Sector-Specific Financial Sentiment Lexicons for Portuguese: Evidence from Brazilian Stock News

Antônio Artur de Souza[1], **Mateus Binda**[1] **Adriana S. Pagano**[2]

[1] CEPEAD – Centro de Pós-Graduação e Pesquisa em Administração –
Faculdade de Ciências Econômicas – Universidade Federal de Minas Gerais (UFMG)
[2]Faculdade de Letras – Universidade Federal de Minas Gerais (UFMG)

artur@face.ufmg.br, tetatrader22334@gmail.com, apagano@ufmg.br

Abstract. *This paper investigates whether sector-specific financial sentiment lexicons improve the analysis of Brazilian stock-market news in Portuguese. We use 19,460 news items about 38 firms in the financial, electric, and manufacturing sectors, published between 2020 and 2025. Starting from OpLexicon, we build sector-specific lexicons through unsupervised expansion based on PMI and conditional word-polarity association. The lexicons are evaluated by lexical heterogeneity and through LSTM price forecasting. Results show sectoral variation and moderate classification performance, but limited predictive gains. Sentiment improves forecasting only in manufacturing, while other sectors do not benefit from its inclusion. The findings indicate that sector-specific lexicons are useful resources, although their predictive value are context dependent.*

Keywords: Sentiment analysis; financial lexicons; Portuguese NLP; domain adaptation; stock-market news.

1. Introduction

The increasing digitalization of financial markets has substantially expanded the availability of unstructured textual data, including news articles, corporate disclosures, reports, and social media content. In this setting, sentiment analysis has become a relevant strategy for converting qualitative information into quantitative indicators that may support the analysis of market behavior. Prior work has shown that textual sentiment can be associated with prices, returns, and volatility, but the evidence remains mixed and strongly dependent on language, source, time horizon, and modeling choices [Loughran and McDonald 2011, Ranco et al. 2015, Deveikyte et al. 2022, Janková 2023].

For Portuguese, the gap is even more evident. General-purpose lexicons are available, but specialized financial resources remain limited, especially for Brazilian data and sector-specific language use. This limitation matters because the meaning of sentiment-bearing words may change across sectors and market contexts. A term that is positive in one segment may be neutral or even negative in another, which reduces the adequacy of generic lexical resources for financial NLP in Portuguese [Kos et al. 2019, Peres et al. 2019, de Melo 2022].

This paper addresses that gap by investigating whether sector-specific financial sentiment lexicons in Portuguese can better represent Brazilian stock-market news and whether the resulting sentiment signal provides useful downstream information. We focus on three sectors traded on B3[1]: financial, electric, and manufacturing. From a methodological perspective, we start from a seed lexicon and expand it automatically using an unsupervised procedure based on PMI and conditional polarity association. From an evaluation perspective, we assess the lexicons both intrinsically, through classification-oriented evidence, and extrinsically, through a downstream stock-forecasting task with LSTM models.

Our main contributions are threefold. First, we compile and analyze a Portuguese corpus of 19,460 specialized financial news items related to 38 Brazilian listed companies. Second, we propose sector-specific lexicon expansion for Portuguese financial text and show that the resulting vocabularies display measurable heterogeneity across sectors. Third, we show that the usefulness of the derived sentiment signal is conditional: sentiment is associated with market variables only weakly, and downstream predictive gains are not uniform across sectors. This shifts the discussion away from the binary question of whether sentiment is useful or not, toward the more precise question of when, where, and under what conditions sentiment provides useful information.

2. Related Work

Lexicon-based sentiment analysis remains an important approach when interpretability is desirable and large labeled datasets are not available. However, domain mismatch is a central limitation. Loughran and McDonald [Loughran and McDonald 2011] showed that general-purpose dictionaries often misclassify financial language because many terms common in business discourse do not carry the same polarity they have in everyday language. This observation motivated the construction of domain-specific lexical resources and automatic lexicon-expansion methods for finance and other specialized domains.

