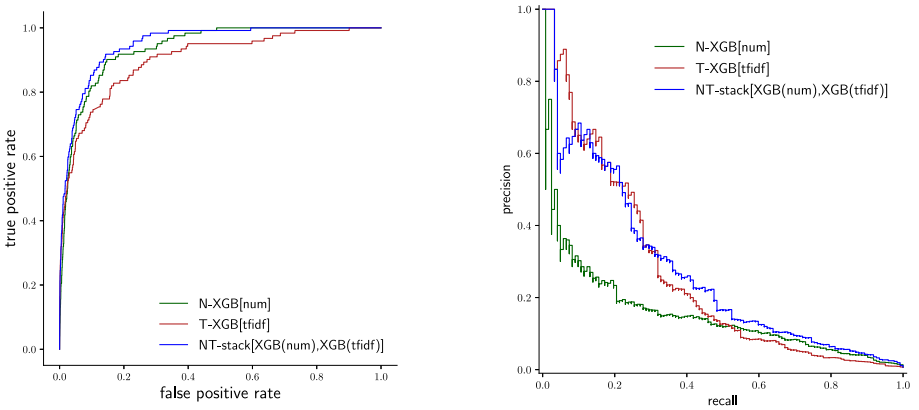
keyword-based features (`tfidf`) excel in terms of `AP` and `recall@100`, compared to the models trained without them. In the next section, we expand on this.

5.2.2 The predictive value of the numerical and textual data

Feature-specific ranking performance. In Fig. 3, we show the receiver operating curves (ROC-curves) and the precision-recall curves (PR-curves) for a selection of our models. From the ROC-curves we can see that the models trained with numerical features (`num`) recall *most* positives (e.g. 85%) at a lower false positive rate compared to the model trained without numerical features. Furthermore, at low recall values (e.g. 10%), the PR-curves from the models trained with keyword-based features (`tfidf`) show higher precision values than the corresponding PR-curve from the model trained without keyword-based features. This means that the keyword-based models are capable of detecting a relatively small fraction of the positives (e.g. 10%), without generating many false positives, while this is not the case for the models trained without keyword-based features.

The performance metrics from Tables 2, 3, 4 and 5 and the analysis of the performance curves from Fig. 3 suggest that (1) including numerical features (`num`) in a model leads to a ranking of the 10K filings where *most* positives (10K filings of failing companies) are positioned highly in the ranking, but not necessarily at the top; and (2) including keyword features (`tfidf`) in a model leads to a ranking of the 10K filings where *a subset* of positives (10K filings of failing companies) are positioned at the top, but the rest of the positives not necessarily highly in the ranking. In order to clarify this finding, we expand on this with an illustration in Appendix E.

The distinctive role of the numerical features and the keywords for business failure prediction can also be analysed by looking at the fitted parameters of our stacking classifier. The β_1 and β_2 values (from the equation presented in Sect. 1) of the `NT-stack[XGB(num),XGB(tfidf)]` classifier are (on average) 1.17 and 0.31, respec-



(a) receiver operating curves (b) precision-recall curves

Fig. 3 The performance curves for a selection of our models trained with numerical features (`num`), keyword-based features (`tfidf`) or a combination of both

tively. This further highlights the importance of the numerical features (financial ratios) for business failure prediction.

Keyword-based business failure prediction. Now that we have an understanding of the reported performance metrics and the corresponding performance curves, we question why a subset of 10K filings appear 'easy-to-classify' with keyword-based features (`tfidf`). To this end, we inspected the features with the highest coefficients—retained after the L1-regularisation—from the `T-LR[tfidf]` model. These includes terms such as *'chapter case'*, *'reorganization'* and *'recovery'*, enabling the model to assign a high likelihood of business failure to 10K filings that contain passages such as *"...it may be necessary for us to seek protection from creditors under Chapter 11 of the U.S. Bankruptcy Code..."*. Consequently, the strong performance of the models trained on these features (in terms of `AP` and `recall@100`) can be attributed to their ability to detect 10K filings where the management explicitly states that they consider filing for bankruptcy.

