Neural Computing and Applications
https://doi.org/10.1007/s00521-024-09474-8

ORIGINAL ARTICLE

Interpretability of deep learning models in analysis of Spanish financial text

César Vaca[1] · Manuel Astorgano[1] · Alfonso J. López-Rivero[3] · Fernando Tejerina[2] · Benjamín Sahelices[1]

Received: 10 February 2023 / Accepted: 14 January 2024

Abstract

Artificial intelligence methods based on deep learning (DL) have recently made significant progress in many different areas including free text classification and sentiment analysis. We believe that corporate governance is one of these areas, where DL can generate very valuable and differential knowledge, for example, by analyzing the biographies of independent directors, which allows for qualitative modeling of their profile in an automatic way. For this technology to be accepted it is important to be able to explain how it generates its results. In this work we have developed a six-dimensional labeled dataset of independent director biographies, implemented three recurrent DL models based on LSTM and transformers along with four ensembles, one of which is an innovative proposal based on a multi-layer perceptron (MLP), trained them using Spanish language and economics and finance terminology and performed a comprehensive test study that demonstrates the accuracy of the results. We have also performed a complete study of explainability using the SHAP methodology by comparatively analyzing the developed models. We have achieved a mean error (MAE) of 8% in the modeling of the open text biographies, which has allowed us to perform a case study of time analysis that has detected significant variations in the composition of the Standard Expertise Profile (SEP) of the boards of directors, related to the crisis of the period 2008–2013. This work shows that DL technology can be accurately applied to free text analysis in the finance and economic domain, by automatically analyzing large volumes of data to generate knowledge that would have been unattainable by other means.

Keywords Machine learning · Deep learning · Boards of directors profiling · Interpretability · Artificial intelligence · Finance · Economics

C. Vaca, M. Astorgano, A. J. López-Rivero, F. Tejerina and B. Sahelices have contributed equally to this work.

✉ César Vaca
cesar.vaca@uva.es

Manuel Astorgano
manuel.astorgano@estudiantes.uva.es

Alfonso J. López-Rivero
ajlopezri@upsa.es

Fernando Tejerina
ftejerina@uva.es

Benjamín Sahelices
benjamin.sahelices@uva.es

[1] GIR GCME. Department of Informatics, Universidad de Valladolid, Valladolid, Spain

[2] Department of Financial Economics and Accounting, Universidad de Valladolid, Valladolid, Spain

[3] Faculty of Informatics, Universidad Pontificia de Salamanca, Salamanca, Spain

Published online: 26 February 2024

1 Introduction

Independent directors represent a very significant fraction of the decision-making capacity on the boards of directors of companies listed on the Spanish stock exchange. Each year, the Annual Corporate Governance Report is published on the website of the National Securities Market Commission (CNMV). This report is made available to the investors and scientific community as part of the open data repertoire, but its format is unstructured free text which greatly complicates its analysis. An important part of this report are the biographies of the independent directors. In order to model and analyze the operation of these companies and to understand the reasons behind their main strategic decisions, it is very important to know the background and professional career of the independent directors. But the analysis of this information must be done manually, and therefore, requires large amounts of resources given the large number of people, companies and years involved. The development of learning enhancement techniques in the last years has greatly improved the performance of deep learning (DL) models and, as a consequence, the fields in which their application is a success story have increased significantly. In particular, the use of DL recurrent networks generates very good results in the field of text classification [1–3] and sentiment analysis [4–6]. Models based on recurrent neural networks, such as BERT [7] and LSTM [8] provide very promising results in natural language processing and open the door to a large number of applications in many different areas [9–11]. We believe that corporate governance is an area, where DL-based models can generate valuable insight. In this work we propose to analyze the biographies of independent directors to carry out an automated qualitative modeling of their profile.

The excellent ability of DL models to interpret complexity through a generic training stage is the main advantage and technological breakthrough of DL. The main reason for this success is the huge amount of parameters it is able to train and manage, which gives it great adaptability and learning capacity in very complex environments. This great advantage is also the origin of its main drawback, which is the difficulty in the reasoning capacity for explaining the decision-making process employed, i.e., its interpretability. The ability to interpret the decisions made by a model is as important as its ability to model complex systems. DL models should not be used as black boxes but should be provided with analytical tools to justify their decision generation process.

In this paper we propose to take advantage of the excellent natural language analysis features of DL models based on recurrent networks to analyze the biographies of independent directors. For this purpose, we propose three recurrent DL models and four ensembles, one of them in turn an MLP. These seven models generate a conceptual classification based on regression of the professional profiles of the boards of directors. To feed these models we have created a six-dimensional labeled dataset of independent director biographies. The models were initially trained with texts in the Spanish language together with specific books in the financial field and business management. Finally, the models were trained with our six-dimensional dataset and a test study was carried out on a dataset not used in the training phase. This study shows the statistical characteristics of our proposal and serves as the basis for the case study carried out. An important contribution of this work is the interpretability analysis performed based on [12]. In this analysis, a framework is developed to link the responses generated by our models with the main elements that have been taken into consideration to generate these responses. Finally, using the results of our DL models on a wide dataset, a case study has been carried out from the perspective of time evolution of the Standard Expertise Profile (SEP) in corporate management which reveals very interesting trends in the pre-crisis, crisis and post-crisis periods of the period (2004, 2020).

The main contributions of this work are the creation of our own dataset with more than a thousand labeled entries of biographies of independent directors, the training and adjustment of three state-of-the-art DL models together with four ensembles one of which is an innovative MLP-based proposal that significantly improves model accuracy, the test study carried out to characterize our proposal, an interpretability study that facilitates understanding the mechanism of generation of results of our model and, finally, the case study carried out with our best model in which the SEP index is proposed to represent the influence of the biographies of the independent directors in the companies management.

The remainder of this article is organized as follows. In Sect. 2 we show other scientific proposals close to ours, in Sect. 3 we present our dataset, the DL models, we describe the training phase and perform a test study, in Sect. 4 we make an interpretability study of our models, in Sect. 5 we show the results of our best model applied to a full business administration dataset and, finally, the conclusions are shown in Sect. 6.

2 Related work

Statistical language modeling seeks to assign probabilities to different linguistic levels such as word, sentence, paragraph or document. In this area, multiple works have been proposed in recent years in a wide range of application areas. The models whose results are most promising are those based on DL, such as LSTM [8] and BERT [7, 13]. Conneau et al. [1] proposes to improve the basic LSTM model by adding a deep convolutional network that demonstrates a clear improvement in the problem of text classification, categorization and sentiment analysis. Similarly, Lyu and Liu [3] combines bidirectional RNNs to obtain word representation information that is analyzed by a CNN to extract features about the importance of each word. The results show a clear advantage over pure RNN models.

In the area of sentiment analysis and automatic sentiment-biased report generation, Radford et al. [4] proposes a multiplicative LSTM model that significantly improves training times and reduces the amount of labeled data required to achieve results within the state-of-the-art. In a similar application domain, Mnassri et al. [5] combines BERT models with LSTM and CNN to classify hate-sentiment text on Twitter. Their results are especially promising when combining the basic models into ensembles with different heuristics. Similarly, [14] performs a detailed study of the performance of transformer-based pre-trained language model for classification of depression intensity using Tweets.

In the Fintech field, [15] reviews the most influential technologies in this area, including AI and DL. Similarly, [16] reviews the literature related to ML and DL in the context of the financial services industry. [17] shows a practical application of the DL *word2vec* model to conduct a comprehensive study of corporate culture of different institutions. To train their models they use the earning calls of multiple companies over the period 2001–2018. The results with *word2vec* show good correlation with external test data which validates this DL technique in the field of corporate culture analysis.

In the context of DL, the trend in scientific work is a progressive increase in the size and complexity of the models, giving rise to an increasing problem of interpretability of their results. Tjoa and Guan [18] carries out a survey of interpretability and explainability analysis methods on complex DL models and categorizes them in the medical field. Lundberg and Lee [12] proposes SHapley Additive exPlanations (SHAP) as a unified framework for interpretability analysis based on assessing the individual importance of each token involved in the prediction. This proposal allows for an intuitive and generic mechanism for interpretability analysis in complex systems. Lapuschkin et al. [19] analyzes the interpretability of DL models and proposes the spectral relevance analysis to characterize and validate this type of models generating also accurate information about reliability of the predictions.

3 Methods

In this paper we propose several DL models to automatically characterize professional profiles based on the description of the free text biographies of the boards of directors of Spanish listed companies. To this end, this section describes the datasets used, the models developed, the training procedure and characteristics and, finally, a test analysis that characterizes the quality of the models developed.

3.1 Datasets

The data used were taken from the Annual Corporate Governance Report (ACGR) and correspond to 10,000+ biographies written in Spanish language of the independent directors of Spanish listed companies in the period 2004–2020. Each biography consists of natural language free text with variable size and no common rules, which describes the curriculum vitae of each director each year. To extract information from the biographies of public reports, web scraping and complex analysis of PDF files were used. Each of the 10,000+ biographies obtained was written differently from all the others. Since there are no rules to describe the professional profile, the generation of the descriptive text of each biography has an important non-deterministic component. Therefore, there is a great variability in the writing, some profiles are described with a few words, others occupy paragraphs of more than 30 lines of text and the prose style is very different between them.

Figure 1 shows the details of the dataset, its composition, treatment and relationship with the training, validation, test and exploitation process. Firstly, the repertoire of biographies was preprocessed to eliminate unaccepted characters, writing errors, empty entries, erroneous formatting, nested character sequences, information added in the extraction process and other common situations arising when accessing open and highly unstructured data. Next, the data were cleaned. To do this, firstly those entries with errors such as lack of information, exchange of fields or the existence of erroneous fields were identified and corrected and, secondly, those entries with repeated information due to alias problems were corrected, that is, different entries referring to the same people but who had not been identified as such. As a result of this data cleaning process, 6564 validated entries were obtained with information on the professional profiles of independent directors. In addition to the CV, also the gender, the company name and the business sector were identified.

To characterize the professional profile of the biographies, a six-dimensional representation system was designed. That is, for each profile a numerical value is

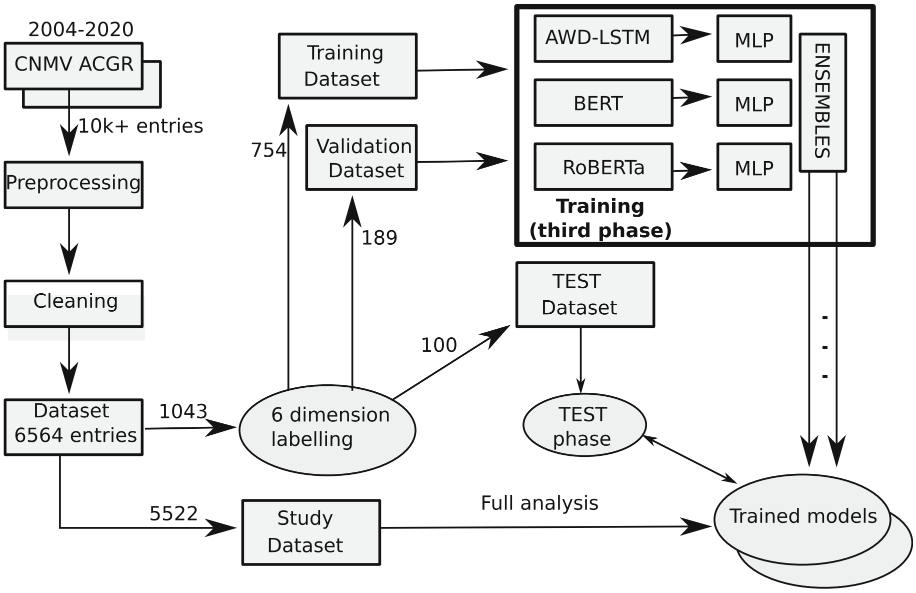


Fig. 1 Description of the dataset created in relation to the training and test of the models

assigned for each dimension that is linked to the intensity of the professional experience in said dimension. Six dimensions were identified according to their special relevance in the financial and business environment: Financial (F), Executive/Consultant (E/C), Audit/Tax/Accountant (A/T/A), Legal (L), Political (P) and Academic (Ac). The justification from the financial and business management point of view of the use of these six dimensions is made in Sect. 5. The main details of the dataset are explained next.

