electronics

MDPI

Article

Concatenation Augmentation for Improving Deep Learning Models in Finance NLP with Scarce Data

César Vaca [1,*], Jesús-Ángel Román-Gallego [2], Verónica Barroso-García [3,4], Fernando Tejerina [5] and Benjamín Sahelices [6]

[1] Department of Informatics, University of Valladolid, Paseo de Belén s/n, 47011 Valladolid, Spain
[2] Escuela Politécnica Superior de Zamora, University of Salamanca, Avda Requejo 33, 49022 Zamora, Spain; zjarg@usal.es
[3] Biomedical Engineering Group, Department of Theory of Signal and Communications and Telematic Engineering, University of Valladolid, Paseo de Belén s/n, 47011 Valladolid, Spain; veronica.barroso@uva.es
[4] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), 47011 Valladolid, Spain
[5] Department of Financial Economics and Accounting, University of Valladolid, Paseo de Belén s/n, 47011 Valladolid, Spain; ftejerina@uva.es
[6] Grupo de Caracterización de Materiales y Dispositivos Electrónicos (GIR-GCME), Department of Informatics, University of Valladolid, Paseo de Belén s/n, 47011 Valladolid, Spain; benjamin.sahelices@uva.es
* Correspondence: cesar.vaca@uva.es

Abstract: Nowadays, financial institutions increasingly leverage artificial intelligence to enhance decision-making and optimize investment strategies. A specific application is the automatic analysis of large volumes of unstructured textual data to extract relevant information through deep learning (DL) methods. However, the effectiveness of these methods is often limited by the scarcity of high-quality labeled data. To address this, we propose a new data augmentation technique, Concatenation Augmentation (CA). This is designed to overcome the challenges of processing unstructured text, particularly in analyzing professional profiles from corporate governance reports. Based on Mixup and Label Smoothing Regularization principles, CA generates new text samples by concatenating inputs and applying a convex additive operator, preserving its spatial and semantic coherence. Our proposal achieved hit rates between 92.4% and 99.7%, significantly outperforming other data augmentation techniques. CA improved the precision and robustness of the DL models used for extracting critical information from corporate reports. This technique offers easy integration into existing models and incurs low computational costs. Its efficiency facilitates rapid model adaptation to new data and enhances overall precision. Hence, CA would be a potential and valuable data augmentation tool for boosting DL model performance and efficiency in analyzing financial and governance textual data.

Keywords: board of directors profiling; concatenation augmentation; data augmentation; deep learning; finance; long short-term memory

Academic Editor: Chunping Li

Received: 31 March 2025
Revised: 27 May 2025
Accepted: 3 June 2025
Published: 4 June 2025

Citation: Vaca, C.; Román-Gallego, J.-Á.; Barroso-García, V.; Tejerina, F.; Sahelices, B. Concatenation Augmentation for Improving Deep Learning Models in Finance NLP with Scarce Data. *Electronics* **2025**, *14*, 2289. https://doi.org/10.3390/electronics14112289

1. Introduction

In the digitalization and artificial intelligence era, financial institutions and their governing bodies face the challenge of adapting to new methodologies that enhance decision-making and optimize investment strategies [1]. Among these innovations, natural language processing (NLP) has emerged as an essential tool, enabling analysts and researchers to efficiently examine large volumes of textual information, such as professional profiles or biographies (CVs), financial reports, and relevant publications [2–5]. This capability is

particularly valuable for uncovering the impact of independent directors on governance and corporate strategy [6,7].

A key resource in Spain for this task is the set of Annual Corporate Governance Reports [8] issued by the National Securities Market Commission (Comisión Nacional del Mercado de Valores, CNMV), which is available at https://www.cnmv.es/ (date of access to data 10 November 2022). These freely accessible reports provide a detailed explanation of the structure and operation of the entity's governing bodies. Among other content, it includes information about the shareholder composition and the management bodies, related-party transactions, risk control systems, and the operating of the board, as well as the CVs of the members of the board of directors and senior management. Among the different types of directors, we focus on the independent ones. Since these directors are typically selected for their expertise in specific areas, their knowledge has the potential to significantly impact company decisions. Therefore, analyzing their profiles is crucial to understanding how their experience and skills can influence the strategies adopted by the company and identifying patterns or trends in decision-making (e.g., directors with a strong financial background could lean toward strategies that prioritize financial stability, while those with an academic profile could promote investment in innovation and development).

One of the issues is that these reports are usually formed by unstructured and heterogeneous information. When processing this type of unstructured texts, we find variability in its format and content, which makes the standardization process difficult [2]. Relevant information is often implicit or scattered across various sections of the CV, requiring advanced methods to extract and consolidate this data. Additionally, CVs can contain typos, inconsistent terminology, and even contextual and semantic ambiguities, adding another layer of complexity to the analysis process. All of this makes the extraction of information a complex task, which is why a system is needed that helps automatically analyze the information contained in these CVs and extract that which is truly relevant.

In this regard, deep learning (DL), an approach that has revolutionized the field of NLP, allows for more sophisticated modeling and analysis of these reports, extracting complex patterns and correlations that go beyond the limited capabilities of traditional methods [9]. Its ability to identify and learn data representations at multiple abstraction levels makes DL techniques, such as Bidirectional Encoder Representations of Transformers (BERT) and Long Short-Term Memory (LSTM) neural networks [10–12], widely used to improve the contextual and semantic understanding of the texts. These techniques usually adopt significant modeling flexibility, featuring a wide range of parameters. With such extensive modeling capabilities, DL models require efficient regularization methods to prevent overfitting and promote generalization [13]. Consequently, techniques such as weight decay [14], dropout [15], batch normalization [16], and data augmentation schemes [17] are commonly used for this purpose. However, one of the most significant limitations of its application in this context is the scarcity of high-quality labeled data [18]. Unlike other fields where data is abundant and well-structured, data from CVs is heterogeneous and limited, making it challenging to effectively train and optimize DL models.

In order to address this inherent limitation, several data augmentation techniques have been explored to enrich the available dataset and thus improve the robustness and performance of the models [19–21]. Token-level and sentence-level augmentation applied in NLP include the use of synonym and word replacement, translation and back-translation, and paraphrasing, compositional augmentation, conditional generation, and saliency map-based token span merge-replacement [22–28]. Adversarial data augmentation uses adversarial training to generate text with greater variability but is semantically indistinguishable [29]. Finally, hidden-space augmentation generates variations in the text by manipulating its representation in internal space, for example by adding noise or

interpolations between different data points [30]. While these techniques are valuable, they have very high implementation costs. Additionally, they are domain-specific [18], which makes automation difficult and results in significant variability in their applicability and effectiveness across different contexts. These limitations highlight the ongoing need to continue developing and refining data augmentation methods that can tackle this problem and enhance the quality and diversity of the dataset, thereby ensuring more effective and accurate training of DL models in analyzing the professional profiles of independent directors and financial research in general.

In this study, we propose a new data augmentation technique. We start from the hypothesis that Mixup [31] and Label Smoothing Regularization (LSR) [32] methods, which are commonly used in image processing and share similar goals with our approach, can be extrapolated to the NLP context to accurately create text samples. Our objective is thus to design, implement, and evaluate a new data augmentation method that, inspired by these techniques, generates variability from the available data and significantly improves DL model training. Accordingly, this research addresses the following main questions:

- Can the proposed method enhance the performance and generalization of DL models in financial NLP tasks under data scarcity conditions?
- To what extent does it outperform or complement existing comparable augmentation methods, such as Wmix and LSR, in these same tasks and conditions?

To answer these questions, we introduce the new method and conduct extensive experiments in the context of corporate governance disclosures. Therefore, the main contributions of our study are the following:

- Development of the Concatenation Augmentation (CA) method. This new data augmentation technique generates new samples by concatenating inputs and applying a convex additive operator to generate labels, maintaining spatial and semantic coherence.
- Enhancement of the accuracy and generalization capabilities of DL models for automatically extracting key data about independent directors from corporate reports.

Thereby, it is intended to contribute to the advancement of NLP techniques in finance, promoting the development of more accurate and robust models capable of handling the inherent variability and uncertainty of existing unstructured text data.

A key feature of our technique is its low computational cost, which positions it as a promising contribution within the broader context of sustainable development [33,34] and responsible AI. As concerns over the environmental impact of large-scale AI systems grow, particularly those requiring extensive computational resources, techniques that minimize energy consumption and hardware demands play a critical role in promoting more sustainable practices in machine learning. By enabling model enhancement without reliance on computationally intensive architectures or augmentation pipelines, our approach supports efforts to reduce the carbon footprint of AI workflows, thereby aligning with sustainability goals outlined in global agendas such as the United Nations' Sustainable Development Goals (SDGs).

Furthermore, this work contributes to the discourse on ethical and inclusive AI governance by making advanced data augmentation more accessible to institutions or research groups with limited computational resources. This supports broader participation in AI development and enhances institutional resilience, particularly in low-resource settings. Integrating efficiency-focused techniques like ours into standard ML practices not only advances performance metrics but also addresses long-term challenges related to equity, environmental responsibility, and the democratization of AI technologies.

