Mining Discursive and Interactional Signals in Public Teleconsultation Platforms: A Deep Learning Approach to Referral Conduct Analysis

Diego S. N. dos Santos[1], Alvaro B. Cunha[1], Gustavo Penna[2], Frederico Coelho[3], Carlos H. G. Ferreira[4], Luiz Carlos B. Torres[4]

[1]Department of Computing and Systems - Universidade Federal de Ouro Preto
[2]Department of Internal Medicine - Universidade Federal de Minas Gerais
[3]Department of Electrical Engineering - Universidade Federal de Minas Gerais
[4]Department of Computing - Universidade Federal de Ouro Preto

{diego.nere, alvaro.braz}@aluno.ufop.br
gustavocpenna@gmail.com, fredgfc@ufmg.br
chferreira@ufop.edu.br, luiz.torres@ufop.edu.br

Abstract. *Digital platforms used in public healthcare systems generate large volumes of interactional and discursive data. In the Brazilian Unified Health System (SUS), teleconsultation platforms mediate communication between primary care professionals and specialists, producing textual dialogues that encode referral requests, diagnostic uncertainty, collaborative decision-making, and operational patterns of care coordination. This paper investigates whether such digital and discursive signals can be mined to classify patient conduct as Retention in primary care or Referral to specialized care. Using 40,506 de-identified teleconsultation records from a Brazilian telehealth center, we evaluate Transformer-based models adapted to Portuguese clinical text, combining BioBERTpt, Head+Tail truncation, Focal Loss, and validation-based threshold selection. The final model achieved a Macro F1-Score of 0.61 and an accuracy of 71.2%, preserving predictive performance while shifting the validation-optimized threshold closer to the default probability boundary. Beyond aggregate performance, we conduct a post-hoc analysis across medical specialties, revealing statistically significant heterogeneity in model errors. These findings suggest that interprofessional teleconsultation logs can serve as a valuable source of digital and discursive signals for operational public health analysis, while also highlighting the need for subgroup evaluation, calibration-specific metrics, and external validation before any prospective deployment.*

1. Introduction

Societal digitalization generates vast datasets from online interactions and socio-technical systems. In public health, these data reveal patterns in behavior and discourse that reflect individual risks and systemic care coordination, resource distribution, and access organization [Scherbakov et al. 2025, Molenaar et al. 2024]. This has motivated computational approaches for mining digital and social signals from heterogeneous environments, including platforms where health-related decisions are negotiated through text.

Within the context of public healthcare, digital communication platforms linking primary care professionals to specialists represent a critical socio-technical ecosystem.

Although often viewed as administrative routing tools, these platforms mediate daily interprofessional interactions in which clinical doubts are formulated, diagnostic uncertainty is expressed, and referral decisions are negotiated. The unstructured text generated in these environments therefore goes beyond symptom reporting: it captures discursive and interactional traces of care coordination, including requests for specialist support, justifications for retention, and differences in decision patterns across medical specialties. For this reason, interprofessional teleconsultation dialogues constitute a rich source of digital signals for operational public health analysis.

The application of Natural Language Processing (NLP) to extract health-related signals from unstructured digital interactions has gained significant traction. Prior studies have used NLP to identify signals in social media, unstructured feedback, telemedicine narratives, and patient-provider communications [Feizollah et al. 2025, Andrade et al. 2024, Raff et al. 2024, Cho et al. 2024, Cunha Reis 2025]. However, less attention has been given to interprofessional communication platforms in public healthcare systems, where the relevant signals are symptoms or diagnoses, and also discursive markers of uncertainty, referral negotiation, and operational demand for specialized care.

Despite these advancements, mining discursive and interactional signals from interprofessional teleconsultation platforms presents unique challenges. The language is hybrid, combining formal medical terminology, abbreviations, local documentation practices, and conversational exchanges. Moreover, the target conduct is naturally imbalanced: most cases are retained in primary care, while a smaller proportion requires referral to specialized services. This imbalance is not merely a statistical artifact, but a reflection of the operational structure of public health systems, where referral decisions directly affect specialist workload and access to care.

