ITSM-BERT: Classificação de Tickets de Service Desk via Fine-Tuning do BERTimbau Integrado com Sistema GLPI

Anderson Maia de Lima Carvalho[1,4], Fábio Manoel França Lobato[1,2], Antonio Fernando Lavareda Jacob Junior[1], Omar Andres Carmona Cortes[1,3]

[1]Universidade Estadual do Maranhão (UEMA) – São Luís, MA – Brasil
[2]Universidade Federal do Oeste do Pará (UFOPA) – Santarém, PA – Brasil
[3]Instituto Federal do Maranhão (IFMA) – São Luís, MA – Brasil
[4]Tribunal de Justiça do Estado do Maranhão (TJMA) – São Luís, MA – Brasil

amlcarvalho@tjma.jus.br, fabio.lobato@ufopa.edu.br,
antoniojunior@professor.uema.br, omar@ifma.edu.br

Abstract. *This article proposes the ITSM-BERT, a model based on BERTimbau, fine-tuned to classify Service Desk tickets in Brazilian Portuguese, using data enrichment. The approach was evaluated with 80,879 real records from the Maranhão Court of Justice, comparing frequency-based, vector-based, and Transformer methods (SVM-BoW, TF-IDF, Word2Vec, and BERT-SVM) through Friedman and Nemenyi tests. In autonomous classification, ITSM-BERT outperforms all baselines (F1-macro = 0.840) and reaches 0.997 with semi-structured attributes from the service catalog (+15.7 percentage points). Although statistically equivalent to SVM-TF-IDF in the assisted scenario, it establishes itself as a robust solution for mitigating semantic ambiguities in raw descriptions, with potential for use by other companies and public agencies that utilize ticketing systems, such as GLPI.*

Resumo. *Este artigo propõe o ITSM-BERT, um modelo baseado no BERTimbau, com ajuste fino para a classificação de chamados de Service Desk em português brasileiro, utilizando dados enriquecidos. A abordagem foi avaliada com 80.879 registros reais do Tribunal de Justiça do Maranhão, comparando métodos de frequência, vetoriais e Transformers (SVM-BoW, TF-IDF, Word2Vec e BERT-SVM) por meio de testes de Friedman e de Nemenyi. Na classificação autônoma, o ITSM-BERT supera todos os baselines (F1-macro = 0,840) e atinge 0,997 com atributos semiestruturados do catálogo de serviços (+15,7 p.p.). Embora estatisticamente equivalente ao SVM-TF-IDF no cenário assistido, consolida-se como uma solução robusta para mitigar ambiguidades semânticas em descrições brutas, com potencial para ser utilizada em outras empresas e órgãos públicos que utilizam sistemas de chamados, como o GLPI.*

1. Introdução

O gerenciamento de serviços de TI, amplamente conhecido pelo termo em inglês *Information Technology Service Management* (ITSM), constitui um pilar estratégico para a entrega de valor e a continuidade operacional nas organizações, sendo orientado por *frameworks* consolidados, como o *Information Technology Infrastructure Library* (ITIL) [Axelos 2019] e o *Control Objectives for Information and Related Technologies* (COBIT) [ISACA 2019]. Nesse contexto, o *service desk* atua como o ponto único de contato entre os usuários e as equipes de suporte, utilizando sistemas de gerenciamento de chamados para registrar e acompanhar a resolução de incidentes e requisições [Fuchs et al. 2026]. Contudo, a digitalização dos serviços, associada à expansão de canais de atendimento (portais e aplicações móveis), tem gerado um crescimento exponencial no

volume de demandas, resultando em sobrecarga das equipes de suporte e atrasos na resolução [Ferdinand et al. 2025].

Este cenário revela uma contradição estrutural: enquanto as interfaces de atendimento se modernizam, processos críticos, como a triagem e o roteamento de chamados, permanecem predominantemente manuais [Jain et al. 2025]. A classificação inicial torna-se ineficiente por depender do conhecimento tácito dos analistas [Liu et al. 2025]. Consequentemente, falhas nessa etapa provocam a redistribuição de chamados entre múltiplos especialistas, comprometendo a produtividade e o cumprimento dos *Service Level Agreements* (SLAs) [Alsaç et al., 2025]. Na prática, uma das principais limitações desse processo manual manifesta-se na inconsistência da interpretação do contexto semântico das solicitações. Por exemplo, um chamado descrito como *"não consigo protocolar o processo"* pode ser classificado de forma ambígua como *"Acesso a Sistema"*, *"Erro de Aplicação"* ou *"Suporte ao Usuário"*, dependendo da experiência do operador. Essa inconsistência taxonômica eleva o tempo médio de atendimento devido a roteamentos inadequados e compromete a base de treinamento dos modelos preditivos.

Com o avanço das técnicas de Processamento de Linguagem Natural (PLN), diversos estudos têm investigado métodos para a classificação automática de chamados de *service desk* [Paket et al. 2024]. Abordagens tradicionais baseadas em *Support Vector Machines* (SVM), combinadas com *Bag-of-Words (BoW)* e *Term Frequency-Inverse Document Frequency (*TF-IDF), têm apresentado desempenho satisfatório nessa tarefa [Boonklay e Jinarat 2025]. Especificamente nesse domínio, estudos demonstram que o SVM supera algoritmos como Regressão Logística, *Naive Bayes* e *K-Nearest Neighbors* (KNN) [Paramesh e Shreedhara 2019], além de apresentar desempenho superior aos de *Random Forests* [Dlodlo e Sibanda 2023]. Apesar do desempenho satisfatório do SVM, sua dependência de representações esparsas limita a captura de relações semânticas complexas.

Nesse sentido, as arquiteturas baseadas em *Transformers* redefiniram o estado da arte em PLN [Vaswani et al. 2017], com destaque para o *Bidirectional Encoder Representations from Transformers* (BERT) [Devlin et al. 2019], capaz de capturar relações contextuais complexas. Para o português, o BERTimbau fornece representações pré-treinadas adequadas ao contexto nacional [Souza et al. 2020]. Apesar desses avanços, sua aplicação no Judiciário brasileiro enfrenta desafios relacionados à especificidade terminológica, uma vez que o vocabulário jurídico combina jargões técnicos e estruturas linguísticas próprias, o que exige modelos orientados ao domínio [Carmo et al. 2023].

