Fine-tuned model evaluation on Transformer Fragments for Identifying Idiomatic Expressions in Portuguese

Ricardo Gomes de Oliveira[1], Laila Pereira Mota Santos[1], Lílian Teixeira de Sousa[2], Marcos Adriano Pereira dos Santos[1], Daniela Barreiro Claro[1], Rerisson Cavalcante de Araújo[2]

[1]FORMAS - Centro de Pesquisa em Dados e Linguagem Natural
Instituto de Computação – Universidade Federal da Bahia (UFBA)
Salvador – BA – Brasil

[2]FORMAS - Centro de Pesquisa em Dados e Linguagem Natural
Instituto de Letras – Universidade Federal da Bahia (UFBA)
Salvador – BA – Brasil

{gomesricardo, laila.pereira, lilian.sousa, marcosaps,

dclaro,rerisson.cavalcante}@ufba.br

Abstract. *This work addresses the challenge of identifying Idiomatic Expressions (IEs) in Portuguese, a problem whose main challenge lies in the semantic non-compositionality and ambiguity of these structures. The scarcity of annotated data and the limitations of the models in capturing idiomaticity motivated the construction of an annotated corpus and the proposal of a method based on the use of Transformer fragments to identify IEs in sentences. The method uses attention weights from the BERTimbau model, focusing on a specific head sensitive to relevant syntactic relations in IEs, and integrates linguistic heuristics to penalize literal uses. The results demonstrate high precision (1.0) with no false positives, and a recall of 66.7%, resulting in an F1 score of 0.8. Furthermore, the work compares the results with methods already used in the literature using other fine-tuned BERT architecture models.*

Resumo. *O presente trabalho aborda o desafio da identificação de Expressões Idiomáticas (EIs) em língua portuguesa, um problema cujo principal desafio está na não composicionalidade semântica e ambiguidade dessas estruturas. A escassez de dados anotados e limitações dos modelos em capturar a idiomaticidade motivaram a construção de um corpus anotado e a proposta de um método que se baseia no uso de fragmentos de Transformer para identificação das EIs em sentenças. O método utiliza pesos de atenção do modelo BERTimbau, focando em uma cabeça específica sensível a relações sintáticas relevantes em EIs e integra heurísticas linguísticas para penalizar usos literais. Os resultados demonstram alta precisão do método (1.0) sem falsos positivos, e uma revocação de 66,7%, resultado em uma pontuação de F1 de 0.8. Além disso o trabalho compara os resultados com métodos já utilizados na literatura de outros modelos de arquitetura BERT ajustados.*

1. Introdução

As Expressões Multipalavras (EM), particularmente as Expressões Idiomáticas (EI), são um desafio recorrente na área de Processamento de Linguagem Natural (PLN). A difi-

culdade reside principalmente na não composicionalidade semântica dessas estruturas, uma vez que o significado da expressão não deriva da combinação dos seus constituintes, mas também na ambiguidade, uma expressão pode adotar um sentido literal ou figurado a depender do contexto.

Em linguística, assume-se, muitas vezes, que a interpretação de expressões complexas é determinada pelos significados dos seus constituintes e pelo modo como estão combinados, o que é conhecido como Princípio da Composicionalidade. Expressões idiomáticas, no entanto, fazem parte de um conjunto de fenômenos linguísticos em que o significado de uma expressão não é propriamente composicional. Uma definição bastante recorrente é a de "uma lexia complexa indecomponível, conotativa e cristalizada em um idioma pela tradição cultural" [Xatara 2001].

Segundo a literatura especializada [Tagnin 2013, Barreto et al. 2018], expressões idiomáticas são caracterizadas por estrutura sintática, complexidade semântica e composicionalidade. Do ponto de vista sintático, expressões idiomáticas apresentam sempre algum grau de cristalização, com restrições de flexão e em relação à inserção de elementos. Nesses casos, a alteração pode levar à perda do sentido figurado, como pode ser observado em "João bateu as botas uma contra a outra". Neste caso, "bater as botas"não é interpretado como morrer. As Expressões Idiomáticas correspondem a unidades semânticas e as palavras perdem seu sentido independente quando combinadas na expressão. Por outro lado, em alguns casos as expressões podem ser ambíguas entre uma interpretação idiomática e literal (ex. "arregaçar as mangas", "lavar as mãos", "abrir os olhos" etc.), o que está relacionado ao grau de idiomaticidade.

A identificação automatizada das expressões idiomáticas observa o contexto prévio para auxiliar no reconhecimento de uma expressão literal ou figurada. Além dos elementos sintáticos e lexicais, o conhecimento de mundo, aspectos culturais e pragmáticos, também influenciam no processamento deste tipo de expressão.

