Beyond Efficiency: The Impact of Instance Selection on Semantic Locality in Transformer-based Text Classification

Karina Batista da Silva[1], Davila de Carvalho Meireles[1], Samai Soares Pereira[1], Washington Cunha[2], Liziane Soares[3], Marcos André Gonçalves[1]

[1] Universidade Federal de Minas Gerais (UFMG)
karinabatista12@gmail.com, davila.meireles@gmail.com, samaisoares@gmail.com, mgoncalv@dcc.ufmg.br
[2] Universidade Estadual de Campinas (UNICAMP)
wcunha@unicamp.br
[3] Universidade Federal de Viçosa (UFV)
liziane.soares@ufv.br

Abstract. Transformer-based models achieve high performance in text classification tasks. In parallel, Instance Selection (IS) techniques have been widely adopted to reduce training sets by removing redundant or noisy examples, lowering computational cost while preserving predictive effectiveness. However, little is known about how these techniques affect the semantic organization of the representation space learned by Transformer models. This work investigates the impact of two recent IS methods, E2SC and biO-IS, on the semantic locality structure of BERT representations and on locality properties previously associated with the quality of model explanations. Models were trained on the original AGNews dataset and on reduced training sets produced by each IS method. Semantic locality was characterized using CLS embeddings from the last BERT layer, average neighborhood similarity and class entropy computed over k-Nearest Neighbors (kNN), complemented by a qualitative analysis based on neighborhood word clouds. The results reveal a dissociated effect of Instance Selection on locality: while both methods increase neighborhood similarity, only biO-IS consistently increases the concentration of instances in highly homogeneous regions. These findings indicate that different Instance Selection strategies modify the geometry of the learned representation space in distinct ways, suggesting that their effects extend beyond computational efficiency and may influence locality properties relevant to explanation methods.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Supervised learning by classification.*

Keywords: BERT, explainability, instance selection, semantic locality, text classification

1 Introdução

O avanço dos modelos baseados em *Transformers* transformou o Processamento de Linguagem Natural (PLN), permitindo alcançar elevados níveis de desempenho em tarefas como classificação de texto. Modelos como BERT [Devlin et al. 2019] e RoBERTa [Liu et al. 2019] aprenderam representações contextuais capazes de capturar relações semânticas complexas, tornando-se predominantes em diversas aplicações. Entretanto, seu treinamento demanda elevado custo computacional e resulta em modelos cuja tomada de decisão permanece pouco interpretável.

Nesse contexto, duas linhas de pesquisa têm recebido crescente atenção. A primeira busca reduzir o custo de treinamento desses modelos por meio de técnicas de Seleção de Instâncias (IS), que removem exemplos redundantes ou ruidosos do conjunto de treinamento. Trabalhos recentes [Cunha et al. 2023; Cunha et al. 2025] mostram que a IS reduz significativamente o tamanho dos conjuntos de treinamento e o tempo de processamento, preservando a efetividade dos modelos. Vale ressaltar que esse ganho restringe-se à fase de treinamento: como a IS atua apenas na filtragem do conjunto de

Agradecimentos: Este trabalho foi financiado e apoiado pelo INCT-TILDIAR (# 408490/2024-1), Google, NVIDIA, CIIA-Saúde, Ana Health, SEBRAE, EMBRAPII, FINEP, CAPES, CNPq, FAPEMIG e FAPESP.

treino, sem alterar a arquitetura ou o número de parâmetros do modelo, o custo de inferência permanece inalterado. A segunda concentra-se em Inteligência Artificial Explicável (XAI), desenvolvendo mecanismos capazes de tornar as decisões dos modelos mais compreensíveis.

Entre essas abordagens, métodos de XAI baseados em localidade assumem que instâncias semanticamente próximas compartilham características relevantes e tendem a produzir explicações mais consistentes. Estudos recentes mostram que regiões do espaço representacional caracterizadas por alta similaridade e baixa entropia estão associadas a maior potencial explicativo [Soares et al. 2026]. Apesar disso, permanece em aberto como técnicas de Seleção de Instâncias afetam a organização semântica desse espaço. Embora tradicionalmente utilizadas apenas para reduzir custo computacional, alterações na geometria das representações aprendidas podem influenciar diretamente propriedades de localidade frequentemente exploradas por métodos de explicabilidade.

