dialect2vec: Um método baseado em vetores para transcrição dialetal do português a partir de questionários do ALiB

Laila Mota[1], Daniela Barreiro Claro[1]
Eloize R. Marques Seno[2], Rerisson Cavalcante de Araújo[3]
[1]FORMAS Centro de Pesquisa em Dados e Linguagem Natural –
Instituto de Computação – Universidade Federal da Bahia (UFBA) – Salvador - Bahia - Brasil
[2]Instituto Federal de São Paulo – Campus São Carlos – Área de Computação
[3]FORMAS Centro de Pesquisa em Dados e Linguagem Natural –
Instituto de Letras – Universidade Federal da Bahia (UFBA) – Salvador - Bahia - Brasil
{laila.pereira,dclaro}@ufba.br, eloize@ifsp.edu.br, rerissoncavalcante@gmail.com

Resumo

A modelagem da variação dialetal enfrenta desafios quando dependente de modelos de linguagem baseados em sub-palavras, que frequentemente falham ao processar a complexidade de transcrições fonéticas devido a restrições de vocabulário e vieses semânticos. Este trabalho introduz o dialect2vec, um método para capturar a diversidade dialetal do Português Brasileiro. Nossa proposta adota o modelo *token-free* ByT5 para codificar sequências do Alfabeto Fonético Internacional (IPA) ao nível de *byte*, mitigando a perda de informação causada por tokens desconhecidos. Os experimentos foram realizados com dados do Atlas Linguístico do Brasil (ALiB), em que a dimensão fonética isolada demonstrou viabilidade em tarefas de agrupamento não supervisionado, com desempenho próximo do estado da arte léxico (BERTimbau), comprovando que arquiteturas baseadas em *bytes* podem recuperar estruturas dialetais complexas exclusivamente através de pistas fonológicas, oferecendo um mapeamento mais granular das fronteiras linguísticas.

1 Introdução

A representação vetorial de palavras (*word embeddings*) é uma ferramenta fundamental no processamento de linguagem natural, sendo eficaz na captura de relações semânticas e sintáticas entre palavras. Entretanto, existem tarefas linguísticas que dependem da materialidade sonora da língua, como o caso da distinção dialetal. Nesses cenários, modelos focados apenas na semântica podem ser insuficientes, exigindo representações que consigam capturar a similaridade fonética, ou seja, aproximar vetores de palavras que "soam" parecidas, e não apenas aquelas que significam a mesma coisa.

Embora existam avanços na literatura sobre *embeddings* fonéticos, muitas abordagens ainda dependem de tabelas de características articulatórias manuais ou enfrentam dificuldades técnicas ao processar transcrições fonéticas complexas. Abordagens consideradas tradicionais, que tentam processar texto ortográfico e fonético com a mesma arquitetura baseada em sub-palavras, frequentemente sofrem com a perda de informação. Isso ocorre porque símbolos específicos do Alfabeto Fonético Internacional (IPA) podem não existir no vocabulário do modelo, resultando na substituição por *tags* de desconhecido (`[UNK]`) ou na fragmentação excessiva da sequência.

Para enfrentar esse desafio e capturar a complexidade da variação dialetal, este trabalho propõe o dialect2vec, um método capaz de capturar a diversidade dialetal da língua portuguesa. O presente trabalho avança o estado da arte por não tratar a transcrição fonética como texto comum. O método *dialect2vec* integra três dimensões distintas: a léxica (processada pelo BERTimbau), a geolocalização (baseada na localidade do informante do ALiB) e a fonética (baseada na transcrição fonética do ALiB).

A principal contribuição desta abordagem reside no tratamento da dimensão fonética através do modelo ByT5(Xue et al., 2022). Este modelo opera diretamente ao nível de *bytes*, com uma abordagem *token-free*, sendo capaz de processar a sequência bruta de símbolos do IPA sem depender de vocabulários fixos. Isso evita o problema de palavras fora do vocabulário (OOV) e permite que o modelo capture nuances de pronúncia diretamente da sequência de caracteres, preservando a integridade da informação sonora.

Com o intuito de validar o método proposto, utilizamos dados do Projeto Atlas Linguístico do Brasil (ALiB)(Cardoso and Mota, 2014), focando nas variedades dialetais cujas localizações são: Salvador (BA), Três Rios (RJ) e Curitiba (PR). Os experimentos concentraram-se em demonstrar que a representação fonética baseada em *bytes*, mesmo isolada de pistas semânticas ou geográficas, pos-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 800–807

sui robustez suficiente para recuperar estruturas de agrupamento dialetal, oferecendo uma alternativa viável para a compreensão computacional de não apenas "o que" se diz, mas "como" e "onde" se diz. De acordo com o nosso conhecimento, esse é o primeiro trabalho que incorpora a variação dialetal em vetores para a língua portuguesa.