Linking a free text describing a person's biography to one or more of these dimensions is not an error-free mathematical process. On the contrary, it is a qualitative procedure that must be performed by an expert. A total of 1043 entries in the training set were labeled by a human expert and each entry was assigned a value between 0 and 1 with one significant digit for each of the six dimensions, quantitatively approximating an expert's qualitative assessment. It is possible that the same person can have the maximum value in more than one dimension, which means that he/she has reached the maximum categorization in more than one area. The DL techniques in this work use this dataset to implement a regression model that mimics the human model for characterizing biographies.

Of the 1043 labeled biographies, 100 are separated to carry out the test study, which is essential to know the behavior and generalization capacity of our model. This dataset is called the test set. The remaining 943 entries were used in the training iterations, 754 for training and 189 for validation and decision making on hyperparameters at the end of each epoch.

3.2 Models

In this work we use deep learning models for text analysis which aim to estimate the probability distribution of the words that conform a directors' CV and output the regression values for each individual profile. In order to perform a comparative analysis of interpretability, three different DL models have been constructed. In this way it is possible to generate different explanations for the inferences generated by each model in order to compare the different interpretations. The interpretability study allows to know the strengths and weaknesses of each model and thus improves the understanding and explanatory capability of the regression results obtained.

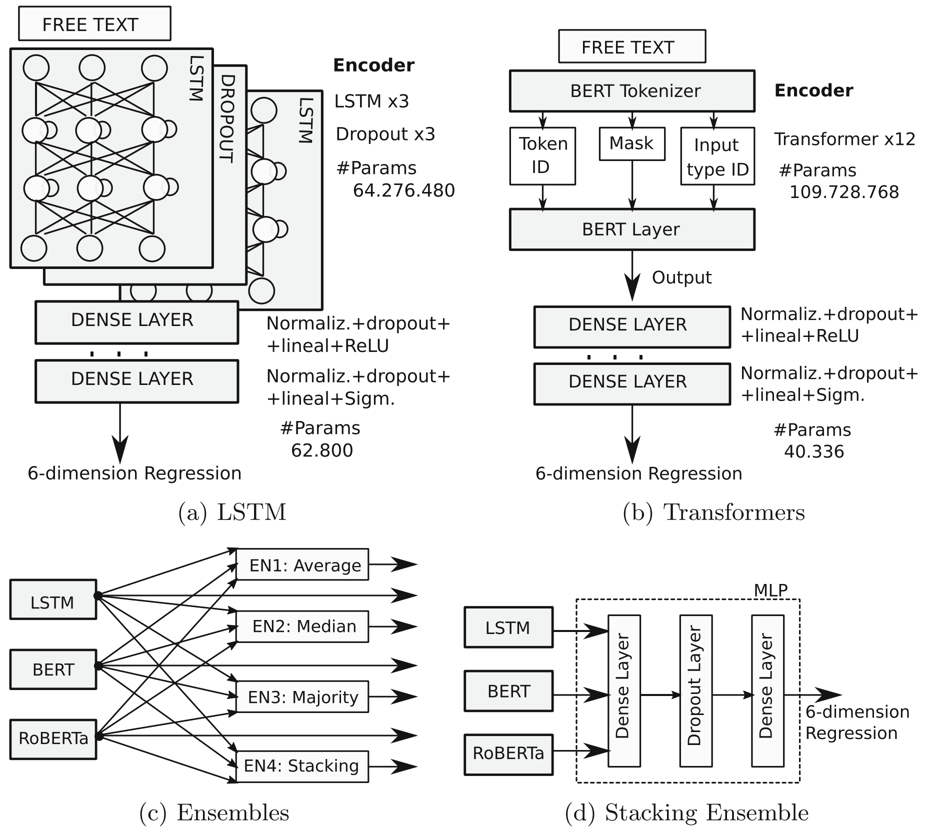
Two different architectures have been used, ASGD Weight-Dropped LSTM neural networks (AWD-LSTM) and transformers. AWD-LSTM neural networks [8] bases on a classic RNN-LSTM which implements the concept of *sequence*, while introducing new regularization techniques that mitigate the tendency of LSTMs to overfit to the training data. The transformer architecture was first introduced in [13]. Here, we use two variations of this architecture, BERT and RoBERTa. BERT [7, 20] architecture has been pre-trained with a large Spanish corpus [21]. Transformers in BERT allow for both a deeper training due to parallelization and better context inference thanks to their attention mechanism. RoBERTa [22] is an enhanced version of BERT with improved training methods. RoBERTa main improvements are the dynamic modification of the mask patterns applied in the training phase by replicating the training data, the modification of the next sentence prediction (NSP) loss function and the training with larger sequences and for more epochs with bigger batches. Figure 2 shows the main characteristics of the models used in this work.

Based on the three models developed, four ensembles were also proposed to reduce the negative worst-case effects of each model:

- *Average*: this is the simplest model and is based on the average of the three values.
- *Median*: takes the central value of the ordered values obtained from the three models.
- *Majority*: takes the inferences of the three models, analyzes them qualitatively and generates the democratic majority as its response. In other words, a response is generated based on the majority trend of the three models. In the case of three qualitatively incompatible responses, one model is given priority over the others.
- *Stacking*: a simple MLP has been added that takes as input the outputs of the three models and generates a new prediction based on the selection of specific components according to criteria learned in the training process. The training process takes, for each epoch, the outputs generated by the three models together with the label generated by the human expert, thus generating an internal representation of the quality of each model with respect to each of the six dimensions.

The last ensemble described, stacking, is actually a new MLP model that is trained to adjust the outputs of the three main models, LSTM, BERT and RoBERTa. In the training process, this new model learns the biases that each of the three base models have with respect to each of the six dimensions described in Sect. 3.1, namely F, E/C, A/T/A, L, P and Ac. As a result, this ensemble selects the best of the three models using the best characteristics that each of the base models has for each of the dimensions.

Depending on the complexity and resources used in the design of the stacking ensemble, a wide range of results can be obtained. Those shown in this work have been achieved with a simple version of the MLP, using a single hidden layer. In order to know the hypothetical maximum optimization upper bound, a study of the best possible stacking ensemble has been carried out. This is a model that reproduces the behavior of a stacking ensemble when it has an ideally large number of assigned resources. This variant of the stacking ensemble is called *Lower bound* and its value is represented and explained in the Sect. 3.4. This is an ideal approximation that cannot be achieved in practice and is included only for comparative purposes.



(a) LSTM

(b) Transformers

(c) Ensembles

(d) Stacking Ensemble

Fig. 2 Architecture of the LSTM and transformed-based models

3.3 Training

The training phase has been done for the Spanish language, which is the language used in the reports of listed companies. Training is made up of four phases (see Fig. 3), in which the first three have been designed based on [2]. The first phase corresponds to the basic training in Spanish language to obtain a language model that adjusts the probability distribution of words in specific syntactic structures. This phase is trained by self-supervised learning using a randomly chosen section of Wikipedia due to computational limitations. This first phase of training has been performed only on the AWD-LSTM model since we have used pre-trained models in Spanish for BERT and RoBERTa [21] which are public and free access.

In the second phase of training, academic financial and management textbooks were selected for self-supervised learning. The objective was to improve the language model and adjust it to the specific vocabulary and syntactic structures of the financial and management field. This process is commonly known as fine tuning. The third phase of training consists of a regression model with the labeled dataset described in Sect. 3.1. A set of fully-connected layers with six regression outputs for the six labeling dimensions described was added to each model. The training set consisting of 943 professional profiles of directors was divided into two groups, 80% dedicated to training and the remaining 20% to validation at the end of each epoch. This validation phase is used to modify the hyperparameters of the model for the next iteration. Finally, the fourth phase has been added to mitigate the lack of training data. To achieve this, state-of-the-art generative models are used that allow us to generate variations on the text that describes the professional profiles of the directors. This phase is described in Sect. 3.3.1.

Figure 4 shows the four phases of training performed for the AWD-LSTM model. The weights of this model, prior to the first training phase, were randomly initialized. The plots show the loss function for each epoch over the training set and over the validation set. The first two phases were trained for 10 epochs due to computational limitations. The overfitting detected in the first two phases of self-supervised learning is significantly reduced in phase three when the regression is trained. Similarly, Figs. 5 and 6 show the training for the BERT and RoBERTa models, starting in phase two since weights were initialized from generic pre-trained models for the Spanish language.

All the results have been obtained using Python and additional libraries such as PyTorch, Scikit-learn and SHAP. The three models developed have very low computational and storage requirements. The BERT and RoBERTa models were trained for 24 h to obtain the results shown. The AWD-LSTM model required 96 h of training due to the first phase of language modeling on Wikipedia being more computationally expensive. All training was performed on a single NVIDIA Tesla P100 GPU with 16GB of memory.

3.3.1 Generative data augmentation

During the development of this work, artificial intelligence services based on deep learning have rapidly become popular. Among them is the GPT-4 (Generative Pre-trained Transformer [23, 24]) system which is a conversation-oriented transformer-based Large Language Model (LLM). GPT-4 relies on transformers trained on a huge scale to give rise to a very precise LLM. Additionally, it has been trained through reinforcement learning to improve its human interaction capabilities and comply with political and social regulations and restrictions. Another widely popular LLM which was recently updated is Llama2 [25]. This one, developed by Meta, is open source and its three versions (which differ in their number of parameters) are available for download.

Since the labeling of the dataset used in phase three of the training has been carried out by a human expert in the financial and business field, the volume of data available is small compared to what would be recommended for good training. It is for these reasons that the use of state-of-the-art LLMs that allow applying data augmentation through generative techniques is proposed. This is what is represented in phase 4 of Fig. 3.

