Fusão de Rankings Baseada em Confiança: Um Estudo sobre Calibração de Scores e Fusão de Rankings para Classificação de Texto Multi-Classe

Gabriel Franco Jallais
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil
gabrieljallais@dcc.ufmg.br

Marcos André Gonçalves
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil
mgoncalv@dcc.ufmg.br

Celso França
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil
celsofranca@dcc.ufmg.br

***Abstract*—Pipelines de recuperação que combinam recuperadores esparsos (BM25) e densos (baseados em transformadores) têm se mostrado promissores para tarefas de Classificação de Texto Multi-Rótulo Extrema (*Extreme Multi-Label Text Classification* – XMTC), explorando a complementaridade entre correspondências lexicais e semânticas. Contudo, os *scores* produzidos por esses recuperadores não são calibrados, isto é, não representam probabilidades reais de relevância. Neste trabalho, investiga-se a aplicação de métodos de calibração de *scores* – incluindo *Platt Scaling*, Regressão Isotônica e um método proposto denominado *QueryFeature Calibration* – ao *pipeline* xCoRetriev. Para permitir uma análise controlada utilizando métricas de calibração binária padrão, os experimentos foram conduzidos no contexto de Classificação de Texto Multi-Classe (MCTC), onde cada documento possui um único rótulo relevante. Experimentos em três *benchmarks* (REUTERS, ACM e TWITTER) demonstram que, embora a calibração isolada não melhore significativamente métricas de ranking em 2 dos 3 datasets, os métodos de calibração aplicados apresentam boa capacidade de calibrar os *scores* retornados pelos recuperadores, em especial, o método *QueryFeature Calibration* apresentou um desempenho muito bom, abrindo perspectivas para abordagens contextuais de fusão de *rankings* aplicáveis tanto a MCTC quanto a cenários XMTC.**

***Index Terms*—classificação de texto multi-classe, calibração de scores, fusão de rankings, *Platt Scaling*, regressão isotônica, recuperação de informação, pipelines de recuperação**

I. Introdução

A classificação automática de documentos é uma tarefa central em sistemas de organização de informação, desde a categorização de produtos em plataformas de *e-commerce* até a anotação de artigos científicos em bases bibliográficas [1], [2]. Em cenários reais, um mesmo documento frequentemente pertence a múltiplas categorias simultâneas, caracterizando o problema de classificação multi-rótulo. Quando o número de categorias possíveis atinge milhares ou milhões, essa tarefa recebe a denominação de **Classificação de Texto Multi-Rótulo Extrema** (*Extreme Multi-Label Text Classification* — XMTC).

A escala extrema do espaço de rótulos em XMTC introduz desafios que tornam inviáveis as abordagens convencionais de classificação. Uma solução promissora, proposta em trabalhos recentes, é reformular o problema sob a ótica de Recuperação de Informação (RI), utilizando *pipelines* que combinam recuperadores esparsos (BM25) e densos (baseados em transformadores). Este trabalho investiga um aspecto ainda pouco explorado nesse contexto: a calibração de *scores* de relevância produzidos por esses sistemas de recuperação.

A motivação central é que, em *pipelines* modernos baseados em fusão de recuperadores, os *scores* retornados não possuem interpretação probabilística — um *score* de 0,8 não significa 80% de probabilidade de relevância. Essa descalibração pode comprometer tanto a combinação de múltiplos recuperadores quanto a tomada de decisão baseada em limiares de confiança.

Para permitir uma análise controlada de calibração, utilizando métricas de calibração binária padrão (ECE, MCE, Brier Score), este trabalho foca em Classificação de Texto Multi-Classe (MCTC), onde cada documento possui um único rótulo relevante. Essa simplificação preserva a arquitetura e metodologia de *pipelines* desenvolvidos para XMTC, enquanto viabiliza o uso de calibradores e métricas tradicionais. Os resultados e métodos propostos são diretamente aplicáveis ao contexto mais amplo de XMTC.

A. Definição Formal do Problema

Formalmente, seja $\mathcal{X}$ o espaço de documentos de texto e $\mathcal{L} = \{l_1, l_2, \ldots, l_L\}$ o espaço de rótulos com $|\mathcal{L}| = L$ rótulos. No problema geral de XMTC, dado um documento $x \in \mathcal{X}$, o objetivo é prever o conjunto de rótulos relevantes $Y \subseteq \mathcal{L}$, onde $|Y| \geq 1$. No caso especial de Classificação

Multi-Classe (MCTC), cada documento possui exatamente um rótulo relevante, isto é, $|Y| = 1$. Matematicamente:

$$f : \mathcal{X} \to \mathcal{L} \quad \text{(MCTC)}, \quad f : \mathcal{X} \to 2^{\mathcal{L}} \quad \text{(XMTC)} \qquad (1)$$

Na prática, a função f é frequentemente implementada como um ranqueador: ao invés de produzir diretamente um subconjunto de rótulos, o sistema retorna uma lista ordenada $(l_{\pi(1)}, l_{\pi(2)}, \ldots, l_{\pi(L)})$ onde π é uma permutação que ordena os rótulos por *score* de relevância decrescente. O rótulo predito (em MCTC) ou conjunto de rótulos (em XMTC) é então obtido selecionando os *top-k* ou aplicando um limiar aos *scores*. Essa formulação como problema de ranking é particularmente relevante para o presente trabalho, pois a qualidade dos *scores* — e não apenas a ordenação — torna-se crítica quando se deseja interpretar a confiança das predições.

B. Por que XMTC é Difícil?

A dificuldade de XMTC não reside apenas no número de categorias, mas na interação de fatores que invalidam pressupostos de métodos tradicionais de classificação.

O primeiro fator é a **inviabilidade computacional de abordagens exaustivas**. Em um classificador convencional *one-vs-all*, cada categoria requer um modelo independente; com $L = 500.000$ categorias, isso implica treinar e armazenar 500.000 modelos. Redes neurais com camadas de saída densas enfrentam problema análogo: uma camada final com 500.000 neurônios torna o treinamento proibitivo. Soluções arquiteturais como partições hierárquicas [2] ou classificadores lineares distribuídos [1] mitigam esse custo, mas não o eliminam.

O segundo fator é a **escassez de exemplos para a maioria das categorias**. Em bases reais, a frequência de rótulos segue distribuições de cauda longa: enquanto algumas categorias aparecem em milhares de documentos, a maioria ocorre em menos de dez. Essa assimetria — formalizada pela métrica de *propensity* [3] — cria um viés sistemático: modelos aprendem a predizer categorias frequentes com alta confiança, mas falham em generalizar para categorias raras. Do ponto de vista de calibração, isso significa que *scores* altos para categorias frequentes não têm o mesmo significado que *scores* altos para categorias raras.

O terceiro fator é a **disparidade semântica entre documentos e rótulos**. Documentos podem conter milhares de palavras; rótulos são tipicamente frases curtas ou termos isolados. Essa diferença de granularidade dificulta o mapeamento direto, especialmente quando o rótulo é ambíguo fora de contexto (e.g., "Java" pode referir-se a uma linguagem de programação, uma ilha ou um tipo de café).

C. Recuperação de Informação como Paradigma para XMTC

Uma abordagem promissora para lidar com os fatores acima é reformular XMTC sob a ótica de Recuperação de Informação (RI). Nessa perspectiva, cada documento a ser classificado é tratado como uma consulta, e as categorias possíveis como documentos em uma coleção a ser pesquisada. Essa reformulação permite aplicar técnicas consolidadas de RI, como recuperadores esparsos (e.g., BM25 [4]), que exploram correspondências lexicais exatas, e recuperadores densos, que utilizam representações vetoriais de modelos de linguagem [5] para capturar similaridades semânticas.

