A Brazilian Dataset for Financial Sentiment Analysis: An Active Learning-Based Annotation Framework

Ramon Abilio[1,2], Guilherme Palermo Coelho[2], Ana Estela Antunes da Silva[2]

[1] Instituto Federal de São Paulo (IFSP), Brazil
ramon.abilio@ifsp.edu.br
[2] Faculdade de Tecnologia (FT), Universidade Estadual de Campinas (Unicamp), Brazil
{gpcoelho,aeasilva}@unicamp.br

Abstract. This paper introduces a new expert-annotated dataset for financial sentiment analysis in Brazilian Portuguese, constructed from earnings call transcripts of the banking sector. The dataset was developed using a two-stage active learning pipeline designed to reduce annotation cost while preserving data quality. In the first stage, clustering and active learning were combined to filter relevant sentences from noisy transcripts. In the second stage, a Query by Committee strategy combining the transformer models BERTimbau and DeB3RTa was employed to select informative samples for expert annotation. The resulting dataset contains 719 annotated sentences with agreement analysis across multiple annotators. We further evaluated BERTimbau, DeB3RTa, and the external transformer model BERTuguês under three experimental settings, including fine-grained eight-class classification and complementary three-class scenarios. In the eight-class setting, BERTuguês and BERTimbau achieved comparable F1-macro scores (0.43 ± 0.03 and 0.43 ± 0.01, respectively). In the sentiment-only three-class setup, BERTimbau achieved the best performance (0.68 ± 0.06). These results suggest that the dataset supports both fine-grained and coarse-grained sentiment analysis while remaining useful for models not involved in the active learning process.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Language resources**; **Active learning settings**.

Keywords: deep learning, machine learning, natural language processing, transformer

1. INTRODUCTION

The availability of high-quality annotated datasets has played an important role in the advancement of sentiment analysis and other Natural Language Processing (NLP) tasks [Mostafavi et al. 2026]. In the financial domain, several datasets for sentiment analysis have been proposed in different languages, covering sources such as news articles, banking documents, social media posts, and customer reviews [Malo et al. 2014; Raicu 2019; Lan et al. 2025; Alawaji and Aloraini 2025].

Although some resources exist for Brazilian Portuguese [da Silva et al. 2020; Januário et al. 2022; Santos et al. 2023; Kakimoto et al. 2024], they mainly rely on news or social media posts, while datasets based on Brazilian corporate disclosures remain scarce. Among corporate disclosure sources, earnings call transcripts are particularly relevant but still underexplored. Earnings calls are formal events in which company executives discuss financial results and future expectations, followed by questions from analysts and investors. As a result, they provide information directly related to business performance, risks, opportunities, and strategic decisions, however they also contain non-informative passages that can introduce substantial noise.

In the Brazilian context, building datasets from earnings call transcripts is even more challenging. Earnings call transcripts are typically released after quarterly earnings announcements, and they are

This study was financed in part by the Coordenaç ao de Aperfeiçoamento de Pessoal de N'ivel Superior - Brasil (CAPES) - Finance Code 001, and supported by the Programa de Pós-Graduação em Tecnologia (PPGT).

usually distributed as PDF files across individual investor relations websites rather than stored in a centralized public repository. In addition, file structures may vary across institutions and over time, which makes text extraction harder and increases the presence of fragmented or noisy sentences. This reinforces the need for careful preprocessing and explicit relevance filtering before sentiment annotation.

Another practical challenge is annotation cost. Labeling sentence-level financial data usually requires domain expertise, which limits scalability. In addition, high-quality annotation often benefits from multiple annotators and agreement analysis to reduce individual biases and improve reliability [Pustejovsky and Stubbs 2013]. Active learning is an iterative sampling strategy in which a model selects the most informative unlabeled examples for annotation instead of relying solely on random sampling [Li et al. 2025]. By focusing annotation effort on more informative cases, this strategy can reduce manual effort while improving dataset quality.