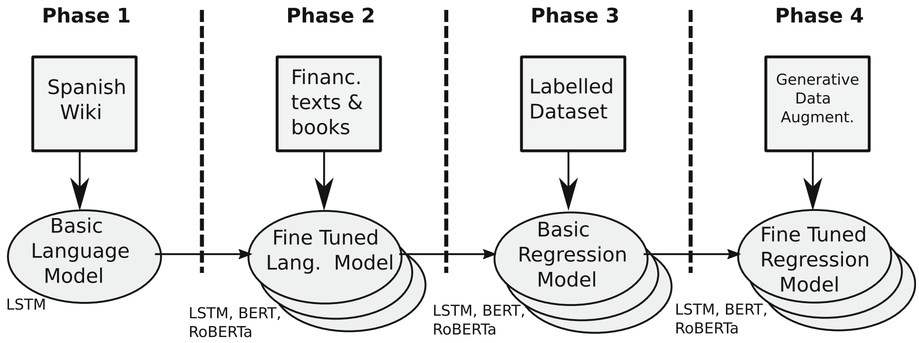
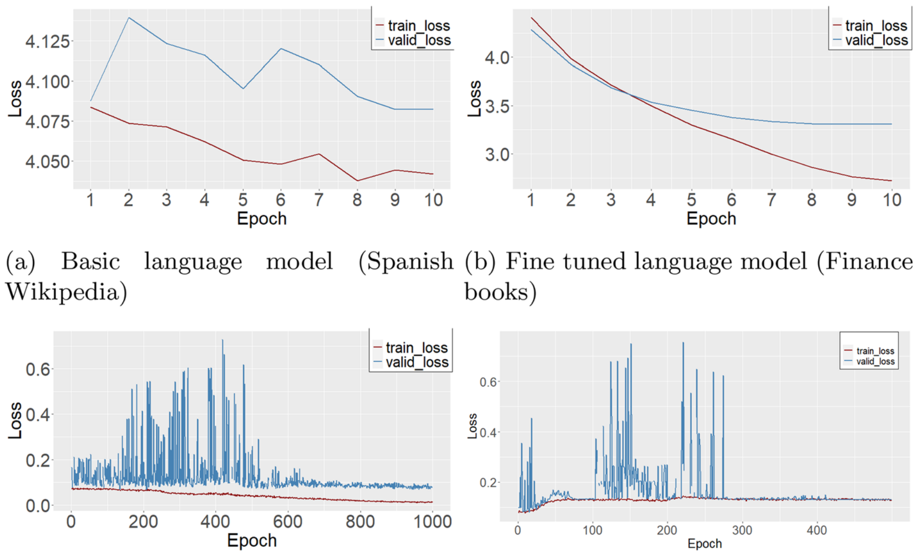


Fig. 3 Description of the four training phases

Fig. 4 AWD-LSTM training metrics for the four phases

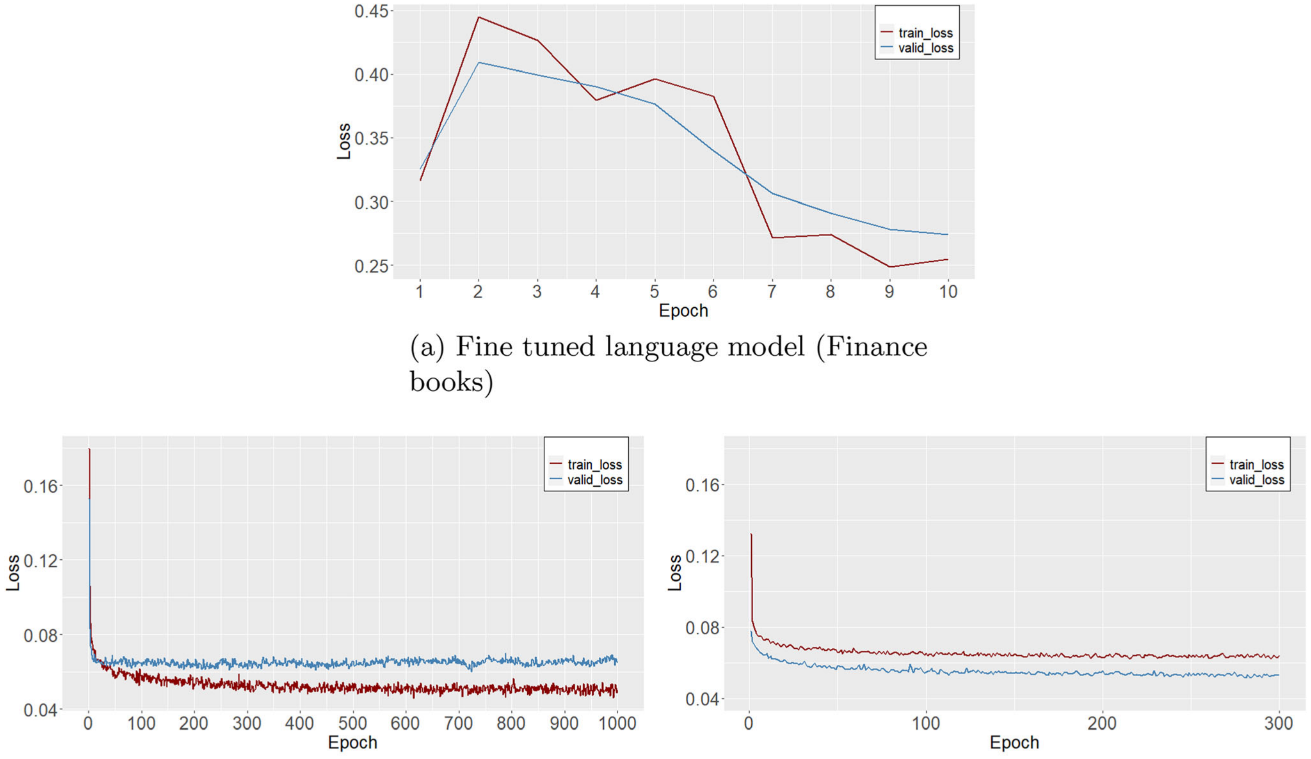


(a) Basic language model (Spanish Wikipedia)

(b) Fine tuned language model (Finance books)

(c) Basic regression model

(d) Generative data augmentation



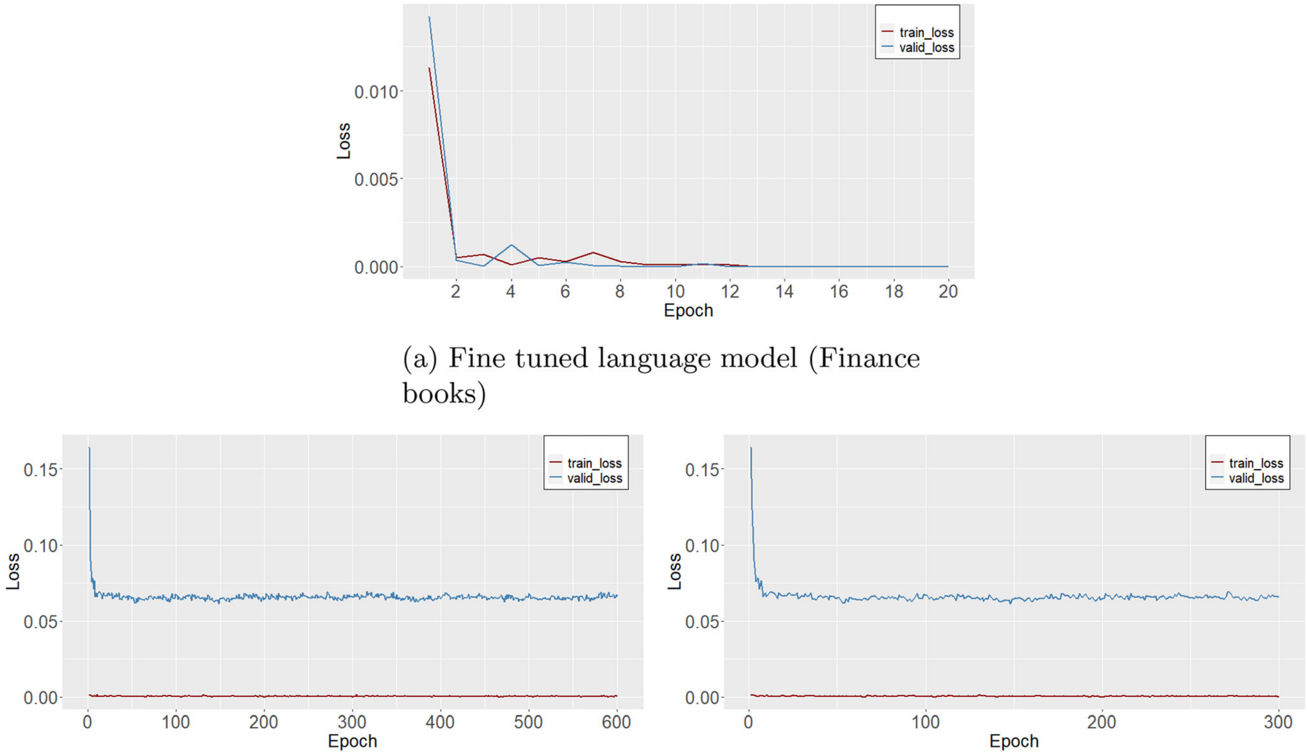
(a) Fine tuned language model (Finance books)

(b) Basic regression model

(c) Generative data augmentation

Fig. 5 BERT training metrics for phases two, three and four

The innovation that we propose in this work consists of using the excellent text generative capacity of LLMs to introduce variability in our dataset by generating iso-entropic sentences. For each sentence of each biography in our dataset, a variable number of iso-entropic sentences are generated that contain the same information but expressed in very different ways. This set of sentences will be used as a generative data augmentation technique to improve training, reduce computational requirements, minimize overfitting and maximize the accuracy of the resulting models. To the best of the authors' knowledge, this is the first time that LLM models have been used to improve the training of finance specialized DL systems such as the one proposed in this work.



(a) Fine tuned language model (Finance books)

(b) Basic regression model

(c) Generative data augmentation

Fig. 6 RoBERTa training metrics for phases two, three and four

Both ChatGPT and Llama2 7B were used as paraphrasing engines for the CV dataset. Both models were instructed with a system prompt to rewrite each biography as differently as possible from the original one, while maintaining the whole meaning and context. This provided us with a new dataset, effectively tripling the size of the original one and taking advantage of the two LLMs writing styles.

Finally, the three models were trained with this augmented dataset in a final phase. Figures 4d, 5c and 6c show the training metrics for each model with generative data augmentation, where early stopping was used in order to avoid overfitting. The AWD-LSTM was the one which benefited the most from the data augmentation, as the train and valid loss converge more quickly (Fig. 4d) and has less overfitting than the basic regression model (Fig. 4c). BERT follows with a slight enhancement, while RoBERTa shows little to none improvement in the training phase.