We acknowledge that our proposed model is not as effective in detecting these cases (where bankruptcy is mentioned) compared to the models trained with keyword-based features. Instead, our model tends to focus on a wider range of topics in the disclosure texts that are intuitively linked to business failure and that are potentially interesting from an interpretability perspective. While this is an important limitation of our model, a stacking approach (as discussed in Sect. 4.3) can help overcome this issue. By combining the predictions of our proposed model (`NT-att[num,sent-emb-300]`) and the best uni-modal baselines (`N-XGB[num]` and `T-XGB[tfidf]`), we obtain a classifier (denoted `NT-stack[all]`) that attains the best results overall (reported in Table 7).

Additionally, we observe that the textual baseline trained on document embeddings (`T-LR[emb]`) performs worse than all other models trained with textual data across every metric. This suggest that the embeddings capture very little information that is useful for the task. It is worth noting that the encoder was not trained specifically for business failure prediction and that it had to encode extremely long documents (over 6,000 words on average) into a single representation.

Table 7 The results of a stacking classifier that combines the predictions of our proposed model (NT-att[num,sent-emb-300]) and those of the best uni-modal baselines (N-XGB[num] and T-XGB[tfidf]). This model attains the best results on every metric, although the multi-modal baselines perform equally well on AP and recall@100

Model	ROC-AUC	AP	recall@100	CAP ratio
NT-stack[all]	95.34% (± 0.07%)	26.68% (± 1.07%)	28.28% (± 1.42%)	90.68% (± 0.13%)

6 Conclusion

Corporate disclosures are growing in complexity, evolving into lengthy documents containing both numerical accounting data and substantial amounts of text (Dyer et al., 2017). Concurrently, the advancements in artificial intelligence (AI), and large language models (LLMs) specifically, have equipped us with the means to develop high performing business failure prediction models. However, given that such models are intended for high-stakes domains like banking, insurance and investing, model interpretability is vital for their practical adoption and offers a way to instill user trust (Bauer et al., 2023).

In this study, we have proposed a new neural network architecture for business failure prediction from corporate disclosures, with an interpretable link between the numerical and textual features. The model combines financial ratios and sentence-level textual features through a well known attention mechanism, commonly used in transformer models (Vaswani et al., 2017), and achieves competitive results on a variety of performance metrics (including ROC-AUC). The sentences that receive the highest attention weight can be intuitively linked to business failure, covering topics such as debt renegotiations, breaching loan covenants or declining product demand, among others.

Moreover, our work highlights the importance of traditional financial ratios for business failure prediction. These numerical features are the key drivers behind our best performing models. Keyword-based textual features are mainly useful to detect specific cases where the management explicitly states that they are considering to file for bankruptcy, making the prediction task based on text trivial. Additionally, we have evaluated the performance of commercially available, pretrained document embeddings in the context of our task. Their poor results suggest that these embeddings capture little information relevant to predict business failure and that tailored textual features are more suited to this end.

Finally, ensemble learning plays a crucial role in our findings. Through a stacking approach, we demonstrate that combining predictions from models trained on different data modalities can significantly improve performance. This ensemble method leverages the complementary strengths of each modality, resulting in our best-performing business failure prediction classifier. Additionally, we have used stacking to address the main limitation of our proposed model, which struggles to detect cases where bankruptcy is explicitly mentioned in the text. By incorporating predictions from models better suited for these cases, we can effectively create a classifier that retains the interpretability of our proposed model—highlighting sentences intuitively linked to business failure—while also being able to detect the more straightforward cases where the possibility of bankruptcy is mentioned.

Appendix A: Financial ratios included in our experiments

The financial ratios included in our experiments (the numerical features denoted as num) are given in Table 8. These ratios were selected from the study of Mai et al. (2019). Specifically, we only include those that can be constructed from the accounting information in a 10K filing. The ratios used by Mai et al. (2019) that require market information (such as stock prices) are discarded.

Table 8 This table presents the financial ratios included in our experiments (num) along with their formulas in Compustat. These ratios were selected from the study of Mai et al. (2019). Specifically, we only include those that can be constructed from the accounting information in a 10K filing. The ratios used by Mai et al. (2019) that require market information (such as stock prices) are discarded

