INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Extração e classificação automática de conteúdos textuais de publicações periódicas em Portugal

Humberto Carlos Pestana Fernandes

Mestrado em Ciência de Dados

Orientadores:
Doutor Fernando Manuel Marques Batista, Professor Associado, ISCTE – Instituto Universitário de Lisboa

Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Associado, ISCTE – Instituto Universitário de Lisboa

Outubro, 2025

Departamento de Métodos Quantitativos para Gestão e Economia

Departamento de Ciências e Tecnologia da Informação

Extração e classificação automática de conteúdos textuais de publicações periódicas em Portugal

Humberto Carlos Pestana Fernandes

Mestrado em Ciência de Dados

Orientadores:
Doutor Fernando Manuel Marques Batista, Professor Associado, ISCTE – Instituto Universitário de Lisboa

Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Associado, ISCTE – Instituto Universitário de Lisboa

Outubro, 2025

Uma vida curta, mas cheia de sentido.

Para ti, Ana Teresa, foste o caminho que me

levou a descobrir o significado das palavras ***a paz de Cristo****.*

Que Deus te abençoe, filha querida.

Agradecimento

Os últimos dois anos foram desafiantes em diversos aspetos; não recordo outro período da vida em que tenha vivido tantas emoções em tão curto espaço de tempo. Michelle, obrigado por estares ao meu lado, por me ouvires e por me suportares nos momentos mais difíceis. Gracias, amor; este triunfo también es tuyo.

Amada Beatriz Valdez, tinhas de estar presente aqui, desta vez em português. Bem sabes a influência que tiveste neste percurso. Obrigado por me contagiares com o teu entusiasmo pela programação e pela análise de dados. A minha gratidão será sempre eterna.

Agradeço igualmente a todos os que, na ERC, apoiaram este projeto, não apenas em termos financeiros, mas também na sua concretização. Um agradecimento especial à Vanda Cruz: há alguns anos, dificilmente imaginaríamos que aquela simples ideia de reorganizar as categorias daria origem a uma dissertação. Também o meu reconhecimento à Carla Martins, ao Telmo Gonçalves e à Professora Helena Sousa pelos contributos prestados a este trabalho. Obrigado.

Não podia deixar de expressar o meu agradecimento aos Professores Fernando Batista e Ricardo Ribeiro. Sei que um trabalho deste género exige uma grande entrega por parte do aluno, mas vocês, enquanto orientadores, cheios de paciência, dedicação, disponibilidade e criatividade, foram parte essencial para que o projeto fosse culminado.

Resumo

A classificação temática das publicações periódicas constitui um desafio para a Entidade Reguladora para a Comunicação Social (ERC), responsável pelo registo e categorização destas publicações. O sistema de categorização em vigor apresenta limitações, o que tem levado a própria ERC a reconhecer a necessidade de o reformular. Neste contexto, o presente estudo abordou o desafio da classificação temática das publicações periódicas portuguesas com presença digital, com o propósito de avaliar o desempenho de modelos de linguagem pré-treinados baseados em *transformers*. Desenvolveu-se um sistema de categorias temáticas em colaboração com a ERC e aplicaram-se técnicas de *web scraping* para extrair o conteúdo textual das publicações. O estudo avaliou o desempenho de modelos *transformer* baseados em arquiteturas *encoder*, nomeadamente o mBERT, XLM-RoBERTa, BERTimbau e Albertina, tendo sido definido o modelo *decoder* EuroLLM como referência para efeitos de comparação. Foram igualmente testadas diferentes abordagens de extração de dados, bem como o impacto do pré-processamento e da modelação de textos longos. Os resultados indicam que as arquiteturas do tipo *encoder* apresentaram um melhor desempenho, com métricas de F1-*score* até 84,2% na classificação multiclasse e 96,0% na classificação binária. Verificou-se ainda que os modelos *encoder* obtiveram melhores resultados com a extração integral do texto dos sítios eletrónicos das publicações periódicas e sem recurso a pré-processamento, o que simplifica a fase de preparação dos dados. Adicionalmente, a modelação de textos longos, através da agregação das probabilidades, revelou-se a estratégia mais eficaz para contornar a limitação do comprimento máximo de sequência de entrada dos modelos *encoder*.

Palavras-chave: classificação de texto, modelos de linguagem pré-treinados, transformadores, extração de dados, publicações periódicas.

Abstract

The thematic classification of periodicals is a challenge for the Portuguese Regulatory Authority for the Media (ERC), which is responsible for the registration and categorisation of these media outlets. The current categorisation system has limitations, which has led the ERC itself to recognise the need to reformulate it. In this context, the present study addressed the challenge of thematic classification of Portuguese periodicals with a digital presence, aiming to evaluate the performance of pre-trained transformer-based language models. A system of thematic categories was developed in collaboration with ERC, and web scraping techniques were applied to extract textual content from the publications. The study assessed the performance of transformer models based on encoder architectures, namely mBERT, XLM-RoBERTa, BERTimbau, and Albertina, with the EuroLLM decoder model defined as a baseline for comparison. Different data extraction approaches were also tested, as well as the impact of pre-processing and long-text modeling. The results confirmed the superiority of the encoder architectures, with F1-score metrics reaching up to 84.2% in multiclass classification and 96.0% in binary classification. It was also found that the encoder models achieved better results with the full-text extraction from the periodicals' websites and without pre-processing, which simplifies the data preparation phase. Additionally, long-text modeling, through the aggregation of probabilities, proved to be the most effective strategy to overcome the maximum input sequence length limitation of the encoder models.

Keywords: text classification, pre-trained language models, transformers, data extraction, periodicals

Índice

Agradecimento i
Resumo iii
Abstract v
Índice vii
Índice de Figuras ix
Índice de Tabelas xi
Glossário de Siglas xiii
CAPÍTULO 1 1
Introdução 1
1.1. Objetivos de Investigação 2
1.2. Motivação 3
1.3. Estrutura do Documento 3
CAPÍTULO 2 5
Enquadramento Teórico 5
2.1. A Evolução da Classificação de Texto 6
2.2. Modelos de Linguagem Pré-treinados Baseados em *Transformers* 8
2.2.1. EuroLLM 10
2.2.2. BERT 11
2.2.3. XLM-RoBERTa 11
2.2.4. BERTimbau 12
2.2.5. Albertina 12
2.3. Classificação Temática das Publicações Periódicas em Portugal 13
CAPÍTULO 3 15
Revisão de Literatura 15
CAPÍTULO 4 19
Abordagem Metodológica 19
4.1. Classificação Manual 19
4.2. Extração de Dados 21
4.3. Base de Dados 23
4.4. Experiências de Classificação de Texto 27
CAPíTULO 5 31
Resultados e Discussão 31
5.1. Classificação Multiclasse 32

5.2. Classificação Binária 36
5.3. Discussão dos Resultados 39
CAPÍTULO 6 41
Conclusões 41
Fontes 43
Referências Bibliográficas 45
Anexo A 49
Anexo B 53
Anexo C 55
Anexo D 57

Índice de Figuras

Figura 2.1: Arquitetura dos Modelos *Transformers* GPT e BERT 10
Figura 2.2: Publicações Periódicas por Classificação Temática da ERC 14
Figura 4.1: Esquema Geral do Processo Metodológico Aplicado à Classificação Temática das Publicações Periódicas 19
Figura 4.2: Publicações Periódicas Classificadas Manualmente 21
Figura 4.3: Publicações Periódicas por Categoria Temática – Base de Dados para a Classificação Multiclasse 26
Figura 4.4: Experiências de Classificação de Texto 27
Figura 4.5: Instrução para a Classificação através do Modelo EuroLLM 28
Figura 5.1: Desempenho (F1-*score*) dos Modelos *Encoder* na Classificação Multiclasse, por Abordagem de Extração 33
Figura 5.2: Desempenho (F1-*score*) dos Modelos *Encoder* na Classificação Multiclasse, por Pré-processamento dos Dados 34
Figura 5.3: Matriz de Confusão do Modelo mBERT 35
Figura 5.4: Matriz de Confusão do Modelo XLM-RoBERTa 35
Figura 5.5: Desempenho (F1-*score*) dos Modelos *Encoder* na Classificação Binária, por Abordagem de Extração 37
Figura 5.6: Desempenho (F1-*score*) dos Modelos *Encoder* na Classificação Binária, por Pré-processamento dos Dados 38
Figura 5.7: Matriz de Confusão do Modelo BERTimbau 38

Índice de Tabelas

Tabela 4.1: Sistema de Categorias Temáticas 20
Tabela 4.2: Extração por Etiquetas e Relevância da Informação em 150 Sítios Eletrónicos de Publicações Periódicas 22
Tabela 4.3: Estatísticas Descritivas dos Textos Extraídos por Etiquetas e de Forma Integral, Antes e Após Pré-Processamento – Base de Dados para a Classificação Multiclasse 25
Tabela 4.4. Estatísticas Descritivas dos Textos Extraídos por Etiquetas e de Forma Integral, Antes e Após Pré-Processamento – Base de Dados para a Classificação Binária 26
Tabela 5.1: Melhores Resultados por Experiências na Classificação Multiclasse 32
Tabela 5.2: Melhores Resultados por Experiências na Classificação Binária 37

Glossário de Siglas

BERT – *Bidirectional Encoder Representations from Transformers*

CNN – *Convolutional Neural Network*

ERC – Entidade Reguladora para a Comunicação Social

GPT – *Generative Pre-trained Transformer*

MLM – *Masked Language Modeling*

NER – *Named Entity Recognition*

NLP – *Natural Language Processing*

NSP – *Next Sentence Prediction*

RNN – *Recurrent Neural Networks*

XLM-R – XLM-RoBERTa

CAPÍTULO 1

Introdução

O crescimento exponencial de dados não estruturados provenientes de fontes tão diversas como redes sociais, portais de notícias, bibliotecas digitais, sistemas de informação organizacionais e outras plataformas representa um recurso de enorme valor, mas coloca desafios no que respeita à sua gestão, tratamento e análise.

A complexidade, heterogeneidade e ausência de formato predefinido deste tipo de dados dificultam a sua representação e processamento, exigindo abordagens capazes de transformá-los em informação útil e analisável. No caso dos dados textuais, a escala atual torna a classificação manual uma tarefa impraticável, o que reforça a necessidade de recorrer a processos automáticos e contínuos que permitam organizar, filtrar e tornar a informação acessível de forma eficiente e sistemática (Allam et al., 2025; Minaee et al., 2021).

Neste contexto, as abordagens do campo da mineração de texto (*text mining*) assumem particular relevância. Trata-se de uma área interdisciplinar que conjuga métodos de aprendizagem automática (*machine learning*), processamento de linguagem natural (*Natural Language Processing* – NLP), estatística e recuperação de informação (*information retrieval*) com o propósito de extrair conhecimento a partir de dados textuais (Aggarwal, 2018; Ignatow & Mihalcea, 2018).

Entre as técnicas mais amplamente aplicadas no âmbito de *text mining* encontra-se a classificação de texto, que procura atribuir automaticamente uma categoria a um documento com base no seu conteúdo, a partir de exemplos previamente categorizados (Ignatow & Mihalcea, 2018; Sebastiani, 2002).

Recentemente, a introdução de modelos de linguagem pré-treinados (*Pre-trained Language Models* – PLM), baseados na arquitetura transformador (*transformer*), marcou um novo paradigma na classificação de texto. Estes modelos são pré-treinados com um enorme volume de dados textuais, o que lhes permite adquirir um conhecimento profundo sobre a estrutura e a semântica da linguagem. Tal conhecimento pode posteriormente ser reutilizado em tarefas específicas, como a classificação, potenciando a compreensão contextual e o desempenho dos modelos (Chen et al., 2022; Wu & Wan, 2025).

Os modelos de linguagem pré-treinados baseados em *transformers* alcançam bons resultados sem depender de grandes volumes de dados categorizados (X. Han et al., 2021; Santos et al., 2021). Esta característica torna-os particularmente adequados em línguas com poucos recursos, como o português europeu, que, apesar da sua ampla difusão, continua a dispor de recursos linguísticos limitados no campo do NLP (Garcia et al., 2024; Ramos et al., 2024; Santos et al., 2021).

Neste enquadramento, o presente projeto de investigação contribui para o avanço da aplicação de técnicas de *text mining* a conjuntos de dados em português europeu, abordando o desafio da classificação temática das publicações periódicas com presença digital registadas em Portugal. Pretende-se explorar a automatização deste processo através de modelos de linguagem pré-treinados baseados em *transformers*, incluindo modelos multilingues como o mBERT e o XLM-RoBERTa, e modelos especializados em língua portuguesa, como o BERTimbau e o Albertina, sobre o conjunto de dados extraídos dos próprios sítios eletrónicos das publicações periódicas.

Para compreender a relevância e o enquadramento deste desafio, importa considerar o contexto em que se insere o objeto de estudo. As publicações periódicas, enquanto meios de comunicação social, estão sujeitas à supervisão e regulação da Entidade Reguladora para a Comunicação Social (ERC). De acordo com os dados da ERC, a 6 de fevereiro de 2025, encontravam-se registadas 1 680 publicações periódicas.

No âmbito das suas atribuições legais, a ERC procede à classificação manual destas publicações segundo os critérios definidos pela Lei de Imprensa. Para além da classificação estipulada pela Lei, a ERC também realiza a classificação temática das publicações. Este processo é conduzido manualmente pelos serviços da entidade, tendo em conta o conteúdo publicado, o estatuto editorial, o público-alvo e a entidade proprietária de cada publicação.

Contudo, o sistema atual de categorização temática apresenta uma série de limitações, uma vez que as categorias combinam dimensões distintas, como a periodicidade (ex.: diário, não diário), o tipo de publicação (ex.: jornal, revista) e o conteúdo temático (ex.: turismo, desporto). Em consequência, a ERC opta por não divulgar publicamente estes dados, reconhecendo as limitações da classificação existente e a necessidade de rever o atual sistema de categorias temáticas.

Neste sentido, a investigação procura responder à seguinte questão: É possível automatizar de forma eficaz a classificação temática das publicações periódicas portuguesas com base no texto disponível nos seus *websites*, recorrendo a modelos *transformers* pré-treinados? Para tal, o projeto visa desenhar um novo sistema de categorias temáticas, bem como extrair e preparar os dados textuais das publicações com presença digital e determinar qual a combinação de modelo, metodologia de extração, tratamento de dados e estratégia de modelação de textos longos que maximiza o desempenho dos modelos de linguagem pré-treinados na tarefa de classificação automática.

1.1. Objetivos de Investigação

O objetivo principal desta investigação consiste em automatizar a classificação temática das publicações periódicas portuguesas com presença digital, identificando a categoria predominante de

cada publicação a partir do conteúdo textual disponibilizado nos respetivos sítios eletrónicos, mediante a aplicação de modelos de linguagem pré-treinados baseados na arquitetura *transformer*.

Para concretizar este objetivo, delinearam-se diferentes objetivos específicos que estruturaram o desenvolvimento da investigação:

- Desenhar um sistema de categorias temáticas das publicações periódicas, que constitua a base para as tarefas de classificação supervisionada.
- Extrair e estruturar os dados textuais das publicações periódicas com presença digital, aplicando técnicas de extração de dados da internet (*web scraping*).
- Avaliar o desempenho de diferentes modelos *transformer* na tarefa de classificação automática considerando as abordagens de extração de texto, o tratamento dos dados, as estratégias de modelação de textos longos e a eficácia comparativa entre modelos multilingues (mBERT e XLM-RoBERTa) e modelos especializados na língua portuguesa (BERTimbau e Albertina).

1.2. Motivação

Este estudo reveste-se de relevância institucional e aplicada, ao procurar responder a uma necessidade concreta da ERC. A criação de um sistema automatizado de classificação temática das publicações periódicas pode ser um contributo para a organização, consulta e acessibilidade da informação que permita caracterizar o setor da imprensa no país.

Por outro lado, este projeto pretende constituir um contributo na área das Ciências de Dados e nas Ciências da Computação em Portugal. A presente investigação assume particular pertinência no atual contexto de desenvolvimento de modelos de linguagem pré-treinados em português, ao testar o desempenho de modelos de última geração, como o Albertina (versão PT-PT), e ao promover a disponibilização de bases de dados textuais em português europeu.

1.3. Estrutura do Documento

A presente dissertação organiza-se em seis capítulos. Após esta introdução, o segundo capítulo apresenta o enquadramento teórico do estudo, abordando os aspetos conceptuais da classificação de texto e dos modelos de linguagem pré-treinados baseados em *transformers*, bem como descreve a classificação temática das publicações periódicas em Portugal com base na legislação nacional e nos dados disponibilizados pela ERC.

O terceiro capítulo apresenta a revisão da literatura sobre a aplicação de *transformers* pré-treinados à classificação de texto em português.

O quarto capítulo descreve a abordagem metodológica adotada, que compreende a definição de um novo sistema de classificação temática, a identificação das publicações com presença digital e a sua classificação manual, a extração e preparação dos dados, a construção das bases de dados e a definição das experiências de classificação automática.

O quinto capítulo apresenta e discute os resultados obtidos, relacionando-os com a questão de investigação e os objetivos definidos.

Por fim, o sexto e último capítulo reúne as conclusões do estudo e as recomendações para trabalhos futuros.

CAPÍTULO 2

Enquadramento Teórico

A classificação de texto, também conhecida como categorização de texto, é o processo que consiste em associar a um texto uma ou mais categorias previamente definidas. Em termos formais, tendo uma representação R de um texto T e um conjunto fixo de categorias $C = \{C_1, C_2, \ldots, C_n\}$, a tarefa foca-se em determinar uma correspondência (mapeamento) de R para uma das categorias em C. Este processo é habitualmente concretizado através da aprendizagem de como criar tais correspondências, baseando-se num conjunto de textos que já foram previamente classificados (Ignatow & Mihalcea, 2018).

A tarefa de classificação de texto pode ser binária, se os textos forem categorizados numa de duas classes exclusivas, ou multiclasse, quando a categorização ocorre numa única classe dentro de um universo de três ou mais categorias. Quando um documento pode pertencer a múltiplas categorias simultaneamente, o processo designa-se por classificação múltipla.

Segundo Sebastiani (2002), a abordagem computacional a este problema assenta em duas premissas:

1. As categorias funcionam como simples etiquetas simbólicas, sem qualquer significado ou conhecimento associado.
2. Não se utiliza conhecimento externo para a classificação. O processo de classificação deve basear-se exclusivamente na informação interna, extraída do conteúdo dos próprios documentos. Isto significa que não se conta com qualquer informação externa ou metadatos como data de publicação, tipo de documento ou fonte (p. 3).