Até onde sabemos, este é o primeiro trabalho que investiga sistematicamente como métodos de Seleção de Instâncias modificam a topologia semântica do espaço representacional aprendido por modelos *Transformer* e como essas mudanças se refletem em propriedades de localidade frequentemente exploradas por métodos de explicabilidade, apresentando evidência empírica inicial no cenário AGNews–BERT.

Para responder a essa questão, investigamos os impactos dos métodos E2SC e biO-IS sobre a localidade semântica de modelos BERT treinados no conjunto AGNews. A análise utiliza os *embeddings* do *token CLS* da última camada do BERT para caracterizar vizinhanças por meio de métricas de similaridade e entropia baseadas em *k-Nearest Neighbors* (kNN), complementadas por análises qualitativas utilizando nuvens de palavras.[1]

Especificamente, este trabalho busca responder às seguintes perguntas de pesquisa (**PPs**):

—**PP1:** Os conjuntos reduzidos pelos métodos de Seleção de Instâncias considerados estado-da-arte preservam a eficácia dos modelos no cenário avaliado, estabelecendo uma base válida para a análise de localidade?

—**PP2:** A Seleção de Instâncias altera a localidade semântica do espaço representacional, em termos de similaridade e homogeneidade das vizinhanças?

—**PP3:** O efeito sobre a localidade difere entre os métodos E2SC e biO-IS, refletindo os diferentes mecanismos de remoção de redundância e de ruído empregados por cada método?

No cenário avaliado (AGNews com BERT), a Seleção de Instâncias preserva a eficácia com reduções de 25% a 31% do treino (PP1) e torna o espaço mais compacto, porém com vizinhanças menos homogêneas em média (PP2). O efeito difere entre os métodos: o biO-IS concentra mais instâncias em regiões de alta homogeneidade, enquanto o E2SC reduz essa concentração (PP3).

2 Trabalhos Relacionados

Este trabalho situa-se na interseção entre Seleção de Instâncias, explicabilidade baseada em localidade e caracterização de espaços representacionais em *Transformers*.

2.1 Seleção de Instâncias em Classificação de Texto

A Seleção de Instâncias (IS) reduz o conjunto de treinamento removendo exemplos redundantes ou ruidosos, diminuindo o custo computacional sem comprometer a efetividade. A identificação de componentes informativos também possui antecedentes em recuperação de informação, incluindo métodos para estimar a importância de blocos estruturais em páginas Web [Fernandes et al. 2007]. Cunha et al. [Cunha et al. 2023] propuseram o E2SC, baseado em estimativas de confiança de um classificador auxiliar (i.e., *weak-classifier*) para identificar redundância em classificação de texto com *Transformers*. Em seguida, Cunha et al. [Cunha et al. 2025] apresentaram o biO-IS, que combina remoção de redundância e de ruído por duas taxas independentes, com acelerações de até 2,46× sem perdas de efetividade. A redução de conjuntos de treinamento também é estudada sob os rótulos de *data pruning*

[1]Código e notebooks disponíveis em `https://github.com/karinabatista/projeto-PLN-IS-XAI`.

e *coreset selection*, com estratégias que vão da cobertura geométrica do espaço de representação [Sener and Savarese 2018] à pontuação de instâncias pela dinâmica de treinamento [Swayamdipta et al. 2020], combinadas em trabalhos recentes que equilibram dificuldade e diversidade [Maharana et al. 2024]. Diferentemente desses trabalhos, cujo foco é a relação entre taxa de poda e efetividade preditiva, o presente estudo investiga como a seleção altera a organização do espaço representacional resultante.

2.2 Explicabilidade baseada em Localidade

A organização semântica de documentos a partir de suas representações tem sido explorada em diferentes tarefas de PLN, incluindo a identificação de estruturas hierárquicas de tópicos [Viegas et al. 2020]. Na área de Inteligência Artificial Explicável (XAI), abordagens pioneiras baseadas em localidade, como o LIME [Ribeiro et al. 2016], assumem que instâncias semanticamente próximas no espaço vetorial compartilham características e, portanto, fornecem explicações mais consistentes [Bilal et al. 2025]. Soares [Soares et al. 2026] investigou empiricamente o papel da localidade na explicabilidade de tarefas de classificação de texto, caracterizando vizinhanças por meio de métricas de similaridade média e entropia da distribuição de classes, calculadas com o algoritmo *k-Nearest Neighbors* (kNN) sobre *embeddings* do BERT. A autora demonstrou que localidades mais homogêneas (alta similaridade e baixa entropia) geram nuvens de palavras mais compreensíveis e fiéis à decisão do modelo. O presente trabalho adota diretamente a metodologia de caracterização de localidade proposta por Soares [Soares et al. 2026], estendendo-a para avaliar o efeito da Seleção de Instâncias.