Financial ratio	Compustat	Financial ratio	Compustat
$\frac{\text{current assets}}{\text{current liabilities}}$	$\frac{\text{ACT}}{\text{LCT}}$	$\frac{\text{current liabilities}}{\text{sales}}$	$\frac{\text{LCT}}{\text{SALE}}$
$\frac{\text{accounts payable}}{\text{sales}}$	$\frac{\text{AP}}{\text{SALE}}$	$\frac{\text{total liabilities}}{\text{total assets}}$	$\frac{\text{LT}}{\text{AT}}$
$\frac{\text{cash and short term investments}}{\text{total assets}}$	$\frac{\text{CHE}}{\text{AT}}$	log(total assets)	log(AT)
$\frac{\text{cash}}{\text{total assets}}$	$\frac{\text{CH}}{\text{AT}}$	log(sales)	log(SALE)
$\frac{\text{cash}}{\text{current liabilities}}$	$\frac{\text{CH}}{\text{LCT}}$	$\frac{\text{net income}}{\text{total assets}}$	$\frac{\text{NI}}{\text{AT}}$
$\frac{\text{(EBIT + depreciations and amortisations)}}{\text{total assets}}$	$\frac{\text{EBIT+DP}}{\text{AT}}$	$\frac{\text{net income}}{\text{sales}}$	$\frac{\text{NI}}{\text{SALE}}$
$\frac{\text{EBIT}}{\text{total assets}}$	$\frac{\text{EBIT}}{\text{AT}}$	$\frac{\text{operating income after depreciations}}{\text{total assets}}$	$\frac{\text{OIADP}}{\text{AT}}$
$\frac{\text{EBIT}}{\text{sales}}$	$\frac{\text{EBIT}}{\text{SALE}}$	$\frac{\text{operating income after depreciations}}{\text{sales}}$	$\frac{\text{OIADP}}{\text{SALE}}$
$\frac{\text{total debt}}{\text{total assets}}$	$\frac{\text{DLC+0.5}\times\text{DLTT}}{\text{AT}}$	$\frac{\text{current assets}-\text{inventory}}{\text{total current liabilities}}$	$\frac{\text{ACT}-\text{INVT}}{\text{SALE}}$
$\frac{\text{inventory decrease}}{\text{inventory}}$	$\frac{\text{INVCH}}{\text{INVT}}$	$\frac{\text{retained earnings}}{\text{total assets}}$	$\frac{\text{RE}}{\text{AT}}$
$\frac{\text{inventory}}{\text{sales}}$	$\frac{\text{INVT}}{\text{SALE}}$	$\frac{\text{retained earnings}}{\text{current liabilities}}$	$\frac{\text{RE}}{\text{LCT}}$
$\frac{\text{current liabilities}-\text{cash}}{\text{total assets}}$	$\frac{\text{LCT}-\text{CH}}{\text{AT}}$	$\frac{\text{sales}}{\text{total assets}}$	$\frac{\text{SALE}}{\text{AT}}$
$\frac{\text{current liabilities}}{\text{total assets}}$	$\frac{\text{LCT}}{\text{AT}}$	$\frac{\text{total equity}}{\text{total assets}}$	$\frac{\text{SEQ}}{\text{AT}}$
$\frac{\text{current liabilities}}{\text{total liabilities}}$	$\frac{\text{LCT}}{\text{LT}}$	$\frac{\text{working capital}}{\text{total assets}}$	$\frac{\text{WCAP}}{\text{AT}}$

Appendix B: Training details

Proposed model

As explained in Sect. 4.1, the sentence embeddings of the MD&A's in the dataset with fewer than 300 or 500 sentences were padded. We do not allow padded sentences to participate in the attention operation by setting the dot product of the query vector and the key vector to $-\infty$. In order to handle the class imbalance, we treat the fraction of negative over positive examples in each batch as a hyperparameter that is adjusted by oversampling the positives. This is optimised, along with model-specific hyperparameters, in a grid search procedure where ROC-AUC on the validation set is the maximisation objective. After the optimal hyperparameters are selected, the model is retrained on the combined training and validation set and evaluated on the test set. Furthermore, to account for the randomness caused by weight ini-

tialisation and data sampling, we repeated the entire training process five times with different seed values, and report the mean results along with the standard deviations in our results tables. The model was trained using mini-batch gradient descent with an Adam optimiser and an unweighted binary cross entropy loss. As preprocessing steps, the numerical features were mean-imputed, centered and scaled to unit variance. For a complete overview of the hyperparameter optimisation and data preprocessing steps, we refer to our GitHub page.[6]