Este autor assinala que a dependência exclusiva da informação interna dos documentos implica que o processo de classificação se assenta na semântica; por conseguinte, e visto que a semântica é uma noção subjetiva, a atribuição de uma categoria não é determinística (p. 3).

Nesse sentido, à semelhança da classificação manual, a classificação automática de texto não está isenta de erro, refletindo a incerteza inerente aos critérios interpretativos. Esta inconsistência é notória mesmo entre especialistas. Cleverdon (1984) argumentou que investigações sobre o processo de armazenamento e recuperação de informação revelaram que, se dois indexadores experientes indexarem um mesmo documento utilizando um dado tesauro, apenas 30% dos termos de indexação podem ser comuns aos dois conjuntos de termos (p. 38).

Ainda assim, nas últimas décadas a classificação automática evoluiu ao ponto de alcançar níveis de eficácia iguais ou superiores aos de profissionais especializados, em particular com o recurso a aprendizagem automática e, mais especificamente, a aprendizagem profunda (*deep learning*). Entre

as principais vantagens salientam-se a capacidade de processar grandes volumes de dados com rapidez, a portabilidade e adaptabilidade a novos domínios, bem como maior consistência, objetividade e precisão (Allam et al., 2025; Minaee et al., 2021; Sebastiani, 2002).

Atualmente, a classificação de texto tem diversas aplicações como a análise de sentimentos, a deteção de *spam*, a análise de opinião pública, a identificação de notícias falsas, a deteção de linguagem ofensiva, a classificação de notícias ou artigos publicados por jornais, revistas e páginas *web* de meios de comunicação social, entre outras aplicações. O enorme volume de conteúdos digitais divulgados diariamente, torna impraticável a sua classificação manual, o que reforça a importância de soluções automatizadas na gestão da informação.

2.1. A Evolução da Classificação de Texto

A utilização de computadores para auxiliar na tarefa de classificação de texto começou a ser explorada nas décadas de 1950 e 1960 (Fields et al., 2024; Sebastiani, 2002), destacando-se o trabalho de Maron (1961), que formulou o problema sob a designação de "indexação automática" (*automatic indexing*). Maron definiu-o como "o problema de decidir, por meios mecânicos, a que categoria (assunto ou domínio de conhecimento) pertence um determinado documento" (p. 404, tradução própria). Este cientista introduziu aspetos fundamentais da forma moderna de classificação de texto, como a atribuição de categorias temáticas a documentos através de estatísticas de ocorrência de palavras e do cálculo de probabilidades condicionais.[1]

No entanto, a abordagem dominante até ao final da década de 1980 foi a dos sistemas baseados em regras, os quais eram construídos através de especialistas com conhecimento na área. Estes sistemas consistiam em combinações de palavras que permitiam determinar a categoria de um texto e o seu desempenho era, em geral, elevado. Uma limitação desta abordagem era a sua manutenção, pois o processo de construção de regras era manual, o que tornava dispendiosa a alteração de categorias ou a introdução de novas (Ignatow & Mihalcea, 2018; Sebastiani, 2002).

A partir do início da década de 1990, a abordagem à classificação de texto sofreu um ponto de inflexão, impulsionada pela ascensão da aprendizagem automática e das técnicas de NLP[2] (Allam et al., 2025; Sebastiani, 2002). Este novo paradigma introduziu um processo indutivo, no qual um classificador automático é treinado para aprender as características de um conjunto de documentos previamente classificados, com o objetivo de categorizar novos documentos ainda não vistos.

[1] Melvin E. Maron exerceu funções como matemático na RAND Corporation entre 1959 e 1966, no Departamento de Ciências da Computação, onde se dedicou ao desenvolvimento de métodos estatísticos para a automatização da recuperação de informação (UC Berkeley School of Information, 2017).

[2] O processo ou a capacidade de uma máquina ou programa compreender texto ou fala em linguagem natural (ou humana) (Ignatow & Mihalcea, 2018).

No contexto da aprendizagem automática, esta metodologia é um exemplo de aprendizagem supervisionada, a qual se caracteriza por duas etapas essenciais para construir um modelo capaz de prever a categoria de um documento. Na primeira, procede-se ao pré-processamento dos textos (limpeza, normalização e preparação dos dados). Na segunda, realiza-se o processamento, em que o texto é transformado em vetores numéricos que podem ser interpretados por algoritmos de aprendizagem. É nesta última fase que se extraem e selecionam os atributos relevantes que o modelo utilizará para fazer as suas previsões.

A partir de 2010, a utilização das redes neuronais mostrou grandes avanços na tarefa de classificação de texto (X. Han et al., 2021; Minaee et al., 2021). O sucesso desta abordagem foi impulsionado pela combinação de vários fatores, como a evolução dos sistemas de *hardware* mais potentes e acessíveis, a disponibilidade de grandes conjuntos de dados e a necessidade de analisar o crescente volume de informação não estruturada.[3] Esta abordagem, conhecida como aprendizagem profunda, insere-se na aprendizagem automática, mas distingue-se pela sua capacidade de extrair características complexas dos documentos de forma automática. Ao contrário dos métodos tradicionais de aprendizagem automática, onde o processamento é uma etapa manual e separada, nos modelos de aprendizagem profunda esta extração está integrada no próprio processo de aprendizagem.

Em anos recentes, os modelos de linguagem pré-treinados baseados em arquiteturas de redes neuronais avançadas *transformers*, marcaram o novo paradigma na tarefa de classificação de texto (Fields et al., 2024; X. Han et al., 2021; Jurafsky & Martin, 2025; Wu & Wan, 2025). A capacidade de treinar estes modelos com um enorme volume de dados permite-lhes capturar relações contextuais complexas, o que fornece uma base de conhecimento que pode ser posteriormente transferida e ajustada (*fine-tuning*) para domínios de aplicação específicos.

A principal vantagem prática desta abordagem é a sua menor dependência de grandes bases de dados categorizados. Ao contrário de arquiteturas anteriores, como as redes neuronais recorrentes (*Recurrent Neural Networks* – RNN) ou convolucionais (*Convolutional Neural Network* – CNN), um modelo de linguagem pré-treinado baseado em *transformers* não necessita de grandes volumes de dados anotados específicos da tarefa para atingir um bom desempenho. Esta característica permite obter bons resultados mesmo com conjuntos de dados mais limitados, mitigando o risco de sobreajustamento (*overfitting*) (X. Han et al., 2021).

[3] São dados não estruturados os que não obedecem a um esquema pré-definido, como tabelas ou bases de dados relacionais; exemplos típicos incluem texto livre, imagens e áudio.

2.2. Modelos de Linguagem Pré-treinados Baseados em *Transformers*

Em 2017, Vaswani et al. publicaram o artigo "*Attention Is All You Need*", no qual foi apresentada a arquitetura *transformer*, uma rede neuronal que depende exclusivamente de mecanismos de atenção, descartando as estruturas das redes tradicionais como as RNN e as CNN utilizadas até então.

Nos *transformers*, o mecanismo de atenção permite que cada *token*[4] da sequência seja relacionado com todos os outros, gerando múltiplas representações contextuais em paralelo. Cada uma dessas representações pode enfatizar um tipo distinto de relação semântica ou sintática entre *tokens* (Silva Barbon & Akabane, 2022).

Esta arquitetura organiza-se segundo um modelo codificador-descodificador (*encoder–decoder*). O codificador (*encoder*) transforma a sequência de entrada em representações contextuais, enquanto o descodificador (*decoder*) utiliza essas representações para gerar a sequência de saída. Tanto no *encoder* como no *decoder* empilham-se blocos formados por camadas de auto-atenção multi-cabeça (*multi-head self-attention*), redes de propagação direta posicionais (*feed-forward*), ligações residuais (*residual connections*) e normalização de camadas (*layer normalization*), o que garante maior profundidade e estabilidade do treino (Vaswani et al., 2017).

A capacidade de processamento paralelo e de captação eficaz de dependências de longo alcance tornou os *transformers* adequados para tarefas de NLP, como a classificação de texto. A partir desta arquitetura emergiram os modelos de linguagem pré-treinados, também conhecidos como grandes modelos de linguagem (*Large Language Model* – LLM) (Fields et al., 2024).

A base destes modelos assenta num processo de duas fases, que define o paradigma de aprendizagem por transferência (*transfer learning*):

1. A fase de pré-treino, em que o modelo é inicialmente treinado de forma autosupervisionada[5] sobre um volume extenso de texto não categorizado proveniente da internet, de livros e de outras fontes. Durante esta fase, o modelo aprende a prever palavras em falta ou a prever a próxima palavra numa frase. O objetivo não é executar uma tarefa específica, mas sim forçar o modelo a aprender representações linguísticas profundas sobre a estrutura da linguagem, como a gramática, a semântica e os diferentes contextos de uma palavra.

[4] Unidade mínima de texto que pode corresponder a uma palavra inteira ou a uma parte de uma palavra.
[5] Nesta técnica a aprendizagem não depende da categorização manual dos dados, mas sim da própria sequencia das palavras (Jurafsky & Martin, 2025).

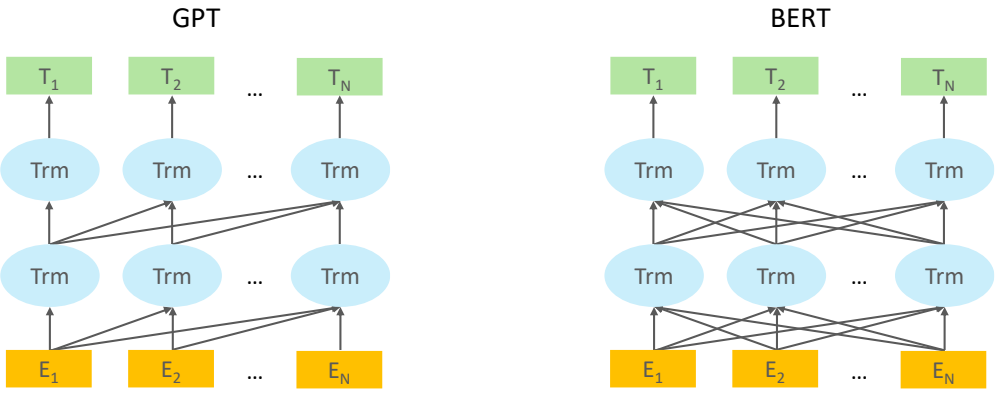
2. A fase de *fine-tuning*, na qual o conhecimento adquirido pelo modelo na fase anterior é transferido e especializado para uma tarefa concreta, como a classificação de texto, a tradução, a geração de resumos, entre outras. Este ajuste é realizado com um conjunto de dados categorizados, específico para a tarefa, que é tipicamente de uma dimensão muito inferior à dos dados de pré-treino. O modelo adapta os seus pesos para se otimizar na nova tarefa, sem perder a base de conhecimento que obteve na primeira fase.

Os modelos de linguagem pré-treinados baseados em *transformers* podem ser agrupados, de modo geral, em duas categorias: modelos generativos autorregressivos e modelos de autocodificação (X. Han et al., 2021; Minaee et al., 2021).

Os modelos generativos autorregressivos são constituídos exclusivamente por camadas de *decoder*, razão pela qual também são designados por modelos *decoder*. Tratam-se de modelos unidirecionais, nos quais a previsão de cada palavra depende das previsões anteriores, gerando a sequência de texto de forma sequencial da esquerda para a direita (ou da direita para a esquerda), este processo designa-se por *Causal Language Modeling* (CLM). Pela sua natureza, estes modelos são sobretudo aplicados em tarefas de geração de texto, como redação automática, sumarização ou tradução. Exemplos representativos incluem os modelos proprietários, como o ChatGPT (OpenAI), o Gemini (Google) e o Claude (Anthropic), bem como modelos de acesso aberto, como o EuroLLM, Mistral (Mistral AI) e o LLaMA (Meta), entre outros.

Já os modelos de autocodificação são compostos apenas por camadas de *encoder*, pelo que são igualmente conhecidos como modelos *encoder*. Estes distinguem-se por serem bidirecionais e por utilizarem como objetivo de treino o procedimento de *Masked Language Modeling* (MLM), no qual uma parte dos *tokens* de entrada é mascarada e, em vez de prever a palavra seguinte, a rede procura reconstruir os *tokens* ocultos a partir do contexto completo, tanto à esquerda como à direita. Esta abordagem de treino possibilita o condicionamento bidirecional, fundamental para que o modelo consiga inferir relações contextuais complexas entre as palavras de uma sequência. Entre as suas principais aplicações encontram-se a classificação de texto, o reconhecimento de entidades mencionadas (*Named Entity Recognition* - NER) e resposta automática a perguntas (*Question Answering* - QA), entre outras. São exemplos de modelos *encoder* o BERT (Google), o RoBERTa (Meta), o XLM-RoBERTa (Meta), o ALBERT (Google e Toyota Technological Institute), o BERTimbau (Neuralmind e Universidade Estadual de Campinas) e o Albertina (Universidade de Lisboa e Universidade do Porto).

Na Figura 2.1, observa-se as diferenças na arquitetura entre os modelos *decoder* e *encoder*, especificamente entre o modelo GPT (*Generative Pre-trained Transformer*) e o modelo BERT (*Bidirectional Encoder Representations from Transformers*), onde se observa que o modelo GPT utiliza um *transformer* unidirecional de esquerda para a direita, enquanto o BERT utiliza um *transformer* bidirecional.



Fonte: Devlin et al. (2019)

Figura 2.1: Arquitetura dos Modelos *Transformers* GPT e BERT

No âmbito da presente dissertação, serão descritos de forma sucinta os modelos de linguagem pré-treinados selecionados para o projeto, destacando as suas principais características.

2.2.1. EuroLLM

Desenvolvido em 2024 com o apoio da União Europeia, o modelo *decoder* EuroLLM integra um projeto cujo objetivo central consiste no desenvolvimento de modelos de linguagem de código aberto aptos a compreender e gerar texto em todas as línguas oficiais da União Europeia e noutras línguas relevantes (nomeadamente árabe, japonês, hindi, coreano, turco, ucraniano, russo e chinês), contribuindo para a promoção da inclusão linguística e da soberania digital europeia (Martins, 2024).

O projeto foi levado a cabo por um consórcio composto por universidades e centros de investigação, designadamente o Instituto Superior Técnico (Portugal), o Instituto de Telecomunicações (Portugal), a Universidade de Edimburgo (Reino Unido), a Universidade Paris-Saclay (França), a Universidade de Amesterdão (Países Baixos) e a Sorbonne Université (França), bem como por empresas do setor tecnológico, nomeadamente a Unbabel (Portugal), a Aveni (Reino Unido) e a Naver Labs (Coreia do Sul).

O EuroLLM-9B-Instruct, utilizado nesta dissertação, é uma versão ajustada para seguir instruções (*instruction-tuned large language model*) do modelo base EuroLLM-9B, composto por 9,15 mil milhões de parâmetros, isto é, variáveis ajustáveis ou pesos utilizados para aprender padrões e relações na linguagem durante o treino. Apresenta uma arquitetura do tipo *transformer* com 42 blocos e 32 cabeças de atenção. Cada estado oculto possui dimensão de 4 096 unidades e a camada *feed-forward* (camada totalmente ligada em cada bloco *transformers*) amplia os vetores para 12 288 unidades. O comprimento máximo da sequência de entrada é de 4 096 *tokens*.

O modelo foi pré-treinado com 4 mil milhões de *tokens* em dados multilingues provenientes de diversas fontes, incluindo dados da *web*, dados paralelos (frases ou documentos com as respetivas traduções noutra língua) e conjuntos de dados de alta qualidade (Wikipedia, a base de dados de artigos científicos arXiv, a coleção de livros BookCorpus e a coleção de documentos médicos Apollo), tendo sido posteriormente ajustado com o conjunto de dados EuroBlocks, com foco no seguimento de instruções e na tradução automática (Martins, 2024; Martins et al., 2025).

2.2.2. BERT

Em 2018, a Google AI Language apresentou o modelo BERT. Ao contrário de abordagens baseadas em modelos de linguagem unidirecionais – como o GPT, também apresentado em 2018 pela OpenAI –, o BERT foi desenhado com o propósito de "pré-treinar representações bidirecionais profundas a partir de texto não categorizado, condicionando em simultâneo o contexto à esquerda e à direita em todas as camadas" (Devlin et al., 2019, p. 1, tradução própria).

De acordo com os autores, a relevância do BERT assenta na possibilidade de o modelo pré-treinado ser adaptado a diferentes tarefas com a simples adição de uma camada de saída, alcançando resultados de referência em domínios como a resposta automática a perguntas e a inferência textual (Devlin et al., 2019).

A versão do modelo utilizada neste trabalho, *multilingual uncased* (mBERT), foi concebida para lidar com múltiplas línguas, incluindo o português, sem distinção de maiúsculas e minúsculas, e apresenta um total de 110 milhões de parâmetros. A arquitetura da rede *transformer* é composta por 12 camadas (blocos *transformers*) e 12 cabeças de atenção. A dimensão dos estados ocultos é de 768 unidades, enquanto a dimensão da camada *feed-forward* corresponde a 3 072. O comprimento máximo de sequência para a entrada é de 512 *tokens*. Esta configuração estabeleceu-se como padrão de referência, servindo de base para o desenvolvimento de diversos modelos subsequentes, alguns dos quais serão descritos nas secções seguintes.

O processo de pré-treino baseou-se em duas tarefas não supervisionadas. A primeira foi a tarefa de MLM, que consiste em mascarar aleatoriamente 15% dos *tokens* de entrada e treinar o modelo para prever os *tokens* originais com base no contexto bidirecional. A segunda foi a tarefa de *Next Sentence Prediction* (NSP), que visa treinar o modelo a identificar se uma frase segue logicamente outra no texto de origem. O pré-treino foi realizado com dados de Wikipedia em 102 línguas (Devlin et al., 2019; Hugging Face, 2022).

2.2.3. XLM-RoBERTa

O modelo XLM-RoBERTa (XLM-R) foi desenvolvido pela equipa do Facebook AI (Meta) em 2019. Segundo os autores, o principal objetivo do modelo é melhorar as capacidades de transferência entre

línguas (*cross-lingual transfer*) e ultrapassar limitações de modelos multilingues anteriores, como o mBERT, especialmente em contextos de línguas de poucos recursos (Conneau et al., 2020).

