

Proposta de algoritmo de classificação automática de papéis semânticos em português no âmbito do modelo Abstract Meaning Representation

Proposal of an algorithm for automatic classification of semantic roles in Portuguese within the Abstract Meaning Representation model

Jackson Wilke da Cruz Souza *¹, **Pedro Semcovici** †² e **Thiago Alexandre Salgueiro Pardo** ‡³

¹*Universidade Federal da Bahia, Instituto de Ciência, Tecnologia e Inovação, Camaçari, BA, Brasil.*

²*Universidade de São Paulo, Escola de Artes, Ciências e Humanidades, São Paulo, SP, Brasil.*

³*Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, São Carlos, SP, Brasil.*

Resumo

O nível semântico em Processamento de Linguagem Natural (PLN) apresenta desafios significativos devido à complexidade dos fenômenos, que são menos suscetíveis a descrições objetivas. Nem todas as abordagens linguísticas, como o modelo teórico de papéis semânticos proposto por Cançado e Amaral (2017), são facilmente implementáveis em sistemas computacionais devido à sua variabilidade terminológica e metodológica. O modelo Abstract Meaning Representation (AMR) (Banarescu *et al.*, 2013; Weischedel *et al.*, 2013) tem se destacado por oferecer uma representação clara da estrutura argumental, proporcionando explicabilidade tanto para humanos quanto para sistemas computacionais sobre como o sentido se organiza em sentenças de línguas naturais. Baseando-se no AMR, desenvolvemos um classificador automático de papéis semânticos. Utilizando técnicas de Aprendizado de Máquina, nosso classificador foi treinado e testado em um corpus multigênero em Português do Brasil. Realizamos dois experimentos: o primeiro comparando Argumentos 0 e 1, e o segundo comparando Argumentos de 0 a 4, obtendo melhores resultados no primeiro experimento. Os resultados ressaltam a importância da aplicação de modelos semânticos em PLN para o português e abrem possibilidades para novas iniciativas de pesquisas.

Palavras-chave: Papéis semânticos. Abstract Meaning Representation. Processamento de Linguagem Natural.

Abstract

Semantic level in Natural Language Processing (NLP) presents significant challenges due to the phenomena's complexity, which are less amenable to objective description. Not all linguistic approaches, such as the semantic role theory proposed by Cançado and Amaral (2017), can be easily implemented in computational systems due to their terminological and methodological variability. The Abstract Meaning Representation (AMR) model (Banarescu *et al.*, 2013; Weischedel *et al.*, 2013) has gained prominence for providing a clear representation of argument structure, offering explicability both for humans and computational systems on how meaning is organized in sentences of natural languages. Based on AMR, we developed an automatic semantic role classifier. Using Machine Learning techniques, our classifier was trained and tested on a multi-genre corpus in Brazilian Portuguese. We conducted two experiments: the first comparing Arguments 0 and 1, and the second comparing Arguments 0 to 4, achieving better results in the former. The results highlight the importance of applying semantic models in NLP for Portuguese and open possibilities for new research initiatives.

Keywords: Semantic roles. Abstract Meaning Representation. Natural Language Processing.



DOI: 10.1590/1983-3652.2025.55346

Seção:
Artigos

Autor Correspondente:
Jackson Wilke da Cruz Souza

Editor de seção:
Daniervelin Pereira
Editor de layout:
João Mesquita

Recebido em:
18 de outubro de 2024

Aceito em:
19 de novembro de 2024
Publicado em:
23 de fevereiro de 2025

Esta obra tem a licença
"CC BY 4.0".



*Email: jackcruzsouza@gmail.com

†Email: pedrosemcovici@usp.br

‡Email: tasparo@icmc.usp.br

1 Introdução

As pesquisas em Processamento de Linguagem Natural (PLN), de maneira geral, buscam “investigar e propor métodos e sistemas de processamento computacional da linguagem humana” (Caseli; Nunes; Pagano, 2023, p. 10). Nesse ínterim, as pesquisas podem desenvolver ferramentas, recursos e/ou aplicações, que abarcam fenômenos linguísticos entre os níveis fonético/fonológico e discursivo. A depender do nível linguístico a ser analisado, os fenômenos podem ser mais ou menos complexos e abstratos do ponto de vista de representação formal e de processamento automático. Assim, quanto mais próximo do nível discursivo, mais subjetivos os fenômenos tendem a ser, ao passo que, quanto mais próximo do nível fonético/fonológico, mais objetivos.

Quanto ao processamento automático do nível semântico, em particular, tem-se desafios complexos, já que há fenômenos com menos possibilidades de serem descritos objetivamente. Com efeito, nem todas as propostas de metodologias e teorias que nascem nos estudos linguísticos são passíveis de pronta implementação computacional, dado o desafio de modelar computacionalmente fenômenos e perspectivas não discretas. Isso significa dizer que, diante da ausência de traços característicos que possibilitem distinção entre classes observáveis, por exemplo, sistemas computacionais podem não atingir bom desempenho.

Apesar do desafio de modelar computacionalmente o sentido, é possível que utilizar aspectos sintáticos contribua positivamente para essa tarefa, tal como a perspectiva de papéis temáticos. Cançado e Amaral (2017, p. 39) definem esse conceito como noções semânticas que “apresentam relações diretas com estruturações e propriedades sintáticas”. Considere as sentenças ilustradas abaixo.