The organization of this document is as follows: Section 2 provides a review of conventional data augmentation techniques (including LSR, Mixup, and its variants) and

more advanced LLM-based approaches. Section 3 provides a description of our proposed CA method, detailing its application and benefits. In Section 4, we present the conducted experiments, which include a description of the initial dataset, the base model used, and the experimental methodology. Section 5 presents the results obtained. Section 6 discusses the practical implications of our findings and offers recommendations for implementing CA in real-world financial environments. Finally, Section 7 summarizes the main findings and conclusions of this study.

2. Related Work

In this section, we provide an overview of data augmentation techniques used in NLP, categorizing them into two main groups: conventional techniques, discussed in Section 2.1, and more recent, sophisticated approaches that leverage complex models and/or large language models (LLMs), covered in Section 2.2.

The technique proposed in this work belongs to the first category of conventional methods. It is characterized by its low computational cost and independence from specific model architectures, making it highly practical and scalable. Among the existing techniques, the most closely related to our approach are Label Smoothing (LSM) and WMix, which are detailed in Section 2.1.1 and Section 2.1.2, respectively.

2.1. Conventional Data Augmentation Techniques

Data augmentation has been widely and effectively utilized in computer vision [32,35,36] and speech recognition [37–39]. Many of these techniques rely on human expertise to transform data in a label-invariant manner, such as scaling, flipping, and rotation of images. However, unlike images, slight alterations to a word within a sentence can drastically change its meaning. Consequently, there is not a straightforward rule for conducting label-invariant transformations in NLP [18].

As previously mentioned, conventional data augmentation methods in NLP focus on techniques such as transforming text by substituting words with synonyms sourced from manually crafted ontologies [22,23,25], translating samples to a foreign language and back [24] to rephrase the inputs. However, these techniques are not without limitations. The use of synonyms and word substitution can alter the meaning of the original texts, while translation and back-translation can introduce errors and unwanted variations that do not accurately reflect the original content. Paraphrasing, although useful, can result in the loss of important details or changes in context that affect the interpretation of the text. Moreover, all these techniques have a very high implementation cost, a high domain dependency, and cannot be easily automated. Another category of techniques, such as SSMix [27] and DropMix [28], relies on mixing and replacing tokens based on their saliency maps. Although they are simpler to apply than those mentioned above, they are tied to a specific type of problem, sentiment analysis, so their adaptation to other types of problems is not immediate, requiring a very careful analysis of their applicability.

In this study, our main interest lies in the development of free-text data augmentation techniques that can be easily automated, do not require external data or models, and have a low computational cost. We will draw inspiration from the Mixup [31] and LSR [32] techniques to adapt or develop variants that can be applied to our problem. At a conceptual level, the motivation behind applying LSR is to incorporate into the network the prior knowledge of continuity and imprecision in the results. On the other hand, Mixup generates variability by superimposing information and labels, but has the difficulty of not being directly reproducible in NLP. There are other proposals in the state of the art that have already been discussed in Section 1, such as token- and sentence-level augmentation, adversarial augmentation and hidden-space augmentation. However, we base our proposal

on LSR and Mixup, instead of using other existing techniques, since it should be easily generalizable and not depend on a specific domain or model.

Below we provide a brief overview of two techniques that share similar objectives with CA. These will be used to establish a comparison basis of results and to highlight the advantages of our proposal.

2.1.1. Label Smoothing

LSR [32] is a well-known regularization technique that is commonly used in classification tasks. Its main aim is to reduce the confidence of the training model with the desired goal of preventing the largest logit from becoming much larger than all others. In order to achieve this, LSR generates new samples by adding a uniform distribution to the original label distribution. As a result, given the training set x and its labels $k \in \{1 \dots K\}$, LSR is defined based on the smoothing parameter ϵ and a distribution over the labels $\mu(k)$ that is independent of the training set. The distribution of labels $q(k \mid x) = \delta_{k,x}$ is replaced by $q^{'}(k \mid x) = (1-\epsilon)\delta_{(k,x)} + \epsilon\mu(x)$; that is, it is replaced by the union of the original distribution with $\mu(k)$, independent of the training set and with weights $(1-\epsilon)$ and ϵ. For example, if the uniform distribution $\mu(k) = \frac{1}{k}$ is used, then the distribution of labels takes the form $q(k \mid x) = (1-\epsilon)\delta_{k,x} + \frac{\epsilon}{k}$. Thus, this method creates controlled variability in the labels that results in a reduction in overfitting of the training process [32].

The adaptation of this technique to our particular case of regression instead of classification is straightforward due to the granularity present in the labeled data, consisting of values between 0 and 1 with increments of 0.1 between them. For this reason, a loss function based on RMSE has been used instead of the more standard Cross Entropy, since the latter is better adapted to classification tasks. This change does not affect the data augmentation technique since it only alters the training of the DL model to make it more effective in regression problems. In fact, it is possible to generalize this adaptation to any regression problem where there is a non-negligible precision limit in the resulting values.

Thus, LSR successfully combines effectiveness in increasing the variability of training data with simplicity. However, it is limited, by design, in the amount of variability it can generate in the dataset. For this reason, it loses effectiveness when the amount of available data is small and therefore must be combined with other approaches.

2.1.2. Mixup and Derivatives: WordMixup (WMix)

Mixup [31] is a data augmentation technique that has yielded excellent results in the field of computer vision, despite its simplicity [40–42]. It involves generating new inputs by interpolating inputs and outputs. Thereby, new virtual training images and their corresponding labels are constructed by combining pairs of images with each other using a parameter λ according to the following equations [31]:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda)x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j \end{aligned} \tag{1}$$

where (x_i, y_i) and (x_j, y_j) are samples from the training set, $(\tilde{x}, \tilde{y})$ is the new sample generated by means of Mixup, and the parameter λ that manages the interpolation degree is sampled from the statistical $Beta(\alpha, \alpha)$ distribution, whose hyperparameter α controls the dispersion of the generated values. Several tests were conducted for different values of α and, finally, we chose the value $\alpha = 1$ since it was the one that provided the best results for our dataset.

However, generalizing Mixup to the NLP field is problematic due to the inputs consisting of variable-length character sequences, so there is no direct way to perform the interpolation operation between them. To deal with this limitation, ref. [43] proposes two

models, wordMixup (WMix) and senMixup, where interpolation is performed at the word embedding level (zero-padded) or at the sentence representation level, respectively. There are also more recent variants, such as Mixup-Transformers [44]. However, this technique is specifically designed for Transformer-based models. Because Transformers and LSTM models have fundamental differences in how they process sequential information, using this technique would require significant adaptations and is not suitable for our model. Thus, we chose WMix for comparison purposes since it is easier to implement in our model and has similar performance to senMixup [43].

In this regard, LSR and our proposed CA can be applied directly to the data, without the need to modify the model code, allowing rapid prototyping to test improvements. However, WMix requires modification of the model code since the interpolation occurs in internal layers [43]. On the other hand, WMix, like Mixup, uses the hyperparameter α that must be tuned to achieve optimal performance [43]. In contrast, LSR and CA are designed to work effectively with default settings or minimal adjustment and do not have hyperparameters, simplifying their implementation. As a result, LSR and CA are more work-friendly and less prone to issues related to hyperparameter sensitivity, such as overfitting.

2.2. Advanced Data Augmentation Techniques

Recent advances in NLP have given rise to sophisticated data augmentation methods that go beyond traditional heuristics. Notable techniques include LLM-based paraphrasing, embedding-based augmentation using Transformer models, adversarial augmentation, and contextual augmentation.

- Large language models like GPT-4 [45] and Llama [46] have been widely adopted for paraphrase generation. This method enables the creation of diverse training samples that help models generalize better to different linguistic expressions.
- Another advanced technique leverages the latent space of Transformer-based models such as BERT [47] or RoBERTa [48]. In this approach, text inputs are encoded into embeddings, and augmentation is performed in the continuous embedding space—either by interpolating between vectors (e.g., Mixup), adding noise, or sampling nearby points. The modified embeddings are then decoded or used directly for training, depending on the task. This strategy allows for smooth and semantically meaningful transformations that are often difficult to achieve with surface-level text manipulations.
- Adversarial data augmentation involves generating examples that are intentionally challenging for a model—often by introducing minimal but semantically significant perturbations. Techniques such as HotFlip [49] craft adversarial inputs by substituting words or characters while preserving the original meaning. These examples are used to fine-tune models, making them more resilient to subtle variations and adversarial attacks.
- Contextual augmentation uses masked language models (e.g., BERT [50]) to replace selected words with contextually appropriate alternatives. Unlike static synonym replacement, this method considers the surrounding text to generate grammatically and semantically coherent augmentations, preserving the integrity of the original sentence.
- Finally, a promising area of ongoing research is the development of semi-supervised models designed to support—or potentially replace—the manual labeling process, like [51].