A versão utilizada neste trabalho, XLM-R *base*, integra aproximadamente 270 milhões de parâmetros. O processo de pré-treino seguiu a abordagem de MLM, recorrendo ao corpus CC-100, construído a partir do CommonCrawl, o qual abrange 100 línguas (entre as quais se encontra o português) e reúne aproximadamente 2,5 TB de dados. Os autores salientam que o XLM-R supera significativamente o mBERT em diferentes tarefas avaliadas por testes de referência (XNLI, MLQA, CoNLL-2002, CoNLL-2003 e GLUE)[6] (Conneau et al., 2020).

2.2.4. BERTimbau

O modelo BERTimbau foi proposto pela equipa da Neuralmind (Brasil) e da Universidade Estadual de Campinas (Brasil) em 2020, com o objetivo de disponibilizar representações linguísticas pré-treinadas para o português do Brasil, melhorando o desempenho em tarefas de NLP com *transfer learning* (Souza et al., 2023).

Para o presente projeto utilizou-se o modelo BERTimbau *base*, versão que integra cerca de 110 milhões de parâmetros. O pré-treino do modelo seguiu a abordagem do BERT, recorrendo às técnicas de MLM e NSP, e utilizou como base de dados o brWaC (*Brazilian Web as Corpus*) (Filho et al., 2018), um vasto conjunto de páginas *web* brasileiras composto por 3,53 milhões de documentos.

O BERTimbau demonstrou ganhos significativos em tarefas de similaridade textual, inferência textual e NER na língua portuguesa, comparativamente ao mBERT (Souza et al., 2023).

2.2.5. Albertina

O modelo Albertina foi criado em 2023 por uma equipa da Universidade de Lisboa e da Universidade do Porto, no âmbito do projeto PORTULAN CLARIN, com o objetivo de disponibilizar um modelo *encoder* de última geração especificamente concebido para o português europeu e para o português do Brasil. Segundo os autores, a principal motivação consistiu em colmatar a ausência de *encoders* públicos e competitivos para o português europeu, assegurando "um contributo para o avanço da investigação e da inovação em tecnologias da linguagem para o português" (Rodrigues et al., 2023, p. 1, tradução própria).

[6] Os testes referidos avaliam diferentes tarefas de NLP. O XNLI testa a inferência textual multilingue (relação lógica entre premissas e hipóteses); o MLQA avalia a capacidade de resposta automática a perguntas com base em textos em múltiplas línguas; os CoNLL-2002 e CoNLL-2003 são utilizados para reconhecimento de entidades mencionadas (pessoas, locais, organizações, entre outras); e o GLUE integra um conjunto diversificado de tarefas, como inferência textual, análise de sentimentos, paráfrase e identificação de perguntas duplicadas, com o intuito de medir a compreensão linguística geral dos modelos.

A versão utilizada nesta dissertação, Albertina PT-PT *Base*, integra 100 milhões de parâmetros e o seu pré-treino baseou-se no modelo DeBERTa, recorrendo a um conjunto diversificado de dados em português europeu, nomeadamente o OSCAR, o European Parliament Proceedings Parallel Corpus (Europarl), o Digital Corpus of the European Parliament (DCEP) e o ParlamentoPT, perfazendo cerca de 8 milhões de documentos.

De acordo com os resultados reportados, o Albertina PT-PT estabeleceu um novo estado da arte em testes de referência adaptados ao português, demonstrando desempenho superior ao BERTimbau em tarefas como inferência textual e paráfrase (Rodrigues et al., 2023).

2.3. Classificação Temática das Publicações Periódicas em Portugal

Em Portugal, as publicações periódicas enquadram-se no conceito de imprensa, conforme o estipulado na Lei de Imprensa (1999). Este diploma define-as como reproduções de textos ou imagens, disponibilizadas ao público em série contínua, publicadas com um título fixo, uma periodicidade definida e em formato impresso e/ou digital. Integram este conjunto os jornais, as revistas, os boletins e os anuários editados em fascículos ou volumes sucessivos, independentemente do suporte utilizado para a sua divulgação.

Enquanto meios de comunicação social, as publicações periódicas estão sujeitas à supervisão e à regulação da ERC, a quem compete o registo de todas as entidades editoras (Assembleia da República, 2005). O processo de registo implica a classificação de cada publicação segundo os critérios definidos no artigo 10.º da Lei de Imprensa (1999): a origem da edição (portuguesa ou estrangeira); o âmbito ou interesse geográfico do tema (nacional, regional ou destinadas às comunidades estrangeiras); e o seu conteúdo (doutrinária, de informação geral e de informação especializada).

Para uma caracterização mais completa das publicações periódicas, a ERC recolhe um conjunto de dados para além dos legalmente exigidos, como o tipo de suporte (digital, impresso ou ambos), a periodicidade, o *website*, a temática principal, entre outros. Visto que estes elementos não correspondem a uma categorização definida na lei, compete aos serviços da ERC estabelecer os critérios para a sua classificação.

A Lei de Imprensa (1999), apesar de não prever uma lista formal de categorias temáticas, fornece no seu artigo 13.º uma base para a diferenciação das publicações através do seu conteúdo. Este artigo distingue as publicações doutrinárias das informativas. As primeiras destinam-se à divulgação de uma ideologia ou credo. As segundas, focadas na difusão de informações ou notícias, dividem-se em duas modalidades: de informação geral e de informação especializada. Uma publicação é de informação geral quando o seu conteúdo predominante são notícias ou informações não especializadas. Por outro

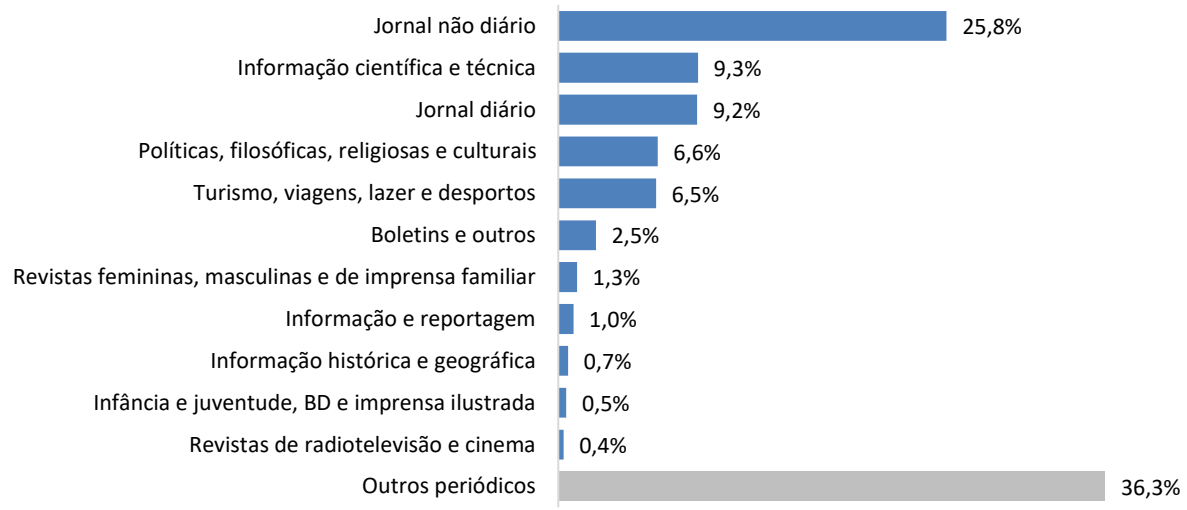
lado, é considerada de informação especializada aquela que dedica grande parte do seu conteúdo a um domínio específico, como a ciência, a literatura, a arte ou o desporto.

De acordo com a base de registos da ERC, a 6 de fevereiro de 2025, encontravam-se ativas 1 680 publicações periódicas. Relativamente à natureza do seu conteúdo, 49,7% destas publicações eram de informação especializada, 44,6% de informação geral e os restantes 5,7% correspondiam a publicações doutrinárias.

A análise da classificação temática, efetuada pelos serviços da ERC, revelou que a categoria mais representativa era a de "Outros periódicos" (36,3%), seguida de "Jornal não diário" (25,8%). As restantes categorias apresentavam valores inferiores a 10%, conforme se ilustra na Figura 2.2.

O sistema de classificação temática atualmente utilizado pela ERC apresenta diversas limitações, uma vez que as categorias combinam dimensões distintas, como a periodicidade (ex.: diário, não diário), o tipo de publicação (ex.: jornal, revista) e o conteúdo temático (ex.: turismo, desporto). Esta sobreposição conduz a uma categorização ambígua, dificultando a análise da realidade do setor da imprensa em Portugal. Em consequência destas limitações, a ERC não procede à divulgação pública dos dados com base nesta categorização.

Figura 2.2: Publicações Periódicas por Classificação Temática da ERC*



N = 1 680
Fonte: ERC.
* A ERC reconhece as limitações do seu atual sistema de classificação temática, bem como a existência de classificações inadequadas em diversas publicações periódicas. Por este motivo, a entidade opta por não divulgar estes dados.

CAPÍTULO 3

Revisão de Literatura

Apesar de ser uma das línguas mais faladas no mundo, o português[7] é considerado uma língua com poucos recursos no campo do NLP (Garcia et al., 2024; Ramos et al., 2024; Santos et al., 2021). Isto deve-se à escassez de conjuntos de dados de grande dimensão anotados e de modelos pré-treinados específicos, quando comparado com o inglês. A maioria das investigações e dos recursos disponíveis foca-se no português do Brasil. Esta disparidade leva a que muitas investigações em português europeu que envolvam a aplicação de modelos *transformers*, recorram a modelos multilingues ou modelos treinados na variante brasileira.

O modelo BERTimbau surge como *encoder* de referência em estudos com dados em português. Por seu turno, o mBERT é utilizado como termo de comparação quando se pretende avaliar o comportamento de um modelo multilingue em conjuntos de dados em português.

Embora tenham sido disponibilizados recentemente modelos específicos para o português europeu, como a família Albertina, a sua adoção na investigação encontra-se numa fase inicial. No estudo de Ramos et al. (2024), o modelo não foi incluído porque, à data da investigação, apenas estava disponível a versão *large*, de elevado custo computacional, e os ensaios iniciais não indicaram melhorias no desempenho (p. 101379).

A aplicação de *transformers* pré-treinados em classificação de texto em português europeu ganhou expressão em diversos contextos, como o da saúde, do jurídico-legislativo, da educação, das redes sociais e conteúdos mediáticos, entre outros, demonstrando um desempenho superior quando comparado com abordagens clássicas de aprendizagem automática e de aprendizagem profunda. Sem pretender realizar uma revisão exaustiva da literatura, nos parágrafos seguintes detalham-se alguns estudos para ilustrar a aplicação destes modelos em cada um dos domínios mencionados.

Na saúde, Yang et al. (2022) compararam arquiteturas de *deep learning* com o modelo BERTimbau na classificação automática de chamadas para a linha de saúde do Serviço Nacional de Saúde de Portugal (SNS24). O objetivo consistiu em prever o percurso clínico mais adequado com base na descrição textual dos sintomas, visando melhorar o desempenho do serviço de triagem telefónica. Nos testes preliminares, que consideraram a configuração predefinida de todas as redes, a arquitetura baseada em *transformers* obteve o melhor desempenho, com uma acurácia de 78,2%, superando as redes CNN (acurácia = 76,6%) e RNN (acurácia = 75,9%).

[7] Segundo o Instituto Camões (2023), o português é a quarta língua mais falada no mundo como língua materna, a seguir ao mandarim, inglês e espanhol.

Posteriormente, os autores desenvolveram um processo estruturado de ajuste dos hiperparâmetros, concebido para otimizar o desempenho do modelo *encoder* e reduzir o custo computacional do treino. Esse processo incidiu em três parâmetros: a taxa de aprendizagem (*learning rate*), o número de épocas (*epochs*) e o tamanho do lote (*batch size*), ajustados de forma sequencial para identificar a melhor combinação. Numa primeira fase, foram testadas várias taxas de aprendizagem com os otimizadores Adam e a Descida do Gradiente Estocástico (*Stochastic Gradient Descent* - SGD), de modo a observar o impacto na convergência do modelo. Numa segunda fase, o número de épocas foi aumentado progressivamente, avaliando-se o ponto de sobreajustamento a partir das curvas de função de perda e acurácia. Por fim, procedeu-se à afinação do tamanho do lote, testando diferentes valores para maximizar a estabilidade e a generalização do modelo. O melhor desempenho foi alcançado com o otimizador Adam, uma taxa de aprendizagem de 1×10^{-6} e um tamanho do lote de 16, configuração que permitiu atingir uma acurácia do modelo BERTimbau de 80,6%.

No contexto jurídico-legislativo, o estudo de Costa et al. (2024), focado na identificação automática de encargos administrativos em documentos legislativos portugueses, evidenciou a importância da adaptação dos modelos pré-treinados baseados em *transformers* ao domínio ou à tarefa em causa, como alternativa perante a escassez e/ou desequilíbrio de dados. O *corpus* trabalhado pelos autores compreendeu 19 465 passagens de texto, das quais apenas 5,89% continham encargos, evidenciando um acentuado desequilíbrio entre classes. Inicialmente, o modelo BERTimbau, apenas ajustado com os dados classificados, resultou num desempenho de F1 igual a 50,0%.

Para melhorar o processo de classificação, os autores introduziram uma fase adicional de pré-treino com 262 documentos legais não anotados, na qual o modelo foi submetido a um treino não supervisionado durante 100 épocas, utilizando o otimizador AdamW e uma taxa de aprendizagem de 5×10^{-5}. Esta etapa, baseada na técnica MLM, permitiu ao modelo aprender a estrutura sintática e o vocabulário jurídico português, melhorando a sua capacidade de representação linguística no domínio legislativo. Após esta fase, o modelo foi ajustado à tarefa de classificação binária, alcançando um F1 de 79,4%.

Por seu lado, Santos et al. (2021), no campo da educação, demonstraram como os *transformers* pré-treinados podem auxiliar na criação de materiais pedagógicos adequados ao nível de proficiência dos estudantes. Este estudo, focado no ensino de português como segunda língua, aplicou modelos como GPT-2 e RoBERTa para classificar textos segundo os níveis do Quadro Europeu Comum de Referência para as Línguas (QECR), estabelecendo um novo estado da arte para esta tarefa, mesmo com conjuntos de dados de dimensão reduzida. O modelo GPT-2, já pré-treinado e posteriormente adaptado do inglês para português, foi diretamente ajustado ao *corpus* anotado de 500 textos fornecido pelo Camões I.P. Enquanto o RoBERTa foi treinado de raiz, pelo que os autores recorreram

a 10 milhões de frases em português e 10 milhões em inglês provenientes do *corpus* multilingue OSCAR, antes de ser também ajustado à tarefa de classificação. O GPT-2 revelou-se o modelo com melhor desempenho com uma acurácia em torno de 75%. Segundo os autores, os resultados obtidos através de modelos pré-treinados baseados em *transformers* são "boas notícias para as línguas sub-representadas que não possuem as ferramentas necessárias para a extração de características" (p. 723, tradução própria) exigida por outras abordagens.

Os modelos *transformer* têm ganho relevância como alternativa na moderação de conteúdos em redes sociais, especialmente na deteção de discurso de ódio, destacando-se os estudos de Fialho et al. (2024) e de Ramos et al. (2024), ambos desenvolvidos no projeto kNOwHATE, financiado pela União Europeia. O primeiro estudo, aplicado a comentários em YouTube, foca-se na classificação de um comentário como "neutro", "discurso de ódio" ou "contradiscurso", considerando sempre o comentário anterior como contexto. Para tal, Fialho et al. (2024) compararam o desempenho de quatro modelos *transformer*: o mBERT, o XLM-R, o BERTimbau e um modelo pré-treinado em discurso de ódio em português o HateBERTimbau. Os modelos foram treinados com o otimizador Adam, com uma taxa de aprendizagem de 1×10^{-5}, durante 10 épocas e com um tamanho do lote de 16. Os resultados indicaram que em termos gerais o XLM-R apresentou o melhor desempenho médio (F1 = 73,2%), ainda assim a análise por categoria destacou o modelo BERTimbau na categoria "contradiscurso" (F1 = 41,7%), o modelo HateBERTimbau na categoria "discurso de ódio" (F1 = 78,6%) e o XLM-R na categoria "neutro" (F1 = 76,2%).

Por sua vez, o estudo de Ramos et al. (2024) abordou a tarefa de classificação binária de deteção de discurso de ódio, comparando o desempenho de modelos *decoder* e *encoder* em comentários do YouTube e Twitter (atualmente X). Os autores testaram os modelos *decoder* GPT-3.5, GPT-4, Gemini-Pro e Mistral-7B-Instruct-v0.3, introduzindo contexto ou não na instrução (*prompt*), e os modelos *encoder* BERTimbau, BERTimbau-hatebr, mDeBERTa-hatebr e HateBERTimbau. Apesar dos modelos BERTimbau-hatebr e mDeBERTa-hatebr serem treinados com dados relativos ao discurso de ódio, os autores realizaram um pré-treino com os dados utilizados no estudo, o que resultou em melhorias de desempenho. O processo de treino para a tarefa de classificação seguiu as recomendações de Devlin et al. (2019), com um tamanho do lote de 32, taxa de aprendizagem de 2×10^{-5} e três épocas. Para o conjunto de dados do YouTube, o modelo HateBERTimbau apresentou o melhor desempenho (F1 = 87,1%), corroborando a ideia de que os modelos *transformers* pré-treinados para a tarefa tendem a ser mais eficazes; quanto aos dados da plataforma Twitter, o modelo *decoder* GPT-3.5 sem contexto destacou-se (F1 = 50,2%).

No que diz respeito aos estudos sobre a categorização de conteúdos mediáticos em português europeu, Yanchuk et al. (2022) testaram o desempenho de modelos de aprendizagem automática e de aprendizagem profunda, incluindo o modelo BERTimbau, na classificação binária de notícias como

“estigmatizantes” ou “literais”, em referência aos termos esquizofrenia e psicose. Os resultados indicaram que o modelo BERTimbau apresentou valores de acurácia de 92,0%, ultrapassando as redes neuronais de memória de longo e curto prazo bidirecional (*Bidirectional Long Short-Term Memory* – Bi-LSTM) (acurácia = 91,3%), CNN (acurácia = 87,8%) e de memória de longo e curto prazo (*Long Short-Term Memory* – LSTM) (acurácia = 87,2%).

