

Opinion Mining on Reddit: Data Engineering and NLP for Analyzing Precarious Work in Post-Pandemic Brazil

Gabriel Bernalle Barbosa Matozinhos
gabrielbbmatozinhos@gmail.com
CEFET-MG

Daniel Hasan Dalip
hasan@cefetmg.br
CEFET-MG

Danilo Boechat Seufitelli
danilobs@cefetmg.br
CEFET-MG

Gabriel P. Oliveira
gabrielpdoliveira@gmail.com
PUC-MG

ABSTRACT

In post-pandemic Brazil, the rise of “pejotização” (hiring individuals as corporate entities) and micro-entrepreneurship fuels debates on labor precarization. Despite official narratives of innovation, rising labor lawsuits (up 57% in 2024) and fiscal warnings highlight a growing crisis. To investigate this dichotomy, we built a web data pipeline to analyze Reddit discussions on working conditions over time. Orchestrated via Mage.ai, our ETL flow ingests unstructured API data and applies NLP deep learning models (XLM-RoBERTa) for sentiment analysis and Zero-Shot classification. We also use dynamic topic modeling (BERTopic) to visually compare the discourse’s temporal evolution against macroeconomic indicators. Processing over 338,000 records, the results confirmed the pipeline’s viability and revealed predominantly negative sentiments (59.40%). These findings suggest that micro-entrepreneurship tends to function less as a voluntary endeavor and more as a survival mechanism tied to inflation and household debt, offering quantitative grounding for current legal debates.

KEYWORDS

Data Engineering, Natural Language Processing, Pejotização, Labor Precarization

1 INTRODUCTION

The Brazilian labor market faces paradoxes in the post-pandemic period. Recent economic indicators point to a formal heating of the economy (with the unemployment rate falling to 5.4% in October 2025 [17]). However, this apparent statistical stability coexists with structural precarization. The informality rate remains high and affects around 38.1% of the employed population (approximately 40 million workers) [11, 12]. The youth also faces particular vulnerability, as one in five Brazilians between 15 and 29 years old neither studies nor works [5]. In response to the scarcity of formal jobemployers often situate these new forms in the gray areas of legislation.

Consequently, phenomena like *pejotização* (hiring individuals as corporate entities to mask employment) and *Uberization*¹ consolidated themselves as workforce management strategies following the 2017 Labor Reform [15]. This dynamic generates intense institutional conflicts. The Supreme Federal Court (STF) favored unrestricted outsourcing [19], but the Superior Labor Court (TST) recorded a 57% increase in hiring fraud litigation [18]. The Attorney General’s Office (AGU) warns of fiscal risks to Social Security and the “hollowing out of labor rights” [2]. Sociologically, this reflects a “necessity entrepreneurship” [3, 13] rather than Schumpeterian innovation². The dismantling of social safety nets compels workers to become “entrepreneurs of the self” [14].

Traditional macroeconomic indicators fail to capture the subjectivity and qualitative perception of this workforce amidst the impasse between the official modernization narrative and the signs of precarization. We identify a gap in understanding how workers feel and express this transition (whether as an achievement of autonomy or as a survival strategy marked by anxiety and indebtedness). Furthermore, capturing these qualitative nuances on a large scale requires computational methods that exceed the capabilities of traditional demographic surveys.

To fill this gap, we propose a Data Engineering approach applied to the Social Sciences. We aim to develop a data pipeline orchestrated by the Mage.ai tool³. This pipeline collects, processes, and analyzes large volumes of unstructured data from the social network Reddit⁴. We employ advanced Natural Language Processing (NLP) techniques, including BERT models for contextual classification and dynamic topic modeling (BERTopic), to monitor the evolution of the discourse on labor in Brazil.

We chose Reddit as the primary source because its relative anonymity mitigates the ‘social desirability’ bias common in strictly professional networks like LinkedIn. This characteristic grants access to authentic accounts of actual working conditions and perceptions of precarization. Furthermore,

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia’2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

¹**Uberization:** A term used to denote the transformation of labor relations mediated by digital platforms, replacing formal employment with on-demand work. Available at: <https://www.academia.org.br/nossa-lingua/nova-palavra/uberizacao>

²Reference to Joseph Schumpeter’s concept of “Creative Destruction”, defining entrepreneurship as the engine of innovation, differentiating it from mere self-employment service provision.

³<https://www.mage.ai/>

⁴<https://www.reddit.com/>

the platform offers a structural organization conducive to analysis by dividing content into thematic communities (subreddits). To define the scope of the analysis, we apply a filter targeting 18 Brazilian communities dedicated to economics, labor, innovation, and technology (including specific computing forums such as *r/brdev*). We then cross-reference the posts with keywords related to the new hiring dynamics.

Based on this objective, we seek to answer the following research problem: **How can NLP techniques applied to social media data complement official macroeconomic indicators in characterizing the precarization of work?** Specifically, we define the following research questions (*RQs*):

- RQ1.** What are the latent topics and predominant sentiments in discussions about *pejotização* and entrepreneurship on the Brazilian Reddit?
- RQ2.** To what extent have language models (BERT) proven effective in distinguishing narratives of labor autonomy from those marked by precarization and necessity?
- RQ3.** How does the temporal evolution of the extracted social narratives visually track the fluctuations of government metrics for unemployment and inflation?

Finally, we characterize the mapped social sentiment regarding work precarization in post-pandemic Brazil. We then contrast these findings with official structured government data. Through visualizations and comparisons, we investigate whether the deterioration of traditional macroeconomic indicators (such as unemployment, inflation, indebtedness, and labor litigation) precedes or tracks the increase in discourses about vulnerability and fiscal risk. Thus, we offer a layer of qualitative intelligibility to the official numbers. These numbers, in isolation, fail to capture the worker's subjectivity.