A common characteristic of these techniques is their reliance on advanced, computationally intensive models, which differs from the simplicity and low computational cost of

our proposed method. Our study includes a comparison with LLM-based paraphrasing, as a representative example of this class of techniques.

3. Material and Methods

3.1. Concatenation Augmentation (CA)

We propose a new method called Concatenation Augmentation (CA), which generates high variability into the training dataset through simple and straightforward operations. CA is based on concatenating training information, with an operation defined by the additivity of the assigned labels, reflecting prior embodied knowledge. The heuristics that justify it are based on the union of knowledge of free-text information that is expressed through the new additive operation.

3.2. Algorithm

Our proposal, CA, combines the best characteristics of LSR and Mixup in a data augmentation technique with very reduced computational requirements. CA significantly increases the variability provided by LSR by using a Mixup variant that consists of concatenating instead of directly mixing the information. CA adapts very well to environments in which the availability of trained data is very limited. Our proposal generates new samples according to the following equations:

$$\begin{aligned} \tilde{x} &= x_i \oplus x_j \\ \tilde{y} &= \Phi(y_i, y_j) \end{aligned} \tag{2}$$

where (x_i, y_i) and (x_j, y_j) are samples from the training set, $(\tilde{x}, \tilde{y})$ is the new sample generated by CA, $\oplus$ represents the concatenation operator over strings (in the implementation used, no special string processing was performed, except for appending sentence terminators to strings where they were missing), and Φ is a function that represents the combination of outputs and depends on the kind of activation function used by the last layer. In our case, with the logistic function as the activation function, Φ is defined by the following:

$$\Phi(y_i, y_j) = \sigma\Big(\sigma^{-1}(y_i) + \sigma^{-1}(y_j)\Big) = \frac{y_i y_j}{1 - y_i(1 - y_j) - y_j(1 - y_i)} \tag{3}$$

where $\sigma^{-1}(y) = \ln \frac{y}{1-y}$ is the inverse logistic function used to recover the original preactivation values.

In other neural network architectures or output configurations, distinct activation functions may be used, requiring a different inverse transformation. Alternatively, if access to the model's internal code is available, the output of the layer preceding the activation function can be obtained directly, eliminating the need for this mathematical inversion.

3.3. Advantages of CA

Like LSR, CA is a data augmentation technique with very low computational cost, does not require additional data or models, and can be easily automated. It is equivalent to increasing the training dataset with new labeled data, as it generates new pairs $(\tilde{x}, \tilde{y})$ in a robust and reliable way. In order to achieve this, its heuristic relies on Mixup modifying its logic so that it can be used in NLP, avoiding the loss of information since it adapts to the sequential nature of written language. Based on all the above, it can be said that CA is a combination of LSR and Mixup that takes advantage of the best features of each of them to significantly increase the variability generated in the training dataset. This variability allows us, in turn, to regularize the weights of the trained model to improve its adaptation

to unknown data during the training phase. In short, CA should improve the training of models by increasing its accuracy and reducing overfitting.

The implementation of CA is very simple and can be easily included in the training data processing pipeline. It is also directly usable in all types of NLP-based regression problems, provided that some minimum requirements based on the convex additivity of the regression values are met. Due to its ability to combine any pair of training data, it allows us to greatly expand the amount of available training data, thus adapting to problems where the scarcity of data limits its performance. This provides a new hyperparameter that allows adjusting the training stages according to the needs of variability and dataset size. To assess overfitting reduction and accuracy, a complete set of experiments has been designed, described, and analyzed in the following section.

4. Experimentation

4.1. Dataset

In this section, we describe the initial dataset, the base DL model used, the characteristics of the training performed, and a validation analysis of the base model.

The dataset created in this work is made up of the biographies of the independent directors of companies listed on the Spanish stock exchange in the period 2003–2020. This section explains the main aspects of the creation of the dataset, its justification, the models, and the training. For a more detailed description, please see our previous studies [6,7]. The first step to generate the dataset was to obtain the CVs, which are openly available but were written as free text and without any type of predefined structure. As a result, there is great variability, both in length and style, as well as in vocabulary (e.g., the length varies between 5 and 4052 characters). These features make it very difficult to apply NLP algorithmic techniques. However, recurrent DL neural network models adapt very well to this type of problem due to their ability to analyze arbitrarily long data series [10].

The data acquisition, processing, modeling, and training are described in Figure 1. The data was taken from the Annual Corporate Governance Reports (ACGRs) [8], which are open reports generated by the CNMV (National Securities Market Commission) but have the problem that they are published in unstructured data formats. In order to extract them, web scraping, pre-processing, cleaning, and filtering had to be carried out, thus obtaining a total of 6561 biographies of independent directors. The generated dataset is robust and reliable because its creation required the work and supervision of human experts in finance, computing, and DL. However, to train a supervised learning DL model with good guarantees, it is essential to achieve reliable labeling of the data.

Of the 6561 biographies obtained, 1042 were selected to carry out data labeling, leaving 5519 available to perform the full analysis once a trained model was obtained. To achieve a reliable dataset, the labeling must be performed by a human expert. Although from the point of view of the processing capacity of a DL model, there is little data, from the perspective of the human expert, generating reliable labels for each of the 1042 biographies is a very time-consuming task. Considering financial and business criteria, we decided to carry out the labeling in six dimensions: Financial (F), Executive/Consultant (E/C), Audit/Tax/Accountant (A/T/A), Legal (L), Political (P), and Academic (Ac). The description of each one is as follows:

- Financial (F): This refers to 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): This refers to 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

- 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): This refers to directors who have held or are holding public offices of various kinds, especially political posts.
- Academic (Ac): Finally, this refers to directors with academic experience.

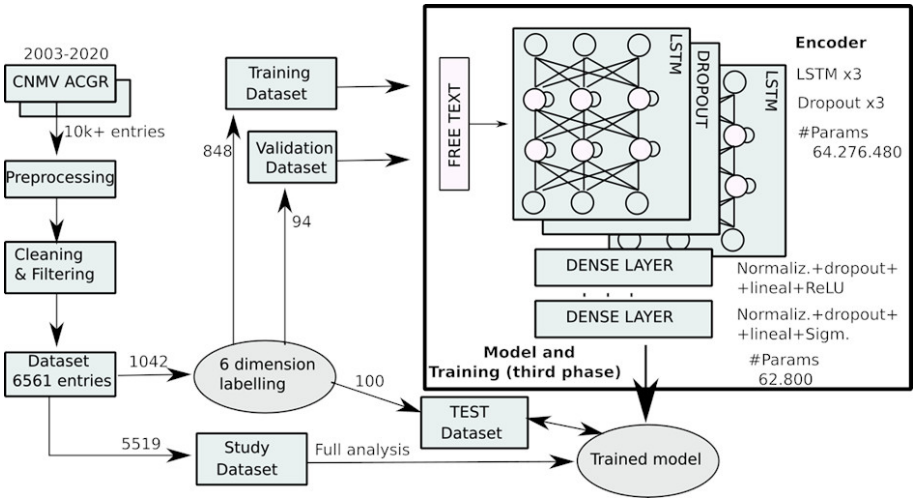

Figure 1. Data acquisition, processing, modeling, and training. All tables and figures are the authors' own creation.

Almost all the skills identified in the biographies 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. Refs. [52,53] document that board members with financial expertise significantly influence firm financing and investment decisions. Likewise, according to the literature, independent directors help improve firm value with their industry experiences [54], and ref. [55] documents the positive market response to the appointment of successful executives from other companies. In addition, firms with an accounting expert sitting on their audit committees show a stronger accounting conservatism [56]. Regarding the legal dimension, its importance is highlighted by works such as that of [57], which shows that the presence of directors with a legal background is associated with a higher quality of financial reporting. Ref. [58] finds that the cost of bank loans is significantly lower for companies with board members with political ties and [59] points to the prevalence of directors with a political background for companies with significant government contracts. Finally, some articles highlight that academic directors play an important governance role through their advisory and supervisory functions, leading to increased R&D performance and investment [60].

In order to carry out the labeling, the human expert generates a qualitative assessment for the six dimensions using the CVs. This assessment is not exclusive; i.e., the same independent director can have the highest qualification in the Financial and Academic dimensions, for example. This assessment is then quantitatively approximated by a six-dimensional numerical vector where each value is represented in the interval [0, 1] with a single decimal digit, that is, 11 possible values for each dimension, minimum value (0.0), maximum (1.0), and nine intermediate values (0.1, 0.2, . . . , 0.9).

Generating six-dimensional labeling of 1042 CVs by a human expert is a labor-intensive process that produces robust and reliable results, as it involves manual, one-by-one annotation. However, it does not generate a sufficient number of values to completely and reliably train a DL model. The reason is that, first of all, 100 entries are set aside to generate the test dataset. This dataset is used once a trained model has been obtained to evaluate the generalization capacity of the model when working with data that it has never previously used in the training stage. With the remaining 942 entries, a 90–10% distribution is made between the training and validation sets, leaving a total of 848 biographies with which to train the model. The validation set is used at the end of each epoch to evaluate the state of the training and make decisions about the values to apply to the hyperparameters in the next epoch with the aim of improving the training.

