Bridging Citizens and Public Services: Improving Service Association with Retrieval-Augmented Generation (RAG) Labels

Ticiana L. Coelho da Silva[1, 4], Celso França[2], Marcos André Gonçalves[2], Leonardo Rocha[3], Leonardo Alamy[4], Fernando Sola Pereira[4], Eduardo Soares de Paiva[4]

[1]Federal University of Ceará, Brazil
[2]Federal University of Minas Gerais, Brazil
[3]Federal University of São João del-Rei, Brazil
[4]Brazilian Office of the Comptroller General, Brazil
ticianalc@insightlab.ufc.br
{celsofranca, mgoncalv}@dcc.ufmg.br
lcrocha@ufsj.edu.br
{leonardo.martins, fernando.pereira, eduardo.paiva}@cgu.gov.br

Abstract

Linking citizen complaints to the public services they concern remains a major challenge in the Brazilian federal administration. In 2025, over 1.2 million manifestations were submitted across 328 agencies, yet only about 1.8% are currently associated with a specific government-offered service, limiting large-scale monitoring and evidence-based management. We cast this task as an extreme multi-class text classification problem marked by severe class imbalance and strong lexical–semantic gaps between citizen language and official service descriptions. Building on recent work that reframes the task as information retrieval, we combine sparse retrieval with BM25 over representative complaint corpora and dense retrieval enriched with RAG-labels: semantically expanded label descriptions generated via Retrieval-Augmented Generation and Small Language Models. This approach markedly reduces vocabulary mismatch and semantic ambiguity, yielding substantial gains over direct text or embedding matching. To our knowledge, this is the first Portuguese-language application of RAG-labels for service–complaint association. In real operational data from the Federal Ombudsman Office, our method can automatically assign plausible services to roughly 73% of previously unlabeled cases, improving coverage and supporting more effective public service evaluation.

1 Introduction

Linking citizen submissions to the public services they reference is a central challenge in the Brazilian federal administration. In 2025, over 1.2 million submissions were recorded across 328 federal agencies, spanning complaints, reports, suggestions, and compliments. [1] Although the Ministry of Management and Innovation (MGI) maintains a structured catalog of more than 5,000 federal services [2], only about 1.8% of submissions registered in the Fala.BR platform are currently associated with a specific service (Tribunal de Contas da União, 2025). This extremely low linkage rate undermines service monitoring, limits evidence-based policy planning, and prevents systematic identification of structural issues. Prior NLP-based analyses of citizen submissions (Jiao et al., 2024; Esperança et al., 2025) highlight the need for automated solutions that operate at scale while accommodating the linguistic variability of citizen discourse.

The task can be formulated as large-scale Extreme (Multi-Class) Text Classification (XMTC), with hundreds or thousands of service labels. As shown by (França et al., 2025), XMTC is characterized by two main challenges: (i) a severe lexical–semantic gap between inputs and class descriptions and (ii) extreme class imbalance. Vocabulary mismatch arises because citizen language diverges substantially from the terminology used in official service descriptions. At the same time, semantic ambiguity emerges when common terms (e.g., "registration") map to multiple services depending on agency and context. These properties closely mirror the dynamics of a public-service association, in which submissions often contain sparse or ambiguous cues about the underlying service.

To address these challenges, (França et al., 2025) recast XMTC as an Information Retrieval (IR) problem, treating the input as a query and the classes as documents to be ranked. Their sparse component retrieves representative texts linked to each class, producing (text, class) pairs whose aggregated scores mitigate vocabulary mismatch. Their dense component introduces

[1]https://centralpaineis.cgu.gov.br/visualizar/resolveu

[2]https://www.gov.br/pt-br/servicos

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 131–140

RAG-labels, expanded class descriptions generated via Retrieval-Augmented Generation (Lewis et al., 2020b), which enrich each class's semantic space and better approximate citizen language. Although developed for general XMTC settings, this IR-driven formulation and its reliance on expanded label semantics align naturally with the linguistic properties of public-service submissions.