A literatura recente demonstra que combinar recuperadores esparsos e densos via fusão de *rankings* produz resultados superiores aos de cada recuperador isolado [6]. A intuição é que os dois tipos de recuperadores cometem erros em situações distintas: recuperadores esparsos falham quando o vocabulário do documento não coincide com o do rótulo; recuperadores densos falham quando a relação documento-rótulo depende de termos específicos ausentes no espaço semântico aprendido. A fusão explora essa complementaridade.

Contudo, a fusão de *rankings* pressupõe que os *scores* dos recuperadores sejam comparáveis — o que raramente é verdade. Este trabalho investiga justamente essa lacuna: como a calibração dos *scores* afeta a qualidade da fusão?

D. O Problema de Calibração: Motivação Central

Apesar da eficácia da fusão de recuperadores heterogêneos, um problema fundamental permanece inexplorado: os *scores* de relevância produzidos não são calibrados. Em termos técnicos, um *score* é calibrado quando pode ser interpretado como uma probabilidade: se um modelo é perfeitamente calibrado, entre todas as predições com *score* 0,8, exatamente 80% deveria corresponder a eventos positivos (rótulos relevantes).

Na prática, os *scores* de recuperadores não possuem essa propriedade. Um valor de 0,8 retornado pelo BM25 representa uma pontuação arbitrária cuja escala depende da coleção e da consulta. Modelos densos baseados em transformadores frequentemente exibem padrões de sobre-confiança (*overconfidence*): predizem com alta confiança mesmo quando incorretos [7].

Essa ausência de calibração introduz três dificuldades específicas para a fusão de *rankings*. Primeiramente, *scores* de diferentes recuperadores residem em escalas incompatíveis, dificultando combinação aritmética direta. Além disso, um recuperador sistematicamente sobre-confiante pode dominar indevidamente o resultado fusionado, mesmo quando suas predições são menos precisas. Por fim, sem calibração adequada, não é possível usar limiares de confiança para decisões como "aceitar apenas rótulos com probabilidade ≥ 0.7".

A hipótese central deste trabalho é que calibrar os *scores* antes da fusão pode melhorar tanto a qualidade do ranking fusionado quanto a interpretabilidade das predições. Para testá-la, investigaram-se métodos clássicos de calibração e foi proposta uma abordagem contextual que considera características da consulta.

E. Objetivos e Contribuições

Este trabalho investiga a aplicação de métodos de calibração de *scores* a *pipelines* de recuperação para classificação de texto, utilizando MCTC como cenário experimental controlado. Os objetivos são quatro: (i) avaliar métodos clássicos de calibração — *Platt Scaling* e Regressão Isotônica — quando aplicados a *scores* de recuperadores esparsos e densos; (ii) propor o método *QueryFeature Calibration*, uma

abordagem contextual que considera características da consulta para produzir probabilidades calibradas; (iii) investigar variantes de fusão ponderada por confiança que utilizam os *scores* calibrados como pesos de combinação; e (iv) avaliar experimentalmente o impacto da calibração na qualidade da fusão em três *benchmarks* de classificação multi-classe.

As principais contribuições incluem: uma análise sistemática de métodos de calibração para *pipelines* de recuperação; o método *QueryFeature Calibration*; duas variantes de fusão ponderadas por confiança (CombMNZ-Conf e CombMULT-Conf); e evidências experimentais de que calibração melhora a qualidade probabilística dos *scores* mas não impacta significativamente as métricas de ranking nos cenários avaliados.

F. Organização do Trabalho

O restante deste trabalho está organizado da seguinte forma: a Seção II apresenta trabalhos relacionados sobre classificação de texto, abordando tanto métodos de Classificação Multi-Classe (MCTC) quanto de XMTC, além de técnicas de calibração de *scores* e fusão de *rankings*; a Seção III descreve a metodologia proposta, incluindo o *pipeline* de calibração e as variantes de fusão; a Seção IV detalha a configuração experimental; a Seção V apresenta e discute os resultados obtidos; e a Seção VI conclui o trabalho apontando limitações e direções futuras.

II. Trabalhos Relacionados

Esta seção revisa a literatura relevante para os pilares deste trabalho: métodos de Classificação de Texto Multi-Classe (MCTC), sua extensão para cenários extremos (XMTC), técnicas de calibração de *scores* e algoritmos de fusão de *rankings*. A discussão destaca como o presente trabalho se posiciona em relação a essas linhas de pesquisa.

A. Classificação de Texto Multi-Classe

A Classificação de Texto Multi-Classe (*Multi-Class Text Classification* — MCTC) é uma tarefa fundamental em Processamento de Linguagem Natural (PLN), onde cada documento deve ser atribuído a exatamente uma categoria dentre um conjunto predefinido [8]. Esta tarefa constitui a base sobre a qual variantes mais complexas, como a classificação multi-rótulo, são construídas.

Métodos tradicionais de MCTC incluem representações baseadas em *bag-of-words* combinadas com classificadores como *Naive Bayes*, SVM e Regressão Logística [9]. A introdução de *word embeddings* (Word2Vec, GloVe) e modelos neurais impulsionou avanços significativos. Kim [10] propôs o uso de Redes Neurais Convolucionais (CNNs) para classificação de sentenças, demonstrando que arquiteturas simples com filtros convolucionais sobre representações pré-treinadas alcançam resultados competitivos. Abordagens baseadas em redes recorrentes, como LSTMs e GRUs, capturam dependências sequenciais e tornaram-se populares para tarefas de classificação [11].

Mais recentemente, modelos baseados em *transformers* pré-treinados, como BERT e RoBERTa, estabeleceram novos patamares de desempenho em MCTC [12]. Esses modelos aprendem representações contextualizadas que capturam relações semânticas complexas, permitindo *fine-tuning* eficiente para tarefas específicas com quantidades moderadas de dados rotulados.

B. Classificação de Texto Multi-Rótulo Extrema

Quando o número de categorias atinge milhares ou milhões, e cada documento pode pertencer a múltiplas categorias simultaneamente, a tarefa assume características de XMTC [8]. Este cenário extremo introduz desafios computacionais e estatísticos que tornam inviáveis as abordagens convencionais de MCTC.

Métodos clássicos como DiSMEC [1] adotam classificadores lineares independentes para cada rótulo, escalando via paralelismo distribuído. Abordagens mais recentes utilizam árvores hierárquicas de rótulos [2] para reduzir o espaço de busca de forma hierárquica. O problema do desbalanceamento motivou o desenvolvimento de métricas ponderadas por propensão [3], que atribuem maior recompensa a predições corretas de rótulos raros.

Recentemente, França et al. [6] propuseram o xCoRetriev, reformulando XMTC como tarefa de Recuperação de Informação. O sistema combina recuperadores esparsos e densos em um *pipeline* de dois estágios com fusão de *rankings*. O presente trabalho utiliza o cenário de MCTC como ambiente controlado para investigar se a calibração dos *scores* pode melhorar a qualidade da fusão, com perspectivas de aplicação em XMTC.

C. Calibração de Scores

Conforme discutido na Introdução, calibração refere-se à propriedade de que *scores* de confiança correspondam a probabilidades reais [13]. Um classificador é perfeitamente calibrado se, para qualquer *score* p, a proporção de instâncias verdadeiramente positivas é exatamente p. A seguir, descrevemos os principais métodos de calibração avaliados neste trabalho.

