Noise in Brazilian Clinical Anamnesis: An Empirical Study

Leandro A. Carvalho[1], Thiago Q. Oliveira[2], Flávio R. C. Sousa[1], João B. F. Filho[1]

[1]MDCC - Universidade Federal do Ceará (UFC)
Campus do Pici – Bloco 952 – CEP 60.440-900 – Fortaleza – CE – Brazil

[2]Instituto Federal do Ceará – Fortaleza, CE – Brazil

leandroalmeida@alu.ufc.br, thiago.queiroz@ifce.edu.br,

flaviosousa@ufc.br, bosco@dc.ufc.br

Abstract. Research Context: *The lack of representative data can limit the development of robust clinical Natural Language Processing (NLP) models, as models trained on idealized data can perform poorly on noisy real-world Electronic Health Records (EHRs).* ***Scientific and/or Practical Problem:*** *A performance gap exists when these NLP models are deployed on noisy, real-world clinical text. This issue can be found in less-resourced languages, such as Brazilian Portuguese, where the scarcity of data can limit the development of effective clinical information systems.* ***Proposed Solution and/or Analysis:*** *This study addresses this challenge by presenting a systematic approach to identify and quantify textual noise patterns found in Brazilian Portuguese clinical narratives.* ***Related IS Theory:*** *Based in Task-Technology Fit (TTF) Theory, this study investigates the misalignment between the task of reliable information extraction from noisy EHRs and the technology of NLP models, which can presuppose clean data.* ***Research Method:*** *A multi-stage methodology was employed to identify textual noise. Starting with a classification stage to flag candidate tokens likely representing typos and abbreviations, followed by a lexicon-based validation executed to refine this selection, ensuring that only authentic noise instances were selected.* ***Summary of Results:*** *The analysis of a dataset of clinical anamneses revealed not only a high incidence of textual noise, but also a consistent recurrence of specific noisy tokens across the dataset, demonstrating the widespread nature of data quality issues in this domain.* ***Contributions and Impact to IS area:*** *A taxonomy of textual noise, complemented by two JSON files that structurally map the noisy tokens, establishing an empirical benchmark for Brazilian Portuguese clinical text and formalizing the data quality challenges that must be overcome for successful NLP implementation.*

1. Introduction

Natural Language Processing (NLP) in healthcare has potential to improve medical technologies by possibly revealing significant clinical information hidden in unstructured Electronic Health Records (EHRs) [Juhn and Liu 2020]. By enabling large-scale analysis of data from sources such as clinical notes and EHRs, NLP technologies can offer opportunities to increase diagnostic accuracy and accelerate medical research [Cai et al. 2016]. On the other hand, a significant language gap in academic research and the widespread "noise" that can be found in clinical data [Liu et al. 2012], can still be a challenge that needs to be addressed to increase the potential of NLP use in medical researches.

In this context, this "noise" refers to the range of lexical and syntactic irregularities that deviate from standardized language. This textual noise is a defining characteristic of clinical documentation and manifests in various forms, including typographical errors and non-standard abbreviations. For NLP models, this "noise" presents challenges to accurately extract meaningful clinical information from the anamnesis dataset.

Initially, the field of clinical NLP was marked by a significant linguistic disparity. The overwhelming majority of benchmark datasets and pretrained models are concentrated on high-resource languages, primarily English [Hasan et al. 2024]. This linguistic asymmetry can create a significant resource disparity for other major world languages, including Brazilian Portuguese, which is spoken by over 200 million people. There is a lack of extensible and accessible resources related to clinical data in languages other than English [Névéol et al. 2018] that can cause issues for scientific progress and affect the creation of healthcare technologies related to NLP tools designed to medical area.

Second, aside from the language barrier, a widespread yet often overlooked issue is the "clean world" assumption that forms the basis for creating and training NLP models [Shickel et al. 2017]. These models are typically trained and evaluated on curated, well-formed texts, which contrasts with the reality of EHRs [Leaman et al. 2015]. Real-world EHRs are usually "noisy" and can present a broad range of deviations from standard grammatical and orthographic conventions. This noise is not completely random, it can comprise systematic patterns of potential abbreviations and potential typos born from the high-pressure, time-constrained environment of clinical practice. The resulting mismatch between clean training data and noisy real-world inputs can be a critical vulnerability, as numerous studies have demonstrated that model performance degrades significantly in the presence of even minor textual perturbations, a brittleness that persists even in modern transformer-based architectures [Moradi and Samwald 2021].

This study contends that before this noise can be mitigated, it must be systematically characterized. While the existence of noise in clinical text is widely acknowledged [Nguyen and Patrick 2016], its specific forms and frequencies in Brazilian Portuguese clinical documentation have not been formally quantified. Such an analysis can provide an empirical foundation for developing targeted automated text-normalization pipelines and more realistic synthetic data augmentation methods. To achieve this goal, the results were demonstrated using two artifacts:

1. **Brazilian Clinical Noise Patterns Dataset:** 2 datasets of textual noise patterns, `typos.json`[1] and `abbreviations.json`[2], systematically identified and cataloged from a de-identified source corpus of 8,411 clinical anamnesis notes. This high-quality, structured resource was designed to be a foundational benchmark to support research on noise-aware model development, addressing a critical prerequisite for advancing NLP in medical context.
2. **A Taxonomy of Clinical Textual Noise:** A detailed quantitative and qualitative taxonomy of the noise patterns found is presented, establishing possibly the first empirical benchmark for documentation errors and shortcuts in this domain for

[1] `https://github.com/sbsi2026-paper/SBSI-2026/blob/main/potential_typos.json`

[2] `https://github.com/sbsi2026-paper/SBSI-2026/blob/main/potential_abbreviations.json`

Brazilian Portuguese.