Como abordagens para o processamento de EM, estudos focaram na distinção entre o potencial de uma EM ser idiomática (classificação a nível de tipo) [Cook et al. 2007, Cordeiro et al. 2016], em sua real utilização idiomática em uma sentença (classificação a nível de token) [King and Cook 2018, Rohanian et al. 2020, Hashempour and Villavicencio 2020, Zeng and Bhat 2021, Tayyar Madabushi et al. 2021, Tayyar Madabushi et al. 2022] ou ambas [Garcia et al. 2021].

A utilização de modelos distribucionais com embeddings contextualizados, como o caso de modelos como BERT, trouxeram outras possibilidades para a desambiguação de EM em contexto. O trabalho de [Hashempour and Villavicencio 2020] demonstrou que os embeddings contextualizados, quando aplicado o "Princípio do Idioma", que trata a EM como um token único no treinamento, foi capaz de organizar as EM em agrupamentos semanticamente distintos (literal/idiomático) no espaço semântico.

Apesar dos avanços em termos de embeddings contextualizados, alguns trabalhos questionam a profundidade com que esses modelos distribucionais compreendem a idiomaticidade. Autores em [Garcia et al. 2021] desenvolveram o dataset NCTTI e concluíram que as representações geradas pelos modelos à época não capturavam adequadamente a variabilidade da idiomaticidade em diferentes sentenças, quando comparados

com anotadores humanos. Utilizando outra abordagem, [Zeng and Bhat 2021] propuseram o modelo DISC, endereçando a questão de detecção de EI não vistas, por meio de uma arquitetura neural que utiliza um fluxo de atenção que une informações lexicais e contextuais, avaliando a compatibilidade semântica de uma EI com seu contexto.

Assim como em outras áreas do PLN, a escassez de dados, principalmente anotados, para EM é um problema a ser superado. Autores em [Tayyar Madabushi et al. 2021] apresentam um corpus para avaliar modelos em configurações zero-shot, one-shot e few-shot, sobre o qual demonstraram que, embora os modelos pré-treinados tenham dificuldade no cenário zero-shot, apresentam um bom desempenho em cenários one-shot ou few-shot. Essa observação também é observada em outros trabalhos que exploraram métodos de aprendizado eficientes em amostras para a detecção e representação de EM em contextos de poucos dados [Phelps et al. 2022]. A avaliação no SemEval-2022 Task 2 confirmou a eficácia dessas abordagens, embora destaque desafios no desempenho em línguas que não o inglês [Tayyar Madabushi et al. 2022].

Nesse contexto, o presente trabalho se insere na tarefa de detecção e classificação de EM por meio de representações contextualizadas, analisando o uso dos modelos BERT fine-tunados e os fragmentos de *transformers*. Este trabalho difere dos demais visto que utiliza estratégias no *transformers* para análise das expressões idiomáticas, quais sejam fine-tuning e os fragmentos de *transformers*. Estes últimos visam modelar explicitamente as sentenças com EM em sentido literal e idiomático a partir de um corpus anotado, aplicando a penalização ou reforço na atenção do modelo para identificação da EI. Os resultados alcançados são promissores, visto que o método proposto demonstra capacidade significativa na identificação de expressões idiomáticas, especialmente em língua portuguesa. Além disso, os achados oportunizam o desenvolvimento de novas pesquisas aplicadas às tarefas de tradução automática, análise semântica e ensino de línguas assistido por tecnologia.

O presente artigo está organizado em seções, como segue: Seção 2 descreve o método proposto. Seção 3 descreve os experimentos realizados para identificar as expressões idiomáticas. Seção 4 descreve os resultados alcançados de acordo com as métricas definidas, e, por fim as Considerações Finais na Seção 5

2. Detecção Automática de Expressões Idiomáticas em PT-BR

O método proposto para detecção automática de expressões idiomáticas em português brasileiro baseia-se no modelo BERTimbau[Souza et al. 2020] e no fine-tuning de modelos de arquitetura BERT [Devlin et al. 2019] para o português. O método percorre sentenças em busca de sequências de palavras que correspondam às expressões idiomáticas prévias, avaliando em seguida se essas ocorrências representam uso literal ou figurado por meio de duas abordagens: i) Modelos ajustados via fine-tuning de modelos pré-treinados e ii) aprendizado de máquina com heurísticas linguísticas, conforme descrito na Figura 1.