2.3 Lacuna e Posicionamento

A literatura mostra, por um lado, que a IS reduz dados preservando efetividade [Cunha et al. 2025] e, por outro, que a homogeneidade da localidade favorece a explicabilidade [Soares et al. 2026]. Permanece em aberto a relação entre essas duas frentes. A hipótese deste trabalho é que a aplicação prévia de métodos de IS altera ativamente a topologia do espaço latente, concentrando instâncias em regiões de alta homogeneidade. Este trabalho aborda essa lacuna, posicionando metódos de IS não apenas como ferramenta de eficiência, mas como um possível mecanismo de influência sobre a estrutura semântica.

3 Metodologia e Projeto Experimental

O desenho metodológico deste estudo caracteriza-se como um experimento comparativo controlado. A adoção integral do arcabouço de caracterização de localidade proposto por Soares [Soares et al. 2026]: operacionaliza o construto *localidade* seguindo o trabalho previamente proposto, tratando a IS como o fator em estudo e preservando a comparabilidade com os resultados de base. O fluxo analítico estrutura-se em três etapas: redução do conjunto de treinamento, *fine-tuning* das representações e caracterização topológica das vizinhanças com extração de explicações; mais detalhes a seguir.

3.1 Dataset

Os experimentos foram conduzidos utilizando o *dataset* AGNews [Zhang et al. 2015], composto por notícias distribuídas em quatro categorias temáticas: *World, Sports, Business* e *Sci/Tech*. A escolha é coerente com o desenho controlado do estudo: por ser o mesmo conjunto adotado por Soares [Soares et al. 2026], de onde derivam as representações de base, seu uso preserva a comparabilidade direta da caracterização de localidade; suas quatro classes bem definidas e seus textos de comprimento moderado favorecem ainda a legibilidade das nuvens de palavras. Foram utilizados as mesmas partições previamente definidos pelo *benchmark*, disponibilizadas pelos autores [Soares et al. 2026], contendo conjuntos de treinamento, validação e teste. Mais especificamente, AGNews conta com 91872 instâncias de treinamento, 25520 de validação e 10208 de teste; os mesmos conjuntos de validação e teste foram mantidos em todos os experimentos, garantindo comparações sobre as mesmas instâncias.

3.2 Seleção de Instâncias

Os métodos de IS foram empregados para reduzir o conjunto de treinamento pela remoção de exemplos redundantes e potencialmente ruidosos. Foram avaliados dois métodos recentes da literatura: E2SC e biO-IS, ambos desenvolvidos especificamente para tarefas de Classificação Automática de Texto e

considerados, atualmente, o estado-da-arte.

Antes da aplicação dos algoritmos de seleção, os documentos foram representados utilizando *TF-IDF* (*Term Frequency-Inverse Document Frequency*). Usou-se o `TfidfVectorizer` do *Scikit-Learn* [Pedregosa et al. 2011] com *stopwords* em inglês, unigramas (*n-gram range* (1,1)), normalização L2, *smooth idf* e *sublinear tf* ativados e limite de 50.000 atributos, seguindo a metodologia presente em [Cunha et al. 2025; Zanotto et al. 2021; Cunha et al. 2020].

O método E2SC (*Effective, Efficient and Scalable Confidence-Based Instance Selection*) identifica e remove instâncias redundantes. O algoritmo utiliza estimativas de confiança produzidas por um classificador auxiliar (*weak classifier*) para selecionar exemplos representativos e reduzir a redundância presente no conjunto de treinamento. Já o método biO-IS (*Noise-Oriented and Redundancy-Aware Instance Selection*) realiza a remoção conjunta de instâncias redundantes e exemplos potencialmente ruidosos. O método utiliza duas taxas independentes de redução, uma para redundância e outra para ruído, produzindo conjuntos de treinamento mais compactos e semanticamente consistentes.

Em ambos os métodos, o classificador auxiliar utilizado para estimar as medidas de confiança foi uma Regressão Logística (*Logistic Regression*) com os parâmetros padrão da biblioteca *Scikit-Learn*.