The remainder of this article is organized as follows. Section 2 discusses related work. Section 4 details the data pipeline and NLP methodology. Section 5 validates the applied models. Section 6 presents the results and visual comparisons. Finally, Section 7 concludes the paper.

2 RELATED WORK

The sociological literature discusses the transformation of the Brazilian labor market through the lens of structural precarization and technological evolution [7]. Damasceno et al. [9] argue that legislative flexibilization following the Labor Reform distorted employment relations. Employers compel individuals to accept corporate entity (PJ) status to reduce costs. In the context of formal job scarcity, Abilio [1] defines digital platform workers as “subordinate self-managers” who bear economic risks under algorithmic control. Coutinho and Chalfin [8] map how precarization affects worker identities and subjectivity. These qualitative studies provide insights into the meanings attributed to work but face scale limitations. Qualitative methods cannot monitor the narratives' evolution in real time across large population volumes.

To overcome the scale constraints of qualitative methods, the economic literature adopts Nowcasting to monitor socioeconomic conditions in real time [4]. Researchers use high-frequency data flows to anticipate official statistics and assist

decision-makers. Approaches relying on unstructured web data like search volumes face a semantic limitation because they miss ironies and sentiment polarity. Advances in Natural Language Processing (NLP) offer solutions to this semantic gap. Xu et al. [20] demonstrate that Transformer-based models like BERTopic outperform probabilistic approaches like LDA in clustering social media discourses. Recent literature demonstrates the effectiveness of combining topic modeling with Large Language Models (LLMs) to semantically organize data for strategic decision-making [6]. Other studies apply deep learning and sentence embeddings to extract human perceptions from Location-Based Social Networks, which confirms the viability of scalable social sensing [16]. Furthermore, researchers combine machine learning models with lexicons such as LIWC to effectively identify subjective states and depression symptoms in user-generated content [10].

The literature reveals a methodological gap. Sociological approaches [1, 8] investigate structural precarization but lack tools for massive analysis. Nowcasting techniques [4] anticipate economic scenarios but neglect discourse nuances. We propose a convergence of these perspectives to address this limitation. We utilize a Data Engineering pipeline and language models like BERT and BERTopic to bring scalability to qualitative analysis. Our approach qualifies the debate through NLP, whereas monitoring methods only measure post volume. We distinguish whether the digital noise reflects a narrative of labor autonomy or structural precarization.

3 DATASET CHARACTERIZATION

We source our empirical corpus from 18 thematic Brazilian subreddits dedicated to economics, technology, and labor (e.g., *r/brdev*, *r/antitrampo*). This purposive selection concentrates discussions about “*pejotização*” and platform-based work, directly aligning the corpus with our research objectives. We extracted 338,773 records spanning from 2018 to 2025. This initial corpus contains substantial digital noise, including spam, cross-posted commercial content, ambiguous acronym usage, and discussions unrelated to labor market dynamics.

We implemented a multi-stage filtering protocol that prioritizes semantic relevance over corpus size. Table 1 summarizes each stage of the data reduction process. This protocol removes non-relevant discussions and preserves only records that provide sufficient contextual evidence for downstream analysis. The reduction from 338,773 to 1,697 validated records illustrates the challenges of social sensing with unstructured social media data rather than a loss of meaningful information. By applying strict confidence thresholds, we ensure that sentiment analysis and topic modeling rely exclusively on semantically robust texts while minimizing the noise commonly found in platform-derived datasets. The final dataset contains 1,697 records, which we classify into two primary thematic axes: *Pejotização* and *Uberization*.

After characterizing the composition and quality of our empirical corpus, we now describe the pipeline orchestration and the NLP techniques in Section 4. That section explains how our ETL pipeline transforms these validated documents into the quantitative evidence analyzed in the following sections.

Table 1: Data Processing Funnel

Stage	Document Count
Raw Extraction	338,773
Filtered (Regex)	8,481
Validated (Zero-Shot)	1,697

4 METHODOLOGY

This work is classified as applied and exploratory research with a quantitative approach, grounded in the intersection between Data Engineering and Computational Social Sciences (*Social Sensing*). We aim to establish a methodological bridge between qualitative sociological analyses and the massive processing of textual data, enabling scalability to investigate the precariousness of labor in the post-pandemic scenario.

We structured the method into four macro-stages to execute this investigation. The first stage involves the extraction of unstructured Reddit texts and the ingestion of official macroeconomic/legal indicators (IBGE, BCB, TST). The second phase focuses on preprocessing and semantic cleaning via Regular Expressions (Regex). In the third stage, we apply an advanced NLP flow with Transformer models (Zero-Shot classification, XLM-RoBERTa, and BERTopic). Finally, we store the consolidated data in a Google BigQuery Data Warehouse for comparative analysis. Figure 1 illustrates the end-to-end architecture orchestrated via Mage.ai.

Architecture and Orchestration. We isolate the computational environment using Docker, encapsulating dependencies for Mage.ai and Python. Given the computational cost of Transformer inference, we enable GPU acceleration (NVIDIA CUDA 11.8+). Our design adheres to *Modern Data Stack* (MDS) pillars: **modularity** (segregation of domains), **scalability** (cloud-native processing), **reproducibility** (containerization), and **auditability** (intermediate checkpoints). We prefer Mage.ai over alternatives due to its native inspectability, block-oriented modularity, and low operational overhead.

Data Ingestion. We segment the extraction into two fronts: unstructured social data (Reddit) and structured government data (IBGE, BCB, TST). For Reddit, we use the PRAW library on 18 Brazilian communities (*r/investimentos*, *r/brdev*, *r/antitrampo*), applying purposive sampling with 18 keywords (e.g., “MEI”, “PJ”). We manage API limits via *ThreadPoolExecutor*. For official data, we ingest: (1) Unemployment Rate (IBGE) to assess workforce absorption; (2) Cumulative Inflation (IPCA) to measure purchasing power erosion; (3) Household Indebtedness (BCB) to map financial fragility; and (4) New Labor Lawsuits (TST) as a proxy for litigation. To align these sources, we convert all series to monthly granularity, allowing for visual trend analysis.