To address this gap, this work is characterized as a retrospective development and internal validation study. The primary goal is to evaluate whether completed dialogues can be mined as digital traces of interprofessional decision-making to support operational public health analysis. We evaluate whether domain-adapted models combined with Focal Loss [Lin et al. 2020] can support the classification of these outcomes. By treating logs as traces of care coordination, this study provides insights for specialty demand monitoring and subgroup auditing.

2. Dataset and Exploratory Analysis

For this study, the dataset comprises 40,506 records extracted from the Telehealth Center of the School of Medicine of UFMG (Nutel/FM UFMG) teleconsulting system, including structured metadata and unstructured textual dialogues. These records represent interactions mediated by a public digital health platform, where primary care professionals and specialists exchange information and define conduct. Each individual record consist of isolated messages rather than complete cases. We reconstructed the interactional unit of analysis by grouping records by conversation ID and concatenating messages into cohesive dialogues. Consequently, the primary features extracted for our NLP pipeline include the medical specialty (*ESPECIALIDADE*), the International Classification of Diseases or Primary Care codes (*CID/CIAP*), the patient's brief clinical history (*HISTORIA*), and the final concatenated message thread exchanged between the primary care professional and the specialist (*txt*). In this framing, the message thread is a clinical narrative and also

a discursive record of how uncertainty, referral requests, and specialist recommendations are expressed within the platform. To preserve patient confidentiality and adhere to ethical guidelines, the data were anonymized in accordance with the Brazilian General Personal Data Protection Law (LGPD, Law No. 13,709/2018) [Brasil 2018]. Consequently, the raw dataset provided had all Personally Identifiable Information (PII), such as the names of patients and doctors, already masked with a [NOME] tag.

2.1. Digital, Discursive, and Interactional Signals

In this study, we define digital signals as structured and unstructured traces generated through the teleconsultation platform. Structured signals include specialty, diagnostic coding, and the final conduct label. Discursive signals include the textual formulation of clinical doubts, expressions of uncertainty, referral requests, and specialist recommendations. Interactional signals refer to the organization of the dialogue itself, including the sequence of question and answer, the presence of concluding recommendations, and the way decision-making is recorded across professional roles. This framing allows the task to be interpreted as text classification and a computational analysis of how referral conduct is represented in a public digital health ecosystem.

Table 1. Example of a chat record from the dataset. Only the features relevant to the model's input context are displayed. Original text is kept in Portuguese.

Feature	Content
Specialty	General Pulmonology (*Pneumologia Geral*)
Patient Demographics	Male, born in 1948
Diagnostic Code	CIAP: R99 (Respiratory Diseases)
Conduct	Referral
Clinical History (*HISTORIA*)	*"Paciente trabalhador rural, analfabeto, hipertenso. Em uso de captopril [...] Hoje comparece à consulta sem queixas, trazendo resultado da TC de Tx realizada dia 07/10/2021. [...]"*
Message (*txt*)	**Question (Primary Care):** *"Qual o possível diagnóstico do paciente? Devo evoluir propedêutica? Encaminho à referência de pneumologia?"* **Answer (Specialist):** *"Olá [NOME], precisa que vc encaminhe ao pneumologista, para que verifique a evolução das imagens. Você também pode solicitar a espirometria, para a HD de DPOC. Att."*

It is important to emphasize that, in this retrospective setting, the complete message thread may include the specialist's final recommendation. Therefore, the classification task should be interpreted as retrospective conduct classification from completed teleconsultation dialogues, not as prospective triage before specialist response. This distinction is relevant for public health applications: the proposed pipeline is primarily intended to support operational analysis, auditing of referral patterns, and identification of specialty-level heterogeneity. Future work should evaluate pre-response prediction using only the initial primary care request and patient history.