2.1. Modelos Ajustados

O modelo ajustado da arquitetura BERT pré-treinado para a tarefa de detecção de idiomaticidade teve duas configurações: *one-shot*, em que o ajuste fino contém um exemplo de EM idiomático e um exemplo composicional; *few-shot*, contendo entre 6 e 10 exemplos

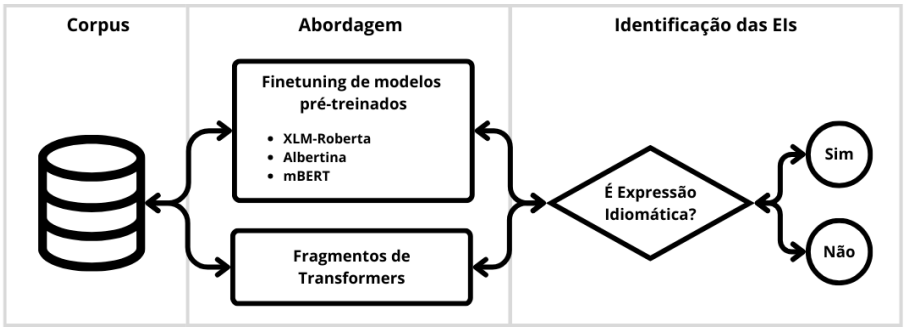


Figura 1. Arquitetura do modelo de Identificação de Expressões Idiomáticas.

por EM, sendo metade idiomáticos e metade composicionais, de acordo com a disponibilidade do corpus.

Os modelos selecionados para ajuste e comparação foram: XLM-RoBERTa [Conneau et al. 2019], Albertina PT-BR [Rodrigues et al. 2023], e mBERT [Pires et al. 2019]. Inicialmente foram adotados os mesmos valores dos hiperparâmetros de [Tayyar Madabushi et al. 2021] Task 1, subtask A: 9 épocas, *seed* de [1,5], *batch* de 32, taxa de aprendizado de $2e^{-5}$ e sem a utilização de *dropout* e *weight decay*.

2.2. Fragmentos de Transformer

A abordagem baseada *Fragmentos de Transformer* utiliza de forma direcionada os pesos de atenção de uma única cabeça do modelo BERTimbau, selecionada com base em sua capacidade de capturar relações sintáticas relevantes. Em vez de explorar toda a arquitetura, concentra-se nos fragmentos mais informativos do mecanismo de *self-attention*, buscando maximizar interpretabilidade e eficiência. Esse uso seletivo se apoia em evidências de que diferentes cabeças de atenção apresentam especializações funcionais distintas [Clark et al. 2019], o que permite extrair sinais linguísticos úteis mesmo a partir de um subconjunto do modelo.

A partir de experimentos prévios foi identificado que a cabeça 3 da segunda camada demonstra sensibilidade específica a dependências gramaticais lineares, especialmente entre verbos e seus objetos diretos — uma configuração frequentemente presente em expressões idiomáticas do tipo verbo + objeto. Por exemplo, na sentença *"Ela chutou o balde"*, observou-se atenção significativamente elevada entre os tokens *"chutou"* e *"o balde"* nessa cabeça.

Desta forma, para este trabalho limitamos a análise ao subespaço gerado por esta núcleo de atenção. Para cada expressão candidata s, calculamos a média dos pesos de atenção mútuos entre todos os pares de tokens em s, denotada por $\text{att}(s)$. Esse valor funciona como um indício de coesão interna da expressão, sob a hipótese de que tokens fortemente interconectados nessa cabeça formam uma unidade semântica coesa — potencialmente uma EI.

A escolha por uma cabeça de atenção intermediária se justifica teoricamente: camadas inferiores tendem a capturar relações sintáticas locais, enquanto camadas superiores refletem abstrações semânticas mais amplas [Tenney et al. 2019]. No caso das expressões idiomáticas, que muitas vezes preservam estrutura gramatical comum mas implicam significados não-composicionais, essa atenção sintática localizada mostra-se eficaz.

Além disso, foi integrada ao modelo uma penalização baseada em heurísticas de literalidade. A função $\text{lit}(s)$ avalia o contexto da expressão s em busca de indícios de uso literal. São utilizados padrões léxico-sintáticos comuns, como a presença de objetos concretos ou ações físicas explícitas. Por exemplo, a sentença *"O garoto puxou o saco de batatas"* apresenta elementos que sugerem interpretação literal, e nesse caso $\text{lit}(s)$ retorna um valor elevado, que reduz o *score* final da expressão, evitando os falsos positivos.

Esse componente heurístico cobre desde padrões regulares de uso literal (por meio de expressões regulares) até verificações baseadas em coocorrência com substantivos concretos e predicados físicos. Com isso, o modelo equilibra indícios estatísticos internos (atenção) com conhecimento linguístico explícito, resultando em uma classificação mais robusta e interpretável.

2.3. Métrica de Avaliação

Durante o fine-tuning, o modelo foi ajustado para atribuir um escore de idiomaticidade a potenciais EMs em cada sentença, indicando a probabilidade de a sequência de palavras corresponder a um idiomatismo não-composicional em vez de uma combinação literal de palavras. Esse ajuste fino foi orientado por aprendizado supervisionado: sentenças contendo expressões idiomáticas conhecidas foram fornecidas ao modelo com sinalização das posições das EMs, enquanto sentenças sem idiomatismos (ou com usos estritamente literais das palavras) atuaram como exemplos negativos.