O presente trabalho está organizado como segue: a Seção 2 revisa a literatura sobre as representações vetoriais fonéticas e trabalhos relacionados. A Seção 3 detalha a metodologia proposta, descrevendo a arquitetura *dialect2vec* e a estratégia de modelagem baseada em *bytes*. A Seção 4 apresenta os experimentos utilizando dados do projeto ALiB, seguido da análise e discussão dos resultados. Por fim, a Seção 5 traz as considerações finais e aponta direções para pesquisas futuras.

2 Trabalhos Relacionados

A representação vetorial de palavras é amplamente utilizada na captura de relações semânticas e sintáticas. No entanto, tarefas que utilizam da sonoridade da língua, como a distinção dialetal, exigem representações que capturam a similaridade fonética.

Embeddings fonéticos buscam aproximar vetores de palavras que "soam" de forma semelhante. Silfverberg et al. (2018) demonstraram que representações vetoriais de fonemas, aprendidas via redes neurais recorrentes (RNNs), conseguem capturar analogias fonológicas complexas, evidenciando que há uma estrutura explorável em sequências fonéticas puras.

Mortensen et al. (2016) introduziram o PanPhon, uma base de dados que mapeia segmentos do Alfabeto Fonético Internacional (IPA) para vetores de características articulatórias (como [±sonoro], [±nasal]). Trabalhos subsequentes, como o de Parrish (2017) e Sharma et al. (2021), utilizam essas características para construir *embeddings* que capturam similaridade entre pares de palavras dadas as sequência e características fonéticas.

Recentemente, Zouhar et al. (2024) propuseram o PWESuite, um framework de avaliação para *embeddings* fonéticos, demonstrando que modelos treinados com supervisão de distância articulatória superam abordagens baseadas apenas em caracteres ou fonemas brutos. No entanto, essas abordagens frequentemente dependem de tabelas de *features* pré-definidas, o que pode limitar a generalização para variações dialetais não mapeadas.

Chen et al. (2018) argumentam que palavras semanticamente próximas (ex: "irmão´´ e "irmã´´) podem ser foneticamente distantes, assim como palavras fonéticamente próximas podem ser semanticamente distantes (ex: "pato´´ e "prato´´). Eles propuseram uma arquitetura de dois estágios para Recuperação de Conteúdo Falado, onde *embeddings* fonéticos e semânticos são aprendidos separadamente e depois fundidos. Os autores demonstram que a combinação das duas modalidades permite recuperar documentos tanto pela similaridade sonora quanto pela relevância do tópico.

No contexto de línguas com poucos recursos, Hu et al. (2020) e Yang and Hirschberg (2019) exploraram o treinamento conjunto de *embeddings* acústicos (baseados em áudio) e escritos. Seus trabalhos utilizam métodos para aproximar representações de palavras que compartilham a mesma raiz ou pronúncia.

Adicionalmente, Fang et al. (2020) investigaram o uso de representações fonéticas para aumentar a robustez de modelos de NLP a erros de Reconhecimento Automático de Fala (ASR). Eles concluíram que *embeddings* baseados em fonemas são mais resilientes a ruídos de transcrição do que *embeddings* de caracteres ou palavras, uma vez que erros de ASR tendem a preservar a estrutura fonética da palavra original.

Embora os trabalhos citados avancem na representação fonética, a maioria depende de características articulatórias manuais (PanPhon) ou de arquiteturas complexas de alinhamento áudio-texto. O presente método propõe integrar *embeddings* semânticos (BERTimbau), a geolocalização e a representação fonética aprendida de forma *token-free* com o modelo ByT5 (Xue et al., 2022). Ao processar transcrições IPA em nível de *byte*, nosso método evita a dependência de dicionários de features fixos e o problema de vocabulário desconhecido (OOV), permitindo capturar informações dialetais diretamente da sequência de símbolos fonéticos, juntamente com as transcrições grafemáticas e respectivas geolocalizações.