Pre-processing and Semantic Cleaning. Textual data from social networks is inherently noisy. We perform structural normalization (removing line breaks/HTML tags) and apply validation filters (language detection and length constraints). We reduce false positives (e.g., “DAS” vs. preposition+article)

using context-sensitive Regex filters, which remove over 97% of digital noise. Table 2 summarizes these patterns.

NLP Pipeline. We use three encoder-based Transformer models. First, we use **mDeBERTa-v3** for *Zero-Shot* classification of Reddit posts into six socioeconomic categories. We apply a confidence threshold of 0.55 to prioritize precision. Next, we use **XLM-RoBERTa** for sentiment analysis. Finally, we apply *Guided BERTopic* with *paraphrase-multilingual-mpnet-base-v2* to extract and track temporal narrative evolution. These results are persisted in BigQuery for trend comparison.

5 QUALITY ASSESSMENT OF NLP MODELS

This section details the experimental protocol and validation criteria used to assess the pipeline’s reliability. Before interpreting the results, we establish quality control metrics for the NLP inferences (thematic categories and sentiment predictions). We also define the confidence thresholds that support the integrated time-series analysis.

5.1 Sensitivity Analysis and Confidence Threshold Optimization

Zero-Shot probabilistic models classify text and assign a confidence score (between 0 and 1) to each prediction. Since social media texts contain extensive slang and ambiguity, the model often produces low-confidence predictions. This characteristic increases the rate of false positives.

This analysis quantifies the *trade-off* between two competing metrics: the model’s global average confidence and the **data retention rate**. The retention rate defines the percentage of Reddit comments that pass the filter. Figure 2 illustrates the impact of the cutoff on these variables.

Results indicate that a 0.55 threshold represents the ideal inflection point for the architecture. As Figure 2 shows, this cutoff guarantees a global average confidence of 71.09%. This methodological choice retains 20.64% of the initial data volume. We deliberately prioritize quality over quantity. This strict filter ensures that only highly reliable predictions feed the subsequent macroeconomic cross-references.

We then evaluate how this cutoff impacts the six sociological categories (Survival and Necessity, Fiscal Risk and Debt, Pejotização and Subordination, Precarization of Rights, Management and Opportunity, and Junk/Off-Topic). We analyze the confidence score distribution for each class to assess the model’s performance across different themes. Figure 2 illustrates these distributions.

Reddit’s informal vocabulary naturally concentrates the model’s probability mass below 0.50 for most classes (Figure 3). In this context, the 0.55 cutoff acts as a strict methodological filter. This threshold suppresses noise and ambiguity entirely. It preserves only the data in the upper confidence deciles. Core categories, such as “Survival and Necessity” and “Fiscal Risk and Debt”, retain a sufficient density of high-certainty data. These results demonstrate the accurate isolation of precarization discourses.

We label documents below this threshold as “Discarded due to Low Confidence” and exclude them from the database. This

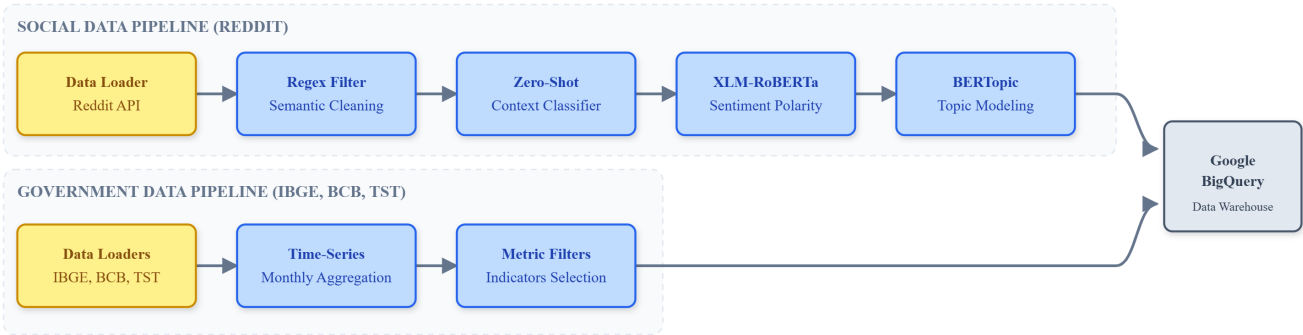


Figure 1: End-to-end data pipeline architecture orchestrated via Mage.ai. The diagram depicts the parallel processing of unstructured social media data (Reddit) and structured government indicators, both converging into a centralized Data Warehouse for analysis on Google Big Query.

Table 2: Regex patterns for contextual disambiguation.

Target Context	Applied Regex Pattern	Capture Examples
DAS Tax	<code>\b(?:o do no pagar guia boleto valor atrasado)\s+das\b</code>	“pagar das”
Taxation	<code>(?:\b(?:pagar pago o qual)\s+imposto\b \bimposto\s+(?:de do mei))</code>	“imposto do mei”
Pejotização	<code>(?:\b(?:vaga contrato trabalhar sou)\s+pj\b \bpj\s+(?:ou versus clt))</code>	“vaga pj”

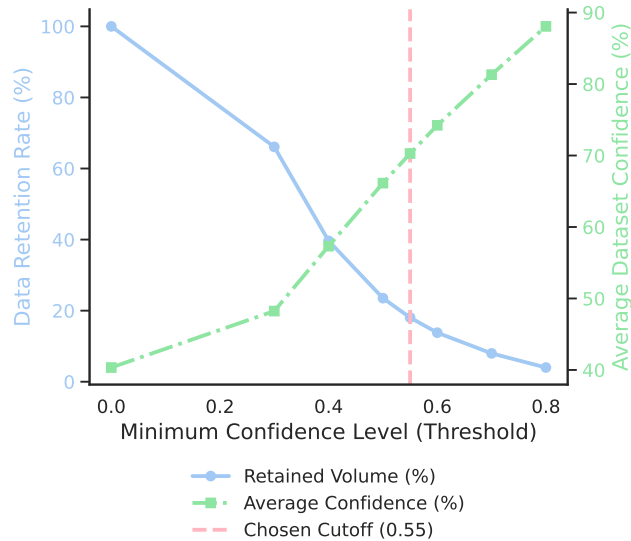


Figure 2: Methodological Trade-off: Volume Retention versus Zero-Shot Model Average Confidence.

step yields a purified dataset, free of predictive distortions and statistically robust for subsequent models.

