MINING SENTIMENT IN AGRIBUSINESS NEWS: AN ANALYSIS OF THE CORRELATION WITH PRICES IN THE POULTRY CLUSTER OF BASTOS-SP

MINERAÇÃO DE SENTIMENTO EM NOTÍCIAS DO AGRONEGÓCIO: UMA ANÁLISE DA CORRELAÇÃO COM PREÇOS NO CLUSTER AVÍCOLA DE BASTOS-SP

ANÁLISIS DEL SENTIMIENTO DEL MERCADO EN LAS NOTICIAS AGROINDUSTRIALES: UN ANÁLISIS DE LA CORRELACIÓN CON LOS PRECIOS EN EL CLÚSTER AVÍCOLA DE BASTOS-SP

https://doi.org/10.56238/sevened2026.019-001

Miguel Guimarães Morassuti[1], Mario Mollo Neto[2]

ABSTRACT

This study investigates the correlation between the sentiment of specialized media and price formation in the poultry cluster of Bastos-SP, a hub responsible for approximately 11% of Brazilian egg production. The central objective of the work is to validate whether the information flow acts as a leading indicator or whether the local market operates decoupled from digital expectations due to physical fundamentals. The applied methodology, of an exploratory and quantitative nature, employs Natural Language Processing (NLP) techniques through the BERTimbau Deep Learning model to analyze a corpus of headlines collected between 2023 and 2026. The data were subjected to semantic filtering and weekly resampling using the Pandas library's ".resample('W')" method in Python. The results demonstrate that source curation and the removal of technical noise tripled the observed Pearson correlation (from r = 0.04 to r = 0.12), indicating that egg prices are marginally sensitive to macroeconomic factors rather than to zootechnical information. The weak, but positive, correlation (r = 0.12) observed after the inclusion of macroeconomic sources suggests that local producers are marginally more sensitive to structural trends (grain costs, external environment) than to zootechnical technical factors, which proved statistically irrelevant for short-term price forecasting. Therefore, it is concluded that the weak magnitude of the correlation confirms the dominance of the spot market and the biological nature of the asset, positioning media sentiment predominantly as a coincident indicator of market conditions.

Keywords: Sentiment Mining. BERTimbau. Agribusiness. Price Formation. Bastos-SP.

[1] Master's Student in Agribusiness and Development. Universidade Estadual Paulista. Faculdade de Ciências e Engenharia. E-mail: miguel.morassuti@unesp.br Orcid: https://orcid.org/0009-0004-3728-135X Lattes: https://lattes.cnpq.br/8974205660452073

[2] Dr. in Agricultural Engineering. Universidade Estadual de Campinas (UNICAMP). Universidade Estadual Paulista. Faculdade de Ciências e Engenharia. E-mail: mario.mollo@unesp.br Orcid: https://orcid.org/0000-0002-8341-4190 Lattes: http://lattes.cnpq.br/6037463340047597

RESUMO
Este estudo investiga a correlação entre o sentimento da mídia especializada e a formação de preços no *cluster* avícola de Bastos-SP, polo responsável por cerca de 11% da produção brasileira de ovos. O objetivo central do trabalho é validar se o fluxo informacional atua como um indicador antecedente ou se o mercado local opera descolado das expectativas digitais devido a fundamentos físicos. A metodologia aplicada, de caráter exploratório e quantitativo, emprega técnicas de Processamento de Linguagem Natural (PLN) por meio do modelo *Deep Learning* BERTimbau para analisar um *corpus* de manchetes coletadas entre 2023 e 2026. Os dados foram submetidos à filtragem semântica e reamostragem semanal por meio do método ".resample('W')" da biblioteca *Pandas* com programação em *Python*. Os resultados demonstram que a curadoria de fontes e a remoção de ruídos técnicos triplicaram a correlação de Pearson observada (de r = 0,04 para r = 0,12), evidenciando que o preço do ovo é marginalmente sensível a pautas macroeconômicas em detrimento de informações zootécnicas. A correlação fraca, porém, positiva (r=0.12), observada após a inclusão de fontes macroeconômicas, sugere que o produtor local é marginalmente mais sensível a tendências estruturais (custos de grãos, cenário externo) do que a pautas técnicas zootécnicas, que se mostraram estatisticamente irrelevantes para a previsão de preços de curto prazo. Conclui-se, portanto, que a magnitude fraca da correlação ratifica a dominância do mercado à vista (*spot market*) e a natureza biológica do ativo, posicionando o sentimento midiático predominantemente como um indicador coincidente das condições de mercado.