É nesse contexto que organizações com grandes volumes de chamados em português, como tribunais de justiça, constituem um ambiente propício à aplicação dessas técnicas. No âmbito do Poder Judiciário, o Tribunal de Justiça do Maranhão (TJMA) mantém uma central de *service desk* que registra mais de 126.000 chamados por ano por meio do *Gestionnaire Libre de Parc Informatique* (GLPI), um sistema *open source* para gerenciamento de chamados. Esse sistema centraliza o registro das demandas, mas não dispõe, de forma nativa, de mecanismos de classificação automática com base em PLN. Os chamados são descritos em linguagem natural, geralmente curta

e técnica, e classificados manualmente por analistas, que interpretam a descrição textual e encaminham ao grupo técnico solucionador correspondente.

Nesse cenário de variabilidade semântica, o enriquecimento de dados (*data enrichment*) surge como uma estratégia para melhorar a qualidade dos dados ao incorporar informações adicionais, ampliando a capacidade preditiva dos modelos de classificação [Barchilon et al. 2024]. Em PLN, essa abordagem combina metadados contextuais com representações textuais para capturar nuances que descrições isoladas não explicitariam [Chi et al. 2024]. A eficácia dessa estratégia é evidenciada pelo modelo Ticket-BERT, que alcançou F1-score de 98,24% ao incorporar contextos auxiliares às descrições de chamados, contra 85,90% na configuração textual pura, o que evidencia o impacto do enriquecimento contextual na redução da ambiguidade semântica [Liu et al. 2023].

Este trabalho propõe e avalia o ITSM-BERT, um modelo baseado no BERTimbau, com ajuste fino, para a classificação automática de chamados de ITSM em português no contexto do Judiciário. A avaliação é conduzida por meio de um estudo de caso, utilizando um *corpus* de 80.879 chamados reais do sistema GLPI do TJMA e comparando-o a abordagens tradicionais, incluindo SVM-BoW, SVM-TF-IDF, SVM-Word2Vec e uma abordagem híbrida BERT-SVM. Adicionalmente, investiga-se o impacto da incorporação de atributos semiestruturados do catálogo de serviços como variáveis de entrada no processo de enriquecimento de dados. Nesse contexto, considerando a delimitação do escopo ao estudo de caso investigado, este trabalho busca responder às seguintes Questões de Pesquisa (QP):

QP1 – O uso de modelos de linguagem com ajuste fino específico para o domínio de gestão de serviços de TI resulta em ganhos de desempenho na classificação automática de chamados em português brasileiro?

QP2 – Qual é o impacto do enriquecimento de dados com o catálogo de serviços no desempenho de classificadores automáticos de chamados de ITSM?

Os resultados experimentais mostram que o ITSM-BERT alcança F1-macro de 99,70%, um ganho de 15,7 pontos percentuais em relação à configuração de entrada textual pura com a descrição do usuário, superando os demais modelos, com validação estatística por meio dos testes de Friedman e do *post-hoc* de Nemenyi. Esses achados confirmam que a combinação entre o ajuste fino de domínio e o enriquecimento com dados do catálogo de serviços constitui uma estratégia eficaz para a classificação automática de chamados de ITSM em português brasileiro.

Em síntese, as principais contribuições técnico-científicas deste artigo são: (i) a proposição e avaliação de um modelo baseado em *Transformer* com ajuste fino para classificação de chamados ITSM em português brasileiro, enfrentando desafios de ambiguidade semântica, conhecimento tácito e desbalanceamento em um grande volume de textos técnicos; (ii) análise comparativa entre abordagens tradicionais e baseadas em *Transformer* com validação estatística; e (iii) avaliação do impacto do enriquecimento de dados com atributos semiestruturados do catálogo de serviços, evidenciando a influência da qualidade taxonômica no desempenho dos classificadores.

O restante deste artigo encontra-se organizado como segue. Na Seção 2, são discutidos trabalhos relacionados à presente proposta, evidenciando as lacunas na literatura apontadas na introdução. Os materiais e métodos são detalhados na Seção 3. Os resultados empíricos são apresentados e discutidos na Seção 4. Por fim, na Seção 5, são apresentadas as conclusões, as ameaças potenciais à validade do estudo e as sugestões de trabalhos futuros.

2. Trabalhos Relacionados

A classificação automática de chamados de suporte técnico tem sido amplamente estudada, com abordagens que evoluíram de modelos clássicos para arquiteturas baseadas em *Transformers*. Inicialmente, as pesquisas empregaram técnicas clássicas com representações vetoriais esparsas. Paramesh e Shreedhara (2019) avaliaram SVM, *Naive Bayes* e *Random Forest* para a classificação de chamados de ITSM em *corpus* em língua inglesa, demonstrando que o SVM com TF-IDF alcançou os melhores resultados. Este achado de pesquisa também é corroborado pelo estudo de Dlodlo e Sibanda (2023), igualmente conduzido em inglês, no mesmo domínio de aplicação.

O BERT [Devlin et al. 2019], um modelo de linguagem pré-treinado baseado em *Transformers* [Vaswani et al. 2017], representou um avanço significativo em PLN ao gerar representações contextuais bidirecionais. Liu et al. (2023) apresentaram o Ticket-BERT, um modelo com *fine-tuning* para a classificação de *tickets* de incidentes em inglês, alcançando desempenho superior em comparação com *Naïve Bayes* e *Logistic Regression* com TF-IDF e BoW. No contexto brasileiro, Souza et al. (2020) disponibilizaram o BERTimbau, um modelo BERT pré-treinado em português brasileiro que alcançou o estado da arte em similaridade textual e no reconhecimento de entidades nomeadas, superando o BERT multilíngue. O modelo tem sido aplicado com sucesso em diversas tarefas de PLN, como a classificação de texto em português [Souza e Souza Filho 2023], porém, não foram identificadas aplicações no domínio do ITSM.