Contudo, o estudo destaca-se por demonstrar que o BERTimbau não conseguiu superar as abordagens clássicas de aprendizagem automática. Os classificadores Naive Bayes e Regressão Logística, utilizando uma representação vetorial baseada na técnica Frequência do Termo-Frequência Inversa do Documento (*Term Frequency-Inverse Document Frequency* – TF-IDF), alcançaram valores de acurácia superiores, de 93,4% cada. Este resultado evidencia que, em cenários específicos como o analisado, um modelo *encoder*, mesmo sem uma fase de pré-treino com dados do domínio, demonstra uma elevada competitividade, alcançando um desempenho próximo das abordagens clássicas otimizadas.

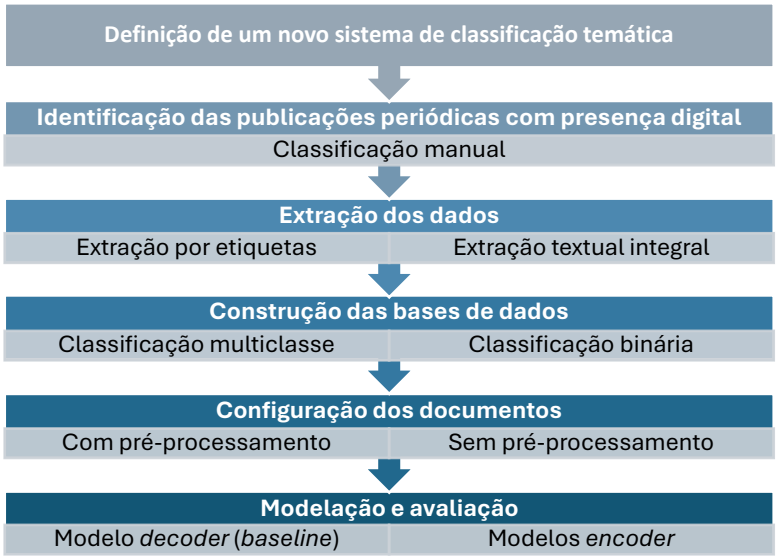
Quanto à classificação temática de conteúdos mediáticos em português, sobressaem os estudos aplicados a notícias em português do Brasil. A título de exemplo, Garcia et al. (2024) realizaram uma análise comparativa entre métodos clássicos de aprendizagem automática, redes neuronais e o modelo *transformer* BERTimbau, avaliando o desempenho de classificação e os custos computacionais. O BERTimbau, ajustado aos dados previamente classificados, apresentou os melhores resultados nas diferentes bases de dados testadas, ainda que com tempos de processamento superiores às restantes abordagens. O classificador Máquina de Vetores de Suporte (*Support Vector Machine* – SVM) com TF-IDF revelou-se competitivo quando o custo computacional é determinante, combinando bom desempenho com baixa exigência de recursos. Ainda assim, os autores salientam a simplicidade operacional do BERTimbau por eliminar etapas de pré-processamento, pois os resultados foram obtidos com dados em bruto, e recomendam-no quando a acurácia é o fator crítico, apesar do maior custo computacional (p. 12).

CAPÍTULO 4

Abordagem Metodológica

O presente capítulo descreve a metodologia aplicada para o desenvolvimento do processo de classificação temática automatizada das publicações periódicas registadas na ERC com presença digital. O objetivo central consistiu em identificar, para cada publicação, uma única categoria temática, isto é, o tema dominante, a partir do conteúdo textual disponibilizado nos respetivos sítios eletrónicos.

O processo envolveu várias etapas, como se observa na Figura 4.1, que compreenderam a definição de um novo sistema de classificação temática, a categorização manual das publicações, a extração de dados, a construção das bases de dados segundo o tipo de classificação, a preparação dos documentos, e, por fim, o treino e avaliação das experiências realizadas com os modelos pré-treinados baseados em *transformers,* ajustados para a tarefa de classificação de texto.



Fonte: Elaboração própria.

Figura 4.1: Esquema Geral do Processo Metodológico Aplicado à Classificação Temática das Publicações Periódicas

4.1. Classificação Manual

Antes da classificação manual das publicações periódicas, elaborou-se um sistema de categorias temáticas, para o qual contribuíram membros do Conselho Regulador, da Unidade de Registo e técnicos da ERC. O sistema, ilustrado na Tabela 4.1, compreende nove categorias temáticas, organizadas por tipo de conteúdo. A sua conceção baseou-se nos sistemas de categorização de publicações periódicas aplicados pelo Instituto Nacional de Estatística (INE) e pela Associação Portuguesa para o Controlo de Tiragem e Circulação (APCT), os quais se apresentam no Anexo A.

Tabela 4.1: Sistema de Categorias Temáticas

Conteúdo	**Categoria temática**
Informação geral	Geral
Informação especializada	Desporto Economia, finanças, gestão e negócios Artes e cultura Estilo de vida e lazer Ciência, educação e história Política, administração pública, direito e sociedade
Doutrinária	Religião Doutrinas políticas/sociais

Fonte: Elaboração própria.

A classificação temática de cada publicação teve como critério principal a análise do seu conteúdo. No entanto, foram igualmente ponderados outros elementos, como o estatuto editorial, as descrições da publicação e da entidade proprietária, e a definição do público-alvo, caso estivesse disponível no sítio eletrónico da publicação periódica. Para apoiar a aplicação dos critérios de classificação manual e assegurar a sua coerência interna, definiram-se subcategorias temáticas específicas para cada categoria (ver Anexo A).

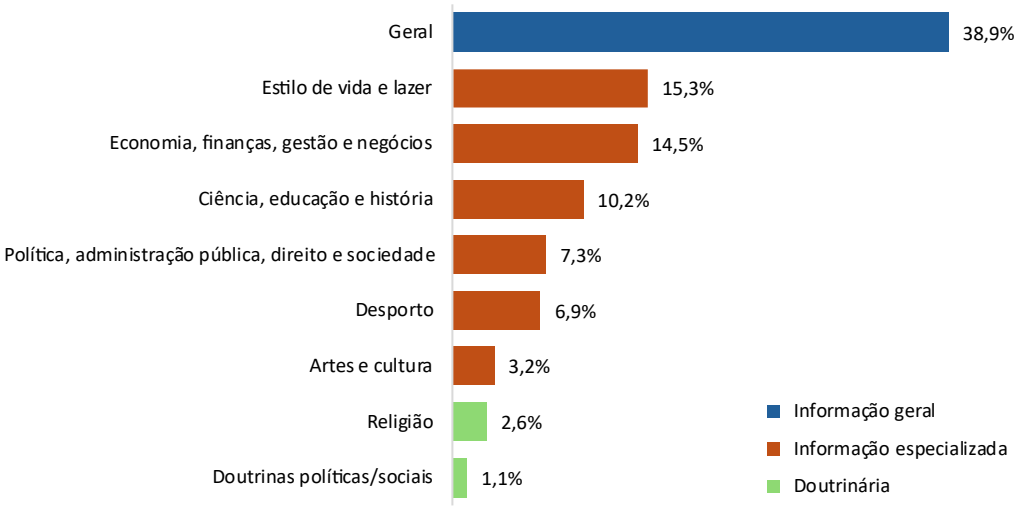
Identificou-se como objeto de estudo as publicações periódicas com presença digital, exclusiva ou também divulgada em papel. Estas publicações podem ser um documento digital ou um sítio eletrónico. Nos registos da ERC, a 6 de fevereiro de 2025, totalizavam 1 185 títulos com estas características, o que correspondia a 70,5% do universo de publicações registadas; as restantes eram publicações divulgadas exclusivamente em formato impresso.

Do total de publicações com presença digital, excluíram-se 197 publicações por motivos técnicos ou metodológicos. As exclusões técnicas deveram-se à ausência de registo do sítio eletrónico na base da ERC (49 publicações) ou à impossibilidade de definir um tema dominante a partir da informação disponível no *website* da publicação (38 publicações). A exclusão metodológica abrangeu 110 publicações periódicas, previamente definidas como de informação geral. Esta exclusão justifica-se por se considerar que o número de publicações classificadas como "Geral" era já suficiente para os objetivos do estudo.

Resultou deste processo a classificação do tema dominante de 988 publicações periódicas. Como se observa na Figura 4.2, a maioria (57,4%) enquadra-se nas categorias definidas para as publicações de informação especializada, seguindo-se as publicações de informação geral (38,9%) e, por fim, as de conteúdo doutrinário (3,7%).

Dentro do grupo de informação especializada, as categorias com maior expressão são as de "Estilo de vida e lazer" (15,3%) e "Economia, finanças, gestão e negócios" (14,5%). No que respeita às

publicações doutrinárias, a temática religiosa representa 2,6% da amostra total, enquanto a de "Doutrinas políticas/sociais" corresponde a 1,1%.



N = 988
Fonte: Elaboração própria.

Figura 4.2: Publicações Periódicas Classificadas Manualmente

Considerando que as publicações de informação especializada e as doutrinárias difundem conteúdos maioritariamente centrados num único tema, enquanto as de informação geral abordam uma diversidade temática, optou-se por numa primeira fase, realizar a extração de dados e a classificação do tema dominante das publicações de informação especializada e doutrinária (classificação multiclasse). Numa segunda fase, procedeu-se à extração de dados das publicações de informação geral e à aplicação da classificação em articulação com os dados previamente extraídos das publicações especializadas ou doutrinárias (classificação binária).

4.2. Extração de Dados

Considerando o volume de publicações periódicas, recorreu-se a técnicas de *web scraping* com o objetivo de extrair o conteúdo textual disponibilizado nos respetivos sítios eletrónicos. Para este processo utilizou-se a biblioteca de Python BeautifulSoup,[8] que permitiu recolher e estruturar os dados extraídos numa base de dados destinada à posterior classificação automática.

Nos casos em que as publicações estavam apenas disponíveis em formato de documento digital transferível, a extração não incidiu sobre o ficheiro em si. Em alternativa, procedeu-se à extração do texto a partir da página *web* que continha a informação mais relevante para a definição do tema da publicação periódica.

Definiram-se duas abordagens distintas de extração:

[8] Consultar: https://www.crummy.com/software/BeautifulSoup/

1. Extração textual integral, baseada no método *get_text()*, que permite obter apenas o texto visível em cada elemento HTML (*HyperText Markup Language*), removendo código, atributos e elementos aninhados.
2. Extração por etiquetas, que consiste na recolha seletiva de texto associado a um conjunto de etiquetas HTML previamente identificadas como potencialmente informativas.

Na segunda abordagem, avaliou-se previamente a relevância do texto contido nos diferentes atributos HTML para classificar o tema da publicação. Para esse efeito, realizou-se uma experiência exploratória em 150 sítios eletrónicos de publicações periódicas de informação especializada ou doutrinária, selecionados de forma aleatória.

Numa primeira etapa, examinou-se o código HTML de dez sítios, identificando-se os principais atributos passíveis de conter texto relevante: cabeçalhos (<h1>, <h2>, <h3>, <h4>, <h5> e <h6>), parágrafos (<p>), âncoras com hiperligação (<a>), elementos de formatação (<b>, , , <i>, <small>, <mark>, <ins>), citações curtas e longas (<q>, <blockquote>), tabelas (<td>) e listas (<li>). Posteriormente, procedeu-se à extração do conteúdo dessas etiquetas, assim como à sua avaliação qualitativa.

Tabela 4.2: Extração por Etiquetas e Relevância da Informação em 150 Sítios Eletrónicos de Publicações Periódicas

Elemento HTML (Etiqueta)	**Extração de texto**		**Texto relevante para a classificação**	
	Nº de sítios	**%**	**Nº de sítios**	**%**
a	139	92,7%	112	80,6%
p	125	83,3%	88	70,4%
li	134	89,3%	70	52,2%
h3	101	67,3%	47	46,5%
h2	105	70,0%	48	45,7%
h4	63	42,0%	19	30,2%
h5	25	16,7%	5	20,0%
h1	77	51,3%	8	10,4%
td	12	8,0%	1	8,3%
strong	54	36,0%	3	5,6%
h6	18	12,0%	1	5,6%
b	20	13,3%	0	0,0%
em	11	7,3%	0	0,0%
small	11	7,3%	0	0,0%
blockquote	6	4,0%	0	0,0%
i	5	3,3%	0	0,0%
mark	1	0,7%	0	0,0%
ins	0	0,0%	0	-
q	0	0,0%	0	-

Fonte: Elaboração própria.

Dos 150 sítios testados, em apenas 11 não foi possível recolher dados; nos restantes, obteve-se texto em um ou mais dos atributos identificados, destacando-se as etiquetas <a>, <li> e <p>, presentes em 92,7%, 89,3% e 83,3% dos sítios, respetivamente, como se observa na Tabela 4.2. A análise do conteúdo recolhido demonstrou que 80,6% do texto contido em <a> e 70,4% do texto em <p> apresentaram informação relevante para a classificação temática das publicações. Em contrapartida, a relevância do texto extraído de atributos como listas, cabeçalhos, tabelas e do elemento de ênfase foi menor, com valores que se situaram entre 5,6% e 52,2%. Por seu lado, as etiquetas <b>, , <small>, <blockquote>, <i>, <mark>, <ins> e <q> revelaram utilidade nula para este fim. Assim, concluiu-se que as etiquetas <a> e <p> constituem as fontes mais presentes e consistentes de dados textuais relevantes, razão pela qual a extração por etiquetas incidiu sobre estas duas.

Importa referir que os dados extraídos correspondem a informação pública, não pessoal, disponível nos sítios eletrónicos das publicações. Não foi necessária autenticação ou criação de conta para aceder à informação, tendo-se respeitado os casos em que o acesso era tecnicamente negado. Avaliaram-se igualmente os potenciais danos que o processo de *scraping* poderia causar aos utilizadores individuais e às plataformas, os quais se demonstraram nulos, e implementaram-se técnicas de limitação da taxa de pedidos de forma a reduzir a carga de extração e a evitar instabilidades nos sítios eletrónicos. Assim, a recolha de dados, conduzida de forma responsável e com finalidade exclusivamente científica, encontra-se alinhada com os padrões éticos reconhecidos para a investigação com recurso a *web scraping* (Brown et al., 2024).

A extração de dados textuais foi realizada de forma faseada, numa primeira fase para as publicações especializadas ou doutrinárias e, posteriormente, para as publicações classificadas como "Geral". Depois da extração, removeram-se os registos duplicados, as publicações que negaram o acesso e os casos em que o conteúdo se encontrava vazio. Cada sítio eletrónico extraído foi arquivado com indicação da data e da hora da extração.

No conjunto de publicações de informação especializada ou doutrinária, a abordagem orientada na extração exclusiva de texto visível permitiu obter dados de 583 publicações periódicas, enquanto a abordagem baseada na extração de texto contido em etiquetas específicas forneceu informação de 574 publicações. No caso das publicações de informação geral, ambas as abordagens resultaram na recolha de dados de 373 publicações cada.

4.3. Base de Dados

O processo de classificação de texto foi desenvolvido em duas etapas. Na primeira, realizou-se a classificação multiclasse do conjunto de publicações periódicas previamente categorizadas nas oito classes temáticas definidas para conteúdos especializados ou doutrinários. Na segunda, procedeu-se

à classificação binária, distinguindo entre as publicações periódicas de informação geral e as publicações de informação especializada ou doutrinária. Para este fim, construíram-se duas bases de dados, ambas compostas pelos indicadores apresentados no Anexo B, incluindo o texto extraído segundo o tipo de abordagem descrita na secção 4.2.

Com vista a garantir a comparabilidade, consideraram-se apenas as publicações periódicas com dados extraídos por ambas as abordagens. Assim, eliminaram-se os casos em que apenas uma das abordagens de extração forneceu dados. Este processo de uniformização resultou numa base de dados de 570 publicações para a classificação multiclasse e de 736 publicações para a classificação binária.[9]

Para identificar a forma mais adequada de tratar os documentos, avaliou-se o desempenho da classificação entre os dados em bruto e os dados após o pré-processamento. Nesse sentido, foi aplicado um conjunto de transformações com vista à limpeza e normalização do texto. Esta etapa de pré-processamento envolveu os seguintes procedimentos:

- A correção de erros de codificação de caracteres, assegurando que todos os textos fossem convertidos para um formato uniforme (UTF-8). Simultaneamente, eliminaram-se elementos de ruído resultantes da presença de caracteres inválidos ou não reconhecidos no padrão Unicode.
- A eliminação de ligações, endereços de correio eletrónico, números de telefone, fax e telemóvel, bem como identificadores de depósito legal, IBAN ou números extensos sem relevância semântica. Do mesmo modo, retiraram-se referências a moradas e códigos postais.
- A remoção de termos associados à navegação e à estrutura dos sítios eletrónicos (por exemplo: "*home*", "contactos", "política de privacidade", "subscreva a nossa *newsletter*", "ler mais", entre outros), de forma a preservar apenas o conteúdo editorial relevante.
- A identificação e eliminação de frases e sequências de palavras duplicadas.
- A remoção de caracteres especiais que não acrescentavam significado semântico (por exemplo, símbolos gráficos ou marcas tipográficas descontextualizadas), bem como dos sinais de pontuação que apareciam de forma isolada ou desarticulada do enunciado. Do mesmo modo, eliminaram-se combinações redundantes de pontuação (como múltiplos pontos, vírgulas ou hífenes consecutivos). A pontuação remanescente foi normalizada, de modo a respeitar as convenções ortográficas e a estrutura sintática das frases, assegurando a legibilidade e a coerência gramatical do texto.

[9] Após o processo de uniformização do conjunto de dados extraído das publicações de informação geral, obteve-se um total de 368 publicações. Para a construção da base de classificação binária, selecionou-se aleatoriamente uma amostra estratificada composta por 368 publicações de informação especializada ou doutrinária, que foram integradas com as publicações de informação geral.

- A conversão do texto para minúsculas para garantir a uniformidade lexical.

Na Tabela 4.3, apresentam-se as medidas descritivas dos textos extraídos pelas duas abordagens consideradas, segundo o estado de pré-processamento, para a base de dados final constituída pelas publicações periódicas de informação especializada ou doutrinária. Observa-se que a extração do texto integral tende a ter, em média, um volume superior de palavras por documento (1 321,7 antes do pré-processamento e 1 167,6 depois do pré-processamento) face à extração por etiquetas (1 093,2 antes e 801,6 depois).

Do mesmo modo, em ambas as técnicas de extração, verifica-se uma grande variabilidade nos tamanhos dos documentos, mas a extração textual integral apresenta uma dispersão superior (2 282,5 antes do pré-processamento e 2 138,9 depois), quando comparada com a extração por etiquetas (1 693,4 e 1 366,9, respetivamente). A mediana, indicador menos sensível a valores extremos, corrobora estas análises.