5.2 Polarity Inference Confidence

We also evaluate the XLM-ROBERTa model’s prediction confidence. This algorithm classifies the emotional polarity of each validated text as Negative, Neutral, or Positive. Figure 4

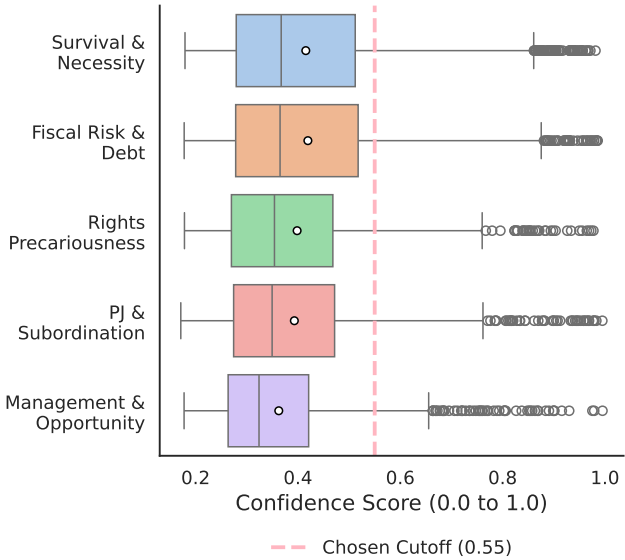


Figure 3: Zero-Shot Model Confidence Distribution by Semantic Category.

shows the confidence distribution score. The model interprets informal text effectively: the negative sentiment (the majority class) achieved the highest median precision. Its first quartile remains above 0.80. The global average confidence reached 84.39%. This high score guarantees reliable qualitative analysis and accurate visual comparisons against macroeconomic default rates.

The confidence scores vary significantly across the three sentiment classes. The model classifies negative texts with a major confidence (median near 0.90). Positive texts, however, show lower median confidence and greater variance. This variance reflects the linguistic nature of the dataset. Users express dissatisfaction through direct and explicit vocabulary. Conversely, positive narratives in this context often contain irony or ambiguous terms. This architectural behavior validates the model's ability to isolate genuine precarization and other complaints with higher precision.

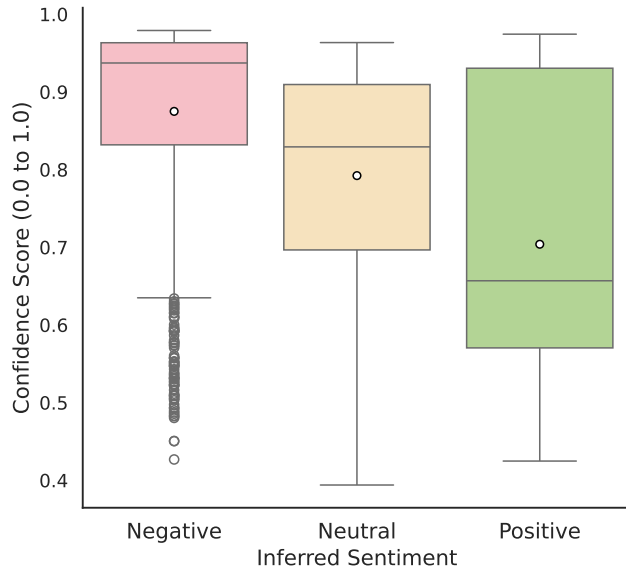


Figure 4: XLM-RoBERTa Model Confidence by Sentiment.

5.3 Topic Modeling Density Assessment (BERTopic)

The pipeline's final stage clusters narratives via BERTopic. Traditional predictive models force the classification of all inputs. BERTopic, however, relies on the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm. This approach dynamically identifies and excludes outliers.

Density metrics demonstrate a highly cohesive model performance. The algorithm grouped 97.3% of the documents (1,638 texts) into semantically valid topics and isolated only 2.7% (46 documents) as noise or Topic -1 (outliers). The average clustering confidence reached 94.87%. This high metric confirms the success of our prior sanitization and disambiguation steps (Regex and Zero-Shot threshold). These filters produced a dataset without structural distortions. Consequently, such a dataset allows us to extract highly representative narratives for our exploratory analysis. Table 3 summarizes the final clustering architecture performance.

This high confidence metric confirms the success of our prior sanitization and disambiguation steps (Regex and Zero-Shot threshold). These filters produced a dataset without

Table 3: BERTopic Clustering Performance Metrics

Clustering Metric	Result
Total Input Documents	1,684
Clustered Documents	1,638 (97.3%)
Identified Outliers (Topic -1)	46 (2.7%)
Global Average Confidence	94.87%

structural distortions. Consequently, it allows us to extract highly representative narratives for our exploratory analysis.

6 RESULTS AND DISCUSSION

This section presents the results generated by our pipeline. We structure the analysis to directly address the research questions (RQ1, RQ2, and RQ3). The discussion transitions from the pipeline's quantitative validation to a sociological and macroeconomic interpretation of digital discourse regarding labor precarization.

