ENGENHARIAS, TECNOLOGIA, GESTÃO E TURISMO
ENGINEERING, TECHNOLOGY, MANAGEMENT AND TOURISM
INGENIERÍA, TECNOLOGÍA, ADMINISTRACIÓN Y TURISMO

millenium

Millenium, 2(Edição Especial Nº23)

pt

PROCESSAMENTO DE LINGUAGEM NATURAL E MACHINE LEARNING NA CATEGORIZAÇÃO DE NOTÍCIAS
NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING IN NEWS CATEGORIZATION
PROCESAMIENTO DEL LENGUAJE NATURAL Y APRENDIZAJE AUTOMÁTICO EN LA CATEGORIZACIÓN DE NOTICIAS

José Varanda[1]
Cristina Lacerda[1,2] *https://orcid.org/0000-0002-8921-4747*
Joana Fialho[1,3] *https://orcid.org/0000-0002-3910-8292*

[1] Instituto Politécnico de Viseu, Viseu, Portugal
[2] CISeD - Centro de Investigação em Serviços Digitais, Viseu, Portugal
[3] CI&DEI – Centro de Estudos em Educação e Inovação, Viseu, Portugal

José Varanda – jvaranda00@outlook.pt | Cristina Lacerda - cwanzeller@estgv.ipv.pt | Joana Fialho - jfialho@estgv.ipv.pt

Autor Correspondente:
Joana Fialho
Campus Politécnico
3504-510 – Viseu
jfialho@estgv.ipv.pt

RECEBIDO: 27 de maio de 2026
REVISTO: 20 de junho de 2026
ACEITE: 29 de junho de 2026
PUBLICADO: 23 de julho de 2026

DOI: https://doi.org/10.29352/mill0223e.47447

RESUMO

Introdução: O rápido aumento de conteúdos digitais, especialmente na área de notícias, tornou mais difícil organizar e analisar grandes volumes de informação não estruturada. A classificação automática de textos é uma solução importante, apoiada pelos avanços em Processamento de Linguagem Natural e *Machine Learning*, que tornam a categorização de notícias mais eficiente e precisa.
Objetivo: Desenvolver e avaliar métodos de categorização automática de notícias em português que combinem bom desempenho com eficiência computacional, comparando técnicas tradicionais a modelos avançados de aprendizagem profunda.
Métodos: Realizou-se uma investigação experimental, com base em revisão da literatura, análise exploratória de dados e na implementação de diferentes modelos de classificação. Foram comparadas abordagens clássicas de *Machine Learning* e de Mineração de Texto com modelos recentes de Processamento de Linguagem Natural, incluindo arquiteturas baseadas em *Transformers*, como o BERT. A avaliação foi efetuada com base em métricas de desempenho (*Accuracy*, *Precision*, *Recall*, *F1-Score*) e no custo computacional.
Resultados: Os modelos baseados em *Transformers* demonstraram maior capacidade de compreensão contextual e desempenho superior na classificação de notícias. No entanto, apresentam requisitos computacionais mais elevados. Em contrapartida, modelos mais leves evidenciam um melhor equilíbrio entre eficiência e precisão, revelando-se adequados para cenários com recursos computacionais limitados.
Conclusão: Soluções eficazes para a categorização automática de notícias em português podem equilibrar desempenho e eficiência computacional. O estudo ressalta a importância de selecionar o modelo adequado ao contexto, o que favorece o desenvolvimento de sistemas de classificação mais robustos e acessíveis.

Palavras-chave: processamento de linguagem natural; *machine learning*; classificação de texto; categorização de notícias; BERT; inteligência artificial

ABSTRACT

Introduction: The rapid growth of digital content, especially in the news domain, has made more difficult to organize and analyze large volumes of unstructured information. Automatic text classification is an important solution, supported by advances in Natural Language Processing and Machine Learning, which make news categorization more efficient and accurate.
Objective: To develop and evaluate methods for automatic categorization of news in Portuguese that combine strong performance with computational efficiency, comparing traditional techniques with advanced deep learning models.
Methods: An experimental study was conducted, based on literature review, exploratory data analysis, and the implementation of different classification models. Classical Machine Learning and Text Mining approaches were compared with recent Natural Language Processing models, including Transformer-based architectures such as BERT. The evaluation was carried out using performance metrics (Accuracy, Precision, Recall, F1-Score) and computational cost.
Results: Transformer-based models demonstrated greater contextual understanding and superior performance in news classification. However, they require higher computational resources. In contrast, lighter models showed a better balance between efficiency and accuracy, proving suitable for scenarios with limited computational resources.
Conclusion: Effective solutions for automatic news categorization in Portuguese can balance performance and computational efficiency. The study highlights the importance of selecting the appropriate model for the context, supporting the development of more robust and accessible classification systems.

Keywords: natural language processing; machine learning; text classification; news categorization; BERT; artificial intelligence

RESUMEN

Introducción: El rápido aumento del contenido digital, especialmente en el ámbito informativo, ha dificultado la organización y el análisis de grandes volúmenes de información no estructurada. La clasificación automática de texto es una solución importante, respaldada por los avances en el procesamiento del lenguaje natural y el aprendizaje automático, que hacen que la categorización de notícias sea más eficiente y precisa.
Objetivo: Desarrollar y evaluar métodos de categorización automática de noticias en portugués que combinen un buen rendimiento con eficiencia computacional, comparando técnicas tradicionales con modelos avanzados de aprendizaje profundo.
Métodos: Se llevó a cabo una investigación experimental basada en la revisión de la literatura, el análisis exploratorio de datos y la implementación de diferentes modelos de clasificación. Se compararon enfoques clásicos de aprendizaje automático y minería de texto con modelos recientes de procesamiento del lenguaje natural, incluyendo arquitecturas basadas en Transformer como BERT. La evaluación se realizó en función de métricas de rendimiento (precisión, exactitud, exhaustividad, puntuación F1) y coste computacional.
Resultados: Los modelos basados en Transformer demostraron una mayor comprensión contextual y un rendimiento superior en la clasificación de noticias. Sin embargo, presentan mayores requisitos computacionales. En contraste, los modelos más ligeros muestran un mejor equilibrio entre eficiencia y precisión, demostrando ser adecuados para escenarios con recursos computacionales limitados.
Conclusión: Las soluciones eficaces para la categorización automática de notícias en portugués logran un equilibrio entre rendimiento y eficiencia computacional. El estudio subraya la importancia de seleccionar el modelo adecuado para cada contexto, lo que favorece el desarrollo de sistemas de clasificación más robustos y accesibles.