These observations motivate three research questions: **RQ1**: To what extent do citizen-aligned RAG-labels improve service-association performance over technical service descriptions alone? **RQ2**: Which strategy for selecting the K contextual submissions yields the most effective RAG-labels? **RQ3**: What is the magnitude of the lexical–semantic gap between citizen submissions and technical service descriptions?

Empirically, we find that direct lexical overlap between complaints and technical descriptions is just above 1%, and embedding-based similarity (Chandrasekaran and Mago, 2021) between the two spaces yields weak MRR and Top-1 results. These findings confirm the substantial linguistic misalignment between administrative and citizen registers. Incorporating IR-driven methods significantly reduces this gap. On the sparse side, BM25 retrieval over representative submissions improves lexical coverage. On the dense side, the most effective representations combine (i) a citizen-oriented description generated by a Large Language Model (LLM) enriched with RAG-labels and (ii) the official technical description. This hybrid representation produces consistent gains across ranking metrics, improving MRR by 13% and Top-1 by 25% relative to the strongest dense baseline.

To our knowledge, this is the first application of RAG-labels in Portuguese for a public-service association. Beyond methodological contributions, the approach delivers substantial operational value: more than 90% of dataset submissions lack service linkage, and our model correctly associates approximately 80% of them. By enabling large-scale, semantically informed linkage, the proposed approach reduces informational asymmetry and strengthens institutional capacity to monitor and improve public service delivery.

This paper is organized as follows. Section 2 covers related work. Section 3 describes our methodology. Section 4 details the experimental protocol. Section 5 presents and discusses our experimental results. Section 6 concludes the paper.

2 Related Work

Research on large-scale text classification has evolved from classical machine-learning pipelines to deep neural architectures and, more recently, to hybrid retrieval–generation approaches. This trajectory reflects shifts not only in available computational tools but also in how the field conceptualizes representation learning, scalability, and label-space modeling in Multi-Class Text Classification (XMTC).

Early approaches relied on traditional machine-learning algorithms such as Support Vector Machines (SVM) (Sun et al., 2009), Logistic Regression (Cunha et al., 2023), and Random Forest (Zhou et al., 2016), typically combined with sparse TF–IDF representations. These methods offered interpretability and low computational cost, but their dependence on manual feature engineering and inability to capture deeper semantic structure limited performance in settings with large vocabularies, heterogeneous labels, and high lexical variability (Zhou et al., 2016).

The introduction of Small Language Models (SLMs)—e.g., BERT (Devlin et al., 2019), RoBERTa (Sy et al., 2024), BART (Lewis et al., 2020a), and XLNet (Yang et al., 2019)—marked a significant shift toward dense contextual representations. Comparative evaluations such as (Cunha et al., 2023) show substantial Macro-F1 gains when replacing classical models with SLMs. Nevertheless, their high inference cost and limited scalability to thousands of classes motivated the development of architectures that restructure the label space.

Hierarchical and clustering-based models emerged as a response to this scalability challenge. Methods such as Match-XML (Ye et al., 2024), LightXML (Jiang et al., 2021), AttentionXML (You et al., 2019), and XR-Transformer (Zhang et al., 2021) reduce decision complexity by organizing labels into learned clusters and traversing them with attention-based mechanisms. While effective on benchmarks, these methods depend heavily on cluster quality and are less adaptable in domains where labels are dynamic, semantically overlapping, or weakly structured—conditions typical of real-world public-service taxonomies.

The latest wave of work leverages LLMs, including LLaMA (Touvron et al., 2023), Mistral (Jiang, 2024), BloomZ (Muennighoff et al., 2022), and DeepSeek (Guo et al., 2025). LLMs bring powerful generalization and in-context learning

capabilities. In XMTC, they have been used either to generate high-quality embeddings or to enrich label semantics by conditioning class descriptions or alternative label formulations. Recent evidence (de Andrade et al., 2024) shows that tuned LLMs can outperform SLMs under extreme imbalance, though often at substantial computational cost.