1) Platt Scaling: Proposta originalmente para calibrar *scores* de SVMs [14], *Platt Scaling* ajusta uma função sigmoide aos *scores* brutos:

$$P(y = 1|s) = \frac{1}{1 + \exp(-(As + B))} \tag{2}$$

onde s é o *score* original e A, B são parâmetros aprendidos via regressão logística no conjunto de validação. O método destaca-se pela simplicidade e eficiência computacional, requerendo poucos dados para ajuste. Porém, assume que a distorção dos *scores* segue uma forma sigmoide, o que nem sempre é válido.

2) Regressão Isotônica: A Regressão Isotônica [15] é um método não-paramétrico que ajusta uma função monotonicamente não-decrescente aos *scores*:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (y_i - f(s_i))^2 \tag{3}$$

sujeito a $f(s_i) \leq f(s_j)$ sempre que $s_i \leq s_j$, onde $\mathcal{F}$ é o conjunto de funções em escada. A principal vantagem é a flexibilidade: o método não assume forma funcional específica e preserva a ordenação original. Entretanto, é propenso a *overfitting* em conjuntos pequenos e pode produzir platôs indesejados na função calibrada.

Para mitigar o *overfitting*, implementou-se uma variante com *binning* adaptativo e suavização de Laplace, onde os dados são agrupados em *bins* e probabilidades suavizadas são calculadas:

$$\hat{p}_{\text{bin}} = \frac{n_+ + \epsilon}{n_{\text{total}} + 2\epsilon} \tag{4}$$

onde n_+ é o número de positivos no *bin*, n_{total} o total de amostras e ϵ o parâmetro de suavização.

3) Temperature Scaling: Temperature Scaling [7] é uma técnica simples que escala os *logits* por um parâmetro de temperatura T:

$$P(y = 1|z) = \sigma\left(\frac{z}{T}\right) \tag{5}$$

onde $z = \log(s/(1-s))$ é o *logit* correspondente ao *score* s e σ é a função sigmoide. O parâmetro T é otimizado minimizando a *negative log-likelihood* no conjunto de validação. A simplicidade de ter um único parâmetro torna o método eficaz para redes neurais modernas, embora apenas reescale os *scores* sem alterar a forma da distribuição.

4) Calibração Gaussiana: A Calibração Gaussiana [16] generaliza *Platt Scaling* incluindo um termo quadrático:

$$g_\phi(s) = \sigma(as^2 + bs + c) \tag{6}$$

onde $\phi = \{a, b, c\}$ são parâmetros aprendidos sob restrição de monotonicidade ($2as + b > 0$ para todo s no domínio). O método captura distorções quadráticas nos *scores*, porém requer otimização restrita e possui mais parâmetros que *Platt Scaling*.

5) Calibração Gamma: A Calibração Gamma é particularmente adequada para *scores* com distribuição assimétrica:

$$g_\phi(s) = \sigma(a\log(s) + bs + c) \tag{7}$$

onde a transformação logarítmica captura padrões comuns em *scores* de recuperação. É particularmente adequada para *scores* positivos com distribuição assimétrica, embora exija que os valores sejam estritamente positivos.

6) Beta Calibration: A *Beta Calibration* [17] utiliza a função de distribuição acumulada (CDF) da distribuição Beta:

$$P(y = 1|s) = F_{\text{Beta}}(s; \alpha, \beta) \tag{8}$$

onde α e β são parâmetros de forma estimados via máxima verossimilhança. O método é naturalmente adequado para *scores* no intervalo $[0, 1]$ e pode capturar assimetria na distribuição, embora assuma distribuição Beta e possa falhar para padrões mais complexos.

D. Calibração em Recuperação de Informação

Trabalhos recentes investigam calibração especificamente no contexto de *Learning-to-Rank* (LTR). Yan et al. [16] propõem *loss functions* de ranking calibradas que otimizam simultaneamente qualidade do ranking e calibração de *scores*. A motivação é permitir o uso de *scores* calibrados em aplicações que requerem probabilidades, como leilões de anúncios e sistemas de recomendação com limiares de confiança.

Cohen et al. [18] investigam incerteza e calibração em modelos de recuperação neural, propondo um *framework* Bayesiano eficiente que estima distribuições posteriores sobre relevância. Jeon et al. [19] estudam calibração em recomendação sequencial, propondo aprendizado disjunto de calibração e relevância para evitar interferência entre objetivos.

A principal distinção deste trabalho é o foco em calibração como etapa intermediária para fusão de *rankings* em MCTC, investigando como *scores* calibrados podem beneficiar a combinação de recuperadores heterogêneos com escalas e comportamentos distintos.

E. Fusão de Rankings

A fusão de *rankings* (*rank aggregation* ou *data fusion*) é uma técnica fundamental para combinar resultados de múltiplos sistemas de recuperação. O objetivo é produzir um ranking único que aproveite as forças complementares de cada sistema individual [20].

1) Métodos Clássicos de Fusão: Os métodos de fusão de *rankings* podem ser categorizados em duas famílias: métodos baseados em *scores* e métodos baseados em posições [20]:

CombSUM: Soma simples dos *scores* normalizados:

$$\text{score}_{\text{fused}}(d) = \sum_{i=1}^{n} s_i(d) \tag{9}$$

onde $s_i(d)$ é o score do documento d no sistema i.

CombMNZ [21]: Multiplica a soma por um fator que favorece documentos retornados por múltiplos sistemas:

$$\text{score}_{\text{fused}}(d) = N(d) \cdot \sum_{i=1}^{n} s_i(d) \tag{10}$$

onde $N(d)$ é o número de sistemas que retornaram d.

RRF (Reciprocal Rank Fusion) [20]: Utiliza apenas posições, sendo invariante a escalas de *scores*:

$$\text{score}_{\text{fused}}(d) = \sum_{i=1}^{n} \frac{1}{k + r_i(d)} \tag{11}$$

onde $r_i(d)$ é a posição do documento d no ranking do sistema i e k é um parâmetro (tipicamente $k = 60$).

2) Normalização de Scores: Antes da fusão baseada em *scores*, é comum normalizar os valores para uma escala comum. Métodos tradicionais incluem Min-Max (escala para o intervalo $[0, 1]$), Z-Score (subtração da média e divisão pelo desvio padrão), ZMUV (*Zero Mean Unit Variance*, específica para fusão [22]) e Softmax (conversão em distribuição de probabilidade).

3) Fusão Não-Supervisionada com Aprendizado: Abordagens mais recentes utilizam aprendizado para otimizar a fusão. O método UDLF (*Unsupervised Distance-based Lazy Fusion*) [23] aprende afinidades entre documentos baseando-se em padrões de coocorrência em múltiplos *rankings*, utilizando propagação de similaridades para reranquear resultados.

F. Lacunas e Motivação deste Trabalho

A revisão da literatura revela que, embora existam métodos sofisticados tanto para calibração quanto para fusão de *rankings*, a interseção dessas duas áreas permanece inexplorada no contexto de *pipelines* de recuperação para classificação de texto. Não foram encontrados estudos avaliando se calibrar os *scores* de recuperadores antes da fusão beneficia a qualidade final do ranking. Além disso, métodos de calibração tradicionais são globais (aplicam a mesma transformação independentemente da consulta), ignorando que a confiabilidade de um *score* pode depender do contexto. A metodologia apresentada na próxima seção endereça essas lacunas, utilizando MCTC como cenário experimental controlado.

III. Metodologia

Com base nas lacunas identificadas na Seção II, esta seção apresenta a metodologia proposta. Primeiramente, descreve-se o *pipeline* base (xCoRetriev) sobre o qual este trabalho se constrói. Em seguida, detalha-se a etapa de calibração inserida no *pipeline*, incluindo o método proposto *QueryFeature Calibration*. Por fim, apresentam-se as variantes de fusão ponderada por confiança.