Palabras clave: procesamiento del lenguaje natural; aprendizaje automático; clasificación de texto; categorización de notícias; BERT; inteligencia artificial

DOI: https://doi.org/10.29352/mill0223e.47447

INTRODUÇÃO

O rápido aumento da informação digital, impulsionado pelo crescimento de plataformas *online*, redes sociais e outros meios digitais, gera grandes volumes de texto, especialmente em notícias, tornando mais difícil organizar, filtrar e aceder a informações relevantes, o que evidencia a necessidade de soluções automáticas para lidar com dados não estruturados (Zhang & Wallace, 2017; Rezaeenour et al., 2023).

Neste cenário, o Processamento de Linguagem Natural (PLN) e o *Machine Learning* (ML) são fundamentais para transformar textos em dados estruturados para análise automática. Estas tecnologias ajudam a extrair conhecimento a partir de grandes volumes de texto e são usadas em tarefas como análise de sentimentos, sistemas de recomendação e, especialmente, na classificação automática de notícias (Rezaeenour et al., 2023; Lauriola et al., 2022).

Entre estas tarefas, a classificação de texto é uma das principais funções do PLN, pois atribui automaticamente categorias pré-definidas a documentos. No caso das notícias, esta tarefa ajuda a organizar conteúdos em áreas fulcrais como a política, a economia ou o desporto, facilitando a gestão da informação e a tomada de decisões em ambientes com grande volume e rapidez na produção de conteúdo (Gupta, 2024).

A classificação de texto baseia-se tradicionalmente em modelos clássicos de ML, como o *Naive Bayes*, o *Support Vector Machine* (SVM) e a Regressão Logística, geralmente com representações vetoriais como o *Term Frequency-Inverse Document Frequency* (TF-IDF). Embora sejam eficientes e fáceis de implementar, estes métodos têm limitações para compreender o contexto e o significado da linguagem, pois tratam o texto de forma estatística, sem captar adequadamente o contexto ou o significado semântico das palavras (Taha et al., 2024; Sun et al., 2024).

Com o avanço das técnicas de Aprendizagem Profunda, surgiram modelos mais sofisticados, capazes de capturar relações semânticas complexas e dependências contextuais no texto. As arquiteturas baseadas em redes neuronais, como as *Long Short-Term Memory* (LSTM) e as *Convolutional Neural Networks* (CNN), introduziram melhorias relevantes na modelação de sequências textuais, permitindo considerar a ordem das palavras e padrões linguísticos locais (Sarikaya et al., 2024).

Mais recentemente, a introdução da arquitetura *Transformer* representou um avanço significativo no domínio do PLN, ao permitir o processamento paralelo de sequências e a modelação eficiente de dependências de longo alcance. Os modelos baseados nesta arquitetura, como o BERT (*Bidirectional Encoder Representations from Transformers*), revolucionaram a classificação de texto ao introduzirem representações contextuais bidirecionais, nas quais o significado das palavras é determinado em função do contexto global (Vaswani et al., 2017; Devlin et al., 2019).

Apesar do seu elevado desempenho, os modelos baseados em *Transformers* apresentam requisitos computacionais significativos, nomeadamente em termos de memória e tempo de processamento, o que pode limitar a sua aplicação em cenários com recursos limitados. Neste sentido, têm sido propostas variantes mais eficientes, como o *DistilBERT*, que procuram manter níveis elevados de desempenho enquanto reduzem o custo computacional (Sanh et al., 2020).

Neste enquadramento, torna-se essencial analisar comparativamente as diferentes abordagens de classificação de texto, considerando não apenas o desempenho preditivo, mas também o compromisso entre precisão e eficiência computacional. Esta análise é particularmente relevante no contexto da língua portuguesa, em que a disponibilidade de recursos e de estudos comparativos ainda é mais limitada do que a de outras línguas de maior difusão.

Desta forma, o presente artigo tem como objetivo contribuir para o desenvolvimento de soluções eficazes de categorização automática de notícias em língua portuguesa, por meio da análise e da comparação de diferentes abordagens de PLN e de ML. A investigação é motivada por lacunas identificadas na literatura, nomeadamente a escassez de estudos comparativos abrangentes para a língua portuguesa, a falta de padronização metodológica e a necessidade de analisar o equilíbrio entre o desempenho preditivo e o custo computacional dos modelos. Neste contexto, procura-se avaliar o desempenho e a eficiência destas abordagens em tarefas de categorização textual, orientando o estudo pela seguinte questão de investigação: "Que abordagens de PLN e ML apresentam melhor desempenho e eficiência na categorização automática de notícias em língua portuguesa?". Pretende-se, desta forma, identificar estratégias que permitam conciliar um elevado desempenho com a eficiência computacional, promovendo o desenvolvimento de sistemas mais robustos, acessíveis e aplicados a contextos reais.

1. ESTADO DE ARTE

A classificação automática de texto tem evoluído significativamente nas últimas décadas, acompanhando os avanços nas áreas do PLN e do ML. Esta evolução reflete uma transição progressiva de abordagens baseadas em regras e métodos estatísticos para modelos progressivamente mais complexos, contextuais e semanticamente robustos (Jurafsky & Martin, 2026; Aggarwal, 2018).

Os primeiros modelos de classificação de texto baseavam-se em técnicas estatísticas que representavam os documentos como vetores de frequência de termos. O modelo *Bag-of-Words* (BoW) e o *Term Frequency-Inverse Document Frequency* (TF-IDF) foram amplamente utilizados devido à sua simplicidade e eficiência computacional. No entanto, estas representações apresentam limitações significativas, uma vez que ignoram a ordem das palavras e não captam relações semânticas entre termos (Manning et al., 2008).

Os algoritmos clássicos de Machine Learning, como o *Naive Bayes*, o *Support Vector Machine* (SVM) e a Regressão Logística, têm

DOI: https://doi.org/10.29352/mill0223e.47447

sido aplicados com sucesso em tarefas de classificação de texto, especialmente quando combinados com representações TF-IDF. Estes modelos destacam-se pela sua simplicidade, rapidez e baixo custo computacional, sendo particularmente eficazes em cenários com dados bem estruturados e categorias claramente definidas (Aggarwal, 2018). No entanto, a sua principal limitação reside na incapacidade de captar relações semânticas profundas e dependências contextuais entre palavras, o que compromete o desempenho em textos mais complexos ou ambíguos (Manning et al., 2008).

Com o avanço das técnicas de Aprendizagem Profunda, surgiram modelos baseados em redes neuronais capazes de aprender representações mais ricas a partir dos dados textuais. As RNN e, particularmente, as suas variantes LSTM e as *Gated Recurrent Unit* (GRU), introduziram a capacidade de modelar dependências sequenciais e contextuais de médio e longo alcance em textos (Hochreiter & Schmidhuber, 1997). Paralelamente, as CNN demonstraram eficácia na identificação de padrões locais, como n-gramas e expressões características de categorias específicas (Sarikaya et al., 2024). Apesar destas vantagens, estes modelos ainda apresentam limitações na captação de dependências de longo alcance e exigem maior capacidade computacional e mais tempo de treino quando comparados com abordagens mais recentes.