3 Método dialect2vec

Com o objetivo de capturar a complexidade multidimensional da variação dialetal, o presente trabalho propõe o *dialect2vec*, uma representação vetorial (embedding) semântico-fonético-geográfico. Diferente de abordagens monolíticas, que utilizam uma única arquitetura para processar diferentes tipos de dados (em particular, texto ortográfico e

transcrição fonética) e podem sofrer com a perda de informação no processamento de transcrições fonéticas complexas, o método proposto individualiza o processamento semântico do fonológico, utilizando modelos especializados em cada tipo de dado antes da união vetorial.

A arquitetura processa uma entrada da tripla = $x = (w, p, l)$, onde:

- w representa o item lexical;
- p é a sequência da transcrição fonética segundo a representação IPA;
- l é a localização geográfica do informante do qual a transcrição foi obtida.

Nossa proposta se baseia na hipótese de que a representação final do termo em contextos dialetais tem melhor desempenho por uma função de composição dessas três dimensões. O embedding final resultante da fusão dos vetores extraídos (E_{hybrid}) é definido por:

$$E_{hybrid} = \mathcal{F}(v_w, v_p, v_l) \quad (1)$$

A arquitetura geral e o fluxo do processamento são apresentados na Figura 1 e o detalhamento da extração e modelagem dos componentes do embedding E_{hybrid} estão nas subseções seguintes.

3.1 Dimensão Léxica (w)

A dimensão léxica tem como objetivo capturar o semântico e sintático das palavras transcritas. Para esta dimensão utilizamos o modelo BERTimbau (Souza et al., 2020), um modelo baseado na arquitetura BERT pré-treinado para a língua portuguesa.

O modelo BERTimbau gera *embeddings* contextuais para cada sub-palavra de uma sentença de entrada. Em nossa proposta, para um token alvo t_i, extraímos o seu vetor correspondente v_w. Esta escolha é justificada pela capacidade do modelo em capturar a semântica do português brasileiro em comparação a modelos multilíngues.

$$v_w = \text{Pool}(\text{Encoder}_{BERT}(t_i)) \quad (2)$$

3.2 Dimensão Fonética (p)

Para capturar as variações de pronúncia dos tokens alvo (t_i), utilizamos o modelo ByT5 (Xue et al., 2022). O modelo ByT5 opera diretamente ao nível de *bytes* (caracteres), o que é importante para o processamento de representações fonéticas segundo o padrão do Alfabeto Fonético Internacional (IPA) no qual cada símbolo importa e não existem palavras desconhecidas (*out of vocabulary* - OOV).

A princípio, as transcrições fonéticas no padrão do IPA são processadas pelo encoder do ByT5. Em seguida, utilizamos a média dos estados ocultos da última camada do encoder (*mean pooling*) para obter uma representação vetorial fixa de cada transcrição:

$$v_p = \text{Pool}(\text{Encoder}_{ByT5}(\text{IPA}(t_i))) \quad (3)$$

3.3 Dimensão Geográfica (l)

Na dimensão geográfica, para representar a região de um token alvo (t_i), utilizamos a localidade de origem do informante (l_i). Essa dimensão, diferente das dimensões anteriores, não requer o uso de modelos profundos, sendo assim, a modelagem é feita diretamente através de um vetor esparso utilizando a técnica *One-Hot Encoding*. Assim, a dimensionalidade desta representação está condicionada à quantidade de localidades presentes nos dados.

$$v_l = \text{OneHot}(l_i) \quad (4)$$

3.4 Embedding Híbrido

O embedding híbrido E_{hybrid} consiste na integração das três dimensões (v_w, v_p, v_l) em um único vetor denso. A dimensão do vetor léxico (d_w) é 768, neste trabalho, pois utilizamos um modelo base pré-treinado para língua portuguesa. O vetor fonético possui dimensão (d_p) de 1472, enquanto o vetor geográfico possui dimensão (d_l) de N, sendo N a quantidade total de localidades.

A representação final, então, é obtida a partir da concatenação dos vetores v_w, v_p e v_l. O vetor resultante possui sua dimensionalidade igual à soma das partes ($d_{hybrid} = d_w + d_p + N$) e tem por resultado o embedding definido por:

$$E_{hybrid} = [v_w \oplus v_p \oplus v_l] \quad (5)$$

4 Experimentos e Resultados

Com o intuito de investigar o potencial da abordagem híbrida, *dialect2vec*, proposta em cenários de variação dialetal, estruturamos a apresentação dos experimentos em três etapas. Inicialmente, descrevemos a composição e o tratamento dos dados. Em seguida, detalhamos as configurações experimentais. Por fim, apresentamos uma discussão dos resultados obtidos, comparando o desempenho do