A recent advance in XMTC is xCoRetriev (França et al., 2025), which models the task as a two-stage retrieval and fusion pipeline designed to address the fundamental challenges of volume, skewness, and label-quality degradation. xCoRetriev introduces three core innovations: (i) dynamic slicing of the large label space to improve scalability; (ii) a fusion mechanism that jointly leverages sparse and dense retrievers while explicitly favoring tail labels; and (iii) the enrichment of the label space through Retrieval-Augmented Generated (RAG) labels to mitigate quality issues. Across four public XMTC benchmarks containing hundreds of thousands of documents and labels, xCoRetriev achieves substantial gains – up to 48% on propensity-scored metrics compared to the strongest baselines – while remaining among the most efficient methods in both training and inference. These results highlight its balanced strengths in effectiveness, scalability, and robustness to noise.

3 Methodology

Figure 1 presents the overall workflow adopted in this study. Although the Federal Ombudsman's database contains millions of citizen submissions collected via the Fala.Br platform,[3] only about 1.8% are associated with a public service. The steps below detail the methodological components used to construct a reliable evaluation setting and to generate, enrich, and compare service representations.

Dataset Construction

We extracted all citizen submissions registered between January 2023 and July 2025 and filtered the corpus to retain only those entries explicitly linked to a public service. This procedure resulted in a curated dataset of 64,231 instances referencing a total of 227 distinct public services. Restricting the analysis to a single federal administration mitigates structural changes caused by government transitions—such as ministerial reorganizations, agency creation and extinction, and catalog updates—that would otherwise introduce inconsistencies in the mapping between submissions and services. The resulting dataset provides a stable and trustworthy ground truth for evaluation.

Embeddings and Service Descriptions

Submission embeddings were generated using a pretrained transformer-based SLM. For each submission, its embedding is compared with the service description's embedding. We experiment with three types of descriptions: (i) technical, (ii) citizen-oriented, and (iii) hybrid.

Technical descriptions. These consist of the concatenation of the service name, formal description, popular names, and keywords, all retrieved from the MGI API [4]. This process yields structured information for approximately 5,350 services.

Citizen-oriented descriptions (RAG-labels). Since citizen submissions differ markedly from administrative language, we adopt the RAG-label strategy (França et al., 2025) to generate citizen-aligned descriptions. A LLM receives (i) a prompt, (ii) a set of K submissions previously associated with the service, and (iii) its technical description. The LLM then produces a concise, natural-language description reflecting how citizens typically refer to the service. The prompt is provided in Listing 1.

Hybrid descriptions. We concatenate the technical and citizen-oriented descriptions to obtain a representation that combines administrative precision with citizen phrasing. All three variants are embedded using the same SLM.

Listing: Prompt used for generating the citizen-oriented service description

You are an NLP model specialized in public-service language understanding.
Your task is to generate a **representative citizen-oriented description** of the service "{service_name}" based on real submissions related to it. This description will be used to match citizen submissions to the correct service and must reflect **how citizens refer to and understand this service**.
Instructions:

- Write in simple, natural language, as

[3] https://falabr.cgu.gov.br/web/home

[4] https://www.gov.br/conecta/catalogo/apis/api-de-servicos

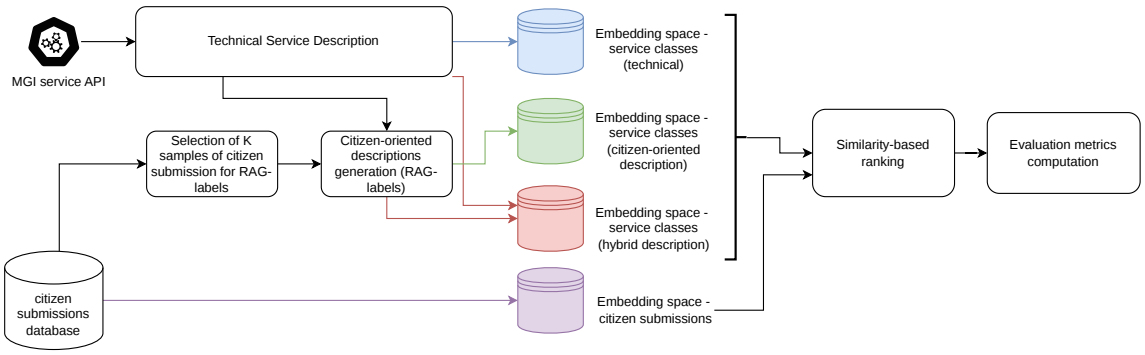

Figure 1: Overall methodological pipeline.

citizens would explain a service to one another.