O avanço mais significativo na classificação textual ocorreu com a introdução da arquitetura *Transformer*, proposta por Vaswani et al. (2017). Esta arquitetura utiliza o mecanismo de atenção para permitir a contextualização dinâmica e bidirecional de cada termo no texto, substituindo o paradigma recursivo das RNN e LSTM. Ao contrário das abordagens anteriores, o *Transformer* processa a sequência de forma paralela e atribui pesos relacionais entre todas as palavras, permitindo captar dependências de longo alcance de forma mais eficiente (Vaswani et al., 2017).

Os modelos baseados na arquitetura *Transformer*, com destaque para o BERT, desenvolvido por Devlin et al. (2019), revolucionaram o domínio do PLN. O BERT distingue-se pela capacidade de compreender o contexto das palavras de forma bidirecional, utilizando *embeddings* contextuais que produzem representações específicas para cada ocorrência de uma palavra numa sequência textual. O seu pré-treino baseia-se em duas tarefas principais: o *Masked Language Modeling* (MLM), que oculta palavras aleatoriamente para que o modelo as preveja com base no contexto, e o *Next Sentence Prediction* (NSP), que permite ao modelo aprender a identificar se duas frases são consecutivas no texto original (Devlin et al., 2019). Esta abordagem permite captar nuances linguísticas, ambiguidade semântica e relações complexas entre termos, aspetos particularmente relevantes no domínio das notícias, onde a linguagem pode ser diversa e contextual.

No contexto da língua portuguesa, foram desenvolvidas variantes específicas destes modelos, como o *BERTimbau*. Este modelo foi pré-treinado em grandes volumes de texto em português, permitindo capturar particularidades linguísticas da língua e alcançar melhorias significativas de desempenho em tarefas de classificação textual quando comparado com modelos multilingues genéricos.

Apesar do seu elevado desempenho, os modelos baseados em Transformers apresentam custos computacionais elevados, tanto no treino como na inferência, o que pode limitar a sua aplicação em sistemas com restrições de recursos. Neste contexto, têm sido propostas variantes otimizadas, como o *DistilBERT*, desenvolvido por Sanh et al. (2020) através da técnica de destilação de conhecimento. Esta abordagem permite reduzir significativamente o número de parâmetros (cerca de 40% menos) e o tempo de inferência (cerca de 60% mais rápido), mantendo cerca de 97% do desempenho do modelo original. O *DistilBERT* surge, assim, como uma alternativa viável que permite um compromisso entre o desempenho e a eficiência computacional (Sanh et al., 2020).

Para sintetizar as principais características das abordagens de classificação de texto descritas, a Tabela 1 apresenta uma comparação entre os modelos clássicos de ML, as arquiteturas de Aprendizagem Profunda e os modelos baseados em *Transformers*, destacando as respetivas vantagens e limitações no contexto da categorização automática de notícias.

Tabela 1 – Comparação entre abordagens de classificação de texto

Abordagem	Tipo	Vantagens	Limitações
Clássicos (SVM, NB)	ML Tradicional	Baixo custo; rápido; eficaz em tarefas de classificação textual simples (Aggarwal, 2018).	Sem contexto; manual (Manning et al., 2008)
TF-IDF + ML	Vetorização	Simples; eficaz em textos básicos.	Sem semântica/contexto.
LSTM / CNN	Deep Learning	Padrões sequenciais/locais; capta dependências contextuais (Hochreiter & Schmidhuber, 1997; Sarikaya et al., 2024).	Custo moderado; complexo.
Transformers	BERT	Alta precisão; compreensão contextual bidirecional (Devlin et al., 2019).	Alto custo de *hardware (Chitty-Venkata et al., 2022).*
Otimizados	*DistilBERT*	Eficiência vs. Performance; 40% menos parâmetros (Sanh et al., 2020).	Pequena perda de precisão.

A análise da literatura evidencia uma progressão clara no desempenho dos modelos clássicos em relação às arquiteturas Transformer. Contudo, observa-se a ausência de estudos que realizem comparações sistemáticas e abrangentes entre estes paradigmas no contexto da língua portuguesa, lacuna esta que motiva e fundamenta o desenvolvimento da presente investigação.

DOI: https://doi.org/10.29352/mill0223e.47447

2. MÉTODOS

O presente estudo adotou uma abordagem quantitativa de natureza experimental, centrada na análise comparativa de diferentes modelos de classificação automática de texto no domínio da categorização de notícias. A metodologia foi estruturada de modo a permitir a avaliação do desempenho e da eficiência computacional de abordagens distintas de PLN e de ML, garantindo o rigor e a reprodutibilidade dos resultados.

2.1. Questão de Pesquisa

A presente investigação centra-se na análise comparativa de diferentes abordagens de classificação automática de texto no domínio das notícias em língua portuguesa, tendo em consideração a evolução das técnicas de PLN e de ML, bem como a crescente utilização de modelos baseados em aprendizagem profunda. Neste contexto, procura-se avaliar o desempenho e a eficiência destas abordagens em tarefas de categorização textual, orientando o estudo pela seguinte questão de investigação: "Que abordagens de PLN e ML apresentam melhor desempenho e eficiência na categorização automática de notícias em língua portuguesa?". A formulação desta questão permite analisar e comparar diferentes modelos quanto à sua capacidade preditiva e aos recursos computacionais necessários, com vista à identificação de soluções eficazes aplicáveis a contextos reais.

2.2. Critérios de Elegibilidade

Foram definidos critérios de inclusão e exclusão com o objetivo de garantir a consistência e relevância dos dados utilizados no estudo. Como critérios de inclusão, consideraram-se conjuntos de dados constituídos por notícias em língua portuguesa, textos devidamente categorizados em classes temáticas, dados com qualidade textual adequada ao processamento automático e amostras representativas de diferentes domínios noticiosos. Por outro lado, como critérios de exclusão, foram considerados dados incompletos ou com ruído excessivo, textos não classificados ou com categorias ambíguas, conteúdos não noticiosos (como *blogs*, fóruns ou redes sociais) e dados redundantes ou duplicados. A Tabela 2 sintetiza os critérios definidos.

Tabela 2 – Critérios de inclusão e exclusão

Critérios de Inclusão	Critérios de Exclusão
Notícias em língua portuguesa	Textos não noticiosos
Dados com categorias definidas	Dados sem classificação
Qualidade textual adequada	Conteúdo com ruído elevado
Diversidade temática	Dados duplicados ou redundantes

2.3. Fontes, Informação e Estratégia de Pesquisa

Os dados utilizados no presente estudo foram obtidos a partir do conjunto de dados *CC News PT v2*, disponibilizado publicamente por Garcia (2021) no repositório *Hugging Face Datasets*. Este conjunto é composto por notícias em língua portuguesa, recolhidas a partir de fontes jornalísticas digitais, abrangendo um conjunto diversificado de domínios temáticos. Para este estudo, foi selecionada uma amostra controlada de cerca de 100.000 notícias, distribuídas por seis categorias temáticas: Política, Economia, Desporto, Cultura, Tecnologia e Outros. O texto de cada notícia resulta da concatenação do título e do corpo do artigo, de forma a preservar o máximo de contexto semântico relevante para a tarefa de classificação.

