The Impact of Text Representations on Classification Complexity and Predictive Performance

Filipe Araújo Dantas[1], **Eraylson Galdino da Silva**[1]

[1] Escola Politécnica de Pernambuco
Universidade de Pernambuco - UPE – Recife, PE – Brasil
fad@ecomp.poli.br, egs@ecomp.poli.br

Abstract. *This work investigates the impact of text representations on data complexity and classification performance. Five representation techniques (TF-IDF, Word2Vec, BERT, SBERT, and E5) and five classifiers were evaluated on three Portuguese text classification datasets from distinct domains. Beyond traditional performance metrics, complexity measures based on Instance Hardness were employed to analyze the quality of the generated feature spaces. The results show a consistent relationship between lower complexity and higher predictive performance. TF-IDF achieved the best overall results in most scenarios, while the effectiveness of contextual embeddings varied according to the characteristics of the dataset. The findings indicate that complexity metrics provide valuable insights for understanding and evaluating text representations.*

1. Introdução

O Processamento de Linguagem Natural (PLN) busca transformar a linguagem humana em representações estruturadas que possam ser processadas computacionalmente. Nesse contexto, a classificação textual constitui uma das tarefas mais relevantes, sendo utilizada para identificar categorias, tópicos, sentimentos, intenções e outros padrões presentes em documentos. Sua aplicação abrange diversos domínios, incluindo saúde [Nam et al. 2023], jurídico [Noguti et al. 2020], financeiro [Shen and Zhang 2024], educacional [Almatrafi and Johri 2022] e industrial [Harsha et al. 2022], apoiando processos de análise de informações e tomada de decisão.

A evolução da classificação textual está diretamente relacionada aos avanços nas técnicas de representação de textos. Abordagens tradicionais baseadas em TF-IDF e classificadores lineares, como SVM, alcançaram resultados expressivos, mas dependiam fortemente de pré-processamento e apresentavam limitações na captura de relações semânticas mais complexas [Choi et al. 2026]. Posteriormente, modelos de embeddings densos, como Word2Vec, GloVe e FastText, passaram a representar palavras em espaços vetoriais contínuos, incorporando informações semânticas e sintáticas [Patil et al. 2023]. Mais recentemente, arquiteturas Transformer estabeleceram o estado da arte ao produzir representações contextualizadas capazes de modelar dependências de longo alcance e diferentes significados das palavras conforme o contexto [Dong et al. 2021].

Apesar dos avanços recentes, ainda há uma compreensão limitada sobre como diferentes técnicas de representação vetorial influenciam a estrutura do espaço de características e, consequentemente, o desempenho da classificação textual. Embora modelos baseados em Transformers e LLMs representem o estado da arte, seu impacto na separabilidade das classes e na redução da complexidade das instâncias ainda não está totalmente esclarecido. Além disso, representações contextuais podem apresentar fenômenos como a

anisotropia, caracterizada pela concentração dos vetores em regiões específicas do espaço vetorial, o que pode reduzir o poder discriminativo entre instâncias semanticamente distintas e dificultar a separação entre classes [Ethayarajh 2019].

Dessa forma, permanece em aberto se a evolução de representações estatísticas para modelos contextuais realmente reduz a dureza das instâncias (*Instance Hardness*) e a sobreposição entre classes, ou se os ganhos observados decorrem principalmente da maior capacidade de modelagem dos classificadores empregados.

Diante dessa lacuna, este trabalho propõe uma avaliação complementar às métricas tradicionais de desempenho, incorporando medidas de complexidade baseadas em *Instance Hardness* para analisar o impacto das representações textuais na estrutura do espaço de características. Especificamente, são investigadas as métricas kDN (*k-Disagreeing Neighbors*), Class Likelihood, Margin, Entropy e Dataset Instance Hardness [Smith et al. 2014, Lavi et al. 2026]. O objetivo é comparar representações estatísticas, estáticas e contextuais em diferentes bases de classificação textual, avaliando não apenas seu desempenho preditivo, mas também sua capacidade de reduzir a sobreposição entre classes e a dificuldade intrínseca das instâncias. Em síntese, as principais contribuições deste estudo são:

- Evidenciar empiricamente a existência de uma relação consistente entre métricas de complexidade das instâncias e desempenho preditivo, indicando que espaços vetoriais menos complexos tendem a produzir melhores resultados de classificação;
- Demonstrar que representações textuais distintas geram níveis significativamente diferentes de complexidade intrínseca, sendo esse efeito dependente do domínio e das características da tarefa;
- Mostrar que embeddings mais recentes e sofisticados não garantem, necessariamente, melhor desempenho ou menor complexidade, com abordagens tradicionais como TF-IDF apresentando resultados superiores em diferentes cenários avaliados.