Os hiperparâmetros adotados em ambos os métodos correspondem aos melhores valores reportados por [Cunha et al. 2025]. Para o E2SC, foi aplicada uma redução de redundância de $\beta = 25\%$. Já no conjunto biO-IS, foram aplicadas uma redução de redundância de $\beta = 25\%$ e uma redução de ruído de $\gamma = 50\%$. A avaliação conjunta dos dois métodos é deliberada e responde diretamente à PP3: contrastar a remoção isolada de redundância (E2SC) com a remoção combinada de redundância e ruído (biO-IS) permite identificar qual mecanismo de seleção altera a organização latente das instâncias. As taxas de redução resultantes são reportadas na Tabela I (Seção 4.1).

3.3 Treinamento e Caracterização da Localidade

Os três cenários (*baseline*, E2SC e biO-IS) usaram como base o modelo pré-treinado `bert-base-uncased` da *Hugging Face Transformers* [Wolf et al. 2020], com *fine-tuning* para a classificação do AGNews. Optou-se pelo BERT [Devlin et al. 2019] em vez do RoBERTa [Liu et al. 2019] para manter a comparabilidade direta com o trabalho de base [Soares et al. 2026], do qual derivam as representações e a metodologia de caracterização de localidade, com eficácia comparável entre os dois modelos, dada a alta separabilidade das representações do BERT [de Andrade et al. 2023; Cunha et al. 2021]. O treinamento (*batch size* 32/64, *learning rate* $2{\times}10^{-5}$, 2 épocas, 64 tokens, FP16, AdamW) usou PyTorch com a biblioteca *Transformers*, com a validação monitorando a perda; os três modelos foram avaliados sobre o mesmo teste, preservando a comparabilidade. Como representação de cada documento, adotou-se o *embedding* do *token* `[CLS]` da última camada (dimensão 768), que o BERT emprega como representação agregada da sequência para a classificação, preservando a ordem das instâncias entre cenários.

A caracterização da localidade segue integralmente Soares [Soares et al. 2026], decisão que mantém o instrumento de medida idêntico ao trabalho de base e garante que as diferenças observadas reflitam o efeito da IS. Para cada instância de teste, recuperam-se seus $k = 30$ vizinhos por kNN com distância do cosseno, valor de k empiricamente selecionado por otimizar a efetividade da classificação por kNN em partição de validação [Soares et al. 2026], ajustando o kNN sobre o treino de cada cenário e consultando o teste completo. Cada vizinhança é descrita pela entropia de Shannon da distribuição de classes (homogeneidade) e pela distância média aos vizinhos (similaridade). Ambas as métricas são normalizadas (*min–max*) por cenário, exceto na Tabela I, que reporta as médias em escala original. Das mesmas vizinhanças, extraem-se os 25 termos mais frequentes para a análise qualitativa por nuvens de palavras.

4 Resultados e Análises

Os resultados são apresentados em três dimensões: eficácia e eficiência (Seção 4.1), impacto na localidade semântica (Seção 4.2) e análise qualitativa via nuvens de palavras (Seção 4.3).

Cenário	Redução	Acurácia	Macro-F1	Entropia	Distância
Baseline	–	0,944	0,945	0,041	0,070
E2SC	25,0%	0,960	0,962	0,141	0,061
biO-IS	31,4%	0,956	0,957	0,083	0,046

Tabela I: Eficácia, redução do treino e médias de localidade (escala original) nos três cenários.

Faixa de entropia	Baseline	E2SC	biO-IS
Muito baixa [0,0–0,1)	91,9	71,3	81,4
Baixa [0,1–0,3)	4,9	19,7	12,8
Média [0,3–0,5)	1,7	6,5	3,7
Alta [0,5–0,7)	1,4	2,2	1,8
Muito alta [0,7–1,0]	0,1	0,3	0,2
Região alta homog. (%)	65,4	50,2	**70,5**

Tabela II: Distribuição das instâncias de teste por faixa de entropia (%); última linha: proporção em alta homogeneidade.