O conjunto de dados foi dividido em subconjuntos de treino, validação e teste, na proporção de 70%, 15% e 15%, respetivamente, com estratificação por classe para preservar a distribuição original das categorias. Para os modelos clássicos de ML, foi ainda aplicada uma estratégia de validação cruzada estratificada com 5 partições, de modo a garantir maior consistência estatística na avaliação dos algoritmos e reduzir o risco de sobreajuste.

A implementação experimental foi desenvolvida num sistema equipado com um processador *Intel Core i7-10750H*, 16 GB de RAM e uma GPU NVIDIA *GeForce* GTX 1650 Ti (4 GB de VRAM). Os modelos clássicos foram executados exclusivamente na CPU utilizando a biblioteca *scikit-learn*, enquanto os modelos de Aprendizagem Profunda e os baseados em arquiteturas *Transformer* utilizaram a aceleração da GPU na *framework PyTorch*.

Numa fase inicial, procedeu-se ao pré-processamento dos dados, com o objetivo de preparar o texto para a análise computacional. Este processo incluiu a limpeza dos dados, especialmente a remoção de caracteres especiais e a pontuação, a normalização do texto através da conversão para minúsculas, a remoção de palavras irrelevantes (*stopwords*) e a tokenização. Sempre que aplicável, foi ainda considerada a utilização de técnicas de lematização, de forma a reduzir as palavras às suas formas base.

Após o pré-processamento, os textos foram convertidos em representações numéricas adequadas aos modelos de Aprendizagem Automática. Para os modelos clássicos de ML, foi utilizada a representação TF-IDF, reconhecida pela sua eficiência em tarefas de classificação de texto. Por outro lado, para os modelos baseados em Aprendizagem Profunda, foram utilizadas representações distribuídas (*embeddings*), nomeadamente os *embeddings* contextuais no caso dos modelos baseados em arquiteturas *Transformer*.

Para a fundamentação teórica e identificação do estado da arte, foi realizada uma pesquisa bibliográfica em bases de dados

DOI: https://doi.org/10.29352/mill0223e.47447

científicas, nomeadamente o *Scopus*, o *Web of Science*, o *IEEE Xplore*, o *ACM Digital Library* e o *Google Scholar*. A pesquisa utilizou as seguintes palavras-chave: *Natural Language Processing, Machine Learning, text classification, news categorization* e *BERT*, combinadas com o descritor *Portuguese* sempre que aplicável. Foram considerados artigos publicados entre 2017 e 2026, nos idiomas de inglês e português. Os critérios de inclusão foram os seguintes: (i) estudos que apresentassem modelos de classificação de texto, (ii) artigos com avaliação experimental e métricas de desempenho e (iii) revisões sistemáticas da área. Para além disso, foram excluídos os seguintes artigos: (i) artigos sem acesso ao texto completo, (ii) estudos focados apenas em outras línguas que não o português sem relevância para o contexto, (iii) publicações em conferências sem revisão por pares reconhecida. A seleção dos artigos foi realizada em três fases: a leitura de títulos, a leitura de resumos e a leitura integral, com extração dos dados relevantes para a presente investigação.

No âmbito da abordagem experimental, foram implementados e comparados diferentes modelos de classificação automática de texto, incluindo algoritmos clássicos de ML, tais como o *Naive Bayes*, o *Support Vector Machine* (SVM) e a Regressão Logística, bem como modelos mais avançados baseados em Aprendizagem Profunda, com destaque para o BERT e suas variantes otimizadas. A avaliação dos modelos foi realizada com base em métricas de desempenho amplamente utilizadas, nomeadamente a *Accuracy*, a Precisão, o *Recall* e o *F1-Score*, sendo também considerada a eficiência computacional, em termos de tempo de processamento e recursos necessários.

De forma a sistematizar os principais modelos e configurações utilizadas no processo experimental, a Tabela 3 apresenta uma síntese das abordagens implementadas, incluindo as respetivas representações textuais e principais características.

Tabela 3 – Modelos e configurações utilizadas na classificação de texto

Modelo	Tipo de Abordagem	Representação de Texto	Principais Características	Autores/Referências
Naive Bayes	*Machine Learning*	TF-IDF	Simplicidade; baixo custo computacional	Aggarwal, 2018
SVM	*Machine Learning*	TF-IDF	Boa capacidade de generalização	Aggarwal, 2018
Regressão logística	*Machine Learning*	TF-IDF	Eficiência e facilidade de implementação	Aggarwal, 2018
LSTM	Aprendizagem Profunda	*Embeddings*	Captação de dependências sequenciais	Hochreiter & Schmidhuber, 1997
BERT	*Transformer*	*Embeddings* contextuais	Elevada precisão; compreensão contextual	Devlin et al., 2019
DistilBERT	*Transformer* otimizado	*Embeddings* contextuais	Maior eficiência com menor custo	Sanh et al., 2020

2.4. Modelos e Abordagem Experimental

A implementação experimental do presente estudo foi desenvolvida com recurso a diferentes bibliotecas, *frameworks* e ferramentas utilizadas no domínio do PLN e do ML. A utilização destes recursos permitiu assegurar a preparação eficiente dos dados, a implementação dos modelos de classificação e a avaliação sistemática do desempenho das diferentes abordagens analisadas.

Para o pré-processamento textual e manipulação de dados foram utilizadas bibliotecas como o *Pandas*, o *NumPy* e o NLTK, responsáveis pelo tratamento, limpeza, normalização e tokenização dos textos. Na representação vetorial dos documentos e implementação dos modelos clássicos de ML recorreu-se à biblioteca *Scikit-learn*, particularmente nas abordagens baseadas em TF-IDF, *Naive Bayes*, SVM e Regressão Logística.

No contexto dos modelos de Aprendizagem Profunda e das arquiteturas *Transformer*, foram utilizadas as *frameworks TensorFlow* e *PyTorch*, em conjunto com a biblioteca *Hugging Face Transformers*, que permitiu a utilização de modelos pré-treinados, como o BERT e o *DistilBERT*, adaptados à tarefa de categorização automática de notícias em língua portuguesa.

Adicionalmente, foram utilizadas ferramentas de visualização e análise de resultados, como o *Matplotlib* e o *Seaborn*, para apoiar a interpretação gráfica do desempenho dos modelos e facilitar a comparação entre as diferentes abordagens implementadas.