2. Trabalhos Relacionados

A relação entre a complexidade dos dados e o desempenho dos classificadores tem sido amplamente investigada na literatura. Lorena et al. [Lorena et al. 2019] apresentam uma revisão abrangente das medidas de complexidade para problemas de classificação, organizando-as em diferentes categorias e destacando métricas como *Instance Hardness* (IH), *k-Disagreeing Neighbors* (kDN), margem e entropia. Os autores demonstram que essas medidas fornecem informações complementares às métricas tradicionais de desempenho, permitindo caracterizar melhor a dificuldade dos problemas de classificação.

No contexto das representações textuais, Ethayarajh [Ethayarajh 2019] analisa a geometria dos embeddings contextuais produzidos por modelos como BERT, GPT-2 e ELMo. O estudo evidencia que esses modelos apresentam forte anisotropia nas camadas superiores, alterando significativamente a estrutura do espaço vetorial e impactando a interpretação da similaridade semântica entre instâncias.

Mais recentemente, Farhangian et al. [Farhangian et al. 2025] propuseram o framework DRES para detecção de notícias falsas, combinando seleção dinâmica de

representações e classificadores. Utilizando métricas como IH e kDN, os autores demonstraram que a complexidade das instâncias pode auxiliar na escolha da representação textual mais adequada e melhorar o desempenho preditivo.

Embora estudos anteriores tenham investigado tanto a complexidade dos dados quanto as propriedades dos embeddings, ainda são escassos os trabalhos que analisam diretamente como diferentes representações textuais alteram a complexidade intrínseca dos conjuntos de dados e, consequentemente, afetam o desempenho dos classificadores. Além disso, essa investigação permanece pouco explorada para textos em língua portuguesa e para domínios especializados. Nesse contexto, este trabalho avalia diferentes embeddings em bases de classificação textual de distintos domínios, empregando métricas de complexidade como IH, kDN, Margin, Entropy e Class Likelihood para analisar a relação entre representação vetorial, dificuldade das instâncias e desempenho preditivo.

3. Metodologia

Esta seção apresenta a metodologia experimental adotada no estudo, incluindo os conjuntos de dados, as técnicas de representação textual, os classificadores avaliados e as métricas utilizadas para analisar o desempenho preditivo e a complexidade das instâncias.

3.1. Base de Dados

Os experimentos foram conduzidos utilizando a versão em português do conjunto de dados ExtraGLUE [Scheible-Schmitt et al. 2025], considerando especificamente as bases SST-2 e MultiRC. O SST-2 (*Stanford Sentiment Treebank*) é um conjunto de dados para análise de sentimentos, no qual a entrada consiste em uma sentença em linguagem natural e a saída corresponde a um rótulo binário indicando sentimento positivo ou negativo. Esse conjunto é utilizado para avaliar a capacidade dos modelos de capturar informações semânticas relevantes para tarefas de classificação textual [Kumar et al. 2024]. O MultiRC (*Multi-Sentence Reading Comprehension*) é uma base de dados voltado à compreensão de leitura, em que cada exemplo é composto por um contexto textual, uma pergunta e uma resposta candidata. O objetivo é predizer se a resposta é suportada pelo contexto fornecido. Esse conjunto permite avaliar a capacidade dos modelos de integrar informações distribuídas ao longo do texto e realizar inferências contextuais [Atanasova et al. 2022].

Os conjuntos de dados foram obtidos a partir do repositório do *Hugging Face*[1], que disponibiliza apenas os subconjuntos *train* e *validation*, não fornecendo uma partição específica para teste. Dessa forma, os dois subconjuntos foram inicialmente unificados, resultando em 68.172 instâncias para o SST-2 e 32.050 instâncias para o MultiRC. Em seguida, cada conjunto de dados foi reparticionado em três subconjuntos mutuamente exclusivos, destinando-se 70% das instâncias para treinamento, 15% para validação e 15% para teste.

Além dos conjuntos de dados amplamente utilizados na literatura, este trabalho também investiga o desempenho das abordagens em um domínio específico do contexto jurídico. Para isso, foi utilizada uma base de dados composta por 11.251 processos judiciais, classificados quanto ao seu grau de relevância em duas classes (relevante e irrelevante). Os rótulos foram atribuídos por especialistas da Procuradoria-Geral do Estado

[1]https://huggingface.co/datasets/PORTULAN/extraglue

de [Informação anonimizada], incluindo Procuradores e Assessores vinculados ao Núcleo de Sucessões e Doações (NSD). A inclusão dessa base permite avaliar a capacidade de generalização dos modelos em um cenário real de classificação documental, caracterizado por vocabulário especializado e elevada complexidade semântica. De forma análoga aos demais conjuntos de dados, a base foi dividida em três subconjuntos mutuamente exclusivos, destinando-se 70% das instâncias para o treinamento, 15% para validação e 15% para o teste.