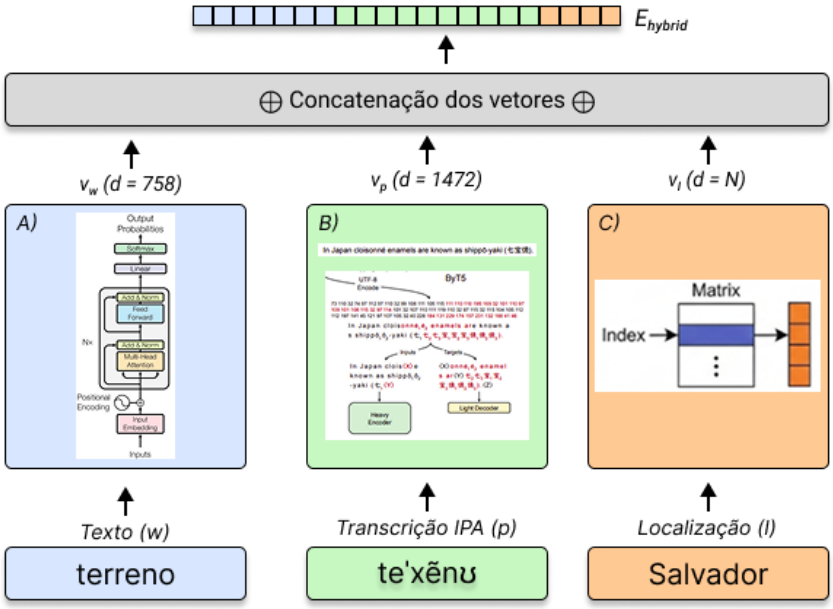

Figura 1: **Visão geral do *dialect2vec*.** O modelo processa três dimensões: léxico (BERTimbau), fonético (ByT5) e geográfico (One-Hot). Os vetores são concatenados para formar o embedding híbrido.

modelo proposto com modelos *baseline* estabelecidos na literatura (mBERT (Devlin et al., 2018) e BERTimbau (Souza et al., 2020)).

4.1 Dados

Os dados utilizados para validar a capacidade do modelo em capturar variações dialetais são provenientes do Projeto Atlas Linguístico do Brasil (ALiB) (Cardoso and Mota, 2014), especificamente, as transcrições das respostas aos questionários fonético-fonológicos de quatro informantes entrevistados para cada uma das localidades aqui analisadas: Salvador (BA), Três Rios (RJ) e Curitiba (PR). Esse conjunto de dados foi transcrito e revisado manualmente por diversos linguistas especialistas, o que garante confiabilidade aos dados já validados por anotadores humanos.

Inicialmente os dados foram organizados em um pipeline de pré-processamento para gerar as entradas (w, p, l) do nosso método. Após a construção dessa base de dados o texto foi normalizado utilizando a Normalização Unicode (NFC), que garante que caracteres similares tenham o mesmo código binário. Em seguida, foram removidas entradas (linhas) cujo conteúdo estivesse duplicado e mantidas apenas entradas em que w ocorresse em todas as N localidades, resultando em um conjunto de dados de 180 observações.

4.2 Configurações de ambiente

Em relação aos experimentos, estes foram conduzidos em ambiente de execução em nuvem (Google Colab), com implementação Python, com suporte das bibliotecas PyTorch e Transformers (Hugging Face) para a manipulação dos modelos de linguagem. Para garantir a reprodutibilidade dos resultados, utilizamos a semente aleatória (*random seed*) em 42 em todas as etapas estocásticas.

Para validar o desempenho da arquitetura proposta, utilizamos os seguintes modelos pré-treinados como *baseline* comparativo:

- mBERT (multilíngue): Utilizado para avaliar o desempenho de um modelo genérico, que compartilha espaço entre várias línguas, não especializado no português, consegue capturar variações dialetais.
- BERTimbau (monolíngue): Estado da arte para o português brasileiro, referência para a dimensão léxica.

Para a dimensão fonética, adotamos o ByT5 *Small*(Xue et al., 2022). A escolha deste modelo é justificada por sua arquitetura *token-free*, baseada em *bytes*, pois permite o processamento integral das transcrições em IPA sem perda de informação, eliminando a substituição de símbolos desconhecidos pela tag `[UNK]`. A Figura 2 evidencia essa vantagem, contrastando a preservação total da informação pelo ByT5 com a fragmentação e perda

de caracteres observada nos modelos baseados em sub-palavras, que dependem de um vocabulário fixo e podem falhar ao processar caracteres fonéticos raros.