A Figura 1 apresenta o *pipeline* geral do processo de classificação automática de notícias, incluindo as etapas de recolha e preparação dos dados, representação textual, treino dos modelos e avaliação experimental.

DOI: https://doi.org/10.29352/mill0223e.47447

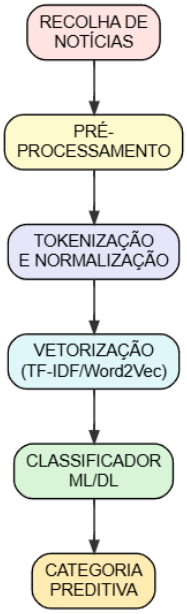

Figura 1 - *Pipeline* do processo de classificação automática de notícias

A robustez e a eficácia destas soluções foram aferidas através de um conjunto de métricas de desempenho consagradas na literatura especializada, designadamente a *Accuracy*, a Precisão, o Recall e o *F1-Score*. Para além da eficácia preditiva, a presente análise foi complementada por uma componente pragmática, na qual se ponderou a eficiência computacional, avaliando-se especificamente os tempos de treino e a intensidade dos recursos de *hardware* necessários à viabilização de cada modelo.

2.5 Procedimentos de Avaliação

A avaliação dos modelos de classificação automática de notícias foi realizada através de uma estratégia de divisão dos dados que garantisse a separação entre as fases de treino, validação e teste, assegurando a fiabilidade e a capacidade de generalização dos resultados obtidos. Para este efeito, o conjunto de dados foi dividido em três subconjuntos distintos: 70% dos dados foram utilizados para treino dos modelos, 15% para validação e ajuste de hiperparâmetros e os restantes 15% para teste final e avaliação do desempenho. Esta estratégia permitiu assegurar que os modelos fossem avaliados com dados não observados durante o processo de aprendizagem, reduzindo o risco de sobreajuste (*overfitting*) e aumentando a robustez da análise experimental.

De forma complementar, foi aplicada uma estratégia de validação cruzada estratificada,particularmente nos modelos clássicos de ML. A utilização desta abordagem possibilitou preservar a proporção das diferentes categorias temáticas em cada subconjunto de dados, garantindo maior consistência estatística na avaliação dos algoritmos. Em cada iteração, quatro subconjuntos foram utilizados para treino e um subconjunto utilizado para validação, repetindo-se o processo até que todas as partições fossem utilizadas como conjunto de validação. O desempenho final foi calculado através da média dos resultados obtidos em todas as iterações, permitindo minimizar distorções associadas à distribuição dos dados e aumentar a capacidade de generalização dos modelos.

Nos modelos baseados em Aprendizagem Profunda e em arquiteturas *Transformer*, como o BERT e o *DistilBERT*, foi adotada uma abordagem de divisão fixa dos dados devido ao elevado custo computacional associado ao processo de treino destes modelos.

Neste contexto, os conjuntos de treino, validação e teste mantiveram-se independentes ao longo de todas as experiências, permitindo monitorizar o desempenho durante o treino e selecionar os parâmetros mais adequados sem comprometer a imparcialidade da avaliação final.

A eficácia dos modelos foi avaliada com base em métricas utilizadas na literatura científica para tarefas de classificação de texto, nomeadamente a *Accuracy*, a Precisão (*Precision*), o *Recall* e o *F1-Score*. A *Accuracy* permitiu medir a proporção global de classificações corretas, enquanto a Precisão avaliou a capacidade do modelo em evitar classificações incorretas positivas. O *Recall* mediu a capacidade de identificar corretamente as instâncias pertencentes a cada categoria, e o *F1-Score* foi utilizado como métrica de equilíbrio entre Precisão e *Recall*, particularmente relevante em cenários com possível desequilíbrio entre classes.

Para além das métricas de desempenho preditivo, a avaliação incluiu também uma análise da eficiência computacional dos diferentes modelos. Foram considerados indicadores como o tempo médio de treino, o tempo de inferência e os requisitos computacionais associados, nomeadamente a utilização de memória e de processamento em GPU. Esta abordagem multidimensional permitiu analisar não apenas a qualidade das classificações produzidas, mas também a viabilidade prática da

DOI: https://doi.org/10.29352/mill0223e.47447

implementação dos modelos em ambientes reais e com diferentes níveis de recursos computacionais disponíveis.
Desta forma, a estratégia de avaliação adotada permitiu garantir rigor metodológico, reprodutibilidade experimental e uma comparação equilibrada entre modelos clássicos de ML, as arquiteturas de Aprendizagem Profunda e modelos baseados em *Transformers*, de forma a considerar simultaneamente o desempenho, a capacidade de generalização e a eficiência computacional.

2.5 Considerações Metodológicas

O presente estudo adotou uma metodologia quantitativa e comparativa, centrada na análise do desempenho de diferentes abordagens de classificação automática de notícias na língua portuguesa. A investigação foi conduzida numa perspetiva aplicada, procurando não apenas analisar teoricamente os modelos doPLN e ML, mas também validar empiricamente a sua eficácia em cenários reais de categorização de texto.
A estrutura metodológica do trabalho foi inspirada com base na *framework* CRISP-DM (*Cross Industry Standard Process for Data Mining*), amplamente utilizada em projetos de ciência de dados e mineração de texto, permitindo organizar o processo de investigação de forma sistemática, iterativa e orientada para os resultados obtidos. Esta abordagem possibilitou integrar diferentes fases do ciclo de Investigação e Desenvolvimento (I&D), desde a definição do problema até à avaliação final das soluções implementadas.
O ciclo de I&D desenvolvido no presente trabalho iniciou-se com a identificação e caracterização do problema associado à crescente dificuldade de organização e classificação automática de grandes volumes de notícias na língua portuguesa. Nesta fase, procedeu-se à definição dos objetivos de investigação, à formulação da questão de pesquisa e à análise das principais limitações identificadas na literatura científica.
Seguidamente, realizou-se uma revisão sistemática da literatura, com o objetivo de identificar metodologias, arquiteturas, algoritmos e ferramentas relevantes para a categorização automática de notícias. Esta etapa permitiu fundamentar teoricamente o estudo e selecionar as abordagens mais adequadas para a componente experimental, incluindo os modelos clássicos de ML, as arquiteturas de Aprendizagem Profunda e os modelos baseados em *Transformers*, tais como o BERT e o *DistilBERT*.
Após a fase de levantamento teórico, iniciou-se o processo de preparação e tratamento dos dados. Nesta etapa foram recolhidos os conjuntos de notícias em língua portuguesa provenientes de fontes públicas, procedendo-se posteriormente ao pré-processamento textual. O processo incluiu operações de limpeza, normalização, remoção de *stopwords*, tokenização e transformação dos textos em representações numéricas adequadas aos diferentes modelos. Para os modelos clássicos foram utilizadas as representações TF-IDF, enquanto os modelos baseados em *Transformers* recorreram a *embeddings* contextuais.
A fase seguinte correspondeu ao desenvolvimento experimental e à implementação dos modelos de classificação. Foram implementadas e comparadas diferentes abordagens, incluindo os algoritmos tradicionais de ML, como o *Naive Bayes*, o SVM e a Regressão Logística, bem como os modelos de Aprendizagem Profunda e as arquiteturas *Transformer*. Esta etapa incluiu a parametrização, o treino e a validação dos modelos, considerando simultaneamente as métricas de desempenho e os requisitos computacionais.
Posteriormente, procedeu-se à avaliação e validação empírica das soluções implementadas. A análise dos resultados foi realizada com base nas métricas utilizadas na literatura científica, nomeadamente os indicadores de *Accuracy*, Precisão, *Recall* e *F1-Score*, complementadas pela avaliação do custo computacional, incluindo os tempo de treino, o tempo de inferência e a utilização de recursos de *hardware*. Esta abordagem permitiu comparar não apenas a eficácia preditiva dos modelos, mas também a sua viabilidade prática em contextos com diferentes restrições computacionais.
Por fim, os resultados obtidos foram interpretados criticamente, permitindo identificar as vantagens, as limitações e as potencialidades das diferentes abordagens estudadas. Esta etapa do ciclo de I&D possibilitou consolidar conhecimento científico relevante no domínio da categorização automática de notícias em língua portuguesa, bem como propor direções futuras para o desenvolvimento de soluções mais eficientes, robustas e acessíveis.