Palavras-chave: Mineração de Sentimento. BERTimbau. Agronegócio. Formação de Preços. Bastos-SP.

RESUMEN
Este estudio investiga la correlación entre el sentimiento de los medios especializados y la formación de precios en el clúster avícola de Bastos-SP, un centro que representa aproximadamente el 11% de la producción brasileña de huevos. El objetivo principal del trabajo es validar si el flujo informativo actúa como un indicador adelantado o si el mercado local opera desvinculado de las expectativas digitales debido a factores físicos fundamentales. La metodología aplicada, de carácter exploratorio y cuantitativo, emplea técnicas de procesamiento del lenguaje natural (PLN) mediante el modelo de aprendizaje profundo BERTimbau para analizar un corpus de titulares recopilados entre 2023 y 2026. Los datos fueron sometidos a filtrado semántico y remuestreo semanal utilizando el método ".resample('W')" de la biblioteca Pandas con programación en Python. Los resultados demuestran que la curación de fuentes y la eliminación del ruido técnico triplicaron la correlación de Pearson observada (de r = 0,04 a r = 0,12), lo que indica que los precios de los huevos son marginalmente sensibles a factores macroeconómicos más que a información zootécnica. La débil pero positiva correlación (r = 0,12) observada tras la inclusión de fuentes macroeconómicas sugiere que los productores locales son ligeramente más sensibles a las tendencias estructurales (costes de los cereales, entorno externo) que a los factores técnicos zootécnicos, los cuales resultaron estadísticamente irrelevantes para la previsión de precios a corto plazo. Por lo tanto, se concluye que la débil magnitud de la correlación confirma el predominio del mercado spot y la naturaleza biológica del activo, posicionando el sentimiento mediático predominantemente como un indicador coincidente de las condiciones del mercado.

Palabras clave: Análisis de Sentimiento. BERTimbau. Agroindustria. Formación de Precios. Bastos-SP.

1 INTRODUCTION

The municipality of Bastos-SP stands out as the largest poultry hub in the state of São Paulo, being responsible for approximately 45% of São Paulo's production and 11% of national egg production (MORAIS, 2025).

Recognized as the 'Capital of the Egg', according to Rodarte (2025), the region has an estimated daily production of 22 million units, which makes the financial sustainability of local producers dependent on efficiency in marketing and the ability to anticipate market volatilities.

The egg market, however, operates under *commodity* dynamics, characterized by narrow margins, high price volatility, and strong dependence on external factors, such as input costs and variations in demand (Dos Reis Filho et al., 2020). In this context, qualified information becomes an input as strategic as animal nutrition.

The poultry farmer's decision-making, historically based on experience and subjective intuition, today faces the challenge of predicting the price in a market increasingly dependent on external variables, requiring decisions based on economic and conjunctural analyses to the detriment of mere accumulated experience (EMBRAPA, 2024). In this scenario, the flow of digital news can influence trade expectations even before the facts materialize on the shelves.

In the Agriculture 4.0 scenario, data science has offered new perspectives for market intelligence. In a previous study, Morassuti and Neto (2025) demonstrated the feasibility of using *Google Trends* as a demand thermometer for the egg market in Bastos, showing that the volume of digital searches has a certain correlation with the local price dynamics. However, while *Google Trends* captures active user interest, there remains a gap in understanding how the provision of passive information, specifically the "tone" or sentiment of the specialized media, impacts asset pricing.