In its original form, the dataset categorized referrals into two distinct tiers: secondary care and tertiary care. For this study, these two tiers were merged into a single "Referral" class (Class 1), representing cases whose completed teleconsultation record indicates the need for specialized care outside the primary care unit. Cases managed locally were grouped as "Retention" (Class 0). Thus, the binary label captures an operational conduct outcome relevant to care coordination and specialist demand, rather than

a direct diagnosis of disease severity. After this aggregation and initial data cleaning removing administrative questions and non-clinical interactions, the refined dataset exposed a significant natural class imbalance. Approximately 73.5% of the records resulted in Retention, while only 26.5% required Referral. This distribution reflects the standard triage in public health, where the vast majority of cases are resolved at the primary care level.

Furthermore, exploratory analysis revealed that this class imbalance is not uniform across medical specialties. As illustrated in Figure 1, the proportion of referrals varies substantially depending on the specialty. This variation is itself an operational signal: it may reflect differences in specialty demand, availability of procedures in primary care, documentation practices, and the degree of uncertainty involved in each domain. For example, specialties that depend on procedures or visual assessment may exhibit different referral patterns from areas in which cases are more frequently resolved through guidance to primary care. This heterogeneity motivates both the use of imbalance-aware learning strategies and the post-hoc subgroup analysis presented later in this paper.

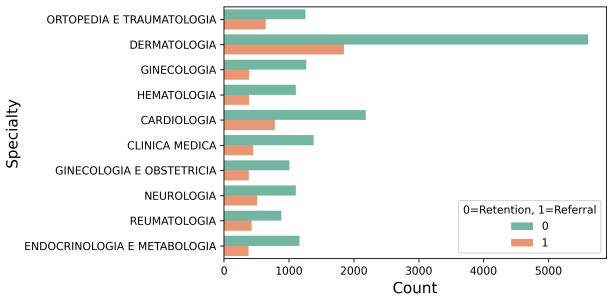

Figure 1. Distribution of clinical conduct (Retention vs. Referral) across the top 10 most frequent medical specialties.

3. Methodology and Pre-processing

The foundation of an NLP implementation relies heavily on the quality, structure, and integrity of the data representation. To model completed teleconsultation dialogues as digital traces of interprofessional decision-making, we implemented dataset formatting and token engineering procedures designed to preserve both structured operational context and unstructured discursive content.

3.1. Text Pre-processing

The raw textual data obtained from the communication database contains numerous artifacts that introduce unwanted noise, potentially degrading the performance of contextual language models. Pre-processing included regex-based URL removal to prevent bias from external hyperlinks. Additionally, clinical records often contain platform-specific artifacts and systemic metadata intertwined with the physician's narrative, which are stripped to normalize spacing. Crucially, while the raw dataset was supplied pre-anonymized our pre-processing pipeline explicitly removes these tags. This step optimizes the model's limited context window, preventing the attention mechanism from wasting valuable token space on repetitive, non-informative markers and ensuring that the sequence length is dedicated to clinically and interactionally meaningful content.

3.2. Input Architecture

To provide the Transformer model with both operational and discursive context, we separated the inputs into structured and unstructured components. The structured fields include the medical specialty queried (‘ESPECIALIDADE‘) and the International Classification of Diseases or Primary Care code (‘CID/CIAP‘), which provide signals about the service domain and clinical area involved. The unstructured fields include the antecedent patient background (‘HISTORIA‘) and the textual dialogue between the primary care unit and the specialist (‘MENSAGEM‘), which encode uncertainty, referral requests, and specialist recommendations.

Instead of independently embedding these features, we linearly concatenated ‘ESPECIALIDADE‘, ‘CID/CIAP‘, and ‘HISTORIA‘ into a single sequence representing the structured-operational context. This sequence was then joined with ‘MENSAGEM‘ using the native ‘[SEP]‘ token defined within the BERT tokenizer structure. Inputs were formatted as ‘[CLS] Context [SEP] Message [SEP]‘, allowing the model to jointly process specialty metadata, patient background, and the discursive exchange recorded in the platform.

3.3. The Head+Tail Truncation Strategy

Processing raw teleconsultation dialogues with BERT architectures is constrained by a 512-subword token limit, which is frequently exceeded by complex interprofessional threads. Traditional ”Head-Only” truncation typically discards the end of the sequence, where the interaction often contains the final recommendation or operational conduct. To preserve both the initial formulation of the clinical doubt and the concluding segment of the dialogue, we implemented a Dynamic Head+Tail Truncation Strategy.