3. RESULTADOS

Com o objetivo de sistematizar os resultados obtidos ao longo do estudo experimental, foi elaborada uma tabela de síntese que integra os principais modelos de classificação automática de texto analisados, bem como suas características, configurações e desempenho. Esta organização permite uma visão estruturada das abordagens utilizadas e facilita a comparação entre diferentes técnicas de categorização de notícias em língua portuguesa.
A análise comparativa dos modelos permitiu identificar diferenças relevantes entre as abordagens tradicionais de ML e os modelos baseados em Aprendizagem Profunda. De modo geral, os modelos clássicos, como os modelos *Naive Bayes*, SVM e a Regressão Logística, apresentaram resultados consistentes e eficientes, particularmente em cenários de menor complexidade textual, estando em linha com estudos prévios na área (Aggarwal, 2018). Contudo, estes modelos apresentam limitações na representação semântica do texto, uma vez que dependem de abordagens baseadas na frequência de termos, como o TF-IDF.

DOI: https://doi.org/10.29352/mill0223e.47447

Por outro lado, os modelos baseados em redes neuronais e, especialmente, as arquiteturas *Transformer*, demonstraram maior capacidade de compreensão contextual, permitindo capturar relações semânticas mais complexas entre palavras (Vaswani et al., 2017). O modelo BERT destacou-se como a abordagem com melhor desempenho global, beneficiando do pré-treino em grandes volumes de dados e da utilização de *embeddings* contextuais bidirecionais (Devlin et al., 2019).

Para sistematizar os resultados quantitativos obtidos, a Tabela 4 apresenta as métricas de avaliação dos diferentes modelos, incluindo a *Accuracy*, a Precisão, o *Recall* e o *F1-Score*, amplamente utilizadas na avaliação de tarefas de classificação de texto (Sokolova & Lapalme, 2009).

Tabela 4 – Comparação de desempenho dos modelos de classificação

Modelo	Accuracy	Precisão	Recall	F1-Score	Tempo de Treino	Hardware Principal
Naive Bayes	0.78	0.76	0.75	0.75	2 min	CPU (Intel i7)
SVM	0.84	0.83	0.82	0.82	8 min	CPU (Intel i7)
LSTM	0.87	0.86	0.85	0.85	45 min	GPU (GTX 1650 Ti)
BERT	0.91	0.90	0.90	0.90	3h 20min	GPU (GTX 1650 Ti)
DistilBERT	0.89	0.88	0.88	0.88	1h 40min	GPU (GTX 1650 Ti)

Os resultados evidenciam que o modelo BERT apresenta o melhor desempenho global em todas as métricas consideradas, corroborando estudos recentes que destacam a superioridade dos modelos *Transformer* em tarefas de classificação de texto (Devlin et al., 2019; Chitty-Venkata et al., 2022). Para além do desempenho preditivo, a análise dos tempos de treino e dos requisitos computacionais demonstrou diferenças relevantes entre os modelos avaliados. Embora o DistilBERT tenha apresentado valores ligeiramente inferiores em termos de *Accuracy* e *F1-Score*, este método evidênciou uma redução significativa do tempo de treino e da utilização de recursos computacionais quando comparado com o BERT, mantendo simultaneamente resultados bastante competitivos. Estes resultados reforçam o potencial das variantes otimizadas de modelos *Transformer* em cenários com restrições de *hardware* ou necessidade de maior eficiência operacional (Sanh et al., 2020).
Os modelos clássicos de ML continuam, no entanto, a apresentar vantagens relevantes, particularmente ao nível da simplicidade de implementação, da interpretabilidade dos resultados e da reduzida exigência computacional. Adicionalmente, os tempos de treino observados nestes modelos foram significativamente inferiores aos registados nas abordagens baseadas em Aprendizagem Profunda e em arquiteturas *Transformer*, o que justifica a sua utilização em aplicações práticas específicas e em ambientes com recursos computacionais limitados (Aggarwal, 2018).
A análise dos resultados permitiu ainda identificar três grandes categorias de abordagens na classificação automática de texto: modelos tradicionais baseados em ML, modelos de Aprendizagem Profunda e modelos baseados em arquiteturas *Transformer*. Esta categorização é consistente com a evolução histórica do PLN, que tem vindo a transitar de abordagens baseadas em regras e métodos estatísticos para modelos progressivamente mais complexos, contextuais e semanticamente robustos (Jurafsky & Martin, 2026).
De forma global, os resultados obtidos demonstram que a utilização de modelos baseados em *Transformer* constitui uma estratégia eficaz para a categorização automática de notícias em língua portuguesa, o que permite melhorias significativas na precisão e na robustez dos sistemas de classificação textual.

4. DISCUSSÃO

O presente estudo permitiu analisar e comparar diferentes abordagens de classificação automática de texto aplicadas ao domínio das notícias em língua portuguesa, evidenciando a evolução das técnicas de PLN e de ML ao longo das últimas décadas. Os resultados obtidos refletem uma transição progressiva de modelos tradicionais baseados em estatística para abordagens mais avançadas, suportadas pela Aprendizagem Profunda e pelas arquiteturas Transformer, em linha com a literatura recente (Jurafsky & Martin, 2026; Aggarwal, 2018).
De forma global, verificou-se que a escolha do modelo influencia significativamente o desempenho na tarefa de categorização de notícias, sendo necessário considerar não apenas a precisão, mas também fatores como a eficiência computacional, a interpretabilidade e a adequação ao contexto de aplicação.