The contributions presented in this research may provide an essential foundation for advancing the development of robust, equitable, and clinically effective NLP technologies in Brazilian Portuguese.

2. Related Work

2.1. NLP for Healthcare Documentation

The maturation of clinical NLP as a field has been intrinsically linked to the availability of large-scale de-identified datasets. In the English-speaking world, foundational resources such as the MIMIC-III and MIMIC-IV databases have become the standard [Johnson et al. 2016, Johnson et al. 2023]. Another great example of these studies is i2b2-2010, which contains patient reports annotated with various types of relations linking medical problems to treatment entities [Uzuner et al. 2011]. By providing vast quantities of data for training and evaluation, these databases have catalyzed the development of sophisticated models and driven innovation across a range of clinical tasks [Johnson et al. 2018].

On the other hand, this progress can be linguistically uneven. Since NLP research remains highly concentrated in English [Lopes et al. 2019], a lack of resources for other languages, such as Brazilian Portuguese, can be noticed [Pereira 2021], and this inconsistency can present a challenge, as the lack of domain-specific data for these languages can delay scientific advancement. Due to this, the automation of clinical tasks via NLP has proven less effective for non-English languages, largely because of these foundational data deficits [Crema et al. 2022].

To address these gaps found in Brazilian Portuguese, researchers have been developing valuable, specialized works that includes SemClinBr, a semantically annotated corpus of 1,000 multi-specialty clinical notes [Oliveira et al. 2022] and also BRAX, a dataset of radiology reports [Reis et al. 2022]. While these resources are crucial, a need for a large-scale corpus of general clinical notes, such as anamneses, is still required, to capture the unfiltered language of routine medical practice.

2.2. Studies on Noise and Non-Standard Language in Clinical Text

Although structured data entry is encouraged, clinical documentation is a unique linguistic domain and a lot of vital information within an EHR is often captured in free-text narratives [Johnson et al. 2008]. This text is not formally structured but is shaped by the need for efficient communication in high-pressure, time-constrained environments.

The linguistic properties of this type of clinical text have been studied, particularly in the English language. A large-scale analysis of Veteran Affairs (VA) clinical notes, for example, revealed a variance in document structure and language use across different notes [Zeng et al. 2011]. This “noise” is not just a simple linguistic curiosity but one of the main causes of performance decrease in NLP models [Sheikhalishahi et al. 2019]. The fragility of high-performing systems is well documented, their accuracy often decreases when inputs contain even minor deviations from the clean text on which they were trained [Smith 2025].

The characterization and mitigation of noise are active research areas in clinical AI, which can cover different modalities, from the impact of acoustic noise on AI scribes [Draper et al. 2025] to the effect of label noise on the training of models [Wei et al. 2024]. On the other hand, the primary goal of this study is to categorize and label noises presented in clinical data to help avoid a decrease in NLP performance in models trained using real-world data. While the characteristics of English clinical text have been explored [Zeng-Treitler et al. 2007], a systematic, quantitative taxonomy of noise for Brazilian Portuguese clinical notes remains unaddressed [de Oliveira et al. 2022]. In order to adress this gap, a methodology for noise pattern extraction is proposed in this study, providing an empirical foundation for building noise-robust NLP tools for Brazilian Portuguese language.

3. Dataset Overview

The source data for this study is composed by 8,411 entries in the EHR systems of Brazilian hospitals, the focus was exclusively on the free-text field labeled queixa_principal (chief complaint), which contains the clinical anamnesis. This narrative field can have great clinical value as it documents the patient medical history [Lopes et al. 2019]. However, its unstructured nature presents significant challenges in its analysis.

3.1. Data Composition and Statistics

This study utilized a corpus of secondary data derived from pre-existing clinical text records, representing a diverse range of healthcare facilities across multiple Brazilian states and ensuring a broad and heterogeneous representation of clinical documentation practices. This compositional diversity is crucial for developing models that can perform reliably across different environments. The following section presents the principal statistics of the dataset.

3.1.1. Geographic Representation

A significant challenge in developing NLP models for Brazilian Portuguese is accounting for the country's vast geographic expanse and sociocultural heterogeneity. These regional distinctions often manifest in linguistic variations, including distinct dialects and terminologies. A model trained on data from a single region may therefore present regional bias and fail to generalize effectively when deployed in a different region. Consequently, a geographically balanced dataset is important for this study.

As illustrated in Figure 1, the source dataset is composed for entries from different states in Brazil. While not yet perfectly balanced, this composition ensures a degree of geographic diversity, enhancing the corpus's external validity and serving as a foundation for building more robust and equitable models.

3.1.2. Distribution of High-Urgency Records

The context of clinical urgency is a critical factor that influences the style and structure of medical documentation. Although this information was not annotated across the entire corpus, Figure 2 confirms that a portion of the records originated from emergency

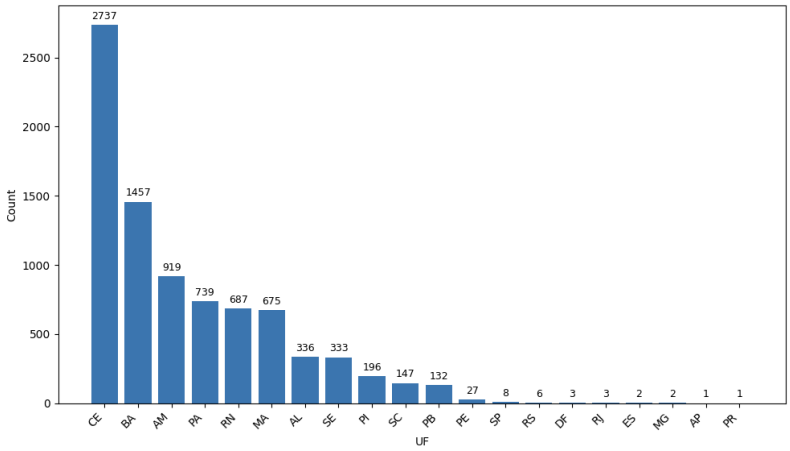

Figure 1. Geographic Distribution of the Dataset by Federative Unit (UF)

situations, wich N/A represents the records without any information related to this. This component is methodologically significant for two reasons: first, it ensures the dataset captures the unique noise patterns and linguistic shortcuts common in high-urgency scenarios. And second, it provides a crucial resource for developing and evaluating NLP models designed to be robust enough for deployment in the most challenging clinical environments.

