When Modernity Enhances Tradition and Specialization: A BERT Ensemble for Temporal Financial Argument Detection

Leonardo Martinho[1‡], Hugo Dutra[1‡], Gabriel Assis[1], Jonnathan Carvalho[2], Aline Paes[1]

[1] Universidade Federal Fluminense, Brazil
{leonardoalvesmartinho, hugo_dutra, assisgabriel}@id.uff.br,
alinepaes@ic.uff.br
[2] Instituto Federal Fluminense, Brazil
joncarv@iff.edu.br

Abstract. Temporal references in financial texts, such as earnings conference calls (ECCs), are essential for interpreting corporate discourse and guiding informed investment decisions. However, accurately identifying these references remains a challenge due to domain-specific language and vague temporal cues. In this work, we investigate whether combining distinct BERT-based models can improve performance on temporal financial argument detection. We evaluate an ensemble approach built upon three complementary models: BERT, FinBERT, and ModernBERT. Experiments on the FinArg-2 ECC dataset show that, while ModernBERT performs best individually, the ensemble of all three models — using a soft-voting strategy — sets a new state of the art. These results highlight the potential of BERT-based ensembles for more accurate and robust temporal reasoning in financial NLP tasks.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**; **Ensemble methods**.

Keywords: BERT, Ensembles, Finance, Temporal Argument Classification

1. INTRODUCTION

In today's financial landscape, access to vast volumes of textual data — from Earnings Conference Calls (ECCs) to investor forums — has reshaped how market participants track performance and assess corporate narratives [Chen et al. 2025]. These sources often contain time-sensitive claims that, while not always explicit, carry significant weight in driving investment decisions. Understanding *when* a financial assertion holds true — its scope, relevance, and duration — has become just as critical as understanding the claim itself [Chen et al. 2018; Chen et al. 2025]. This is particularly true in the context of high-stakes communications like ECCs, where companies frame events in strategic ways, using temporal ambiguity to emphasize gains or even to soften the perception of setbacks [Crawford Camiciottoli 2017].

However, pinpointing temporal information in financial texts remains a persistent challenge [UzZaman et al. 2013]. Unlike structured data, these narratives rely heavily on implicit cues and domain-specific language that obscure precise time anchoring [Dutra et al. 2025]. The difficulty is compounded by the need to distinguish between short-term fluctuations and long-term trends [Chen et al. 2025], especially in fast-moving environments like stock markets. Accurate temporal interpretation plays a central role in evaluating the potential impact of reported events and in determining the shelf life of

The authors acknowledge the support of Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), National Institute of Artificial Intelligence (INCT IAIA), Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), under grant numbers SEI-260003/002930/2024 and SEI-260003/000614/2023, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Financial Code 001. ‡ Equal contribution.

investment opinions [Chen et al. 2018]. Without clear temporal grounding, forecasts lose practical value, and critical market signals may be misread, highlighting the importance of developing robust tools to navigate this complexity.

The widespread adoption of Transformer-based Language Models (LMs) has transformed natural language processing (NLP), also extending its impact to domain-specific applications such as finance [Araci 2019; Erfina and Le-Hong 2025; Nandam et al. 2025]. In this context, encoder-based models like BERT [Devlin et al. 2019] have proven highly effective, supporting tasks ranging from sentiment classification [Araci 2019] to the detection of temporal arguments in financial discourse itself [Chen et al. 2025]. Their bidirectional attention mechanism allows for fine-grained text encoding, which, when properly fine-tuned, often results in performance that surpasses that of larger models (LLMs) like GPT-4o [OpenAI 2024] in specialized tasks, while incurring significantly lower costs [Dutra et al. 2025; Chen et al. 2025; Chen et al. 2025]. Even among BERT-based models, notable performance differences can be observed. Although these models share the same foundational principles, their variants often differ in key aspects such as minor architecture changes, training strategies, and pretraining corpora. These differences give each model its own unique characteristics. In this scenario, several studies have explored the use of different BERT variants as backbones for financial NLP tasks. For instance, Dutra et al. [2025] experimented with DeBERTa [He et al. 2021], Nandam et al. [2025] and Erfina and Le-Hong [2025] used the standard BERT; and You et al. [2025] employed FinBERT [Araci 2019] and ModernBERT [Warner et al. 2024] for their applications.