3.2. Representação Vetorial

Com o objetivo de analisar o impacto da representação textual na classificação e na complexidade das instâncias, foram selecionadas abordagens que representam diferentes gerações de técnicas de representação vetorial: modelos estatísticos, embeddings distribuídos e representações contextualizadas. Essa seleção permite investigar como o aumento da capacidade semântica e contextual afeta a estrutura do espaço de características.

O TF-IDF foi utilizado como representante das abordagens estatísticas tradicionais baseadas em frequência de termos. Os vetores foram gerados considerando unigramas e bigramas. A técnica foi escolhida por sua ampla utilização como baseline em tarefas de classificação textual e por fornecer uma representação esparsa que não incorpora informações semânticas explícitas [Sparck Jones 1972].

O Word2Vec foi empregado como representante dos embeddings distribuídos estáticos. Diferentemente do TF-IDF, essa abordagem projeta palavras em um espaço vetorial contínuo capaz de capturar relações semânticas por meio de padrões de coocorrência. Sua inclusão permite avaliar o impacto da introdução de informações semânticas sem dependência de contexto [Song et al. 2019].

O BERTimbau [Souza et al. 2020] foi utilizado para representar modelos contextualizados baseados em Transformers. A abordagem gera representações dependentes do contexto em que as palavras aparecem, permitindo capturar relações semânticas e sintáticas mais complexas. Sua escolha se deve ao amplo uso em tarefas de PLN para o português e ao seu papel como referência entre os modelos contextualizados.

O SBERT foi selecionado para avaliar representações contextualizadas otimizadas para similaridade semântica entre textos. Diferentemente do BERT tradicional, o modelo produz diretamente embeddings de sentenças e documentos, permitindo investigar o impacto de espaços vetoriais explicitamente estruturados para aproximação semântica [Reimers and Gurevych 2019].

Por fim, o E5 foi incluído por representar uma geração recente de modelos treinados com aprendizado contrastivo. Seu objetivo é produzir embeddings altamente discriminativos, aproximando textos semanticamente relacionados e afastando conteúdos distintos no espaço vetorial. Dessa forma, possibilita avaliar o efeito de representações otimizadas para separabilidade semântica na complexidade das instâncias e no desempenho dos classificadores [Vaičiukynas et al. 2026].

3.3. Modelos de Classificação

Foram selecionados classificadores amplamente utilizados em tarefas de classificação textual e frequentemente empregados como referência em estudos comparativos

[Das et al. 2025, Dhingra et al. 2022]. A escolha buscou contemplar diferentes famílias de algoritmos, permitindo analisar como distintas representações vetoriais influenciam modelos com características de aprendizagem distintas.

A Regressão Logística e o LinearSVC foram incluídos como representantes dos modelos lineares, tradicionalmente reconhecidos pelo bom desempenho em tarefas de classificação de texto [Hazim and Ata 2024]. O Random Forest e o XGBoost representam métodos baseados em árvores e ensembles, capazes de modelar relações não lineares entre os atributos [Domladovac 2021, Shao et al. 2024]. Por fim, o Multi-Layer Perceptron (MLP) foi selecionado como representante das redes neurais, permitindo avaliar o comportamento de um classificador com maior capacidade de modelagem [MadhuMohan et al. 2024].

Modelos baseados em Transformers não foram considerados, uma vez que realizam simultaneamente o aprendizado da representação textual e da função de classificação. Como o objetivo deste trabalho é analisar o impacto dos embeddings sobre o desempenho preditivo e a complexidade das instâncias, foram avaliados apenas classificadores que operam diretamente sobre representações vetoriais previamente geradas.

Os hiperparâmetros foram ajustados por meio de busca em grade (*Grid Search*) combinada com validação cruzada estratificada. Para Regressão Logística, LinearSVC, Random Forest e XGBoost foi utilizada validação cruzada com $k = 5$ partições, enquanto para o MLP adotou-se $k = 2$ devido ao maior custo computacional de treinamento. O F1-score foi utilizado como critério de seleção dos hiperparâmetros, por apresentar maior robustez em cenários com potencial desbalanceamento entre classes. A acurácia foi empregada como métrica complementar na avaliação dos resultados. Os espaços de busca foram definidos com base em recomendações da literatura e práticas consolidadas para cada algoritmo. A Tabela 1 apresenta os hiperparâmetros e os respectivos valores avaliados.