Motivated by these challenges, the main goal of this work is to introduce a new gold-standard, expert-annotated dataset for financial sentiment analysis in Brazilian Portuguese based on earnings call transcripts from the banking sector, built with a two-stage active learning pipeline. The dataset adopts sentence-level annotation and explicitly separates sentence relevance from sentiment polarity and intensity. The construction process also includes inter-annotator agreement analysis to support annotation quality. In addition, we evaluate the dataset under multiple experimental settings, positioning it as both a training resource and a benchmark for financial sentiment modeling in Portuguese.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed methodology. Section 4 presents the experiments and discusses the results. Section 5 concludes the paper and suggests directions for future work.

2. RELATED WORK

Representative studies on financial sentiment analysis in Brazilian Portuguese include da Silva et al. [2020], who created a corpus of stock market tweets in Portuguese with 4,517 manually annotated instances focused on emotion detection rather than sentiment polarity. Using a different source, Januário et al. [2022] compiled a dataset of 828 financial news headlines about Brazilian listed companies, labeled at text level as positive or negative from an investor perspective, although no inter-annotator agreement was reported. Later, Santos et al. [2023] proposed a labeled subset of 503 financial news headlines annotated into positive, negative, and neutral classes, with expert annotation and high Krippendorff's Alpha agreement ($\alpha = 0.88$). More recently, Kakimoto et al. [2024] introduced the Brazilian Tweet Banks Dataset (B2T), composed of 1,096 tweets related to Brazilian banks annotated into positive, negative, and neutral classes, with Cohen Kappa agreement scores ranging from moderate (0.56) to substantial (0.67). Recent studies have also explored automatic annotation and sentiment interpretation in financial texts using large language models [Romero et al. 2025].

Beyond differences in data source and label scheme, prior studies also vary in annotation design, including volunteer-based, crowdsourced, and expert-driven approaches, with heterogeneous reporting of inter-annotator agreement. Most work relies on random sampling or manually defined subsets for data selection. Although active learning can optimize annotation by prioritizing informative instances, it has been explored mainly in simulated settings rather than as part of real dataset construction [Ashrafi Asli et al. 2020; Liebenlito et al. 2024; Kaseb and Farouk 2023].

Overall, prior studies focus primarily on tweets or financial news, while sentence-level resources based on corporate disclosures remain scarce. This limitation is particularly relevant given the mixture of informative and non-informative content in financial documents. In addition, the use of active learning in financial sentiment dataset construction in Brazilian Portuguese remains limited, particularly for filtering relevant texts for sentiment analysis. To address these gaps, we introduce a sentence-level dataset based on earnings call transcripts, constructed using a two-stage active learning

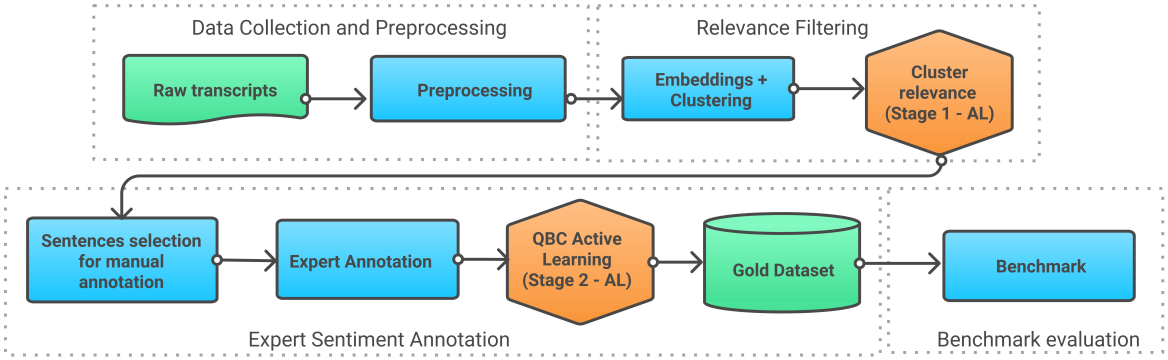

Fig. 1. Methodology overview

pipeline with expert annotation and agreement analysis.