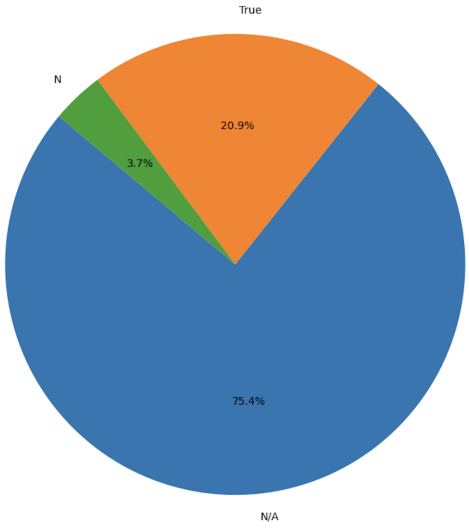

Figure 2. Distribution by Emergency Status

3.1.3. CID Code Distribution

Another critical dimension for dataset heterogeneity is the diversity of clinical diagnoses, as codified by the International Classification of Diseases (CID). The patient's underlying condition profoundly influences the vocabulary, syntax, and overall structure of the anamnese. For instance, documentation related to cardiology will employ a distinct lexicon and format compared to notes from oncology. Therefore, a broad representation of CID codes

is essential for developing robust NLP models that can generalize across different medical specialties.

Figure 3 illustrates the frequency distribution of these codes within the corpus, highlighting the 20 most prevalent classifications. The chart demonstrates a wide variety of diagnoses, while the "Others" category consolidates the long tail of less frequent CIDs, underscoring the dataset's breadth.

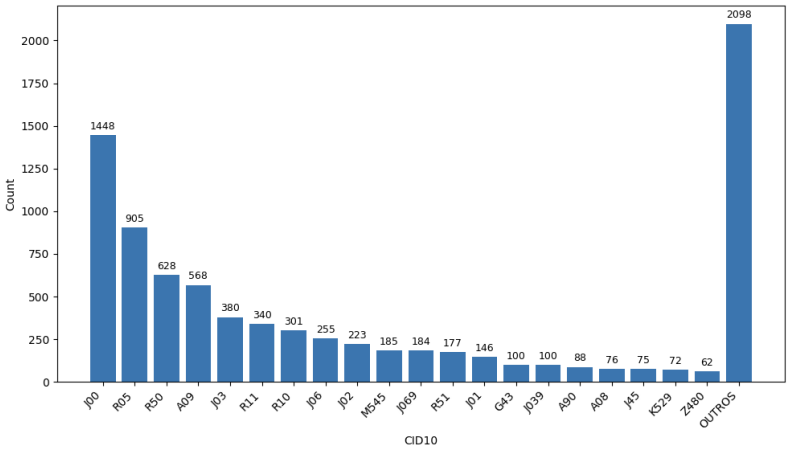

Figure 3. Frequency Distribution of the 20 Most Prevalent CIDs

To provide clinical context, Figure 4 offers the corresponding definitions for these common codes, giving insight into the primary medical domains represented.

CID10	Definição	Contagem
J00	Nasofaringite aguda (resfriado comum)	1448
R05	Tosse	905
R50	Febre de origem desconhecida	628
A09	Diarreia e gastroenterite de origem infecciosa presumível	568
J03	Amigdalite aguda	380
R11	Náusea e vômito	340
R10	Dor abdominal e pélvica	301
J06	Infecções agudas das vias aéreas superiores de localizações múltiplas e não especificadas	255
J02	Faringite aguda	223
M545	Lombalgia	185
J069	Infecção aguda das vias aéreas superiores, não especificada	184
R51	Cefaleia	177
J01	Sinusite aguda	146
G43	Enxaqueca	100
J039	Amigdalite aguda, não especificada	100
A90	Dengue (clássica)	88
A08	Infecções virais intestinais, não especificadas	76
J45	Asma	75
K529	Gastroenterite e colite não-infecciosas, não especificadas	72
Z480	Cuidados a curativos e suturas cirúrgicas	62

Figure 4. Definitions of the Most Prevalent CIDs Identified in the Dataset

3.2. Ethical and Legal Framework

Brazil's legal requirements for healthcare area research ensure patient privacy by operating under Brazil's General Data Protection Law (Lei Geral de Proteção de Dados - LGPD), Law No. 13.709/2018 [Presidência da República 2024], providing a clear legal basis for the use of health data in academic researches. In order to be in compliance with it, all data user were anonymized prior to processing, which is defined as data pertaining to a subject who can no longer be identified through reasonable technical means, falling outside the direct scope of the LGPD.

4. Methodology

To create a comprehensive and empirically grounded taxonomy, a three-stage methodology of data analysis was required, combining the scalability of automated analysis with human review, which can be essential for distinguishing intentional clinical shorthand from unintentional errors. The whole process is illustrated on Figure 5.