In Portuguese, OpLexicon is one of the most widely used lexical resources for opinion mining [Souza et al. 2011]. Even so, it was not originally designed for financial language. Recent work has therefore explored automatic strategies for domain adaptation and lexicon construction in Portuguese, including unsupervised expansion procedures based on distributional and probabilistic signals [de Melo 2021, de Melo 2022, de Sousa and Fernandes 2023]. For the Brazilian financial domain, previous studies have already indicated that lexical methods can extract useful information from news and market-related texts, although performance remains moderate and sensitive to preprocessing and representation choices [Peres et al. 2019, Januário et al. 2022].

The second relevant line of work concerns the relationship between textual sentiment and market variables. Some studies report statistically significant associations between sentiment and returns, prices, or volatility, as well as gains in predictive tasks when text-derived features are incorporated into machine-learning models [Huang and Liu 2020, Hsu et al. 2021, Gao et al. 2022, Zhao et al. 2025]. Others, however, emphasize that these gains are unstable and context-dependent [Ranco et al. 2015, Janková 2023]. This suggests that sentiment should not be treated as a universally informative signal.

Our work lies at the intersection of these two lines of research. We contribute to Portuguese NLP by building sector-specific lexicons for Brazilian financial news and evaluating them not only as lexical resources but also through a downstream forecasting task.

[1]A B3 é a bolsa de valores brasileira e seu nome é uma abreviação para Brasil, Bolsa, Balcão.

3. Corpus and Method

3.1. Corpus construction

The corpus combines financial news retrieved via the GNews API and market data obtained through BRAPI. We collected news about firms listed on B3 and belonging to three sectors: electric, financial, and manufacturing. The sectoral classification follows the institutional structure used by B3 [B3 2025]. After consolidation, auditing, source filtering, and duplicate removal, the final dataset contained 19,460 news items published between November 11, 2020 and November 11, 2025.

The final corpus is distributed across 38 firms and three sectors: 4,286 news items from the electric sector, 11,053 from the financial sector, and 5,112 from manufacturing. The total is not the sum of the sectoral subsets because duplicate cross-sector items were removed from the general corpus. For the downstream forecasting analysis, we selected one representative firm per sector based on market relevance and liquidity: Eletrobras (electric), B3 (financial), and Weg (manufacturing). The retained news sources were specialized financial outlets, namely InfoMoney, Economia UOL, Investing, Valor Investe, and Valor Econômico.

3.2. Seed lexicon and PMI-based title labeling

The starting point for lexicon construction is OpLexicon V3.0 [Souza et al. 2011], a Portuguese opinion lexicon containing lexical items annotated for semantic polarity. Although not financial by design, it offers a suitable seed resource for domain adaptation.

Following Shapiro, Sudhof, and Wilson [Shapiro et al. 2022], we compute PMI over news titles. Titles were chosen because they concentrate the main informational framing of each article and provide a comparable textual unit for the initial labeling stage. Based on the distribution of PMI scores, we assigned titles to three groups using interquartile thresholds: titles at or below the first quartile were labeled as negative, titles at or above the third quartile were labeled as positive, and titles in between were labeled as neutral. This procedure yields balanced positive and negative subsets for subsequent unsupervised expansion.

3.3. Unsupervised sector-specific lexicon expansion

After title labeling, we expanded the seed lexicon using an unsupervised procedure inspired by de Melo [de Melo 2022]. The core idea is to estimate the association between each word and positive or negative classes based on its conditional frequency in previously labeled titles. For each candidate word w_i, we compute positive and negative conditional probabilities and derive a polarity score from their difference. Words with sufficient evidence are then incorporated into the corresponding sector-specific lexicon.

This procedure yields one expanded lexicon for each sector. In practice, it preserves the initial semantic basis provided by the seed lexicon while incorporating vocabulary that is characteristic of sector-specific Brazilian financial news.