- Capture the main themes, needs, and concerns present in the submissions.
- Maintain a neutral, informative tone.
- Produce 3–6 short sentences.
- Do not copy text from submissions or the technical description.
- Do not use first-person expressions.
- Ensure the description reflects the service's meaning and purpose from the citizens' perspective.

Related submissions: `{manifestation_text}`
Technical description: {technical_description_service}
Now write the **citizen-oriented description** of {service_name}.

Selecting the Sample for RAG-Labels

To generate RAG-labels, each service is paired with K representative submissions. We compare two sampling strategies: (i) random selection and (ii) MMR-based selection (Goldstein and Carbonell, 1998), which promotes both relevance and diversity to expose the LLM to broader linguistic variation.

Figure 2 shows the distribution of submissions per service. Only 229 services had at least seven associated submissions; among them, 227 were active in the MGI API. Consequently, we set $K = 7$ for all experiments. The seven submissions used to construct the citizen-oriented description of each service were excluded from the test set.

Similarity-Based Ranking

For each submission, we compute the cosine similarity between its embedding and the ones of services within the same government agency. This yields a ranked list of candidate services. Ranking is produced independently for each of the four descriptions: (i) technical, (ii) citizen-oriented (random K), (iii) citizen-oriented (MMR K), and (iv) hybrid.

Evaluation Metrics

Ranking quality is assessed using standard Information Retrieval metrics. The main metric is Mean Reciprocal Rank (MRR), defined as:

$$\text{MRR} = \frac{1}{|M|} \sum_{i=1}^{|M|} \frac{1}{\text{rank}_i}, \tag{1}$$

where M is the set of submissions and rank_i is the position of the correct service in the ranking for submission i.

We report Top-1, Top-2, Top-5, and Top-10 accuracy, which indicate whether the correct service is retrieved among the top k ranked candidates. In addition, we report MRR to capture the position of the correct service within the ranking.

Together, these metrics assess both ranking quality and operational usefulness: while MRR and Top-1 emphasize precise ranking, Top-k accuracy reflects the system's ability to narrow the candidate space to a manageable set, supporting efficient human validation in platforms such as Fala.Br.

4 Experimental Protocol

Our experiments address the three research questions introduced earlier: the impact of citizen-oriented descriptions on classification performance (**RQ1**); the comparative effectiveness of sampling