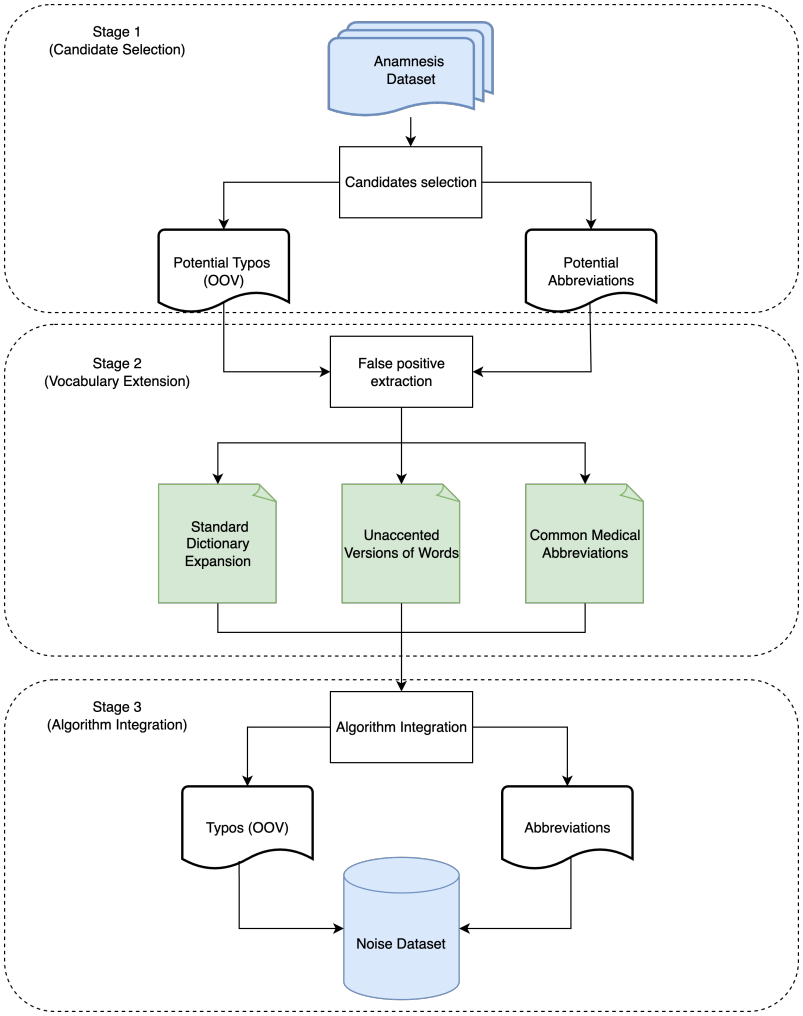

Figure 5. Three-stage Hybrid Methodology for Data Analysis

1. Stage 1 involves an analysis of the entire dataset to generate a comprehensive list of potential noise candidates.
2. Stage 2 represents a refinement of these noise candidates to classify and quantify the definitive noise patterns, extending the tool vocabulary to remove the false positives.
3. Stage 3 consists on integrate the results from stage 2 to refine the process and avoid false positive results.

4.1. Candidate Selection

A Python algorithm, `categorizer.py`[3], was developed to identify and flag tokens that executed an automated analysis of the dataset composed of 8,411 anamnesis in the

[3] `https://github.com/sbsi2026-paper/SBSI-2026/blob/main/categorizer.py`

first stage. The algorithm integrates two primary detection strategies, each targeting a different category of textual noise, allowing for a more robust and comprehensive initial examination of potential errors before the second stage.

A key characteristic of this stage 1 is its high recall but relatively low precision. It successfully captures a large pool of non-standard tokens but does not distinguish between true misspellings and lack of tool resources. Therefore, these preliminary results serve as input for a second, more detailed stage of analysis designed to refine this list and isolate the genuine typographical errors.

4.1.1. Potential Typos

The initial phase of noise characterization involved a lexical analysis to identify potential typographical and orthographic errors. This was accomplished using a dictionary-based, out-of-vocabulary (OOV) detection approach, implemented with the Pyspellchecker library configured for a standard Brazilian Portuguese dictionary [Barrus 2025]. In this stage, each token in the corpus was systematically cross-referenced with the dictionary's lexicon. Tokens that were not found were flagged as candidate errors, providing a foundational layer for identifying words that could compromise the performance of downstream NLP tasks like named-entity recognition or text classification. This stage is effective at capturing a wide spectrum of anomalies beyond simple typos, including:

1. **Genuine Typographical Mistakes**: Common typing errors (e.g., hipogastro instead of hipogástrio).
2. **Orthographic Errors**: Missing diacritics, which is a frequent issue in informally typed Portuguese text (e.g., nauseas vs. náuseas).

Figure 6 illustrate an example of how typos are selected:

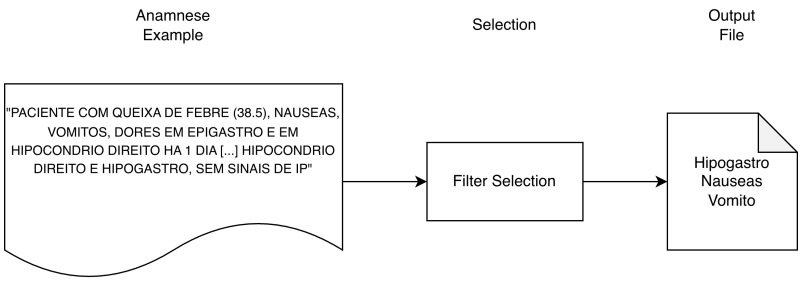

Figure 6. Failed Typo Classification in a Sample Anamnesis

The word hipogastro was flagged as a potential typo, wich is the correct assumption, but Nauseas and Vomitos were also flagged, identifying them as orthographic errors due to the missing diacritics. This classification will be correctly addressed on stage 2.

4.1.2. Potential Abbreviation

A rule-based heuristic using regular expressions was employed to identify potential non-standard abbreviations. Tokens with 1-5 characters that were not present in a standard

Portuguese list were flagged, capturing prevalent clinical shorthand (e.g., pcte for paciente) that is not present in general-purpose dictionaries, as illustrated in Figure 7. This approach favors high recall to ensure comprehensive candidate selection, with the understanding that any false positives (i.e., valid short words) will be filtered during the Stage 2.