Tabela 1. Hiperparâmetros para os modelos avaliados

Modelo	**Hiperparâmetros**
Regressão Logística	$C \in \{0.01, 0.1, 1, 10\}$; *solver* $\in \{lbfgs, saga\}$; *max_iter* $\in \{500, 1000\}$
Random Forest	$n_estimators \in \{100, 200\}$; *max_depth* $\in \{None, 10, 20\}$
LinearSVC	$C \in \{0.1, 1, 10\}$; *max_iter* $\in \{1000, 2000\}$
MLP	*hidden_layer_sizes* $\in \{(50), (100), (100, 50)\}$; *max_iter* $\in \{300, 500\}$
XGBoost	$n_estimators \in \{100, 200\}$; *learning_rate* $\in \{0.05, 0.1\}$; *max_depth* $\in \{3, 6\}$

3.4. Métricas de Complexidade das Instâncias

Além das métricas tradicionais de desempenho, foram empregadas medidas de complexidade com o objetivo de analisar como diferentes representações vetoriais influenciam a dificuldade de classificação das instâncias. Para isso, foram utilizadas métricas amplamente adotadas na literatura de *Instance Hardness* e análise de complexidade de classificação.

A métrica *k-Disagreeing Neighbors* (kDN) quantifica o grau de sobreposição local entre as classes por meio da proporção de vizinhos mais próximos que pertencem a classes diferentes da instância analisada [Sha et al. 2022, Smith et al. 2014]. Valores baixos indicam boa separação entre as classes, enquanto valores elevados sugerem regiões

de maior ambiguidade e dificuldade de classificação.

A dificuldade individual das instâncias foi avaliada por meio da métrica *Instance Hardness* (IH), que estima o quão difícil é classificar corretamente uma instância [Smith et al. 2014]. Valores próximos de zero indicam exemplos facilmente classificáveis, enquanto valores próximos de um representam instâncias mais difíceis.

Adicionalmente, foram analisadas medidas derivadas das distribuições de probabilidade preditiva dos classificadores. Para reduzir a influência de modelos individuais, as probabilidades foram previamente alinhadas em um espaço comum de classes e agregadas por meio da média das probabilidades produzidas pelos modelos avaliados.

A métrica *Class Likelihood* (CL) corresponde ao grau de confiança atribuído à classe correta da instância. Valores elevados indicam maior confiança na predição e, consequentemente, menor complexidade.

A *Margin* mede a diferença entre as probabilidades associadas às duas classes mais prováveis. Valores elevados indicam maior separação entre as classes concorrentes e classificações mais confiantes, enquanto valores reduzidos sugerem maior incerteza.

Por fim, a *Entropia* quantifica o grau de dispersão da distribuição preditiva [Lavi et al. 2026]. Valores baixos indicam decisões mais confiantes, enquanto valores elevados refletem maior incerteza do classificador.

Em síntese, menores valores de kDN, IH e Entropia, bem como maiores valores de CL e Margin, indicam espaços de representação mais favoráveis à classificação. Em conjunto, essas métricas permitem avaliar como diferentes embeddings influenciam a separabilidade das classes, a dificuldade das instâncias e a confiança dos classificadores.

4. Resultados e Discussão

Esta seção apresenta os resultados de complexidade e desempenho obtidos pelos embeddings e classificadores avaliados. As Tabelas 2, 3 e 4 apresentam as métricas de complexidade calculadas a partir das probabilidades agregadas dos classificadores no conjunto de teste. Na sequência, são apresentados os resultados de desempenho preditivo das diferentes combinações de embeddings e classificadores, considerando o mesmo conjunto de teste.

Com base na Tabela 2, base STT-2, é possível observar que BERT, E5 e TF-IDF apresentaram menores valores para o IH e níveis de Entropy, e os maiores valores para o Class Likelihood e a Margin quando comparados ao SBERT e ao Word2Vec. Esses resultados indicam que essas representações geraram espaços vetoriais com menor complexidade intrínseca e maior separação entre as classes. Por outro lado, em relação ao SBERT e o Word2Vec, é apresentada uma maior dificuldade na discriminação das classes, produzindo espaços vetoriais menos favoráveis ao processo de classificação.

Tabela 2. Métricas de Complexidade Intrínseca – SST-2 (Média $\pm$ Desvio Padrão)

Métrica	BERT	E5	SBERT	TF-IDF	Word2Vec
Instance Hardness	$0{,}134 \pm 0{,}280$	$0{,}131 \pm 0{,}278$	$0{,}194 \pm 0{,}280$	$0{,}132 \pm 0{,}252$	$0{,}279 \pm 0{,}338$
KDN	$0{,}229 \pm 0{,}264$	$0{,}194 \pm 0{,}255$	$0{,}190 \pm 0{,}264$	$0{,}208 \pm 0{,}269$	$0{,}347 \pm 0{,}281$
Class Likelihood	$0{,}782 \pm 0{,}201$	$0{,}779 \pm 0{,}192$	$0{,}666 \pm 0{,}148$	$0{,}785 \pm 0{,}180$	$0{,}635 \pm 0{,}181$
Margin	$0{,}644 \pm 0{,}255$	$0{,}628 \pm 0{,}256$	$0{,}386 \pm 0{,}221$	$0{,}636 \pm 0{,}222$	$0{,}376 \pm 0{,}249$
Entropy	$0{,}413 \pm 0{,}179$	$0{,}426 \pm 0{,}176$	$0{,}587 \pm 0{,}101$	$0{,}434 \pm 0{,}151$	$0{,}581 \pm 0{,}133$