4.2. Base Model

The base DL model used is an ASGD Weight-Dropped LSTM (AWD-LSTM) recurrent neural network [10], whose main features are DropConnect regularization and NT-ASGD optimization. DropConnect regularizes by randomly discarding certain information learned by the network, but instead of discarding the outputs of each stage, it does so with the weights of the neurons. NT-ASGD is a variant of ASGD that reduces the influence of randomness in the training process. Thus, AWD-LSTM has demonstrated excellent learning capacity and has generated robust results in multiple applications in the NLP field [61,62]. It is true that there are other models with good results in NLP, such as those based on Transformers [63,64]. These models have not been used in this work because the purpose of this research is to mitigate the negative effects of the scarcity of reliable data in DL model training in the NLP field, not to make a comparison between different state-of-the-art models. By applying the proposed data augmentation techniques to a single model, the effectiveness of these techniques can be easily established [65].

The training was carried out using the universal language model fine-tuning approach described in [66]. A more detailed description of the training carried out can be found in [7]. The training consists of three phases, as described in Figure 2. The first phase consists of a random initialization of all the parameters of the AWD-LSTM model and the training of a Spanish-language model through self-learning using Wikipedia articles. In this way, the DL model captures the main features of the language. The dataset for this first phase comprises 63,900 Wikipedia articles, containing a total of 65 million words. In the second phase, academic texts from the financial, economic, and business fields were selected, with a total length of 2.6 million words. By using texts specific to the scope of application, the model is adapted to specific linguistic, formal, and vocabulary characteristics. The third phase consists of training the network to achieve the established objective, which in our case is a regression in six dimensions. To this end, the model was augmented with two additional fully connected levels, including the corresponding normalization, dropout, and activations in each new level [7]. As previously explained, the dataset used for training is made up of 848 CVs plus 94 additional ones for the validation stage.

4.3. Transformer-Based Model (Ensemble)

To evaluate the model-independence of the CA technique, we employed a second model that differs from the primary one. This alternative model, introduced in a previous study [7], is an ensemble architecture composed of the main model (AWD-LSTM) and two Transformer-based models: BERT [47] and RoBERTa [48]. All three models were trained on the same dataset described in earlier sections to ensure a fair comparison.

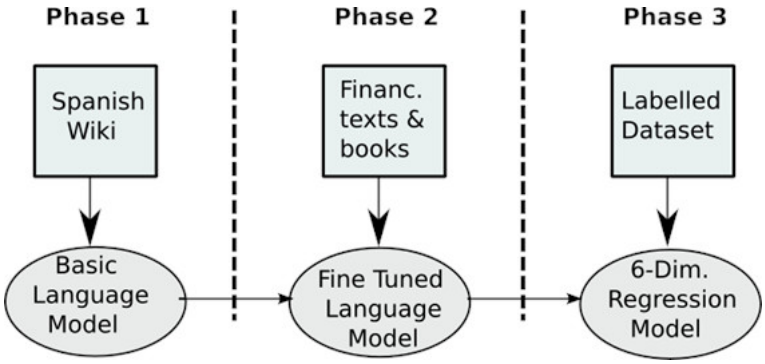

Figure 2. Training phases.

The ensemble employs a stacking strategy, in which a simple multi-layer perceptron (MLP) takes as input the outputs of the three base models and produces a final prediction. This MLP learns to select and weight specific components based on patterns identified during training. Specifically, for each epoch, the ensemble receives the outputs from the three models alongside the labels provided by human experts, allowing it to build an internal representation of the relative performance of each model across the six evaluation dimensions.

This setup enables us to assess whether the effectiveness of the CA technique generalizes across different model architectures, encompassing both traditional (RNN-based) and Transformer-based approaches.

4.4. Experimental Methodology

A series of experiments was conducted using the two models described in Sections 4.2 and 4.3, each with different variants based on the data augmentation techniques applied. In the case of Concatenation Augmentation, a human expert reviewed the generated datasets to ensure semantic and labeling consistency.

4.4.1. Base Model

Three variants of the base model presented in Section 4.2 have been created, differing only in the data augmentation technique used: (i) Concatenation Augmentation (CA), the technique proposed and developed in this paper; (ii) LSR; and (iii) WMix, with the latter two described in Section 2.1.1 and Section 2.1.2, respectively.

In the case of WMix, the parameter λ was randomly sampled from a $Beta(\alpha, \alpha)$ distribution. Several tests were conducted for different values of α and, finally, we chose the value $\alpha = 1$ since it is the one that provided the best results.

Each variant of the base model was trained by modifying the number n of additional data generated by the corresponding data augmentation method. Particularly, we conducted trials with $n = 0, 1000, 2000, 3000, 4000, 5000$. Thus, 16 models (the model obtained with the value $n = 0$ (no augmentation) is the same for the three models) were trained, inferences were made from them, and they were evaluated in the test dataset. All models were trained with the same hyperparameters and number of epochs (50), except for the model with $n = 0$ (no augmentation) in which the number of epochs was increased to 100 in order to ensure better convergence.

To ensure statistical significance, bootstrapping techniques (300 replicates) were applied to our model, and subsequent analyses were conducted to compute 95% confidence intervals (± 2 standard deviations) for key variables and error metrics.

In the evaluation, the usual statistical metrics of a regression model were obtained (MAE—Mean Absolute Error, RMSE—Root Mean Standard Error, correlation coefficient). We also extended the study to a category model, since the metrics obtained (hit rate,

confusion matrix) are more informative about the goodness of fit of each model. To accomplish this, the values of each dimension were discretized into four categories corresponding to the intervals [0, 0.3), [0.3, 0.5), [0.5, 0.7), and [0.7, 1.0], which are indexed by the integers 0, 1, 2, and 3, respectively [6].

4.4.2. Alternative (Ensemble) Model

For the alternative model, whose primary purpose is to evaluate the independence of the proposed Concatenation Augmentation (CA) technique from the underlying model architecture, a reduced set of variants, tests, and results was considered. In terms of data augmentation, two techniques were evaluated: (i) Concatenation Augmentation (CA) and (ii) LLM-based paraphrasing. For the LLM-based paraphrasing approach, ChatGPT-4 and Llama2 external models were used to generate 1000, 2000, and 3000 additional inputs. These inputs consisted of semantically equivalent rephrasings of the original biographies, expressed using different wording while preserving their original meaning. The evaluation employed the same metrics as in the previous subsection; however, this analysis did not extend to the category-level model.

5. Results

We present the results of our experiments, evaluating the impact of different data augmentation techniques on the training and performance of our models. We focus on two main aspects: the improvement in training and the trade-off between precision and computational cost. Firstly, we analyze the training improvement by comparing the training loss of the models with and without data augmentation. Subsequently, we assess the precision of the models in relation to the computational cost, examining how the number of additional inputs affects the performance metrics. Lastly, we provide a detailed breakdown of results by category, highlighting the effectiveness of our proposed technique, CA, in achieving superior and consistent outcomes across all professional profiles.

5.1. Training Improvement

To illustrate the improvements in training, Figure 3 displays the training loss graphs of the model without augmentation and of the models with 3000 additional inputs for each augmentation technique considered. The value of 3000 was chosen because it is the minimum amount of data at which the model achieves the best results, as shown in the next subsection. The three models with data augmentation showed faster convergence and better results than the model without augmentation. It should be noted that this last model was trained with twice as many epochs. Of the models with data augmentation, WMix exhibited a slower convergence than the other two. Additionally, it can be seen by examining the test loss curve that CA had a smoother training than the other methods.

5.2. Precision vs. Computational Cost

The application of data augmentation techniques has an associated computational cost, since a greater number of inputs implies longer training times. Therefore, it is important to evaluate the different data augmentation techniques to characterize the performance of the model with respect to the number of additional inputs. Thus, Figure 4 presents the evolution of two metrics related to the precision of the model with respect to the number of additional inputs. Since there are six dimensions (categories), we show only their average value. The error bars representing the 95% confidence intervals, obtained through bootstrapping, are displayed above each data point. In the graph on the left, the chosen metric is RMSE for the regression model, which measures the average magnitude of the errors between predicted and actual values. On the right, the metric used is the hit rate for the categorized model, as described in Section 4.4. This metric quantifies the proportion of

correctly predicted instances out of the total, providing insight about the overall precision of the model in classification tasks.