4.1 Eficácia e Eficiência

Esta seção responde à PP1. Diferentemente das demais, esta pergunta não busca um achado inédito, a preservação de eficácia pela IS já está estabelecida na literatura [Cunha et al. 2023], mas verifica que os conjuntos reduzidos aqui produzidos reproduzem esse comportamento, condição necessária para que as diferenças de localidade observadas nas seções seguintes sejam atribuíveis à seleção e não a uma degradação dos modelos. A Tabela I consolida eficácia (acurácia e Macro-F1 no teste) e eficiência (redução do treino). Os modelos com IS mantiveram desempenho equivalente ou superior ao *baseline* apesar da redução de 25 a 31%, corroborando a literatura [Cunha et al. 2025]: a remoção de redundância e ruído não compromete a eficácia. O ganho de eficiência correspondente, com *speedup* de até 2,46×, é documentado por Cunha et al. [Cunha et al. 2025]. Não se comparam aqui os métodos de IS a amostragens aleatórias de mesmo tamanho: E2SC e biO-IS são métodos estado-da-arte cuja superioridade sobre seleção aleatória já foi estabelecida em avaliação extensiva [Cunha et al. 2023; Cunha et al. 2025].

4.2 Impacto na Localidade Semântica

Esta seção responde às PP2 e PP3. Inicialmente analisam-se as métricas globais de localidade, seguidas da distribuição das instâncias por níveis de entropia, da concentração na região de alta homogeneidade e, por fim, do efeito da IS sobre diferentes tipos de instância.

4.2.1 *Métricas globais da localidade.* As duas últimas colunas da Tabela I apresentam as médias globais de entropia e distância em escala original, única exceção à normalização adotada no restante do trabalho. Ambos os métodos reduzem a distância média (maior similaridade), mas aumentam a entropia média (vizinhanças mais heterogêneas). O biO-IS produz a menor distância (0,046), enquanto o E2SC apresenta a maior entropia (0,141), caracterizando um efeito dissociado da IS sobre as duas dimensões da localidade. Esse resultado é coerente com os métodos: a IS remove sobretudo redundância intraclasse, então os 30 vizinhos passam a incluir mais exemplos de outras classes (elevando a entropia) enquanto o espaço fica mais compacto (reduzindo a distância), efeito mais acentuado no E2SC, voltado só à redundância.

4.2.2 *Distribuição das instâncias por faixa de entropia.* As médias não revelam como as instâncias se distribuem ao longo da escala de entropia. A Tabela II mostra que o *baseline* concentra 91,9% das instâncias na faixa de entropia muito baixa, percentual que cai para 81,4% no biO-IS e 71,3% no E2SC, indicando um deslocamento para faixas intermediárias de entropia, mais acentuado no E2SC.

4.2.3 *Concentração em alta homogeneidade.* A Figura 1 apresenta os mapas de calor da homogeneidade (entropia *versus* distância) das instâncias corretamente classificadas (HIT). Cada célula traz a quantidade de instâncias em uma combinação de entropia e distância; a região inferior esquerda (entropia *e* distância $< 0{,}1$) corresponde à máxima homogeneidade. Embora a entropia média aumente, a Figura 1 e a última linha da Tabela II mostram comportamentos distintos entre os métodos: o biO-IS aumenta a concentração em alta homogeneidade (70,5% contra 65,4% do *baseline*), enquanto o E2SC a reduz para 50,2%, respondendo à PP3. Como isso combina baixa entropia *e* baixa distância, o re-

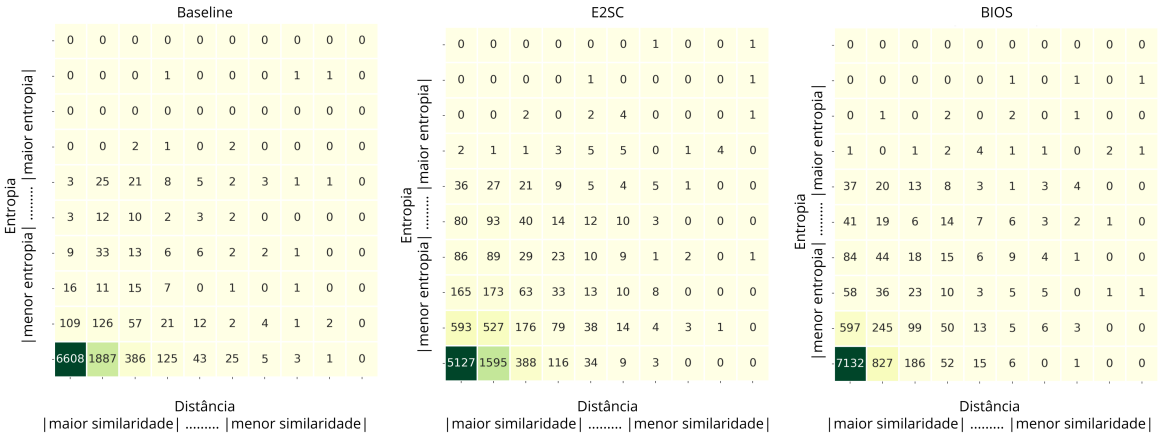

Figura 1: Mapas de homogeneidade (entropia × distância) nos três cenários; apenas instâncias corretamente classificadas.