Advancing in this line of investigation, the present work is based on the premise that the media acts as a vector of expectations. Tetlock (2007) established the theoretical basis that the 'tone' of the news has predictive power over the financial market. Recently, Xu and Hsu (2022) validated this approach in the agricultural sector, demonstrating computationally that news-based sentiment indicators can capture price anomalies in agricultural products that escape traditional econometric models.

In view of this scenario, this article seeks to fill an informational gap on the efficiency of the physical market of perishable commodities, investigating the interdependence relationship between the sentiment expressed in the specialized media and the price fluctuation in the poultry cluster of Bastos-SP.

The originality of this research lies in the application of *Deep Learning* architectures (BERTimbau) in an over-the-counter market highly dependent on physical fundamentals, testing the boundary between media noise and the economic reality of the *analyzed cluster* . In addition to this introduction, the work is structured as follows: Section 2 details the methodological flow and data curation; Section 3 presents the results and the theoretical discussion in the light of the informational economy; and Section 4 summarizes the conclusions, limitations, and implications for agribusiness governance.

2 MATERIALS AND METHODS

To investigate the relationship between the sentiment expressed in agribusiness news and price fluctuation in the poultry sector of Bastos-SP, this study will employ an exploratory and quantitative approach, applying Natural Language Processing (NLP) techniques. As highlighted by Finatto, Lopes and Ciulla (2015), NLP seeks to create solutions to specific problems related to language recognition and reproduction, being a facilitating tool in the processing of large volumes of information.

The spatial cut comprises the production *cluster* of Bastos-SP, the largest poultry hub in the state of São Paulo, and the time frame covers the period from April 2023 to December 2026. The methodological flow was structured in four sequential stages: (i) Data Collection; (ii) Pre-processing; (iii) Sentiment Modeling via *Deep Learning*; and (iv) Statistical Correlation Analysis

2.1 DATA ACQUISITION AND CURATION

For the composition of the dependent variable (price), the daily historical series of the extra egg, white, in Bastos-SP, from the Center for Advanced Studies in Applied Economics (CEPEA, 2026), were extracted. These series allow us to monitor price dynamics and understand the sensitivity of the local market to external variables, whose daily variations reflect the direct influence of factors such as input costs and fluctuations in demand (CEPEA, 2026).

For the independent variable (feeling), a textual *corpus* was constituted from the collection of headlines from four news portals of high sectoral relevance: Avisite, Avinews, Canal Rural and Globo Rural (AVISITE, [n.d.]; AVINEWS, [n.d.]; CANAL RURAL, [n.d.]; GLOBO RURAL, [n.d.]). Unlike strictly automated approaches, we opted for manual data extraction and curation, aiming to establish a 'Gold Standard' of quality for the training and testing of NLP architecture.

This deliberate procedure allowed for rigorous semantic filtering through the standardized search term 'eggs', ensuring that the raw *dataset* of 636 records was free of institutional noise, social columns, or merely administrative edicts, elements that often compromise the accuracy of unsupervised *web scraping* algorithms.

2.2 SEMANTIC FILTERING AND NOISE TREATMENT

In order to ensure semantic relevance and data quality, an algorithmic filtering pipeline was applied. To this end, the *Pandas* library was used, which provides data structures and tools for the efficient manipulation of tables and time series, and is widely used in economic and market analysis (MCKINNEY; PANDAS DEVELOPMENT TEAM, 2026).

Through this tool, records containing terms associated with corporate events, sanitary management, or irrelevant niches, such as "vaccine", "pet", "congress", and "technology", were excluded. At the same time, an inclusion filter was applied to headlines containing economic markers, such as "price", "market", "export", "cost" and "supply".

This process refined the *original corpus*, resulting in 135 headlines of high economic relevance (N = 135).

2.3 SENTIMENT MODELING (BERT)

The polarity classification of the headlines was performed using the BERTimbau model (COSTA et al., 2020).