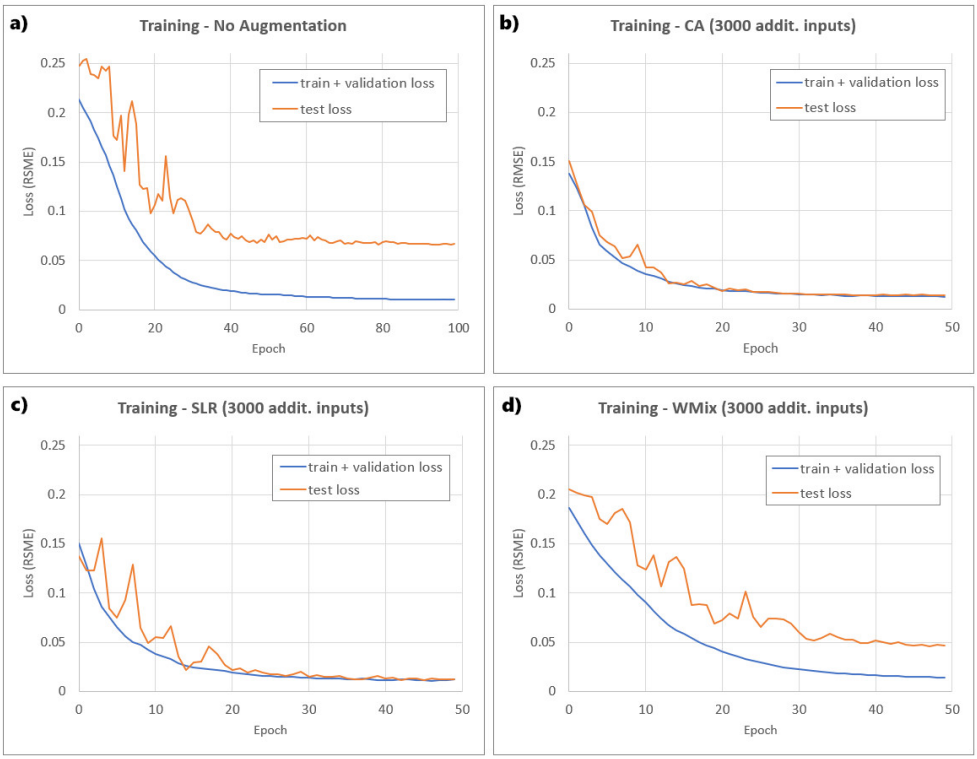

Figure 3. Training results obtained (**a**) without data augmentation, (**b**) with samples generated by CA, (**c**) with samples generated LSR, and (**d**) with samples generated by WMix.

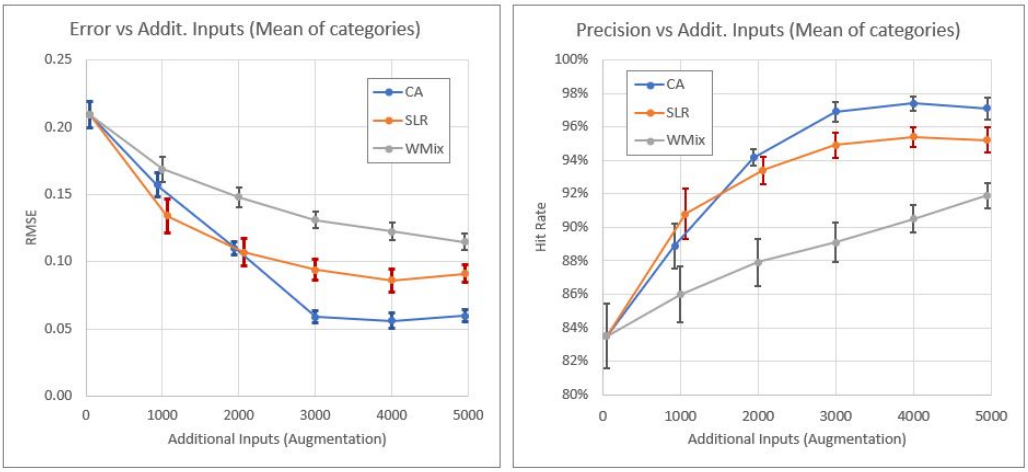

Figure 4. Performance of the base model with respect to computational cost.

The graphs show, as expected, an increase in the precision of the model when the number of additional inputs increases. This rate of improvement is lower for WMix. In the case of CA and LSR, the precision peaks between 3000 and 4000 additional entries, after which it either plateaus or slightly decreases.

5.3. Alternative Model and LLM-Based Augmentation

The alternative model (Ensemble) described in Section 4.4.2 was evaluated using two data augmentation techniques: Concatenation Augmentation (CA) and LLM-based paraphrasing, each tested with varying numbers of additional inputs. The results (average

RMSE across categories) are presented in Figure 5, and include, for comparison purposes, the performance of the base model with CA augmentation.

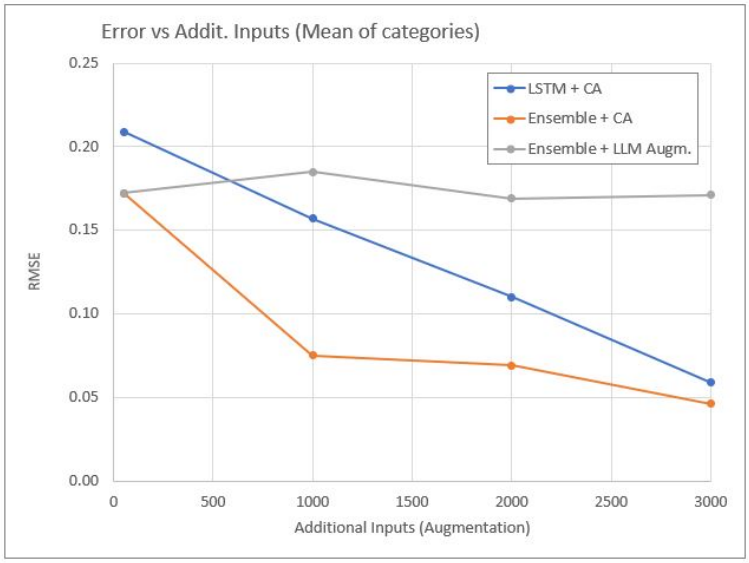

Figure 5. Performance of the alternative model with respect to computational cost.

Figure 5 shows that CA also leads to performance improvements in the alternative Transformer-based model. Notably, the most significant gains occur with a smaller number of additional data points, likely due to the model's higher capacity and sophistication. In contrast, the results for the LLM-based paraphrasing technique indicate that this form of augmentation does not improve model performance. A plausible explanation is that the paraphrased biographies do not contribute substantial new information. The changes are primarily syntactic, which may have limited impact on the model's predictions—particularly in a task where the presence of specific keywords is likely more influential than sentence structure.

5.4. Results by Category

In addition to the overall performance metrics, it is essential to analyze the results achieved in each professional category to understand the impact of CA in this context. Thus, Table 1 shows the statistical metrics (mean of bootstraped data) obtained for the base model without data augmentation and for the three variants when the number of additional inputs is $n = 3000$. As explained in Section 4.4, the hit rate was calculated on the discretized model. According to these results, CA achieved the best overall performance, closely followed by LSR. Moreover, CA obtained lower squared and absolute errors, as well as higher correlation and precision than LSR.

In order to further explore the comparison between CA and LSR, Figure 6 shows the confusion matrices of both methods for all categories for a representative run. As can be seen, the values are more clustered around the main diagonal with our proposal.

An analysis was conducted on the 3.2% of entries misclassified in the confusion matrices for the test case "CA with 3000 additional inputs". Of these misclassifications, 64% were attributable to biographies lacking sufficient or high-quality information (e.g., excessively short texts, typographical errors). Another 26% involved biographies containing relevant terms (such as company names, job titles, or academic credentials) whose low frequency resulted in their exclusion from the token dictionary. The remaining 10% represented genuine misclassifications by the model.

Table 1. Test set results for each model variant.