3. METHODOLOGY

Building a gold-standard dataset from earnings call transcripts presents both annotation and data selection challenges. Not all sentences in these documents are useful for sentiment analysis. In addition to financially relevant statements, transcripts also include greetings, moderator instructions, procedural remarks, and other non-informative passages. As a result, manually labeling the full corpus would require substantial expert effort, while simple random sampling could waste annotation time with redundant, easy, or non-relevant examples.

Active learning (AL) is an iterative annotation strategy in which a model is first trained on a small labeled set and then used to select the most informative unlabeled examples for manual annotation. After each annotation cycle, the labeled set is expanded and the model is retrained. This strategy was employed in the relevance filtering and expert sentiment annotation steps of our annotation workflow (Figure 1): first, to identify relevant sentences, and second, to prioritize challenging instances for transformer fine-tuning.

Figure 1 presents an overview of the proposed workflow, composed of four steps: i) data collection and preprocessing; ii) relevance filtering; iii) expert sentiment annotation; and iv) benchmark evaluation.

3.1 Data Collection and Preprocessing

The raw corpus was obtained from the BraFiNER dataset, which contains 118,411 sentences extracted from 384 earnings call transcripts of ten Brazilian banks, published quarterly between 2006 and 2023 [Abilio et al. 2024]. In the context of the present study, we observed duplicated sentences, short fragments, English passages, timestamps, speaker markers, and other noisy patterns. To address these issues, we applied manual revision and automatic rules while preserving financially meaningful content. After this preprocessing step, the corpus contained 100,245 sentences.

3.2 Relevance Filtering

To reduce the effort required to identify relevant and non-relevant sentences in the corpus, we employed clustering to group similar sentences, followed by active learning to train a binary classifier. Clusters labeled as not relevant were filtered out, while relevant clusters were retained for the next step.

In this pipeline, sentence embeddings were generated using the BGE-m3 model[1], which produces 1024-dimensional representations, and clustered using UMAP[2] for dimensionality reduction and HDBSCAN[3] for density-based grouping. Hyperparameters were selected based on the number of clusters, proportion of outliers, and manual inspection of sampled clusters, prioritizing downstream usefulness for relevance filtering.

After clustering the sentence embeddings, an initial seed set of approximately 10% of the clusters was manually labeled as *Relevant* or *Not Relevant* by the research team. Each cluster was represented by the centroid of its embeddings and used to train a Logistic Regression classifier within a pool-based active learning framework. At each cycle, the most uncertain clusters were selected for manual annotation and added to the labeled set, averaging 13 clusters per cycle. The procedure was conducted for 40 cycles.

Label stability across cycles was used to identify clusters with consistent predictions. A randomized sample of stable clusters was manually reviewed using a 95% confidence level and 5% margin of error [Cochran 1977], and validated labels were incorporated into the final training set. The final classifier was then applied to assign relevance labels to all clusters, which were propagated to the corresponding sentences.

The resulting subset of relevant sentences was used in the second step of the workflow, where experts performed sentiment annotation.

3.3 Expert Sentiment Annotation

This step aimed to assign sentiment labels from the perspective of the bank and its shareholders, based on whether the described events were expected to improve or worsen the bank's financial fundamentals rather than its short-term stock price. For example, increased profits and reduced personnel expenses were considered positive, whereas reduced net interest margins were considered negative. Because this task requires domain interpretation, annotation was conducted by four domain experts with academic or professional experience in finance, economics, accounting, and business. To expand an initial labeled dataset and improve efficiency, we adopted a second active learning stage. While many active learning settings rely on a single model, we employed a Query by Committee (QBC) strategy, using two transformer models with different training backgrounds to reduce dependence on a single classifier during data selection. Disagreement between these complementary models was used to identify informative cases for manual review, while diversity constraints helped avoid repeatedly selecting highly similar sentences.

Initially, a subset of 500 sentences was selected based on the sentence clusters obtained in the previous step to ensure thematic diversity. Before annotation, a joint training session was held to align concepts and labeling criteria. Following this preparation, each sentence received one of the following labels: Strong Negative, Moderate Negative, Weak Negative, Neutral, Weak Positive, Moderate Positive, Strong Positive, or Not Relevant. This scheme also allowed experts to reject sentences that remained irrelevant after the previous filtering step.