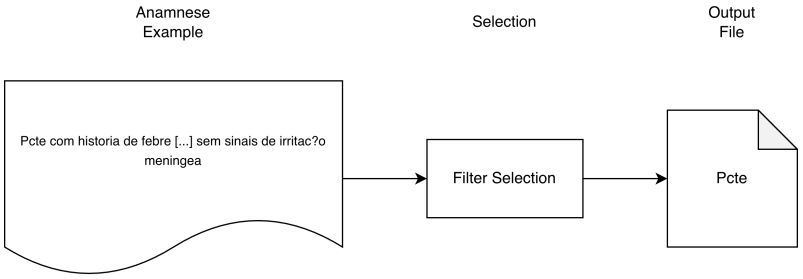

Figure 7. Abbreviation Classification in a Sample Anamnesis

4.2. Vocabulary Extension

One of the challenges in this process was to refine the detection algorithm to make it more accurate, specifically by reducing the high number of false positives correctly spelled and specialized terms that spell-checker incorrectly flagged as errors.

The standard dictionary used by the algorithm was expanded. A curated list of words containing 2,582 common medical terms was created and added to the algorithm, the 2 files containing false positives for typos and abbreviations were combined in one file to increase the vocabulary, as represented in Figure 8. This step effectively increased the algorithm classification method, ensuring that correct terms such as “taquicardia" (tachycardia) were not classified in any of the categories.

Also, the algorithm was adapted to handle the inability to properly render accented characters, words such as “saúde" (health) were often typed as “saude". To prevent these from being flagged as typos, the custom glossary was populated with the unaccented versions of words. This adjustment improved the algorithm’s precision by making it aware of this specific type of systemic noise.

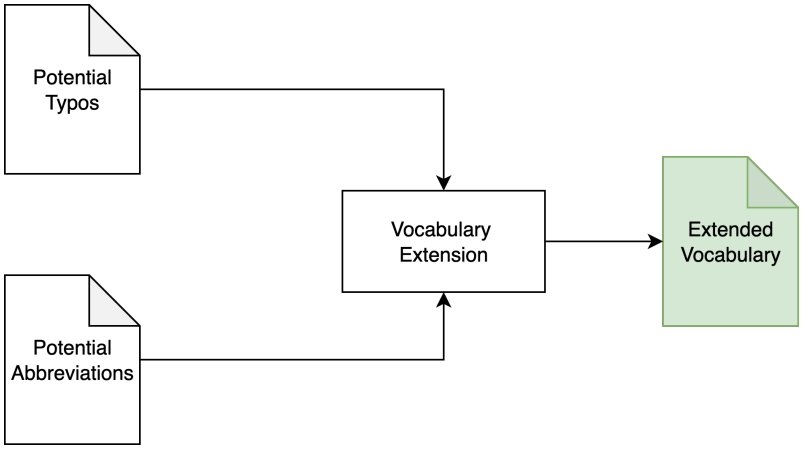

Figure 8. Construction of the Extended Vocabulary from Source Files

Finally, to prevent another common source of false positives, a predefined list of common medical abbreviations (e.g., “otosc" for “Otoscopia") was integrated as an array into the code. Due to this, the algorithm recognizes these abbreviations as valid tokens, avoiding categorizing them as typos and improving the accuracy of the entire noise detection process.

4.3. Algorithm Integration

The final step of this methodology consists of integrating the outputs from the previous stages to create a highly refined analysis engine, ensuring that the final taxonomy was not only accurate but also contextually aware of the specific nuances of clinical texts. To achieve this, the lexical resources created in the second stage were directly incorporated into the algorithm, which was enhanced to a pipeline capable of a much deeper level of linguistic analysis. This refined process ensured that the resulting error taxonomy was robust and genuinely reflected the true error patterns present in the clinical anamneses. Figure 9 illustrates how the algorithm identified the “noise” after the refinement added in the previous stage.

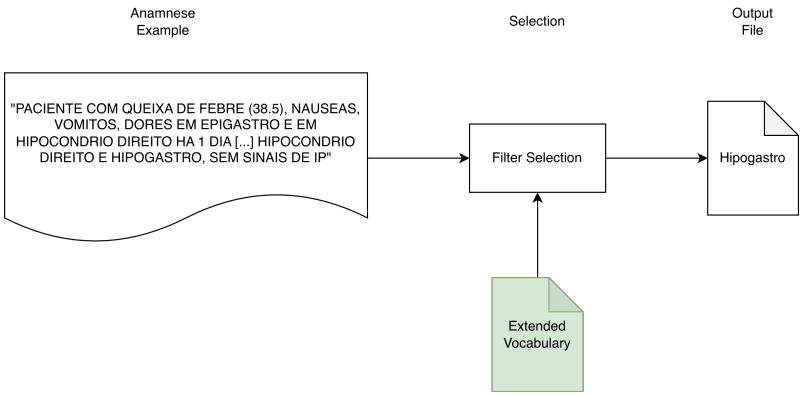

Figure 9. Extended Vocabulary Integration Stage

5. Results

Based on an extensive analysis of 8,411 clinical notes comprising 640,477 tokens, this investigation reveals that textual noise is a pervasive phenomenon in real-world clinical documentation. The findings confirm that lexical and orthographic deviations, such as typos and non-standard abbreviations are not occasional anomalies but rather an intrinsic and systemic feature of this data type, likely reflecting the high-pressure, time-constrained environments in which medical information is recorded.