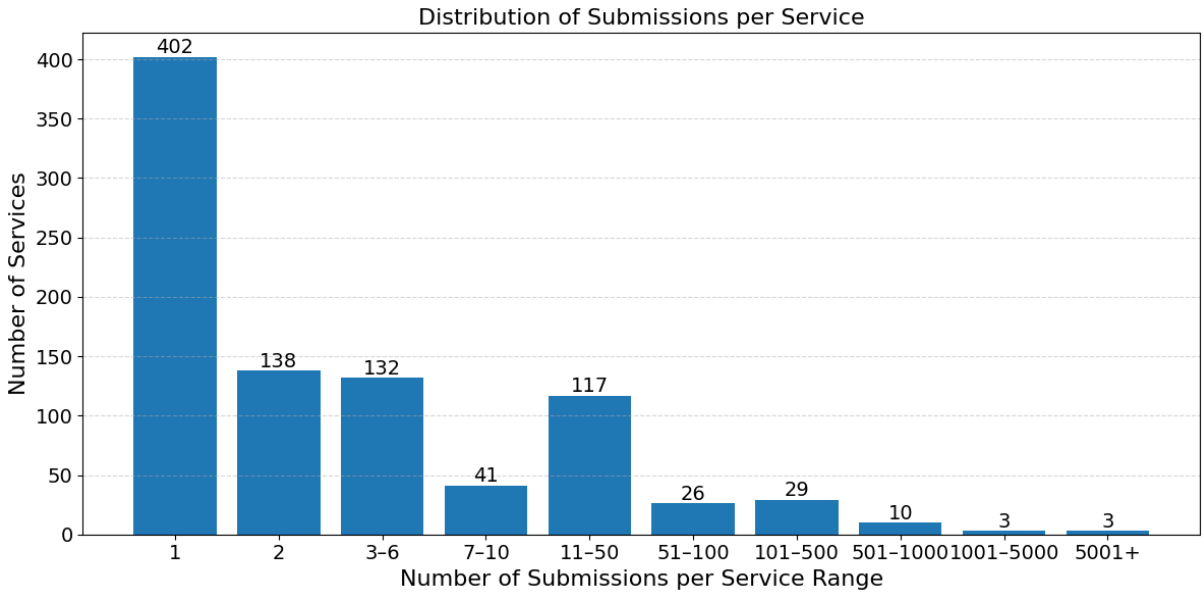

Figure 2: Distribution of submissions per service.

strategies for selecting the K submissions used in *RAG-labels* (**RQ2**); and the magnitude of the lexical–semantic gap between submissions and service descriptions (**RQ3**).

Models

We evaluate several pretrained embedding models with full support for Portuguese. Specifically, we employ three variants of Qwen3-Embedding (Zhang et al., 2025) (0.6B, 4B, and 8B), as well as the `paraphrase-multilingual-MiniLM-L12-v2` model from Sentence Transformers (Reimers and Gurevych, 2019). For generating citizen-oriented service descriptions (*RAG-labels*), we use `Qwen3-30B-Instruct` (Team, 2025), which provides high-quality generative behavior for downstream rephrasing tasks.

Computational Setup

All experiments were executed on a workstation equipped with an Intel Core i9-12900K (12th Gen, 3.20 GHz), 64 GB RAM, and an NVIDIA RTX A4000 GPU with 16 GB VRAM. This infrastructure served as the backend for both embedding generation and execution of the Qwen3-30B model.

Ranking Procedure

After computing embeddings for all submissions and services, we perform similarity-based ranking. For each submission in the test set—comprising more than 60,000 instances with known service labels—we compute cosine similarity between its embedding and the embeddings of all services belonging to the same government agency. The resulting scores are sorted in descending order, yielding a per-submission ranking of candidate services.

We repeat this process independently for each service-representation variant: (i) technical descriptions; (ii) citizen-oriented descriptions generated using randomly sampled K submissions; (iii) citizen-oriented descriptions generated via MMR-based sampling; and (iv) hybrid descriptions combining technical and citizen-oriented text. This setup allows us to isolate and quantify each representation strategy's contribution to overall retrieval performance.

5 Experimental Results

The experiments address the three research questions introduced earlier and follow the protocol described in the previous section. We evaluate (i) the impact of citizen-oriented descriptions on classification performance (RQ1), (ii) the effect of different sampling strategies for generating *RAG-labels* (RQ2), and (iii) the magnitude of the lexical–semantic gap between citizen submissions and service descriptions (RQ3). Together, these experiments allow us to assess not only whether *RAG-labels* improve service association effectiveness, but also why they do so and under which representational conditions they are most beneficial.

5.1 Effectiveness of RAG-Labels

Before evaluating RAG-labels, we first identify the most suitable embedding model for our dataset.

Table 1 reports MRR and Top-k accuracy for four embedding models using only technical descriptions. The evaluation includes more than 60,000 submissions mapped to 227 services, a setting characterized by high thematic diversity and strong heterogeneity in linguistic expression.

The Sentence Transformer model, while computationally efficient, performs the worst across all metrics, indicating that simpler architectures struggle to represent the nuanced and context-dependent phrasing used by citizens. All Qwen3-Embedding variants substantially outperform it, demonstrating stronger semantic modeling capacity. Qwen3-Embedding-4B and Qwen3-Embedding-8B present very similar performance, with the 4B model providing the best balance between computational cost and accuracy. This result is particularly relevant for large-scale government platforms, where deployability and processing time are critical constraints. Wilcoxon tests (Wilcoxon, 1992) confirm that differences across models are statistically significant ($\rho < 0.05$), reinforcing the robustness of these conclusions.