O *score* EM calculado com base nas saídas do modelo ajustado foi definido para calibrar a decisão final. Esse *score* combina informações de atenção interna do BERTimbau com penalizações ou reforços heurísticos. Os parâmetros de controle foram definidos empiricamente: Os parâmetros principais do modelo foram definidos, empiricamente, como $\alpha = 0.1$, $boost = 0.10$ e $\tau = 0.75$. O parâmetro α controla o peso da penalização por uso literal no cálculo do escore final, enquanto $boost$ representa um reforço aditivo aplicado as sequências de palavras que coincidem exatamente com expressões idiomáticas da lista de referência. O limiar τ, por sua vez, define o valor mínimo necessário para que uma sequência candidata seja aceita como expressão idiomática detectada.

Com base nesses parâmetros, o escore final atribuído a uma sequência candidata s é descrito na Equação 1.

$$\texttt{score}(s) = \texttt{att}(s) + boost \cdot I[\texttt{match}(s)] - \alpha \cdot \texttt{lit}(s) \tag{1}$$

onde $\texttt{att}(s)$ representa uma medida derivada da atenção do modelo para os tokens em s, $I[\texttt{match}(s)]$ é uma função indicadora que vale 1 se s corresponde exatamente a uma expressão da lista de referência (e 0 caso contrário), e $\texttt{lit}(s)$ é uma função de literalidade que indica o grau de evidência de uso literal. Caso $\texttt{score}(s) \geq \tau$, a sequência s é marcada como uma expressão idiomática detectada; caso contrário, é ignorada.

3. Experimentos

3.1. Corpus e Anotação de Referência

O conjunto experimental utilizado neste estudo é composto por 405 sentenças em português brasileiro, extraídas de dois corpora: o CETENFolha[1] e o Carolina[2]

[1]Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo. Disponível em: `https://www.linguateca.pt/cetenfolha/index_info.html`

[2]Versão 1.2, taxonomia dat.

[Crespo et al. 2023]. As sentenças foram compiladas manualmente com o objetivo de representar uma ampla gama de contextos sintáticos e discursivos, contemplando tanto registros formais quanto informais da linguagem.

Para selecionar quais das EMs coletadas seriam utilizadas neste trabalho, foram definidos os seguintes critérios:

1. A sequência de palavras que forma a EM precisa ser potencialmente ambígua, isto é, deve possuir significados tanto idiomático quanto composicional.
2. As EMs precisam ser encontradas nos corpora em sentenças tanto na forma idiomática quanto composicional.
3. A EM deve ter no mínimo 50 ocorrências no corpus. Esse critério foi adotado devido à desproporcionalidade entre ocorrências idiomáticas e composicionais nos corpora. Durante as buscas das sentenças, almejou-se encontrar exatamente 10 ocorrências idiomáticas e 10 ocorrências composicionais para cada EM, nos casos em que não foi possível, foi usado a quantidade encontrada com um mínimo de 2 ocorrências de cada tipo.

No corpus, então, cada sentença contém, no mínimo, uma ocorrência de uma expressão idiomática previamente conhecida e registrada em uma lista de referência composta por 129 construções, elaborada por linguistas. A anotação manual foi realizada por especialistas em linguística computacional, que verificaram se as expressões estavam sendo empregadas em sentido idiomático ou literal, construindo assim um *gold standard* de referência para a avaliação dos métodos automáticos. Esse corpus anotado constitui a base para os experimentos de detecção e classificação das expressões idiomáticas, permitindo comparar as saídas do método com os julgamentos humanos em situações reais de uso da linguagem.

3.2. Configuração Experimental

O modelo utilizado na abordagem com fragmentos de *transformers* foi o BERTimbau *base* [Souza et al. 2020], com extração explícita dos pesos de atenção da camada 2, cabeça 3.

As expressões candidatas foram geradas via *sliding window* sobre a sequência de tokens, com tamanho n variando de 2 a 3 palavras consecutivas. Para cada expressão s, computou-se o vetor médio de embedding contextualizado ($\vec{s}$) e a média dos pesos de atenção mútuos entre os tokens que a compõem, denotada por $\texttt{att}(s)$. O vetor médio da sentença foi usado como referência para cálculo de similaridade cosseno, $\texttt{sim}(s)$. O escore final foi definido como:

$$\texttt{score}(s) = \alpha \cdot \texttt{att}(s) + (1 - \alpha) \cdot \texttt{sim}(s) \tag{2}$$

com $\alpha = 0,30$. Adicionalmente, o escore de s recebe incremento de $+\texttt{boost} = 0,20$ se a sequência coincide lexicalmente com uma expressão da lista de referência. Penalizações de $-0,10$ são aplicadas caso s seja identificado como um uso literal. Expressões com $\text{score}(s) \geq \tau = 0,55$ foram consideradas detecções positivas.