All experts annotated the initial subset independently, and inter-annotator agreement was measured with Krippendorff's Alpha (α) [Krippendorff 2011] for nominal and ordinal scales (considering sentiment intensity from -3 (*Strong Negative*) to 3 (*Strong Positive*), excluding *Not Relevant*), and with Fleiss' Kappa (κ) [Fleiss 1971]. Using majority voting (at least three concordant labels), we built the initial labeled dataset.

The initial labeled dataset was divided into training and validation sets using an 80/20 split for

[1]https://huggingface.co/BAAI/bge-m3
[2]https://umap-learn.readthedocs.io/
[3]https://hdbscan.readthedocs.io/

model fine-tuning and early stopping. It was then expanded through the active learning stage, in which the committee consisted of two transformer-based models: BERTimbau (base, 110M parameters) [Souza et al. 2020] and DeB3RTa (small, 70M parameters) [Pires et al. 2025], representing general-purpose and finance-adapted language models for Brazilian Portuguese.

At each cycle, both models were fine-tuned for up to 10 epochs, with early stopping (patience of two) based on the macro F1-score. Class-weighted loss functions were adopted to mitigate class imbalance during model fine-tuning. After fine-tuning, both models were used to classify the remaining unlabeled sentences (the pool). For each sentence, disagreement was measured using the Jensen–Shannon divergence [Lin 1991] between the predictive distributions of the two models. Candidate sentences were ranked by disagreement and then filtered to increase diversity across the sentence clusters obtained in the previous step. At each cycle, approximately 20 sentences were selected and submitted to expert annotation.

Newly labeled sentences with consensus or majority agreement were incrementally added to the training and validation sets while preserving the original 80/20 split, and the models were fine-tuned from scratch in the next cycle. The annotation cycles were conducted for a predefined period of 12 weeks, while convergence indicators such as F1-score stability on newly annotated sentences and predictive uncertainty reduction across consecutive cycles were monitored throughout the process. This process resulted in an expert-annotated dataset, which was evaluated in the next step.

3.4 Benchmark Evaluation

To assess the usefulness and consistency of the proposed dataset, we considered three transformer-based models: BERTimbau, DeB3RTa, and BERTuguês (110M parameters)[Zago and Pedotti 2024]. The first two models were also employed in the second active learning stage, while BERTuguês was included as an external baseline independent from the annotation pipeline. This comparison allows us to assess whether the resulting dataset remains useful for models not involved in data selection.

Given the limited dataset size, models were evaluated using stratified 5-fold cross-validation, and we report macro-averaged F1 score. Macro F1 was calculated across the eight predefined classes, assigning zero to the F1 of the absent *Strong Negative* class. Transformer models were fine-tuned with class-weighted loss functions to mitigate class imbalance across sentiment categories.

Experiments were conducted under three settings: i) the original eight-class scheme; ii) a three-class setting obtained by grouping sentiment intensities into Positive, Neutral, and Negative categories, while merging *Neutral* and *Not Relevant* labels; and iii) a sentiment-only three-class setting obtained by excluding *Not Relevant* instances. All models were trained and evaluated using the same configuration across settings to ensure fair comparison.

4. RESULTS AND DISCUSSION

This section presents and discusses the results obtained in each step of the proposed workflow (Figure 1), covering relevance filtering, annotation quality, and benchmark evaluation.

4.1 Relevance filtering

After preprocessing, the resulting corpus of 100,245 sentences was used in the relevance filtering step of the proposed workflow (Figure 1).

Using this corpus, we conducted exploratory experiments to define suitable parameters for UMAP and HDBSCAN. Applying UMAP ($n\_neighbors = 5$, $n\_components = 512$) improved cluster coherence compared to using the original embeddings. For HDBSCAN, smaller minimum cluster sizes

increased coverage by reducing outliers while preserving cluster specificity. We therefore adopted a minimum cluster size of 5, resulting in 2,792 clusters, with outliers assigned to the noise label (-1).