The raw extraction from all communities and their comment trees produces 338,773 text lines. We applied pre-processing and Regex semantic disambiguation. We discarded 330,292 noisy or uncontextualized records (e.g., links, digital artifacts, and non-labor terms). The pipeline retained 8,481 filtered lines for the Zero-Shot classifier. Next, we applied the 0.55 confidence threshold and removed 261 'Junk/Off-Topic' records. This filtering yielded a final volume of 1,697 validated documents. This reduction highlights the prevalence of digital noise on social media platforms. By removing irrelevant content, we ensure that the analysis focuses exclusively on authentic labor-related discussions. Table 4 summarizes this extraction process.

Table 4: Data Processing Funnel

Stage	Records	Action
Raw Extraction	338,773	Initial Volume
Regex Filtered	8,481	Semantic Sanitization
Zero-Shot Validated	1,697	Confidence Cutoff (0.55)

We stored these high-certainty records in the Google BigQuery Data Warehouse. This final empirical base serves as the foundation for the following analyses, starting with the investigation of labor discourse dynamics (RQ1).

6.1 Semantic Characterization of Social Discourse

To answer RQ1 ("What are the latent topics and predominant sentiments regarding *pejotização* and entrepreneurship on Brazilian Reddit?"), we submitted the validated dataset to Topic Modeling via BERTopic. Of the 1,697 documents approved in the polarity stage, 1,684 provided sufficient vector density for clustering. The algorithm performed consistently. It grouped 97.3% of the texts (1,638 documents) and labeled only 2.7% (46 documents) as noise (outliers). The model isolated two central macro-semantic groups: "Pejotização and

Precarization of Employment Relationships" and "Uberization and Platform Work." Volumetric distribution highlights the dominance of the *pejotização* debate. This topic accounts for 79.1% (1,295 documents) of extracted narratives, whereas platform work represents 20.9% (343 documents).

Figures 5a and 5b detail the morphological composition of these clusters. The vertical axis represents the relevance score extracted via c-TF-IDF. The *Pejotização* cluster (Figure 5a) reveals a lexicon focused on bureaucratic and fiscal burdens. Key terms such as *imposto* (0.065), *mei* (0.051), *pj* (0.047), and *pagar* (0.037) show that workers in this regime primarily discuss tax obligations and the costs of the 'fake company' status rather than entrepreneurship. The term *clt* (0.031) benchmarks these experiences against lost social protections. Conversely, the Uberization narrative (Figure 5b) centers on survival-oriented operations, evidenced by terms such as *uber* (0.144), *motoboy* (0.080), and *ifood* (0.048).

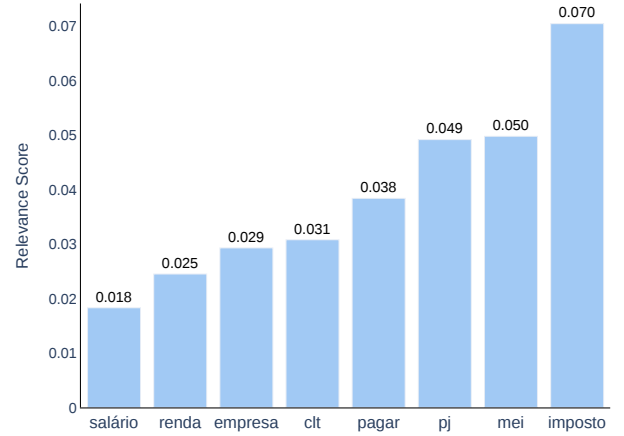
Dynamic Topic Modeling (DTM) allows us to track these narratives over time (Figure 6). The Uberization discussion (orange line) gained traction in 2021 and reached a high-frequency plateau between 2023 and 2025. This timeline coincides with the post-pandemic economic contraction. Conversely, the *Pejotização* narrative (blue line) exhibits longer latency. It began an exponential climb in 2024, which mirrors the sharp rise in labor litigation [18].

6.2 Polarity Inference and Predominant Sentiments

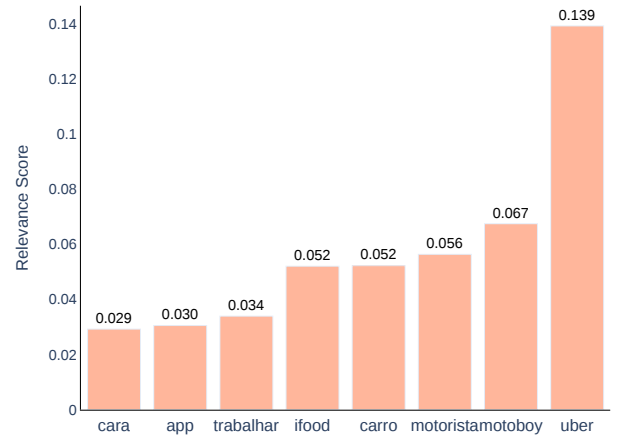
To address RQ2 ("To what extent have language models proven effective in distinguishing narratives of labor autonomy from those marked by precarization and necessity?"), we process the 1,697 validated documents using the XLM-RoBERTa model. The results categorize 59.40% (1,008) as negative and 38.30% (650) as neutral. Only 2.30% (39) expresses positive polarity.

We cross-reference these polarities with the Zero-Shot thematic categories (Figure 8). This analysis highlights high levels of dissatisfaction across informality narratives. The 'Precarization of Rights' category records 80.4% negative sentiment.

The predominance of negative sentiment may be partially attributed to the functional nature of the platform. On Reddit, discourse frequently focuses on venting, troubleshooting, and seeking support regarding professional grievances; thus, negative polarity in discussions concerning 'pejotização' and entrepreneurship likely reflects this platform-specific bias toward problem-solving rather than solely representing the objective labor experience. Neutral classifications, conversely, predominantly consist of technical queries—such as tax obligations or bureaucratic compliance rather than an absence of opinion. While the minimal positive polarity (2.30%), it primarily highlights that these digital forums operate as hubs for crisis management and peer support, rather than spaces for professional endorsement.