Uma função específica foi implementada para detecção de contextos que indicam uso literal de uma expressão. A heurística verifica:

- padrões linguísticos definidos via expressões regulares (e.g., “quebrou o galho da árvore”), associados a objetos concretos;

- presença de entidades factuais (produto, localização, data), identificadas via análise morfossintática com `spaCy`;
- ausência de verbos subjetivos e predominância de substantivos concretos, sugerindo ação literal e não figurada.

3.3. Critérios de Avaliação

A avaliação foi realizada considerando Verdadeiros Positivos (TP), EIs esperadas e corretamente detectadas; Falsos Positivos (FP), EIs detectadas que não correspondem a nenhuma entrada na lista de referência ou são usadas literalmente; Falsos Negativos (FN), EIs presentes na lista de referência mas não detectadas automaticamente. A partir desses valores as métricas de desempenho calculadas foram Precisão, Revocação e F1.

$$\text{Precisão} = \frac{TP}{TP + FP}, \quad \text{Revocação} = \frac{TP}{TP + FN}, \quad F_1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

3.4. Exemplo Ilustrativo

Dois exemplos do comportamento do método diante de diferentes contextos sintáticos e semânticos envolvendo expressões idiomáticas candidatas são apresentados a seguir.

> Sentença 1: *Talvez nessa situação as pessoas abrirão os olhos para uma comunidade que se sente oprimida.*

Na Sentença 1, o método identificou corretamente a expressão idiomática "abrirão os olhos", atribuindo-lhe um *score* de $0{,}802$. A detecção foi confirmada com base na presença da expressão na lista de referência, em sua estrutura sintática típica (verbo + objeto direto), e na ausência de padrões indicativos de uso literal. A atenção média entre os tokens "abrirão", "os" e "olhos" foi alta, e a similaridade com o vetor da sentença contribuiu positivamente para o score final.

> Sentença 2: *Quando abri os olhos, percebi que estava deitada no chão e que meu marido estava me sacudindo violentamente para me acordar.*

Nesta sentença 2, apesar de expressão "abri os olhos" também estar presente na lista de referência como potencial EI, nenhuma expressão da sentença foi marcada como idiomática, mesmo que "no chão" tenha recebido um score de $0,769$, e "os olhos" tenha alcançado $0,683$. O módulo de filtragem de literalidade identificou forte sinal de uso concreto e físico (e.g., "deitada no chão", "me sacudindo", "me acordar"), levando à exclusão da hipótese idiomática. Assim, a ausência de interpretação figurada foi corretamente inferida, demonstrando o funcionamento eficaz do componente de heurística contextual. As Figuras 2(a) e 2(b) descrevem as sentenças e os reconhecimentos pelo método proposto.

4. Resultados

Os resultados obtidos demonstram a eficácia do método proposto para a detecção automática de expressões idiomáticas em português brasileiro. O experimento foi conduzido sobre um conjunto de 405 sentenças, contendo 261 sentenças contendo expressões idiomáticas, anotadas previamente por especialistas, servindo como referência para a avaliação automática, conforme Tabela 1.

```
Avaliando sentença: Quando abri os olhos, percebi que estava deitada no chão e que meu marido estava me sacudindo violentamente para me acordar
Camada 2, Cabeça 3, Alpha=0.30, Boost=0.20

      expressão base              | média | freq | POS                | Confiável  | Lista ext. | Var
-------------------------------------------------------------------------------------------------
(2,3) no chão                     | 0.769 |  1   | ADP NOUN           |    Não     |    Não     | no chão
(2,3) que estava                  | 0.747 |  1   | SCONJ AUX          |    Não     |    Não     | que estava
(2,3) estava me                   | 0.715 |  1   | AUX PRON           |    Não     |    Não     | estava me
(2,3) os olhos,                   | 0.683 |  1   | DET NOUN PUNCT     |    Não     |    Não     | os olhos,
(2,3) que estava deitada          | 0.671 |  1   | SCONJ AUX ADJ      |    Não     |    Não     | que estava deitada
(2,3) me acordar                  | 0.671 |  1   | PRON VERB          |    Não     |    Não     | me acordar
(2,3) estava me sacudindo         | 0.656 |  1   | AUX PRON VERB      |    Não     |    Não     | estava me sacudindo
(2,3) meu marido                  | 0.651 |  1   | DET NOUN           |    Não     |    Não     | meu marido
(2,3) no chão e                   | 0.651 |  1   | ADP NOUN CCONJ     |    Não     |    Não     | no chão e
(2,3) marido estava me            | 0.647 |  1   | NOUN AUX PRON      |    Não     |    Não     | marido estava me

Nenhuma Expressão Idiomática identificada
```