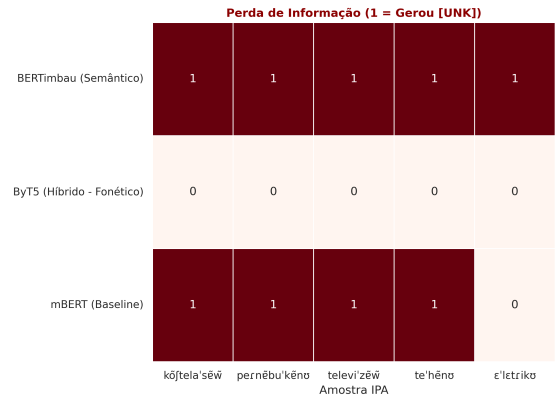

Figura 2: **Exemplos de tokenização de entradas fonéticas.** Demonstra a perda de informação pela presença de [UNK] em oposição à preservação dos caracteres pela codificação em *bytes*.

Embora a arquitetura proposta seja relacionada a concatenação de três dimensões, nos experimentos de agrupamento apresentados focamos na validação da dimensão fonética (p) e removemos explicitamente os vetores v_w e v_l durante os experimentos de agrupamento, para evitar o vazamento de dados e interferência nos resultados. Como o objetivo do experimento é verificar se as características fonéticas são suficientes para distinguir as variedades dialetais, a inclusão das variáveis poderia influenciar erroneamente os resultados, induzindo o algoritmo a agrupar baseando-se apenas no léxico ou rótulo da localidade em vez de aprender os padrões da língua.

Desta forma, como o objetivo é verificar se a representação vetorial induz uma separação natural entre variações dialetais, sem que o modelo seja explicitamente treinado para classificar as regiões, utilizamos dois algoritmos de agrupamento não supervisionado:

1. KMeans: Configurado com $k = N$, induzindo o particionamento em regiões convexas.

2. HDBSCAN (Campello et al., 2013): Baseado em densidade, utilizado para identificar grupos de formato arbitrário e detectar ruído (pontos que não pertencem a nenhuma variedade dialetal).

Na quantificação da qualidade dos agrupamentos gerados, inclusive frente aos rótulos de localidade, adotamos as seguintes métricas:

- *Silhouette Score* (Sil): Avalia a coesão intra-grupo e a separação inter-grupo, medindo o quão próximo um objeto está de seu grupo em relação aos demais grupos. (-1 = provavelmente atribuído ao grupo errado, 0 = limiar entre grupos, +1 = bem agrupado). Essa métrica avalia sem o conhecimento dos rótulos dos grupos.

- *Adjusted Rand Index* (ARI): Mede a similaridade entre o agrupamento predito e a divisão real das localidades, ajustando a pontuação para o acaso (0 = aleatório, 1 = perfeito).

- *F1-Score*: Calculado após o agrupamento para as classes, permitindo uma interpretação do agrupamento como uma tarefa de classificação. Utilizada para medir a capacidade do algoritmo de agrupamento em manter pares de pontos semelhantes no mesmo grupo e separar pontos diferentes.

4.3 Resultados e Discussão

Analisando o comportamento dos modelos na tarefa de agrupamento dialetal, é importante destacar que, para garantir a isonomia do experimento, todos os modelos foram alimentados exclusivamente com as transcrições fonéticas para geração dos vetores (v_p). A avaliação quantitativa, representada na Tabela 1, apresenta informações sobre o desempenho das arquiteturas ao organizar o espaço dialetal. As abordagens utilizando os modelos baseline apresentam resultados robustos, especialmente o modelo mBERT combinado ao algoritmo HDBSCAN, que obteve maior F1-Score (0.863) e Silhueta (0.716). Esses valores sugerem que, embora o mBERT não seja um modelo especializado em transcrições fonéticas, ele consegue projetar os dados em conjuntos densos e bem separados. Entretanto, a alta pontuação de Silhueta pode sugerir que o modelo está agrupando com base em estruturas sintáticas, mais do que nas características dialetais.

O modelo BERTimbau, por outro lado, demonstrou estabilidade em conjunto com o algoritmo KMeans, atingindo o maior ARI do experimento (0.622). O ARI é uma métrica que penaliza acertos ao acaso e mede a concordância estrutural com as classes reais. Isso sugere que a robustez do tokenizador do BERTimbau permite capturar informações na sequência fonética, mesmo operando fora do domínio ortográfico nativo auxiliando na capacidade do modelo em representar a variação regional.