Instead of focusing on selecting the best individual model, as is commonly done, this study poses a different question: *Can combining distinct BERT-based models in an ensemble improve performance in temporal financial argument detection?* We build on prior research that highlights the potential of BERT-based ensembles in other domains [Amorim et al. 2024], and evaluate whether models with complementary characteristics can be jointly leveraged for improved results. Our experiments use three models that reportedly perform well in the task [Chen et al. 2025]: the original BERT, FinBERT — pretrained specifically on financial texts — and ModernBERT, which integrates recent advances in Transformer architecture into the standard encoder pipeline. We evaluate our approach on the FinArg-2 ECC dataset [Chen et al. 2025], a challenging standard benchmark for temporal reference identification in financial discourse. Our findings show that, individually, ModernBERT achieves the strongest results. Pairwise combinations with BERT and FinBERT lead to further improvements, but the best performance is attained by the full ensemble of all three models. This configuration surpasses the current state of the art [Erfina and Le-Hong 2025] — based on the traditional BERT — by 3.10 percentage points in micro-F1 and 3.30 points in macro-F1, demonstrating the effectiveness of our proposed ensemble approach.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the proposed method, including the data and the models employed, and Section 4 reports and discusses the results. Finally, Section 5 presents conclusions and avenues for future work.

2. RELATED WORK

This section reviews related work focused on temporal argument references in finance. Particularly, temporal relationships in financial arguments may lose their validity rapidly, as they are shaped by the dynamic nature of the domain. This contrasts with other fields, where such references may remain relevant over longer periods [Chen et al. 2021]. Consequently, several studies have concentrated on the classification and detection of temporal references in this context [Chen et al. 2025].

Chen et al. [2018] proposed a taxonomy for interpreting numerals in financial data, identifying categories such as monetary values, percentages, quantities, and — importantly — temporal information. This taxonomy emphasizes the need to account for the temporal dimension in financial texts, as numerals associated with specific time periods can significantly affect the interpretation of forecasts

and predictions. In the same vein, [Lin et al. 2024] conducted an argument-based sentiment analysis aimed at identifying and assessing forward-looking statements in financial documents. Their study highlights the challenges posed by such texts.

The importance of accurately detecting temporal references in financial arguments has motivated the creation of a dedicated shared task: FinArg-2 — Detection of Argument Temporal References in ECCs [Chen et al. 2025]. Among the approaches applied to this task, encoder-based models such as BERT have shown considerable promise. For instance, You et al. [2025] employed ModernBERT for classification, concatenating all available features as input to the LM. Erfina and Le-Hong [2025] achieved the highest performance reported on the task to date. Their approach combined a traditional BERT with a TF-IDF (term frequency–inverse document frequency, a weighting scheme that reflects the importance of terms in relation to specific documents) based representation, in which temporal information — such as the year and financial quarter — was encoded using one-hot vectors and concatenated with the LM embeddings on the main text for final prediction.

Also within the FinArg-2 shared task, Nandam et al. [2025] report strong performance by simply fine-tuning a BERT model for three-class classification, demonstrating that such encoder models can yield effective results even under relatively simple configurations. The authors also experimented with an enhanced temporal expression knowledge base, though this resulted in lower performance. Dutra et al. [2025] explored a similar three-label setup using DeBERTa variants, along with strategies for data manipulation similar to those of You et al. [2025]. In addition, they investigated a cascaded approach, dividing the problem into two binary classification tasks and also experimented with prompt engineering using LLMs such as GPT-4o. Nonetheless, their findings confirmed that encoder-based classifiers still achieved the best performance. Close results are reported by Chen et al. [2025], who contrast a RoBERTa-based [Liu et al. 2020] model with a GPT-4o instance specifically fine-tuned for the task, highlighting the superior performance of the first.

Although previous work has demonstrated the strong performance of BERT-based models on temporal argument reference tasks in finance, these approaches typically involve selecting a single model as the backbone of the solution. In contrast, our work proposes an ensemble of distinct BERT-based models for this task, aiming to investigate whether architectural and training-related differences among them can lead to complementary performance gains.