(a) Exemplo de expressão não identificada, corretamente.

```
Avaliando sentença: Talvez nessa situação as pessoas abrirão os olhos para uma comunidade que se sente oprimida
Camada 2, Cabeça 3, Alpha=0.30, Boost=0.20

      expressão base              | média | freq | POS                | Confiável  | Lista ext. | Var
-------------------------------------------------------------------------------------------------
(2,3) se sente                    | 0.823 |  1   | PRON VERB          |    Não     |    Não     | se sente
(2,3) abrirão os olhos            | 0.802 |  1   | VERB DET NOUN      |    Sim     |    Sim     | abrirão os olhos
(2,3) os olhos                    | 0.788 |  1   | DET NOUN           |    Não     |    Não     | os olhos
(2,3) as pessoas                  | 0.782 |  1   | DET NOUN           |    Não     |    Não     | as pessoas
(2,3) nessa situação              | 0.763 |  1   | ADP NOUN           |    Não     |    Não     | nessa situação
(2,3) que se sente                | 0.739 |  1   | SCONJ PRON VERB    |    Não     |    Não     | que se sente
(2,3) uma comunidade              | 0.732 |  1   | DET NOUN           |    Não     |    Não     | uma comunidade
(2,3) que se                      | 0.723 |  1   | SCONJ SCONJ        |    Não     |    Não     | que se
(2,3) para uma comunidade         | 0.686 |  1   | ADP DET NOUN       |    Não     |    Não     | para uma comunidade
(2,3) situação as pessoas         | 0.663 |  1   | NOUN DET NOUN      |    Não     |    Não     | situação as pessoas

Expressão Idiomática mais provável: abrirão os olhos
```

(b) Exemplo de expressão identificada, corretamente.

Figura 2. Comparação entre exemplos com identificação correta e incorreta de expressões idiomáticas.

Os modelos fine-tunados obtiveram uma maior revocação, elevando a medida F1 para 92%. Por outro lado, a aplicação do modelo BERTimbau *base*, com extração de atenção na camada 2, cabeça 3, e parâmetros $\alpha = 0,30$, $\texttt{boost} = 0,20$, penalização por literalidade de $-0,10$ e limiar de decisão $\tau = 0,55$, resultou nos seguintes indicadores quantitativos apresentados na Tabela 1.

Esses números evidenciam que o fragmento do *transformer* observa que todas as detecções realizadas corresponderam às expressões idiomáticas válidas no contexto, sem falsos positivos. Isso comprova a robustez da estratégia combinada baseada em atenção, similaridade e penalização de uso literal, que atuou como filtro eficaz contra ambiguidades lexicais. O comparativo com o método de ajuste dos modelos pré-treinados apresentados na Tabela 1 mostra a capacidade em detecção das EIs, em relação a técnicas já conhecidas.

Por outro lado, a revocação de $66,7\%$ indica que aproximadamente um terço das expressões idiomáticas esperadas não foi identificado. A análise qualitativa dos falsos negativos revelou as seguintes causas possíveis:

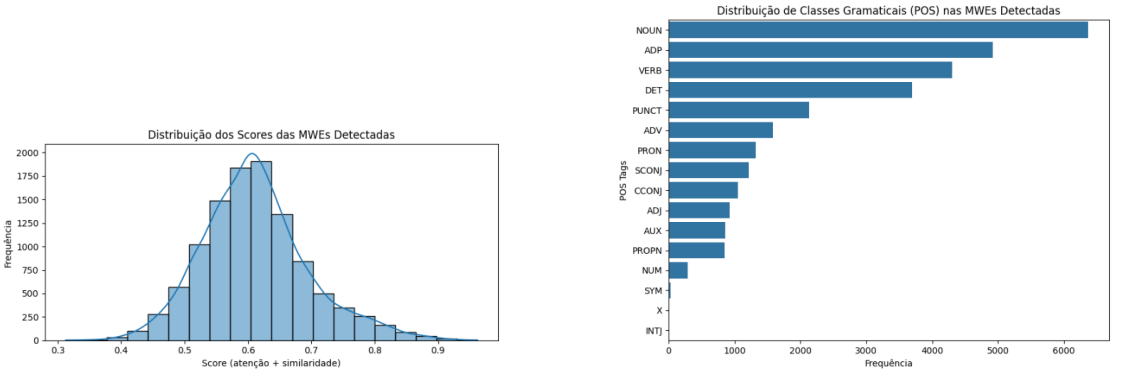
1. variações sintáticas (como inserções adverbiais ou troca de ordem canônica);
2. formas morfológicas pouco frequentes;
3. fragmentações estruturais que dificultaram a coesão atencional entre os tokens da EM.