No cenário brasileiro, estudos recentes reforçam que a eficácia da classificação textual em instituições públicas reside na especialização representacional e na adaptação ao respectivo domínio. Enquanto Brandão et al. (2023) demonstram, no contexto de licitações, que a representação semântica prevalece sobre técnicas de pré-processamento, Silva et al. (2024) evidenciam que a composição do *corpus* de adaptação para o modelo BERTimbau é crítica para o desempenho em textos governamentais. Nessa linha, Sant'Anna et al. (2024) ratificam a viabilidade dessas abordagens com base em dados reais de produção ao classificarem os registros de dívida ativa. Coletivamente, essas evidências corroboram a premissa de que a captura de nuances terminológicas específicas constitui um diferencial para superar as limitações dos modelos generalistas em sistemas de alta criticidade.

A revisão da literatura evidencia duas lacunas: (i) estudos sobre classificação de chamados ITSM restringem-se a *corpora* em inglês [Paramesh e Shreedhara 2019, Dlodlo e Sibanda 2023, Liu et al. 2023], sem aplicações de *Transformers* com *fine-tuning* em português nesse domínio; e (ii) o impacto do enriquecimento de dados com o catálogo de serviços na classificação de chamados permanece inexplorado, inclusive em trabalhos correlatos no cenário brasileiro [Brandão et al. 2023, Silva et al. 2024]. Este trabalho se distingue por abordar simultaneamente ambas as lacunas,

propondo o ITSM-BERT, um modelo baseado no fine-tuning do BERTimbau para chamados de ITSM em português, com avaliação experimental abrangente que compara cinco abordagens em duas configurações de entrada e com validação estatística rigorosa por meio dos testes de Friedman e dos *post-hoc* de Nemenyi. A metodologia adotada para a condução desse experimento é detalhada na seção seguinte.

3. Metodologia

Este estudo adotou a metodologia *CRoss Industry Standard Process for Data Mining* (CRISP-DM), organizada em seis fases interativas para guiar projetos de ciência de dados [Wirth e Hipp 2000], no contexto de um estudo de caso aplicado. A Figura 1 apresenta o *framework* experimental desenvolvido para a construção do *pipeline* de classificação automática de chamados.

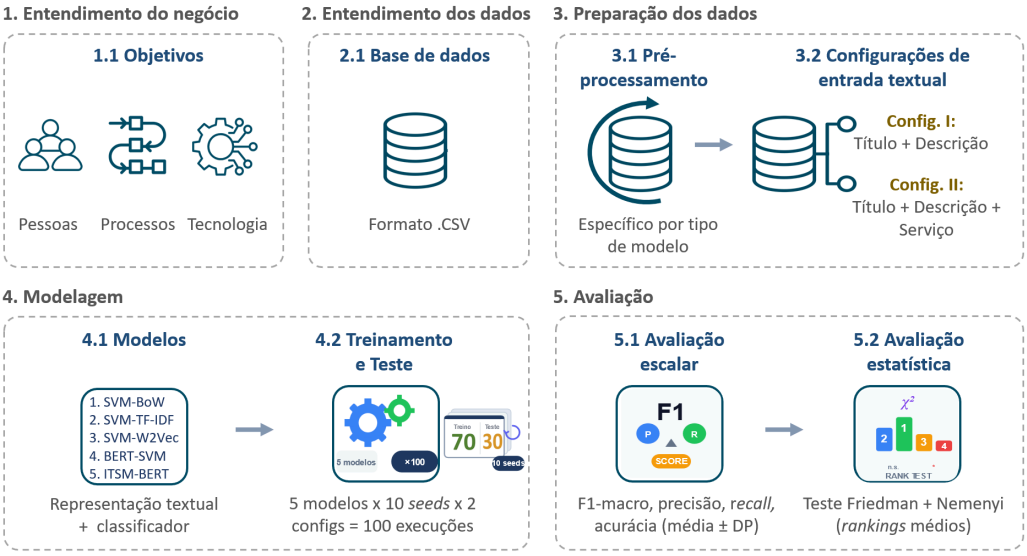

Figura 1. *Framework* experimental para classificação de chamados.

Esse experimento está estruturado em cinco etapas: (i) Entendimento do Negócio, com análise do contexto e dos requisitos do problema; (ii) Entendimento dos Dados, dedicada à exploração e caracterização do *corpus*; (iii) Preparação de Dados, com limpeza, pré-processamento e configuração das entradas textuais; (iv) Modelagem, abrangendo seleção de modelos, treinamento e ajuste de hiperparâmetros; e (v) Avaliação, com mensuração do desempenho dos classificadores por meio de métricas escalares e testes estatísticos. A etapa de implantação está fora do escopo deste estudo. Cada etapa é detalhada nas subseções a seguir.

3.1. Entendimento do Negócio

A fase de entendimento do negócio foi conduzida a partir de três dimensões: entrevistas com gestores (pessoas), mapeamento do fluxo de triagem e escalonamento (processos) e análise do sistema GLPI (tecnologia). O *service desk* do TJMA opera em três níveis: L1 (suporte inicial para demandas simples), L2 (suporte intermediário) e L3 (casos complexos, investigativos ou mudanças no ambiente). O processo atual de classificação depende de conhecimento tácito dos analistas e de catálogos de palavras-chave mantidos

manualmente, o que tende a gerar inconsistências na atribuição dos grupos solucionadores, retrabalho e aumento do tempo de triagem.

Diante disso, o objetivo deste estudo é investigar o uso de modelos de PLN, desde representações vetoriais tradicionais até arquiteturas neurais do tipo *Transformer*, para automatizar a classificação multiclasse de chamados, atribuindo-os automaticamente ao grupo técnico responsável. Como variável-alvo, adotou-se o grupo solucionador final do chamado, independentemente de eventuais redirecionamentos durante o atendimento**.** O critério de sucesso dessa proposta está associado à redução do tempo de triagem e dos erros de escalonamento. Com o problema negocial claramente definido, verificam-se os dados, detalhados a seguir.