3. MATERIALS AND METHODS

This section outlines the investigated models, presents the data and the preprocessing procedures applied, describes the model application designs — including the ensemble strategy —, and specifies the evaluation metrics used in the experiments.

3.1 Investigated Models

We explore three LMs, namely (i) BERT [Devlin et al. 2019], (ii) FinBERT [Araci 2019], and (iii) ModernBERT [Warner et al. 2024]. Beyond their prior successful applications to financial temporal reference tasks [You et al. 2025; Erfina and Le-Hong 2025; Nandam et al. 2025; Chen et al. 2025], these models were selected for their distinct characteristics.

The first, BERT, serves as the foundation for subsequent developments in both bidirectional encoders and the Transformer architecture itself. FinBERT, in turn, is pretrained on curated financial data, offering domain-specific specialization in terminology and financial jargon. Finally, ModernBERT incorporates recent advances in attention mechanisms and neural layer design within the standard Transformer-based encoder. This selection not only enables the assessment of each model's individual performance but also introduces model diversity, allowing for an evaluation of their combined impact within the proposed ensemble strategy.

3.2 Dataset

We conduct our experiments using the NTCIR-18 FinArg-2 shared task dataset on ECCs [Chen et al. 2025]. Specifically, ECCs refer to quarterly corporate meetings intended to communicate a company's past financial performance and outline forward-looking business projections [Erfina and Le-Hong 2025]. We selected this dataset as it is a well-established benchmark in temporal reference detection, extensively documented and cited in the recent literature, which supports direct comparison with state-of-the-art approaches.

The data format is illustrated in Figure 1. Each instance in the dataset consists of a claim (a financial argument) and a list of premises (statements supporting it). Temporal features are also provided, specifically the quarter (Q1–Q4) and year of publication. In addition, each instance is labeled according to the type of temporal reference found in the claim: class 0 indicates *no temporal reference*, class 1 denotes a *long past reference* (more than half a year before publication), and class 2 denotes a *short past reference* (less than half a year). We preserve the original training (600 instances), validation (150 instances), and test (84 instances) splits provided with the dataset to allow for consistent comparison with previous works.

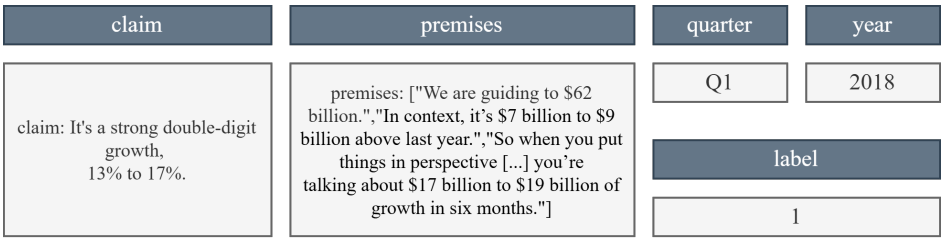

Fig. 1. Visualization of a NTCIR-18 FinArg-2 ECC dataset instance.

3.3 Data Preprocessing

We adopted an approach that combines all attributes of a data instance into a single textual input for the LM. We automatically merged claims and premises into unified paragraphs, keeping the original order of premises. The claim appeared either at the beginning — if the first premise started with a lowercase letter — or after the premises, if it began with a connector like "`And`".

Additionally, the publication date was incorporated at the end of each input to clarify the temporal context. For example, if an argument came from `Q1 2018`, we appended the sentence "`This publication is from Q1 of 2018.`" We chose this sentence-based structure given the known advantage of LMs in processing contextualized textual data over purely numerical inputs [Cunha et al. 2024]. Figure 2 illustrates the preprocessing pipeline.

Fig. 2. Data preprocessing for constructing model inputs.

3.4 Model Application Designs

In this section, we detail the application of the models introduced in Section 3.1 to the task of classifying financial temporal argument references.