A. Arquitetura de Recuperação Utilizada

Este trabalho utiliza como base a arquitetura proposta por França et al. [6] o pipeline xCoRetriev. No primeiro estágio, o documento de entrada é submetido a dois sistemas de recuperação independentes:

Recuperador Esparso (BM25): Implementação clássica do algoritmo BM25 [4], que pontua rótulos com base na frequência de termos compartilhados entre documento e rótulo. Os *scores* BM25 não são limitados ao intervalo $[0, 1]$ e dependem do tamanho da coleção e da distribuição de termos, o que motiva a necessidade de calibração.

Recuperador Denso: Utiliza um modelo de linguagem pré-treinado (RoBERTa e BERT) para gerar representações vetoriais de documentos e rótulos. A similaridade é computada via produto interno no espaço de *embeddings*. Esses *scores* também não possuem interpretação probabilística e frequentemente exibem sobre-confiança [7].

Cada recuperador produz um ranking de rótulos candidatos, ordenados por *score* de relevância. O segundo estágio combina esses rankings via algoritmos de fusão. A questão central que este trabalho investiga é: *a calibração dos scores antes da fusão melhora o resultado final?*

B. Pipeline de Calibração

Para cada combinação de recuperador (BM25 ou denso) e partição (cabeça ou cauda), um calibrador é treinado no conjunto de validação. O particionamento do espaço de rótulos em **cabeça** e **cauda** segue o critério proposto por França et al. [6], que define rótulos de cauda como aqueles cuja frequência no conjunto de treino está abaixo de um limiar. O processo inicia-se com a extração de pares (*score*, rótulo): para cada consulta q e cada rótulo candidato l recuperado, obtém-se o *score* $s_{q,l}$ e o indicador binário $y_{q,l} \in \{0, 1\}$ de relevância. Os métodos de calibração (Platt, Isotônica, entre outros) são então ajustados a esses pares. Na fase de teste, os *scores* brutos são transformados em probabilidades calibradas.

Formulação Multi-Classe: Conforme descrito na Introdução, este trabalho adota a formulação de **Classificação de Texto Multi-Classe (MCTC)**, onde cada documento possui um único rótulo relevante. Essa escolha metodológica permite: (i) uso direto de métricas de calibração binária padrão (ECE, MCE, Brier Score); (ii) aplicação de calibradores tradicionais sem necessidade de adaptação; e (iii) interpretação clara do indicador de relevância $y_{q,l} \in \{0, 1\}$. Os *datasets* utilizados (REUTERS, ACM, TWITTER) foram processados para garantir exatamente um rótulo por documento. A extensão para cenários verdadeiramente multi-rótulo é discutida como trabalho futuro.

C. QueryFeature Calibration

Propõe-se o método *QueryFeature Calibration*, uma abordagem de calibração contextual que considera não apenas o *score* bruto, mas também características extraídas da consulta e da distribuição de *scores*. A motivação é que a confiabilidade de um *score* depende do contexto: um *score* de 0,7 pode ser altamente confiável para consultas "fáceis" (onde a maioria dos *scores* são pequenos) mas menos confiável para consultas "difíceis".

1) Extração de Características: Para cada par (consulta, *score*), extraem-se três categorias de características. A primeira categoria compreende indicadores de dificuldade da consulta, derivados de estatísticas da distribuição de *scores* retornados: média, desvio padrão, mínimo, máximo, mediana, assimetria (*skewness*), curtose (*kurtosis*) e amplitude. A intuição é que consultas "fáceis" produzem distribuições de *scores* com alta assimetria positiva (poucos candidatos com *scores* altos, a maioria com *scores* baixos) e alta curtose (distribuição concentrada), enquanto consultas "difíceis" produzem distribuições mais uniformes. A segunda categoria abrange características de especificidade, quantificadas pela entropia $H = -\sum_i p_i \log(p_i)$ e pelo índice de Gini $G = 1 - \sum_i p_i^2$, onde p_i são os *scores* normalizados. Consultas específicas têm baixa entropia (poucos candidatos dominão). A terceira categoria contém indicadores contextuais do *score* individual, incluindo sua posição relativa no ranking, o Z-score $z = (s - \mu_q)/\sigma_q$ e o percentil na distribuição.

2) Modelo de Calibração: Utiliza-se um classificador *Gradient Boosting* [24] para mapear as características extraídas à probabilidade de relevância:

$$P(y = 1|s, q) = \text{GBM}(\mathbf{f}(s, q)) \tag{12}$$

onde $\mathbf{f}(s, q)$ é o vetor de características. O modelo foi treinado com os seguintes hiperparâmetros: 100 árvores, profundidade máxima 5, taxa de aprendizado 0,1 e *subsampling* de 0,8. Esses valores foram selecionados via validação cruzada 5-*fold* no conjunto de validação.

O *QueryFeature Calibration* satisfaz os critérios formais de um calibrador probabilístico: produz *outputs* no intervalo $[0, 1]$, interpretáveis como $P(\text{relevância} = 1|s, q)$; é treinado via minimização do *log-loss*, correspondente à maximização da verossimilhança sob modelo de Bernoulli; e é uma função monotônica.

A diferença fundamental em relação a métodos tradicionais é a natureza **condicional** da calibração. Enquanto *Platt Scaling* e Regressão Isotônica aprendem mapeamentos globais $s \mapsto P(y = 1)$, assumindo relação estacionária entre *score* e probabilidade, o *QueryFeature Calibration* aprende mapeamentos condicionados ao contexto: $(s, \mathbf{f}(q)) \mapsto P(y = 1)$. Essa abordagem reconhece que a mesma magnitude de *score* pode ter significados distintos dependendo da "dificuldade" da consulta (capturada pelo desvio padrão) e da "especificidade" da recuperação (capturada pela entropia).

D. Fusão Ponderada por Confiança

Propõem-se duas variantes de fusão que utilizam os *scores* calibrados como pesos de confiança:

1) CombMNZ com Confiança (CombMNZ-Conf): Adapta-se a fórmula clássica de CombMNZ [21] para incorporar confiança:

$$\text{score}_{\text{fused}}(d) = \left(\sum_i c_i \cdot s_i(d) \right) \cdot N(d) \tag{13}$$

onde c_i é a confiança calibrada do recuperador i para o documento d, $s_i(d)$ é o *score* original e $N(d)$ é o número de recuperadores que retornaram d.

2) CombMULT com Confiança (CombMULT-Conf): Variante multiplicativa:

$$\text{score}_{\text{fused}}(d) = \prod_i (1 + c_i \cdot s_i(d)) \tag{14}$$

Esta formulação amplifica documentos com alta confiança em múltiplos recuperadores.

E. Fusão via UDLF-RLSIM

Além das variantes de fusão por confiança propostas, este trabalho avalia o método UDLF (*Unsupervised Distance-based Lazy Fusion*) [23], especificamente a variante RLSIM (*Reciprocal kNN Late Similarity Fusion*). Este método foi incluído como *baseline* de fusão avançada para comparar com as abordagens baseadas em calibração.

Funcionamento do UDLF-RLSIM: Diferentemente de métodos de fusão baseados em *scores* (como CombMNZ), o UDLF opera sobre as **posições** dos itens nos rankings. O algoritmo constrói um grafo de afinidade entre rótulos baseado em padrões de coocorrência: se dois rótulos frequentemente aparecem próximos nas listas de múltiplos recuperadores, eles são considerados similares. Essa matriz de similaridade é então utilizada para propagar relevância: rótulos bem posicionados "emprestam" relevância a seus vizinhos no grafo.

A variante RLSIM utiliza a distância recíproca de kNN para computar afinidade:

$$\text{RLSIM}(d_i, d_j) = \sum_k \frac{1}{r_k(d_i) + r_k(d_j)} \tag{15}$$

onde $r_k(d)$ é a posição do documento d no ranking do sistema k.