Tabela 3. Métricas de Complexidade Intrínseca – MultiRC (Média ± Desvio Padrão)

Métrica	BERT	E5	SBERT	TF-IDF	Word2Vec
Instance Hardness	$0{,}467 \pm 0{,}340$	$0{,}473 \pm 0{,}310$	$0{,}442 \pm 0{,}440$	$0{,}438 \pm 0{,}357$	$0{,}449 \pm 0{,}388$
KDN	$0{,}483 \pm 0{,}224$	$0{,}487 \pm 0{,}236$	$0{,}482 \pm 0{,}233$	$0{,}491 \pm 0{,}237$	$0{,}491 \pm 0{,}240$
Class Likelihood	$0{,}511 \pm 0{,}097$	$0{,}507 \pm 0{,}081$	$0{,}509 \pm 0{,}058$	$0{,}530 \pm 0{,}176$	$0{,}516 \pm 0{,}116$
Margin	$0{,}157 \pm 0{,}115$	$0{,}133 \pm 0{,}093$	$0{,}108 \pm 0{,}045$	$0{,}302 \pm 0{,}192$	$0{,}189 \pm 0{,}138$
Entropy	$0{,}674 \pm 0{,}027$	$0{,}680 \pm 0{,}017$	$0{,}686 \pm 0{,}005$	$0{,}626 \pm 0{,}075$	$0{,}665 \pm 0{,}037$

Tabela 4. Métricas de Complexidade Intrínseca – LIA (Média ± Desvio Padrão)

Métrica	BERT	E5	SBERT	TF-IDF	Word2Vec
Instance Hardness	$0{,}355 \pm 0{,}385$	$0{,}364 \pm 0{,}393$	$0{,}447 \pm 0{,}246$	$0{,}255 \pm 0{,}361$	$0{,}326 \pm 0{,}381$
KDN	$0{,}440 \pm 0{,}264$	$0{,}448 \pm 0{,}269$	$0{,}441 \pm 0{,}260$	$0{,}408 \pm 0{,}268$	$0{,}436 \pm 0{,}257$
Class Likelihood	$0{,}566 \pm 0{,}145$	$0{,}561 \pm 0{,}145$	$0{,}524 \pm 0{,}098$	$0{,}671 \pm 0{,}235$	$0{,}587 \pm 0{,}159$
Margin	$0{,}269 \pm 0{,}172$	$0{,}267 \pm 0{,}166$	$0{,}171 \pm 0{,}107$	$0{,}520 \pm 0{,}260$	$0{,}304 \pm 0{,}197$
Entropy	$0{,}640 \pm 0{,}060$	$0{,}642 \pm 0{,}058$	$0{,}673 \pm 0{,}021$	$0{,}503 \pm 0{,}154$	$0{,}624 \pm 0{,}079$

Já a Tabela 3, base MultiRC, observa-se que as diferenças entre os embeddings são menos pronunciadas do que aquelas observadas na SST-2. Todas as representações apresentaram elevados valores de IH e KDN, acompanhados de baixos valores de Margin e altos níveis de Entropy, indicando uma elevada complexidade intrínseca do problema independentemente da técnica de representação adotada. Apesar desse comportamento geral, o TF-IDF apresentou os maiores valores de Class Likelihood e Margin, além do menor valor de Entropy, sugerindo um espaço vetorial ligeiramente mais favorável à separação entre as classes. Em contrapartida, o SBERT apresentou o menor valor de Margin e o maior nível de Entropy, indicando maior dificuldade na discriminação das classes. Em resumo, esses resultados indicam que a complexidade da MultiRC está mais associada às características da própria tarefa do que à escolha do embedding, embora o TF-IDF tenha produzido a representação mais favorável entre as alternativas avaliadas.

Na Tabela 4, base LIA, o TF-IDF apresentou os melhores resultados em todas as métricas de complexidade analisadas, obtendo os menores valores de IH, KDN e Entropy, além dos maiores valores de Class Likelihood e Margin, sugerindo que essa representação produziu um espaço vetorial com menor complexidade intrínseca e maior capacidade de separação entre as classes, ou seja, de representar melhor características discriminativas dos documentos jurídicos do que as outras analisadas, produzindo um espaço vetorial mais adequado ao processo de classificação.