3.4.1 *Base Models: Direct 3-Label Classification.*

As an initial setup, we performed traditional fine-tuning of the LMs [Paes et al. 2024]. Specifically, we attached linear classification layers on top of the LM's encoder stack and fine-tuned all parameters, including those of the LM itself. This setup allows the model to directly predict one of the three temporal reference classes given an input. For training, we used the dataset's training split, applying the transformation described in Section 3.3. Experimental settings are provided in Section 3.5.

3.4.2 *BERT-based Ensembles.*

In our ensemble approach, we adopted soft-voting [Zhou 2012] to combine the outputs of BERT-based models. Each model produced a probability distribution over the target classes, and we computed the average probability per class across all models. The final prediction corresponds to the class with the highest average probability. Formally, the class c selected is:

$$c_x = \arg\max_c \left(\frac{1}{m} \sum_{i=1}^{m} p_c^i(x) \right) \tag{1}$$

where x is the input instance and m is the number of models in the ensemble.

Unlike hard-voting, which treats all final and individual predictions equally, soft-voting considers the confidence of each model in its predictions. By averaging probabilities rather than final class labels, the ensemble can make more informed decisions — often resulting in better overall results [Zhou 2012; Amorim et al. 2024]. To form our ensemble, we reuse the same models trained for the baseline experiments in the previous section.

3.5 Evaluation and Experimental Setup

Micro- and macro-F1 scores are used as metrics for evaluating model performance, as they are the official metrics adopted by the NTCIR-18 FinArg-2 task [Chen et al. 2025]. While micro-F1 may be less reliable in imbalanced scenarios such as the NTCIR-18 FinArg-2 ECCs dataset, we report it to enable comparison with prior work. Nonetheless, our main analysis centers on the macro-F1 score.

All experiments were conducted using the Hugging Face Transformers [Wolf et al. 2020] framework running on Google Colab[1] T4 TPUs, with the `random seed` fixed at 42 for reproducibility. We rely on prior preliminary cross-validation experiments to determine the hyperparameter setup [Dutra et al. 2025]. Specifically, we set the `learning_rate` to 2e-5, `epoch` to 12, and `batch_size` to 8 for model fine-tuning, while keeping the remaining configurations at their default values.

4. RESULTS AND DISCUSSION

This section presents and discusses the experimental results. First, Table I presents the results on the test set for the models described in Section 3. The best individual scores are underlined, while the overall best results are highlighted in **bold**. We observe that, individually, ModernBERT performs the strongest, tying with FinBERT in terms of micro-F1, while achieving the highest macro-F1. The

[1] `https://colab.google`

gap between micro and macro scores suggests that, although both models perform similarly in general, ModernBERT yields more balanced results across classes. Furthermore, the advantage of the domain-specific FinBERT and the more recent ModernBERT over the traditional BERT highlights how both domain specialization and architectural advancements can positively impact final performance.

Table I. Classification performance of individual models and model ensembles.

Model / Ensemble	Micro-F1	Macro-F1
Individual Models		
BERT	0.726	0.701
FinBERT	0.762	0.722
ModernBERT	0.762	0.750
Ensembles ($\pm\delta$ from best individual)		
BERT + FinBERT	0.726 (-4.7%)	0.696 (-7.2%)
BERT + ModernBERT	0.774 (+1.6%)	0.762 (+1.6%)
FinBERT + ModernBERT	0.786 (+3.1%)	0.765 (+2.0%)
BERT + FinBERT + ModernBERT	**0.798** (+4.7%)	**0.776** (+3.5%)

The ensemble results in Table I further highlight the strength of ModernBERT. All ensembles that include this model outperform the best individual result, while the ensemble excluding it shows a decline in performance. The best overall results are achieved through the combination of all three base models, yielding gains of 3.5% in macro-F1 compared to ModernBERT alone. These findings demonstrate the potential of ensemble approaches to enhance the classification of temporal financial argument references.

Figure 3 presents the model performance metrics along with statistical confidence intervals. We used paired bootstrap with replacement [Efron and Tibshirani 1994], with $B = 10,000$ iterations. The `BERT + FinBERT + ModernBERT` ensemble outperforms the other models by approximately one percentage point. Although the confidence intervals exhibit some degree of overlap across models, the ensemble's consistent superiority across all metrics suggests a non-negligible improvement, particularly in the context of temporal reference detection in finance, where small error reductions can prevent misinterpretation of critical dates and timeframes in corporate reports. These seemingly modest gains may, in fact, carry substantial economic significance, as even minor improvements in temporal accuracy can enhance the reliability of risk models, improve the timing of financial forecasts, and reduce errors in trading strategies that rely on precise chronological information. This further reinforces the value of ensemble methods for this task.