Motivação para inclusão: O UDLF-RLSIM representa uma abordagem ortogonal à calibração: enquanto a calibração busca tornar os *scores* interpretáveis, o UDLF ignora os *scores* completamente e opera apenas sobre ordenações. Comparar essas abordagens permite avaliar se a informação contida nos *scores* calibrados agrega valor além da simples ordenação.

Implementação: Utilizou-se a biblioteca pyUDLF [23], que implementa diversos métodos de fusão não-supervisionada. Os rankings de cada recuperador foram convertidos para o formato esperado pela biblioteca, e o método RLSIM foi aplicado com parâmetros padrão ($k = 20$ vizinhos).

IV. Configuração Experimental

A. Datasets

Os experimentos foram conduzidos em três *benchmarks* de classificação de texto multi-classe com características distintas:

TABLE I
Estatísticas dos Datasets (Formulação Multi-Classe)

Dataset	Documentos	Rótulos	Rót./Doc
REUTERS	13.327	90	1,00†
ACM	24.897	11	1,00†
TWITTER	6.997	6	1,00†

† Cada documento possui exatamente um rótulo relevante.

1) REUTERS: O dataset REUTERS-21578 é um *benchmark* clássico de classificação de texto, composto por artigos de notícias da agência Reuters. Apesar de seu tamanho relativamente pequeno (cerca de 13 mil documentos), é amplamente utilizado devido à qualidade das anotações e disponibilidade de partições padronizadas. Os textos são curtos a médios, com linguagem jornalística formal e rótulos temáticos bem definidos. A distribuição de rótulos é desbalanceada, e o tamanho reduzido do dataset aumenta o risco de *overfitting* nos calibradores.

2) ACM: O dataset ACM é derivado da ACM Digital Library, contendo abstracts de artigos científicos associados a categorias do ACM Computing Classification System. Apresenta textos técnicos com vocabulário especializado e estrutura formal de abstracts acadêmicos. Os principais desafios incluem: forte desbalanceamento entre rótulos frequentes e raros, e vocabulário técnico que pode favorecer o recuperador esparso para rótulos específicos.

3) TWITTER: O dataset TWITTER consiste em *tweets* associados a categorias de sentimento ou tópicos, tratados como classificação multi-classe. Os textos são muito curtos (limite de 280 caracteres), com linguagem informal, alta variabilidade de estilo e presença de ruído (URLs, menções, emojis). Essa natureza ruidosa desafia tanto recuperadores esparsos (vocabulário não padronizado) quanto densos (contexto limitado).

B. Métricas de Avaliação

Para avaliação de ranking, utilizaram-se Precision@k (P@k), que mede a fração de rótulos relevantes nas *top-k* predições, sua versão ponderada por propensão (ps-Precision@k), nDCG@k (*Normalized Discounted Cumulative Gain*) e psnDCG@k. Para avaliação de calibração, empregaram-se três métricas: ECE (*Expected Calibration Error*) [13], definida como ECE $= \sum_{b=1}^{B} \frac{|B_b|}{n} |\text{acc}(B_b) - \text{conf}(B_b)|$, que mede o erro médio ponderado entre confiança e acurácia por *bin*; MCE (*Maximum Calibration Error*), que captura o máximo erro de calibração entre *bins*; e *Brier Score* [25], BS $= \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2$, que mensura o erro quadrático médio das probabilidades.

C. Baselines e Configurações

Foram avaliadas quatro configurações de calibração: *Baseline* (fusão sem calibração, CombMNZ padrão), *Platt* (calibração via *Platt Scaling*), *Isotonic* (calibração via Regressão Isotônica) e *QueryFeature* (calibração via método proposto). Para cada configuração, testaram-se os métodos de fusão CombMNZ, CombMNZ-Conf, CombMULT-Conf e UDLF-RLSIM. Métodos adicionais de calibração (*Temperature Scaling*, *Gaussian*, *Gamma*, *Beta*) foram avaliados preliminarmente, mas não apresentaram resultados competitivos e são omitidos da análise principal. Os resultados reportados são médias e intervalos de confiança de 95% sobre 10 *folds*.

V. Resultados e Discussão

A. Qualidade da Calibração

A Tabela II apresenta métricas de calibração para o recuperador BM25 no dataset REUTERS. Resultados similares foram observados para o recuperador denso e demais datasets.

TABLE II
Qualidade de Calibração – BM25 REUTERS (Head)

Método	ECE ↓	MCE ↓	Brier ↓
Baseline (sem calib.)	0,342	0,512	0,187
Platt Scaling	0,089	0,156	0,142
Isotonic	0,076	0,143	0,138
QueryFeature	**0,052**	**0,098**	**0,125**

Observa-se que todos os métodos de calibração melhoram significativamente a qualidade em relação ao *baseline* não calibrado. O método *QueryFeature Calibration* apresenta os melhores resultados em todas as métricas, confirmando que a consideração do contexto da consulta beneficia a calibração.

A Tabela III apresenta os resultados para o recuperador denso (baseado em RoBERTa).

TABLE III
Qualidade de Calibração – Denso REUTERS (Head)

Método	ECE ↓	MCE ↓	Brier ↓
Baseline (sem calib.)	0,289	0,478	0,165
Platt Scaling	0,078	0,142	0,128
Isotonic	0,065	0,128	0,122
QueryFeature	**0,048**	**0,092**	**0,118**

Observa-se que o recuperador denso apresenta menor erro de calibração inicial (ECE de 0,289 vs 0,342), possivelmente devido à natureza mais suave das representações densas. Após calibração, ambos os recuperadores atingem níveis similares de qualidade.

Os diagramas de confiabilidade (Figura 1) confirmam visualmente esses resultados. O *baseline* apresenta forte descalibração, com confiança sistematicamente maior que a acurácia real (sobre-confiança). Após calibração, a curva aproxima-se da diagonal ideal.

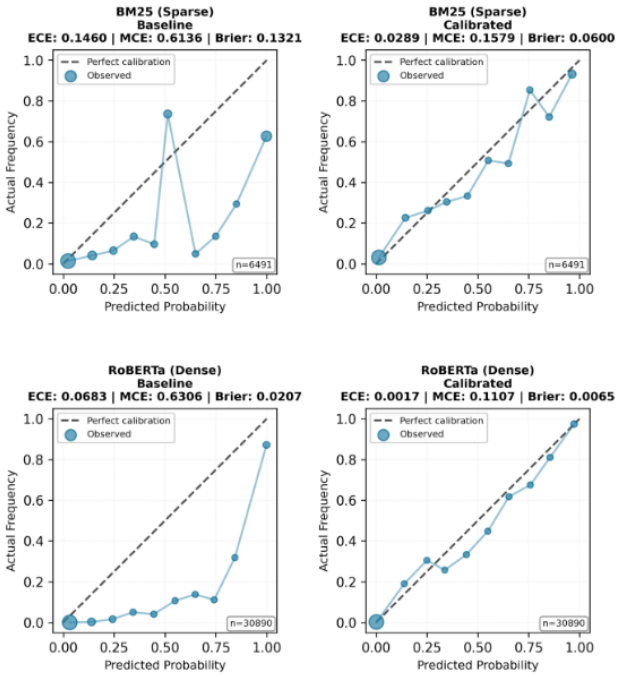

Fig. 1. Diagrama de confiabilidade – REUTERS

B. Impacto na Fusão de Rankings

A Tabela IV apresenta os resultados usando fusão CombMNZ, que soma os *scores* normalizados multiplicando pelo número de sistemas que retornaram cada documento. Os resultados estão separados por rótulos frequentes (*head*) e raros (*tail*).