A Figura 3(a) mostra a distribuição dos scores atribuídos às expressões candidatas. Observa-se uma separação clara entre candidatos fortes (score $\geq 0{,}75$) e fracos (score $< 0{,}50$), reforçando empiricamente a adequação do limiar $\tau = 0{,}55$ adotado.

A análise morfossintática revelou que cerca de 80% das EMs detectadas pertencem ao padrão de sintagma verbal (verbo + objeto). O desempenho do modelo foi notavel-

Método	Modelo	Precisão	Revocação	F1	TP	FP	FN
One-shot	XLM-Roberta	0.887	0.618	0.729	47	6	29
	mBERT	0.763	0.803	0.782	61	19	15
	Albertina PT-BR	0.899	0.697	0.855	62	7	14
Few-shot	XLM-Roberta	0.953	0.803	0.871	61	3	15
	mBERT	0.873	0.816	0.844	62	9	14
	Albertina PT-BR	0.971	**0.882**	**0.924**	67	2	9
Fragmento de Transformer	BERTimbau	**1.000**	0.667	0.800	86	0	46

Tabela 1. Resumo do desempenho quantitativo do método na detecção de expressões idiomáticas.



(a) Distribuição dos scores atribuídos às expressões candidatas pelo modelo.

(b) Frequência das classes gramaticais (POS tags) nas MWEs detectadas.

Figura 3. Informações de distribuição e frequência das expressões no corpus.

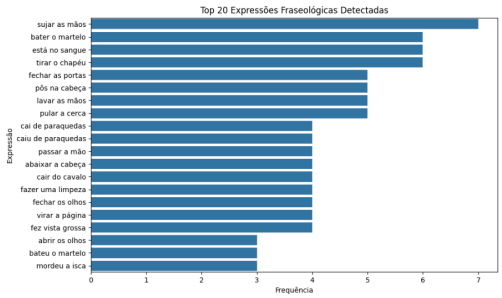
mente superior nesse tipo de construção, em consonância com a especialização da cabeça de atenção selecionada. A Figura 3(b) mostra a distribuição das classes gramaticais (`POS tags`) nas expressões identificadas.

A Figura 4(a) apresenta as expressões mais frequentemente identificadas. Todas pertencem ao conjunto de referência, e são construções altamente consagradas no uso idiomático da língua. Já a Figura 4(b) mostra a dispersão entre atenção média (att(s)) e similaridade (sim(s)), sugerindo que ambas contribuem de forma complementar.

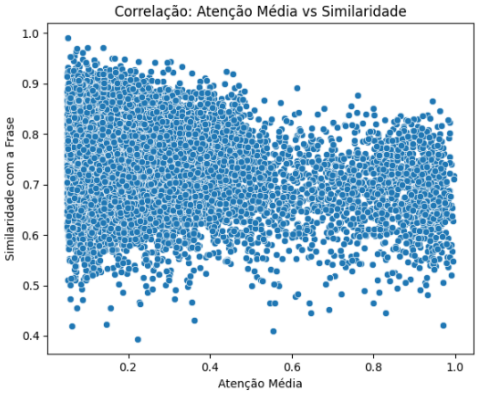
Assim, o método apresentou precisão e alinhamento linguístico. A filtragem heurística de literalidade aliada ao mecanismo de pontuação híbrido (atenção + similaridade) provou-se eficaz para minimizar erros e capturar expressões idiomáticas válidas com alto grau de confiabilidade.

5. Considerações Finais e Trabalhos Futuros

Este trabalho apresentou uma abordagem híbrida para a detecção automática de expressões idiomáticas (EIs) em sentenças do português brasileiro, combinando mecanismos de atenção do modelo BERTimbau com heurísticas linguísticas voltadas à



(a) Top 20 expressões idiomáticas mais frequentemente detectadas pelo método.



(b) Dispersão entre atenção média e similaridade para todas as expressões candidatas.

Figura 4. Análise de frequência e dispersão das expressões idiomáticas identificadas.

identificação de usos figurados. O método proposto fundamenta-se na extração de atenção direcionada (camada 2, cabeça 3), cálculo de similaridade com embeddings da sentença e aplicação de filtros estruturais para penalizar ocorrências literais.

Os resultados experimentais demonstram que a abordagem é precisa, visto que nenhuma expressão foi identificada incorretamente (0% de falsos positivos), o que dá indícios da robustez do componente de filtragem de literalidade. A revocação de $66,7\%$ confirma a capacidade do método em capturar a maioria das expressões idiomáticas de referência, mesmo diante de variações morfossintáticas e em um cenário de dados complexos e heterogêneos. O F_1-score de $0,8$ reflete um equilíbrio satisfatório entre cobertura e precisão, sem comprometer a confiabilidade das detecções.