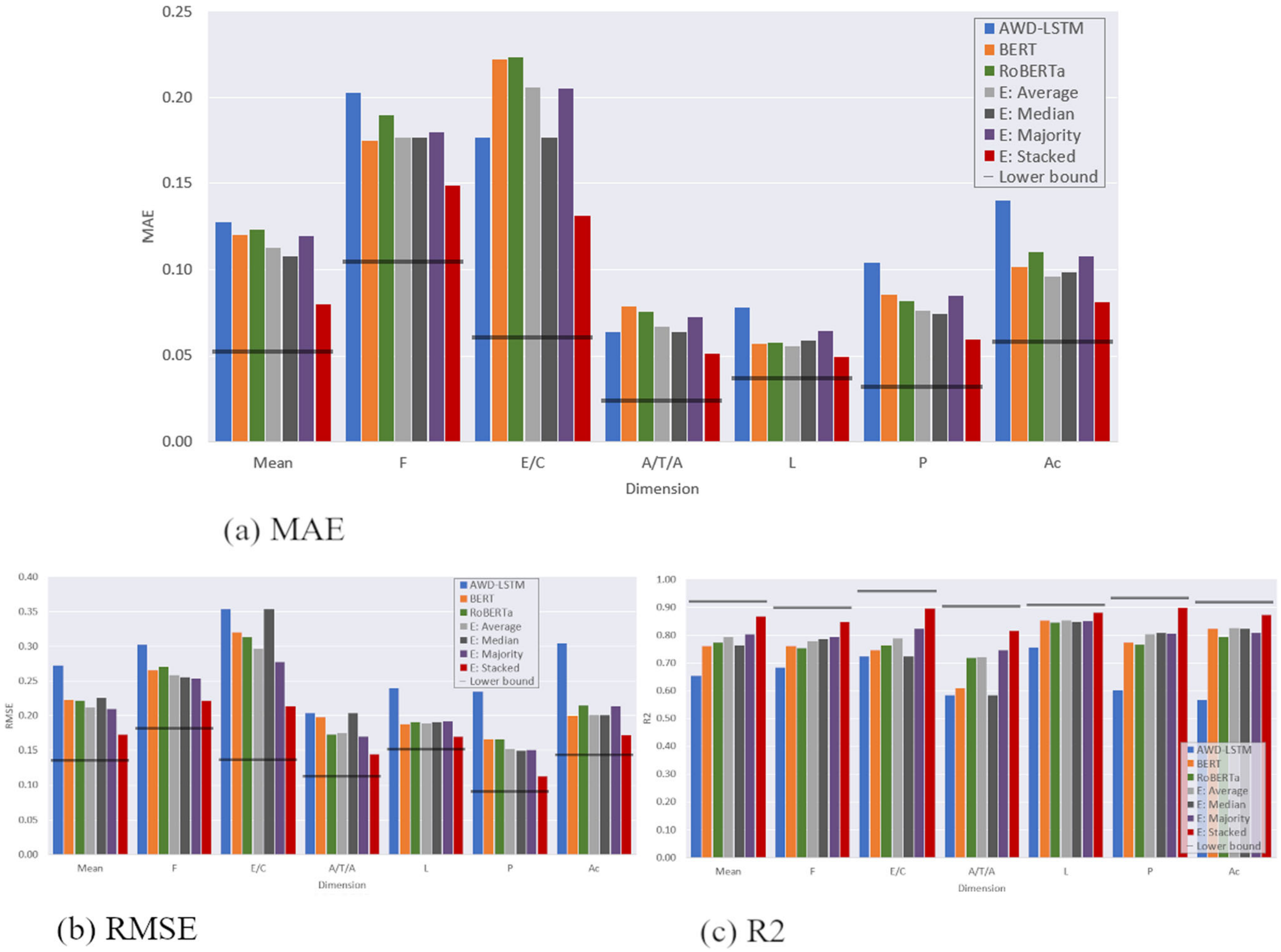
3.4 Test phase

The test dataset consists of 100 profiles that were initially separated from the labeled dataset. This test set was randomly selected maintaining the percentage of significant values among the different dimensions and was not used in the training phase. The aim of this procedure is to ensure that the test study significantly shows the generalization capacity of the models developed in this work. It should be remembered that the labeling carried out in Sect. 3.1 is basically qualitative in nature despite having been expressed in numerical form and that, therefore, this test study should also be carried out qualitatively.

Inference has been performed on 100 profiles with six categories for each of them and the results are compared with the labeling using three different error metrics. The RMSE was the loss function used during training. The MAE is easily interpreted as the average deviation of the predictions from the true values. Finally, the R-Squared is the proportion of explained variance of the data by the model (1 being a perfect fit).

In order to isolate the effect of generative data augmentation technique, firstly, the results achieved in the first three training phases are presented, as described in the Fig. 3. Subsequently, the results achieved by performing fine tuning of the models using the generative data augmentation described in Sect. 3.3.1 are shown.

Figure 7 shows the results for the proposed AWD-LSTM, BERT, RoBERTa and the four ensemble models. Also, the best possible hypothetical ensemble case has also been modeled, the one in which the model closest to the truth represented by the labeling is always chosen. This result is named *Lower bound* and is represented as a mark



(a) MAE

(b) RMSE

(c) R2

Fig. 7 Test study of the three base models and the four ensembles generated

in the results bar. It can be seen that, both for the three base models and the four ensembles, the MAE is higher in the F (Financial) and E/C (Executive/Consultant) dimensions than in the rest. That is, our models identify more accurately the Audit/Tax/Accountant (A/T/A), Legal (L), Political (P) and Academic (Ac) profiles. This result is corroborated in Fig. 7b, where it is observed that the RMSE is lower for these four categories, i.e., there are fewer outliers and/or the magnitude of their deviation is smaller. For the R^2 metric the differences between the dimensions are less notable. As can be seen in Fig. 7c, the categories L, P and Ac are slightly more predictable than the rest of the categories.

The mean value of MAE ranges between 0.08 and 0.13 for the seven models, which shows that our proposal is well adapted to the classification of professional profiles, always taking into account the qualitative nature of this profiling. The AWD-LSTM model is the least accurate in MAE and the one that generates more outliers which results in a lower predictability value than the rest of the models (see Fig. 7). BERT and RoBERTa behave similarly with slight variations in some categories. The ensemble models developed manage to combine the best characteristics of the three base models. It can be observed that the best ensemble is stacking, with an average MAE of 0.08 and the worst is the Majority, with a value of 0.12. Stacking performs consistently better for all six dimensions, even with values very close to the ideal case (*Lower bound*). It is significant that three of the ensembles always improve the best base model, thus corroborating that the correct combination of the different base models works better than those models individually.

The accuracy of our models is limited by the small number of available labeled biographies, 943 available for the training phase and 100 for the testing phase. This is a known issue that limits the learning ability of neural networks. Emerging DL-based conversational systems have excellent language processing capabilities which can be applied to increase the size of our labeled dataset. In this work we have used generative LLM to increase the number of biographies available using its ability to rewrite a text while maintaining its basic meaning.

Using the new generated dataset, the three models, AWD-LSTM, BERT and RoBERTa and the four ensembles, were trained again. Figure 8 shows the comparative MAE only for the best ensemble, stacking. The first bar shows the stacking ensemble with the three base models trained in the first three phases (see Fig. 3) and the second bar shows the results when the models were fine-tuned with generative LLM data augmentation. It is observed that the ensemble with generative data augmentation does not show significant differences in any of the six dimensions. We think that the reason for not achieving a significant improvement is that the rewrite created by generative LLM does not generate enough variability either in syntactic structure or in vocabulary to get a significant improvement in the learning phase. Despite not significantly improving the MAE, it must be remembered that generative augmentation has greatly improved training, specifically by eliminating overfitting and increasing the convergence rate, as observed in Figs. 4 and 5. To achieve better results in the future, a very detailed design of experiments and control of the syntax and vocabulary generated by generative LLM will be required.

Based in the results shown, it can be concluded that the test study carried out demonstrates the feasibility of our proposal to generalize the extraction of board of directors profiles of Spanish listed companies.

4 Interpretability

Deep learning models often are referred to as black boxes due to the difficulty of interpreting their predictions. However, in the financial field of research, interpretability is highly valued as knowing the reasoning behind a prediction provides useful insights in decision making. In this section we perform an interpretability analysis of the DL models used in this work.

To achieve the objectives of interpretability, it is necessary to generate human-comprehensible arguments about how our DL model operates. In this work, the SHAP model is used to obtain values on the relevance of the words and phrases that make up the natural language in the model's decision. In other words, the phrases and phrase fragments that are most relevant in the output generated by the model are identified. This allows for an expert in the area of finance, without requiring deep computational knowledge, to argue and justify about the results obtained in the model in order to include them, with clarifying justifications, in scientific works specific to their area. Specifically, the work on explainability has allowed us to generate force diagrams for each inference made by the model. In these diagrams, for each of the six dimensions that make up the biographies, those fragments of phrases relevant to assigning a greater weight to the dimension are represented versus those that assign a lower weight to said dimension. These force diagrams provide simple and open insights into the DL model decision-making process.

To fully justify interpretability, we begin by presenting the mathematical foundations used in the SHAP analysis, followed by the results obtained with the dataset and the financial models, along with the insights generated.

4.1 Methods

We use the methodology SHAP [12] (SHapley Additive exPlanations) to analyze the interpretability of the three deep learning neural networks. This state-of-the-art explanation method is based both in the Shapley value and LIME [26] (Local Interpretable Model-agnostic Explanations). Shapley values come from the concept of game theory, which needs a *game* and its *players*. In the field of machine learning, we will consider our game as one individual outcome of a model and our players will be the independent variables included in the model. The Shapley

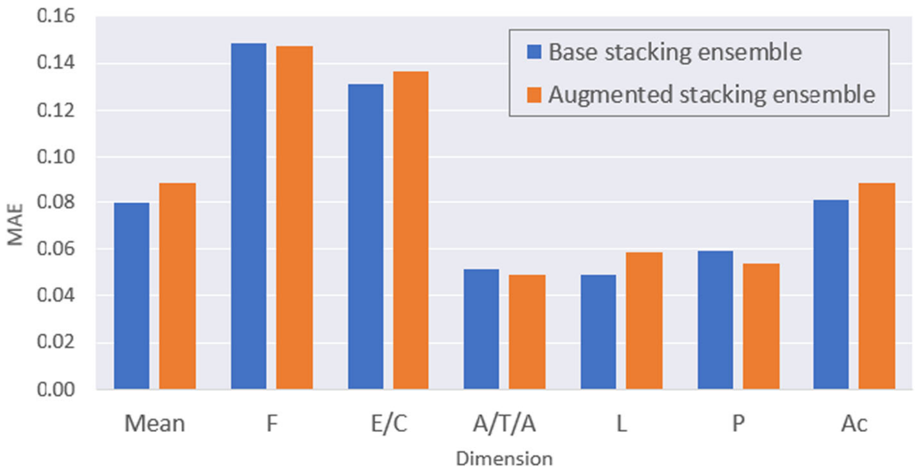


Fig. 8 MAE comparative analysis of the Base Stacking Ensemble vs the Augmented Stacking Ensemble

values are used to quantify the contribution that each independent variable adds to every individual prediction generated by the model. In our models, the independent variables are the individual tokens (words) that conform each biography.

To obtain a measure of the importance of each variable, SHAP generates perturbations on the subsets of input features and analyzes the outputs and how they change based on those perturbations. This output value is called marginal contribution (MC) and is expressed in Eq. 1, where token is the variable to evaluate, set is the subset of input features including the token, set-token is the same subset without token and x is the input we feed to the neural net.

$$\mathrm{MC}_{\mathrm{token,set}}(x) = \mathrm{Pred}_{\mathrm{set}}(x) - \mathrm{Pred}_{\mathrm{set-token}}(x) \tag{1}$$

Let's say we feed an input x of three tokens (or independent variables) to our model: "*Alberto fue profesor*" ("*Albert was a professor*"). Each combination of tokens should be considered to compute the importance of each token. First of all, it is necessary to compute the power set with all the features, which gives us $2^n = 2^3 = 8$ coalitions, where n is the number of independent variables. Now, SHAP needs to infer eight predictive outputs, each one using a different coalition of variables. This can be visualized in Fig. 9, where each node contains the coalitions of tokens and a prediction for the Academic (Ac) attribute of our models. The value in the top part denoted by $\emptyset$ is merely the average value of all the possible inputs, or in other words, the prediction if the biography was empty.

For example, the marginal contribution for a set containing only the token "*profesor*" is obtained with Eq. 2:

$$\begin{aligned}\mathrm{MC}_{\mathrm{profesor},\{\mathrm{profesor}\}}(x) &= \mathrm{Pred}_{\{\mathrm{profesor}\}}(x) \\ &- \mathrm{Pred}_{\emptyset}(x) = 0.7 - 0.2 = 0.5\end{aligned} \tag{2}$$

With the previous results, the marginal contribution of each token can be calculated. For this purpose we have to consider all the edges linking two nodes so that the upper node does not contain the variable and the bottom one does contain it. Generalizing to the result in [12], to compute the SHAP value for any token in an input x we follow Eq. 3, where $|\mathrm{set}|$ is the number of tokens in the set and F is the total number of tokens.