Destaca-se o caso TWITTER: Platt e Isotonic elevam P@1 de 91% para 98.6–98.8% no *head*, um ganho de 7.6 p.p.— comportamento não observado em REUTERS e ACM. No *tail*, a calibração prejudica REUTERS (baseline 86.3% vs

TABLE IV
RESULTADOS DE FUSÃO COM COMBMNZ (MÉDIA ± IC 95%)

			Precision			nDCG		
Dataset	**Tipo**	**Calibração**	@1	@5	@10	@1	@5	@10
REUTERS	Head	Baseline	93.7±0.9	19.9±0.0	10.0±0.0	93.7±0.9	97.2±0.4	97.3±0.4
		Platt	92.7±1.1	19.9±0.0	10.0±0.0	92.7±1.1	96.9±0.5	97.0±0.5
		Isotonic	93.1±1.1	20.0±0.0	10.0±0.0	93.1±1.1	97.0±0.5	97.1±0.5
		QueryFeat	92.9±1.1	19.9±0.0	10.0±0.0	92.9±1.1	96.9±0.5	97.0±0.5
	Tail	Baseline	86.3±2.5	19.5±0.1	9.9±0.0	86.3±2.5	92.6±1.3	93.2±1.2
		Platt	84.8±2.8	19.5±0.1	9.9±0.0	84.8±2.8	92.0±1.2	92.4±1.2
		Isotonic	84.9±1.9	19.5±0.1	9.9±0.0	84.9±1.9	92.1±0.8	92.5±0.8
		QueryFeat	84.5±2.2	19.4±0.1	9.9±0.0	84.5±2.2	91.3±0.9	91.9±1.0
ACM	Head	Baseline	89.4±0.6	20.0±0.0	10.0±0.0	89.4±0.6	95.8±0.2	95.8±0.2
		Platt	89.6±0.5	20.0±0.0	10.0±0.0	89.6±0.5	95.9±0.2	95.9±0.2
		Isotonic	89.7±0.4	20.0±0.0	10.0±0.0	89.7±0.4	95.9±0.2	95.9±0.2
		QueryFeat	90.4±0.5	20.0±0.0	10.0±0.0	90.4±0.5	96.2±0.2	96.2±0.2
	Tail	Baseline	83.8±0.8	19.6±0.1	10.0±0.0	83.8±0.8	91.8±0.5	92.6±0.4
		Platt	82.2±1.2	19.8±0.0	10.0±0.0	82.2±1.2	91.8±0.5	92.2±0.5
		Isotonic	82.2±1.2	19.8±0.1	10.0±0.0	82.2±1.2	91.8±0.5	92.1±0.5
		QueryFeat	83.6±0.9	19.8±0.0	10.0±0.0	83.6±0.9	92.4±0.4	92.7±0.4
TWITTER	Head	Baseline	91.0±0.5	20.0±0.0	10.0±0.0	91.0±0.5	96.7±0.2	96.7±0.2
		Platt	98.6±0.5	20.0±0.0	10.0±0.0	98.6±0.5	99.5±0.2	99.5±0.2
		Isotonic	98.8±0.5	20.0±0.0	10.0±0.0	98.8±0.5	99.5±0.2	99.5±0.2
		QueryFeat	97.1±1.6	20.0±0.0	10.0±0.0	97.1±1.6	98.9±0.6	98.9±0.6
	Tail	Baseline	86.0±1.7	20.0±0.0	10.0±0.0	86.0±1.7	94.0±0.8	94.0±0.8
		Platt	89.2±1.3	20.0±0.0	10.0±0.0	89.2±1.3	95.3±0.6	95.3±0.6
		Isotonic	90.3±1.4	20.0±0.0	10.0±0.0	90.3±1.4	95.9±0.6	95.9±0.6
		QueryFeat	85.8±3.1	20.0±0.0	10.0±0.0	85.8±3.1	94.1±1.2	94.1±1.2

QueryFeat 84.5%) mas beneficia TWITTER com Isotonic (90.3% vs baseline 86.0%). ACM mostra QueryFeature superior no *head* (90.4%) e *tail* (83.6%).

C. Resultados Agregados

A Tabela V apresenta a melhor configuração de fusão e normalização para cada método de calibração. Para cada combinação calibração×dataset, selecionou-se a configuração com maior ps-nDCG@1. As melhores configurações são: *baseline* usa CombMNZ sem normalização; *Platt* e *Isotonic* favorecem CombMNZ-Variant com normalização ZMUV em REUTERS, mas CombMNZ simples em ACM; *QueryFeature* alcança melhores resultados com CombMULT-Conf (REUTERS) ou CombMNZ-Variant (ACM/TWITTER).

As variações entre métodos de calibração são marginais para REUTERS e ACM. TWITTER é exceção: Isotonic atinge 96.0%, superando QueryFeature (93.6%) por 2.4 p.p.—possivelmente a simplicidade do dataset (6 classes) beneficia regressão não-paramétrica.

D. Comparação entre Recuperadores: Esparso vs. Denso

Uma análise complementar interessante é a comparação da qualidade de calibração entre os recuperadores esparso (BM25) e denso (baseado em transformadores). A Tabela VI apresenta as métricas de calibração antes e após a aplicação do método *QueryFeature Calibration*.

Observa-se que: (i) ambos os recuperadores apresentam descalibração significativa sem calibração; (ii) o recuperador denso é ligeiramente melhor calibrado naturalmente, possivelmente devido ao treinamento com objetivo discriminativo; (iii) após calibração, ambos atingem níveis similares de qualidade; (iv) a redução relativa de ECE é maior para BM25 (85%) do que para o recuperador denso (83%).

E. Análise dos Diagramas de Confiabilidade

Os diagramas de confiabilidade (*reliability diagrams*) permitem visualizar o grau de calibração de um modelo. Esses diagramas plotam a confiança prevista no eixo X contra a acurácia observada no eixo Y. Um modelo perfeitamente calibrado produziria uma linha diagonal (linha tracejada nos gráficos).

A Figura 1 mostra quatro diagramas para o dataset REUTERS: BM25 *baseline* (normalização MinMax), recuperador denso *baseline*, BM25 com *QueryFeature Calibration* + ZMUV, e recuperador denso com *QueryFeature Calibration* + ZMUV.

O diagrama evidencia que, após a aplicação da normalização min–max, os scores produzidos tanto pelo BM25 quanto pelo RoBERTa apresentam baixa calibração quando interpretados como probabilidades. Esse efeito manifesta-se pela acentuada distância entre as curvas observadas e a linha da diagonal, indicando um padrão de sobreconfiança em ambos os modelos. Entretanto, após o processo de calibração, observa-se que as curvas passam a alinhar-se de forma consistente com a diagonal, que representa a condição de calibração perfeita. Esse alinhamento demonstra que o método empregado foi eficaz em transformar os scores originais dos recuperadores em estimativas probabilísticas adequadamente calibradas.