To systematically characterize these occurrences, the output of the analysis is a dataset with all identified noise. For each unique noisy token, this catalog provides three key data points: the token itself, its total frequency of occurrence (count), and an example that provides the original context in which the token appeared. This methodology produces detailed entries, as illustrated by the entry for the phonetic misspelling "supercifial":

Token: `supercifial`

Frequency: 15

`M?e relata crianca com febre iniciada ha 1 hora atras [...] INDOLOR A PALPAC?O` **`SUPERCIFIAL`** `E PROFUNDA [...] COM RECEITA.`

This context-aware catalog moves beyond simple error detection, providing a valuable empirical resource, essential for the development and evaluation of noise-robust NLP models that are resilient to the idiosyncrasies of authentic clinical text.

The quantitative analysis provides a concrete statistics on the frequency and distribution of different error types, as it is possible to see on Table 1, allowing a clear measurement of the scale of the problem. Also offers a comprehensive contextualized understanding of the textual noise, detailing not only how widespread the issue is but also what it looks like in practice.

Table 1. Summary of Anamnesis Dataset Analysis

Metric	Value
Total Anamnesis Records Processed	8,411
Total Tokens Analyzed	640,477
Total Noisy Tokens Found	69,046
Records with at least one noisy token	7,700
% of Records with Noisy Token	91.55%
Average noisy tokens per record	8.21
Average noisy tokens per noisy record	8.97
Noisy tokens in most affected anamnesis	50
Overall Noise Percentage	10.78%

5.1. Quantitative Analysis of Noise Prevalence

A total of 69,046 tokens were identified as potential noise instances. This high proportion shows the challenge faced by standard NLP models when processing unfiltered clinical text is required.

The noise was not uniformly distributed across the dataset, 7,700 of 8,411 notes (91.55%) contained at least one noisy token, and each clinical note contained 8.21 noise instances on average. On the other hand, while some records contained only a few noise, the most affected record contained 50 noisy tokens, illustrating a wide variance in documentation quality across the dataset.

A detailed breakdown of the identified noise candidates into their respective categories is presented in Table 2.

5.2. Qualitative Taxonomy of Noise Patterns

To provide a deeper understanding of these categories, samples from each category were qualitatively analyzed.

Table 2. Distribution of Noise Types

Noise Category	Token Count	Percentage of Total Noise	Percentage of All Tokens
Abbreviation	62,543	90.58%	9.77%
Typo	6,503	9.42%	1.02%
Total Identified	**69,046**	**100%**	**10.70%**

5.2.1. Abbreviations Noise

Presenting a total of 62,543 occurrences from 1,900 unique tokens, potential abbreviations were the most common noise in the dataset, reflecting a possible clinical requirement for efficient and speedy documentation of the anamnesis written during medical routines. This introduces significant challenges in data interpretation.

Table 3. Top 10 Most Frequent Clinical Abbreviations and Their Meanings

Rank	Abbreviation	Meaning	Frequency	% of Total
1	RA	Ruídos Adventícios	3,403	5.44%
2	BEG	Bom Estado Geral	3,340	5.34%
3	2T	2 Tempos (cardíacos)	3,286	5.25%
4	ABD	Abdome	3,117	4.98%
5	RCR	Ritmo Cardíaco Regular	2,768	4.42%
6	BNF	Bulhas Normofonéticas	2,341	3.74%
7	MV	Murmúrio Vesicular	2,264	3.62%
8	RHA	Ruídos Hidroaéreos	2,260	3.61%
9	AP	Ausculta Pulmonar	2,224	3.55%
10	ACV	Ausculta Cardiovascular	2,199	3.51%

A small set of abbreviations dominates the corpus, highlighting their critical role in routine clinical documentation, as presented in Table 3. The top 10 most frequent potential abbreviations alone account for 43.46% of all tokens identified as abbreviations in the dataset, possibly suggesting that clinical note-taking heavily relies on a core vocabulary of shorthand to efficiently record common, underscoring the importance of accurately identifying these terms for any text analysis.

These frequently used abbreviations primarily refer to standard language used in physical examination, including terms such as “BEG” (Bom Estado Geral), a summary of the patient’s overall appearance and “RCR” (Ritmo Cardíaco Regular), a key finding in a cardiovascular assessment. The prevalence of these specific terms illustrates how abbreviations are fundamentally tied to the anamnesis text.

5.2.2. Typos Noise (OOV)

Potential typos accounted for 6,503 total occurrences across 3,110 unique tokens. These tokens are composed of typographical errors and phonetic substitutions, when words are

spelled as they sound, directly reflecting the time-sensitive nature of data entry, where speed can often be prioritized over gramatically precision. Common typographical errors are prevalent, as seen in “irratacao” (irritation), and “simmetrico” (symmetric). Phonetic substitution, which represents an error where a word is misspelled based on its pronunciation, often involves swapping letters or syllables that sound similar in Brazilian Portuguese, such as “blumbuergue” (Blumberg).

Unlike the highly concentrated usage of standard abbreviations, the distribution of these errors is much more diffuse, representing a long-tail problem, as the top ten of most frequent typos represent only 8.59% of all instances in this category, indicating a wide variety of low-frequency errors.

The analysis of the most common tokens revealed several distinct error patterns, as presented in Table 4.

Table 4. Top 10 Most Frequent Potential Typos (OOV) and Their Corrections

Rank	Token Example	Correct Term	Frequency	% of Total
1	isofotoreagentes	Fotorreagentes	98	1.51%
2	irratacao	Irritação	57	0.88%
3	expansib	Expansibilidade	55	0.85%
4	urgicontinencia	Urge-incontinência	54	0.83%
5	blumbuergue	Blumberg (sinal de)	54	0.83%
6	fascies	Fácies	52	0.80%
7	normo	(Prefixo: normo-)	50	0.77%
8	menigea	Meníngea	48	0.74%
9	orofanrige	Orofaringe	45	0.69%
10	simmetrico	Simétrico	45	0.69%

6. Discussion and Future Work

The primary finding of this study is that textual noise is not a marginal issue but a pervasive characteristic of real-world Brazilian Portuguese anamnesis. The fact that over 91.55% of individual clinical records contained at least one such token demonstrates that this is a widespread, systemic issue and not one confined to a few poorly documented cases. This incredibly high prevalence is a critical discovery of our study, establishing a firm quantitative baseline for future researches, proving that any attempt to process these texts must begin with the assumption that the data will contain textual noises.