A análise qualitativa mostrou que os casos de não detecção (FN) concentraram-se em expressões com inserções não canônicas, flexões raras ou estrutura sintática interrompida — fenômenos que desafiam a coerência atencional típica entre os componentes da MWE. Além disso, observou-se forte correlação entre o padrão morfossintático das expressões (majoritariamente locuções verbais) e o desempenho do método, com resultados especialmente satisfatórios para estruturas verbo–objeto, onde os mecanismos de atenção demonstraram maior alinhamento semântico.

Como Trabalhos Futuros, observa-se a possibilidade de expansão do corpus e a utilização de múltiplas cabeças de atenção. Além disso, há a possibilidade de avaliar o método com outras línguas, tais como Ingles, Espanhol e Galego.

Agradecimento

Esse trabalho é parcialmente apoiado pela Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) por meio dos processos TIC 0002/2015, CCE 0022/2023 e INCITE PIE0002/2022.

Referências

[Barreto et al. 2018] Barreto, S. d. O. G., Marcilese, M., and de Oliveira, A. J. A. (2018). Idiomaticidade, familiaridade e informação prévia no processamento de expressões

idiomáticas do pb. *Letras de Hoje*, 53(1):119–129.

[Clark et al. 2019] Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. In *Proceedings of ACL*, pages 311–330.

[Conneau et al. 2019] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

[Cook et al. 2007] Cook, P., Fazly, A., and Stevenson, S. (2007). Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In Gregoire, N., Evert, S., and Kim, S. N., editors, *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.

[Cordeiro et al. 2016] Cordeiro, S., Ramisch, C., Idiart, M., and Villavicencio, A. (2016). Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997, Berlin, Germany. Association for Computational Linguistics.

[Crespo et al. 2023] Crespo, M. C. R. M., de Souza Jeannine Rocha, M. L., Sturzeneker, M. L., Serras, F. R., de Mello, G. L., Costa, A. S., Palma, M. F., Mesquita, R. M., de Paula Guets, R., da Silva, M. M., Finger, M., de Sousa, M. C. P., Namiuti, C., and do Monte, V. M. (2023). Carolina: a general corpus of contemporary brazilian portuguese with provenance, typology and versioning information.

[Devlin et al. 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

[Garcia et al. 2021] Garcia, M., Kramer Vieira, T., Scarton, C., Idiart, M., and Villavicencio, A. (2021). Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

[Hashempour and Villavicencio 2020] Hashempour, R. and Villavicencio, A. (2020). Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In Zock, M., Chersoni, E., Lenci, A., and Santus, E., editors, *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80, Online. Association for Computational Linguistics.

[King and Cook 2018] King, M. and Cook, P. (2018). Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 345–350, Melbourne, Australia. Association for Computational Linguistics.

[Phelps et al. 2022] Phelps, D., Fan, X.-R., Gow-Smith, E., Tayyar Madabushi, H., Scarton, C., and Villavicencio, A. (2022). Sample efficient approaches for idiomaticity detection. In Bhatia, A., Cook, P., Taslimipoor, S., Garcia, M., and Ramisch, C., editors, *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 105–111, Marseille, France. European Language Resources Association.

[Pires et al. 2019] Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

[Rodrigues et al. 2023] Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing neural encoding of portuguese with transformer albertina pt-*.

[Rohanian et al. 2020] Rohanian, O., Rei, M., Taslimipoor, S., and Ha, L. A. (2020). Verbal multiword expressions for identification of metaphor. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

[Souza et al. 2020] Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. *Brazilian Conference on Intelligent Systems (BRACIS)*. arXiv preprint arXiv:2009.10683.

[Tagnin 2013] Tagnin, S. E. O. (2013). *O jeito que a gente diz: combinações consagradas em inglês e português*. Disal, Barueri.

[Tayyar Madabushi et al. 2022] Tayyar Madabushi, H., Gow-Smith, E., Garcia, M., Scarton, C., Idiart, M., and Villavicencio, A. (2022). SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In Emerson, G., Schluter, N., Stanovsky, G., Kumar, R., Palmer, A., Schneider, N., Singh, S., and Ratan, S., editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

[Tayyar Madabushi et al. 2021] Tayyar Madabushi, H., Gow-Smith, E., Scarton, C., and Villavicencio, A. (2021). AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Tenney et al. 2019] Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. In *Proceedings of ACL*, pages 4593–4601.

[Xatara 2001] Xatara, C. M. (2001). Tipologia das expressões idiomáticas. *ALFA: Revista de Linguística*, 42(1).

[Zeng and Bhat 2021] Zeng, Z. and Bhat, S. (2021). Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.