The first active learning stage was conducted over 40 cycles, during which 801 clusters were manually annotated. In addition, 1,904 clusters were identified as stable in the final cycles and incorporated into the labeled set after manual validation of a randomized sample of 320 stable clusters. The final classifier was trained on 2,705 labeled clusters and achieved 95% macro F1 and 96% accuracy on a stratified 70/30 train–test split.

Overall, the proposed procedure substantially reduced annotation effort. Instead of manually labeling all 2,792 clusters, the workflow required inspection of 1,121 clusters, corresponding to approximately 40% of the total, while maintaining high classification performance. The final clustering structure contained 1,724 relevant clusters and 1,068 not relevant clusters, corresponding to 84,278 relevant sentences and 15,967 not relevant sentences.

During the active learning cycles, performance metrics showed non-monotonic behavior, with temporary drops when more diverse clusters were introduced. This suggests that uncertainty-based selection may favor local refinement, while diversity contributes to broader generalization. Despite these fluctuations, the overall process provided an efficient relevance filter for the sentiment annotation step.

4.2 Expert Sentiment Annotation

The annotation process, including the initial subset and active learning stage, spanned approximately six months. Annotating the initial 500 sentences took nearly three months, reflecting the task's complexity and the limited availability of domain experts, who were not fully dedicated to annotation.

At the end of this initial phase, 331 sentences reached majority agreement. The label distribution in this initial dataset was imbalanced: i) Strong Negative: 0; ii) Moderate Negative: 4; iii) Weak Negative: 36; iv) Neutral: 76; v) Weak Positive: 126; vi) Moderate Positive: 30; vii) Strong Positive: 5; and viii) Not Relevant: 54. We note that no instances were labeled as *Strong Negative*, which may reflect the communication style in earnings calls, where extreme negative statements are uncommon due to regulatory and reputational considerations.

Regarding inter-annotator agreement, we obtained Krippendorff's Alpha (α) of 0.66 (nominal) and 0.79 (ordinal). According to Krippendorff [2011], values above 0.80 indicate reliable agreement, while values between 0.67 and 0.80 are generally considered acceptable for exploratory analyses. In this context, the ordinal α suggests acceptable agreement when accounting for the ordered nature of sentiment intensity, while the nominal α reflects the difficulty of achieving exact label agreement in a fine-grained annotation scheme. Pairwise agreement analysis showed nominal α values ranging from 0.50 to 0.83 and ordinal α values ranging from 0.69 to 0.90. For comparison, Fleiss' Kappa (κ) reached 0.66, indicating substantial agreement and reinforcing the overall consistency of the annotation despite the fine-grained label scheme.

Compared to prior work [Santos et al. 2023; Kakimoto et al. 2024], which reports higher agreement in simpler settings, the lower agreement observed here is likely due to increased annotation complexity, including fine-grained labels and sentence-level interpretation in earnings call transcripts, which often involve subtle and context-dependent expressions of sentiment.

During the active learning stage, each annotation cycle required approximately three days of expert effort, while model update and sample selection took around 30 minutes. A total of 22 cycles were conducted over 12 weeks. Macro F1-score and predictive uncertainty fluctuated across cycles, reflecting the introduction of diverse and challenging samples. Some sentences were selected repeatedly due to persistent annotator disagreement, suggesting that the QBC strategy effectively identified ambiguous cases. For instance, the sentence "Seguros e previdência praticamente estáveis, 15% ante 16%." ("Insurance and pension services remained practically stable, 15% compared to 16%.") was labeled as

Neutral by two annotators and as *Weak Negative* by two others, illustrating how subtle quantitative changes combined with mitigating language can lead to divergent interpretations.

In addition, a subset of sentences remained consistently labeled across all active learning cycles, suggesting the presence of high-confidence examples that are less sensitive to model updates. At the same time, the proportion of label changes decreased over successive cycles, indicating a progressive stabilization of predictions as more informative samples were incorporated into the training set. These observations suggest that the disagreement-driven selection mechanism initially focuses on uncertain regions and gradually transitions toward a more stable decision boundary.