3.2. Entendimento dos Dados

O conjunto de dados utilizado neste estudo foi extraído do sistema GLPI, compreendendo registros de chamados técnicos abertos pelos usuários do TJMA de janeiro a dezembro de 2025. O *dataset* bruto continha 126.319 registros, compostos por atributos estruturados e não estruturados, organizados conforme a Tabela 1.

Tabela 1. Descrição do *dataset*.

Atributo	Tipo	Formato	Descrição
Data de abertura	E	Temporal	Data e hora do registro do chamado
Data de fechamento	E	Temporal	Data e hora da resolução do chamado
Prazo de resposta	E	Numérico	SLA associado ao tipo de chamado
Grupo solucionador	E	Categórico	Equipe técnica responsável (variável-alvo)
Título	NE	Textual	Resumo textual do chamado
Serviço	NE	Textual	Catálogo de serviços
Descrição	NE	Textual	Relato detalhado do incidente ou solicitação

E = Estruturado; NE = Não estruturado.

Para este estudo, os campos '*Título*', '*Serviço*' e '*Descrição*' são utilizados como variáveis de entrada principais, capturando tanto a estrutura quanto o conteúdo semântico necessários para uma classificação precisa. No cenário supervisionado, definiu-se como variável de saída (*label*) o '*Grupo Solucionador*', que identifica a equipe técnica responsável pela resolução do chamado. A análise exploratória revelou um forte desbalanceamento entre as 12 classes, conforme apresentado na Figura 2.

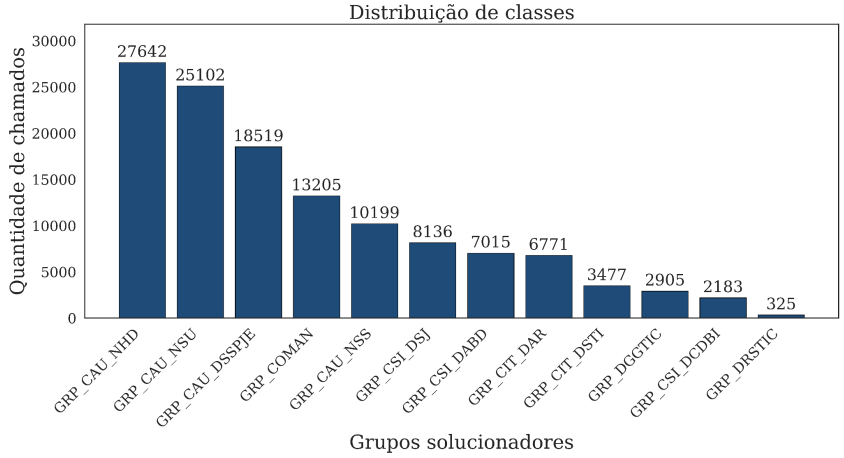

Figura 2. Distribuição de chamados por grupo solucionador.

A distribuição exibe um padrão assimétrico de cauda longa, em que as três classes mais frequentes concentram 56,42% dos chamados, com um *Imbalance Ratio* (IR) de 85,05 entre a classe majoritária e a minoritária. Esse padrão é consistente com a estrutura operacional de ambientes de *service desk*, uma vez que as classes dominantes correspondem aos grupos do primeiro nível de atendimento (L1). A análise textual complementar indicou concentração em descrições curtas, com percentil 95 de 201 palavras. Esse entendimento fundamenta as estratégias de preparação descritas nas subseções seguintes.

3.3. Preparação dos Dados

Nesta fase, o pré-processamento foi estruturado em duas etapas: uma limpeza uniforme aplicada a todo o *corpus*, seguida de um pré-processamento diferenciado, conforme o tipo de representação textual. Na primeira etapa, realizou-se a normalização do *corpus* por meio de expressões regulares, visando eliminar ruídos típicos de *service desk*. As transformações incluíram a extração do conteúdo relevante, a remoção de cabeçalhos de *e-mail*, saudações e assinaturas, a eliminação de *tags* de marcação e a normalização de caracteres especiais. As Informações Pessoais Identificáveis (PII) foram anonimizadas por meio de expressões regulares e de técnicas de *Named Entity Recognition* (NER), em conformidade com a Lei Geral de Proteção de Dados. A deduplicação textual exata entre os campos 'Título' e 'Descrição', realizada antes da divisão treino-teste, resultou na remoção de 39.657 registros redundantes. Por fim, a filtragem por extensão textual (de 5 a 200 palavras) e a exclusão da classe minoritária GRP_DRSTIC (N=289) resultaram em um *dataset* final de 80.879 registros, distribuídos em 11 classes, com média de 40 palavras por chamado e percentil 95 de 91 palavras. Para evitar o vazamento de dados (*data leakage*), esse *dataset* foi imediatamente particionado pelo protocolo *repeated stratified holdout* (proporção 70/30, com 10 sementes pré-definidas distintas).

A segunda etapa foi diferenciada conforme o tipo de representação textual. Para os modelos baseados em representações estáticas (SVM-BoW, SVM-TF-IDF e SVM-Word2Vec), aplicaram-se a conversão para minúsculas e a tokenização, com remoção implícita de pontuação. Optou-se por preservar as *stopwords*, uma vez que termos funcionais podem conter informações relevantes para a classificação de chamados técnicos. Por exemplo, o termo "sem" diferencia chamados como "*notebook sem acesso à rede*" e "*notebook com acesso à rede*", cujas interpretações são substancialmente distintas. Os vetorizadores e o modelo Word2Vec foram ajustados exclusivamente no conjunto de treino, evitando o fenômeno conhecido como vazamento de dados (*data leakage*) [Kapoor e Narayanan 2023]. Para os modelos baseados em representações contextuais (BERTimbau *Feature-Based* e *Fine-Tuned*), adotou-se pré-processamento mínimo, uma vez que o tokenizador pré-treinado do HuggingFace já realiza normalização textual, adição de *tokens* especiais e gerenciamento de truncamento e *padding*. Como resultado, a Figura 3 apresenta a distribuição de *tokens* para cada classe de grupos solucionadores.