Augmentation	Metric	F	E/C	A/T/A	L	P	Ac
None	MAE	0.133	0.211	0.072	0.085	0.095	0.088
	RMSE	0.204	0.316	0.182	0.185	0.167	0.170
	r	0.869	0.759	0.762	0.858	0.781	0.878
	Hit Rate	78.2%	69.4%	91.3%	92.3%	84.2%	84.1%
CA (+3000 inputs)	MAE	0.038	0.027	0.007	0.014	0.020	0.038
	RMSE	0.095	0.090	0.023	0.028	0.036	0.091
	r	0.970	0.979	0.993	0.994	0.987	0.972
	Hit Rate	92.5%	97.5%	99.7%	99.6%	96.5%	92.4%
SLR (+3000 inputs)	MAE	0.076	0.080	0.026	0.021	0.050	0.043
	RMSE	0.146	0.166	0.049	0.047	0.070	0.103
	r	0.940	0.942	0.982	0.989	0.980	0.961
	Hit Rate	88.6%	90.6%	97.7%	97.6%	92.6%	92.5%
WMix (+3000 inputs)	MAE	0.089	0.133	0.046	0.056	0.068	0.067
	RMSE	0.134	0.209	0.111	0.118	0.112	0.128
	r	0.948	0.906	0.953	0.948	0.913	0.935
	Hit Rate	85.8%	78.9%	91.9%	92.9%	88.8%	89.7%

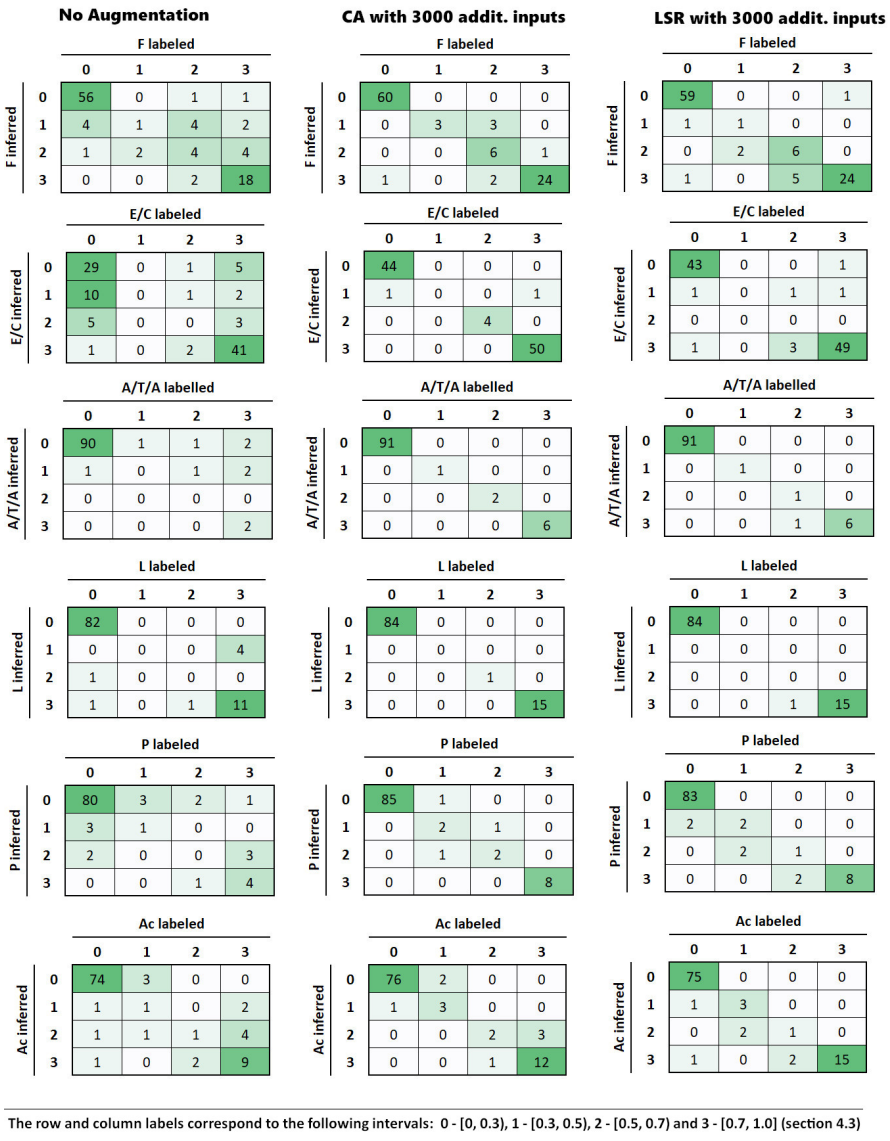

Figure 6. Confusion matrix for each category.

6. Discussion

Throughout our study we have analyzed how the proposed technique, CA, can mitigate the limitations imposed by the scarcity of data, improving the generalization capacity of the models and, therefore, its performance in critical tasks such as the extraction of key information from financial-related corporate reports.

As we saw in the results obtained, the use of data augmentation techniques has a clear positive impact on the model training process. In fact, some important trends were observed. Firstly, it is common to see a rapid decrease in RMSE at the beginning of training (Figure 3), as this indicates that the model is learning important patterns from the data. Afterwards, the loss curve stabilizes, indicating that the model has reached a point where it continues to improve, but at a much slower rate. In this regard, models with data augmentation demonstrated faster convergence and achieved better results compared to the model without data augmentation. This is a significant finding, as the improvement in convergence speed, and thus learning, can translate into a reduction in training time and potentially more efficient use of computational resources. Additionally, models with data augmentation achieved better results even with fewer epochs, which highlights the effectiveness of these techniques in enriching the training set and enhancing the ability of the model to generalize.

Among the data augmentation techniques considered, it was observed that the model using WMix showed slower convergence compared to CA and LSR (Figure 3). This behavior suggests that, although WMix may be beneficial for dataset variability, it might require more careful tuning of its parameter α or more training time to reach an optimal performance. Additionally, it is important to note that if the curve starts increasing again after stabilizing, as seen with WMix, this could be an indication of overfitting, as the model might be memorizing the training data rather than learning generalizable patterns. In this regard, the models trained with data generated through CA and LSR did not show signs of overfitting.

Additionally, the CA technique showed a smoother test loss curve throughout the training, without sharp fluctuations or pronounced peaks (Figure 3). This smoothness in the loss curve indicates lower variability in the performance of the model in the test set, which can be interpreted as greater stability and the ability to generalize well to new data. The smoothness of the curve also suggests that CA might be providing an adequate balance between variability and similarity in the augmented data, facilitating more consistent and robust model learning. This behavior also indicates a greater regularization effect, which would be helping to control model complexity and reduce overfitting.

The results obtained also provided a detailed view of how the addition of supplementary inputs affects model precision (Figure 4). In this regard, it was observed that as the number of additional inputs increases, RMSE decreases, indicating an improvement in the precision of the regression model. This trend was expected, as a larger number of training data generally helps the model better capture the underlying patterns in the data, thereby reducing prediction errors. However, this improvement is not uniform across the data augmentation techniques compared in this study. It is notable that WMix shows a lower improvement rate compared to CA and LSR, regardless of the number of samples incorporated into the process. In contrast, CA exhibits a lower error than the other techniques even before reaching 2000 additional samples. This fact demonstrates that, in addition to being beneficial for training DL models, the effectiveness of our proposal is higher compared to other techniques when it comes to reducing prediction error in a regression model.

Similar to RMSE, the hit rate improved with the increase in the number of additional inputs (Figure 4), reinforcing the notion that more training data contributes to better model performance. For CA and LSR techniques, it was observed that precision peaked when

the number of additional inputs was between 3000 and 4000. Beyond this point, precision tended to stabilize or even slightly decrease, which probably means that the model had reached its best performance. This phenomenon would be related to the saturation point of the model, where adding more data no longer significantly improves precision and could introduce redundancy or even noise that negatively affects performance and computational cost. It is important to highlight that before reaching 2000 additional samples, CA already exhibited a higher hit rate than the other techniques. This not only indicates that our proposal offers significant precision improvements with fewer additional data but also suggests that CA may be particularly effective in the early stages of data augmentation. Thus, its ability to achieve and maintain a high level of precision with fewer additional samples makes it a particularly efficient and robust option, optimizing model performance from the early stages of training and reducing the need for high computational costs to achieve substantial improvements.

The CA technique, as shown in Figure 5, also improves performance in the alternative Transformer-based ensemble model. Interestingly, the most notable gains are observed with a relatively small number of additional inputs, which may be attributed to the model's greater capacity to generalize from limited but well-structured augmented data. By contrast, the LLM-based paraphrasing technique does not appear to yield performance improvements in this setting. One possible explanation is that the paraphrased biographies introduce primarily syntactic variations without enriching the semantic content in a way that benefits the model. Given the nature of the task—where the presence of specific keywords is likely more predictive than sentence structure—such superficial changes may offer limited value. This suggests that not all augmentation strategies are equally effective, and their utility may depend heavily on both the task characteristics and the model architecture.

From a categorical perspective (Table 1), CA achieved better performance than LSR and WMix in all categories. It not only outperformed them in terms of overall precision but also showed lower squared and absolute errors, as well as a higher correlation with the true categories. In this regard, the advantage of CA in error and accuracy metrics suggests that this technique is particularly effective in reducing the discrepancy between predictions and actual values. Additionally, the fact that CA achieves better results in terms of correlation implies that the predictions of the model are more aligned with the actual categories, reinforcing confidence in its ability to generalize. Consequently, CA not only enhances the ability of the model to make more accurate predictions but also provides a more faithful representation of the data compared to other data augmentation techniques.

Thus, CA obtained better results than WMix and LSR in all categories. Additionally, the confusion matrices were examined to further compare CA and LSR (Figure 6). These confusion matrices revealed that the professional profiles obtained with CA are more clustered around the diagonal. This means that when the model fails, it is more likely to do so in a category closer to the true one. Consequently, the concentration of values along the main diagonal suggests that the CA-trained model is more consistent in correctly classifying instances, even in the more challenging categories. This characteristic is crucial in our study, as accurate classification of professional profiles is fundamental and can have a significant impact on the interpretation of results and on financial decision-making derived from them.