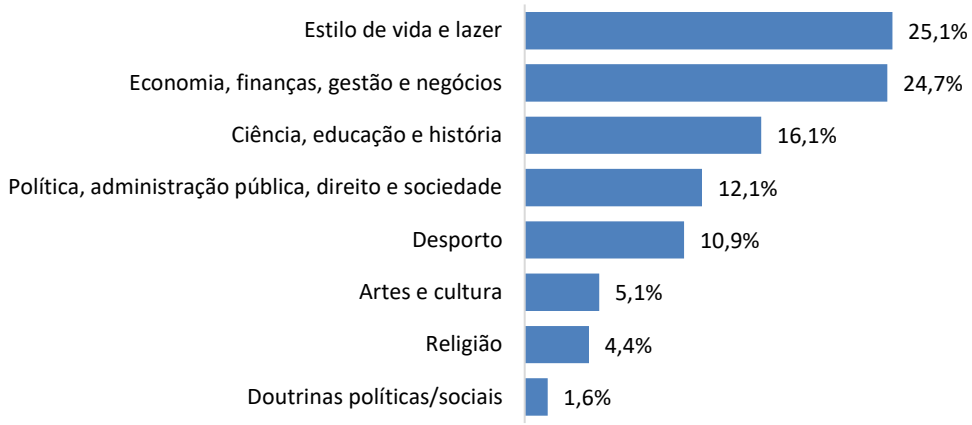
Como esperado, a aplicação das regras de limpeza e normalização resultou numa redução do volume de texto. Verificou-se que a redução média de palavras por documento foi maior na extração por etiquetas, que passou de uma média de 1 093,2 para 801,6 palavras (redução de 26,7%), enquanto na extração textual integral a média diminuiu de 1 321,7 para 1 167,6 (redução de 11,7%).

Tabela 4.3: Estatísticas Descritivas dos Textos Extraídos por Etiquetas e de Forma Integral, Antes e Após Pré-Processamento – Base de Dados para a Classificação Multiclasse

Medidas descritivas por documento	**Extração por etiquetas**		**Extração textual integral**	
	Texto extraído	**Texto pré-processado**	**Texto extraído**	**Texto pré-processado**
Média	1 093,2	801,6	1 321,7	1 167,6
Desvio padrão	1 693,4	1 366,9	2 282,5	2 138,9
Mediana	683,5	478,0	875,0	771,0
Mínimo	12	1	47	28
Máximo	15 613	14 208	41 861	40 234

Fonte: Elaboração própria.

A distribuição das categorias desta base de dados, representada na Figura 4.3, mantém a ordem de importância obtida no processo de categorização manual explicado anteriormente na secção 4.1, ou seja, as categorias “Estilo de vida e lazer” (25,1%) e “Economia, finanças, gestão e negócios” (24,7%) surgem com maior expressão, enquanto as publicações doutrinárias apresentam uma menor percentagem de publicações, especificamente, as de “Religião” 4,4% e as de “Doutrinas políticas/sociais” 1,6%.



N = 570
Fonte: Elaboração própria.

Figura 4.3: Publicações Periódicas por Categoria Temática – Base de Dados para a Classificação Multiclasse

Como se observa na Tabela 4.4, a análise descritiva da base de dados para a classificação binária confirma as tendências previamente identificadas no conjunto de dados para a classificação multiclasse. Também neste caso, a extração textual integral resultou em documentos de maior volume e variabilidade face à extração seletiva por etiquetas. Adicionalmente, o impacto do pré-processamento foi mais acentuado nos textos extraídos por etiquetas, que registaram uma redução de 29,6%, valor que supera em mais do dobro a redução verificada na extração integral (13,2%).

Tabela 4.4. Estatísticas Descritivas dos Textos Extraídos por Etiquetas e de Forma Integral, Antes e Após Pré-Processamento – Base de Dados para a Classificação Binária

Medidas descritivas por documento	**Extração por etiquetas**		**Extração textual integral**	
	Texto extraído	**Texto pré-processado**	**Texto extraído**	**Texto pré-processado**
Média	1 202,1	846,5	1 453,0	1 260,6
Desvio padrão	1 841,8	1 505,4	2 223,1	2 027,3
Mediana	797,5	523,5	1 040,5	911,0
Mínimo	9	3	21	11
Máximo	26 427	21 304	41 861	40 234

Fonte: Elaboração própria.

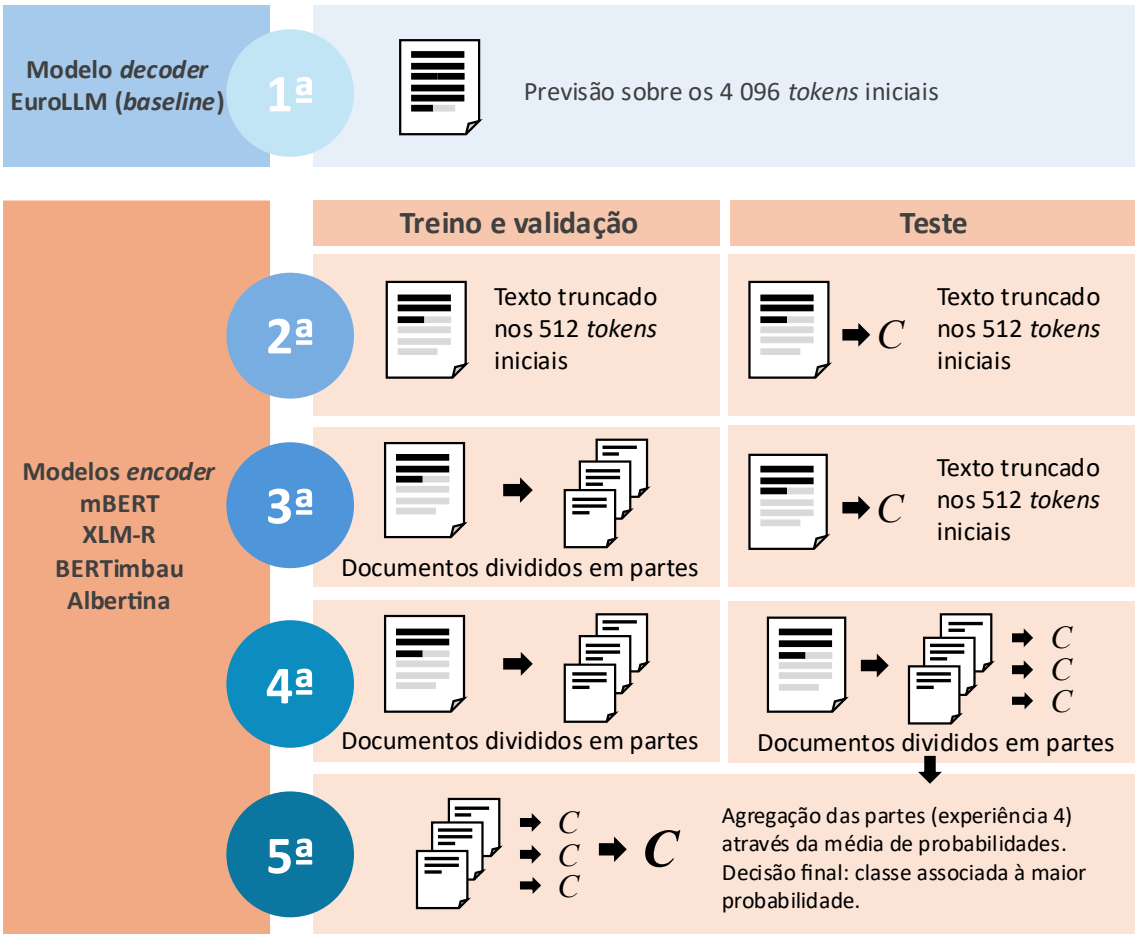
No que diz respeito à distribuição das categorias da base de dados para a classificação binária, 50,0% das publicações periódicas correspondem a "Informação geral" e os restantes 50,0% a "Informação especializada".[10]

[10] Dado o número reduzido e à especificidade temática das publicações doutrinárias, estas foram agregadas a esta categoria.

Para o processo de modelação, ambas as bases de dados foram subdivididas em três conjuntos distintos: um conjunto de treino, correspondendo a 70% dos dados; um conjunto de validação, representando 10%; e um conjunto de teste, relativo aos 20% restantes. Esta divisão considerou a proporção das categorias em cada conjunto.

4.4. Experiências de Classificação de Texto

Definiram-se cinco experiências de classificação automática (ver Figura 4.4). A primeira correspondeu à definição de um modelo de referência (*baseline*), para o qual se selecionou o modelo *decoder* EuroLLM. As restantes experiências foram realizadas com os modelos *encoder* mBERT, XLM-R, BERTimbau e Albertina. Todos os modelos estão disponíveis no repositório Hugging Face,[11] e a sua implementação e avaliação foram realizadas na plataforma Google Colaboratory (Colab).[12]



Fonte: Elaboração própria.

Figura 4.4: Experiências de Classificação de Texto

[11] Consultar: https://huggingface.co/
[12] Consultar: https://colab.research.google.com/

A inclusão de um modelo *decoder* como modelo de referência teve como objetivo estabelecer um ponto de comparação na tarefa de classificação face às arquiteturas *encoder*. Optou-se pelo EuroLLM por se tratar de um modelo multilingue de grande dimensão que inclui o português europeu na fase de treino. A sua arquitetura, composta por 9,15 mil milhões de parâmetros, e o comprimento máximo de sequência de 4 096 *tokens* permitem ao modelo considerar dependências de longo alcance no texto, característica relevante para o processamento de textos longos. A classificação com este modelo foi realizada através de uma instrução sem exemplos ou contexto prévio (*zero-shot classification*), conforme apresentado na Figura 4.5.

```
Analise o texto abaixo e determine a sua categoria temática predominante.
A sua resposta deve ser EXCLUSIVAMENTE o nome exato de UMA das
seguintes categorias, sem qualquer texto adicional, explicação ou pontuação.
    Categorias temáticas válidas: {categorias}
    Texto para análise:''{texto}''
    Categoria predominante: {categoria}
```

Figura 4.5: Instrução para a Classificação através do Modelo EuroLLM

Quanto à seleção dos modelos *encoder*, teve por objetivo avaliar o desempenho entre modelos multilingues, como o mBERT e o XLM-R, e modelos especializados na língua portuguesa, como o BERTimbau e o Albertina.

A opção por realizar diferentes experiências com os modelos *encoder* justificou-se pela limitação do comprimento máximo de sequência de entrada, fixado em 512 *tokens*. Os textos que ultrapassavam esse limite eram truncados, conduzindo à perda de informação potencialmente relevante para a classificação (ver Anexo C).[13] Como resposta a esta limitação, definiu-se uma estratégia destinada a modelar a totalidade dos dados extraídos.

A segunda experiência consistiu na utilização dos 512 *tokens* iniciais de cada documento para os conjuntos de treino, validação e teste. Nas duas experiências seguintes, aplicou-se, nos conjuntos de

[13] Na base de dados para classificação multiclasse, a abordagem de extração por etiquetas com pré-processamento demonstrou os valores mais baixos de documentos truncados, situando-se aproximadamente em 69% para a maioria dos modelos, sendo a exceção o Albertina, que registou 84,4%. Em contraste, o método de extração integral sem pré-processamento registou as maiores taxas de documentos truncados, com valores superiores a 86%. Nas experiências com os modelos *encoder*, o Albertina apresentou a maior percentagem de documentos que ultrapassaram o limite de 512 *tokens*. A mesma tendência verificou-se na base de dados utilizada para a classificação binária, com valores superiores aos observados na base de dados para a classificação multiclasse.
O modelo EuroLLM, por permitir uma sequência de entrada mais extensa (até 4 096 *tokens*), registou taxas de truncamento inferiores, com valores que variaram entre 2% e 16% (ver Anexo C).

treino e validação, a divisão dos documentos em segmentos de 500 *tokens*, com uma sobreposição de 100 *tokens* entre segmentos, de forma a reduzir a perda de contexto.

A diferença entre a terceira e quarta experiência residiu na configuração do conjunto de teste. Como se observa na Figura 4.4, na terceira experiência os modelos previram com base nos 512 *tokens* iniciais de cada documento. Na quarta, a previsão realizou-se sobre os segmentos do conjunto de teste, obtidos através da estratégia de divisão dos documentos anteriormente descrita.

A última experiência dos modelos *encoder* foi construída a partir dos resultados da quarta experiência, introduzindo um método de agregação baseado nas probabilidades atribuídas pelos modelos. Este processo de agregação considerou os *logits* produzidos pelo modelo para cada parte do documento. Estes *logits* foram transformados em distribuições de probabilidade sobre as categorias através da função *softmax*. De seguida, para cada documento, os vetores de probabilidade correspondentes às suas partes foram agrupados e combinados através da média. O vetor resultante representou a probabilidade média do documento em cada categoria. Finalmente, aplicou-se a operação *argmax* a este vetor médio, selecionando como previsão final do documento a categoria associada à maior probabilidade.

Conforme referido anteriormente, todas as experiências foram aplicadas em primeiro lugar ao conjunto de dados definido para a classificação multiclasse e, posteriormente, à base de dados de classificação binária. Em cada um destes conjuntos de dados — multiclasse e binário — avaliaram-se as abordagens de extração de texto, realizadas quer por *etiquetas* quer pela extração textual integral, com e sem aplicação de pré-processamento. No total, testaram-se 136 modelos (68 para cada base de dados), dos quais oito correspondiam aos modelos *baseline* (quatro para cada base de dados), enquanto os restantes 128 resultaram das experiências com os modelos *encoder* (64 para cada base de dados).

No que respeita à configuração dos parâmetros de treino dos modelos *encoder*, considerou-se o otimizador AdamW, uma taxa de aprendizagem de 2×10^{-5} e cinco épocas de treino, com aplicação do método *Early Stopping* para prevenir o sobreajustamento. Relativamente ao tamanho do lote, definiu-se um valor de 16 para os modelos mBERT e BERTimbau, enquanto para os modelos Albertina e XLM-R se optou por um tamanho do lote de oito, devido à escassez de recursos computacionais e à maior exigência de memória GPU associada a estes modelos.

No que se refere ao modelo *decoder*, atendendo a que possui 9,15 mil milhões de parâmetros e que a GPU disponibilizada pela plataforma Google Colab na sua versão gratuita apresenta VRAM limitada, recorreu-se à quantização a 4 bits para possibilitar o carregamento e a execução do modelo.

CAPÍTULO 5

Resultados e Discussão

Neste capítulo apresentam-se os resultados obtidos nas cinco experiências definidas no capítulo anterior. Em primeiro lugar, realiza-se uma análise descritiva das métricas de desempenho obtidas tanto na classificação multiclasse como na classificação binária. Posteriormente, discutem-se os resultados alcançados, relacionando-os com os objetivos definidos para a presente dissertação.

Para a avaliação do desempenho dos modelos de classificação, recorreu-se às seguintes métricas:

- Acurácia (Eq. 4.1): Mede a proporção de previsões corretas sobre o total de previsões realizadas. É uma medida adequada quando as classes estão equilibradas, podendo ser enganadora em contextos de desequilíbrio de classes (J. Han et al., 2012, p. 364).

$$Acurácia = \frac{Verdadeiros\ positivos + Verdadeiros\ negativos}{Total\ de\ amostras} \quad (4.1)$$

- Precisão (Eq. 4.2): Indica, de todas as previsões feitas para uma classe, quantas estavam efetivamente corretas.

$$Precisão = \frac{Verdadeiros\ positivos}{Verdadeiros\ positivos + Falsos\ positivos} \quad (4.2)$$

- Sensibilidade (Eq. 4.3): Mede a capacidade do modelo para identificar corretamente todas as instâncias positivas.

$$Sensibilidade = \frac{Verdadeiros\ positivos}{Verdadeiros\ positivos + Falsos\ negativos} \quad (4.3)$$

- F1-*score* (Eq. 4.4): Corresponde à média harmónica entre precisão e sensibilidade.

$$F1 = 2 \times \frac{Precisão \times Sensibilidade}{Precisão + Sensibilidade} \quad (4.4)$$

Segundo J. Han et al. (2012, p. 370), as medidas de precisão, sensibilidade e F1 proporcionam uma avaliação mais robusta em cenários de desequilíbrio entre classes. Nesse sentido, considerando o desequilíbrio entre as categorias na classificação multiclasse, optou-se por apresentar as métricas de precisão, sensibilidade e F1 ponderadas. Este método consiste em calcular a média das métricas obtidas para cada classe, ponderando-as pela respetiva frequência relativa, de modo que o resultado reflita a distribuição real dos dados. A leitura dos resultados privilegia a métrica F1, que sintetiza a precisão e a sensibilidade num único indicador.

Para a classificação binária, embora as categorias estivessem equilibradas e a acurácia pudesse constituir uma métrica adequada, optou-se por manter o F1 como principal indicador de desempenho, de modo a assegurar a consistência na análise comparativa, uma vez que não se verificaram grandes diferenças em relação à acurácia. Importa ainda referir que, na classificação binária, a classe positiva corresponde à categoria "Informação geral" e que as métricas de precisão, sensibilidade e F1 apresentadas correspondem a esta classe.

Apresentam-se também as matrizes de confusão correspondentes aos melhores modelos de ambos os processos de classificação.

5.1. Classificação Multiclasse

A análise comparativa dos resultados indica um melhor desempenho dos modelos *encoder* em relação ao modelo *decoder* utilizado como referência. O melhor resultado obtido pelo EuroLLM correspondeu a 60,7% de F1, enquanto os modelos mBERT, XLM-R, BERTimbau e Albertina alcançaram valores acima de 81%, conforme evidenciado na Tabela 5.1, onde se apresentam os três melhores resultados de cada experiência com modelos *encoder.*

Tabela 5.1: Melhores Resultados por Experiências na Classificação Multiclasse

Experiência	Modelo	Extração	Pré-processamento	Acurácia	F1	Precisão	Sensibilidade
1	EuroLLM	Etiquetas	Sem pré-processamento	55,7%	56,5%	65,1%	55,7%
	EuroLLM	Etiquetas	Com pré-processamento	60,0%	60,7%	72,1%	60,0%
	EuroLLM	Integral	Sem pré-processamento	56,7%	55,3%	74,5%	56,7%
	EuroLLM	Integral	Com pré-processamento	54,8%	54,6%	65,1%	54,8%
2	Albertina	Integral	Sem pré-processamento	83,3%	82,3%	82,3%	83,3%
	Albertina	Etiquetas	Com pré-processamento	79,8%	78,6%	79,4%	79,8%
	Albertina	Etiquetas	Sem pré-processamento	78,9%	77,5%	79,8%	78,9%
3	BERTimbau	Integral	Com pré-processamento	81,6%	81,2%	82,3%	81,6%
	Albertina	Integral	Sem pré-processamento	81,6%	80,7%	82,3%	81,6%
	BERT	Etiquetas	Com pré-processamento	79,8%	79,8%	81,1%	79,8%
4	mBERT	Integral	Sem pré-processamento	72,4%	72,0%	71,7%	72,4%
	BERTimbau	Integral	Sem pré-processamento	71,6%	71,5%	77,8%	71,6%
	BERTimbau	Etiquetas	Com pré-processamento	69,6%	70,2%	74,0%	69,6%
5	mBERT	Integral	Sem pré-processamento	85,1%	84,2%	83,6%	85,1%
	XLM-RoBERTa	Etiquetas	Sem pré-processamento	85,1%	84,2%	84,4%	85,1%
	Albertina	Integral	Sem pré-processamento	84,2%	83,1%	84,7%	84,2%

Nota: Os modelos estão ordenados pelo valor de F1 de forma decrescente, exceto o *baseline* (experiência 1), para o qual se apresentam todas as abordagens.
Apresentam-se os valores ponderados de F1, precisão e sensibilidade.
Fonte: Elaboração própria.