The dataset grew from 330 to 719 sentences, with an average increment of 17.7 ($\pm$ 4.1) sentences per cycle. The final dataset had the following label distribution: i) Strong Negative: 0; ii) Moderate Negative: 5; iii) Weak Negative: 65; iv) Neutral: 162; v) Weak Positive: 270; vi) Moderate Positive: 52; vii) Strong Positive: 8; and viii) Not Relevant: 157. The final α values were 0.60 (nominal, ranging from 0.40 to 0.80) and 0.77 (ordinal, ranging from 0.68 to 0.86). The final κ was 0.60, indicating moderate agreement. The gold-standard dataset is publicly available [Abilio et al. 2026].

Comparing the agreement between the initial and final datasets, we observe that agreement levels remained within a similar range, despite a slight decrease in nominal α and κ. This behavior is consistent with the introduction of more challenging and ambiguous samples during the active learning process, which can increase annotation difficulty while preserving overall consistency.

4.3 Benchmarking

The eight-class setting represents the most challenging and fine-grained version of the task. Performance remained moderate across models, with BERTuguês and BERTimbau achieving comparable F1-macro scores (0.43 ± 0.03 and 0.43 ± 0.01, respectively), while DeB3RTa showed substantially lower performance (0.28 ± 0.04). These results reflect the difficulty of distinguishing subtle differences in sentence-level sentiment intensity. The similar performance between BERTuguês and BERTimbau suggests that the dataset does not exhibit strong bias toward the models used during construction.

In the three-class settings, performance increased across all models, as expected with reduced class granularity. When *Neutral* and *Not Relevant* were merged, BERTimbau achieved 0.71 ± 0.04 and BERTuguês 0.68 ± 0.05. In the sentiment-only setting, performance decreased slightly but remained higher than in the eight-class setup, with BERTimbau reaching 0.68 ± 0.06 and BERTuguês 0.66 ± 0.02. Across both configurations, BERTimbau achieved the highest scores, while BERTuguês showed more stable results across folds. DeB3RTa remained stable at approximately 0.52 ± 0.04 in both settings.

5. CONCLUSIONS

This work introduced a new expert-annotated dataset for financial sentiment analysis in Brazilian Portuguese, based on earnings call transcripts and built through a two-stage active learning pipeline. The results demonstrate its applicability to fine- and coarse-grained sentiment classification, with comparable performance across models used during data selection and an external baseline, providing preliminary evidence that the dataset can be used beyond the models involved in its construction.

Despite these contributions, some limitations should be acknowledged. The dataset is restricted to the banking sector and reflects a relatively small number of annotated instances, which may limit generalization to other domains. In addition, the use of a fine-grained annotation scheme increases task complexity and leads to moderate inter-annotator agreement, highlighting the inherent difficulty of sentence-level financial sentiment analysis.

Future work includes extending the dataset to additional sectors, exploring alternative annotation strategies to improve agreement, and investigating active learning dynamics, such as label stability and sample difficulty, to better understand their impact on dataset construction and model performance.

REFERENCES

Abilio, R., Coelho, G. P., and da Silva, A. E. A. Evaluating Named Entity Recognition: A comparative analysis of mono- and multilingual transformer models on a novel Brazilian corporate earnings call transcripts dataset. *Applied Soft Computing* vol. 166, pp. 112158, 2024.

Abilio, R., Coelho, G. P., and da Silva, A. E. A. Replication Data for: A Brazilian Dataset for Financial Sentiment Analysis: An Active Learning-Based Annotation Framework. https://doi.org/10.25824/redu/HUQPYK, 2026. Repositório de Dados de Pesquisa da Unicamp.

Alawaji, R. and Aloraini, A. Sentiment Analysis of Digital Banking Reviews Using Machine Learning and Large Language Models. *Electronics* 14 (11): 2125, 2025.