Unlike static truncation methods, our pipeline first calculates a dynamic token budget. We prioritize the encoding of the structural context, including the medical specialty, diagnostic codes, and patient history (HISTORIA), ensuring these fundamental features are preserved in their entirety. The remaining token capacity is then allocated to the message thread (MENSAGEM).

If the message length exceeds this calculated budget, we slice the token list using a 20/80 ratio. We extract the first 20% of the available budget as the ”Head” to capture the initial formulation of the doubt and the final 80% as the ”Tail” to preserve the concluding segment of the interaction. These subsets are concatenated into a single sequence, formatted as [CLS] Context [SEP] Head + Tail [SEP]. This strategy is appropriate for retrospective conduct classification because completed teleconsultation threads often encode relevant decision signals near the end of the exchange. However, it should not be interpreted as a prospective triage design, since future pre-response prediction would require excluding the specialist’s final answer from the input.

4. Experimental Setup

Training a model to classify operational conduct from interprofessional teleconsultation records requires careful experimental design. Due to the hybrid nature of these dialogues, which combine medical terminology, platform-specific language, and conversational decision-making, traditional lexical classifiers may fail to capture the necessary contextual signals. Therefore, this study adopted Transformer-based architectures. The

models were trained for 5 epochs using the AdamW optimizer. Experiments were conducted on a workstation equipped with an AMD Ryzen 9 9950X CPU, and an NVIDIA GeForce RTX 5080 GPU. The following sections describe the transition from a general Portuguese baseline model (BERTimbau) to a domain-adapted model for Portuguese biomedical and clinical text.

4.1. Dataset Splitting

The dataset was partitioned into training (80%), validation (10%), and testing (10%) subsets. To ensure that the subsets reflected the observed distribution of operational conduct, we applied a stratified splitting strategy based on the target class. This preserved the natural imbalance between Retention and Referral across training, validation, and test partitions. Stratification is important because artificial changes in class prevalence could distort threshold selection and the interpretation of referral detection performance.

4.2. Domain Adaptation for Portuguese Clinical and Discursive Signals

The baseline variation of our work attempted to solve the medical retrospective conduct classification using BERTimbau (NeuralMind) [Souza et al. 2020]. As an encoder-only model pre-trained on open domains like Wikipedia and news streams, BERTimbau provides strong representations of Portuguese syntax. However, clinical text is not merely Portuguese; it is a heavily hybridized dialect interwoven with Latin derivatives, pharmacology codes, and highly ambiguous abbreviations.

To adapt the representation space to the target domain, we replaced the underlying weights with **BioBERTpt** (*pucpr/biobertpt-all*) [Schneider et al. 2020]. Because BioBERTpt is adapted to Portuguese biomedical and clinical language, it is expected to better represent abbreviations, disease names, medications, and documentation patterns commonly found in teleconsultation records. In this study, the motivation for domain adaptation is a diagnostic terminology and also the need to represent the hybrid language of public teleconsultation platforms, where clinical descriptions and interactional conduct signals appear in the same textual sequence.

4.3. Loss Function Optimization: Focal Loss for Class Imbalance

The distribution of the target conduct is imbalanced, with approximately 73% of cases labeled as Retention and 27% as Referral. This imbalance reflects the operational role of teleconsultation in public healthcare: most cases are expected to be resolved or managed in primary care, while a smaller fraction requires specialized intervention. In preliminary experiments, we used a native ‘WeightedRandomSampler‘ in the PyTorch ‘DataLoader‘ to increase exposure to the minority class. However, this strategy repeatedly presented the same minority instances during training, increasing the risk of overfitting and reducing the diversity of observed interactional contexts.

Standard Cross-Entropy (CE) loss is inherently biased toward majority classes. As a third improvement, we therefore removed the sampling strategy and replaced the core optimization criterion with the *Focal Loss* [Lin et al. 2020], which reshapes how errors are penalized during backpropagation. Specifically, Focal Loss introduces two modulating factors, Alpha (α) and Gamma (γ), into the cross-entropy formulation, yielding $FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t)$, where p_t denotes the predicted probability assigned to the true class.