Tabela 1: Avaliação de desempenho dos *embeddings*: Comparativo entre Modelos e Algoritmos.

Modelo	Algoritmo	F1-Score	Precisão	Recall	ARI	Silhueta
mBERT	KMeans	0.667	0.673	0.683	0.486	0.424
	HDBSCAN	**0.863**	**0.884**	**0.861**	0.528	**0.716**
BERTimbau	KMeans	0.799	0.800	0.800	**0.622**	0.698
	HDBSCAN	0.791	0.824	0.767	0.515	0.703
Fonético (ByT5)	KMeans	**0.753**	0.769	**0.750**	**0.602**	**0.354**
	HDBSCAN	0.700	**0.847**	0.650	0.013	0.177

O método *dialect2vec* apresentou um desempenho competitivo, especialmente quando utilizado com o algoritmo KMeans. Com um ARI de 0.602, nosso método se aproxima ao desempenho do BERTimbau (0.622), demonstrando que a dimensão fonética agrega a integridade estrutural e os agrupamentos foram preservados.

É importante ressaltar a diferença nas pontuações de Silhueta, pois os modelos baseados em sub-palavras apresentam valores acima de 0.69, enquanto o modelo *token-free* apresentou uma Silhueta de 0.354. Esse resultado sugere que a dimensão fonética (p) possui espaço vetorial complexo. A variação dialetal real pode ser ruidosa, não formando esferas perfeitas e isoladas. Portanto, uma Silhueta menor pode indicar que a representação *token-free* apresenta um mapa mais realista e granular da língua, em vez de simplificá-la em blocos. As Figuras 3 e 4 apresentam uma demonstração visual dos agrupamentos observados em cada modelo com os algoritmos utilizados.

As Matrizes de Confusão (apresentadas na Figura 5) corroboram os dados da Tabela 1 e sugerem uma clara distinção entre as regiões Curitiba x Salvador/Três Rios, indicando que o espaço vetorial que representa as variações dialetais entre as localidades é complexo.

O experimento demonstra que *dialect2vec* distingue melhor as áreas dialetais visto que inclui as transcrições fonéticas, atingindo níveis de concordância (ARI) próximos ao modelo BERTimbau. A vantagem desta abordagem é a incorporação de cada *byte* do alfabeto fonético, sem a perda de informação ou substituição de caracteres, indicando a viabilidade de modelos na compreensão não apenas do que se diz, mas como se diz e onde se diz.

A análise das métricas de agrupamento, especificamente a diferença entre os valores de Silhueta e o desempenho inferior no algoritmo baseado em densidade (HDBSCAN), sugere que o espaço vetorial fonético gerado é complexo e contínuo, refletindo a natureza das fronteiras dialetais reais, em oposição à simplificação em agrupamentos densos. Isso sugere que a arquitetura proposta desenha um mapa mais granular da língua portuguesa.

5 Considerações Finais e Trabalhos Futuros

Este trabalho apresentou o método *dialect2vec* para a construção de representações vetoriais que capturam a complexidade da variação dialetal. A principal contribuição metodológica está na utilização do modelo *token-free* ByT5 para a dimensão fonética, que se mostrou capaz de processar transcrições em IPA ao nível de *byte*. Essa estratégia eliminou o problema de palavras fora do vocabulário (OOV) e a perda de informação causada pela tokenização por sub-palavras, comum em modelos de tokenização por sub-palavra, garantindo a integridade dos dados fonéticos.

Os experimentos realizados focaram na validação da dimensão fonética (v_p), garantindo a isonomia ao remover informações semânticas e geográficas que pudessem influenciar na avaliação. Os resultados demonstraram que o *dialect2vec* representa os dados dialetais de uma maneira apropriada aos objetivos do estudo. Ao atingir um ARI de 0.602 no algoritmo KMeans, o modelo proposto aproximou-se significativamente do desempenho do estado da arte lexical (BERTimbau), sugerindo que as características estritamente fonético-fonológicas preservadas pelo ByT5 são suficientes para recuperar agrupamentos dialetais.

Como trabalhos futuros, uma avaliação extrínseca do *dialect2vec* (E_{hybrid}) será desenvolvida para tarefas de classificação, tais como classificação de dialetos ou a geração de texto dialetal. Além disso, planeja-se expandir o conjunto de dados para incluir mais localidades do Projeto ALiB e investigar novas estratégias referente à sobreposição de áreas dialetais, ampliando a definição das isoglossas dialetais e transpondo-as para os vetores, tais como pode ser observado entre Salvador e Três Rios.