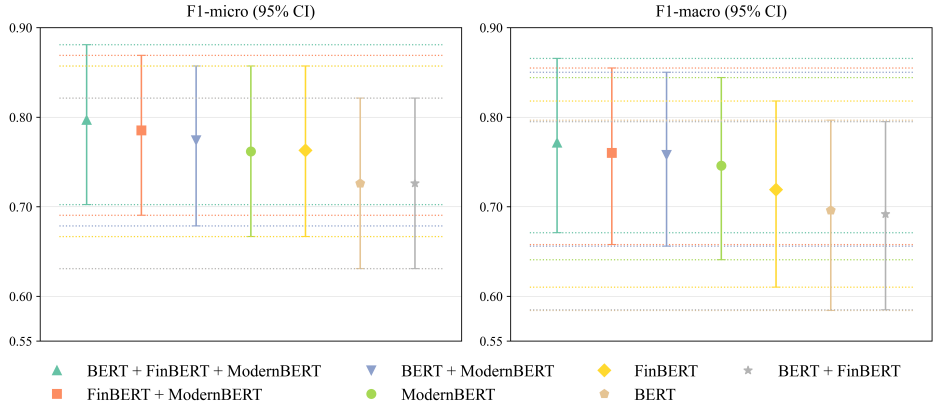

Fig. 3. Performance of the investigated approaches with confidence intervals.

Lastly, Table II compares our best results with previous work that investigated the same dataset. Our BERT-based ensemble of three models outperforms even the current state of the art [Erfina and Le-Hong 2025], a BERT model combined with TF-IDF features, by just over 3%. Again, while this may appear modest, it represents a non-negligible improvement in the financial domain. The table also shows that our individual runs of FinBERT and ModernBERT outperform prior approaches, ranking just below Erfina and Le-Hong [2025]. Notably, our ModernBERT configuration surpasses another ModernBERT implementation by You et al. [2025]. These results suggest that, beyond the effectiveness of our ensemble approach, our data preprocessing procedure may also enhance the standalone performance of LMs.

Table II. Comparison of our best-performing approach with literature.

Model	**Micro-F1**	**Macro-F1**
BERT + FinBERT + ModernBERT (this work)	**0.798**	**0.776**
BERT + TF-IDF [Erfina and Le-Hong 2025]	0.774	0.751
BERT ([Nandam et al. 2025])	0.702	0.679
RoBERTa ([Chen et al. 2025])	0.691	0.671
mDeBERTa ([Dutra et al. 2025])	0.691	0.671
ModernBERT ([You et al. 2025])	0.691	0.661

5. CONCLUSION

In response to the main question posed — *Can combining distinct BERT-based models in an ensemble improve performance in temporal financial argument detection?* — this research concludes that the use of BERT-based ensembles indeed enhances performance on this task. Our results establish a new state of the art. The most effective ensemble is based on soft-voting and combines BERT, FinBERT, and ModernBERT, with the latter standing out not only as a component of the best-performing ensemble but also for improving performance in pairwise combinations with the other models.

As future work, we plan to explore ensemble configurations in other financial classification tasks, such as sentiment analysis [Araci 2019], to assess whether the performance gains observed in temporal argument detection generalize to different settings. Additionally, we aim to investigate alternative ensemble strategies beyond soft-voting, such as stacking [Zhou 2012], which allows for learning how to optimally combine predictions from multiple models. This direction may offer further performance improvements and deeper insight into how model complementarities can be exploited in financial NLP.

Disclaimer The authors acknowledge the use of AI tools, specifically LLMs (ChatGPT-4o and o3), for grammar corrections and assistance with LaTeX and Python code. All outputs were reviewed by the authors, who take full responsibility for the content of this work.