Uni-modal baselines

For the baselines using either numerical or textual features, the fraction of negatives over positives in the training set was also treated as a hyperparameter, which is adjusted by oversampling the positives. For each classifier, we tuned this, along with model-specific hyperparameters, to maximise the `ROC-AUC` on the validation set and retrain the final model on the combined training and validation set. The entire training process was repeated five times with different seed values to account for model-specific variability and data sampling. We report the mean performance metrics and the corresponding standard deviations in our results tables. The numerical features were mean-imputed, centered and scaled to unit variance, except for the XGBoost classifiers where mean-imputation is not performed. The `tfidf` feature vectors have been normalised to unit length using their L2 norm to prevent any feature from disproportionately affecting the regularisation process. For a complete overview of the hyperparameter optimisation and data preprocessing steps, we refer to our GitHub page.

Multi-modal baselines

The early fusion baseline is trained similarly to the uni-modal baselines. The fraction of negatives over positives in the training set is a hyperparameter, tuned along with model-specific hyperparameters, with `ROC-AUC` on the validation set as the maximisation objective. Training is repeated five times with different seed values to account for model-specific variability and data sampling. Mean performance metrics and corresponding standard deviations are reported in our results tables. The numerical features were centered and scaled to unit variance. For a complete overview, we refer to to our GitHub page. The late fusion baseline does not require hyperparameter tuning. The class imbalance is handled by the base models and no regularisation is applied.

Appendix C: Altman Z'-score model

As explained in Sect. 4.2, the Z'-score is a revised version of the original Altman Z-score that can be calculated from the accounting information in 10K filings. Therefore, it relies on different financial ratios compared to the other models included in our experiments. We have

[6] The code to replicate the experiments is available at https://github.com/henriarnoUG/ECL

estimated the parameters of the following model:

$$\text{Z'-score} = \beta_1 \left(\frac{\text{working capital}}{\text{total assets}} \right) + \beta_2 \left(\frac{\text{retained earnings}}{\text{total assets}} \right) + \beta_3 \left(\frac{\text{EBIT}}{\text{total assets}} \right) + \beta_4 \left(\frac{\text{book value of equity}}{\text{book value of total liabilities}} \right) + \beta_5 \left(\frac{\text{sales}}{\text{total assets}} \right)$$

Appendix D: Sentences with high attention weights for randomly sampled 10K filings

In order to better understand the interpretability of the attention weights assigned by our proposed model, we show the sentences with the highest attention weights for 5 randomly sampled 10K filings in the test set (in Table 9 below).

Appendix E: Numerical versus keyword-based features in business failure prediction

In this section, we expand on the role of the numerical features (`num`) and the keyword-based features (`tfidf`) for business failure prediction, as discussed in Sect. 5.2.2, with an illustration. Assume that we have a dataset consisting of 100 positive instances (representing 10K filings of failing companies) and 10,000 negative instances (representing 10K filings from non-failing companies). This implies a class imbalance of 1 positive for every 100 negatives, similar to the class distribution in our dataset. Furthermore, assume that we have trained two models, respectively representing our classifiers trained with numerical features (denoted $\text{model}_{\texttt{num}}$) and those trained with keyword-based features (denoted $\text{model}_{\texttt{tfidf}}$). On the one hand, if we rank the instances in the dataset according to the scores assigned by $\text{model}_{\texttt{num}}$, all 100 positives are evenly distributed in the top 1,000 of the dataset (1 in the top 10, 2 in the top 20, 3 in the top 30, etc.). On the other hand, if we rank them according to the scores assigned by $\text{model}_{\texttt{tfidf}}$, a subset of 20 positives make up the top 20 predictions. For the remaining 80 positives, 60 are distributed evenly in the top 2,500 of the dataset while 20 are distributed evenly in the remaining ($\pm$) 7,500 instances of the dataset. The ranking metrics in this hypothetical setting are shown in Table 10, while the PR-curves and the ROC-curves are shown in Fig. 4.

Since $\text{model}_{\texttt{num}}$ ranks all positives in the top 10% of the dataset, the ROC-curve attains high recall values at a much lower false positive rate, compared to $\text{model}_{\texttt{tfidf}}$. In fact, $\text{model}_{\texttt{tfidf}}$ only recalls the subset of 20 positives quickly, while the remaining 80 positives are only detected after having generated many false positives. By consequence, the `ROC-AUC` of $\text{model}_{\texttt{num}}$ will be significantly higher compared to the `ROC-AUC` of $\text{model}_{\texttt{tfidf}}$. A similar reasoning holds for the CAP-curve and the corresponding `CAP ratio`.