Taken together, these results define Qwen3-Embedding-4B as the default backbone for subsequent experiments. More importantly, they establish a solid baseline relying exclusively on institutional service descriptions, enabling us to isolate the incremental value introduced by *RAG-labels* in the following analyses.

5.2 Impact of Citizen-Oriented Descriptions

We compare three service-representation conditions: (i) technical descriptions; (ii) citizen-oriented descriptions generated via RAG-labels; and (iii) hybrid descriptions combining both sources. Table 2 shows that RAG-labels significantly outperform technical descriptions, improving MRR from 0.7614 to 0.8446 and yielding consistent gains across all Top-k metrics. This indicates that grounding representations in citizen language leads to more faithful alignment between submissions and services.

The hybrid representation achieves the best overall performance, confirming that institutional and citizen-oriented descriptions encode complementary knowledge: while official descriptions provide procedural precision, citizen-oriented ones capture how people conceptualize and verbalize services. Wilcoxon tests confirm statistically significant differences between all description variants ($\rho < 0.05$), validating the complementary nature of both sources. Practically, the improvement in Top-1 accuracy from 0.6480 to 0.8075 represents a substantial reduction in manual analyst workload and significantly increases the likelihood of correct automatic association.

5.3 Sampling Strategies for RAG-Labels

RQ2 evaluates whether the choice of K submissions used as context for generating RAG-labels affects downstream performance. We compare random sampling with Maximal Marginal Relevance (MMR) sampling. As shown in Table 3, both strategies yield strong improvements over technical descriptions alone, confirming that exposing LLMs to real examples of citizen discourse is consistently beneficial.

Random sampling performs slightly better than Maximum Marginal Relevance sampling, achieving the highest MRR (0.8698) and Top-1 accuracy (0.8075). Despite the numerical difference being relatively small, Wilcoxon signed-rank tests confirm that these gains are statistically significant under the same evaluation protocol adopted in previous sections ($\rho < 0.05$). This indicates that submissions associated with the same service tend to form reasonably coherent linguistic clusters, making sophisticated diversity-driven sampling less critical. From an operational perspective, this is particularly encouraging, as random sampling is simpler, scalable, and easier to maintain in production environments.

5.4 Lexical–Semantic Gap Between Submissions and Services

To quantify RQ3, we evaluate both lexical and semantic alignment between submissions and service representations. Lexically, we apply BM25 using, for each service, a corpus of seven randomly sampled submissions. Using official technical descriptions yields extremely low term overlap ($\sim$1%), leading to poor performance and confirming a severe vocabulary mismatch. Even when grounded in citizen submissions, BM25 remains limited (MRR = 0.77, Top-1 = 0.66) and computationally expensive ($\sim$26 minutes to process the test set).

Semantically, embeddings generated from technical descriptions achieve performance comparable to BM25 but with much lower inference time ($\sim$3 minutes). However, technical descriptions still fail to fully reflect the linguistic structure of citizen discourse, resulting in persistent semantic misalignment.

Table 1: Model performance on the service-ranking task

Model	MRR	Top-1	Top-2	Top-5	Top-10
Sentence Transformer	0.4487	0.3088	0.4486	0.6035	0.7185
Qwen3-Emb-8B	0.7557	0.6352	0.8018	**0.9012**	**0.9471**
Qwen3-Emb-4B	**0.7614**	**0.6480**	**0.8023**	0.8977	0.9462
Qwen3-Emb-0.6B	0.6440	0.4819	0.6775	0.8540	0.9313

Table 2: Classification Effectiveness using different types of service descriptions.