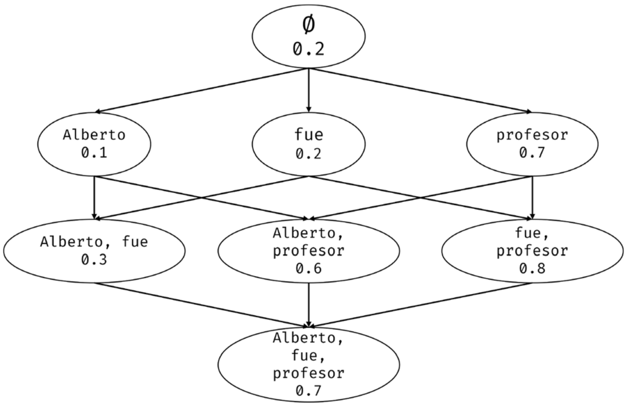


Fig. 9 Coalitions and Ac attribute outputs for the input "*Alberto fue profesor*"

$$\mathrm{SHAP}_{\mathrm{token}}(x) = \sum_{\mathrm{set:token}\in\mathrm{set}} \frac{1}{|\mathrm{set}|\binom{F}{|\mathrm{set}|}} \mathrm{MC}_{\mathrm{token,set}}(x) \tag{3}$$

Applying Eq. 3 to our example we compute the marginal contributions of "*profesor*" and therefore its SHAP value as expressed in Eq. 4:

$$\begin{aligned}\mathrm{SHAP}_{\mathrm{profesor}}(x) &= \frac{1}{1*\binom{3}{1}} * \mathrm{MC}_{\mathrm{profesor},\{\mathrm{profesor}\}}(x) \\ &+ \frac{1}{2*\binom{3}{2}} * \mathrm{MC}_{\mathrm{profesor},\{\mathrm{profesor,fue}\}}(x) \\ &+ \frac{1}{2*\binom{3}{2}} * \mathrm{MC}_{\mathrm{profesor},\{\mathrm{profesor,Alberto}\}}(x) \\ &+ \frac{1}{3*\binom{3}{3}} * \mathrm{MC}_{\mathrm{profesor},\{\mathrm{profesor,fue,Alberto}\}}(x) \\ &= \frac{1}{3} * 0.5 + \frac{1}{6} * 0.6 + \frac{1}{6} * 0.5 \\ &+ \frac{1}{3} * 0.4 = 0.4833\end{aligned} \tag{4}$$

This process is computationally intensive so we have used the algorithm developed in [12] which generates a good approximation with much less computing requirements.

4.2 Results

We have computed the Shapley values for all the CVs in the test set. In Table 1 we show the six most influential words for each attribute and model. It can be seen that the main tokens are directly related to activities compatible with the corresponding attribute. There are some surprising situations such as the influence of the word *profesor* in the attribute P (Politician) or *abogado* (lawyer) in the same attribute. These are indirect relationships which have been identified by some of our models, and that contribute to a more precise characterization of the attribute. In other words, our models have been able to detect that a significant percentage of politicians come from the field of law or academia.

Table 1 Six most important tokens in the test set

	AWD-LSTM	BERT	RoBERTa
F	BANCO	Banco	Banca
	Psicología	Capital	Financiero
	Capital	Inversiones	Inversiones
	Activos	Presidente	Banco
	Telefónica	Inmobiliario	Economista
	Finance	Bolsa	Bank
E/C	Presidente	Abogado	Profesor
	Filial	Inmobiliario	Despacho
	Obra	Banco	Contabilidad
	Energías	Presidente	Abogado
	Innovación	Despacho	Empresario
	Telefónica	Vicepresidente	Socio
A/T/A	FISCAL	Fiscal	Contabilidad
	Despachos	Auditoría	Fiscal
	ITT	Contabilidad	Auditoría
	Materia	Contables	Corporación
	Estate	Economista	Independiente
	Filiales	Tribu	Financiero
L	Abogado	Abogado	Abogado
	Estado	Despacho	Despacho
	Presidente	Derecho	Ejercicio
	Gubernamentales	Fiscal	Legal
	Ejercicio	Ejercicio	Derecho
	Ilustre	Auditoría	Economista
P	Inmobiliario	Ministro	Ministro
	Ministro	Economista	Alcalde
	Ahorros	Abogado	Gabinete
	Liderazgo	Presidenta	Profesor
	Innovación	Nacional	Gobierno
	Telefónica	Diputado	Congreso
Ac	PROFESOR	Profesor	Profesor
	Medicina	Economista	Docente
	Docente	Doctor	Doctor
	Universidades	Rector	Titular
	Doctor	Presidente	Universidad
	Amplia	Universidad	Ejercicio

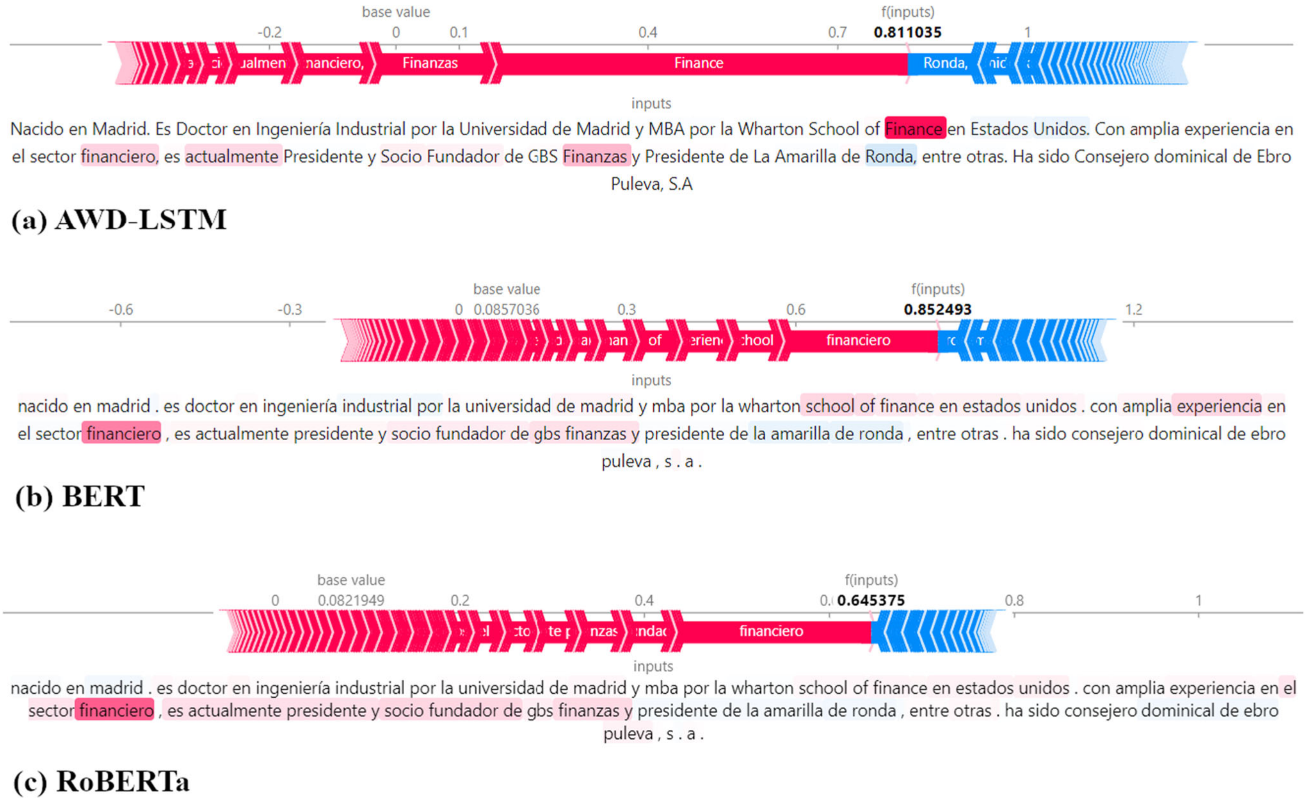
The models developed assign an implicit weight to each token to accurately profile the CV of each person on the board of directors. Using the described techniques it is possible to create force plots to represent the weight that each token has in the identification of an individual profile. This plot shows, for each attribute, the impact of all the tokens in the input and how they "push" the prediction toward 0 (blue) or 1 (red). The base value in the top part is the average value of all the inputs. We also highlight that the transformers inputs are lower cased. We have select several CVs with high values in the individual attributes and compute the SHAP values for all the tokens. Using these values, we can represent their force plots.

For a biography with a high value in attribute F (experience in Finance) as shown in Fig. 10, the models mainly use "finance" and "financial" as well as "experience" and "partner". These outputs seem to be reasonable and very human-like. For the characterization of the attribute E/C or Executive/Consultant (see Fig. 11) tokens such as "deputy", "enterprise", "leader" and "president" are marked in red. However, in the LSTM model, words like "Marshall" should not appear as they belong to a company name and do not imply leadership. As this is the attribute with the most variable biographies, it is expected to find some inaccuracies and slight overfit in the predictions.

Looking now into a directors' CV with an experience in Law (L attribute) the models seem to understand correctly that "lawyer", "practicing" and "law" are important, as shown in Fig. 12. In this case, the LSTM model slightly decreases its prediction due to "energies" being present and not being relevant. The attribute P (Politician) is represented by tokens such as "minister", "parliament" and "spokesperson" which are all reasonable and easily understood by the networks despite the input being long. The force plots are shown in Fig. 13. Finally, using an input with a high value in the Ac (Academic) attribute (see Fig. 14), words like "professor" or "technology" embrace all the importance.

As a summary, the interpretability study allows the identification of relevant phrases and generates the exact value of the positive and negative weights of each word and phrase in the quantification carried out for each biography in each of the six dimensions. In this way it is now possible to accompany each financial study with an annex justifying its results. This annex explains the evaluation carried out by the method through the use of numerical values and specific phrases, and thus facilitates its acceptance and dissemination. Additionally, explainability mitigates the problem previously presented in this work of lack of quality data, fundamentally due to the need for the data to be generated by a human expert. The reason is the increased reliability of the test phase because of the quantitative and specific reports justifying and explaining the results of the model. This can ensure that the model is based on relevant information and not on spurious or circumstantial data.

However, while being extremely valuable in corporative governance, this interpretability study presents a major challenge: scalability. The amount of time required to generate the insights for each token in a biography grows exponentially with the size of both the input and model, so the study is limited to neural networks with a smaller number of parameters in comparison with current LLMs.



Nacido en Madrid. Es Doctor en Ingeniería Industrial por la Universidad de Madrid y MBA por la Wharton School of Finance en Estados Unidos. Con amplia experiencia en el sector financiero, es actualmente Presidente y Socio Fundador de GBS Finanzas y Presidente de La Amarilla de Ronda, entre otras. Ha sido Consejero dominical de Ebro Puleva, S.A

(a) AWD-LSTM

nacido en madrid . es doctor en ingeniería industrial por la universidad de madrid y mba por la wharton school of finance en estados unidos . con amplia experiencia en el sector financiero , es actualmente presidente y socio fundador de gbs finanzas y presidente de la amarilla de ronda , entre otras . ha sido consejero dominical de ebro puleva , s . a .

(b) BERT

nacido en madrid . es doctor en ingeniería industrial por la universidad de madrid y mba por la wharton school of finance en estados unidos . con amplia experiencia en el sector financiero , es actualmente presidente y socio fundador de gbs finanzas y presidente de la amarilla de ronda , entre otras . ha sido consejero dominical de ebro puleva , s . a .