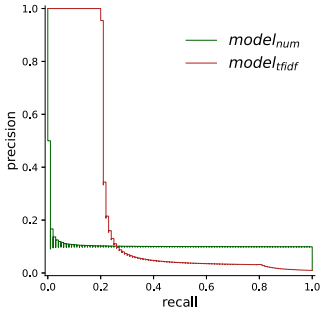
If we turn our attention to the PR-curves, we can see that $\text{model}_{\texttt{num}}$ recalls all positives with a precision of ($\pm$) 10%. This can be attributed to the even distribution of all 100 positives in the top 10% of the dataset. In contrast, $\text{model}_{\texttt{tfidf}}$ can recall the subset of 20 positives at a precision of 100%. This area of the PR-curve significantly contributes to the `AP` of $\text{model}_{\texttt{tfidf}}$, making it much higher than the `AP` of $\text{model}_{\texttt{num}}$, despite the fact that the remaining 80 positives are recalled at a much lower precision. Lastly, it is also the subset

Table 9 For 5 randomly sampled 10K filings in the test set, we give the business failure label, the predicted score by the `NT-att[num,sent-emb-300]` model and the 3 sentences with the highest attention weights

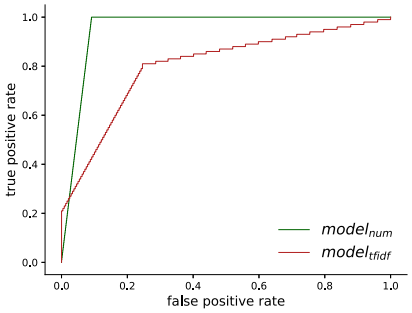
Label	Score	Selected sentence (attention weight)
`false`	0.01	For this year, our SG&A expenses increased 4.6% versus last year. (2.12%)
		The largest factor driving the increase in SG&A expenses (increase year-over-year), is additional administrative expenses attributed to higher incentive-based compensation (approximately 12 million dollars) and performance-based stock compensation (approximately 10 million dollars), as performance targets were met this year to a higher degree than last year. (1.79%)
		Based on current interest rate and debt levels, we expect our aggregate interest expense for next year to be between 6 million dollars and 9 million dollars. (1.78%)
`false`	0.03	In July, we entered into and currently maintain a five-year revolving credit agreement. (1.42%)
		This credit agreement provides for a 250 million dollars senior revolving credit facility which may be increased up to 375 million dollars. (1.21%)
		Last year, we borrowed 100 million dollars under our credit facility that we fully repaid in the next fiscal year. (1.14%)
`false`	0.20	The wholesale segment generally has less risk than the retail segment. (1.65%)
		In the event of a decrease in the company's credit ratings or a disruption in the financial markets, the company may not be able to refinance its maturing debt in the financial markets. (1.35%)
		In such circumstances, the company would be exposed to liquidity risk to the degree that the timing of debt maturities differs from the timing of receivable collections from customers. (1.16%)
`false`	0.04	The company's consolidated financial statements are prepared on the basis of GAAP. (0.71%)
		Management generally measures the company's operating results by examining the company's net income and return on equity, as well as the loss and settlement expense, acquisition expense and combined ratios. (0.65%)
		The following estimates and assumptions are considered by management to be critically important in the preparation and understanding of the company's financial statements and related disclosures. (0.61%)
`false`	0.01	Management believes that net income, as defined by GAAP, is the most appropriate earnings measurement. (1.30%)
		In preparing the Consolidated Financial Statements, management is required to exercise judgment and make assumptions and estimates that may impact the carrying value of assets and liabilities and the reported amounts of revenues and expenses. (1.10%)
		If it is determined that the company is the primary beneficiary of a VIE, the company's Consolidated Financial Statements would include the operating results of the VIE rather than the results of the variable interest in the VIE. (1.09%)

Table 10 The results of the hypothetical models from the illustration. It is clear that $\text{model}_{\texttt{num}}$ does significantly better in terms of `ROC-AUC` and `CAP ratio` compared to $\text{model}_{\texttt{tfidf}}$ while the opposite holds for `AP` and `recall@100`