These results can change the approach required to build effective clinical NLP tools for Brazilian Portuguese, as standard models trained on clean, formal text are insufficient and likely to fail when deployed to work with real-world data. The baseline established in this study serves as a reference for developers to implement specialized strategies to deal with these types of noise, which may include robust pre-processing pipelines designed to normalize and clean the text, and data augmentation techniques to expose models to similar noise during training.

The high concentration of abbreviations, where the top 10 make up over 40% of their category, exemplifies a pattern widely spread across the anamneses. In contrast, the

long tail of typos, where the top 10 account for less than 9% of their category, shows a different pattern of informal variation.

Targeted text normalization and error correction models informed by these specific patterns could provide substantial improvements in data quality, with each noise category requiring a tailored solution.

1. **Abbreviations** require context-aware models to resolve clinical ambiguity.
2. **Typos** require sophisticated normalization to map varied forms to a single concept.

This study not only confirms the literature's call for more diverse, publicly available datasets for less-resourced languages, but also delivers a tangible solution to this challenge. The availability of this resource, fosters a new research trajectory aimed at creating and delivery noise-robust NLP models, allowing the study field to move beyond theoretical models and start to tackle the complexities of authentic clinical text found specifically within the Brazilian healthcare ecosystem.

6.1. Threats to validity

The validity of this study was reinforced by addressing potential threats based on [Wohlin et al. 2012] validation model. Specific mitigation strategies were implemented to address potential internal, external, construct, and conclusion threats to this study and ensure the reliability of the results presented.

1. **Internal Validity**: A limitation of this study is the potential for selection bias, since the dataset is composed of entries originating from various Brazilian regions but this representation is uneven, with an imbalance across geographic locations. To address this limitation, the analysis was conducted on the complete dataset as a single, ensuring that the conclusions are, at a minimum, representative of the specific sample investigated.
2. **External Validity**: This study was based only on Brazilian Portuguese anamnesis, wich may lead for **Lack of Representativeness**. All results presented are specific to this language, and it is not possible to apply the same conclusion to all clinical environments. While this study provides a rigorous foundational analysis, its external validity can only be established through systematic replication by applying our methodology in different settings to test the cross-contextual validity of the results. Therefore, this study serves as a starting point for a broader comparative research agenda to develop a more comprehensive understanding of clinical documentation errors.
3. **Construct Validity**: The use of a automated spell-checker as a dictionary to classify the tokens in each category is a process that can lead to **Inadequate Definitions**. Words not present in the comparison base could be incorrectly classified and change the taxonomy presented in the results. Also, a large dataset with different structures for each record may lead to unnecessary statistical collection. To prevent both threats, all error categories were previously defined, ensuring that the tokens found during the process were classified correctly in these limited categories, which allowed the research to focus only on the main goal, without inflating the results with non-valuable statistics.
4. **Conclusion Validity**:

While automated text analysis provides scalability and speed, it is also **susceptible to misinterpretations** that can lead to significant statistical errors, where the algorithm can incorrectly flag a token as an error. If not mitigated, it can compromise the results, making them inconsistent or reducing their significance. In order to reduce this threat, the methodology applied a manual step that created a list of false-positive results, refined the algorithm to not classify them in any of the categories, and ensured that only real errors were used to define the taxonomy. This multistage process was designed to refine our automated process, ensuring that our final analysis was based only on a verified set of true errors and preventing the algorithm from inflating the count using false positives.

6.2. Limitations

It is crucial to contextualize the limitations of this study to ensure a balanced interpretation of its findings.

First, the analysis was exclusively based on queixa_principal (chief complaint) notes, that has its own distinct, often abbreviated, linguistic style. Consequently, the typos and abbreviation frequencies identified in this study may not be fully generalizable to other, more descriptive forms of clinical documentation.

The second limitation pertains to the scale of the noise dataset. While it represents a significant contribution to Brazilian Portuguese clinical NLP resources, its size is modest when compared to the massive English-language datasets commonly used in machine learning, which may limit its direct utility for training large language models (LLMs) in another language, the result dataset is likely more valuable for fine-tuning existing Brazilian Portuguese models or for developing and evaluating more specialized, task-specific algorithms. On the other hand, the methodology developed in this study is not confined to Brazilian Portuguese. The framework is designed to be largely language-agnostic, meaning it can be adapted for other languages, provided a correspondent dataset.

6.3. Ethical Implications and Warning Against Clinical Use

The findings of this study are intended to advance academic researches, it must be explicitly stated that the noise dataset is a research dataset and not a clinically validated data. The noise and potential biases inherent in these datasets mean that any model trained on them is suitable only for experimental and academic use. A high degree of textual noise and variability was discovered and quantified in these anamneses, deploying a model trained on such data for actual patient diagnosis or clinical decision-making would be irresponsible and would represent a significant ethical risk of patient harm. Rigorous clinical validation, far beyond the scope of this study, is a prerequisite for any real-world application.

7. Conclusion

This study was designed to support critical challenges related to clinical NLP, healthcare data automation, the pronounced scarcity of curated resources for Brazilian Portuguese, and the complex problem of textual noise inherent in real-world EHRs. This study addressed these gaps directly introducing two primary contributions: a noise dataset, a publicly available dataset containing samples of authentic noise patterns, and an accompany-

ing empirically grounded taxonomy of textual noise, providing a structured classification system for the diverse types of noise found in the dataset.