In our implementation, we set the class weighting parameter to $\alpha = 0.73$ for the minority Referral class, corresponding to the complement of its empirical prevalence. This increases the contribution of referral misclassifications to the training objective without requiring explicit duplication of minority samples. Additionally, we use $\gamma = 2.0$ to down-weight easy examples as the model becomes more confident, thereby emphasizing harder and more ambiguous cases. In practice, this shifts the optimization focus toward borderline instances where the distinction between Retention and Referral is less clear, allowing us to address class imbalance directly within the loss function without relying on oversampling.

4.4. Threshold Selection and Ablation Strategy

Continuous model scores must be converted into discrete binary classes using a decision threshold. Rather than optimizing only recall for the Referral class, which could inflate false-positive referrals, we used the *Macro F1-Score* as the validation objective. The optimal threshold was selected exclusively on the validation subset and then applied once to the unseen test set. All reported metrics were calculated with 95% confidence intervals obtained through a bootstrap procedure on the test set. This procedure evaluates the stability of aggregate performance, although calibration-specific metrics such as Brier score or Expected Calibration Error should be included in future work to assess probability calibration directly.

To isolate the contribution of each architectural decision, we designed an ablation study. By systematically enabling and disabling domain adaptation, Focal Loss, and contextual data inputs (Specialty, ICD/CIAP, and Patient History), we evaluate how each component affects aggregate predictive performance and the validation-selected decision threshold. In this paper, threshold movement is interpreted as evidence of changes in score distribution and decision behavior, not as definitive proof of probability calibration.

5. Results and Performance Analysis

The quantitative measurements evaluate both raw predictive accuracy and Macro F1-Score under validation-based threshold selection. Rather than focusing only on absolute metric gains, the analysis examines how different modeling choices affect referral detection, score distribution, and the stability of decision thresholds in an imbalanced operational setting.

5.1. Ablation Study and Decision Threshold Behavior

To evaluate the impact of our proposed framework, we analyzed four distinct configurations, observing their Accuracy, Macro F1-Score, and the optimal Decision Threshold (t) required to balance the predictions. The results are summarized in Table 2.

Table 2. Ablation Study on Framework Components with 95% Confidence Intervals.

Setup	Architecture	Focal Loss	Clinical Context	Accuracy	F1-Macro	Threshold (t)
1. Baseline	BERTimbau	No	No	70.6% ($\pm$1.4)	0.597 ($\pm$0.017)	0.38
2. Domain	BioBERTpt	No	No	71.2% ($\pm$1.3)	0.610 ($\pm$0.017)	0.34
3. Loss Only	BioBERTpt	Yes	No	70.1% ($\pm$1.3)	0.600 ($\pm$0.016)	**0.52**
4. Final	BioBERTpt	Yes	Yes	71.2% ($\pm$1.3)	0.610 ($\pm$0.016)	**0.56**

The results, summarized in Table 2, show that Accuracy and Macro F1-Score remain similar across the four configurations, with overlapping 95% confidence intervals. Therefore, the proposed architectural refinements should not be interpreted as producing a statistically robust improvement in aggregate predictive performance. Instead, the most visible difference appears in the validation-selected decision threshold, suggesting that the loss function and contextual inputs affect the distribution of model scores and the operating point required to balance Retention and Referral classification.

The transition from the general BERTimbau baseline to the domain-adapted BioBERTpt model (Setup 2) resulted in a nominal Macro F1 increase to 0.610, but this difference remains within the uncertainty interval. Both configurations trained with standard Cross-Entropy (Setups 1 and 2) required thresholds below 0.40 to optimize Macro F1, indicating that the default 0.50 threshold would not be appropriate for this imbalanced conduct classification task. From an operational perspective, this reinforces the importance of explicit threshold selection when using model scores to study referral behavior.