REFERENCES

Amorim, A., Assis, G., Oliveira, D., and Paes, A. Multiple Voices, Greater Power: A Strategy for Combining Language Models to Combat Hate Speech. In *Anais do XII Symposium on Knowledge Discovery, Mining and Learning*. SBC, Porto Alegre, RS, Brasil, pp. 121–128, 2024.

Araci, D. FinBert: Financial Sentiment Analysis with Pre-trained Language Models, 2019.

Chen, B.-J., Hsiao, W.-H., Wu, J.-Y., Wu, C.-Y., and Day, M.-Y. IMNTPU at the NTCIR-18 Finarg-2: Fine-Tuning and Prompt-Based Learning for Temporal Argument Detection and Claim Validity Assessment. In *18th NTCIR Conference on Evaluation of Information Access Technologies*. NII Institutional Repository, 2025.

Chen, C.-C., Huang, H.-H., and Chen, H.-H. *From opinion mining to financial argument mining*. Springer Nature, 2021.

Chen, C.-C., Huang, H.-H., Shiue, Y.-T., and Chen, H.-H. Numeral Understanding in Financial Tweets for Fine-Grained Crowd-Based Forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, pp. 136–143, 2018.

Chen, C.-C., Lin, C.-Y., Chiu, C.-C., Huang, H.-H., Alhamzeh, A., Huang, Y.-L., Takamura, H., and Chen, H.-H. Overview of the NTCIR-18 Finarg-2 Task: Temporal Inference of Financial Arguments. In *18th NTCIR Conference on Evaluation of Information Access Technologies*. NII Institutional Repository, 2025.

Crawford Camiciottoli, B. Persuasion in Earnings Calls: A Diachronic Pragmalinguistic Analysis. *International Journal of Business Communication* 55 (3): 275–292, 2017. (Original work published 2018).

Cunha, R., Chinonso, O., Campos, J., Timoney, B., Davis, B., Cozman, F., Pagano, A., and Castro Ferreira, T. Imaginary Numbers! Evaluating Numerical Referring Expressions by Neural End-to-End Surface Realization Systems. In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*. ACL, Mexico City, Mexico, pp. 73–81, 2024.

Devlin, J., Chang, M., Lee, K., and 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*. pp. 4171–4186, 2019.

Dutra, H., Martinho, L., Assis, G., Carvalho, J., and Paes, A. AIDAVANCE at the NTCIR-18 Finarg-2 Task: Making the Most of Small Language Models. In *18th NTCIR Conference on Evaluation of Information Access Technologies*. NII Institutional Repository, 2025.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

Erfina, A. and Le-Hong, P. FTRI at the NTCIR-18 Finarg-2 Task: Identify Temporal Reference in Earnings Conference Calls. In *18th NTCIR Conference on Evaluation of Information Access Technologies*. NII Institutional Repository, 2025.

He, P., Liu, X., Gao, J., and Chen, W. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021.

Lin, C.-Y., Chen, C.-C., Huang, H.-H., and Chen, H.-H. Argument-Based Sentiment Analysis on Forward-Looking Statements. In *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, pp. 13804–13815, 2024.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Nandam, S. S., Dasari, C. S. K. R., and Madasamy, A. K. SCaLAR IT at the NTCIR-18 Finarg-2: Temporal Inference of Financial Arguments. In *18th NTCIR Conference on Evaluation of Information Access Technologies*. NII Institutional Repository, 2025.

OpenAI, T. Gpt-4o system card, 2024.

Paes, A., Vianna, D., and Rodrigues, J. Modelos de linguagem. In *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 ed., H. M. Caseli and M. G. V. Nunes (Eds.). BPLN, Book chapter 17, 2024.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, S. Manandhar and D. Yuret (Eds.). Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 1–9, 2013.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., and Poli, I. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen (Eds.). Association for Computational Linguistics, Online, pp. 38–45, 2020.

You, X.-Y., Liew, D. J., Yeh, W.-C., and Chang, Y.-C. TMUNLPG1 at the NTCIR-18 Finarg-2 Task. In *18th NTCIR Conference on Evaluation of Information Access Technologies*. NII Institutional Repository, 2025.

Zhou, Z.-H. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 2012.