Method	MRR	Top-1	Top-2	Top-5	Top-10
Technical description (baseline)	0.7614	0.6480	0.8023	0.8977	0.9462
Citizen-oriented description (RAG-label)	0.8446	0.7709	0.8704	0.9375	0.9667
Citizen + technical (hybrid)	**0.8698**	**0.8075**	**0.8910**	**0.9493**	**0.9743**

Table 3: Selection of the *K RAG-labels* manifestations (citizen + technical description) - Random vs MMR.

Method	MRR	Top-1	Top-2	Top-5	Top-10
RAG-label (Random)	**0.8698**	**0.8075**	**0.8910**	0.9493	**0.9743**
RAG-label (MMR)	0.8649	0.7992	0.8878	**0.9495**	0.9728

RAG-labels dramatically reduce this gap: citizen-oriented descriptions increase MRR from 0.76 to 0.84 and Top-1 from 0.64 to 0.77. Hybrid descriptions yield the largest gains, improving MRR by 13% and Top-1 by 25% relative to the baseline (we refer to Table 2). Table 4 provides qualitative evidence that RAG-labels capture common phrasings, intents, and communicative styles absent in institutional texts, bringing the representational space of services significantly closer to citizen language.

Overall, the results confirm that the lexical–semantic gap is a central obstacle in public-service association and that *RAG-labels* offer an effective and computationally viable solution to mitigate it.

6 Conclusions and Future Work

This work presented the first Portuguese-language approach for automatically associating citizen submissions to public services, introducing *RAG-labels* to reduce the substantial lexical–semantic gap between institutional descriptions and citizen language on Fala.Br. By generating citizen-oriented reformulations and combining them with official texts in a hybrid representation, our framework delivers consistent effectiveness gains—approximately 13% in MRR and 25% in Top-1 accuracy over a strong baseline. These gains translate into a substantial increase in correct associations in a challenging real-world setting where service annotations are scarce. In the full dataset, only about 2% of submissions are explicitly linked to a service. Among the remaining 98% without service labels, approximately 73% are associated with a public agency, defining the subset of cases in which our approach can be effectively applied.

Beyond improving retrieval accuracy, our findings highlight that public-sector NLP requires bridging linguistic distance rather than simply scaling model capacity. The proposed approach demonstrates that socially grounded intermediate representations can enhance both performance and institutional usability.

Future work will focus on extending coverage to services with little or no historical data, improving robustness under evolving administrative contexts, and exploring richer retrieval–ranking architectures to further advance AI support for public service delivery.

7 Limitations

Despite the strong empirical results achieved by our approach, several limitations remain. First, generating *RAG-labels* requires that each service have a minimum number of associated submissions. This excludes services with sparse or no historical data—often newly created or rarely accessed—and limits full catalog coverage. Although hybrid descriptions reduce vocabulary mismatch, their quality still depends on the availability and representativeness of sampled submissions.

Federal Police		Office of the Comptroller General	
Service	Extend Stay as a Tourist in Brazil	**Service**	Submit an Ombudsman Report on the Fala.BR Platform
Citizen Submission	I am an Argentine tourist [...] is it possible to request another extension?	**Citizen Submission**	Good morning [...] would this reassignment be considered a misuse of function?
Technical Description	Service name: Extend Stay [...]	**Technical Description**	Service name: Submit an ombudsman report [...]
Citizen-Oriented Description	Need to stay longer in Brazil as a tourist? [...]	**Citizen-Oriented Description**	Used to file complaints, reports, compliments, or suggestions [...]

Table 4: Comparison between official service descriptions and RAG-label enriched descriptions

Second, our evaluation is restricted to a single federal administration to avoid inconsistencies introduced by structural changes in government agencies and service catalogs. While this ensures a reliable ground truth, it constrains the temporal generalizability of the findings; linguistic patterns and service structures may evolve across different administrative periods.

Third, although LLMs provide high-quality citizen-oriented descriptions, their outputs remain sensitive to prompt design, sampling strategy, and model-specific biases. Even with MMR and random sampling, generated descriptions may disproportionately reflect dominant linguistic patterns or overlook minority perspectives.

Fourth, our ranking relies on cosine similarity over static embeddings, which may not fully capture the relational or compositional nuances of complex submissions that involve multiple issues or implicit references. More expressive architectures, such as cross-encoders or retrieval-augmented rerankers, could yield further improvements but were not explored due to computational cost.