Ashrafi Asli, S. A., Sabeti, B., Majdabadi, Z., Golazizian, P., Fahmi, R., and Momenzadeh, O. Optimizing Annotation Effort Using Active Learning Strategies: A Sentiment Analysis Case Study in Persian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France, pp. 2855–2861, 2020.

Cochran, W. G. *Sampling Techniques*. Wiley, New York, 1977.

da Silva, F., Roman, N., and Carvalho, A. Stock market tweets annotated with emotions. *Corpora* 15 (3): 343–354, 2020.

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76 (5): 378–382, 1971.

Januário, B. A., Carosia, A. E. d. O., da Silva, A. E. A., and Coelho, G. P. Sentiment Analysis Applied to News from the Brazilian Stock Market. *IEEE Latin America Transactions* 20 (3): 512–518, 2022.

Kakimoto, G., Haddadi, S., Araújo, P., Silva, F., dos Reis, J., and Reis, M. B2T: A dataset of tweets in portuguese language about brazilian banks. In *Proceedings of the VI Dataset Showcase Workshop (DSW)*. Florianópolis, SC, Brazil, pp. 1–11, 2024.

Kaseb, A. and Farouk, M. Active learning for Arabic sentiment analysis. *Alexandria Engineering Journal* vol. 77, pp. 177–187, 2023.

Krippendorff, K. Computing Krippendorff's alpha-reliability. https://repository.upenn.edu/handle/20.500.14332/2089, 2011. Accessed in: 2026-05-04.

Lan, Y., Wu, Y., Xu, W., Feng, W., and Zhang, Y. Chinese fine-grained financial sentiment analysis with large language models. *Neural Computing and Applications* 37 (30): 24883–24892, 2025.

Li, D., Wang, Z., Chen, Y., Jiang, R., Ding, W., and Okumura, M. A Survey on Deep Active Learning: Recent Advances and New Frontiers. *IEEE Transactions on Neural Networks and Learning Systems* 36 (4): 5879–5899, 2025.

Liebenlito, M., Inayah, N., Choerunnisa, E., Sutanto, T. E., and Inna, S. Active learning on Indonesian Twitter sentiment analysis using uncertainty sampling. *Journal of Applied Data Sciences* 5 (1): 114–121, 2024.

Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37 (1): 145–151, 1991.

Malo, P., Sinha, A., Takala, P., Korhonen, P., and Wallenius, J. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. *Journal of the American Society for Information Science and Technology* 65 (4): 782–796, 2014.

Mostafavi, S., Yahyavi, Y., and Ravanmehr, R. Systematic literature review on sentiment analysis using transformers. *International Journal of Data Science and Analytics* 22 (1): 49, 2026.

Pires, H., Paucar, L., and Carvalho, J. P. DeB3RTa: A Transformer-Based Model for the Portuguese Financial Domain. *Big Data and Cognitive Computing* 9 (3), 2025.

Pustejovsky, J. and Stubbs, A. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. O'Reilly, Sebastopol, CA, 2013.

Raicu, I. Financial Banking Dataset for Supervised Machine Learning Classification. *Informatica Economica* 23 (1): 37–49, 2019.

Romero, V., Assis, G., Carvalho, J., Mann, P., and Paes, A. Opportunities vs. Risks: Exploring Automatic Annotation of Financial Polarity Biases via Large Language Models. In *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*. Fortaleza, CE, Brazil, pp. 1–8, 2025.

Santos, L., Bianchi, R., and Costa, A. FinBERT-PT-BR: Análise de Sentimentos de Textos em Português do Mercado Financeiro. In *Anais do II Brazilian Workshop on Artificial Intelligence in Finance*. João Pessoa, PB, Brazil, pp. 144–155, 2023.

Souza, F., Nogueira, R., and Lotufo, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*. Number 12319 in Lecture Notes in Computer Science. Cham, pp. 403–417, 2020.

Zago, R. and Pedotti, L. BERTugues: A Novel BERT Transformer Model Pre-trained for Brazilian Portuguese. *Semina: Ciências Exatas e Tecnológicas* vol. 45, pp. e50630–e50630, 2024.