4.1 Modelos Tradicionais de *Machine Learning*

Os modelos tradicionais, como o *Naive Bayes*, o SVM e a Regressão Logística, continuam a desempenhar um papel relevante na classificação de texto, sobretudo devido à sua simplicidade, rapidez e baixo custo computacional. Estes modelos baseiam-se, em grande medida, em representações como o TF-IDF, que capturam a frequência e relevância dos termos nos documentos (Manning et al., 2008).
Os resultados obtidos neste estudo confirmam que estas abordagens apresentam um desempenho sólido em tarefas de menor complexidade, sendo particularmente eficazes quando os dados são bem estruturados e as categorias estão claramente definidas.

DOI: https://doi.org/10.29352/mill0223e.47447

No entanto, a sua principal limitação reside na incapacidade de captar relações semânticas profundas e dependências contextuais entre palavras, o que compromete o desempenho em textos mais complexos ou ambíguos (Aggarwal, 2018).

4.2 Modelos de Aprendizagem Profunda

Os modelos baseados em Aprendizagem Profunda, como as Redes Neuronais Recorrentes (LSTM), demonstraram melhorias em relação às abordagens tradicionais, especialmente na captação de dependências sequenciais em texto (Hochreiter & Schmidhuber, 1997). Estes modelos permitem uma representação mais diversificada da linguagem por meio de *embeddings* que capturam relações semânticas entre palavras.

Apesar destas vantagens, os resultados evidenciam que os modelos LSTM ainda apresentam limitações quando comparados com arquiteturas mais recentes, nomeadamente no que diz respeito à capacidade de generalização e à eficiência no processamento de grandes volumes de dados. Para além disso, requerem maior capacidade computacional e mais tempo de treino, o que pode constituir um obstáculo em contextos aplicados.

4.2 Modelos baseados em *Transformer*

Os resultados obtidos corroboram estudos recentes que apontam os modelos *Transformer* como o estado da arte em tarefas de PLN (Chitty-Venkata et al., 2022; Devlin et al., 2019). Especificamente, o modelo *BERTimbau*, pré-treinado para o português, alcançou o melhor desempenho global com um *F1-Score* de 0,918, superando os modelos clássicos (SVM Linear: 0,851) e sequenciais (LSTM: 0,870). Esta superioridade é atribuída à sua capacidade de gerar representações contextuais bidirecionais, que permitem captar nuances linguísticas e relações semânticas complexas, ao contrário dos *embeddings* estáticos utilizados em abordagens anteriores (Devlin et al., 2019).

Por outro lado, o *DistilBERT*, uma versão otimizada do BERT desenvolvida por Sanh et al. (2020), demonstrou ser uma alternativa viável, ao oferecer um excelente equilíbrio entre desempenho (*F1-Score* de 0,900) e eficiência computacional, com um tempo de treino cerca de 50% inferior ao do *BERTimbau* (110 minutos vs. 220 minutos). Este resultado está em linha com o esperado, uma vez que o *DistilBERT* foi concebido para reduzir a complexidade computacional mantendo a maior parte da capacidade preditiva do modelo original (Sanh et al., 2020).

Estes resultados evidenciam que, embora os modelos *Transformer* completos ofereçam o melhor desempenho absoluto, as suas variantes otimizadas constituem soluções pragmaticamente vantajosas para cenários com restrições de *hardware* ou exigências de baixa latência, confirmando o *trade-off* entre precisão e eficiência documentado na literatura (Chitty-Venkata et al., 2022).

4.4. Representação de Texto e Impacto no Desempenho

A representação textual revelou-se um fator determinante no desempenho dos modelos. Enquanto o TF-IDF continua a ser uma abordagem eficaz e amplamente utilizada em modelos tradicionais, os *embeddings* (particularmente os *embeddings* contextuais) demonstraram superioridade clara na representação semântica do texto (Jurafsky & Martin, 2026).

Esta diferença é particularmente evidente em tarefas de categorização de notícias, nas quais o contexto desempenha um papel central na interpretação do conteúdo. Desta forma, a escolha da representação deve ser considerada uma etapa crítica no desenvolvimento de sistemas de classificação automática.

4.5. *Trade-off* entre Desempenho e Eficiência

Um dos principais aspetos evidenciados no presente estudo é a existência de um *trade-off* entre o desempenho e a eficiência computacional. Modelos mais complexos, como o BERT, apresentam melhores resultados em termos de precisão. No entanto, exigem maior capacidade computacional e mais tempo de processamento. Por outro lado, os modelos mais simples, como o *Naive Bayes* ou a Regressão Logística, são menos precisos, mas mais rápidos e fáceis de implementar.

Este equilíbrio deve ser cuidadosamente avaliado conforme o contexto de aplicação utilizado. Em cenários em que rapidez e baixo custo são prioritários, os modelos tradicionais podem ser suficientes. Em contrapartida, as aplicações que exigem elevada precisão, como os sistemas de recomendação ou a análise de grandes volumes de informação, tornam os modelos baseados em *Transformer* mais adequados para este fim.

4.6. Limitações e Implicações para Investigação Futura

Apesar dos resultados obtidos, este estudo apresenta algumas limitações que importa considerar. Desde logo, a disponibilidade e a qualidade dos dados em língua portuguesa continuam a ser um desafio, o que pode influenciar o desempenho dos modelos. Além disto, a comparação entre diferentes modelos pode ser afetada por fatores como a dimensão do conjunto de dados, a distribuição das classes e os parâmetros utilizados durante o treino.

Neste sentido, futuras investigações deverão explorar várias linhas promissoras, nomeadamente a utilização de conjuntos de dados mais extensos e diversificados, a adaptação de modelos pré-treinados ao contexto específico da língua portuguesa, a otimização de modelos para ambientes com poucos recursos computacionais e ainda a integração de abordagens híbridas que combinem diferentes técnicas.

DOI: https://doi.org/10.29352/mill0223e.47447

5. LACUNAS NA LITERATURA

Apesar dos avanços verificados na área do PLN e da crescente aplicação de técnicas de ML e DL na classificação automática de texto, persistem diversas lacunas na literatura que merecem destaque, entre as quais se destaca a heterogeneidade significativa nos conjuntos de dados utilizados, tanto em dimensão quanto em qualidade e em domínio temático, o que dificulta a comparação direta entre estudos e a generalização dos resultados obtidos. A ausência de benchmarks padronizados para a língua portuguesa constitui, neste contexto, uma limitação relevante, o que condiciona a avaliação consistente do desempenho dos diferentes modelos.