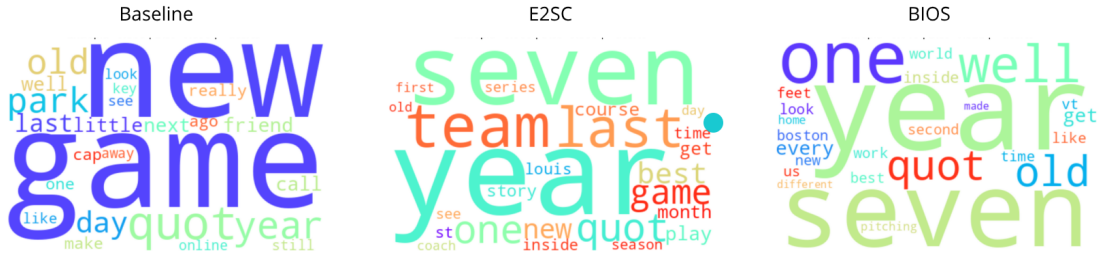

Figura 2: Instância da classe *Sports*, *corretamente* classificada, em região de **baixa** homogeneidade.

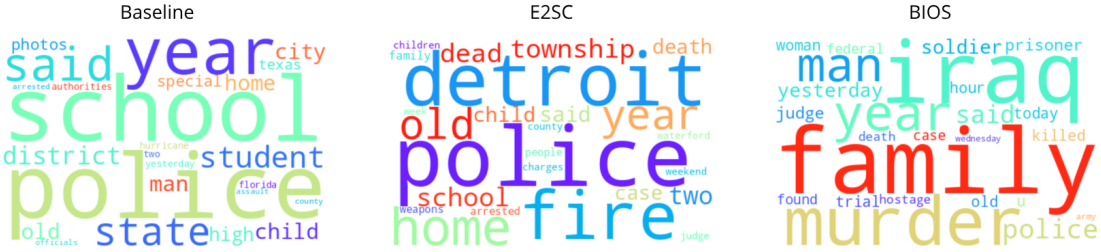

Figura 3: Instância da classe *Sci/Tech*, *incorretamente* classificada, em região de **baixa** homogeneidade.

sultado sugere que apenas o biO-IS aprimora simultaneamente ambas as propriedades, coerentemente com sua remoção conjunta de redundância e ruído.

4.2.4 *Efeito por tipo de instância.* Para entender em quais instâncias a IS atua, analisaram-se as 16 instâncias de Soares [Soares et al. 2026], agrupadas segundo a homogeneidade e o acerto no *baseline*. O efeito concentra-se nas instâncias pouco homogêneas corretamente classificadas (HIT-Low). No E2SC, a entropia da instância da classe *Sports* (cujas nuvens serão apresentadas na Figura 2) caiu de 0,808 para 0,000, e a de *Business*, de 0,671 para 0,184; no biO-IS, as reduções foram mais moderadas (0,646 e 0,368, respectivamente). Em contraste, as instâncias já homogêneas (HIT-High) mantiveram entropia nula nos três cenários. Esses resultados indicam que a IS atua principalmente sobre localidades inicialmente mais heterogêneas (regiões ruidosas), enquanto pouco altera vizinhanças já homogêneas.

4.3 Análise Qualitativa: Nuvens de Palavras

Para avaliar qualitativamente o impacto na explicabilidade, geraram-se nuvens de palavras a partir dos 30 vizinhos de cada instância, extraindo os 25 termos mais frequentes do texto original dos documentos vizinhos, após remoção de *stopwords*, conforme [Soares et al. 2026]. As Figuras 2 e 3 ilustram dois casos exemplos; os demais comportamentos são resumidos no texto.