As highlighted by Costa et al. (2020), BERTimbau is the first language model pre-trained specifically for Brazilian Portuguese, achieving state-of-the-art results in various Natural Language Processing tasks.

The choice of this architecture is based on the robustness of systems based on *Bidirectional Encoder Representations from Transformers* (BERT) for capturing contextual nuances and its successful application in sentiment quantification aimed at modeling and predicting asset prices (CHAUDHRY, 2022).

Specifically, the *fine-tuned* nlptown/bert-base-multilingual-uncased-sentiment version was used, whose effectiveness for the classification of polarities in corpora of digital media and diversified databases was validated in recent comparative investigations, demonstrating high accuracy in identifying opinion trends (SAHOO et al., 2023).

The algorithm assigned a polarity score to each headline, which was later discretized into three classes: Negative (-1), Neutral (0), and Positive (1).

2.4 STATISTICAL ANALYSIS

Given the frequency discrepancy between the price data (daily) and the news (sporadic), the technique of weekly *resampling* (or weekly resampling) was used to harmonize the time series. The procedure was performed using the ".resample('W')" method of the *Pandas* library (MCKINNEY; PANDAS DEVELOPMENT TEAM, 2026), calculating the simple arithmetic mean of the variables for each epidemiological week.

It is noteworthy that weeks in which there was no information flow (news) were disregarded from the correlation analysis to avoid artificial neutrality bias, a procedure performed via the ".dropna()" function.

The validation of the influence hypothesis was performed through Pearson's Correlation Coefficient (), a parameter that measures the intensity and direction of the linear relationship between two quantitative variables, ranging from -1 to 1. Where a value close to 1 indicates strong positive correlation, -1 indicates strong negative correlation, and 0 indicates no linear relationship, testing the strength and direction of the linear relationship between the Weekly Sentiment Index and the Weekly Average Price of the physical asset.

3 RESULTS AND DISCUSSION

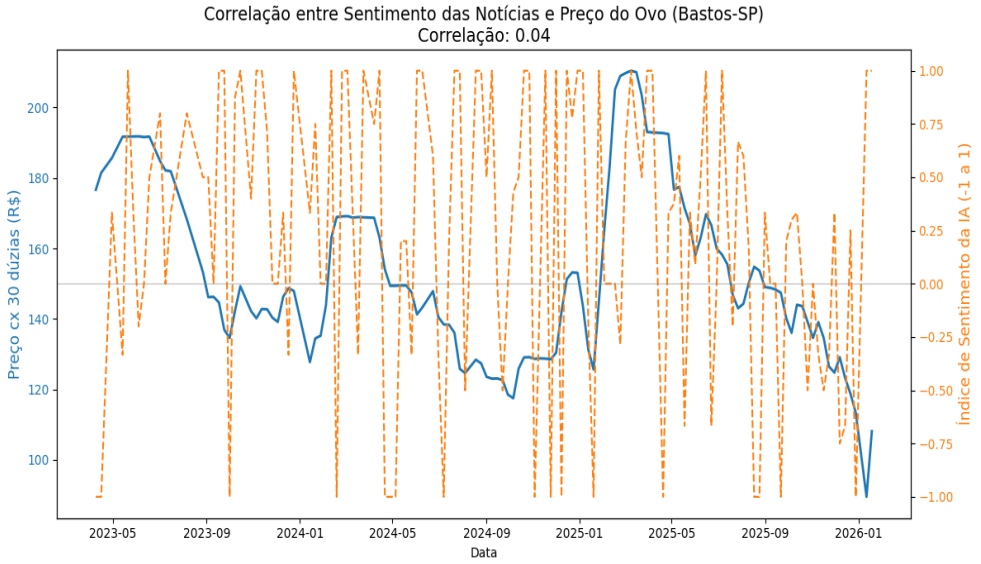
After the data collection stage, the analysis was structured into three distinct test scenarios to validate the consistency of the model.