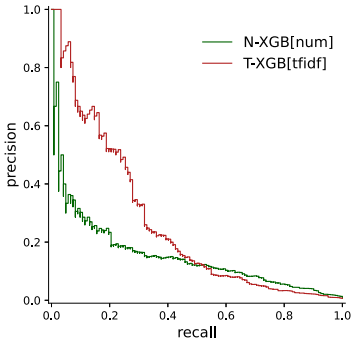
model	`ROC-AUC`	`AP`	`recall@100`	`CAP ratio`
$\text{model}_{\texttt{num}}$	95.45%	10.63%	10.00%	90.90%
$\text{model}_{\texttt{tfidf}}$	80.17%	24.61%	22.00%	60.34%



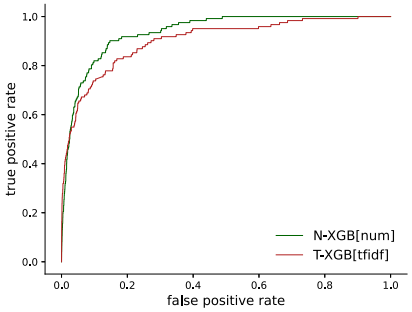
(a) precision-recall curve (illustration)



(b) receiver operating curve (illustration)



(c) precision-recall curve (trained models)



(d) receiver operating curve (trained models)

Fig. 4 Performance curves of the hypothetical models from the illustration and a selection of our models trained with numerical features or keyword-based features (the `N-XGB[num]` model and the `T-XGB[tfidf]` model, respectively)

of 20 correctly classified positives by $\text{model}_{\texttt{tfidf}}$ that causes its `recall@100` to be much higher, than the `recall@100` of $\text{model}_{\texttt{num}}$.

Although this illustration is an idealised representation of the behaviour of our models trained with numerical features (`num`) and those trained with keyword-based features (`tfidf`), it provides the intuition necessary to understand the observed ranking metrics and performance curves. The resemblance between the PR-curves and the ROC-curves of our `N-XGB[num]` and `T-XGB[tfidf]` models and the corresponding curves from the illustration (in Fig. 4) is evident.

Acknowledgements We thank Marijn Verschelde and Matthias Bogaert for useful comments.

Funding This study was funded by Research Foundation Flanders (FWO) under Grant Numbers G006421N and 11Q2C24N.

Declarations

Conflict of interest The authors declare that there are no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of adaboost and neural networks. *Decision Support Systems, 45*(1), 110–122. https://doi.org/10.1016/j.dss.2007.12.002

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609. https://doi.org/10.2307/2978933

Altman, E. (2013). Chapter 17: Predicting financial distress of companies: Revisiting the z-score and zeta models. In *Handbook of research methods and applications in empirical finance*.

Arno, H. , Mulier, K. , Baeck, J., & Demeester, T. (2022). Next-year bankruptcy prediction from textual data: Benchmark and baselines. In *Proceedings of the 4th workshop on financial technology and natural language processing*.

Arno, H. , Mulier, K. , Baeck, J., & Demeester, T. (2023). From numbers to words: Multi-modal bankruptcy prediction using the ECL dataset. In *Proceedings of the 6th workshop on financial technology and natural language processing*.

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(ai)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research, 34*(4), 1582–1602. https://doi.org/10.1287/isre.2023.1199

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research, 4*, 71–111. https://doi.org/10.2307/2490171

Borchert, P., Coussement, K., De Caigny, A., & De Weerdt, J. (2023). Extending business failure prediction models with textual website content using deep learning. *European Journal of Operational Research, 306*(1), 348–357. https://doi.org/10.1016/j.ejor.2022.06.060

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123–140. https://doi.org/10.1007/BF00058655

Cecchini, M., Aytug, H., Koehler, G., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems, 50*(1), 164–175. https://doi.org/10.1016/j.dss.2010.07.012

Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research, 312*(1), 357–372. https://doi.org/10.1016/j.ejor.2023.06.036

Cortés, A., Martínez, G., & Rubio, G. (2007). A boosting approach for corporate failure prediction. *Applied Intelligence, 27*, 29–37. https://doi.org/10.1007/s10489-006-0028-9

Coussement, K., & Benoit, D. (2021). Interpretable data science for decision making. *Decision Support Systems, 150*, 113664. https://doi.org/10.1016/j.dss.2021.113664

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on machine learning*.