The use of Focal Loss changed the operating behavior of the model more clearly than it changed aggregate performance. In Setups 3 and 4, the validation-selected threshold shifted to 0.52 and 0.56, respectively, while Macro F1 remained comparable to the best Cross-Entropy configuration. This suggests that Focal Loss altered the score distribution in a way that reduced the need for very low thresholds when optimizing Macro F1. The addition of structured clinical context in Setup 4 did not produce a measurable gain in Accuracy or Macro F1, but it preserved performance while allowing the model to incorporate specialty, coding, and patient-history signals relevant to operational analysis.

5.2. Operational Interpretation of the Confusion Matrix

To interpret the operational consequences of the final model's predictions, we map the decisions onto a confusion matrix (Figure 2). In this setting, false positives correspond to cases classified as Referral despite being labeled as Retention, potentially overestimating specialist demand. False negatives correspond to cases labeled as Referral but classified as Retention, which is more critical from a care-coordination perspective because it may underestimate demand for specialized services. Therefore, the confusion matrix should be interpreted as a predictive evaluation as well as as an operational lens for understanding the trade-off between specialist workload estimation and referral detection.

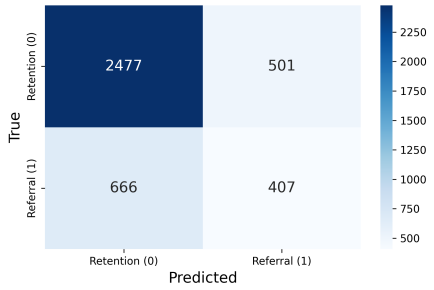

Figure 2. Confusion matrix illustrating the predictive performance of the final model under validation-selected thresholding.

6. Post-Hoc Analysis of Specialty-Level Heterogeneity

Mining digital signals from public teleconsultation platforms requires understanding whether model behavior is consistent across medical specialties. Aggregate performance

may hide subgroup-specific failures, especially when specialties differ in referral prevalence, documentation style, clinical ambiguity, and volume of examples. To evaluate this issue, we conducted a post-hoc analysis of error distributions across medical specialties in the test set.

Because specialties have uneven sample sizes and potentially different error distributions, we used both Analysis of Variance (ANOVA) and the non-parametric Kruskal-Wallis H-test as exploratory statistical tests. The goal was to evaluate whether the model's error magnitude differed significantly across specialty groups. In this analysis, the residual error is treated as a continuous discrepancy between the model score and the ground-truth label, allowing us to examine whether some specialties systematically produce more uncertain or inaccurate predictions than others.

The post-hoc tests on the final model returned a Kruskal-Wallis p-value of 0.002 and an ANOVA p-value of 0.001. Because $p < 0.05$, we reject the null hypothesis of equal error distributions across specialties. This result indicates statistically significant specialty-level heterogeneity in model behavior, suggesting that aggregate metrics alone are insufficient to characterize performance in this teleconsultation setting.

This finding should be interpreted cautiously. Specialty-level heterogeneity may reflect multiple factors, including differences in referral prevalence, linguistic style, documentation practices, sample size, clinical ambiguity, and the availability of explicit conduct markers in the dialogue. Therefore, the observed variation should not be attributed solely to medical complexity or to architectural limitations. From a public health perspective, however, the result is important because it suggests that models trained on digital health interactions may behave unevenly across service domains.

This statistical variance is visually summarized in Figure 3, which illustrates fluctuations in error rates across the top ten most frequent medical specialties. Acknowledging this heterogeneity is critical for public health applications, as it suggests that future systems should include subgroup monitoring, specialty-aware threshold analysis, and external validation before being used to inform operational decisions. In this sense, post-hoc specialty analysis acts as an auditing mechanism for identifying potential inequalities in model reliability across areas of care.

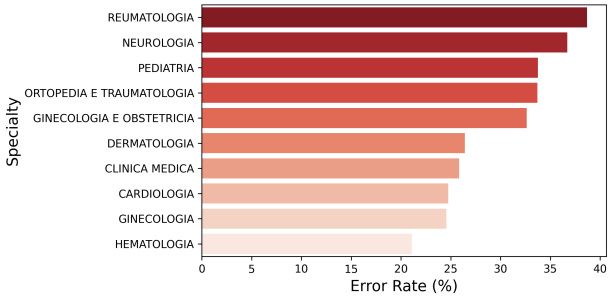

Figure 3. Distribution of model errors across the top 10 medical specialties, highlighting specialty-level heterogeneity in conduct classification.