In the first experimental scenario, which comprised the period from April 6, 2023 to January 16, 2026 using exclusively the technical portals Avinews and Avisite, it processed a raw *corpus* of 499 news items, resulting in a Pearson correlation of r = 0.04.

This result, illustrated in Figure 1, confirms that the flow of mostly technical and productive news has no explanatory power over the formation of prices in the short term for the *Bastos-SP* cluster.

Figure 1

Comparative time series between Average Price (CEPEA) and Sentiment (BERT) for the Avinews and Avisite portals (2023-2026). Correlation of 0.04.



Source: Prepared by the authors.

The second scenario incorporated semantic filtering (removal of institutional noise), reducing the sample to 65 news items, which marginally raised the correlation to 0.07.

Finally, the third scenario expanded the database by integrating macroeconomic sources (*Globo Rural* and *Canal Rural*), after filtering, a final *corpus* of 135 news items was obtained.

This scenario presented the highest correlation between the tests (r=0.12), a value that is still low, but significant compared to the previous scenarios, demonstrating that the curation of sources and semantic filtering contributed decisively to the qualification of the *dataset*.

Table 1 shows the analyzed test scenarios.

Table 1

Test scenarios carried out in the research and their correlations

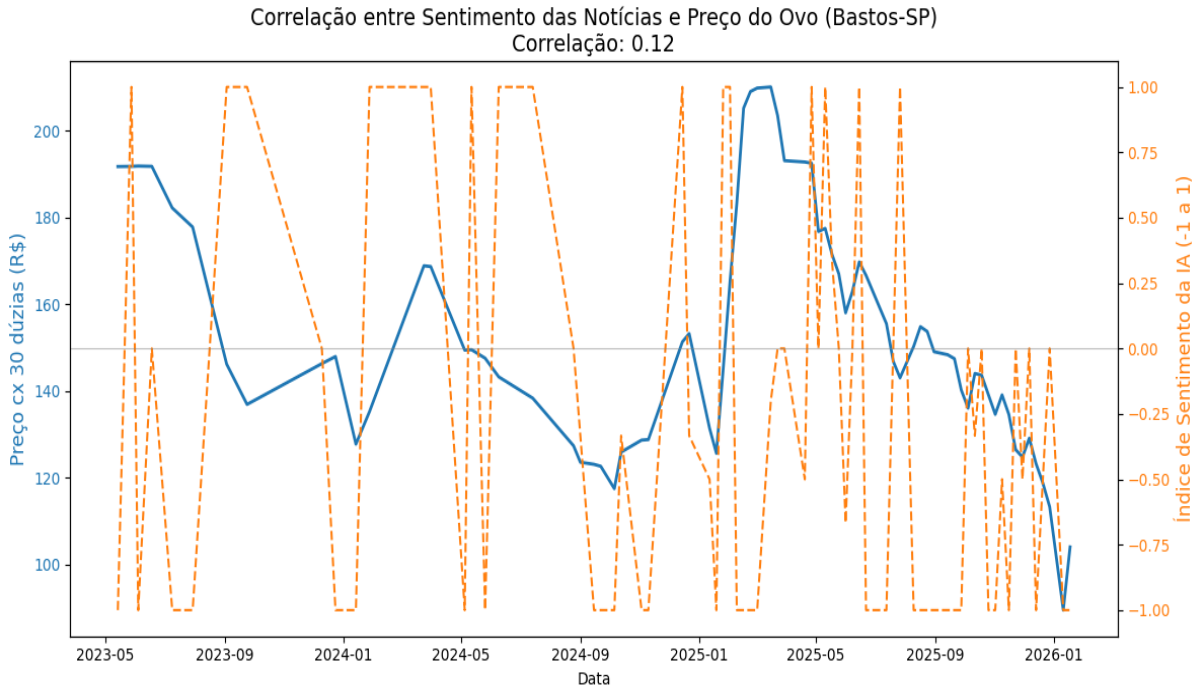
Test Scenario	**News No.**	**Sources Correlation (r)**
Gross	499	0.04 (Null)
Filtered	65	0.07 (Low)
Multi-source	135	0.12 (Best)

Source: Prepared by the authors.