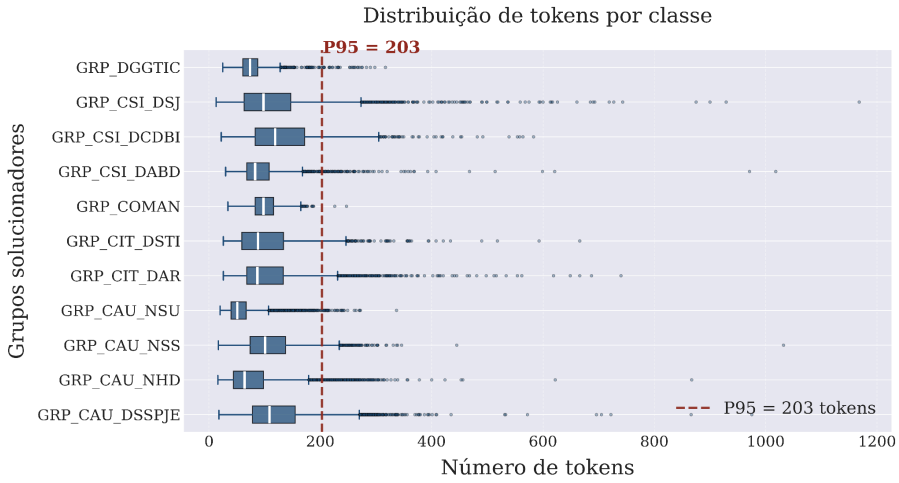

Figura 3. Distribuição de *tokens* por chamado.

A distribuição evidencia uma média de 92 *token*s e um percentil 95 de 203 *tokens*. Com base nesse valor, definiu-se o limite de truncamento de 200 *tokens*, assegurando a preservação do conteúdo com o menor custo computacional possível. Para investigar o impacto da informação contextual na classificação, todos os modelos foram avaliados em duas configurações de entrada de texto. A **Configuração I** utiliza exclusivamente atributos não estruturados ('*Título*' + '*Descrição*'), o que representa o cenário de classificação autônoma baseado apenas no relato textual do usuário. A **Configuração II** incorpora o campo '*Serviço*', resultando na concatenação '*Título*' + '*Serviço*' + '*Descrição*', o que representa um cenário de classificação assistida decorrente do enriquecimento dos dados a partir do catálogo de serviços do GLPI. Com os dados preparados, a seção a seguir descreve os modelos utilizados na classificação.

3.4. Modelagem

A fase de Modelagem consistiu na implementação e comparação de cinco abordagens, organizadas em três categorias funcionais (Tabela 2), para avaliar o impacto de diferentes estratégias de representação textual na classificação de chamados ITSM.

Tabela 2. Modelos avaliados e suas características.

Categoria	Modelo	Representação Textual	Hiperparâmetros Principais
Baselines clássicos	SVM-BoW	Vetores esparsos de contagem (BoW)	5.000 atributos; n-gramas (1,2); min_df=2; SVM linear, C=1,0
	SVM-TF-IDF	Vetores esparsos ponderados (TF-IDF)	5.000 atributos; n-gramas (1,2); tf sublinear; SVM linear, C=1,0
	SVM-Word2Vec	Vetores densos Word2Vec	200 dim.; Skip-gram; janela=5; min_count=2; 50 épocas; SVM linear, C=1,0
BERT *feature-based*	BERT-SVM	Contextual do BERTimbau	Truncamento=200 *tokens*; SVM linear, C=1,0
BERT *fine-tuned*	ITSM-BERT	Contextual do BERTimbau com ajuste fino	Truncamento=200 *tokens*; AdamW, lr=2e-5; *weight_decay*=0,01; 3 épocas; *batch*=16; aquecimento=500 passos; FP16

A primeira categoria compreende representações vetoriais clássicas, esparsas (BoW e TF-IDF) e densas (Word2Vec), combinadas com SVM, adotado como *baseline* por sua superioridade no domínio de ITSM, conforme [Paramesh e Shreedhara 2019, Dlodlo e Sibanda 2023]. A segunda abordagem emprega o BERTimbau como extrator

de características (*feature extractor*) para gerar representações contextuais, sobre as quais o classificador SVM é ajustado, estratégia proposta por Ayyalasomayajula et al. (2024), na qual os pesos do modelo permanecem fixos enquanto apenas o classificador é treinado a partir dos *embeddings* extraídos. A terceira etapa realiza o ajuste fino do BERTimbau, no qual o modelo é inicializado com pesos pré-treinados e acrescido de uma camada de classificação linear conectada à representação do *token* [CLS] [Liu et al. 2023]. Com o ajuste fino, todos os parâmetros da camada de classificação e das camadas internas do *Transformer* são atualizados conjuntamente com base em dados rotulados, permitindo que o modelo adapte suas representações contextuais às especificidades terminológicas e semânticas do domínio ITSM [Devlin et al. 2019].

Cada um dos cinco modelos foi treinado e testado em cada uma das 10 partições do *holdout* estratificado e nas duas configurações de entrada textual, resultando em 100 execuções independentes (5 modelos × 10 sementes × 2 configurações de entrada). Os modelos *baseline* utilizaram SVM com *kernel* linear (C=1.0). Para as representações vetoriais, configuraram-se o *Bag of Words* e o TF-IDF, com 5.000 *features*, *n-grams* (1, 2), frequência mínima de 2 e máxima de 95%. O Word2Vec empregou vetores de 200 dimensões, janela de 5 palavras e arquitetura *Skip-gram*, representando os documentos pela média dos vetores de palavras. O modelo BERT-SVM utilizou o BERTimbau (*bert-base-portuguese-cased*) como extrator de características em modo congelado, empregando a representação do *token* [CLS] (768 dimensões) como entrada para o classificador SVM. O modelo ITSM-BERT realizou ajuste fino completo com *learning rate* de 2×10^{-5}, *batch size* de 16, 3 épocas, comprimento máximo de 200 *tokens*, *weight decay* de 0,01 e precisão mista (FP16).