(c) RoBERTa

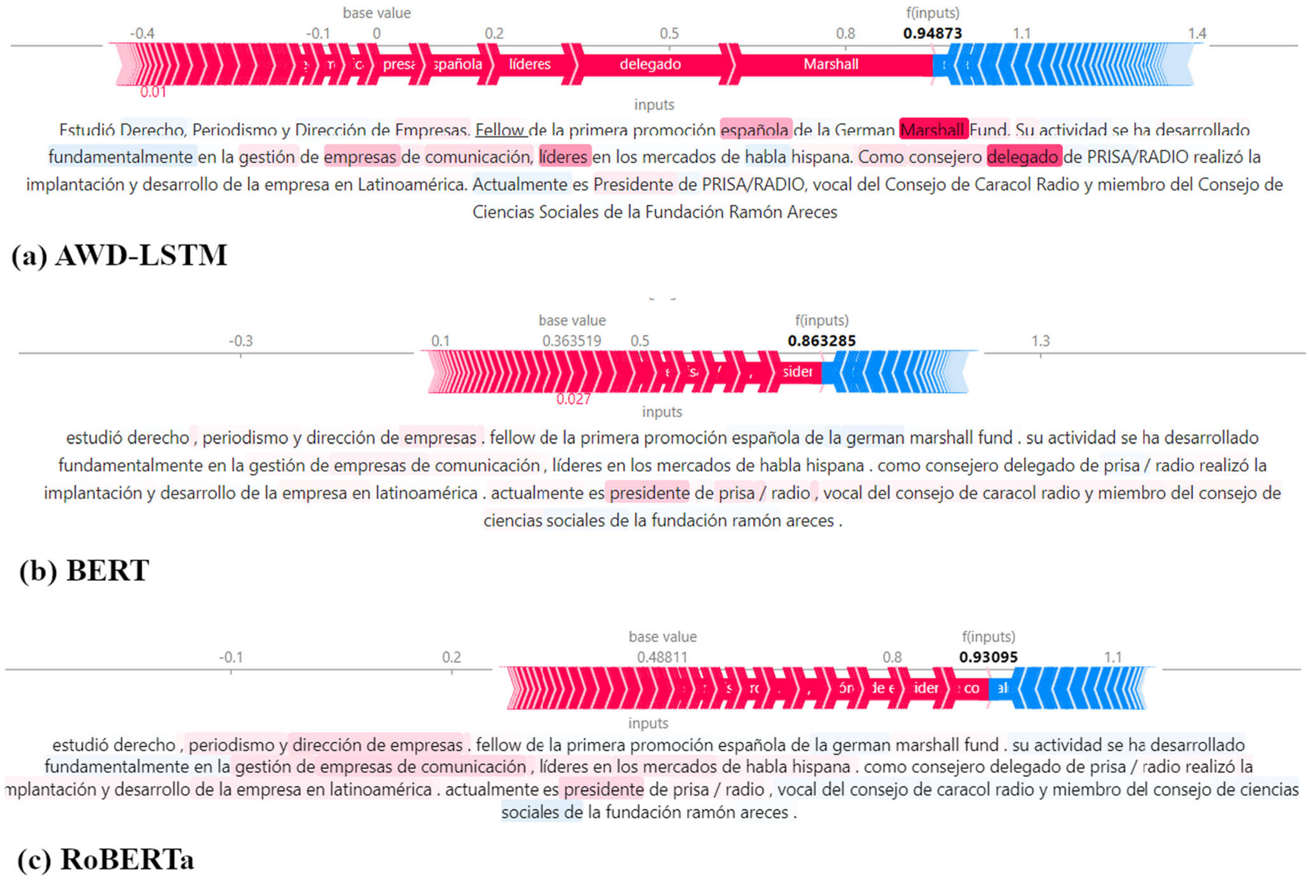
Fig. 10 Financial, F. Force plots for an input with attributes $[F = 1, Rest = 0]$

5 A use case: comparative analysis over time

Next, a case study is presented as an example of the application of the DL models developed to a complete set of 6564 professional profiles of independent directors of companies listed on the Spanish stock exchange. The board of directors is the main corporate governance body, responsible for the main strategic decisions of companies and helps ensure that the interests of shareholders and managers are closely aligned. Their functions include management control and expert advice, which fall to a large extent on the so-called independent directors, those who are neither executives of the company nor represent shareholders. The importance of these directors is reflected in their beneficial influence on company corporate governance and other features of the organization [27]. In fact, regulation, usually in the form of recommendations, encourages a greater presence of independent directors on the board of directors. In Spain, Recommendation 17 of the Good Governance Code of Listed Companies (2015) establishes that independent directors should make up at least half of the total number of directors [28].

In addition, the diversity of the board is one way to build an efficient board of directors because diversity brings more perspectives, knowledge and experience into the boardrooms, which enable the board to better exercise its functions. In this regard, much attention has recently been paid to board diversity in terms of age, gender or tenure. However, the profile or specialization of directors has received less attention, despite its importance in achieving the expert advice function. The reason for this lesser attention lies in the difficulty of obtaining the data, which being qualitative in nature, are not available in the databases used in finance research. This is where the application of DL models is extremely useful for economics and finance research. The analysis conducted allows us to understand the profile of the boards of directors according to the professional experience of each board member. As we pointed out in Sect. 3.1, we defined up to six professional profiles representing key skills for the company. They are as follows:

- Financial (F): Refers to those directors with experience in the financial sector, whether in banking institutions, any type of investment companies or the stock market in general.
- Executive/Consultant (E/C): Directors who have held or are currently holding different types of management positions in other companies or have carried out outstanding advisory tasks. These directors may have experience in different business sectors and management positions.



Fstudió Derecho, Periodismo y Dirección de Fmpresas. Fellow de la primera promoción española de la German Marshall Fund. Su actividad se ha desarrollado fundamentalmente en la gestión de empresas de comunicación, líderes en los mercados de habla hispana. Como consejero delegado de PRISA/RADIO realizó la implantación y desarrollo de la empresa en Latinoamérica. Actualmente es Presidente de PRISA/RADIO, vocal del Consejo de Caracol Radio y miembro del Consejo de Ciencias Sociales de la Fundación Ramón Areces

(a) AWD-LSTM

estudió derecho , periodismo y dirección de empresas . fellow de la primera promoción española de la german marshall fund . su actividad se ha desarrollado fundamentalmente en la gestión de empresas de comunicación , líderes en los mercados de habla hispana . como consejero delegado de prisa / radio realizó la implantación y desarrollo de la empresa en latinoamérica . actualmente es presidente de prisa / radio , vocal del consejo de caracol radio y miembro del consejo de ciencias sociales de la fundación ramón areces .

(b) BERT

estudió derecho , periodismo y dirección de empresas . fellow de la primera promoción española de la german marshall fund . su actividad se ha desarrollado fundamentalmente en la gestión de empresas de comunicación , líderes en los mercados de habla hispana . como consejero delegado de prisa / radio realizó la mplantación y desarrollo de la empresa en latinoamérica . actualmente es presidente de prisa / radio , vocal del consejo de caracol radio y miembro del consejo de ciencias sociales de la fundación ramón areces .

(c) RoBERTa

Fig. 11 Executive/Consultant, E/C. Force plots for an input with attributes $[E/C = 1, Rest = 0]$

- Audit/Tax/Accountant (A/T/A): In this case, these are directors with specific expertise in auditing, tax or accounting.
- Legal (L): Lawyers and legal experts are classified in this category.
- Political (P): Refers to directors who have held or are holding public offices of various kinds, especially political posts.
- Academic (Ac): Finally, this refers to those directors with academic experience.

Almost all the skills identified in the bios can be assigned to one of the six categories. Some are more general (Executive/Consultant) and could be broken down, but we have chosen not to do so because given the sample size, the number of cases would be too small.

The previous literature highlights the importance of the dimensions considered in this paper. Güner et al. [29] and Booth and Deli [30] document that board members with financial expertise significantly influence firm financing and investment decisions. Likewise, according to the literature, independent directors help enhance firm value with their industry experiences [31, 32] documents positive market response to the appointment of successful executives of other companies. Moreover, firms with accounting expert sitting on their audit committees show stronger accounting conservatism ([33]). Regarding legal dimension, its importance is highlighted by works such as that of [34], which shows that the presence of directors with legal backgrounds is associated with higher financial reporting quality. [35] find that the cost of bank loans is significantly lower for companies with board members with political ties and [36] point to the prevalence of directors with background in politics for companies with significant government contracts. Finally, some papers highlight that academic directors play an important governance role through their advisory and supervisory functions, leading to increased R&D performance and investment ([37]).

In our model each entry is a biographical description in free text format as described in Sect. 3.1. The inference process of the three proposed DL models plus the four ensemble models has been performed on this data repertoire. The results obtained should be interpreted qualitatively under the parameters obtained in the test study of Sect. 3.4. In Sect. 4 an interpretability study was performed to know the mechanisms on which each model is based for the inference of free text sentiment describing the professional profile. The test analysis coupled with the