(a) Pejotização and Precarization



(b) Uberization and Platforms

Figure 5: Top Relevance Terms by Thematic Category (c-TF-IDF Score).

6.3 Integrated Analysis and the Precarization Cycle (2018-2025)

We address RQ3 ("To what extent do social narratives correlate with macroeconomic indicators of labor precarization?") through a visual comparison of social discourse and official indicators. This alignment reveals that recent unemployment drops mask a cycle of indebtedness and precarization.

This financial distress can function as a structural catalyst for necessity-based entrepreneurship. In many cases, workers burdened by debt and unable to secure formal employment tend to accept precarious arrangements less as a choice for autonomy and more as a survival mechanism.

Inflation and Household Indebtedness. Post-pandemic inflation increases between 10% and 12%. This economic pressure compelled families to rely on credit for subsistence. Figure 9

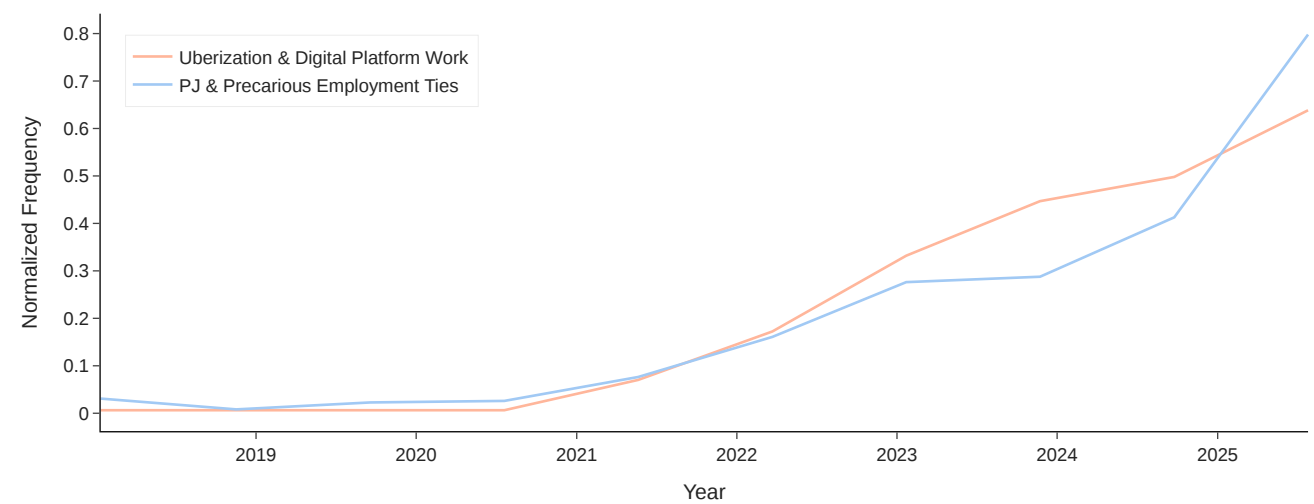


Figure 6: Temporal Evolution of Narratives on Labor Precarization.

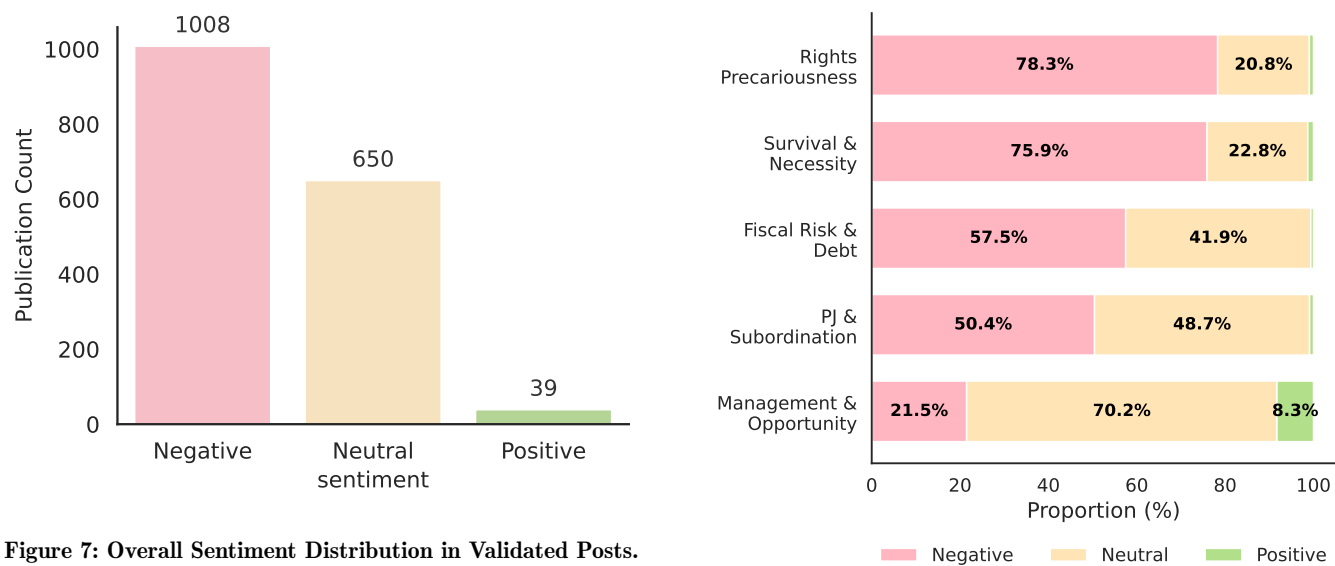


Figure 7: Overall Sentiment Distribution in Validated Posts.

Figure 8: Sentiment Proportion by Thematic Category.

illustrates a sharp rise in household indebtedness from 42% to 50% by 2022.