3.4. Sentiment scoring

Once the sector-specific lexicons were built, we computed sentiment scores for the news corpus. The sentiment-scoring procedure includes text preprocessing, lexical matching against the appropriate sector lexicon, and score computation from the balance between positive and negative words, considering negation and modulation cues, following Zhao et al. [Zhao et al. 2025].

The output of this stage is a daily sentiment index for each sector. We then compare these indices with two market variables: same-day closing prices and daily returns. We report Pearson correlations for the full sample and analyze subsamples corresponding to market upturns and downturns, defined as five consecutive trading days with closing prices moving in the same direction.

3.5. Downstream forecasting with LSTM

To test whether sentiment adds downstream predictive information, we trained LSTM models with and without the daily sentiment index. The target is next-day closing price prediction. The architecture contains two stacked LSTM layers with 64 and 32 units, followed by a dense output layer. We used mean squared error as the loss function, the Adam optimizer, and rolling windows of five days as input sequences.

Model performance was assessed through MAE, MSE, and RMSE. The central question is whether adding sentiment to price-based inputs improves forecasting accuracy.

4. Results and Discussion

4.1. PMI distribution and lexical heterogeneity

The PMI distributions suggest a mild positive bias in all sectors, but with different shapes and dispersion. The electric sector shows the highest mean PMI (0.1521), followed by the financial sector (0.1030) and manufacturing (0.0943). This indicates that the informational tone of the news is not distributed uniformly across domains. The title-labeling procedure also yielded balanced positive and negative sets within sectors, supporting the subsequent expansion stage.

Table 1. Descriptive statistics of PMI scores by sector

Statistic	Electric sector	Financial sector	Manufacturing sector
Mean	0.1521	0.1030	0.0943
Standard deviation	0.1798	0.1617	0.1275
Minimum	-0.5098	-0.5819	-0.3811
First quartile	0.0336	0.0010	0.0084
Median	0.1325	0.0963	0.0885
Third quartile	0.2488	0.1890	0.1791
Maximum	0.7792	0.6878	0.5937
Skewness	0.5021	0.3410	0.0737
Kurtosis	0.5466	0.8018	0.1537

The expanded lexicons confirm sectoral heterogeneity. In the electric sector, expansion processed 4,170 labeled titles and incorporated 782 terms, 303 of them positive and 479 negative. In the financial sector, the expanded lexicon reached 1,531 terms, with 606 positive and 925 negative entries. In manufacturing, the expansion produced 849 terms, 356 positive and 493 negative. These differences indicate that sector-specific vocabulary is not merely a matter of word frequency; it also reflects sector-dependent polarity structure.

Intrinsic evaluation results are moderate but informative. For the electric sector, the F_1-score is 0.57 for the positive class and 0.52 for the negative class. In the financial sector, the corresponding values are 0.53 and 0.50. In manufacturing, they are 0.53 and 0.52. The general corpus, in turn, shows values close to the sector-specific ones, which suggests that sector specialization is not primarily justified by a large intrinsic performance jump. Instead, its main relevance lies in capturing domain-specific lexical organization that may matter for interpretation and downstream use.

4.2. Sentiment and market variables

The daily sentiment indices display different statistical profiles across sectors. In the electric sector, the index has mean 3.43 and standard deviation 7.26. In the financial sector, the mean is 8.5 and the standard deviation is 13.67, indicating stronger positive skew and occasional peaks of optimism. In manufacturing, the mean is 3.92 and the standard deviation is 9.34.

The correlations with market variables are statistically detectable in some cases, but they remain weak overall. In the electric sector, sentiment is positively correlated with same-day closing prices (0.0744, $p = 0.0172$), but not with daily returns. In the financial sector, the same-day correlation with closing prices is 0.1123 ($p < 0.01$), while the correlation with daily returns is not significant. Manufacturing exhibits the opposite pattern: sentiment is not significantly associated with same-day closing prices (-0.0367, $p = 0.2311$), but it is positively correlated with daily returns (0.0912, $p < 0.01$).