Seguindo para as métricas de desempenho, a comparação foi realizada para cada classificador utilizando diferentes embeddings para as bases SST-2, MultiRC e LIA. As Figuras 1, 2 e 3 apresentam, respectivamente, os melhores valores de F1-Macro, Accuracy e Recall-Macro obtidos para cada combinação de dataset e embedding, com o classificador vencedor indicado entre parênteses em cada célula. Os resultados são visualizados por meio de mapas de calor organizados em formato matricial, nos quais o eixo horizontal representa os cinco embeddings avaliados e o eixo vertical representa os três datasets utilizados. Cada célula da matriz é colorida segundo uma escala de intensidade, proporcional ao valor da métrica correspondente, de modo que tonalidades mais escuras indicam desempenhos mais elevados.

Para o STT-2, observou-se que o TF-IDF apresentou os maiores valores de Accuracy, F1-Macro e Recall-Macro para a maioria dos classificadores avaliados. Os melhores

resultados foram obtidos pelas combinações TF-IDF + Linear SVC (F1-Macro = 0,9085) e TF-IDF + Regressão Logística (F1-Macro = 0,9079). Entre os embeddings densos, os melhores desempenhos foram alcançados pelo E5 + MLP (F1-Macro = 0,8946) e pelo BERT + MLP (F1-Macro = 0,8841). Também foi possível observar que o Word2Vec apresentou os menores resultados para a maior parte dos classificadores, com valores de F1-Macro variando entre 0,6745 e 0,7571. Já o SBERT apresentou desempenho intermediário, obtendo resultados próximos aos de BERT e E5 quando combinado a classificadores lineares, mas com reduções mais expressivas quando utilizado com MLP e XGBoost. De forma geral, os resultados evidenciam diferenças consideráveis de desempenho entre as combinações avaliadas para a base SST-2.

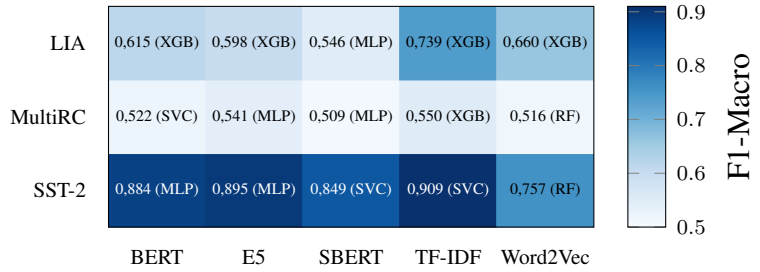


Figura 1. Melhor F1-Macro por dataset e embedding, com o classificador vencedor entre parênteses

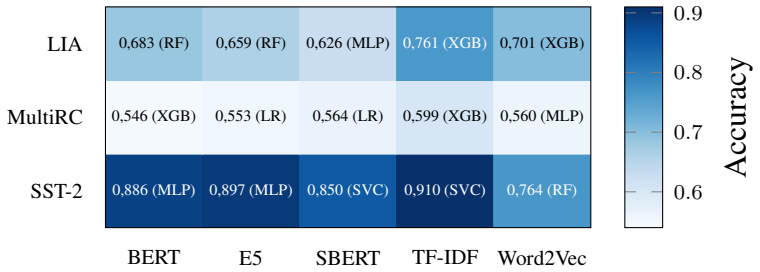


Figura 2. Melhor Accuracy por dataset e embedding, com o classificador vencedor entre parênteses

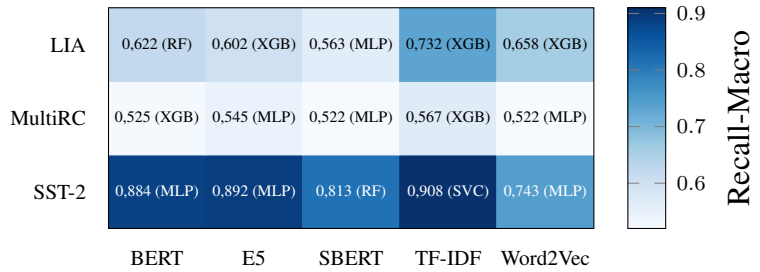


Figura 3. Melhor Recall-Macro por dataset e embedding, com o classificador vencedor entre parênteses

Já para a base MultiRC, os valores de Accuracy, F1-Macro e Recall-Macro foram consideravelmente inferiores aos observados na SST-2 para todas as combinações avali-

adas. O melhor resultado foi obtido pela combinação TF-IDF + XGBoost, alcançando Accuracy de 0,5993, F1-Macro de 0,5495 e Recall-Macro de 0,5673. Além disso, o TF-IDF apresentou os maiores valores de F1-Macro para três dos cinco classificadores avaliados, destacando-se em relação aos demais embeddings. Também é possível observar que BERT e E5 apresentaram desempenhos semelhantes, com valores de F1-Macro concentrados próximos de 0,52. O Word2Vec apresentou resultados intermediários, enquanto o SBERT registrou os menores valores de F1-Macro quando combinado ao Linear SVC, Regressão Logística e Random Forest, apesar de apresentar valores de Accuracy próximos aos demais embeddings. Esse comportamento evidencia diferenças relevantes entre as métricas de avaliação para algumas combinações analisadas na base MultiRC.