Na segunda experiência, que considerou apenas os 512 *tokens* iniciais de cada documento, o modelo Albertina destacou-se ao registar os três melhores resultados. O melhor desempenho foi

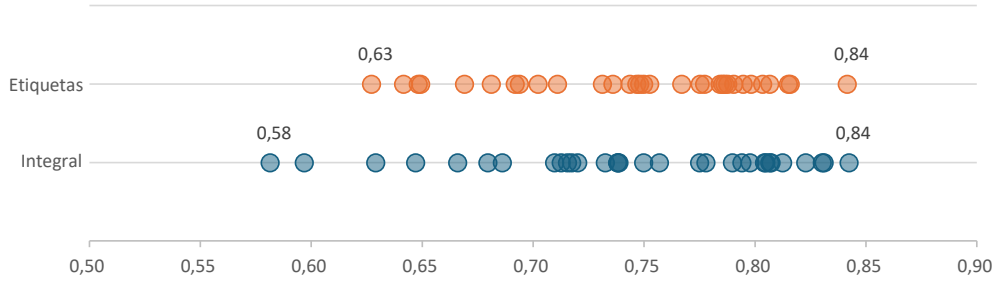
obtido com o conjunto de dados extraídos de forma integral e sem pré-processamento, atingindo um F1 de 82,3%.

Relativamente à terceira experiência, que envolveu o treino dos modelos com a totalidade do texto, mas com a previsão a incidir apenas sobre os 512 *tokens* iniciais, o modelo BERTimbau alcançou o desempenho mais elevado, com um F1 de 81,2%. Foi seguido pelo modelo Albertina, com 80,7%, e pelo modelo mBERT, com 79,8%.

A quarta experiência foi a que registou os valores mais baixos nas métricas de avaliação para os modelos *encoder*. Esta quebra de desempenho pode ser atribuída à metodologia de classificação. Como o processo de classificação foi construído para obter a categoria temática dominante de uma publicação periódica, a divisão em partes pode levar a que segmentos de um mesmo documento sejam classificados em categorias diferentes, comprometendo o seu desempenho. Não obstante, os valores obtidos foram superiores aos da *baseline*. O modelo mBERT destacou-se com um F1 de 72,0%, e o modelo BERTimbau obteve o segundo e terceiro melhores desempenhos desta experiência, com valores de 71,5% e 70,2%, respetivamente.

A quinta experiência resultou diretamente da experiência anterior, agregando as previsões das diferentes partes através da média das probabilidades, o que permitiu obter a categoria dominante por documento. Esta abordagem registou os melhores desempenhos globais na classificação multiclasse com os modelos mBERT e XLM-R a atingirem ambos um F1 de 84,2%. Apesar de apresentarem o mesmo valor de F1, o XLM-R evidenciou ligeira vantagem em termos de precisão, superando o mBERT em 0,8 pontos percentuais (p.p.) nesta métrica. Nesta experiência, o modelo Albertina alcançou também o seu melhor desempenho, com um F1 de 83,1%.

No que respeita à abordagem de extração e à influência do pré-processamento no desempenho, verificou-se que o modelo base (EuroLLM) atingiu o seu melhor resultado com dados pré-processados e extraídos através de etiquetas. Já nas experiências realizadas com os modelos *encoder*, as diferenças entre as duas abordagens de extração revelaram-se menos evidentes, conforme ilustrado na Figura 5.1.

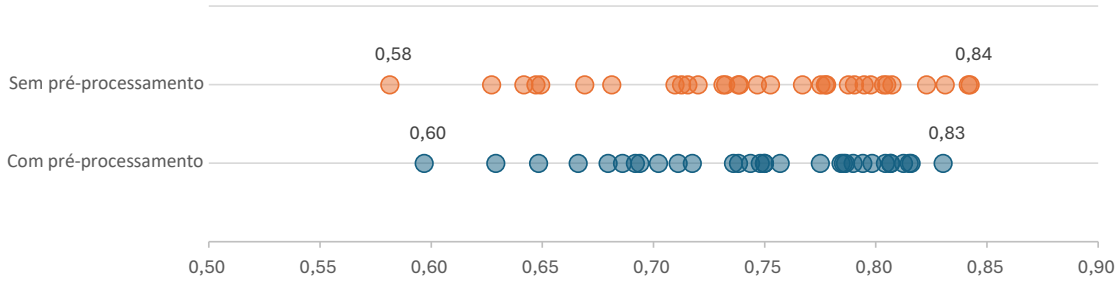


Fonte: Elaboração própria.

Figura 5.1: Desempenho (F1-*score*) dos Modelos *Encoder* na Classificação Multiclasse, por Abordagem de Extração

Ainda assim, a análise dos modelos *encoder* com valores de F1 superiores a 80% mostrou que, dos 16 modelos identificados, 10 utilizaram dados extraídos de forma integral, o que indica uma ligeira vantagem desta abordagem.

Relativamente ao impacto do pré-processamento, observa-se na Figura 5.2 que os modelos treinados com dados sem pré-processamento alcançaram, em geral, melhores métricas de desempenho do que aqueles em que foi aplicada essa fase adicional. Importa salientar que os três modelos *encoder* com melhor desempenho na classificação multiclasse foram obtidos a partir de dados não pré-processados, o que sugere que esta etapa pode eliminar informação relevante para o processo de classificação.



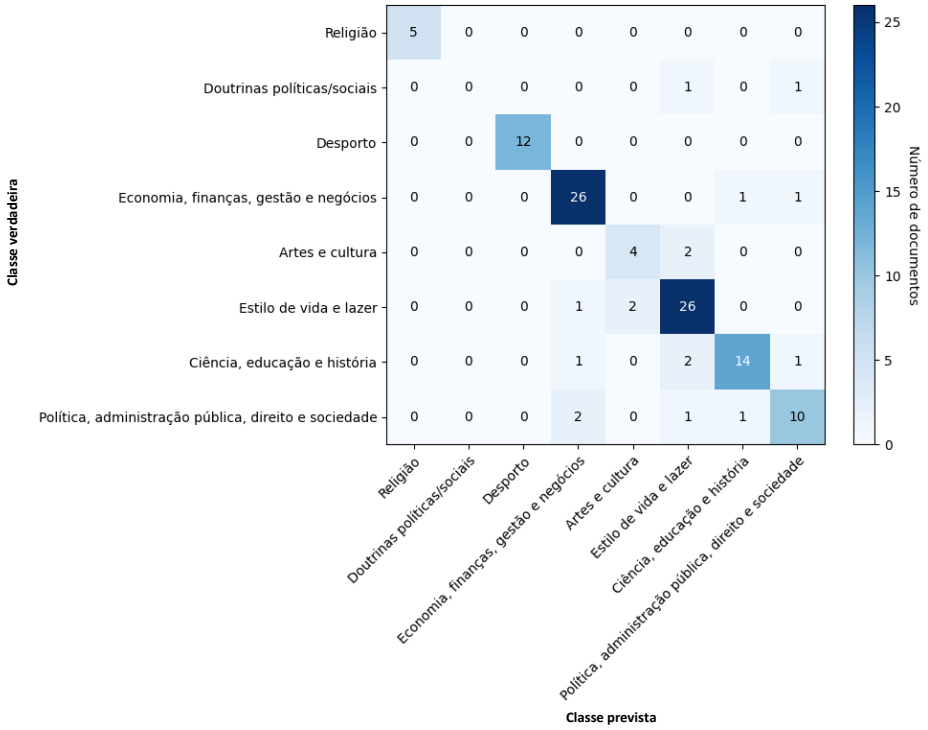
Fonte: Elaboração própria.

Figura 5.2: Desempenho (F1-*score*) dos Modelos *Encoder* na Classificação Multiclasse, por Pré-processamento dos Dados

Para a avaliação do desempenho por categoria, a análise incidiu sobre os modelos mBERT e XLM-R, que alcançaram os melhores resultados na classificação multiclasse. Conforme referido, estes resultados derivam da quinta experiência, em que o modelo mBERT teve o seu melhor desempenho com dados extraídos integralmente, e o XLM-R com dados extraídos por etiquetas. Em ambos os casos, não se aplicou pré-processamento. As matrizes de confusão que detalham o desempenho de cada modelo encontram-se nas Figuras 5.3 (mBERT) e 5.4 (XLM-R). No Anexo D encontram-se as métricas detalhadas por categoria.

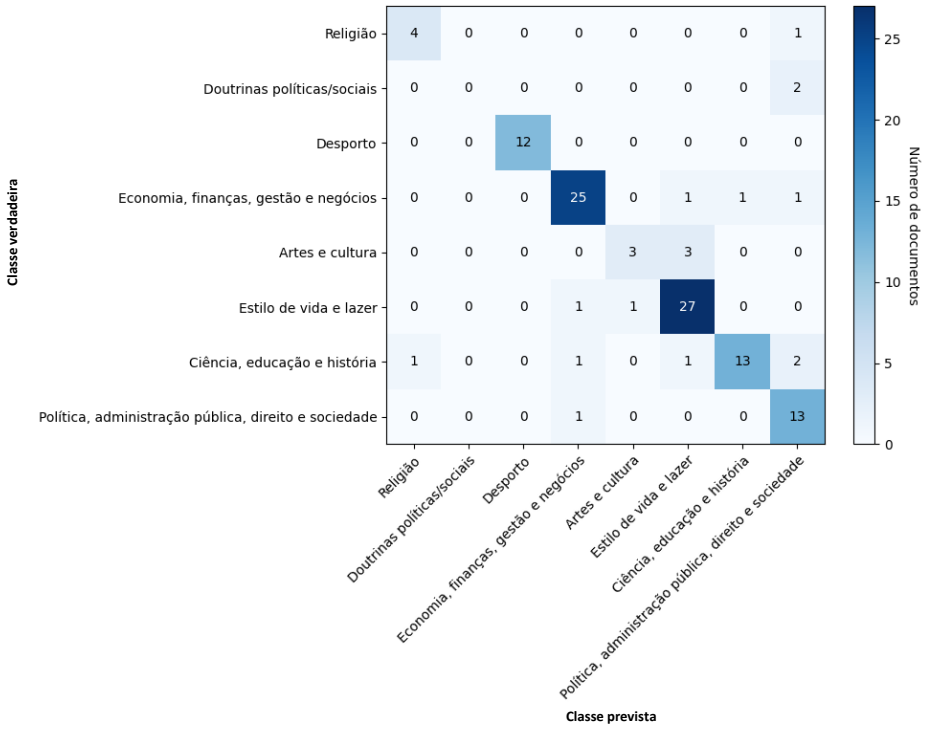
A categoria "Desporto" foi classificada com 100% de acerto por ambos os modelos. Do mesmo modo, a "Religião" também foi classificada corretamente todos os documentos no modelo mBERT; enquanto no modelo XLM-R quatro dos cinco documentos foram corretamente classificados.

As categorias "Economia, finanças, gestão e negócios", "Estilo de vida e lazer" e "Ciência, educação e história" apresentaram bons resultados, com valores de F1 superiores a 80% e um número reduzido de erros. Estas três categorias correspondem também às que possuem maior número de documentos na base de dados, o que poderá ter contribuído para os resultados.



Nota: Modelo obtido na quinta experiência com dados extraídos de forma integral e sem pré-processamento.
Fonte: Elaboração própria.

Figura 5.3: Matriz de Confusão do Modelo mBERT



Nota: Modelo obtido na quinta experiência com dados extraídos por etiquetas e sem pré-processamento.
Fonte: Elaboração própria.

Figura 5.4: Matriz de confusão do modelo XLM-RoBERTa

A categoria "Política, administração pública, direito e sociedade" registou um desempenho moderado, com valores de F1 de 74,1% no modelo mBERT e 78,8% no modelo XLM-R. No entanto, verificaram-se diferenças no tipo de erro: enquanto no modelo mBERT as classificações erradas se dispersaram por várias categorias, no modelo XLM-R concentraram-se apenas num caso, classificado como "Economia, finanças, gestão e negócios", algo que poderá ter sido influenciado por termos como *sustentabilidade*, *sucesso empresarial* ou *economia*.

Por seu lado, a categoria "Artes e cultura" revelou um desempenho baixo em ambos os modelos. Dos seis documentos, apenas quatro foram corretamente classificados através do modelo mBERT (F1 = 66,7%) e três através do modelo XLM-R (F1 = 60,0%). Os erros devem-se, em grande parte, à proximidade lexical com "Estilo de vida e lazer", evidenciada por termos comuns como *festival*, *galeria*, *música* ou *cinema*.

A categoria "Doutrinas políticas/sociais" surge como a mais problemática. Nenhum documento foi corretamente classificado por ambos os modelos. Este resultado relaciona-se com o número muito reduzido de exemplos disponíveis na base de dados (nove documentos) e com o conteúdo ambíguo, frequentemente confundido com a categoria "Política, administração pública, direito e sociedade". Exemplos dessa ambiguidade incluem títulos como "*Bloco de Esquerda questiona o Governo sobre a qualidade do ar em Mirandela*", "*A força política da solidariedade contra as extremas-direitas*" e "*3.ª Marcha da Covilhã pelos Direitos LGBTQIAP+*".

5.2. Classificação Binária

As métricas de desempenho da classificação binária revelaram valores superiores aos observados na classificação multiclasse. O modelo *baseline* EuroLLM alcançou um valor máximo de F1 = 79,3% na classe positiva ("Informação geral") enquanto os modelos *encoder* obtiveram resultados mais elevados, todos acima de 90%, como se observa na Tabela 5.2. O melhor desempenho registado correspondeu a um F1 de 96,0%.

As experiências de maior sucesso foram a agregação de previsões (quinta experiência) e a classificação utilizando o texto truncado (segunda experiência). Na quinta experiência, o modelo BERTimbau, com extração integral do texto e sem pré-processamento, alcançou o valor de F1 máximo de 96,0%. De forma similar, na segunda experiência, o modelo BERTimbau, aplicando a abordagem de extração integral do texto e sem pré-processamento, também apresentou bons resultados, com um F1 de 94,6%.

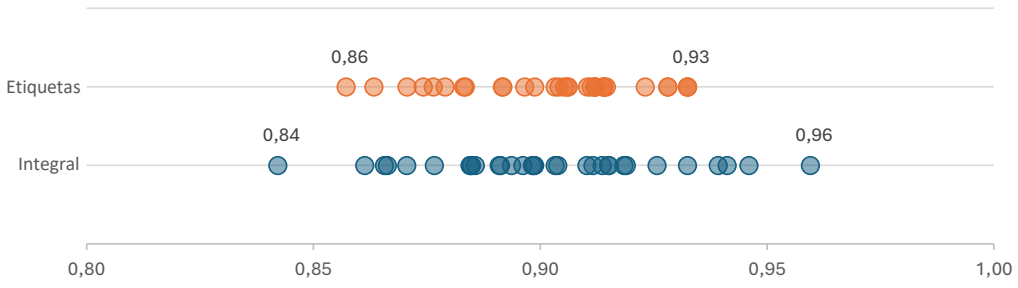
Tabela 5.2: Melhores Resultados por Experiências na Classificação Binária

Experiência	Modelo	Extração	Pre-processamento	Acurácia	F1	Precisão	Sensibilidade
1	EuroLLM	Etiquetas	Sem pré-processamento	73,2%	73,2%	73,4%	73,2%
	EuroLLM	Etiquetas	Com pré-processamento	79,3%	79,3%	79,7%	79,3%
	EuroLLM	Integral	Sem pré-processamento	68,5%	68,4%	68,6%	68,5%
	EuroLLM	Integral	Com pré-processamento	76,1%	76,0%	76,2%	76,1%
2	BERTimbau	Integral	Sem pré-processamento	94,6%	94,6%	94,6%	94,6%
	BERTimbau	Etiquetas	Sem pré-processamento	92,6%	92,8%	89,9%	95,9%
	Albertina	Etiquetas	Sem pré-processamento	92,6%	92,3%	95,7%	89,2%
3	mBERT	Integral	Sem pré-processamento	93,9%	94,1%	91,1%	97,3%
	BERTimbau	Integral	Com pré-processamento	93,2%	93,2%	93,2%	93,2%
	BERTimbau	Etiquetas	Sem pré-processamento	93,2%	93,2%	93,2%	93,2%
4	Albertina	Etiquetas	Sem pré-processamento	89,9%	91,5%	90,0%	92,9%
	mBERT	Etiquetas	Sem pré-processamento	89,8%	91,4%	90,8%	92,0%
	BERTimbau	Etiquetas	Sem pré-processamento	90,0%	91,1%	92,5%	89,8%
5	BERTimbau	Integral	Sem pré-processamento	96,0%	96,0%	96,0%	96,0%
	mBERT	Integral	Sem pré-processamento	93,9%	93,9%	94,0%	93,9%
	BERTimbau	Etiquetas	Com pré-processamento	93,2%	93,2%	93,4%	93,2%

Nota: Os modelos estão ordenados pelo valor de F1 de forma decrescente, exceto o *baseline* (experiência 1), para o qual se apresentam todas as abordagens.
As métricas F1, Precisão e Sensibilidade referem-se ao desempenho na classe positiva: "Informação geral".
Fonte: Elaboração própria.

Na terceira experiência, os modelos mBERT e BERTimbau destacaram-se, alcançando F1 de 94,1% e 93,2%, respetivamente. A quarta experiência voltou a apresentar os desempenhos mais baixos entre os modelos *encoder*, em linha com o observado na classificação multiclasse.