Para mitigar o desbalanceamento entre as classes, aplicou-se a todos os modelos a técnica *Class Balanced Loss* [Cui et al. 2019], que pondera a função de perda com base no número efetivo de amostras por classe. O parâmetro $\beta = 0{,}999$ foi definido para atribuir pesos inversamente proporcionais às frequências de cada classe. No ITSM-BERT, os pesos foram incorporados à função de perda (*cross-entropy* ponderada); nos modelos SVM, foram aplicados como *sample weights* durante o treinamento, favorecendo categorias sub-representadas sem recorrer a técnicas de *oversampling* ou *undersampling*. Com os modelos treinados e testados, a seção a seguir detalha os procedimentos adotados para avaliar o desempenho.

3.5. Avaliação

A fase de avaliação foi conduzida sob um desenho experimental controlado, no qual todos os modelos foram submetidos às duas configurações de entrada textual definidas na Seção 3.3, mantendo-se invariantes os hiperparâmetros, o particionamento dos dados e as métricas adotadas. Essa fase estrutura-se em duas dimensões complementares: a avaliação escalar, baseada em métricas de classificação multiclasse, e a avaliação estatística, voltada a verificar se as diferenças de desempenho observadas entre os modelos são estatisticamente significativas e não atribuíveis ao acaso.

Para a avaliação escalar, adotaram-se as métricas padronizadas de classificação multiclasse: acurácia, precisão, *recall* e F1-macro. Dado o cenário de classes desbalanceadas (IR=85,05), adotou-se o F1-macro como métrica principal. Com relação à avaliação estatística, utilizou-se um protocolo não paramétrico robusto à ausência de

normalidade [Demšar 2006]. Aplicou-se o teste de Friedman (α=0,05) para verificar diferenças globais entre os ranqueamentos dos modelos. Rejeitada a hipótese nula, procedeu-se ao teste *post-hoc* de Nemenyi para identificar pares com diferenças estatisticamente significativas. Os experimentos foram conduzidos com o método *repeated stratified holdout*, proporção 70/30, utilizando 10 sementes pré-definidas distintas, a fim de permitir o controle da variabilidade estocástica, condição essencial para os testes estatísticos. Essa estratégia, adotada em detrimento do *k-fold* pelo custo computacional dos *Transformers*, fundamenta-se em Bouckaert e Frank (2004) que validam esse método e demonstram que a repetição do *holdout* estima a variância e aumenta a replicabilidade dos testes.

4. Resultados e Discussões

Os resultados dos cinco modelos nas duas configurações de entrada são apresentados na Tabela 3, expressos como média ± desvio padrão das 10 execuções independentes. Os valores em negrito indicam os melhores desempenhos.

Tabela 3. Resultados comparativos dos modelos por configuração de entrada textual.

Configuração	Modelo	Acurácia	Precisão	*Recall*	*F1-macro*
(I) Título + Descrição	SVM-BoW	0,805±0,002	0,785±0,003	0,775±0,002	0,779±0,002
	SVM-TF-IDF	0,832±0,002	0,826±0,003	0,802±0,002	0,811±0,002
	SVM-Word2Vec	0,782±0,002	0,784±0,003	0,737±0,003	0,750±0,003
	BERT-SVM	0,769±0,002	0,756±0,002	0,715±0,002	0,727±0,003
	ITSM-BERT	**0,858±0,004**	**0,851±0,010**	**0,835±0,006**	**0,840±0,005**
(II) Título + Descrição + Serviço	SVM-BoW	0,983±0,001	0,977±0,001	0,975±0,002	0,976±0,001
	SVM-TF-IDF	0,986±0,001	0,982±0,001	0,979±0,002	0,980±0,001
	SVM-Word2Vec	0,926±0,001	0,923±0,001	0,903±0,002	0,912±0,001
	BERT-SVM	0,911±0,002	0,893±0,003	0,874±0,002	0,883±0,003
	ITSM-BERT	**0,998±0,001**	**0,997±0,002**	**0,997±0,001**	**0,997±0,001**

Os resultados evidenciam diferenças significativas entre as duas configurações de entrada. Na **Configuração I**, o ITSM-BERT obteve o melhor F1-macro (0,840), seguido pelo SVM-TF-IDF (0,811), enquanto os demais modelos apresentaram desempenho inferior, com destaque negativo para o BERT-SVM (0,727), que, mesmo utilizando o BERTimbau, ficou abaixo dos *baselines* clássicos por operar sem ajuste fino. Na **Configuração II**, a incorporação do campo '*Serviço*' elevou substancialmente o desempenho de todos os modelos: o ITSM-BERT alcançou 0,997, enquanto SVM-TF-IDF e SVM-BoW atingiram 0,980 e 0,976, respectivamente. Dado que a proximidade numérica entre esses modelos não permite afirmar superioridade sem validação formal, aplicou-se o protocolo estatístico da Seção 3.5, cujos resultados são sintetizados no Diagrama de Diferença Crítica (Figura 4).

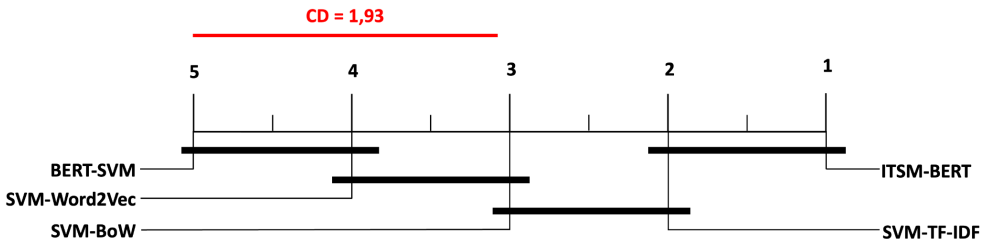

Figura 4. Diagrama da Diferença Crítica para a Configuração II.