De Bock, K. (2017). The best of two worlds: Balancing model strength and comprehensibility in business failure prediction using spline-rule ensembles. *Expert Systems with Applications, 90*, 23–39. https://doi.org/10.1016/j.eswa.2017.07.036

De Bock, K., Coussement, K., De Caigny, A., Słowiński, R., Baesens, B., Boute, R., & Weber, R. (2024). Explainable ai for operational research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research, 317*(2), 249–272. https://doi.org/10.1016/j.ejor.2023.09.026

De Bock, K., & De Caigny, A. (2021). Spline-rule ensemble classifiers with structured sparsity regularization for interpretable customer churn modeling. *Decision Support Systems, 150*, 1–14. https://doi.org/10.1016/j.dss.2021.113523

Doumpos, M., Andriosopoulos, K., Galariotis, E., Makridou, G., & Zopounidis, C. (2017). Corporate failure prediction in the European energy sector: A multicriteria approach and the effect of country characteristics. *European Journal of Operational Research, 262*(1), 347–360. https://doi.org/10.1016/j.ejor.2017.04.024

Du Jardin, P. (2021). Forecasting corporate failure using ensemble of self-organizing neural networks. *European Journal of Operational Research, 288*(3), 869–885. https://doi.org/10.1016/j.ejor.2020.06.020

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from latent Dirichlet allocation. *Journal of Accounting and Economics, 64*(2), 221–245. https://doi.org/10.1016/j.jacceco.2017.07.002

Fernández, A., García, S., Galar, M., Prati, R., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 10). Berlin: Springer.

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the 13th international conference on machine learning*.

García, V., Marqués, A., & Salvador Sánchez, J. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion, 47*, 88–101. https://doi.org/10.1016/j.inffus.2018.07.004

Glikson, E., & Woolley, A. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Holder-Webb, L., & Cohen, J. (2007). The association between disclosure, distress, and failure. *Journal of Business Ethics, 75*, 301–314. https://doi.org/10.1007/s10551-006-9254-7

Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications, 117*(1), 287–299. https://doi.org/10.1016/j.eswa.2018.09.039

Kim, M., & Kang, D. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications, 37*(4), 3373–3379. https://doi.org/10.1016/j.eswa.2009.10.012

Liang, D., Lu, C., Tsai, C., & Shih, G. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research, 252*(2), 561–572. https://doi.org/10.1016/j.ejor.2016.01.012

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research, 274*(2), 743–758. https://doi.org/10.1016/j.ejor.2018.10.024

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge: Cambridge University Press.

Mayew, W., Sethuraman, M., & Venkatachalam, M. (2015). MD&A disclosure and the firm's ability to continue as a going concern. *The Accounting Review, 90*(4), 1621–1651. https://doi.org/10.2308/accr-50983

Odom, M., & Sharda, R. (1990). A neural network model for bankruptcy prediction. In *Proceedings of the international joint conference on neural networks*.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109–131. https://doi.org/10.2307/2490395

Panda, D. , Chakladar, D., & Dasgupta, T. (2020). Multimodal system for emotion recognition using EEG and customer review. In *Proceedings of the global artificial intelligence congress*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery, 8*(4), 1249. https://doi.org/10.1002/widm.1249

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies, 146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Sleeman, W., Kapoor, R., & Ghosh, P. (2022). Multimodal classification: Current landscape, taxonomy and future directions. *ACM Computing Surveys, 55*(7), 1–31. https://doi.org/10.1145/3543848

Sobrie, L., Verschelde, M., & Roets, B. (2024). Explainable real-time predictive analytics on employee workload in digital railway control rooms. *European Journal of Operational Research, 317*(2), 437–448. https://doi.org/10.1016/j.ejor.2023.09.016

Vaswani, A. , Shazeer, N. , Parmar, N. , Uszkoreit, J. , Jones, L. , Gomez, A., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st conference on neural information processing systems*.

Wolpert, D. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Xu, F. , Uszkoreit, H. , Du, Y. , Fan, W. , Zhao, D., & Zhu, J. (2019). Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Proceedings of the 8th natural language processing and Chinese computing conference*.

Xu, K. , Lam, M. , Pang, J. , Gao, X. , Band, C. , Mathur, P., & Xing, E. (2019). Multimodal machine learning for automated ICD coding. In *Proceedings of the 4th machine learning for healthcare conference* (pp. 197–215).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.