Informality and the Unemployment Paradox. The official unemployment rate fell to historical lows (5% – 7%). Our analysis, however, indicates that this decline masks the market’s absorption of informal labor. The convergence of rising informality rates (38.1%) and high default rates confirms that workers migrate to precarious, volatile arrangements. Finally, the surge in labor litigation (Figure 10) demonstrates that the legal system now absorbs the social costs of these atypical contracts.

This trajectory points toward what appears to be a self-reinforcing cycle: economic instability pushes many workers into unprotected labor, and this lack of contractual security

is closely associated with the observed spike in litigation. In this context, the legal system often emerges as a final safety net for a workforce lacking formal stability.

7 CONCLUSION

In this work, we investigated the dynamics of labor precarization in Brazil between 2018 and 2025. We combined Data Engineering and Natural Language Processing to integrate perceptions extracted from social networks with official macroeconomic indicators. Our architecture fulfilled its purpose: we demonstrated that textual data from digital



Figure 9: Household Indebtedness Evolution.

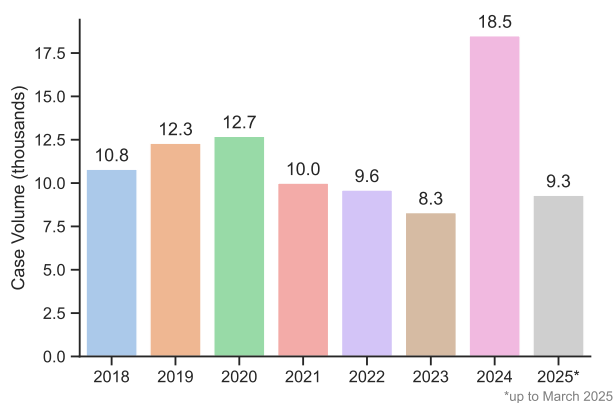


Figure 10: Volume of New Cases by Employment Relationship (TST).

platforms (when properly processed) function as a qualitative indicator complementary to official statistics. These data capture subjective dimensions that IBGE and Central Bank figures cannot reach in isolation. Regarding our research questions, we support the following conclusions:

RQ1.: Using BERTopic modeling, we identified two macro-axes that dominate labor market narratives: *Pejotização* and Precarization of Employment Relationships (79.1% of narratives) and Uberization and Platform Work (20.9%). Our morphological analysis showed that the debate on *pejotização* focuses on tax burdens and the loss of rights (compared to the CLT), while uberization focuses on operational costs and daily survival.

RQ2.: Our polarity inference revealed that these narratives are permeated by high degrees of dissatisfaction. We found that the predominant sentiment was negative in 59.40% of the posts, with only 2.30% positive reports. The “Management and Opportunity” category, which house the “innovative entrepreneurship” discourse, showed 72.4% neutrality associated with operational and bureaucratic doubts. We suggest that, in this dataset, micro-entrepreneurship is portrayed

primarily as an immediate survival arrangement rather than a choice of vocational autonomy.

RQ3.: Through a visual comparison between digital narratives and macroeconomic time series, we revealed a chronological chain consistent with four phases: (i) inflationary shock and debt (2021–2022); (ii) SELIC rate hike and rising default (2022–2023); (iii) migration to informality and the unemployment paradox (2022–2024); and (iv) the explosion of labor litigation (2024–2025). Although unemployment may lead to structural job precarization, further empirical analyses are required to substantiate this claim.

We offer a structural methodological contribution. We found that combining semantic filtering via contextual Regex with *Zero-Shot* classification proved to be an effective strategy for quality control of noisy corpora. By eliminating 97% of the raw volume, we reduced computational costs while maintaining the representativeness of the final corpus. In practical terms, these results qualify the ongoing legal debate.

Limitations. We acknowledge three methodological limitations. First, sample representativeness: Reddit concentrates users with higher education and digital access, excluding traditional informal workers and populations with lower literacy. Second, the analytical nature: our integration between digital discourse and macroeconomic indicators employed only visual trend analysis. Because we did not apply formal statistical tests, such as correlation coefficients or Lag Analysis models, we consider the identified associations to be descriptive; the synchronization of curves configures a temporal hint but does not attest to causality. Third, the model architecture: our strict 0.55 threshold in *Zero-Shot* classification retained only 20.64% of the pre-filtered corpus. Our lack of evaluation of the discarded material prevents the verification of potential thematic biases in the final set. Furthermore, we acknowledge that the XLM-RoBERTa model may misinterpret ironies and regionalisms typical of informal Portuguese.

Future Work. Our improvement idea includes statistical formalization of visual associations. We plan to apply Spearman’s correlation coefficient between the negative sentiment, default, and unemployment series to test the precedence of digital discourse over official indicators. This analysis requires configuring 1- to 6-month lag windows, which we intend to use to convert the current exploratory study into a statistically valid predictive model.

Finally, we recommend expanding the geographic and platform scope. Collecting data from Twitter/X and regional communities in the North and Northeast will allow us to test the robustness of the identified patterns and verify whether the mapped precarization cycle concentrates in large urban centers or constitutes a structural phenomenon across the entire Brazilian labor market.

ACKNOWLEDGMENTS

This research was supported by the Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG).

AI Acknowledgement. The authors utilized GenAI tools for text refinement, language editing, and code optimization. Following the use of these tools, the authors reviewed, validated, and edited the entire content, assuming full responsibility for the integrity and accuracy of the final publication.

Ethical considerations. Not applicable.