When we segment the sample by market regime, the electric sector shows a significant positive correlation only during downturns (0.1116, $p < 0.01$). The financial sector shows significant positive coefficients in both upturns (0.1211, $p < 0.01$) and downturns (0.1037, $p < 0.01$). Manufacturing does not show significant regime-specific effects. Taken together, these results indicate that sentiment carries some information, but its relation to market behavior is weak and sector-dependent.

Table 2. Sentiment associations with prices, returns, and market regimes

Parameter	Electric		Financial		Manufacturing	
	Correlation	p-value	Correlation	p-value	Correlation	p-value
Closing price	0.0744***	0.0172	0.1123***	< 0.01	-0.0367	0.2311
Daily returns	0.0488	0.1184	0.0253	0.3800	0.0912***	< 0.01
Upturn phase	0.4135	0.3641	0.1211***	< 0.01	-0.0611	0.1669
Downturn phase	0.1116***	< 0.01	0.1037***	< 0.01	-0.0045	0.9147

Notes: *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

4.3. Downstream evaluation with LSTM

The downstream forecasting task yields the most important practical result of the paper: sentiment does not improve prediction uniformly. In the electric sector, adding sentiment worsens performance substantially. For Eletrobras, MAE increases from 1.25 to 2.72, MSE from 2.98 to 13.50, and RMSE from 1.72 to 3.67 when sentiment is included. In the financial sector, represented by B3, sentiment also deteriorates the results, though only marginally: MAE changes from 1.5324 to 1.5672, MSE from 4.7055 to 4.9743, and RMSE from 2.1620 to 2.2303.

Manufacturing is the only sector in which sentiment helps. For Weg, the model with sentiment achieves MAE of 2.5473, MSE of 12.6192, and RMSE of 3.5524, compared with 2.7032, 14.0482, and 3.7481 without sentiment. This corresponds to modest improvements of approximately 5.8% in MAE and 4.2% in RMSE.

Table 3. LSTM forecasting results with and without sentiment

Metric	Electric		Financial		Manufacturing	
	With sentiment	Without sentiment	With sentiment	Without sentiment	With sentiment	Without sentiment
MAE (%)	2.72	1.25	1.5672	1.5324	2.5473	2.7032
MSE (%)	13.50	2.98	4.9743	4.7055	12.6192	14.0482
RMSE (%)	3.67	1.72	2.2303	2.1620	3.5524	3.7481
Test sample (days)	241	241	241	241	213	213

The downstream evidence therefore supports restrained interpretation. Sector-specific sentiment lexicons are useful as Portuguese financial NLP resources, but the predictive value of the resulting sentiment index is conditional. In two of the three sectors, adding sentiment introduces noise rather than useful information. Only in manufacturing does the signal appear to complement price history in a consistent way.

5. Conclusion

We presented a study on sector-specific financial sentiment lexicons for Portuguese based on Brazilian stock-market news. Starting from a general-purpose seed lexicon, we built sector-specific lexical resources through unsupervised expansion and evaluated them both intrinsically and extrinsically. The results show that Brazilian financial news exhibits measurable lexical heterogeneity across sectors and that sector-specific lexicons capture part of this structure. However, the usefulness of the resulting sentiment signal is not uniform. Although sentiment is weakly associated with some market variables, downstream forecasting gains appear only in the manufacturing sector.

Overall, the paper contributes to Portuguese NLP by providing evidence that domain and sector adaptation matter for financial sentiment analysis. At the same time, it shows that improved lexical specialization does not automatically translate into robust forecasting benefits. Future work should expand the number of sectors and firms, incorporate manually annotated evaluation data, and compare lexicon-based methods with supervised and transformer-based approaches specialized for Portuguese financial language.