4.3.1 *Homogeneização da localidade.* A Figura 2 apresenta as nuvens de palavras de uma instância da classe *Sports*, classificada *corretamente*. Na nuvem do *baseline* (entropia 0,808) predominam termos genéricos (*new, game, year, day*), sem foco temático claro. Após o E2SC (entropia 0,000), surgem termos característicos do domínio esportivo (*team, season, play, series, coach*), indicando que a IS reorganizou a vizinhança em torno de exemplos semanticamente mais consistentes. Esse é o comportamento previsto na hipótese, indicando que a IS homogeneizou a localidade semântica da instância.

4.3.2 *Limite do método.* A Figura 3 apresenta as nuvens de palavras de uma instância da classe *Sci/Tech* (Ciência e Tecnologia), classificada *incorretamente* nos três cenários, e localizada em uma região de baixa homogeneidade (no *baseline*). Embora a entropia tenha diminuído de 1,000 (*baseline*) para 0,703 (E2SC) e 0,306 (biO-IS), a IS, sobretudo o biO-IS, homogeneíza a vizinhança, mas em torno de um tema (*family, murder, police, killed*) alheio à classe real. Esse resultado mostra que a homogeneização da vizinhança não corrige instâncias mal posicionadas no espaço representacional, apenas torna a região mais coesa, delimitando o alcance do método.

5 Conclusão

Este trabalho investigou como métodos de Seleção de Instâncias influenciam a organização semântica do espaço representacional aprendido por modelos BERT no conjunto AGNews, analisando seus efeitos sobre propriedades de localidade previamente associadas ao potencial explicativo de modelos de classificação de texto. Para isso, comparamos três cenários (*baseline*, E2SC e biO-IS) utilizando métricas de eficácia, similaridade, entropia e análises qualitativas baseadas em nuvens de palavras.

Os resultados respondem às três perguntas propostas. A IS preserva a eficácia do modelo com redução de 25 a 31% do treino (**PP1**) e modifica a estrutura local do espaço representacional, aumentando a similaridade entre vizinhos, mas com efeitos distintos sobre a homogeneidade (**PP2**). O biO-IS aumenta a concentração em regiões de alta homogeneidade (70,5% contra 65,4% do *baseline*), enquanto o E2SC a reduz (50,2%), sugerindo que a remoção conjunta de redundância e ruído favorece uma organização semântica mais consistente (**PP3**).

Mais amplamente, nossos resultados sugerem que técnicas de IS não devem ser vistas apenas como mecanismos para redução do custo de treinamento. Ao modificar a geometria do espaço representacional aprendido pelos modelos, elas também influenciam propriedades de localidade frequentemente exploradas por métodos de Inteligência Artificial Explicável. Esses achados indicam que decisões de pré-processamento podem afetar não apenas a eficiência do treinamento, mas também o potencial explicativo das representações aprendidas, aproximando duas linhas de pesquisa tradicionalmente tratadas de forma independente: Seleção de Instâncias e IA Explicável.

Embora este trabalho não avalie diretamente um método de explicação, as alterações de localidade observadas têm implicações previsíveis para eles. Métodos baseados em perturbação local, como o LIME [Ribeiro et al. 2016], constroem explicações a partir de amostras na vizinhança da instância; vizinhanças mais compactas e homogêneas tendem a produzir modelos substitutos locais mais estáveis entre execuções, enquanto vizinhanças heterogêneas misturam evidências de classes distintas. Nesse sentido, o aumento de concentração em alta homogeneidade promovido pelo biO-IS sugere condições mais favoráveis à estabilidade das explicações, ao passo que o deslocamento para faixas intermediárias de entropia observado no E2SC aponta na direção oposta. A Figura 3 mostra, contudo, que maior homogeneidade não implica maior fidelidade: a vizinhança pode tornar-se coesa em torno de um tema alheio à classe real. Verificar empiricamente essas hipóteses exige aplicar métodos de XAI e medir fidelidade e estabilidade diretamente, o que constitui a continuação natural deste trabalho.

Limitações e trabalhos futuros. O estudo restringiu-se a um único conjunto de dados (AGNews) e a uma única arquitetura (BERT), de modo que a generalização dos resultados requer investigação adicional. Além disso, a análise qualitativa foi conduzida sobre as 16 instâncias propostas por Soares [Soares et al. 2026], não abrangendo toda a diversidade do conjunto de teste. O *speedup* não foi medido diretamente, uma vez que os modelos foram treinados em ambientes computacionais distintos, sendo reportada apenas a redução do conjunto de treinamento [Cunha et al. 2025]. Finalmente, os resultados foram obtidos a partir de uma única execução por cenário; assim, sua confirmação

estatística depende de múltiplas sementes e de testes de significância. Outra limitação é que avaliamos propriedades de localidade previamente associadas à qualidade percebida pelo usuário, e não a qualidade das explicações em si; os resultados devem ser lidos como evidência de aumento do potencial explicativo do espaço representacional, não como demonstração de melhoria da explicabilidade.