O teste de Friedman indicou diferença significativa entre os cinco modelos (χ^2 = 40, p < 0,05), e o *post-hoc* de Nemenyi estabeleceu *Critical Difference* (CD) de 1,93. Em caráter exploratório, a análise da estabilidade dos *rankings* revelou o ITSM-BERT em primeiro lugar (*rank* = 1,0), seguido pelo SVM-TF-IDF (*rank* = 2,0). Como a diferença entre ambos é inferior à CD (Δ = 1,0 < 1,93), os dois modelos são estatisticamente equivalentes. O ITSM-BERT supera significativamente SVM-BoW (*rank* = 3,0), SVM-Word2Vec (*rank* = 4,0) e BERT-SVM (*rank* = 5,0), enquanto o SVM-TF-IDF (*rank* = 2,0) não difere estatisticamente do SVM-BoW. No diagrama, modelos conectados por segmentos horizontais não apresentam diferença estatisticamente significativa entre si, e os de melhor desempenho posicionam-se à direita, por apresentarem os menores *ranks* médios [Demšar 2006]. As subseções seguintes discutem os resultados obtidos em resposta às questões de pesquisa.

4.1. QP1 — Ganhos de Desempenho com Ajuste Fino

Na **Configuração I**, o ITSM-BERT consolidou-se como a abordagem mais eficaz, com F1-macro de 0,840, superando o SVM-TF-IDF (0,811) em 2,9 p.p. e o BERT-SVM (0,727) em 11,3 p.p. O BERT-SVM, mesmo utilizando o BERTimbau, ficou abaixo do SVM-BoW (0,779), evidenciando que representações contextuais, sem especialização ao domínio, são insuficientes para superar abordagens clássicas bem calibradas [Devlin et al. 2019]. O desempenho superior do ITSM-BERT é consistente com os achados do Ticket-BERT [Liu et al. 2023], que atingiu um F1-score de 85,90% com dados de entrada puramente textuais, reforçando a eficácia do ajuste fino do BERT no domínio do ITSM. Na **Configuração II**, o ITSM-BERT permaneceu como o modelo de melhor desempenho absoluto, alcançando um F1-macro de 0,997, confirmando que o ajuste fino ao domínio preserva sua relevância mesmo na presença de dados enriquecidos. Contudo, o teste de Nemenyi não rejeitou a hipótese nula de equivalência entre o ITSM-BERT e o SVM-TF-IDF (0,980), indicando que a diferença observada entre os dois modelos não é estatisticamente significativa ao nível de significância adotado (α = 0,05). Na prática, o SVM-TF-IDF emerge como uma alternativa computacionalmente eficiente ao ITSM-BERT, oferecendo desempenho estatisticamente equivalente, a um custo de complexidade computacional menor, no cenário de classificação assistida.

4.2. QP2 — Impacto do Enriquecimento de Dados com o Catálogo de Serviços

A incorporação do campo '*Serviço*' resultou no maior incremento de desempenho deste estudo, com ganhos consistentes em todos os modelos na Configuração II: ITSM-BERT (+15,7 p.p.), SVM-TF-IDF (+16,9 p.p.), SVM-BoW (+19,7 p.p.) e SVM-Word2Vec (+16,2 p.p.). O comportamento uniforme, independentemente do modelo, indica que o ganho foi atribuível à qualidade da informação adicionada, corroborando a premissa de que o enriquecimento de dados é uma estratégia de amplificação do sinal discriminativo [Barchilon et al. 2024]. Esse resultado pode ser compreendido à luz da ambiguidade semântica característica dos chamados ITSM. O campo '*Serviço*', ao fornecer uma âncora taxonômica pré-estruturada, reduz substancialmente a ambiguidade semântica, sem eliminá-la por completo, pois seu preenchimento está sujeito a imprecisões, variações terminológicas e inconsistências decorrentes da evolução do catálogo. Esse achado é consistente com o Ticket-BERT [Liu et al. 2023], que atingiu um F1-score de 98,24% ao incorporar contexto auxiliar à entrada.

Os resultados das duas questões de pesquisa convergem para uma conclusão prática: a classificação automática de chamados de ITSM em português brasileiro pode ser viabilizada em dois estágios de maturidade organizacional. No estágio inicial, com catálogo inexistente ou pouco confiável, o ITSM-BERT demonstra-se a abordagem mais robusta, evidenciando que a especialização supervisionada é condição necessária para superar representações estáticas e extratores de características genéricas. No estágio avançado, com catálogo consolidado, o enriquecimento de dados emerge como a estratégia de maior impacto, permitindo que modelos clássicos, como o SVM-TF-IDF, alcancem desempenho estatisticamente equivalente ao do ITSM-BERT. Contudo, os ganhos do enriquecimento dependem da consistência taxonômica do catálogo.

5. Conclusões e Trabalhos Futuros

Este estudo investigou a classificação automática de chamados de *service desk* em português brasileiro por meio do ITSM-BERT, um modelo obtido por ajuste fino do BERTimbau ao domínio ITSM, com enriquecimento de dados provenientes do catálogo de serviços. Os resultados respondem às questões de pesquisa formuladas: na configuração autônoma, o ITSM-BERT superou significativamente os *baseline*s clássicos e o BERT-SVM, alcançando F1-macro de 0,840 (QP1); na configuração assistida, a incorporação do campo '*Serviço*' elevou o desempenho para 0,997 (+15,7 p.p.), evidenciando que o enriquecimento de dados ampliou o poder discriminativo dos classificadores (QP2). A equivalência estatística observada na configuração assistida indica que, com um catálogo bem estruturado, o SVM-TF-IDF apresenta desempenho comparável ao do ITSM-BERT, com menor complexidade computacional.