These findings underscore the importance of data augmentation to improve model performance. Thus, through a series of experiments, we have demonstrated that CA not only increases the amount of data available for model training but also preserves the statistical and semantic properties of the original corpus. This translates into significant improvements in the evaluation metrics of DL models, highlighting the effectiveness of our technique in scenarios with limited data. Thereby, CA emerges as a promising data

augmentation option that enhances the precision and predictive capacity of the model while enabling fast convergence and stable training, as well as reducing computational costs.

Despite the promising results obtained, this study presents certain limitations that should be considered. First, the evaluation of the CA technique was conducted within a specific context. It focused on financial NLP tasks and, more particularly, on the extraction of information about independent directors from corporate governance reports written in Spanish. This restriction limits the generalizability of the findings to other languages or financial contexts. Therefore, future validation efforts should consider applying CA to multilingual datasets from diverse domains. Another notable limitation lies in the intensive manual labeling required to create the training dataset, which constrains the scalability of our study. In this regard, future research could explore the adoption of automated or semi-supervised labeling techniques, such as weak supervision, to reduce dependence on manual annotation. Although three data augmentation techniques were analyzed, other emerging approaches may offer additional improvements. Studying their complementarity with CA represents a promising avenue for future work. Finally, it would also be interesting to explore variants of the CA technique for different tasks, such as binary classification or text generation, as well as its integration into more complex model architectures.

7. Conclusions

Although processing unstructured texts presents significant challenges, emerging solutions in NLP and machine learning offer powerful tools to address these difficulties. The application of these techniques for the automatic analysis of CVs in the financial field not only improves the efficiency and accuracy of the process but also opens new possibilities for analysis and decision-making based on that data. In this context, CA has been specifically designed to enhance the training and accuracy of models in text processing problems where quality data is scarce.

This new technique has proven to be particularly effective in the analysis of biographies of independent directors of listed companies, obtaining the best results (average hit rate of 92.5%, 97.5%, 99.7%, 99.6%, 96.5%, and 92.4% for the F, E/C, A/T/A, L, P, and Ac categories, respectively) and outperforming other comparable data augmentation methodologies such as LSR and WMix. Consequently, the results obtained offer a clear and affirmative response to both research questions posed in our study, as CA was shown to enhance the performance and generalization capabilities of DL models in financial NLP tasks and to outperform existing augmentation methods under the same conditions. Moreover, our proposal offers easy integration into existing models, ensuring a smooth implementation process with minimal adjustments. It features low computational cost, allowing for efficient resource use. Additionally, the model trained with CA-generated samples learns quickly, which leads to faster adaptation to new data. This enhances the training process and significantly increases the precision of the model. As a result, all these advantages of CA contribute to improved performance and efficiency in data analysis and modeling. From a practical perspective, CA constitutes a scalable and cost-effective solution for real-world applications, particularly in quality-data-scarce financial environments such as CV evaluation, compliance monitoring, or investment analysis. Its straightforward integration makes it an attractive tool for industry practitioners aiming to improve automated model reliability and efficiency. Moreover, its usefulness could be increased in future research by employing semi-supervised labeling techniques, adapting CA to other NLP tasks, and integrating it into more complex architectures. Additionally, using CA on multilingual datasets or different professional sectors would lead to broader applicability. Given all the above-mentioned points, we consider that CA can be integrated into the collection

of data augmentation techniques to generate variability from available data, facilitate regularization, and enhance the precision of DL models.

Author Contributions: Conceptualization, C.V., J.-Á.R.-G., V.B.-G., F.T. and B.S.; methodology, C.V., J.-Á.R.-G., V.B.-G., F.T. and B.S.; software, C.V. and B.S.; validation, C.V., J.-Á.R.-G., V.B.-G., F.T. and B.S.; formal analysis, C.V., V.B.-G. and B.S.; investigation, C.V., J.-Á.R.-G., V.B.-G., F.T. and B.S.; resources, C.V., V.B.-G., F.T. and B.S.; data curation, C.V., F.T. and B.S.; writing—original draft preparation, C.V., J.-Á.R.-G., V.B.-G., F.T. and B.S.; writing—review and editing, C.V., J.-Á.R.-G., V.B.-G., F.T. and B.S.; visualization, C.V.; supervision, J.-Á.R.-G., V.B.-G., F.T. and B.S.; project administration, J.-Á.R.-G. and B.S.; funding acquisition, J.-Á.R.-G. All authors have read and agreed to the published version of the manuscript.

Funding: Verónica Barroso-García was supported by the projects CPP2022-009735 and PID2023-148895OB-I00, funded by MICIU/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR. Her research was also funded by the "CIBER-Consorcio Centro de Investigación Biomédica en Red" (CB19/01/00012) through "Instituto de Salud Carlos III", co-funded with the European Regional Development Fund.

Institutional Review Board Statement: Not applicable as this study did not involve humans or animals.

Informed Consent Statement: Not applicable as this study did not involve humans.

Data Availability Statement: The data used in this study was obtained from the Annual Corporate Governance Reports issued by the National Securities Market Commission (Comisión Nacional del Mercado de Valores, CNMV), which are freely accessible at https://www.cnmv.es/ (accessed on 20 May 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACGR	Annual Corporate Governance Report
ASGD	Averaged Stochastic Gradient Descent
BERT	Bidirectional Encoder Representations of Transformers
CA	Concatenation Augmentation
CNMV	Comisión Nacional del Mercado de Valores (National Securities Market Commission)
CV	Curriculum Vitae (Professional Profile or Biography)
DL	Deep Learning
LSR	Label Smoothing Regularization
LSTM	Long Short-Term Memory Neural Network
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
NT-ASGD	Non-monotonically Triggered ASGD
RMSE	Root Mean Standard Error
RoBERTa	A Robustly Optimized BERT Pretaining Approach

References

1. Vasile, V.; Manta, O. FinTech and AI as Opportunities for a Sustainable Economy. *FinTech* **2025**, *4*, 10. [CrossRef]
2. Lewis, C.; Young, S. Fad or future? Automated analysis of financial text and its implications for corporate reporting. *Account. Bus. Res.* **2019**, *49*, 587–615. [CrossRef]
3. Wujec, M. Analysis of the Financial Information Contained in the Texts of Current Reports: A Deep Learning Approach. *J. Risk Financ. Manag.* **2021**, *14*, 582. [CrossRef]
4. Katib, I.; Assiri, F.Y.; Althaqafi, T.; AlKubaisy, Z.M.; Hamed, D.; Ragab, M. Hybrid Hunter–Prey Optimization with Deep Learning-Based Fintech for Predicting Financial Crises in the Economy and Society. *Electronics* **2023**, *12*, 3429. [CrossRef]

5. Skondras, P.; Zervas, P.; Tzimas, G. Generating Synthetic Resume Data with Large Language Models for Enhanced Job Description Classification. *Future Internet* **2023**, *15*, 363. [CrossRef]
6. Vaca, C.; Tejerina, F.; Sahelices, B. Board of Directors' Profile: A Case for Deep Learning as a Valid Methodology to Finance Research. *Int. J. Interact. Multimed. Artif. Intell.* **2022**, *7*, 101. [CrossRef]
7. Vaca, C.; Astorgano, M.; López-Rivero, A.J.; Tejerina, F.; Sahelices, B. Interpretability of deep learning models in analysis of Spanish financial text. *Neural Comput. Appl.* **2024**, *36*, 7509–7527. [CrossRef]
8. CNMV. Código de Buen Gobierno de las Sociedades Cotizadas. 2015. Available online: https://www.cnmv.es/DocPortal/Publicaciones/CodigoGov/CBG_2020.pdf (accessed on 3 May 2025).
9. Chowdhary, K.R. Natural Language Processing. In *Fundamentals of Artificial Intelligence*; Springer: New Delhi, India, 2020; pp. 603–649. [CrossRef]
10. Merity, S.; Keskar, N.S.; Socher, R. Regularizing and optimizing LSTM language models. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018. [CrossRef]
11. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186. [CrossRef]
12. Lyu, S.; Liu, J. Convolutional Recurrent Neural Networks for Text Classification. *J. Database Manag. (JDM)* **2021**, *32*, 65–82. [CrossRef]
13. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. Available online: http://www.deeplearningbook.org (accessed on 29 March 2024).
14. Hanson, S.; Pratt, L. Comparing Biases for Minimal Network Construction with Back-Propagation. *Adv. Neural Inf. Process. Syst.* **1988**, *1*, 177–185. [CrossRef]
15. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958. [CrossRef]
16. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning—ICML'15, Lille, France, 6–11 July 2015; Volume 37, pp. 448–456. [CrossRef]
17. Simard, P.Y.; LeCun, Y.A.; Denker, J.S.; Victorri, B. Transformation Invariance in Pattern Recognition—Tangent Distance and Tangent Propagation. In *Neural Networks: Tricks of the Trade: Second Edition*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 235–269. [CrossRef]
18. Kobayashi, S. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*; Walker, M., Ji, H., Stent, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 452–457. [CrossRef]
19. Feng, S.Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; Hovy, E. A Survey of Data Augmentation Approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 968–988. [CrossRef]
20. Shorten, C.; Khoshgoftaar, T.M.; Furht, B. Text Data Augmentation for Deep Learning. *J. Big Data* **2021**, *8*, 101. [CrossRef] [PubMed]
21. Chen, J.; Tam, D.; Raffel, C.; Bansal, M.; Yang, D. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 191–211. [CrossRef]
22. Wang, W.Y.; Yang, D. That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 17–21 September 2015; pp. 2557–2563. [CrossRef]
23. Zhang, X.; Zhao, J.; LeCun, Y. Character-level Convolutional Networks for Text Classification. In *Proceedings of the Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; ACM : Cambridge, MA, USA, 2015; Volume 28. [CrossRef]
24. Edunov, S.; Ott, M.; Auli, M.; Grangier, D. Understanding Back-Translation at Scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 489–500. [CrossRef]
25. Wei, J.; Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Stroudsburg, PA, USA, 3–7 November 2019; pp. 6381–6387. [CrossRef]
26. Andreas, J. Good-Enough Compositional Data Augmentation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 7556–7566. [CrossRef]