Por fim, para a base LIA, o TF-IDF apresentou os maiores valores de Accuracy, F1-Macro e Recall-Macro para todos os classificadores avaliados. O melhor resultado foi obtido pela combinação TF-IDF + XGBoost, alcançando Accuracy de 0,7607, F1-Macro de 0,7388 e Recall-Macro de 0,7322. Além disso, todas as combinações envolvendo TF-IDF apresentaram F1-Macro superior a 0,70, destacando-se em relação aos demais embeddings. Observa-se também que o Word2Vec apresentou o segundo melhor desempenho geral da base, com valores de F1-Macro variando entre 0,6095 e 0,6603. BERT e E5 apresentaram resultados intermediários, enquanto o SBERT registrou os menores desempenhos para a maior parte dos classificadores avaliados, incluindo os menores valores de Accuracy e F1-Macro observados na base. Os resultados também mostram uma maior variação de desempenho entre classificadores para um mesmo embedding quando comparados às bases SST-2 e MultiRC.

Com base nos resultados apresentados, é possível observar que o desempenho preditivo depende da combinação entre o embedding e o classificador adotado. Na SST-2, os melhores resultados foram obtidos pelo TF-IDF combinado ao Linear SVC e pela combinação E5 + MLP. Na MultiRC, o melhor desempenho foi alcançado pelo TF-IDF + XGBoost, enquanto na LIA o TF-IDF apresentou os melhores resultados para todos os classificadores avaliados, com destaque novamente para o XGBoost.

Também é possível observar que não existe um classificador universalmente superior para todas as representações. Enquanto o Linear SVC e a Regressão Logística apresentaram resultados competitivos com TF-IDF, o MLP destacou-se principalmente quando utilizado com embeddings densos, como BERT e E5. Por outro lado, combinações envolvendo SBERT e Word2Vec apresentaram, em geral, os menores desempenhos. Esses resultados mostram que o impacto do classificador está diretamente relacionado às características do espaço vetorial produzido pelo embedding, tornando a escolha conjunta desses componentes um fator determinante para o desempenho final dos modelos.

A partir da análise conjunta das métricas de complexidade intrínseca e dos resultados preditivos, observa-se uma associação consistente entre a qualidade dos espaços vetoriais gerados e o desempenho dos classificadores. De maneira geral, representações associadas a menores valores de Instance Hardness, KDN e Entropy, bem como maiores valores de Class Likelihood e Margin, tenderam a produzir os melhores resultados de Accuracy, F1 e Recall. Esse comportamento foi observado principalmente nas bases SST-2 e LIA, nas quais os embeddings que apresentaram menor complexidade intrínseca também alcançaram os maiores desempenhos preditivos.

A MultiRC apresentou comportamento distinto, caracterizado por elevados níveis de complexidade para todas as representações avaliadas e, consequentemente, desempenhos inferiores aos observados nas demais bases. Esse resultado sugere que, embora a escolha do embedding influencie a qualidade da representação dos dados, as características intrínsecas da tarefa exercem papel determinante na dificuldade do problema. Em conjunto, os resultados indicam que a análise de métricas de complexidade pode fornecer evidências importantes sobre o potencial de desempenho de diferentes representações textuais, permitindo compreender não apenas qual modelo apresenta melhor resultado, mas também os fatores que contribuem para esse comportamento.

5. Conclusão

Este trabalho investigou o impacto de diferentes representações textuais na complexidade dos dados e no desempenho de classificadores supervisionados em tarefas de classificação textual. Os resultados mostraram que existe uma relação consistente entre métricas de complexidade e desempenho preditivo, indicando que espaços vetoriais associados a menores níveis de *Instance Hardness*, *kDN* e *Entropy*, e maiores valores de *Class Likelihood* e *Margin*, tendem a produzir melhores resultados de classificação. Além disso, observou-se que o impacto do embedding varia conforme o domínio e as características da tarefa, não havendo uma combinação universalmente superior entre representação e classificador.

Como limitações, o estudo restringiu-se a três bases de dados e a um conjunto específico de representações e classificadores. Como trabalhos futuros, pretende-se ampliar a avaliação para outros domínios e modelos de linguagem, além de investigar a seleção automática de embeddings e classificadores baseada em métricas de complexidade.