Finally, our experiments assume a single service label per submission. In practice, many submissions reference multiple services or contain ambiguous service boundaries. Extending the framework to multi-label or hierarchical classification remains an open direction for future work.

Ethics Statement

This study uses publicly accessible and institutionally anonymized citizen submissions from the Brazilian Federal Ombudsman. All data were processed to remove residual sensitive content, and no attempt was made to infer personal attributes or behaviors. The proposed approach is intended to support service triage and administrative analysis—not to automate decisions affecting individuals—and should be deployed only with appropriate human oversight and transparency regarding system limitations.

The use of LLMs to generate citizen-oriented descriptions introduces risks of bias amplification and hallucinated content. Although sampling strategies and prompt constraints were designed to promote neutrality, generative outputs may still reflect societal or model-specific biases. Any operational adoption should include monitoring, auditing, and impact assessment, particularly to avoid disadvantaging vulnerable groups or reinforcing inequities in citizens' access to public services.

Acknowledge

This work was (also) supported by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1.

References

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *ACM Computing Surveys*, 54(2):1–37.

W. Cunha and 1 others. 2023. A comparative survey of instance selection methods applied to nonneural and transformer-based text classification. *ACM*.

Claudio de Andrade, Washington Cunha, Davi Reis, Adriana Silvina Pagano, Leonardo Rocha, and Marcos André Gonçalves. 2024. A strategy to combine 1stgen transformers and open llms for automatic text classification. *arXiv preprint arXiv:2408.09629*.

Jacob Devlin and 1 others. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Marco Esperança, Diogo Freitas, Pedro V. Paixão, Tomás A. Marcos, Rafael A. Martins, and João C. Ferreira. 2025. Proactive complaint management in public sector informatics using ai: A semantic pattern recognition framework. *Applied Sciences*, 15(12).

Celso França, Gestefane Rabbi, Thiago Salles, Washington Cunha, Leonardo Rocha, and Marcos André Gonçalves. 2025. Optimizing tail-head trade-off for extreme multi-label text classification (xmtc) with rag-labels and a dynamic two-stage retrieval and fusion pipeline. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 1392–1401, New York, NY, USA. Association for Computing Machinery.

Jade Goldstein and Jaime G Carbonell. 1998. Summarization:(1) using mmr for diversity-based reranking and (2) evaluating summaries. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 181–195.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Fengqing Jiang. 2024. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington.

Ting Jiang and 1 others. 2021. Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. In *AAAI*, volume 35, pages 7987–7994.

Yaran Jiao, Chunming Li, Ziyan Yao, Chen Weng, Anxin Lian, and Rencai Dong. 2024. How can online citizen complaints provide solutions to refine environmental management: A spatio-temporal perspective. *Information Processing & Management*, 61(2):103611.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th ACL*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Aixin Sun, Ee-Peng Lim, and Ying Liu. 2009. On strategies for imbalanced text classification using svm: A comparative study. *Decision Support Systems*, 48(1):191–201.

Christian Y Sy, Lany L Maceda, Mary Joy P Canon, and Nancy M Flores. 2024. Beyond bert: Exploring the efficacy of roberta and albert in supervised multiclass text classification. *International Journal of Advanced Computer Science & Applications*, 15(3).

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tribunal de Contas da União. 2025. Acórdão nº 744/2025-p – plenário. Acesso em: 04 dez. 2025.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer New York, New York, NY.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Hui Ye, Rajshekhar Sunderraman, and Shihao Ji. 2024. Matchxml: An efficient text-label matching framework for extreme multi-label text classification. *IEEE TKDE*, 36(9):4781–4793.

Ronghui You and 1 others. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *NeurIPS*, volume 32, pages 1–11.

Jiong Zhang and 1 others. 2021. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. In *NeurIPS*, volume 34, pages 7267–7280.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Qifeng Zhou, Hao Zhou, and Tao Li. 2016. Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features. *Knowledge-based systems*, 95:1–11.