REFERENCES

- [1] Ludmila Costhek Abilio. 2019. Uberização: Do empreendedorismo para o autogerenciamento subordinado. *Psicoperspectivas* 18, 3 (2019), 1–11. doi:10.5027/psicoperspectivas-vol18-issue3-fulltext-1674
- [2] AGU. 2025. Para advogado-geral da União, pejetização deixa o trabalhador vulnerável e em risco. Gov.br. <https://www.gov.br/agu/pt-br/comunicacao/noticias/para-advogado-geral-da-uniao-pejetizacao-deixa-o-trabalhador-vulneravel-e-em-risco> Acesso em: 04 dez. 2025.
- [3] Ricardo Antunes. 2006. *Adeus ao trabalho?: ensaio sobre as metamorfoses e a centralidade do mundo do trabalho* (11 ed.). Cortez, São Paulo.
- [4] Bañbura, Marta, Giannone, Domenico, Modugno, Michele, Reichlin, and Lucrezia. 2013. *Now-casting and the real-time data flow*. Working Paper Series 1564. European Central Bank, Frankfurt am Main, Germany.
- [5] Vinícius Britto. 2023. Um em cada cinco brasileiros com 15 a 29 anos não estudava e nem estava ocupado em 2022. Agência de Notícias IBGE. <https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/38542-um-em-cada-cinco-brasileiros-com-15-a-29-anos-nao-estudava-e-nem-estava-ocupado-em-2022> Acesso em: 04 dez. 2025.
- [6] Daniel Carvalho, Antônio Pereira, Washington Cunha, Elisa Tuler, Diego Dias, and Leonardo Rocha. 2025. Uma Ferramenta para Organização Semântica de Informações de Helpdesk por meio de Modelagem de Tópicos. In *Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)* (Rio de Janeiro, RJ, Brasil). SBC, Porto Alegre, RS, Brasil.
- [7] Jessé de Hollanda Cordeiro Júnior, Tito Livio Gomes Osorio, Marcelo Genestra, Flávio Campos Carélli, Edney Soares Trindade, Rafael Iacillo Soares, and Rafael Menezes de Castro. 2005. Tecnologia e Desemprego. In *Anais do II Simpósio de Excelência em Gestão e Tecnologia (SEGeT)*. Associação Educacional Dom Bosco, Resende, RJ, Brasil, 684–695.
- [8] Coutinho and Maria Chalfin. 2009. Sentidos do trabalho contemporâneo: as trajetórias identitárias como estratégia de investigação. *Cadernos de Psicologia Social do Trabalho* 12, 2 (2009), 189–202.
- [9] Damasceno, Marina Teixeira, Gonçalves, and Jonas Rodrigo. 2020. Breves considerações sobre a Pejetização e a Reforma Trabalhista. *Anais do Congresso de Políticas Públicas e Desenvolvimento Social da Faculdade Processus* 2, 3 (2020), 55–61.
- [10] Silas Lima Filho, Monica Ferreira da Silva, and Jonice Oliveira. 2024. Depression Symptoms Identification through Social Media Data: Applying Design Science Research to Develop a Classification Model. In *Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)* (Juiz de Fora, MG, Brasil). SBC, Porto Alegre, RS, Brasil.
- [11] IBGE. 2025. PNAD Contínua Trimestral: desocupação cresce em 12 das 27 UFs no primeiro trimestre de 2025. Agência de Notícias IBGE. <https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/43421-pnad-continua-trimestral-desocupacao-cresce-em-12-das-27-ufs-no-primeiro-trimestre-de-2025> Acesso em: 04 dez. 2025.
- [12] IPEA. 2024. Retrato das Desigualdades de Gênero e Raça. Ipea. <https://www.ipea.gov.br/portal/retrato/indicadores/mercado-de-trabalho/desemprego-informalidade-subutilizacao-e-inatividade/apresentacao> Acesso em: 04 dez. 2025.
- [13] Jacob Carlos Lima and Roberto Veras de Oliveira. 2021. O empreendedorismo como discurso justificador do trabalho informal e precário. *Contemporânea* 11, 3 (2021), 905–932. doi:10.4322/2316-1329.2021028
- [14] Attila Magno and Silva Barbosa. 2011. O empreendedor de si mesmo e a flexibilização no mundo do trabalho. *Revista de Sociologia e Política* 19, 38 (2011), 121–140.
- [15] Danielly Silva Pereira. 2017. Reforma Trabalhista na Terceirização. Jusbrasil. <https://www.jusbrasil.com.br/artigos/reforma-trabalhista-na-terceirizacao/524240480> Acesso em: 04 dez. 2025.
- [16] Frances Albert Santos, Thiago Henrique Silva, and Leandro Aparecido Villas. 2024. Urban Perception Extraction from Texts Shared on Social Media: Framework and Applications. In *Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)* (Juiz de Fora, MG, Brasil). SBC, Porto Alegre, RS, Brasil.
- [17] Breno Siqueira. 2025. No trimestre encerrado em outubro, desocupação é de 5,4%, menor taxa desde 2012. Agência de Notícias IBGE. <https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/45272-no-trimestre-encerrado-em-outubro-desocupacao-e-de-54-menor-taxa-desde-2012> Acesso em: 04 dez. 2025.
- [18] TST. 2025. Número de processos julgados pelo TST cresce 57% em cinco anos. TST. <https://www.tst.jus.br/-/nÃzmero-de-processos-julgados-pelo-tst-cresce-57-em-cinco-anos> Acesso em: 04 dez. 2025.
- [19] TST. 2025. Pejetização: STF suspende processos sobre licitude de contratos de prestação de serviços. <https://portal.trt11.jus.br/index.php/comunicacao/10251-pejetizacao-stf-suspende-processos-sobre-licitude-de-contratos-de-prestacao-de-servicos> Acesso em: 04 dez. 2025.
- [20] Weiai Wayne Xu, Jean Marie Tshimula, Ève Dubé, Janice E Graham, Devon Greyson, Noni E MacDonald, and Samantha B Meyer. 2022. Unmasking the Twitter Discourses on Masks During the COVID-19 Pandemic: User Cluster-Based BERT Topic Modeling Approach. *JMIR Infodemiology* 2, 2 (2022), e41198. doi:10.2196/41198