Limitations

This study has some limitations. First, the analysis is restricted to three sectors of the Brazilian stock market and to a specific set of financial news outlets, which may introduce coverage bias. Second, the downstream forecasting analysis relies on a single representative company per sector, which limits firm-level generalization. Third, the lexical approach inherits known limitations of lexicon-based sentiment analysis, including difficulty in handling irony, syntactic scope, semantic shifts, and more subtle contextual cues. Finally, although the corpus spans five years, the time coverage is bounded by the availability of the news API used for data collection.

Acknowledgments

The authors express their gratitude to the anonymous reviewers for their constructive feedback. Adriana Pagano's research is funded by FAPEMIG and the National Council for Scientific and Technological Development (CNPq, grants 404722/2024-5; 313103/2021-6). This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology) and is conducted under the auspices of the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by CNPq (grant 408490/2024-1).

Ethics Statement

This study uses publicly available news content and publicly available financial market data. No personal, private, or sensitive data were collected. The work is intended as research on language technology and financial text analysis. It should not be interpreted as financial advice or as a recommendation for investment decisions. Given the weak and heterogeneous predictive effects observed in the experiments, the paper explicitly cautions against overinterpreting sentiment signals in applied trading contexts.

Generative AI Usage Disclosure

Generative AI tools were used as auxiliary support in programming, debugging, and language revision tasks. All methodological decisions, data-processing choices, result validation, and final scientific claims were defined, checked, and approved by the authors.

References

B3 (2025). Critério de classificação.

de Melo, T. (2021). Sentiprodbr: Building domain-specific sentiment lexicons for the portuguese language. In *Anais do Simpósio Brasileiro de Banco de Dados*, pages 349–354.

de Melo, T. (2022). Sentilexbr: An automatic methodology of building sentiment lexicons for the portuguese language. *Journal of Information and Data Management*, 13(3):439–449.

de Sousa, T. M. and Fernandes, D. S. A. (2023). Expansão automática de léxico para análise de sentimentos de twitter no domínio do mercado financeiro brasileiro. In *Anais da Escola Regional de Informática de Goiás*.

Deveikyte, J., Geman, H., Piccari, C., and Provetti, A. (2022). A sentiment analysis approach to the prediction of market volatility. *Frontiers in Artificial Intelligence*, 5:1–10.

Gao, Y., Zhao, C., Sun, B., and Zhao, W. (2022). Effects of investor sentiment on stock volatility: new evidences from multi-source data in china's green stock markets. *Financial Innovation*, 8(1):77.

Hsu, Y.-J., Lu, Y.-C., and Yang, J. J. (2021). News sentiment and stock market volatility. *Review of Quantitative Finance and Accounting*, 57(3):1093–1122.

Huang, J.-Y. and Liu, J.-H. (2020). Using social media mining technology to improve stock price forecast accuracy. *Journal of Forecasting*, 39(1):104–116.

Janková, Z. (2023). Critical review of text mining and sentiment analysis for stock market prediction. *Journal of Business Economics and Management*, 24(1):177–198.

Januário, B. A. et al. (2022). Sentiment analysis applied to news from the brazilian stock market. *IEEE Latin America Transactions*, 20(3):512–518.

Kos, S. R., Scarpin, J. E., and Pinto, J. S. d. P. (2019). Proposta de um dicionário de termos de emoção em textos financeiros. *Revista de Ciências da Administração*, 21(55):147–164.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66(1):35–65.

Peres, V., Vieira, R., and Bordini, R. (2019). Análises de sentimentos: abordagem lexical de classificação de opinião no contexto mercado financeiro brasileiro. *Anais do Workshop of Artificial Intelligence Applied to Finance*.

Ranco, G. et al. (2015). The effects of twitter sentiment on stock price returns. *PLoS ONE*, 10(9):e0138441.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics*, 228(2):221–243.

Souza, M. et al. (2011). Construction of a portuguese opinion lexicon from multiple resources. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology*.

Zhao, C., Kang, L., Xi, X., Du, S., and Li, J. (2025). Investor sentiment and stock market volatility: exploring the relationship using sentiment analysis of stock bar comments. *Finance Research Open*, 1:100016.