27. Yoon, S.; Kim, G.; Park, K. SSMix: Saliency-Based Span Mixup for Text Classification. *arXiv* **2021**, arXiv:2106.08062.
28. Kong, F.; Zhang, R.; Guo, X.; Mensah, S.; Mao, Y. DropMix: A Textual Data Augmentation Combining Dropout with Mixup. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 890–899. [CrossRef]
29. Morris, J.; Lifland, E.; Yoo, J.Y.; Grigsby, J.; Jin, D.; Qi, Y. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; Liu, Q., Schlangen, D., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 119–126. [CrossRef]
30. Chen, J.; Shen, D.; Chen, W.; Yang, D. HiddenCut: Simple Data Augmentation for Natural Language Understanding with Better Generalizability. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 4380–4390. [CrossRef]
31. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018. [CrossRef]
32. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
33. Manioudis, M.; Meramveliotakis, G. Broad strokes towards a grand theory in the analysis of sustainable development: A return to the classical political economy. *New Political Econ.* **2022**, *27*, 866–878. [CrossRef]
34. Fu, C.; Lu, L.; Pirabi, M. Advancing green finance: A review of sustainable development. *Digit. Econ. Sustain. Dev.* **2023**, *1*, 20. [CrossRef]
35. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
36. Mumuni, A.; Mumuni, F. Data augmentation: A comprehensive survey of modern approaches. *Array* **2022**, *16*, 100258. [CrossRef]
37. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio augmentation for speech recognition. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015; pp. 3586–3589. [CrossRef]
38. Song, X.; Wu, Z.; Huang, Y.; Su, D.; Meng, H. SpecSwap: A Simple Data Augmentation Method for End-to-End Speech Recognition. In Proceedings of the Interspeech 2020, Online, 19 October 2020; pp. 581–585. [CrossRef]
39. Meng, L.; Xu, J.; Tan, X.; Wang, J.; Qin, T.; Xu, B. MixSpeech: Data Augmentation for Low-Resource Automatic Speech Recognition. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Online, 6–11 June 2021; pp. 7008–7012. [CrossRef]
40. Chou, H.P.; Chang, S.C.; Pan, J.Y.; Wei, W.; Juan, D.C. Remix: Rebalanced Mixup. In Proceedings of the Computer Vision—ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Bartoli, A., Fusiello, A., Eds.; Springer: Cham, Switzerland, 2020; pp. 95–110. [CrossRef]
41. Dai, X.; Zhao, X.; Cen, F.; Zhu, F. Data Augmentation Using Mixup and Random Erasing. In Proceedings of the 2022 IEEE International Conference on Networking, Sensing and Control (ICNSC), Online, 20–23 October 2022; pp. 1–6. [CrossRef]
42. Psaroudakis, A.; Kollias, D. MixAugment & Mixup: Augmentation Methods for Facial Expression Recognition. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Los Alamitos, CA, USA, 19–22 June 2022; pp. 2366–2374. . [CrossRef]
43. Guo, H. Nonlinear Mixup: Out-Of-Manifold Data Augmentation for Text Classification. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 4044–4051. [CrossRef]
44. Sun, L.; Xia, C.; Yin, W.; Liang, T.; Yu, P.; He, L. Mixup-Transformer: Dynamic Data Augmentation for NLP Tasks. In Proceedings of the International Conference on Computational Linguistics, Online, 8–13 December 2020; pp. 3436–3440. [CrossRef]
45. Open AI. GPT-4 Technical Report. 2024. Available online: https://cdn.openai.com/papers/gpt-4.pdf (accessed on 3 May 2025).
46. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288.
47. Colasanto, F.; Grilli, L.; Santoro, D.; Villani, G. BERT's sentiment score for portfolio optimization: A fine-tuned view in Black and Litterman model. *Neural Comput. Appl.* **2022**, *34*, 17507–17521. [CrossRef]
48. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
49. Ebrahimi, J.; Rao, A.; Lowd, D.; Dou, D. Hotflip: White-box adversarial examples for text classification. *arXiv* **2017**, arXiv:1712.06751.
50. Wu, X.; Lv, S.; Zang, L.; Han, J.; Hu, S. Conditional BERT Contextual Augmentation. In *Proceedings of the Computational Science—ICCS 2019, Faro, Portugal, 12–14 June 2019*; Rodrigues, J.M.F., Cardoso, P.J.S., Monteiro, J., Lam, R., Krzhizhanovskaya, V.V., Lees, M.H.,Dongarra, J.J., Sloot, P.M., Eds.; Springer: Cham, Switzerland, 2019; pp. 84–95. [CrossRef]

51. Luo, J.; Luo, X.; Chen, X.; Xiao, Z.; Ju, W.; Zhang, M. SemiEvol: Semi-supervised Fine-tuning for LLM Adaptation. *arXiv* **2024**, arXiv:2410.14745.
52. Güner, B.; Malmendier, U.; Tate, G. Financial expertise of directors. *J. Financ. Econ.* **2018**, *88*, 323–354. [CrossRef]
53. Booth, J.; Deli, D. On executives of financial institutions as outside directors. *J. Corp. Financ.* **1999**, *5*, 227–250. [CrossRef]
54. Faleye, O.; Hoitash, R.; Hoitash, U. Industry expertise on corporate boards. *Rev. Quant. Financ. Account.* **2018**, *50*, 441–2479. [CrossRef]
55. Fich, E. Are Some Outside Directors Better than Others? Evidence from Director Appointments by Fortune 1000 Firms. *J. Bus.* **2005**, *78*, 1943–1972. [CrossRef]
56. Qiao, Z.; Chen, K.; Hung, S. Professionals inside the board room: Accounting expertise of directors and dividend policy. *Appl. Econ.* **2018**, *50*, 6100–6111. [CrossRef]
57. Krishnan, J.; Wen, Y.; Zhao, W. Legal Expertise on Corporate Audit Committees and Financial Reporting Quality. *Account. Rev.* **2011**, *86*, 2099–2130. [CrossRef]
58. Houston, J.; Jiang, L.; Lin, C.; Ma, Y. Political Connections and the Cost of Bank Loans. *J. Account. Res.* **2014**, *52*, 193–243. [CrossRef]
59. Agrawal, A.; Knoeber, C. Do Some Outside Directors Play a Political Role? *J. Law Econ.* **2001**, *44*, 179–198. [CrossRef]
60. Xie, Y.; Xu, J.; Zhu, R. Academic Directors and Corporate Innovation. *SSRN Pap.* **2021**. [CrossRef]
61. Briskilal, J.; Subalalitha, C.N. Classification of Idiomatic Sentences Using AWD-LSTM. In *Proceedings of the Expert Clouds and Applications*; Jeena Jacob, I., Gonzalez-Longatt, F.M., Kolandapalayam Shanmugam, S., Izonin, I., Eds.; Springer: Singapore, 2022; pp. 113–124. [CrossRef]
62. Kiran, S.; Shashi, M.; Madhuri, K. Multi-stage Transfer Learning for Fake News Detection Using AWD-LSTM Network. *Int. J. Inf. Technol. Comput. Sci.* **2022**, *14*, 58–69. [CrossRef]
63. Wang, C.; Li, M.; Smola, A. Language Models with Transformers. *arXiv* **2019**, arXiv:1904.09408.
64. Zhang, J.G.; Ping Li, J.; Li, H. Language Modeling with Transformer. In Proceedings of the 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 13–15 December 2019; pp. 249–253. [CrossRef]
65. Aziz, S.; Dowling, M.; Hammami, H.; Piepenbrink, A. Machine learning in finance: A topic modeling approach. *Eur. Financ. Manag.* **2022**, *28*, 744–770. [CrossRef]
66. Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; Gurevych, I., Miyao, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 328–339. [CrossRef]