Referências

Almatrafi, O. and Johri, A. (2022). Improving moocs using information from discussion forums: An opinion summarization and suggestion mining approach. *IEEE Access*, 10:15565–15573.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2022). Diagnostics-guided explanation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10445–10453.

Choi, M., Kim, D., and Kim, J. (2026). Defining abusive news categories: Proposing a detection model for digital media integrity. *Applied Sciences*, 16(7):3190.

Das, A. K., Agrahari, S. K., Bhagat, K., and Pati, P. B. (2025). Enhanced analysis of viva dialogues: Classification of questions, answers, and statements with bert and catboost. In *2025 5th Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–7. IEEE.

Dhingra, M., Dhabliya, D., Dubey, M. K., Gupta, A., and Reddy, D. H. (2022). A review on comparison of machine learning algorithms for text classification. In *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, pages 1818–1823. IEEE.

Domladovac, M. (2021). Comparison of neural network with gradient boosted trees, random forest, logistic regression and svm in predicting student achievement. In *2021*

44th International Convention on Information, Communication and Electronic Technology (MIPRO), pages 211–216. IEEE.

Dong, Y., Cordonnier, J.-B., and Loukas, A. (2021). Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pages 2793–2803. PMLR.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 55–65.

Farhangian, F., Ensina, L. A., Cavalcanti, G. D., and Cruz, R. M. (2025). Dres: Fake news detection by dynamic representation and ensemble selection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20034–20052.

Harsha, T. M., Moukthika, G. S., Sai, D. S., Pravallika, M. N. R., Anamalamudi, S., and Enduri, M. (2022). Automated resume screener using natural language processing (nlp). In *2022 6th international conference on trends in electronics and informatics (ICOEI)*, pages 1772–1777. IEEE.

Hazim, L. R. and Ata, O. (2024). Textual authenticity in the ai era: evaluating bert and roberta with logistic regression and neural networks for text classification. In *2024 International Symposium on Electronics and Telecommunications (ISETC)*, pages 1–6. IEEE.

Kumar, A., Sharma, R., and Bedi, P. (2024). Towards optimal nlp solutions: Analyzing gpt and llama-2 models across model scale, dataset size, and task diversity. *Engineering, Technology & Applied Science Research*, 14(3):14219–14224.

Lavi, T., Shapira, B., and Rappoport, N. (2026). Trajectory-based difficulty scoring for reliable learning on tabular data. *arXiv preprint arXiv:2605.24680*.

Lorena, A. C., Garcia, L. P., Lehmann, J., Souto, M. C., and Ho, T. K. (2019). How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, 52(5):1–34.

MadhuMohan, A., Raghu, H., Priyanka, R., Dsouza, S. M., Teena, K., and Shetty, S. (2024). Automated text extraction and classification from images using multi-layer perceptron (mlp). In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, volume 1, pages 1–6. IEEE.

Nam, J., Choi, J.-W., Shin, Y.-G., and Park, S. (2023). A bert-based artificial intelligence to analyze free-text clinical notes for binary classification in papillary thyroid carcinoma recurrence. In *2023 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–2. IEEE.

Noguti, M. Y., Vellasques, E., and Oliveira, L. S. (2020). Legal document classification: An application to law area prediction of petitions to public prosecution service. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Patil, R., Boit, S., Gudivada, V., and Nandigam, J. (2023). A survey of text representation and embedding techniques in nlp. *IEEe Access*, 11:36120–36146.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Scheible-Schmitt, R., He, H., and Mendes, A. B. (2025). Portbert: Navigating the depths of portuguese language models. In *Proceedings of the Workshop on Beyond English: Natural Language Processing for all Languages in an Era of Large Language Models*, pages 59–71.

Sha, L., Raković, M., Das, A., Gašević, D., and Chen, G. (2022). Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Transactions on Learning Technologies*, 15(4):481–492.

Shao, Z., Ahmad, M. N., and Javed, A. (2024). Comparison of random forest and xgboost classifiers using integrated optical and sar features for mapping urban impervious surface. *Remote Sensing*, 16(4):665.

Shen, Y. and Zhang, P. K. (2024). Financial sentiment analysis on news and reports using large language models and finbert. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pages 717–721. IEEE.

Smith, M. R., Martinez, T., and Giraud-Carrier, C. (2014). An instance level analysis of data complexity. *Machine learning*, 95(2):225–256.

Song, S., Wang, Z., Xu, S., Ni, S., and Xiao, J. (2019). A novel text classification approach based on word2vec and textrank keyword extraction. In *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, pages 536–543. IEEE.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Vaičiukynas, E., Danėnas, P., Ablonskis, L., Šukys, A., Dambrauskas, E., Žitkus, V., Butkienė, R., and Butleris, R. (2026). Comparison of modern multilingual text embedding techniques for hate speech detection task. *Applied Sciences*, 16(10):5099.