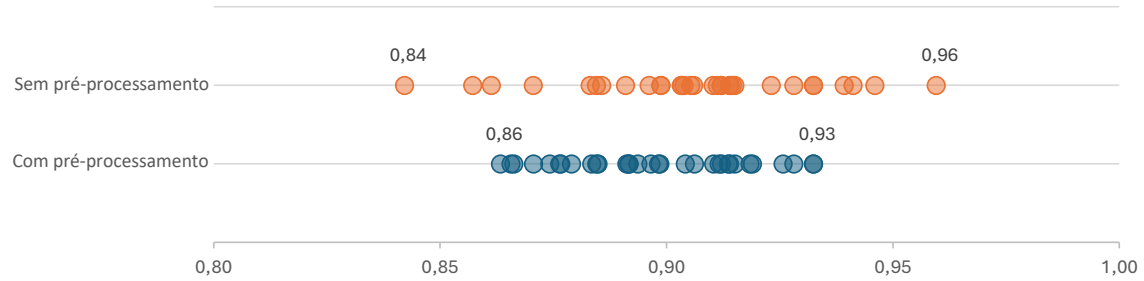
A análise da abordagem de extração mostra que o modelo EuroLLM alcançou os melhores resultados com dados extraídos por etiquetas, enquanto os modelos *encoder* evidenciaram melhor desempenho quando utilizaram dados obtidos de forma integral, conforme ilustrado na Figura 5.5.



Fonte: Elaboração própria.

Figura 5.5: Desempenho (F1-*score*) dos Modelos *Encoder* na Classificação Binária, por Abordagem de Extração

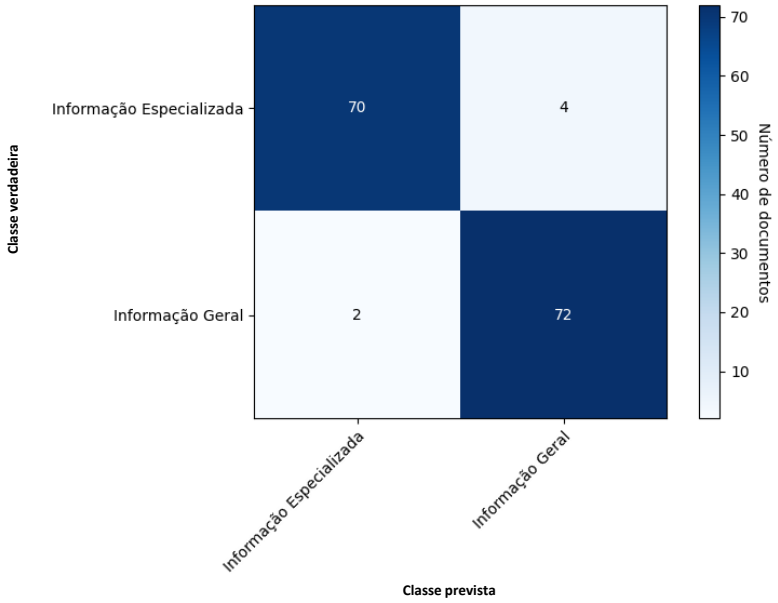
Relativamente à influência do pré-processamento, o modelo base registou uma melhoria no desempenho quando os dados passaram por essa fase. Já nos modelos *encoder*, como se observa na Figura 5.6, os melhores resultados foram alcançados com dados sem pré-processamento.



Fonte: Elaboração própria.

Figura 5.6: Desempenho (F1-*score*) dos Modelos *Encoder* na Classificação Binária, por Pré-processamento dos Dados

No que diz respeito à análise das categorias do modelo com o melhor resultado na classificação binária, o BERTimbau com extração textual integral e sem pré-processamento, a matriz de confusão obtida por este modelo evidencia o equilíbrio da tarefa (ver Figura 5.7).



Nota: Modelo obtido na quinta experiência com dados extraídos de forma integral e sem pré-processamento.
A categoria "Informação especializada" integra as publicações periódicas de conteúdo doutrinário.
Fonte: Elaboração própria.

Figura 5.7: Matriz de Confusão do Modelo BERTimbau

Para a categoria “Informação especializada”, o modelo demonstrou uma precisão de 97,2% (ver Anexo D). Este valor, derivado da correta classificação de 70 das 72 previsões feitas para esta classe. A sensibilidade foi de 94,6%, o que significa que o modelo identificou corretamente 70 dos 74 documentos que pertenciam a esta classe, com quatro casos a serem incorretamente classificados como "Informação geral".

No que respeita à classe “Informação geral”, observou-se um comportamento complementar. A sensibilidade atingiu 97,3%, o que revela uma capacidade notável para identificar quase todos os documentos desta categoria, com apenas dois dos 74 casos a serem classificados de forma errada. A precisão, por sua vez, foi de 94,7%, um valor influenciado pelos quatro documentos de “Informação especializada” que foram incorretamente atribuídos a esta classe.

5.3. Discussão dos Resultados

Os resultados obtidos evidenciam a capacidade dos modelos *encoder* mBERT, XLM-R, BERTimbau e Albertina para classificar de forma eficaz o tema dominante das publicações periódicas portuguesas. As métricas de desempenho observadas indicam um grau de eficácia elevado, comparável ao registado em outros estudos sobre classificação de texto em português aplicados a diferentes domínios (Costa et al., 2024; Fialho et al., 2024; Ramos et al., 2024; Yanchuk et al., 2022; Yang et al., 2022).

Na classificação binária, observou-se um desempenho particularmente elevado, com métricas acima dos 90%, resultado favorecido pela dimensão equilibrada das classes e pela quantidade de exemplos disponíveis. Quanto à classificação multiclasse, embora as métricas se situem abaixo das registadas na classificação binária, os valores em torno dos 80% de F1 confirmam a boa capacidade preditiva dos modelos *encoder*.

A quebra de eficácia na classificação multiclasse explica-se, em parte, pelo desequilíbrio entre categorias, nomeadamente pelo número reduzido de amostras nas classes “Doutrinas políticas/sociais” e “Artes e cultura”, bem como pela proximidade lexical entre categorias, em especial entre “Artes e cultura” e “Estilo de vida e lazer”, e entre “Doutrinas políticas/sociais” e “Política, administração pública, direito e sociedade”.

Importa salientar que a sobreposição semântica é expectável, tendo em conta a abordagem de classificação definida, que atribui a cada documento uma única categoria dominante. Esta metodologia pode colidir com a natureza dos conteúdos mediáticos, nos quais é comum a interligação de diferentes áreas temáticas numa mesma publicação.

Por outro lado, a análise comparativa entre o modelo *decoder* e os modelos *encoder* corrobora as observações de Han et al. (2021) e Minaee et al. (2021), que apontam os modelos *encoder* como mais eficazes na tarefa de classificação de texto. Este desempenho pode ser explicado pela sua arquitetura

bidirecional, que lhes confere a capacidade de capturar relações contextuais complexas entre as palavras. No presente estudo, os melhores resultados do modelo *decoder* EuroLLM corresponderam a 60,7% de F1 na classificação multiclasse e 79,3% na binária, enquanto os melhores modelos *encoder* alcançaram 84,2% e 96,0%, respetivamente. Assim, a aplicação de modelos *encoder* proporcionou ganhos de 23,5 p.p. na classificação multiclasse e de 16,7 p.p. na classificação binária, demonstrando a vantagem da aprendizagem bidirecional no tratamento de dados textuais.

Relativamente à abordagem de extração e ao impacto do pré-processamento, os resultados evidenciaram diferenças entre o modelo *decoder* e os modelos *encoder*. O EuroLLM apresentou melhores resultados com dados pré-processados e extraídos a partir de etiquetas, o que poderá indicar que o texto relevante e a sua limpeza favoreceram o processo de previsão sequencial do modelo. Em contraste, os modelos *encoder* obtiveram resultados superiores com a extração integral do texto e sem pré-processamento, o que sugere que a preservação da estrutura e do contexto original do texto contribuiu para a eficácia da classificação. É possível que as etapas de limpeza, nos modelos *encoder*, eliminem informação contextual útil à reconstrução de palavras mascaradas e, consequentemente, à previsão da categoria temática.

A capacidade dos modelos *encoder* obterem resultados de elevada qualidade a partir de dados em bruto constitui uma vantagem operacional de grande valor, ao simplificar o processo de trabalho e reduzir o tempo necessário para a preparação dos dados. Este aspeto corrobora as conclusões de Garcia et al. (2024) que salientam a simplicidade operacional da arquitetura BERT (no seu caso, o modelo BERTimbau), decorrente da possibilidade de dispensar etapas de pré-processamento.

Quanto à estratégia de modelação de textos longos, aplicada na quinta experiência, revelou-se eficaz para contornar a limitação do comprimento máximo de sequência de 512 *tokens* dos modelos *encoder*. O método de agregação das previsões das diferentes partes de um documento, com base na média das probabilidades, produziu os melhores resultados globais em ambos os processos de classificação. Esta abordagem superou o desempenho da classificação baseada apenas nos 512 *tokens* iniciais em cerca de 2 p.p.

A análise comparativa entre os modelos multilingues e os modelos específicos para a língua portuguesa evidenciou um desempenho variável consoante a tarefa de classificação. Na classificação multiclasse, os modelos multilingues mBERT e XLM-R registaram o desempenho mais elevado. Contudo, o modelo Albertina, desenvolvido para o português europeu, revelou-se competitivo, com uma diferença de apenas 1,1 p.p. face ao melhor resultado, e obteve o melhor desempenho na abordagem com os 512 *tokens* iniciais. Por outro lado, na classificação binária, o modelo BERTimbau, pré-treinado em português do Brasil, destacou-se com o melhor desempenho geral.

CAPÍTULO 6

Conclusões

O presente estudo abordou o desafio da classificação de texto em publicações periódicas portuguesas com presença digital, contribuindo para o avanço da aplicação de técnicas de *text mining* em português europeu. O objetivo central consistiu em identificar a categoria temática dominante de cada publicação com base no conteúdo textual extraído dos respetivos sítios eletrónicos, mediante a aplicação de modelos de linguagem pré-treinados baseados na arquitetura *transformer*.

Para concretizar este objetivo, o trabalho foi estruturado em torno de objetivos específicos, cujas principais conclusões se sintetizam de seguida:

- Em primeiro lugar, foi desenhado um novo sistema de categorias temáticas, em colaboração com a Entidade Reguladora para a Comunicação Social, que serviu de base para a tarefa de classificação supervisionada. Este contributo assume relevância prática e institucional, uma vez que fornece uma estrutura coerente e aplicável à caracterização do setor da imprensa portuguesa.
- Posteriormente, procedeu-se à extração dos dados textuais através de técnicas de *web scraping*, o que permitiu a criação de duas bases de dados, uma para a classificação multiclasse e outra para a classificação binária. Para este efeito, foram testadas duas abordagens de extração: uma integral e outra seletiva, baseada nas etiquetas HTML, as quais se mostraram ambas viáveis para a obtenção dos dados que serviram de base à modelação.
- A avaliação do desempenho de diferentes modelos *transformers* demonstrou a superioridade das arquiteturas *encoder* testadas, como o modelo mBERT, XLM-R, BERTimbau e Albertina, sobre o modelo *decoder* EuroLLM, utilizado como referência. Verificou-se ainda que os modelos *encoder* atingiram, em geral, melhores resultados com dados não processados, o que representa uma vantagem operacional ao simplificar a fase de preparação dos dados.
- Do mesmo modo, a estratégia de modelação de textos longos, que consistiu na divisão dos documentos em partes e na agregação das suas probabilidades, revelou-se a mais eficaz para contornar a limitação de 512 *tokens* dos modelos *encoder*, produzindo os melhores resultados globais.
- A análise comparativa entre modelos multilingues e específicos para a língua portuguesa demonstrou que ambos resultam adequados na classificação de texto em português. Importa salientar que os modelos Albertina e BERTimbau demonstraram ser competitivos, alcançando resultados de destaque em diferentes experiências.

Em conclusão, os resultados obtidos permitem responder afirmativamente à questão de investigação: é possível automatizar de forma eficaz a classificação temática das publicações periódicas portuguesas com base no texto disponível nos seus sítios eletrónicos, recorrendo a modelos *transformers* pré-treinados. O trabalho desenvolvido, ao comparar diferentes modelos, estratégias de extração, tratamentos de dados e metodologias de modelação de textos longos, demonstrou empiricamente que esta automatização é viável e eficaz. Ainda assim, o estudo apresentou algumas limitações que importa reconhecer e que poderão ser consideradas em investigações futuras.

Uma das principais restrições prende-se com os recursos computacionais disponíveis, que condicionaram a seleção dos modelos, impedindo o teste de versões com um maior número de parâmetros (por exemplo, as versões *large*), as quais poderiam, teoricamente, alcançar um desempenho superior. Adicionalmente, o desequilíbrio entre as classes na base de dados para a classificação multiclasse, em particular o número reduzido de exemplos em categorias como "Doutrinas políticas/sociais" e "Artes e cultura", condicionou o desempenho dos modelos nestas categorias.

Nesse sentido, as limitações identificadas abrem caminho para futuras linhas de investigação, entre as quais se destaca a exploração de modelos de linguagem de maior dimensão, capazes de capturar relações contextuais mais complexas. Recomenda-se igualmente a realização de experiências com o ajustamento de hiperparâmetros (taxa de aprendizagem, tamanho do lote, épocas, entre outros), de modo a otimizar o desempenho. Sugere-se, ainda, a expansão da base de dados mediante a inclusão das publicações periódicas em formato impresso, através da sua digitalização.

Outra recomendação passa pelo pré-treino adicional dos modelos com o corpo de textos recolhido dos sítios eletrónicos das publicações periódicas. Esta especialização permitiria adaptar o vocabulário e o conhecimento linguístico dos modelos ao domínio específico da imprensa portuguesa, com potencial para melhorar a precisão da classificação.

Outra linha de investigação pertinente seria a adoção de uma metodologia de classificação múltipla. Tal metodologia permitiria a atribuição de várias categorias temáticas a uma mesma publicação, representando de forma mais exata a sua natureza complexa e a recorrente interligação temática presente em determinados conteúdos mediáticos.

Fontes

Assembleia da República (1999). Lei de Imprensa, Lei n.º 2/99 de 13 de Janeiro (5ª versão), *Diário da República*, Série I-A, (10). https://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=138&tabela=leis

Assembleia da República (2005). Estatutos da ERC, Lei n.º 53/2005 de 8 de Novembro. *Diário da República*, I Série-A, (214). https://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=588&tabela=leis

Referências Bibliográficas

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

Allam, H., Makubvure, L., Gyamfi, B., Graham, K. N., & Akinwolere, K. (2025). Text Classification: How Machine Learning Is Revolutionizing Text Categorization. *Information*, *16*(130). https://doi.org/10.3390/info16020130

Brown, M. A., Gruen, A., Maldoff, G., Messing, S., Sanderson, Z., & Zimmer, M. (2024). *Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations* (No. arXiv:2410.23432v2). arXiv; arXiv. http://arxiv.org/abs/2410.23432

Camões - Instituto da Cooperação e da Língua. (2023, maio 2). Dados sobre a Língua Portuguesa. *Dia Mundial da Língua Portuguesa 2023*. https://www.instituto-camoes.pt/images/img_agenda2023/Dados_sobre_a_l%C3%ADngua_portuguesa_2023.pdf

Chen, X., Cong, P., & Lv, S. (2022). A Long-Text Classification Method of Chinese News Based on BERT and CNN. *IEEE Access*, *10*, 34046–34057. https://doi.org/10.1109/ACCESS.2022.3162614

Cleverdon, C. (1984). Optimizing convenient online access to bibliographic databases. *Information Services & Use*, *4*(1–2), 37–47. https://doi.org/10.3233/ISU-1984-41-204

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale* (No. arXiv:1911.02116). arXiv. https://doi.org/10.48550/arXiv.1911.02116

Costa, V., Coelho, P., & Castelli, M. (2024). Artificial Intelligence for Impact Assessment of Administrative Burdens. *Emerging Science Journal*, *8*(1), 270–282. https://doi.org/10.28991/ESJ-2024-08-01-019

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (No. arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Fialho, P., Ribeiro, R., Batista, F., Ramos, G., Fonseca, A., Moro, S., Guerra, R., Carvalho, P., Marques, C., & Silva, C. (2024). Counter Hate Speech Detection in Youtube Conversations. Em M.-J. Lesot, S. Vieira, M. Z. Reformat, J. P. Carvalho, F. Batista, B. Bouchon-Meunier, & R. R. Yager (Eds.), *Information Processing and Management of Uncertainty in Knowledge-Based Systems* (Vol. 1176, pp. 94–105). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-73997-2_9

Fields, J., Chovanec, K., & Madiraju, P. (2024). A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe? *IEEE Access*, *12*, 6518–6531. https://doi.org/10.1109/ACCESS.2024.3349952

Filho, J. A. W., Wilkens, R., Idiart, M., & Villavicencio, A. (2018). *The brWaC Corpus: A New Open Resource for Brazilian Portuguese*. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. https://aclanthology.org/L18-1686/

Garcia, K., Shiguihara, P., & Berton, L. (2024). Breaking news: Unveiling a new dataset for Portuguese news classification and comparative analysis of approaches. *PLOS ONE*, *19*(1), e0296929. https://doi.org/10.1371/journal.pone.0296929

Han, J., Micheline Kamber, & Pei, J. (2012). *Data mining: Concepts and techniques* (3ª ed). Elsevier/Morgan Kaufmann.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., … Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002

Hugging Face. (2022a). BERT multilingual base model (uncased). *Model Card*. https://huggingface.co/google-bert/bert-base-multilingual-uncased

Hugging Face. (2022b). XLM-RoBERTa (base-sized model). *Model Card*. https://huggingface.co/FacebookAI/xlm-roberta-base

Ignatow, G., & Mihalcea, R. F. (2018). *An introduction to text mining: Research design, data collection, and analysis*. SAGE.