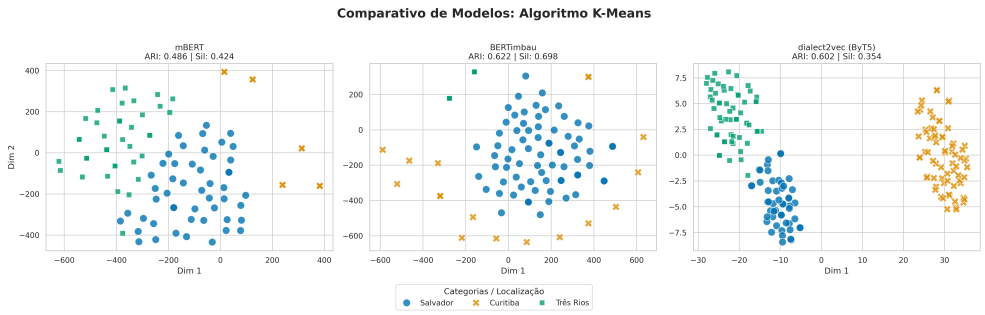

Figura 3: **Visualização dos agrupamentos pelo algoritmo KMeans.** Comparação entre as baselines e o *dialect2vec*.

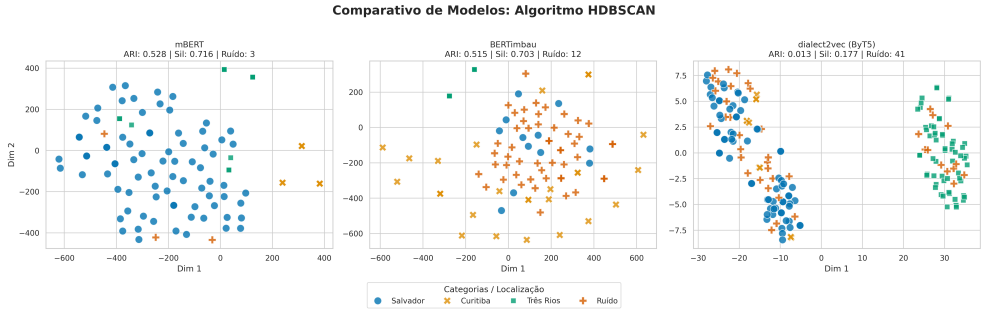

Figura 4: **Visualização dos agrupamentos pelo HDBSCAN.** É possível notar a formação de dois macro-agrupamentos. A alta incidência de pontos de ruído e a fusão das classes explicam o baixo desempenho nas métricas de agrupamento (ARI 0.013).