TABLE V
MELHORES CONFIGURAÇÕES POR CALIBRAÇÃO (MÉDIA ± IC 95%)

Dataset	Método	Precision			psPrecision			nDCG			psnDCG		
		@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
REUTERS	Baseline	92.6±1.1	19.9±0.0	10.0±0.0	90.1±1.3	98.6±0.3	99.5±0.2	92.6±1.1	96.5±0.5	96.6±0.5	90.1±1.3	95.0±0.7	95.3±0.6
	Platt	92.3±1.1	19.9±0.0	10.0±0.0	89.8±1.2	98.8±0.2	99.8±0.1	92.3±1.1	96.5±0.5	96.6±0.5	89.8±1.2	95.0±0.6	95.3±0.6
	Isotonic	92.2±1.0	19.9±0.0	10.0±0.0	89.7±1.1	98.8±0.2	99.7±0.1	92.2±1.0	96.4±0.5	96.6±0.5	89.7±1.1	95.0±0.5	95.3±0.5
	QueryFeat	92.8±1.1	19.9±0.0	10.0±0.0	90.3±1.2	99.0±0.3	99.7±0.1	92.8±1.1	96.7±0.5	96.8±0.5	90.3±1.2	95.3±0.5	95.5±0.5
ACM	Baseline	87.5±0.3	19.8±0.0	10.0±0.0	87.0±0.4	99.0±0.2	100.0±0.0	87.5±0.3	94.3±0.1	94.6±0.1	87.0±0.4	94.1±0.2	94.4±0.1
	Platt	86.8±0.8	19.9±0.0	10.0±0.0	86.3±0.8	99.5±0.1	100.0±0.0	86.8±0.8	94.3±0.3	94.5±0.3	86.3±0.8	94.1±0.3	94.3±0.3
	Isotonic	86.9±0.8	19.9±0.0	10.0±0.0	86.4±0.9	99.6±0.2	100.0±0.0	86.9±0.8	94.3±0.3	94.5±0.3	86.4±0.9	94.1±0.3	94.2±0.3
	QueryFeat	88.0±0.6	19.9±0.0	10.0±0.0	87.4±0.7	99.5±0.1	100.0±0.0	88.0±0.6	94.8±0.2	94.9±0.2	87.4±0.7	94.5±0.2	94.6±0.2
TWITTER	Baseline	91.8±0.5	20.0±0.0	10.0±0.0	90.8±0.4	100.0±0.0	100.0±0.0	91.8±0.5	96.7±0.2	96.7±0.2	90.8±0.4	96.3±0.2	96.3±0.2
	Platt	96.2±0.5	20.0±0.0	10.0±0.0	95.7±0.6	100.0±0.0	100.0±0.0	96.2±0.5	98.4±0.2	98.4±0.2	95.7±0.6	98.2±0.2	98.2±0.2
	Isotonic	96.6±0.5	20.0±0.0	10.0±0.0	96.0±0.6	100.0±0.0	100.0±0.0	96.6±0.5	98.6±0.2	98.6±0.2	96.0±0.6	98.3±0.3	98.3±0.3
	QueryFeat	94.4±1.3	20.0±0.0	10.0±0.0	93.6±1.3	100.0±0.0	100.0±0.0	94.4±1.3	97.8±0.5	97.8±0.5	93.6±1.3	97.4±0.5	97.4±0.5

TABLE VI
COMPARAÇÃO DE CALIBRAÇÃO POR RECUPERADOR – REUTERS

	Antes			Após QueryFeature		
Recuperador	ECE	MCE	Brier	ECE	MCE	Brier
BM25	0,342	0,512	0,187	0,052	0,098	0,125
Denso	0,289	0,478	0,165	0,048	0,092	0,118

F. Resultados Completos — REUTERS

A Tabela VII apresenta todos os resultados experimentais para o dataset REUTERS, com métricas ponderadas por propensão.

Observa-se que: (i) *QueryFeature Calibration* com normalização ZMUV consistentemente produz os melhores resultados (psnDCG@1 = 92.8, psP@1 = 92.8); (ii) a escolha do método de fusão tem impacto menor que a escolha da calibração; (iii) UDLF-RLSIM apresenta resultados competitivos mas ligeiramente inferiores a CombMNZ; (iv) normalização ZMUV supera MinMax para todos os métodos de calibração.

G. Discussão: Quando a Calibração Beneficia o Ranking?

Os resultados deste estudo revelam um padrão interessante: métodos de calibração melhoram consistentemente a qualidade probabilística dos *scores* (ECE reduzido de 0,342 para 0,052), mas seu impacto nas métricas de ranking varia entre *datasets*. Enquanto REUTERS e ACM apresentam ganhos marginais, TWITTER demonstra melhorias expressivas — P@1 salta de 91% para 98,8% com calibração Isotônica. Esta seção analisa os fatores que determinam quando a calibração beneficia o ranking.

Impacto da calibração no ranking: Métricas como P@k e nDCG@k dependem da ordenação relativa dos itens, não dos valores absolutos dos *scores*. Como os métodos de calibração avaliados (Platt, Isotônica, QueryFeature) são transformações monotônicas, a calibração preserva a ordenação de cada *ranking* individual. O ganho na fusão ocorre quando existe assimetria de confiança entre os recuperadores: se um recuperador sistematicamente produz *scores* menores que o outro, suas contribuições são sub-representadas na combinação. A calibração corrige essa assimetria, mapeando os *scores* de ambos os recuperadores para probabilidades comparáveis. Em *datasets* como TWITTER, onde textos curtos e ruidosos podem levar um recuperador a ter confiança muito menor que o outro, a calibração reequilibra a fusão. Em REUTERS e ACM, com textos mais estruturados, os recuperadores podem já apresentar níveis de confiança mais equilibrados, reduzindo o impacto da calibração.

Quando a fusão ponderada agrega valor? Os resultados sugerem que a fusão ponderada por confiança beneficia cenários onde: (i) os recuperadores discordam frequentemente sobre os candidatos mais relevantes; (ii) existe assimetria significativa na qualidade dos recuperadores para diferentes tipos de consultas; e (iii) as escalas originais dos *scores* são incompatíveis. Essas condições variam conforme as características do *dataset*.

Por que UDLF-RLSIM é robusto a calibração? O UDLF opera exclusivamente sobre posições, ignorando *scores*. Isso explica sua robustez: calibrar os *scores* não afeta as posições, portanto não afeta o UDLF. O UDLF atingiu resultados competitivos com CombMNZ, sugerindo que a informação de coocorrência pode ser tão valiosa quanto a informação de *scores*.

Por que QueryFeature produz melhor calibração? O *QueryFeature Calibration* supera métodos tradicionais porque reconhece que o significado de um *score* depende do contexto. Um *score* de 0,7 em uma consulta onde o máximo é 0,75 tem interpretação diferente de 0,7 em uma consulta onde o máximo é 0,99. As características de desvio padrão e entropia capturam essa "dificuldade" da consulta, permitindo ajustes mais precisos.

H. Limitações do Estudo

Este trabalho apresenta limitações que devem ser consideradas na interpretação dos resultados.

Formulação multi-classe: Conforme descrito na Introdução e na Seção III, este estudo foi conduzido no contexto de classificação multi-classe (MCTC), onde cada documento possui um único rótulo relevante. Essa escolha permitiu uma análise controlada de calibração. A extensão para cenários verdadeiramente multi-rótulo (XMTC) requer adaptações nas métricas e métodos, constituindo uma direção futura importante.