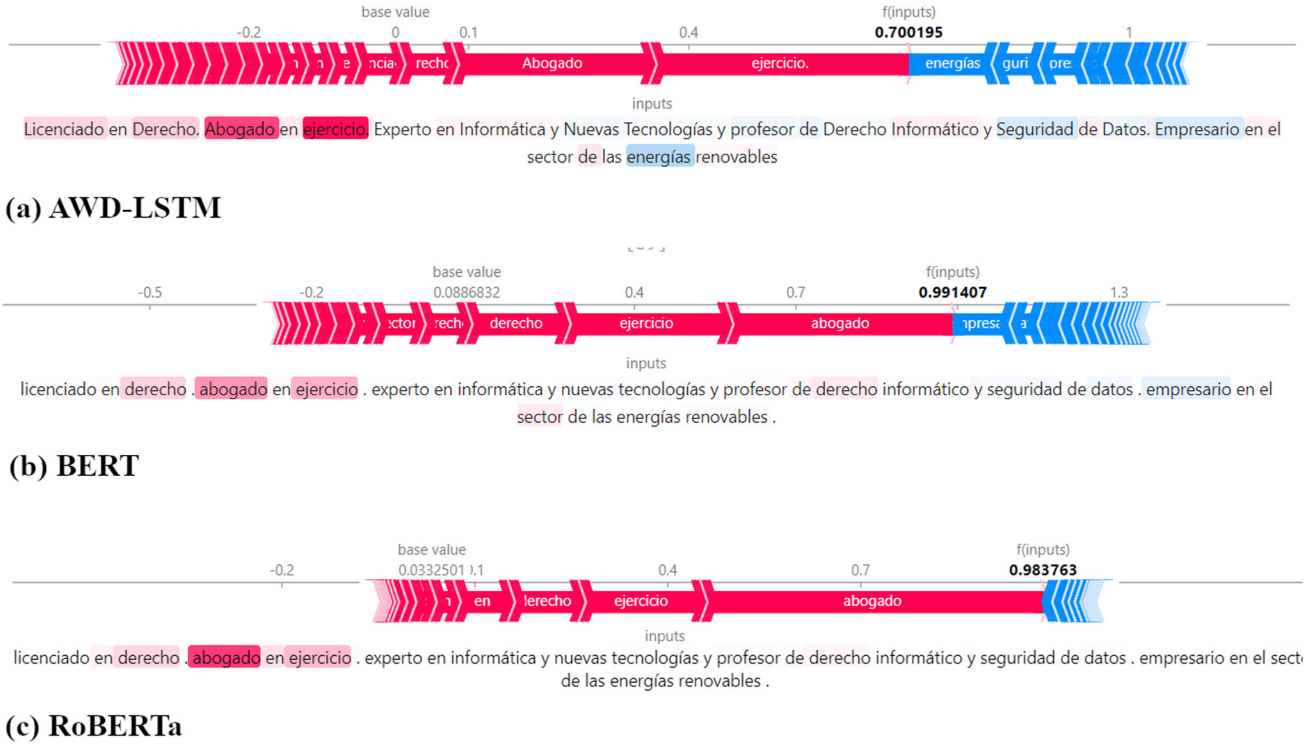


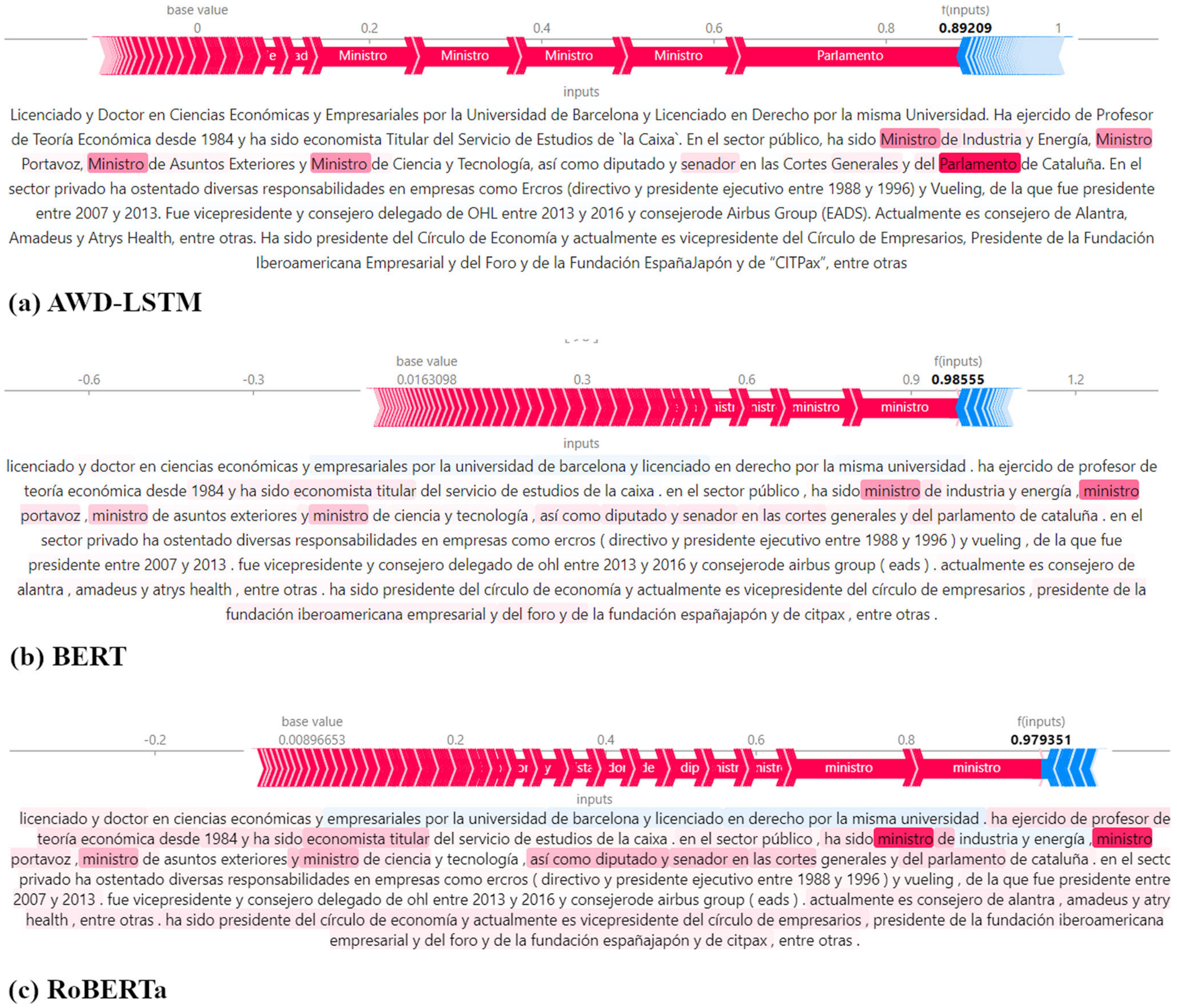
Fig. 12 Legal, L. Force plots for an input with attributes $[L = 1, Ac = 0.6, Rest = 0]$

interpretability study provides tools to improve the understanding of the results obtained.

We propose as the main measurement index the Standard Expertise Profile (SEP) which expresses the expertise in each of the six dimensions of the board of directors for each company. It is calculated as the average of the expertise of each director in each of the six dimensions and, therefore, it is a vector of six positions of numerical values in the interval [0,1]. The SEP index represents the professional orientation of a company's decision-makers and, is therefore, directly linked to its corporate policy strategy. Figure 15 shows the time evolution of the SEP for all companies. Figure 15a plots the yearly mean SEP for all companies over the period (2004, 2020). Figure 15b shows the mean SEP grouped by relevant periods. Three periods have been selected, precrisis in the interval (2004, 2007), crisis in the interval (2008, 2013) and post-crisis in the interval (2014, 2020). These data have been generated using the ensemble described in Sect. 3.2.

The average number of independent directors per company at the beginning of the period analyzed stands at 3.6. After decreasing slightly during the first years, between 2008 and 2013 it experiences a certain stability, before starting an upward trend that continues until the end of the period, in which the value stands at 3.9. Figure 15a shows the average director profile of the companies in each year, represented by the SEP index. There is a very significant increase in the number of expertise units in the E/C profile (directors who have held or are currently holding different types of management positions in other companies or have carried out outstanding advisory tasks), starting with 0.47 in 2004 and reaching more than 0.6 expertise units in 2020. Therefore, companies are demanding profiles that provide management experience and knowledge. Noteworthy, although to a lesser extent, is the increase in the A/T/A profile (directors with specific expertise in auditing, tax or accounting), as well as the decrease in the L profile (lawyers and legal experts), although the latter two are at a much lower level than the former. The rest of the profiles show a certain stability over the period analyzed.

The study period is then grouped into three stages (Fig. 15b): precrisis, crisis and post-crisis. With respect to the first profile, contrary to what might be expected, the financial crisis does not seem to have provided an incentive to increase the specialization in finance of the boards, which hardly increases. However, as indicated above, boards are strengthening their E/C profile and, to a lesser extent, their A/T/A profile. There is also a significant decrease in the legal and political profiles (the latter mainly in the post-crisis period). It seems that companies no longer see the value of maintaining a political profile after the financial crisis. Finally, the academic profile remains constant, with a small increase in the crisis stage, which is maintained thereafter.



Licenciado y Doctor en Ciencias Económicas y Empresariales por la Universidad de Barcelona y Licenciado en Derecho por la misma Universidad. Ha ejercido de Profesor de Teoría Económica desde 1984 y ha sido economista Titular del Servicio de Estudios de `la Caixa`. En el sector público, ha sido Ministro de Industria y Energía, Ministro Portavoz, Ministro de Asuntos Exteriores y Ministro de Ciencia y Tecnología, así como diputado y senador en las Cortes Generales y del Parlamento de Cataluña. En el sector privado ha ostentado diversas responsabilidades en empresas como Ercros (directivo y presidente ejecutivo entre 1988 y 1996) y Vueling, de la que fue presidente entre 2007 y 2013. Fue vicepresidente y consejero delegado de OHL entre 2013 y 2016 y consejerode Airbus Group (EADS). Actualmente es consejero de Alantra, Amadeus y Atrys Health, entre otras. Ha sido presidente del Círculo de Economía y actualmente es vicepresidente del Círculo de Empresarios, Presidente de la Fundación Iberoamericana Empresarial y del Foro y de la Fundación EspañaJapón y de "CITPax", entre otras

(a) AWD-LSTM

licenciado y doctor en ciencias económicas y empresariales por la universidad de barcelona y licenciado en derecho por la misma universidad . ha ejercido de profesor de teoría económica desde 1984 y ha sido economista titular del servicio de estudios de la caixa . en el sector público , ha sido ministro de industria y energía , ministro portavoz , ministro de asuntos exteriores y ministro de ciencia y tecnología , así como diputado y senador en las cortes generales y del parlamento de cataluña . en el sector privado ha ostentado diversas responsabilidades en empresas como ercros (directivo y presidente ejecutivo entre 1988 y 1996) y vueling , de la que fue presidente entre 2007 y 2013 . fue vicepresidente y consejero delegado de ohl entre 2013 y 2016 y consejerode airbus group (eads) . actualmente es consejero de alantra , amadeus y atrys health , entre otras . ha sido presidente del círculo de economía y actualmente es vicepresidente del círculo de empresarios , presidente de la fundación iberoamericana empresarial y del foro y de la fundación españajapón y de citpax , entre otras .

(b) BERT

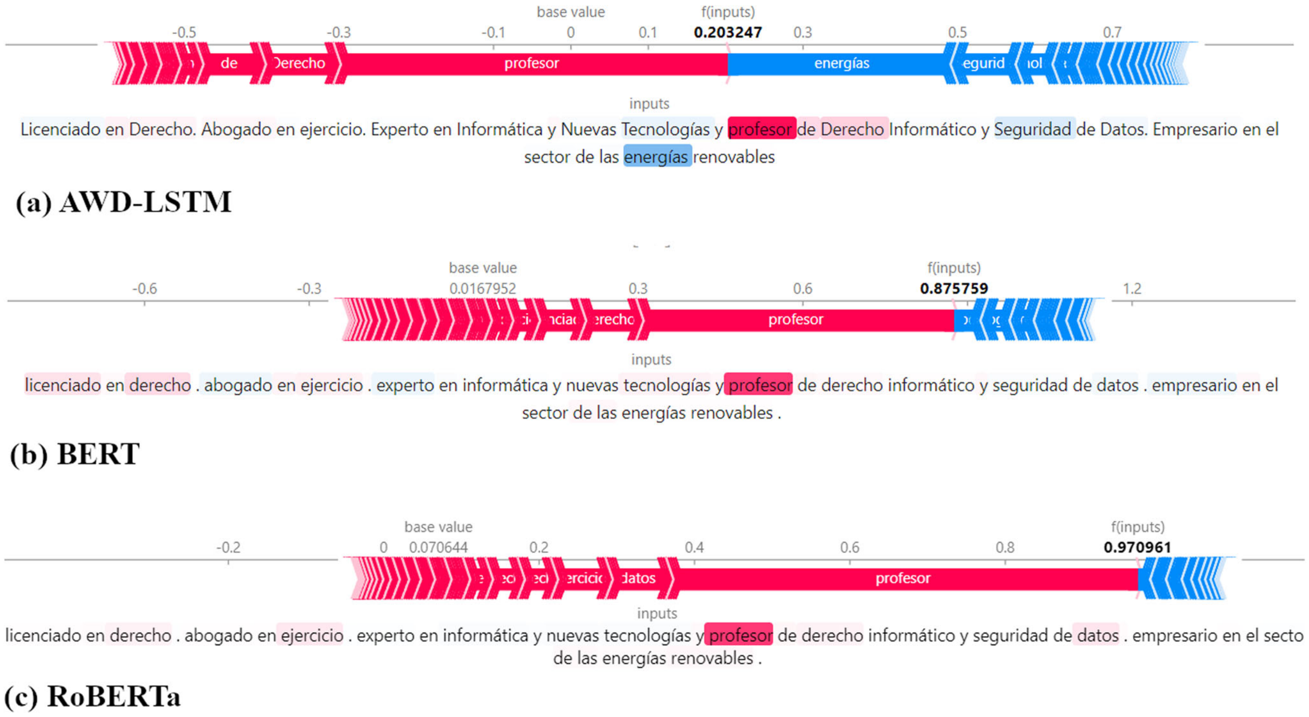
licenciado y doctor en ciencias económicas y empresariales por la universidad de barcelona y licenciado en derecho por la misma universidad . ha ejercido de profesor de teoría económica desde 1984 y ha sido economista titular del servicio de estudios de la caixa . en el sector público , ha sido ministro de industria y energía , ministro portavoz , ministro de asuntos exteriores y ministro de ciencia y tecnología , así como diputado y senador en las cortes generales y del parlamento de cataluña . en el secto privado ha ostentado diversas responsabilidades en empresas como ercros (directivo y presidente ejecutivo entre 1988 y 1996) y vueling , de la que fue presidente entre 2007 y 2013 . fue vicepresidente y consejero delegado de ohl entre 2013 y 2016 y consejerode airbus group (eads) . actualmente es consejero de alantra , amadeus y atry health , entre otras . ha sido presidente del círculo de economía y actualmente es vicepresidente del círculo de empresarios , presidente de la fundación iberoamericana empresarial y del foro y de la fundación españajapón y de citpax , entre otras .