The analysis of the time series, illustrated in Figure 2, reveals the dynamics between the Weekly Sentiment index (orange line) and the Average Price of Extra White Egg in Bastos-SP (blue line) over the period studied.

Figure 2

Comparative time series between the Average Egg Price (CEPEA) and the News Sentiment Index (NLP/BERT) in the period 2023-2025



Source: Prepared by the authors.

Visually, it is observed that the physical egg market presents high seasonal volatility, with characteristic peaks in the pre-Lent periods (February/March), driven by the increase in liturgical demand. The Sentiment index, in turn, demonstrates greater stability over time, reacting with intensity only to extreme events or macroeconomic shocks.

Quantitative statistical analysis corroborated this visual observation.

In the first experimental scenario, using the raw *dataset* (N=499) composed mostly of technical news from the Avisite portal, Pearson's correlation was statistically null (0.04). This indicates that the flow of news about management, health and technology has no explanatory power over price formation in the short term.

However, after the application of semantic filters and the integration of generalist sources (*Canal Rural* and *Globo Rural*), the *refined dataset* (N=135) presented a relevant qualitative change.

Pearson's correlation rose to 0.12 Although the magnitude of this correlation is still considered weak in the statistical literature, its multiplication by three times in relation to the base scenario (0.04-0.12) is a significant finding.

The analysis of the results shows a difference in the sensitivity of the market: the price of eggs in Bastos showed little sensitivity to news of a technical or productive nature ('inside the gate'). On the other hand, there was a slight reaction to macroeconomic variables ('out of the gate'), such as exchange rate fluctuations, input costs and export trends, frequent themes in the generalist portals analyzed.

However, the low correlation observed ($r < 0.20$) in all experimental scenarios ratifies the hypothesis that the Bastos pole operates under a predominantly physical dynamic, structurally diverging from the highly liquid financial markets.

While Tetlock's (2007) seminal literature establishes media sentiment as a leading indicator capable of predicting volatility in financial assets, the results obtained here suggest that, in laying poultry, news flow acts as a coincident indicator.

This distinction is based on the 'spot *market dominance hypothesis*', in which the price formation of highly perishable products is governed by the urgency of physical settlement and daily inventories at the trading counter, rather than by speculative expectations (SHRESTHA et al., 2020).

Unlike financial assets that can be retained based on positive sentiments, the biological nature of the egg imposes a natural barrier to the spread of media noise: the price responds with greater elasticity to physical supply and demand shocks than to digital informational buzz (LI; WANG; DIERSEN, 2024).

Therefore, the weak magnitude of the correlation ($r = 0.12$) in the multisource scenario does not indicate a failure of the NLP model, but rather evidence of the informational efficiency of a market anchored in the physical reality of the gate.

4 FINAL CONSIDERATIONS

The present study sought to fill a gap in the market intelligence literature for agribusiness, applying Natural Language Processing (NLP) techniques to investigate the relationship between specialized media sentiment and price formation in the poultry cluster of Bastos-SP.

The results obtained allow us to conclude that, for the period analyzed (April 6, 2023 to January 16, 2026), the Alta Paulista egg market operates under a pricing dynamic predominantly based on immediate physical variables, demonstrating low elasticity in relation to the news flow.

The weak, but positive, correlation (r=0.12), observed after the inclusion of macroeconomic sources, suggests that the local producer is marginally more sensitive to structural trends (grain costs, external scenario) than to zootechnical guidelines, which proved to be statistically irrelevant for the short-term price forecast.

From a methodological point of view, the work validated the effectiveness of the use of Deep Learning models (BERTimbau) combined with rigorous semantic filters for the structuring of unsupervised data. The "cleaning" of institutional noise has proven essential to avoid false correlations, highlighting the importance of data curation in Big Data projects in Agribusiness.