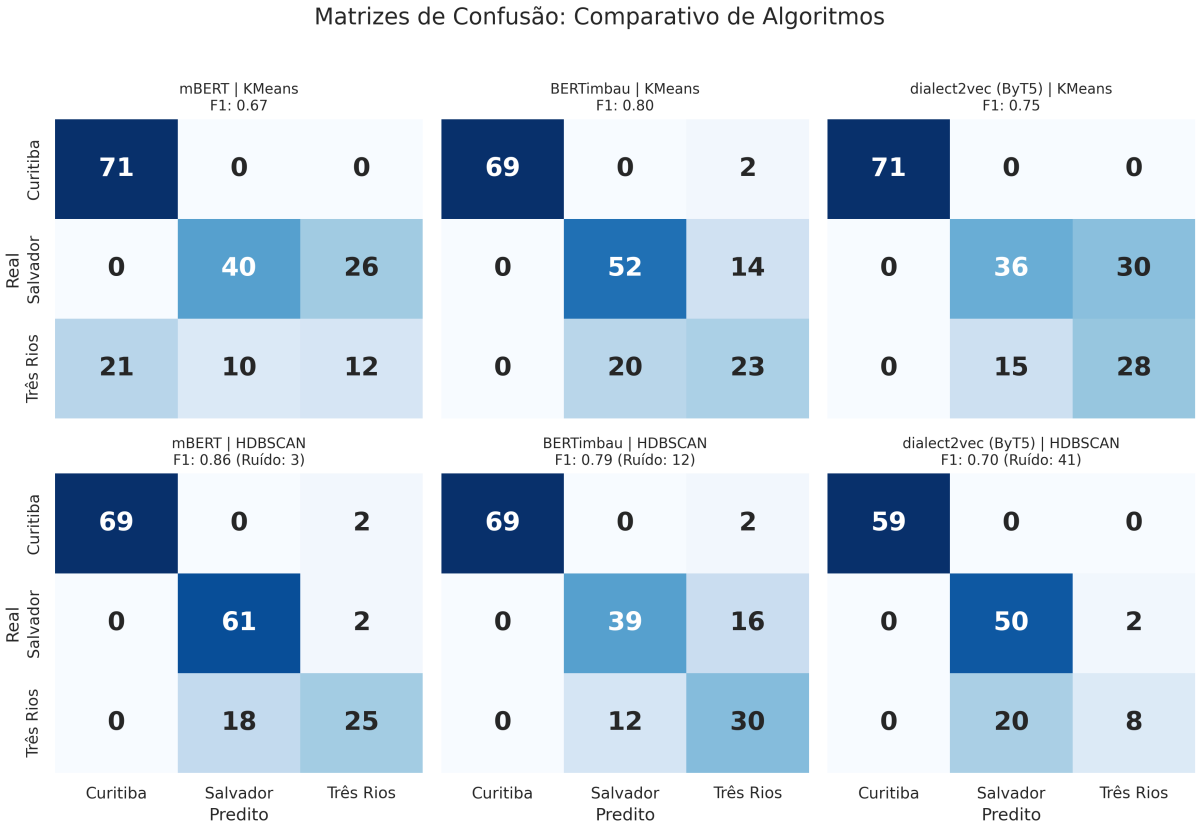

Figura 5: **Matrizes de confusão comparativas entre os modelos.** Destaca-se a capacidade do *dialect2vec* em discriminar a variedade da localidade Curitiba, evidenciando que as pistas fonológicas desta região são preservadas e capturadas pelo método proposto.

6 Limitações

As limitações do *dialect2vec* estão concentradas, principalmente, no recorte do escopo experimental e na granularidade da validação não supervisionada. A aplicação do modelo em um subconjunto restrito do corpus sugere que, em cenários de transição dialetal suave, como o de variedades com maior proximidade fonológica, o espaço vetorial tende a se tornar mais complexo. Essa característica pode ser um desafio para algoritmos de agrupamento baseados em densidade, sugerindo que a distinção entre falares regionais mais sutis pode demandar.

Agradecimentos

Os autores agradecem à FAPESB (084.0508.2025.0000487-58, TIC002/2015 e CCE022/2023), à CAPES, CNPQ e ao INCT-TILDIAR/CNPq (408490/2024-1).

Declaração ética e uso de IA

Durante a preparação deste trabalho, os autores utilizaram o Google Gemini com a finalidade de revisão gramatical e formatação de referências e tabelas, com grau de contribuição mínima. Todo o conteúdo foi revisado e editado pelos autores, que assumem total responsabilidade pelo manuscrito.

References

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Suzana Cardoso and Jacyra Mota. 2014. *Atlas Linguístico do Brasil*. Addison-Wesley Longman Publishing Co., Inc.

Yi-Chen Chen, Sung-Feng Huang, Chia-Hao Shen, Hung-yi Lee, and Lin-shan Lee. 2018. Phonetic-and-semantic embedding of spoken words with applications in spoken content retrieval. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 941–948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Anjie Fang, Simone Filice, Nut Limsopatham, and Oleg Rokhlenko. 2020. Using phoneme representations to build predictive models robust to asr errors. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 699–708, New York, NY, USA. Association for Computing Machinery.

Yushi Hu, Shane Settle, and Karen Livescu. 2020. Multilingual jointly trained acoustic and written word embeddings. In *Interspeech 2020*, pages 1052–1056.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Allison Parrish. 2017. Poetic sound similarity vectors using phonetic features. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 13, pages 99–106.

Rahul Sharma, Kunal Dhawan, and Balakrishna Pailla. 2021. Phonetic word embeddings. *Preprint*, arXiv:2109.14796.

Miikka P. Silfverberg, Lingshuang Mao, and Mans Hulden. 2018. Sound analogies with phoneme embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 136–144.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Preprint*, arXiv:2105.13626.

Zixiaofan Yang and Julia Hirschberg. 2019. Linguistically-informed training of acoustic word embeddings for low-resource languages. In *Interspeech 2019*, pages 2678–2682.

Vilém Zouhar, Kalvin Chang, Chenxuan Cui, Nate B. Carlson, Nathaniel Romney Robinson, Mrinmaya Sachan, and David R. Mortensen. 2024. PWESuite: Phonetic word embeddings and tasks they facilitate. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13344–13355, Torino, Italia. ELRA and ICCL.