TABLE VII
RESULTADOS COMPLETOS — REUTERS: TODAS AS COMBINAÇÕES DE CALIBRAÇÃO, FUSÃO E NORMALIZAÇÃO

Calibração	Fusão	Norm.	psP@1	psP@10	psnDCG@1	psnDCG@10
Baseline (sem calibração)						
baseline	mnz	–	92.6	45.2	92.6	96.6
baseline	udlf_rlsim	–	92.3	45.0	92.3	96.5
baseline	comb_mnz_conf	–	92.4	45.1	92.4	96.5
baseline	comb_mult_conf	–	92.3	45.0	92.3	96.5
Platt Scaling						
platt	comb_mnz_variant	zmuv	92.3	45.1	92.3	96.6
platt	comb_mult_conf	zmuv	92.2	45.0	92.2	96.6
platt	comb_mnz_conf	zmuv	92.3	45.1	92.3	96.6
platt	mnz	zmuv	92.3	45.1	92.3	96.6
platt	udlf_rlsim	zmuv	92.2	45.0	92.2	96.5
platt	mnz	minmax	91.8	44.8	91.8	96.4
Regressão Isotônica						
isotonic	comb_mult_conf	zmuv	92.2	45.0	92.2	96.6
isotonic	comb_mnz_conf	zmuv	92.2	45.0	92.2	96.6
isotonic	comb_mnz_variant	zmuv	92.2	45.0	92.2	96.6
isotonic	mnz	zmuv	92.2	45.0	92.2	96.6
isotonic	udlf_rlsim	zmuv	92.2	44.9	92.2	96.5
isotonic	mnz	minmax	91.9	44.8	91.9	96.4
QueryFeature Calibration						
queryfeature	comb_mult_conf	zmuv	**92.8**	**45.4**	**92.8**	**96.8**
queryfeature	comb_mnz_variant	zmuv	92.7	45.3	92.7	**96.8**
queryfeature	comb_mnz_conf	zmuv	**92.8**	**45.4**	**92.8**	**96.8**
queryfeature	mnz	zmuv	**92.8**	**45.4**	**92.8**	**96.8**
queryfeature	udlf_rlsim	zmuv	92.7	45.3	92.7	96.7
queryfeature	mnz	minmax	92.5	45.2	92.5	96.6

Análises não realizadas: As hipóteses apresentadas na discussão sobre por que a calibração não melhora o ranking não foram testadas empiricamente (e.g., análise de correlação entre confiança e qualidade, análise de concordância entre recuperadores).

Custo computacional: O método *QueryFeature Calibration* requer treinamento de um modelo adicional (Gradient Boosting), o que pode ser custoso para datasets muito grandes.

VI. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho investigou a aplicação de métodos de calibração de *scores* a *pipelines* de recuperação para Classificação de Texto Multi-Classe (MCTC). Foram avaliados métodos clássicos (*Platt Scaling*, Regressão Isotônica) e o método proposto *QueryFeature Calibration*, que incorpora características contextuais da consulta.

Os experimentos em três *benchmarks* (REUTERS, ACM, TWITTER) demonstraram que todos os métodos reduziram substancialmente as métricas de erro de calibração (ECE, MCE, Brier Score), com o *QueryFeature Calibration* produzindo consistentemente os melhores resultados. O impacto nas métricas de ranking variou conforme o *dataset*: enquanto REUTERS e ACM apresentaram ganhos marginais, TWITTER demonstrou melhoria expressiva (P@1 de 91% para 98,8% com calibração Isotônica). Essa variação sugere que a calibração beneficia especialmente cenários onde existe assimetria significativa de confiança entre os recuperadores, corrigindo o desbalanceamento na fusão.

Os resultados indicam que a calibração é uma técnica valiosa para *pipelines* de fusão, com eficácia dependente das características do *dataset*. A arquitetura e métodos desenvolvidos são diretamente aplicáveis a cenários de XMTC, onde estudos adicionais são necessários.

A. Direções Futuras

Diversas linhas de investigação emergem deste trabalho:

Re-ranking baseado em UDLF: Os *scores* do BM25 apresentam alta variância e baixa discriminabilidade. Aplicar o *framework* UDLF como reranqueador individual antes da fusão poderia refinar a ordenação inicial usando padrões de coocorrência.

Calibração *end-to-end*: O *pipeline* atual treina calibradores separadamente da fusão. Um modelo *end-to-end* que aprenda calibração e fusão conjuntamente poderia otimizar diretamente as métricas de ranking.

Fusão híbrida com UDLF: Combinar informação de coocorrência (UDLF) com informação de confiança calibrada poderia produzir fusões superiores, utilizando *scores* calibrados como pesos no grafo de afinidade.

Extensão para XMTC: Investigar calibração em cenários verdadeiramente multi-rótulo requer adaptações nas métricas e métodos para lidar com múltiplos rótulos relevantes por documento.

APPENDIX

A Tabela VIII apresenta a configuração de fusão e normalização que produziu o melhor nDCG@10 para cada combinação de método de calibração e dataset.

TABLE VIII
MELHORES CONFIGURAÇÕES DE FUSÃO POR CALIBRAÇÃO E DATASET

Dataset	Calibração	Fusão	Normalização
REUTERS	Baseline	CombMNZ	nenhuma
	Platt	CombMNZ-Var	ZMUV
	Isotonic	CombMULT-Conf	ZMUV
	QueryFeature	CombMULT-Conf	ZMUV
ACM	Baseline	CombMNZ	nenhuma
	Platt	CombMNZ	nenhuma
	Isotonic	CombMNZ	nenhuma
	QueryFeature	CombMNZ-Var	nenhuma
TWITTER	Baseline	CombMNZ-Conf	nenhuma
	Platt	CombMNZ-Conf	nenhuma
	Isotonic	CombMNZ	nenhuma
	QueryFeature	CombMNZ-Var	nenhuma

Observa-se que apenas REUTERS se beneficiou da normalização ZMUV dos *scores* antes da fusão. ACM e TWITTER obtiveram melhores resultados sem normalização adicional.

REFERENCES

[1] R. Babbar and B. Schölkopf, "Dismec: Distributed sparse machines for extreme multi-label classification," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '17, 2017, pp. 721–729.

[2] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, and S. Zhu, "Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[3] H. Jain, Y. Prabhu, and M. Varma, "Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, 2016, pp. 935–944.

[4] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.

[6] C. França, G. Rabbi, T. Salles, W. Cunha, L. Rocha, and M. A. Gonçalves, "Optimizing tail-head trade-off for extreme multi-label text classification (xmtc) with rag-labels and a dynamic two-stage retrieval and fusion pipeline," in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '25, 2025, pp. 1392–1401.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. ICML '17, 2017, pp. 1321–1330.

[8] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–40, 2021.

[9] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Cunico, "A survey on text classification algorithms: From text to predictions," *Information*, vol. 13, no. 2, p. 83, 2022.

[10] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '14, 2014, pp. 1746–1751.

[11] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[12] L. Galke and A. Scherp, "Are we really making much progress in text classification? a comparative review," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '22, 2022, pp. 2658–2681.

[13] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI '15, 2015, pp. 2901–2907.

[14] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[15] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02, 2002, pp. 694–699.

[16] L. Yan, Z. Qin, X. Wang, M. Bendersky, and M. Najork, "Scale calibration of deep ranking models," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22, 2022, pp. 4300–4309.

[17] M. Kull, T. S. Filho, and P. Flach, "Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. AISTATS '17, vol. 54, 2017, pp. 623–631.

[18] D. Cohen, B. Mitra, O. Lesota, N. Rekabsaz, and C. Eickhoff, "Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21, 2021, pp. 654–664.

[19] H. Jeon, S.-e. Yoon, and J. McAuley, "Calibration-disentangled learning and relevance-prioritized reranking for calibrated sequential recommendation," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, ser. CIKM '24, 2024, pp. 973–982.

[20] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '09, 2009, pp. 758–759.

[21] E. A. Fox and J. A. Shaw, "Combination of multiple searches," *NIST Special Publication*, pp. 243–243, 1994.

[22] M. Montague and J. A. Aslam, "Condorcet fusion for improved retrieval," in *Proceedings of the 11th International Conference on Information and Knowledge Management*, ser. CIKM '02, 2002, pp. 538–548.

[23] S. Vargas, D. C. G. Pedronette, and R. d. S. Torres, "A graph-based ranked-list model for unsupervised distance learning on shape retrieval," vol. 83, 2016, pp. 357–367.

[24] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[25] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.