As limitations, the time window of approximately three years and the exclusive focus on digital news portals are pointed out, which makes it impossible to capture informal flows in social networks that could contain more agile market signals.

In addition, the analysis was based on the processing of headlines; Although these summarize the main fact, the absence of the integral body of the news may omit relevant contextual nuances.

It is also noteworthy that the option for manual curation, although it has established a "gold standard" of quality and ensured the elimination of institutional noise, imposes restrictions on the scalability of the corpus analyzed.

For future studies, it is suggested the evolution of this model to a multivariate architecture, integrating the Sentiment Index developed here with climate variables, future quotations of inputs (corn and soybean) and search volume data (*Google Trends*). It is also recommended the application of time lag analysis to identify the exact interval of information absorption by the physical market of Bastos-SP, aiming at the construction of hybrid predictive models capable of overcoming the accuracy of traditional econometric methods.

REFERENCES

Avinews. (n.d.). Avinews Brasil. https://avinews.com/br/

Avisite. (n.d.). Portal Avisite. https://www.avisite.com.br/

Canal Rural. (n.d.). Portal Canal Rural. https://www.canalrural.com.br/

Centro de Estudos Avançados em Economia Aplicada. (2026). Séries históricas diárias: Ovo tipo extra, branco, Bastos/SP. https://www.cepea.esalq.usp.br/br/indicador/ovo.aspx

Chaudhry, P. (2022). Bidirectional encoder representations from transformers for modelling stock prices. University of Delhi. https://www.researchgate.net/publication/358799138

Costa, W. Y., et al. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In Proceedings of the Brazilian Conference on Intelligent Systems (pp. 403–417). Springer. https://arxiv.org/abs/2009.08144

Reis Filho, I. J., Correa, G. B., Freire, G. M., & Rezende, S. O. (2020). Forecasting future corn and soybean prices: An analysis of textual information. In Proceedings of the Symposium on Knowledge Discovery, Mining and Learning (pp. 113–120). https://doi.org/10.5753/kdmile.2020.11966

Embrapa Suínos e Aves. (2024). Central de inteligência de aves e suínos (CIAS). https://www.embrapa.br/suinos-e-aves/cias/analises

Globo Rural. (n.d.). Portal Globo Rural. https://www.globo.com/globorural/

Li, Z., Wang, Z., & Diersen, M. (2024). Do agricultural commodity price spikes always stem from news? In NCCC-134 Conference on Applied Commodity Price Analysis. https://farmdoc.illinois.edu

McKinney, W., & Pandas Development Team. (2026). Pandas: Powerful Python data analysis toolkit. https://pandas.pydata.org/

Morais, E. (2025, julho 14). Conheça Bastos, a cidade brasileira que bota 290 ovos por segundo. Jornal Correio. https://www.correio24horas.com.br

Morassuti, M. G., & Neto, M. M. (2025). Google trends como ferramenta de análise do mercado de ovos. Revista QualyAcademics, 3(1), 895–900. https://doi.org/10.59283/unisv.v3n2.036

Rodarte, H. (2025, julho 14). 22 milhões de ovos por dia: Conheça a capital do ovo. Agro em Campo. https://agroemcampo.com.br

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Sahoo, A., et al. (2023). Comparative analysis of BERT models for sentiment analysis on Twitter data. In Proceedings of the International Conference on Smart Computing and Communications (pp. 1–6). IEEE.

Souza, F., Nogueira, R., & Lotufo, R. de A. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In Lecture Notes in Computer Science (Vol. 12455, pp. 403–417). Springer. https://doi.org/10.1007/978-3-030-61377-8_28

Shrestha, K., Subramaniam, R., & Thiyagarajan, T. (2020). Price discovery in agricultural markets. American Business Review, 23(1), 53–69.

Xu, J.-L., & Hsu, Y.-L. (2022). The impact of news sentiment indicators on agricultural product prices. Computational Economics, 59(4), 1645–1657. https://doi.org/10.1007/s10614-021-10189-4