By systematically identifying, classifying, refining, and quantifying these noises, this study goes from a general acknowledgment of "data quality issues" to provide a qualitative a quantitative data-driven baseline, allowing researchers to understand not only that noise exists but also precisely what kinds of noise are most prevalent and the frequency at which they were present in the dataset.

Noise dataset and noise taxonomy provide an essential foundation for the next generation of clinical NLP applications in Brazilian Portuguese, serving as an agent for innovation in several areas. First, enabling the development and evaluation of noise-robust models, as researchers can train and benchmark algorithms on data that emulate real-world noise. Second, it provides a background for creating more effective preprocessing pipelines tailored specifically to the patterns observed in Brazilian Portuguese anamnesis. Finally, a detailed characterization of noise facilitates the creation of more realistic synthetic data generation techniques, which can help augment the limited datasets available and further accelerate research in this vital domain. All of these contributions, helps the research community with the fundamental tools needed to build more resilient and effective clinical NLP systems for Brazilian Portuguese.

Acknowledgements

Paperpal, an artificial intelligence-based tool, was used to support the writing and review of English grammar. Sourcely was used to support the finding references. Gemini, the AI Assistant of Google, was used to enhance the text quality. The work was carried out in collaboration with a brazilian health company, which contributed providing the dataset used in this study.

References

Barrus, T. (2025). pyspellchecker. `https://pypi.org/project/pyspellchecker/`. Accessed: July 20, 2025.

Cai, T., Giannopoulos, A. A., Yu, S., Kelil, T., Ripley, B., Kumamaru, K. K., Rybicki, F. J., and Mitsouras, D. (2016). Natural language processing technologies in radiology research and clinical applications. *Radiographics*, 36(1):176–191.

Crema, C., Attardi, G., Sartiano, D., and Redolfi, A. (2022). Natural language processing in clinical neuroscience and psychiatry: A review. *Frontiers in Psychiatry*, 13:946387.

de Oliveira, L. F. A., Pagano, A., e Oliveira, L. E. S., and Moro, C. (2022). Challenges in annotating a treebank of clinical narratives in brazilian portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 90–100, Cham. Springer International Publishing.

Draper, T. C., Leake, J., Lamb-Riddell, K., Cox, T., McCormick, J., Trowell, S., Kiely, J., and Luxton, R. (2025). The impact of acoustic and informational noise on ai-generated clinical summaries. *medRxiv*, pages 2025–03.

Hasan, M. A., Tarannum, P., Dey, K., Razzak, I., and Naseem, U. (2024). Do large language models speak all languages equally? a comparative study in low-resource settings. *arXiv preprint arXiv:2408.02237*.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Johnson, A. E., Stone, D. J., Celi, L. A., and Pollard, T. J. (2018). The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39.

Johnson, S. B., Bakken, S., Dine, D., Hyun, S., Mendonça, E., Morrison, F., Bright, T., Van Vleck, T., Wrenn, J., and Stetson, P. (2008). An electronic health record based on structured narrative. *Journal of the American Medical Informatics Association*, 15(1):54–64.

Juhn, Y. and Liu, H. (2020). Artificial intelligence approaches using natural language processing to advance ehr-based clinical research. *Journal of Allergy and Clinical Immunology*, 145(2):463–469.

Leaman, R., Khare, R., and Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57:28–37.

Liu, F., Weng, C., and Yu, H. (2012). Natural language processing, electronic health records, and clinical research. In *Clinical research informatics*, pages 293–310. Springer London.

Lopes, F., Teixeira, C., and Oliveira, H. G. (2019). Contributions to clinical named entity recognition in portuguese. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233.

Moradi, M. and Samwald, M. (2021). Evaluating the robustness of neural language models to input perturbations. *arXiv preprint arXiv:2108.12237*.

Nguyen, H. and Patrick, J. (2016). Text mining in clinical domain: Dealing with noise. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 549–558.

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., and Zweigenbaum, P. (2018). Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.

Oliveira, L. E. S. E., Peters, A. C., Da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Al Hasan, S., and Moro, C. M. C. (2022). Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1):13.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

Presidência da República (2024). Lei nº 14.874 de 28 de maio de 2024. `https://legislacao.presidencia.gov.br/atos?tipo=LEI&numero=14874&ano=2024`. Accessed: July 20, 2025.

Reis, E. P., De Paiva, J. P., Da Silva, M. C., Ribeiro, G. A., Paiva, V. F., Bulgarelli, L., Lee, H. M., Santos, P. V., Brito, V. M., Amaral, L. T., et al. (2022). Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487.

Sheikhalishahi, S., Miotto, R., and Dudley, J. T. (2019). Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Medical Informatics*, 8(2):e12239.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. (2017). Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604.

Smith, W. (2025). *Applied Deep Learning for Natural Language Processing with AllenNLP: The Complete Guide for Developers and Engineers*. HiTeX Press.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Wei, Y., Deng, Y., Sun, C., Lin, M., Jiang, H., and Peng, Y. (2024). Deep learning with noisy labels in medical prediction problems: a scoping review. *Journal of the American Medical Informatics Association*, 31(7):1596–1607.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., Wesslén, A., et al. (2012). *Experimentation in software engineering*, volume 236. Springer.

Zeng, Q. T., Redd, D., Divita, G., Jarad, S., Brandt, C., and Nebeker, J. R. (2011). Characterizing clinical text and sublanguage: A case study of the va clinical notes. *J Health Med Informat S*, 3(2).

Zeng-Treitler, Q., Kim, H., Goryachev, S., Keselman, A., Slaughter, L., and Smith, C.-A. (2007). Text characteristics of clinical reports and their implications for the readability of personal health records. *Studies in health technology and informatics*, 129(2):1117.