Jurafsky, D., & Martin, J. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. https://web.stanford.edu/~jurafsky/slp3/

Maron, M. E. (1961). Automatic Indexing: An Experimental Inquiry. *Journal of the ACM*, *8*(3), 404–417. https://doi.org/10.1145/321075.321084

Martins, P. H. (2024). Model Card for EuroLLM-9B-Instruct. *Model Card*. https://huggingface.co/utter-project/EuroLLM-9B-Instruct

Martins, P. H., Fernandes, P., Alves, J., Guerreiro, N. M., Rei, R., Alves, D. M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., De Souza, J. G. C., Birch, A., & Martins, A. F. T. (2025). EuroLLM: Multilingual Language Models for Europe. *Procedia Computer Science*, *255*, 53–62. https://doi.org/10.1016/j.procs.2025.02.260

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning-based Text Classification: A Comprehensive Review. *ACM Computing Surveys*, *54*(3). https://doi.org/10.1145/3439726

Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., Guerra, R., Carvalho, P., Marques, C., & Silva, C. (2024). Leveraging Transfer Learning for Hate Speech Detection in Portuguese Social Media Posts. *IEEE Access*, *12*, 101374–101389. https://doi.org/10.1109/ACCESS.2024.3430848

Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., & Osório, T. (2023). *Advancing Neural Encoding of Portuguese with Transformer Albertina PT* (No. arXiv:2305.06721v2). arXiv. https://doi.org/10.1007/978-3-031-49008-8_35

Santos, R., Rodrigues, J., Branco, A., & Vaz, R. (2021). Neural Text Categorization with Transformers for Learning Portuguese as a Second Language. Em G. Marreiros, F. S. Melo, N. Lau, H. Lopes Cardoso, & L. P. Reis (Eds.), *Progress in Artificial Intelligence* (Vol. 12981, pp. 715–726). Springer International Publishing. https://doi.org/10.1007/978-3-030-86230-5_56

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Silva Barbon, R., & Akabane, A. T. (2022). Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. *Sensors*, *22*(21), 8184. https://doi.org/10.3390/s22218184

Souza, F. C., Nogueira, R. F., & Lotufo, R. A. (2023). BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis. *Applied Soft Computing*, *149*, 110901. https://doi.org/10.1016/j.asoc.2023.110901

UC Berkeley School of Information. (2017). *In Memory of Professor Emeritus Melvin "Bill" Maron*. https://www.ischool.berkeley.edu/news/2017/memory-professor-emeritus-melvin-bill-maron?utm_source=chatgpt.com

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (No. arXiv:1706.03762v7). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wu, Y., & Wan, J. (2025). A survey of text classification based on pre-trained language model. *Neurocomputing*, *616*, 128921. https://doi.org/10.1016/j.neucom.2024.128921

Yanchuk, A., Trifan, A., Fajarda, O., & Oliveira, J. L. (2022). Automatic Classification of Stigmatizing Articles of Mental Illness: The Case of Portuguese Online Newspapers. Em S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørvåg, B. Catania, G. Vargas-Solar, & E. Zumpano (Eds.), *New Trends in Database and Information Systems* (pp. 332–343). Springer International Publishing. https://doi.org/10.1007/978-3-031-15743-1_31

Yang, H., Gonçalves, T., Quaresma, P., Vieira, R., Veladas, R., Pinto, C. S., Oliveira, J., Ferreira, M. C., Morais, J., Pereira, A. R., Fernandes, N., & Gonçalves, C. (2022). Clinical Trial Classification of SNS24 Calls with Neural Networks. *Future Internet*, *14*(5), 130. https://doi.org/10.3390/fi14050130

Anexo A

Sistema de Categorias Temáticas

O sistema de categorias temáticas, definido para este projeto, baseou-se na categorização utilizada pelo Instituto Nacional de Estatística (INE) no seu Inquérito às Classificações Periódicas. Este último contempla as 30 categorias seguintes:

1. Generalidades e Reportagem
2. Política, Economia e Finanças
3. Jogos e Desporto
4. Direito, Jurisprudência, Administração Pública e Assistência Social
5. Engenharia e Tecnologia
6. Medicina e Saúde
7. Gestão, Comércio e Comunicação (publicidade, marketing, relações públicas)
8. Religião e Teologia
9. Educação
10. Agricultura, Silvicultura, Caça e Pesca
11. Indústria, Construção e Equipamento
12. História. Biografia
13. Informática
14. Etnologia, Etnografia, Tradições, Folclore e Costumes (moda, vida social)
15. Música e Espetáculos
16. Demografia, Estatística e Sociologia
17. Urbanismo e Arquitetura
18. Filosofia e Psicologia
19. Matemática
20. Ciências do Ambiente e Conservação da Natureza
21. Economia Doméstica (culinária, decoração, body care, familiar)
22. Artes Plásticas, Artes Gráficas, Design e Desenho
23. Línguas, Linguística e Literatura
24. Geografia e Viagens
25. Paleontologia, Biologia, Botânica e Zoologia
26. Jardinagem, Horticultura e Animais
27. Ciências Naturais, Ciências da Terra, Geologia e Meteorologia
28. Pintura e Fotografia

29. Astronomia e Astrofísica
30. Física, Química e Mineralogia

Adicionalmente, ponderou-se o sistema de classificação da APCT, constituído por 24 categorias:

1. Ambiente/Divulgação científica
2. Classificados
3. Crianças
4. Culinária
5. Cultura/Espetáculo
6. Decoração
7. Desporto e veículos
8. Diretório
9. Economia/Negócios e Gestão
10. Femininas/Modas
11. Horóscopo
12. Hotelaria/Restauração
13. Informação geral
14. Juvenis
15. Lazer
16. Lazer, cultura e espetáculo
17. Masculinas
18. Regional
19. Saúde/Educação
20. Sectorial
21. Sociedade
22. Tecnologias da informação
23. Televisão/Jogos
24. Viagens e turismo

A conjugação de ambos os sistemas, com a experiência do corpo técnico da ERC, serviu de base à definição de um conjunto de subcategorias, com o objetivo de otimizar o processo de classificação manual. Este conjunto é apresentado na tabela que se segue:

Conteúdo	Categoria	Subcategorias
Informação geral	Geral	Esta categoria engloba as publicações com predomínio de conteúdo jornalístico ou informativo de âmbito não especializado, que abordam diversas áreas como atualidade, política, cultura, desporto, economia, entretenimento, ciência, entre outras.
Informação especializada	Desporto	
	Economia, finanças, gestão e negócios	Publicidade Marketing Relações públicas Agricultura Pesca (fins comerciais) Mercado bolsista Banca Indústria Construção/Equipamento Desenvolvimento regional Comércio internacional Hotelaria/Restauração Informática/Tecnologia de informação Sustentabilidade/Ambiente/Transição ecológica Associações do setor empresarial Bens e serviços Indústria do turismo Empreendedorismo e *Startups*
	Artes e cultura	Música Arquitetura Desenho Pintura Fotografia Artesanato Tradições, folclore e costumes Património cultural Literatura Artes performativas (teatro, dança) Artes audiovisuais (cinema, TV, videojogos, áudio) Banda desenhada/Ilustração (artes gráficas)
	Estilo de vida e lazer	Vida saudável Vida de figuras públicas Forma física Jardinagem Culinária Vinhos Decoração Videojogos (entretenimento) Áudio/TV/Cinema (entretenimento) Banda desenhada (entretenimento) Turismo/Viagens

		Eventos sociais/culturais ou de lazer Família Carros/Motos *Lifestyle* Moda (vestuário) Adultos (erótico/pornográfico) Infantis
	Ciência, educação e história	Revista/Divulgação científica Revistas académicas Revistas técnicas Matemática Ciências sociais Ciências naturais Ciências humanas Filosofia Línguas/Linguística Biografia
	Política, administração pública, direito e sociedade	Boletins/revistas municipais Jurisprudência Associações/Organizações sociais Apoio social Direitos humanos Urbanismo Justiça e sistema penal Previdência social Proteção civil Defesa
Doutrinária	Religião	Espiritualidade Crenças religiosas
	Doutrinas políticas/sociais	Publicações de partidos políticos (com conteúdo ideológico) Organizações/movimentos sociais (com conteúdo ideológico)

Anexo B

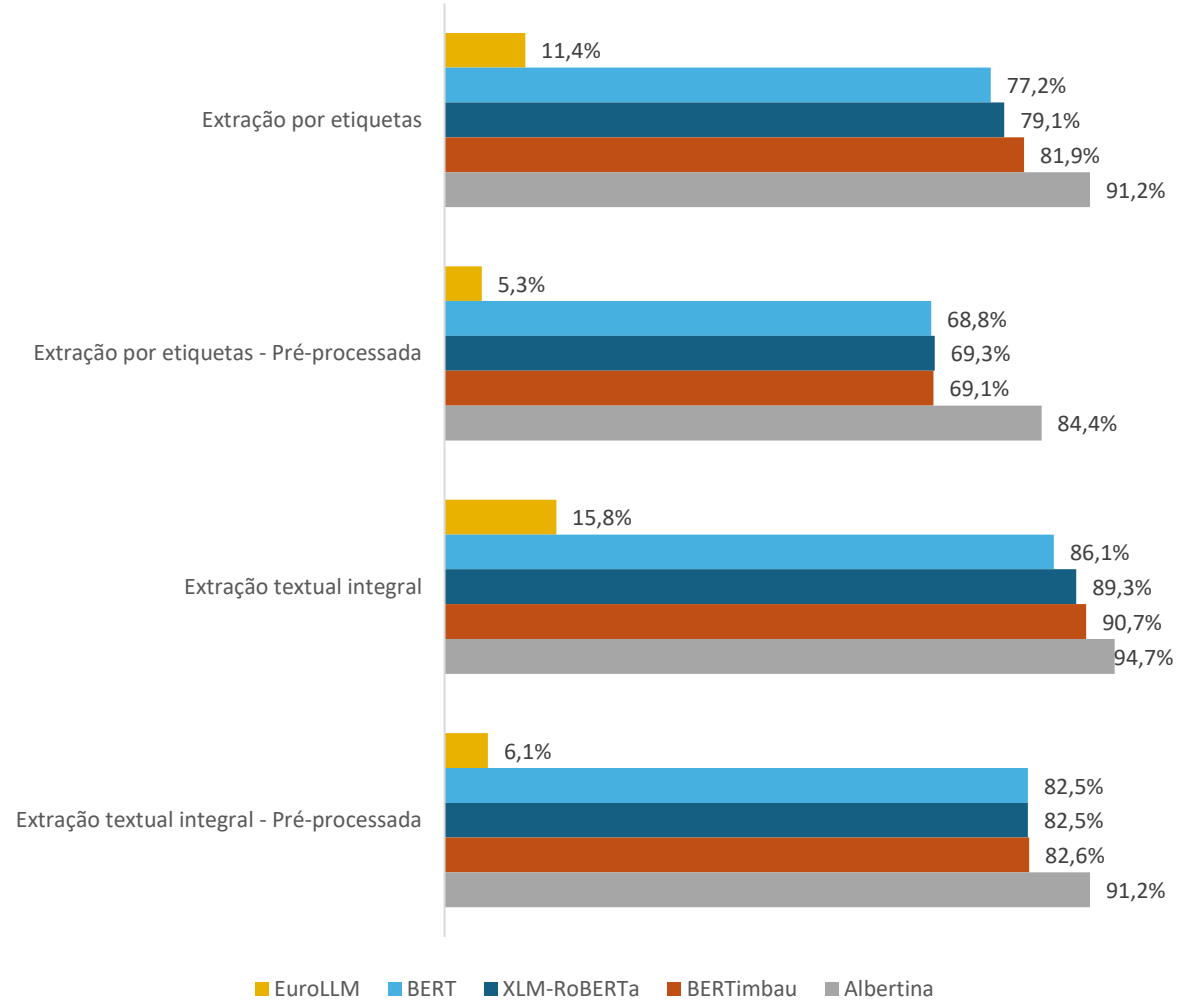
Indicadores das Bases de Dados

Indicador	Definição
Número de registo	Número de identificação utilizado pela ERC no processo de registo.
Denominação	Nome da publicação periódica.
Suporte	Formato da publicação periódica: impressa; digital; impressa/digital.
Conteúdo	Tipo de conteúdo da publicação periódica, segundo o artigo 10.º da Lei de Imprensa (1999).
Temática	Classificação temática na base de dados dos registos da ERC.
Nova classificação	Classificação temática definida para este projeto.
Website	Sítio eletrónico da publicação periódica.
Texto extraído *tags*	Texto extraído nas etiquetas <a> e <p>.
Data extração *tags*	Data da extração.
Arquivo HTML *tags*	Nome do documento HTML da publicação periódica extraída.
Texto extraído *get text*	Texto extraído de forma integral através de *get_text()*.
Data extração *get text*	Data da extração.
Arquivo HTML get text	Nome do documento HTML da publicação periódica extraída.
Texto limpo *tags*	Texto extraído nas etiquetas <a> e <p> pré-processado.
Texto limpo *get text*	Texto extraído de forma integral através de *get_text()* pré-processado.

Fonte: Elaboração própria.

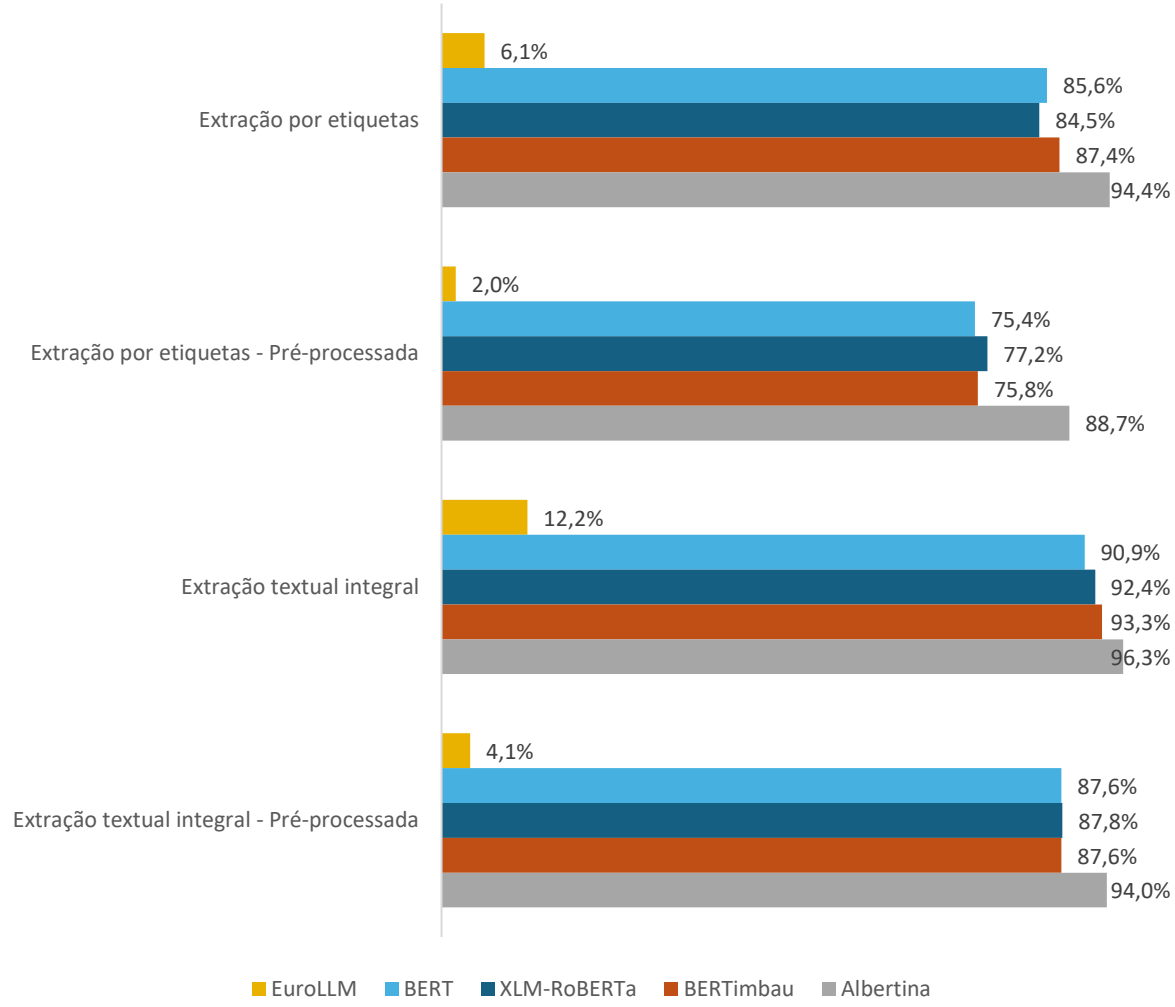
Anexo C

Análise de Documentos Truncados



Fonte: Elaboração própria.

Base de Dados para a Classificação Multiclasse



Fonte: Elaboração própria.

Base de Dados para a Classificação Binária

Anexo D

Métricas de Desempenho por Categoria dos Melhores Modelos Obtidos na Classificação Multiclasse e Binária

Modelo mBERT (dados extraídos de forma integral e sem pré-processamento) – 5ª Experiência – Classificação Multiclasse

Categoria	Precisão	Sensibilidade	F1	Amostras
Desporto	100,0%	100,0%	100,0%	12
Religião	100,0%	100,0%	100,0%	5
Economia, finanças, gestão e negócios	86,7%	92,9%	89,7%	28
Estilo de vida e lazer	81,3%	89,7%	85,2%	29
Ciência, educação e história	87,5%	77,8%	82,4%	18
Política, administração pública, direito e sociedade	76,9%	71,4%	74,1%	14
Artes e cultura	66,7%	66,7%	66,7%	6
Doutrinas políticas/sociais	0,0%	0,0%	0,0%	2

Modelo XLM-RoBERTa (dados extraídos por etiquetas e sem pré-processamento) – 5ª Experiência – Classificação Multiclasse

Categoria	Precisão	Sensibilidade	F1	Amostras
Desporto	100,0%	100,0%	100,0%	12
Economia, finanças, gestão e negócios	89,3%	89,3%	89,3%	28
Estilo de vida e lazer	84,4%	93,1%	88,5%	29
Ciência, educação e história	92,9%	72,2%	81,3%	18
Religião	80,0%	80,0%	80,0%	5
Política, administração pública, direito e sociedade	68,4%	92,9%	78,8%	14
Artes e cultura	75,0%	50,0%	60,0%	6
Doutrinas políticas/sociais	0,0%	0,0%	0,0%	2

Modelo BERTimbau (dados extraídos de forma integral sem pré-processamento) – 5º Experiência – Classificação Binária

Categoria	Precisão	Sensibilidade	F1	Amostras
Informação especializada	97,2%	94,6%	95,9%	74
Informação geral	94,7%	97,3%	96,0%	74

Nota: As categorias estão ordenadas pelo valor de F1 de forma decrescente.
As métricas referem-se ao desempenho na classe positiva: "Informação geral".
A categoria "Informação especializada" integra as publicações periódicas de conteúdo doutrinário.
Fonte: Elaboração própria.