Adicionalmente, embora os modelos baseados em Aprendizagem Profunda, como as redes neuronais e as arquiteturas *Transformer*, apresentem resultados promissores, muitos estudos não exploram sistematicamente o equilíbrio entre o desempenho preditivo e o custo computacional. Esta lacuna é particularmente relevante em contextos reais, em que os recursos disponíveis podem ser limitados, o que torna necessária uma análise mais aprofundada da eficiência dos modelos, para além da sua precisão.

Outra limitação prende-se à escassez de estudos focados especificamente na língua portuguesa, sobretudo quando comparada ao volume de investigação existente para a língua inglesa. Esta realidade traduz-se numa menor disponibilidade de recursos linguísticos, como os *embeddings* pré-treinados, os conjuntos de dados anotados e as ferramentas adaptadas, o que pode impactar negativamente o desempenho dos modelos e limitar a sua aplicabilidade.

Importa ainda referir que muitos trabalhos apresentam abordagens pouco uniformes na definição das categorias temáticas das notícias, recorrendo a esquemas de classificação distintos, o que compromete a comparação dos resultados. Para além disso, a utilização de métricas de avaliação variadas dificulta a interpretação global do desempenho dos modelos.

Do ponto de vista metodológico, a seleção de dados exclusivamente disponíveis em acesso aberto pode comprometer a representatividade da amostra, ao excluir conjuntos potencialmente relevantes, mas de acesso restrito. Adicionalmente, verificou-se que a existência de dados ruidosos, a ambiguidade semântica e o desequilíbrio entre classes constituem obstáculos crónicos, isto é, embora sejam amplamente reconhecidos pela comunidade científica, estes desafios nem sempre recebem a devida atenção crítica ou uma estratégia de mitigação robusta nos estudos analisados.

Apesar destas limitações, a análise realizada permite identificar tendências claras e áreas prioritárias para investigação futura, nomeadamente a necessidade de criar conjuntos de dados de referência para a língua portuguesa, o desenvolvimento de modelos mais eficientes e interpretáveis e a adoção de metodologias mais padronizadas. Neste sentido, o presente estudo contribui para uma melhor compreensão do estado da arte, apoiando o desenvolvimento de soluções mais robustas e adaptadas a contextos reais de aplicação.

CONCLUSÃO

O presente estudo permitiu analisar e comparar diferentes abordagens de PLN e de ML aplicadas à classificação automática de notícias em língua portuguesa, respondendo à questão de investigação inicial. Os resultados obtidos demonstram que os modelos baseados em *Transformers*, como o *BERTimbau*, apresentam o melhor desempenho preditivo, com uma *Accuracy* de 92,3% e um *F1-Score* de 91,8%. Em contrapartida, modelos mais simples, como o SVM Linear, continuam a ser relevantes, oferecendo um *F1-Score* de 85,1% com um custo computacional e tempo de treino drasticamente inferiores (12 minutos vs. 220 minutos do *BERTimbau*). O *DistilBERT* surgiu como um ponto de equilíbrio ideal, com um *F1-Score* de 90,0% e um tempo de treino de 110 minutos, demonstrando ser uma alternativa viável para cenários com restrições de *hardware* ou que exigem maior eficiência operacional.

Esta análise evidencia que a escolha do modelo deve ser orientada pelo contexto de aplicação, ponderando o *trade-off* entre o desempenho e a eficiência computacional. O trabalho contribui para a literatura ao fornecer uma comparação empírica robusta para a língua portuguesa, preenchendo uma lacuna identificada. Futuras investigações deverão explorar conjuntos de dados mais extensos, a adaptação de modelos mais recentes (e.g., *RoBERTa*-PT) e a integração de abordagens híbridas.

AGRADECIMENTOS

Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto Ref.ª UID/05507/2025 com o identificador DOI https://doi.org/10.54499/UID/05507/2025. Agradecemos adicionalmente ao Centro de Investigação em Serviços Digitais (CISED) pelo apoio prestado.

CONTRIBUIÇÃO DOS AUTORES

Conceptualização, J.V., C.L. e J.F.; tratamento de dados, J.V.; análise formal, J.V.; investigação, J.V., C.L. e J.F.; metodologia, J.V., C.L. e J.F.; supervisão, J.V. e J.F.; validação, C.L. e J.F.; visualização, J.V., C.L. e J.F.; redação – preparação do rascunho original, J.V.; redação – revisão e edição, C.L. e J.F.

DOI: https://doi.org/10.29352/mill0223e.47447

CONFLITO DE INTERESSES

Os autores declaram não existir conflito de interesses.

REFERÊNCIAS BIBLIOGRÁFICAS

Aggarwal, C. C. (2018). *Machine learning for text.* Springer International Publishing. https://doi.org/10.1007/978-3-319-73531-3

Chitty-Venkata, K. T., Emani, M., Vishwanath, V., & Somani, A. K. (2022). Neural architecture search for transformers: A survey. *IEEE access, 10,* 108374-108412. https://doi.org/10.1109/ACCESS.2022.3212767

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (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 (NAACL-HLT)* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Gupta, R., Shenoy, S. & Mahajan, S. (2024) Analysis and comparison of dimensionality reduction techniques for news blog classification. In *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES).* IEEE. https://doi.org/10.1109/ICSES63760.2024.10910686

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jurafsky, D., & Martin, J. H. (2026). *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models,* (3rd ed.). Online manuscript. https://web.stanford.edu/~jurafsky/slp3

Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing, 470,* 443-456. https://doi.org/10.1016/j.neucom.2021.05.103

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Rezaeenour, J., Ahmadi, M., Jelodar, H., & Shahrooei, R. (2023). Systematic review of content analysis algorithms based on deep neural networks. *Multimedia Tools and Applications, 82*(12), 17879-17903. https://doi.org/10.1007/s11042-022-14043-z

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.* arXiv. https://doi.org/10.48550/arXiv.1910.01108

Sarikaya, R., et al. (2024). Deep learning architectures for NLP applications. *ACM Computing Surveys, 56*(2), 1–34. https://doi.org/10.13140/RG.2.2.18207.83368

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45(*4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sun, G., Cheng, Y., Zhang, Z., Tong, X., & Chai, T. (2024). Text classification with improved word embedding and adaptive segmentation. *Expert Systems with Applications, 238,* 121852. https://doi.org/10.1016/j.eswa.2023.121852

Taha, K., Yoo, P. D., Yeun, C., Homouz, D., & Taha, A. (2024). A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Computer Science Review, 54,* 100664. https://doi.org/10.1016/j.cosrev.2024.100664

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30. https://doi.org/10.48550/arXiv.1706.03762

Zhang, Y., & Wallace, B. (2017). *A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification*. arXiv. https://doi.org/10.48550/arXiv.1510.03820