Como trabalhos futuros, pretende-se: (i) estender a avaliação para outros conjuntos de dados e arquiteturas; (ii) investigar diferentes combinações das taxas de redundância (β) e ruído (γ) do biO-IS; (iii) analisar a estabilidade das vizinhanças sob perturbações dos *embeddings* [Rahulamathavan et al. 2025]; e (iv) conduzir estudos com usuários [Soares et al. 2026] para verificar se o aumento da homogeneidade observado traduz-se, de fato, em explicações percebidas como mais úteis e compreensíveis.

REFERENCES

Bilal, A., Ebert, D., and Lin, B. Explainable Artificial Intelligence for Text Classification: A Survey. *arXiv preprint* 1 (1): 1–30, 2025.

Cunha, W., Canuto, S., Viegas, F., Salles, T., Gomes, C., Mangaravite, V., Resende, E., Rosa, T., Gonçalves, M. A., and Rocha, L. Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *Information Processing & Management* 57 (4): 102263, 2020.

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S., Resende, E., Viegas, F., Martins, W. S., Almeida, J. M., et al. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management* 58 (3): 102481, 2021.

Cunha, W., Rocha, L., and Gonçalves, M. A. A Noise-Oriented and Redundancy-Aware Instance Selection Framework. *ACM Transactions on Information Systems* 43 (2): 1–33, 2025.

Cunha, W., Viegas, F., França, C., Rosa, T., Rocha, L., and Gonçalves, M. A. A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. *ACM CSUR*, 2023.

Cunha, W., Viegas, F., França, C., Rosa, T., Rocha, L., and Gonçalves, M. A. An Effective, Efficient, and Scalable Confidence-Based Instance Selection Framework for Transformer-Based Text Classification. In *Proceedings of the 46th International ACM SIGIR*. Taipei, Taiwan, pp. 665–674, 2023.

de Andrade, C. M., Belém, F., Cunha, W., França, C., Viegas, F., Rocha, L., and Gonçalves, M. A. On the class separability of contextual embeddings representations – or "the classifier does not matter when the (text) representation is so good!". *Information Processing & Management* 60 (4): 103336, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the NAACL*. pp. 4171–4186, 2019.

Fernandes, D., de Moura, E. S., Ribeiro-Neto, B. A., da Silva, A. S., and Gonçalves, M. A. Computing block importance for searching on web sites. In *CIKM '07*. pp. 165–174, 2007.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint* 1 (1): 1–13, 2019.

Maharana, A., Yadav, P., and Bansal, M. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. In *International Conference on Learning Representations (ICLR)*, 2024.

Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *JMLR* 12 (1): 2825–2830, 2011.

Rahulamathavan, Y., Farooq, M., and Silva, V. D. PLEX: Perturbation-Free Local Explanations for LLM-Based Text Classification. *arXiv preprint* 1 (1): 1–15, 2025.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why Should I Trust You? Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD*. San Francisco, USA, pp. 1135–1144, 2016.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

Soares, L. S. et al. Are all instances equally explainable? a study on the impact of locality on the explainability of automatic text classification tasks. In *Proceedings of the World Conference on xAI*, 2026.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *EMNLP*. pp. 9275–9293, 2020.

Viegas, F., Cunha, W., Gomes, C., Pereira, A., Rocha, L., and Goncalves, M. CluHTM - semantic hierarchical topic modeling based on CluWords. In *Proceedings of the 58th ACL*. Online, pp. 8138–8150, 2020.

Wolf, T. et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online, pp. 38–45, 2020.

Zanotto, B. S., Beck da Silva Etges, A. P., Dal Bosco, A., Cortes, E. G., Ruschel, R., De Souza, A. C., Andrade, C. M., Viegas, F., Luiz, W., et al. Stroke outcome measurements from electronic medical records: cross-sectional study on the effectiveness of neural and nonneural classifiers. *JMIR Med. Inf.* 9 (11): e29120, 2021.

Zhang, X., Zhao, J., and LeCun, Y. Character-Level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*. Montreal, Canada, pp. 649–657, 2015.