No geral, esses resultados validam as contribuições propostas: a combinação entre ajuste fino supervisionado de domínio e enriquecimento semântico com dados do catálogo de serviços constitui uma estratégia promissora para a classificação automática de chamados de ITSM em português brasileiro, sustentada por uma validação estatística. Como limitações, não foram avaliados o impacto do tratamento de desbalanceamento, o desempenho das classes minoritárias, nem as métricas por classe. Além disso, a comparação entre ITSM-BERT e BERT-SVM não permite isolar os efeitos do ajuste fino. Quanto à validade externa, também não foram investigados cenários de transferência entre instituições nem os efeitos da deriva temporal dos dados, o que pode limitar a generalização dos resultados. Diante disso, trabalhos futuros incluem validar a abordagem em outros tribunais e organizações com diferentes catálogos de serviços, investigar o impacto do desbalanceamento e das classes minoritárias, realizar estudos de ablação para isolar os efeitos do ajuste fino, avaliar o uso de *Large Language Models* (LLMs) com mecanismos de anonimização e analisar os custos computacionais.

Agradecimentos e Uso de IA generativa

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - DT-303031/2023-9, pela Financiadora de Estudos e Projetos (FINEP) Pró-Amazônia Nº 2373/2 e pelos Acordos de Cooperação Técnica: Nº 02/2021 (Proc. Nº 38328/2020 - TJMA) e Nº 0056/2025 (Proc. Nº 62666/2024 - TJMA). Declara-se que os modelos de IA generativa Claude e ChatGPT foram utilizados exclusivamente como apoio ao aprimoramento de código e à revisão linguística.

Referências

Alsaç, A., Yılmaz, Ü., Koçoğlu, F. Ö., and Şeker, Ş. E. (2025). Towards intelligent IT service management: A comparative evaluation of machine learning and language models for ticket classification. In *UBMK*, pages 265–270. IEEE.

Ayyalasomayajula, M. M. T., Bussa, S., Kowsalya, S. S. N., Mishra, N., Prasad, S., and Mehra, A. (2024). Leveraging feature extraction and fine-tuning techniques for enhanced detection of fake news using BERT. In *ICAC2N*, pages 1630–1635. IEEE.

Axelos (2019). ITIL Foundation. The Stationery Office, 4th edition.

Barchilon, N., Lopes, H. C. V., Kalinowski, M., and Perez, J. S. (2024). Enriquecimento de dados com base em estatísticas de grafo de similaridade para melhorar o desempenho em modelos de ML supervisionados de classificação. In *Anais do XXXIX Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 220–233. SBC.

Boonklay, J. and Jinarat, S. (2025). Comparative study of machine learning and deep learning algorithms for customer support ticket classification. In *InCIT*, pages 813–819. IEEE.

Bouckaert, R. R. and Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In *PAKDD*, pages 3–12. Springer.

Brandão, M. A., Silva, M. O., Oliveira, G. P., Hott, H. R., Lacerda, A. M., and Pappa, G. L. (2023). Impacto do pré-processamento e representação textual na classificação de documentos de licitações. In *Anais do XXXVIII Simpósio Brasileiro de Banco de Dados (SBBD),* pages 102–114. SBC.

Carmo, F., Serejo, F., Jacob Junior, A., Santana, E., and Lobato, F. (2023). Embeddings jurídico: Representações orientadas à linguagem jurídica brasileira. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, pages 188–199. SBC.

Chi, W. W., Tang, T. Y., Salleh, N. M., Mukred, M., AlSalman, H., and Zohaib, M. (2024). Data augmentation with semantic enrichment for deep learning invoice text classification. *IEEE Access*, 12:57326–57344.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277. IEEE.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Dlodlo, D. and Sibanda, K. (2023). Automated ticket classification for information technology helpdesks using machine learning. In *ZCICT*, pages 1–7. IEEE.

Ferdinand, Y., Lubis, M., and Pratiwi, O. N. (2025). A systematic literature review on AI and NLP applications for customer support automation and digital service. *International Journal of Computer Technology and Science*, 2(4):01–14.

Fuchs, S., Schnellbach, J., Wittges, H., and Krcmar, H. (2026). Human vs. automated data annotation: Labeling the data set for an ML-driven support ticket classifier. Data & Knowledge Engineering, https://doi.org/10.1016/j.datak.2025.102534.

ISACA (2019). COBIT 2019 Framework: Governance and Management Objectives. ISACA.

Jain, S., Gupta, A., and Neha, K. (2025). AI enhanced ticket management system for optimized support. In *AIMLSystems*, pages 1–7. ACM.

Kapoor, S. and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804.

Liu, F., He, X., Zhang, T., Chen, J., Li, Y., Yi, L., Zhang, H., Wu, G., and Shi, R. (2025). TickIt: Leveraging large language models for automated ticket escalation. In *FSE*, pages 343–354. ACM.

Liu, Z., Benge, C., and Jiang, S. (2023). Ticket-BERT: Labeling incident management tickets with language models. *arXiv* preprint arXiv:2307.00108.

Paket, E., Şenerkek, G., Akyol, F. B., and Salman, F. (2024). IT service desk ticket classification via large language models. In *UBMK*, pages 135–140. IEEE.

Paramesh, S. P. and Shreedhara, K. S. (2019). Automated IT service desk systems using machine learning techniques. In *Data Analytics and Learning*, Lecture Notes in Networks and Systems, volume 43, pages 331–346. Springer.

Sant'Anna, Y. F. D., Souza, L. F. C., Alves Neto, A. J., Rodrigues Junior, M. C., Carvalho, A. B., and Gusmão, R. P. (2024). Classificação da dívida ativa do estado de Sergipe. In *Anais do XXXIX Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 562–573. SBC.

Silva, M. O., Oliveira, G. P., Costa, L. G. L., and Pappa, G. L. (2024). Evaluating domain-adapted language models for governmental text classification tasks in Portuguese. In *Anais do XXXIX Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 247–259. SBC.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *BRACIS*, pages 403–417. Springer.

Souza, F. and Souza Filho, J. B. O. (2023). Embedding generation for text classification of Brazilian Portuguese user reviews: From bag-of-words to transformers. *Neural Computing and Applications*, 35(13):9393–9406.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*, volume 30, pages 5998–6008.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, volume 1, pages 29–39.