7. Conclusion

This study investigated how interprofessional teleconsultation records can be mined as digital and discursive signals for operational public health analysis. We evaluated a retrospective conduct classification pipeline designed to identify patterns of Retention and Re-

ferral in completed teleconsultation dialogues. The proposed pipeline utilized BioBERTpt with Head+Tail truncation and Focal Loss to address clinical language and class imbalance. Although the final model did not yield statistically robust gains in aggregate Accuracy or Macro F1-Score over all baselines, it preserved predictive performance while changing the threshold required to optimize Macro F1. Post-hoc analysis further revealed statistically significant heterogeneity across specialties ($p < 0.05$), indicating that model behavior varies across medical domains and should be audited at the subgroup level.

It is essential to contextualize the final accuracy of 71.2%. Because Retention accounts for approximately 73.5% of the cases, a naive majority-class classifier would achieve high accuracy while completely failing to identify referrals. For this reason, Macro F1-Score is a more informative metric for this task. The final model achieved a Macro F1-Score of 0.610, indicating that it captures part of the minority Referral signal despite the natural predominance of Retention. However, this performance should be interpreted as moderate and exploratory, supporting operational analysis rather than autonomous clinical decision-making.

The development of this framework presents research utility for the Brazilian SUS by showing that teleconsultation platforms can be analyzed as sources of digital, discursive, and operational signals. Such signals may support retrospective monitoring of referral flows, identification of specialty-specific demand patterns, and auditing of model reliability across areas of care. Future studies should include external validation in different regions, temporal validation, calibration-specific metrics, specialty-level performance tables, and prospective experiments using only information available before the specialist response. These steps are required before any use in real-time clinical decision support or triage workflows.

References

Andrade, J., Duarte, J., Shimaoka, A., Costa, T., Bandiera-Paiva, P., Junior, A., Silva, G., and Pacheco, E. (2024). Natural language processing for triage of cerebral large-vessel occlusion. *Arquivos de Neuro-Psiquiatria*, 82:1–8.

Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Acesso em: 13 mai. 2026.

Cho, S., Lee, M., Yu, J., Yoon, J., Choi, J.-B., Jung, K.-H., and Cho, J. (2024). Leveraging large language models for improved understanding of communications with patients with cancer in a call center setting: Proof-of-concept study. *J Med Internet Res*, 26:e63892.

Cunha Reis, T. (2025). Artificial intelligence and natural language processing for improved telemedicine: Before, during and after remote consultation. *Atención Primaria*, 57(8):103228.

Feizollah, A., Lin, C.-Y., O'Malley, L., Thompson, W., Listl, S., and Byrne, M. (2025). The use of natural language processing to interpret unstructured patient feedback on health services: Scoping review. *J Med Internet Res*, 27:e72853.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Molenaar, A., Lukose, D., Brennan, L., Jenkins, E. L., and McCaffrey, T. A. (2024). Using natural language processing to explore social media opinions on food security: Sentiment analysis and topic modeling study. *J Med Internet Res*, 26:e47826.

Raff, D., Stewart, K., Yang, M. C., Shang, J., Cressman, S., Tam, R., Wong, J., Tammemägi, M. C., and Ho, K. (2024). Improving triage accuracy in prehospital emergency telemedicine: Scoping review of machine learning–enhanced approaches. *Interact J Med Res*, 13:e56729.

Scherbakov, D. A., Hubig, N. C., Lenert, L. A., Alekseyenko, A. V., and Obeid, J. S. (2025). Natural language processing and social determinants of health in mental health research: Ai-assisted scoping review. *JMIR Ment Health*, 12:e67192.

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., Oliveira, L. E. S. e., Copara, J., Gumiel, Y. B., Oliveira, L. F. A. d., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. M. (2020). BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T., editors, *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 403–417, Berlin, Heidelberg. Springer-Verlag.