(c) RoBERTa

Fig. 13 Politician, P. Force plots for an input with attributes $[P = 1, E/C = 1, Ac = 0.3, Rest = 0]$

6 Conclusions

In this work we have developed and trained three DL models based on recurrent networks (AWD-LSTM) and transformers (BERT) to characterize, in six dimensions, the free text career trajectory of Spanish listed companies' independent directors. Additionally, four ensembles have been implemented, three of them based on standard metrics. The fourth ensemble, stacking, is an innovative proposal based on an MLP trained to choose the best characteristics of each base model in each dimension. A comprehensive test analysis has also been performed to show the validity of our proposal, obtaining an MAE value of 8% for the best ensemble combining the LSTM and BERT models, on a dataset completely independent of the one used in the training phase. Additionally, fine tuning training has been performed using generative LLM data augmentation. The results show that training with augmented data is a promising technique, but requires additional work to achieve significant improvements. In addition, a complete interpretability study has been carried out using the SHAP methodology to identify the basic information used by each model to generate inferences for each biography. This study allows us to know the way in which the models use the tokens in the inference process, thus improving the explainability of the results. Finally, a case study has been constructed using the DL models applied to a corpus of 6564 biographies. This study shows a clear increase in the E/C profile in the period (2004, 2020) both in the study by years and in the study by periods. There is also an increase in the A/T/A profile and a reduction in the L and P profiles. It can be concluded that

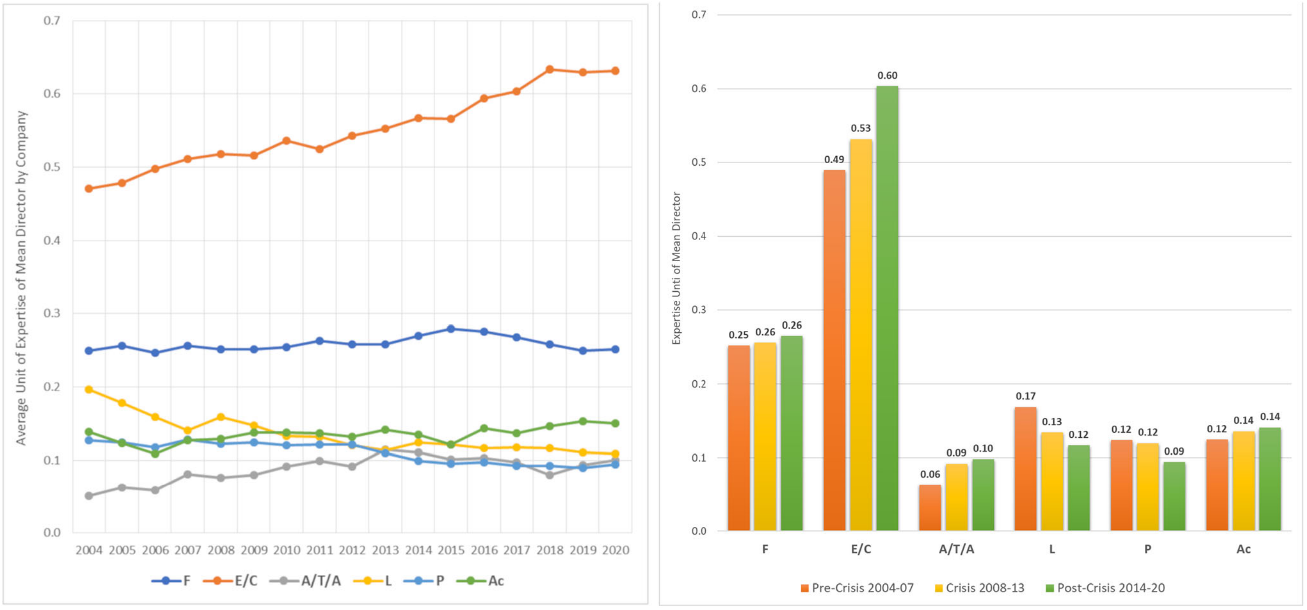


(a) AWD-LSTM

(b) BERT

(c) RoBERTa

Fig. 14 Academic, Ac. Force plots for an input with attributes $[Ac = 0.6, L = 1, Rest = 0]$



(a) Anual

(b) By Periods

Fig. 15 Standard Expertise Profile (SEP) over time. Each SEP item is composed of the numerical values of the six dimensions at a specific time

the work carried out shows the feasibility of DL techniques for the accurate analysis of biographies by demonstrating very reduced MAE and providing mechanisms to improve the explainability of inferences. Our DL techniques generate useful knowledge from unstructured data that could not have been obtained by other means.

Acknowledgements We thank the company AMC Natural Drinks for their help in carrying out this project.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Open access funding provided by Universidad de Valladolid within the CRUE-CSIC Transformative Agreement.

Availability of data and materials The datasets generated and analyzed during the current study are available in the github repository, https://github.com/CesarVacaR/InterpretabilityDLSpanishFinancialText.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Conneau A, Schwenk H, Cun YL, Barrault L (2017) Very deep convolutional networks for text classification. In: Proceedings of 15th conference of the European chapter of the association for computational linguistics 2. https://doi.org/10.18653/v1/e17-1104
2. Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. arXiv:1801.06146
3. Lyu S, Liu J (2021) Convolutional recurrent neural networks for text classification. J Datab Manag 32:65–82. https://doi.org/10.4018/JDM.2021100105
4. Radford A, Jozefowicz R, Sutskever I (2017) Learning to generate reviews and discovering sentiment. arXiv:1704.01444
5. Mnassri K, Rajapaksha P, Farahbakhsh R, Crespi N (2022) BERT-based ensemble approaches for hate speech detection. arXiv:2209.06505
6. Vaca C, Tejerina F, Sahelices B (2022) Board of directors' profile: a case for deep learning as a valid methodology to finance research. Int J Interact Multimedia Artif Intell 7:60. https://doi.org/10.9781/ijimai.2022.09.005
7. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. Naacl-Hlt https://doi.org/10.18653/v1/n19-1423
8. Merity S, Keskar NS, Socher R (2018) Regularizing and optimizing LSTM language models. https://openreview.net/forum?id=SyyGPP0TZ
9. Nkabiti KP, Chen Y (2021) Application of solely self-attention mechanism in CSI-fingerprinting-based indoor localization. Neural Comput Appl. https://doi.org/10.1007/s00521-020-05681-1
10. Xia D, Yang N, Jiang S, Hu Y, Li Y, Li H, Wang L (2022) A parallel NAW-DBLSTM algorithm on spark for traffic flow forecasting. Neural Comput Appl. https://doi.org/10.1007/s00521-021-06409-5
11. Colasanto F, Grilli L, Santoro D, Villani G (2022) BERT's sentiment score for portfolio optimization: a fine-tuned view in Black and Litterman model. Neural Comput Appl. https://doi.org/10.1007/s00521-022-07403-1
12. Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. arXiv:1705.07874
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Łukasz Kaiser, Polosukhin I (2017) Attention is all you need. arXiv:1706.03762
14. Rizwan M, Mushtaq MF, Akram U, Mehmood A, Ashraf I, Sahelices B (2022) Depression classification from tweets using small deep transfer learning language models. IEEE Access 10:129176–129189. https://doi.org/10.1109/ACCESS.2022.3223049
15. Giudici P (2018) Fintech risk management: a research challenge for artificial intelligence in finance. Front Artif Intell. https://doi.org/10.3389/frai.2018.00001
16. Tadapaneni NR (2020) Artificial intelligence in finance and investment. Int J Innov Res Sci Eng Technol (IJIRSET) 9(5)
17. Li K, Mai F, Shen R, Yan X (2021) Measuring corporate culture using machine learning. Rev Financ Stud 34:3265–3315. https://doi.org/10.1093/rfs/hhaa079
18. Tjoa E, Guan C (2021) A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2020.3027314
19. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR (2019) Unmasking clever Hans predictors and assessing what machines really learn. Nat Commun. https://doi.org/10.1038/s41467-019-08987-4
20. Peters ME, Neumann M, Zettlemoyer L, Yih WT (2018) Dissecting contextual word embeddings: architecture and representation. https://doi.org/10.18653/v1/d18-1179
21. Canete J, Chaperon G, Fuentes R, Ho J-H, Kang H, Pérez J (2020) Spanish pre-trained BERT model and evaluation data. Pml4dc at iclr 2020, 1–10
22. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized BERT pretraining approach. arXiv:1907.11692
23. OpenAI: ChatGPT (2023). https://openai.com/blog/chatgpt
24. OpenAI: GPT-4 Technical Report (2023)
25. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux M-A, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T (2023) Llama 2: open foundation and fine-tuned chat models
26. Ribeiro MT, Singh S, Guestrin C (2016) "Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144
27. Hermalin BE, Weisbach MS (2003) Boards of directors as an endogenously determined institution: a survey of the economic literature. FRNBY Policy Rev 9(1):7–26. https://doi.org/10.2139/ssrn.233111
28. CNMV: Código de buen gobierno de las sociedades cotizadas (2015). https://www.cnmv.es/DocPortal/Publicaciones/CodigoGov/CBG_2020.pdf
29. Güner B, Malmendier U, Tate G (2018) Financial expertise of directors. J Financ Econ. https://doi.org/10.1016/j.jfineco.2007.05.009
30. Booth JR, Deli DN (1999) On executives of financial institutions as outside directors. J Corp Finance 5:227–250. https://doi.org/10.1016/S0929-1199(99)00004-8

31. Faleye O, Hoitash R, Hoitash U (2018) Industry expertise on corporate boards. Rev Quant Financ Acc 50:441–2479. https://doi.org/10.1007/s11156-017-0635-z
32. Fich EM (2005) Are some outside directors better than others? Evidence from director appointments by fortune 1000 firms. J Bus 78:1943–1972. https://doi.org/10.1086/431448
33. Qiao Z, Chen KY, Hung S (2018) Professionals inside the board room: accounting expertise of directors and dividend policy. Appl Econ 50:6100–6111. https://doi.org/10.1080/00036846.2018.1489501
34. Krishnan J, Wen Y, Zhao W (2011) Legal expertise on corporate audit committees and financial reporting quality. Account Rev 86:2099–2130. https://doi.org/10.2308/accr-10135
35. Houston JF, Jiang L, Lin C, Ma Y (2014) Political connections and the cost of bank loans. J Account Res 52:193–243. https://doi.org/10.1111/1475-679X.12038
36. Agrawal A, Knoeber CR (2001) Do some outside directors play a political role? J Law Econ 44:179–198. https://doi.org/10.1086/320271
37. Xie Y, Xu J, Zhu R (2021) Academic directors and corporate innovation. SSRN Pap. https://doi.org/10.2139/ssrn.3954290

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