- (1) (1.a) João quebrou a mesa.
 - (1.b) A mesa se quebrou.
 - (1.c) A mesa foi quebrada por João.

Nas sentenças em (1), a função semântica de “a mesa” será sempre de paciente, sendo caracterizada como entidade que sofre o efeito de alguma ação, havendo mudança de estado (Cançado; Amaral, 2017). Entretanto, em (1.a), a função sintática é de complemento, enquanto em (1.b) e em (1.c) é de sujeito. Já “João” exerce a função semântica de agente, que seria o desencadeador de alguma ação, capaz de agir com controle (Cançado; Amaral, 2017); entretanto, em (1.a), sintaticamente exerce a função de sujeito e em (1.c) aparece na posição de adjunção. Nesse sentido, Cançado e Amaral (2017, p. 40) destacam que “apesar de os argumentos estarem em diferentes posições sintáticas, as sentenças não são distintas e sem relação”, pois são assertivas sobre o mesmo evento, diferenciando-se apenas de pontos de vista.

Há diversos autores que apresentam conjuntos de papéis temáticos a partir de descrições linguísticas, como Halliday (1966), Fillmore (1968), Chafe (1970) e Jackendoff (1976) e, para citar estudos do Português do Brasil (PB), Gerald e Ilari (1987) e Cançado e Gonçalves (2016). De maneira mais específica, Cançado (2012) propôs uma tipologia mais abrangente de papéis para o PB, conforme demonstrado no Tabela 1.

A ordem ocupada pelos predicados nas sentenças está contida nas informações semânticas e sintáticas dos itens lexicais verbais, dando origem a outro conceito importante à noção de papéis temáticos, a saber, a *estrutura argumental*, como exemplificado em (2).

- (2) (2.a) João_[Agente] correu
- (2.b) João_[Causa] quebrou o vaso_[Paciente]
- (2.c) João_[Agente] colocou seus pertences_[Tema] na estante_[Locativo]

A estrutura argumental não aponta necessariamente para a ordem com que os predicados ocorrem nas sentenças, nem para as suas possibilidades de flexão morfológica; a estrutura destaca apenas a exigência de cada um dos argumentos. É importante destacar que, no escopo deste trabalho, essa estrutura, apesar de não incluir informações sintáticas, é o que determina a organização dos itens lexicais nas construções linguísticas, incluindo as possibilidades e restrições de alternâncias entre eles. Nesse caso, há uma relação bastante próxima entre estrutura argumental e sintaxe, como destacado por Camacho (1999), já que a ordem dos elementos na sentença contribui para a construção do

Tabela 1. Caracterização dos papéis temáticos para o PB.

PAPEL	DEFINIÇÃO	EXEMPLO
Agente	Desencadeador de alguma ação, capaz de agir com controle	O <i>motorista</i> lavou o carro. O <i>atleta</i> correu.
Causa	Desencadeador de alguma ação, sem controle	As <i>provas</i> preocupam a <i>Maria</i> . O <i>sol</i> queimou a <i>plantação</i> .
Paciente	Entidade que sofre o efeito de alguma ação, havendo mudança de estado	O <i>João</i> quebrou o <i>vaso</i> . O <i>acidente</i> machucou a <i>Maria</i> .
Tema	Entidade transferida, física ou abstratamente, por uma ação	O <i>colega</i> jogou a <i>bola</i> para a <i>menina</i> . O <i>pai</i> deu <i>uma viagem</i> para a <i>filha</i> .
Experienciador	Ser animado que está ou passa a estar em determinado estado mental, perceptual ou psicológico	O <i>namorado</i> pensou na <i>amada</i> . O <i>colecionador</i> viu um <i>pássaro</i> diferente.
Resultativo	Resultado de uma ação, ou seja, alguma entidade que não existia e passa a existir ou vice-versa	As <i>provas</i> preocupam a <i>Maria</i> . O <i>pedreiro</i> construiu a <i>casa</i> . A <i>bruxa</i> comeu a <i>maçã</i> .
Beneficiário	Ser animado que é beneficiado ou prejudicado no evento descrito	O <i>patrão</i> pagou o <i>funcionário</i> . A <i>mujer</i> perdeu a <i>carteira</i> . A <i>bibliotecária</i> emprestou o <i>livro</i> para o <i>aluno</i> .
Objeto estativo	Entidade ou situação à qual se faz referência, sem que esta desencadeie uma ação ou seja afetada por uma ação	O <i>aluno</i> leu <i>um livro do Chomsky</i> . O <i>marido</i> ama a <i>mujer</i> .
Locativo	Lugar de onde algo se desloca, para onde algo se desloca ou em que algo está situado ou acontece	A <i>modelo</i> voltou de <i>Paris</i> . A <i>Sara</i> jogou a <i>bola</i> para o <i>alto</i> . Eu moro em <i>Belo Horizonte</i> . O <i>show</i> aconteceu no <i>teatro</i> .
Instrumento	Instrumento usado por um agente no desencadeamento da ação	<i>Uma tesoura sem ponta</i> cortou as <i>gravuras</i> .

Fonte: Cançado e Amaral (2017).

sentido. A estrutura argumental de (2.a), por exemplo, pode ser usada tanto para o verbo *correr* no pretérito perfeito, como está, ou no imperfeito. O mesmo ocorre em (2.b): tanto na voz ativa quanto na voz passiva a estrutura argumental será idêntica.

Há uma série de limitações que podem dificultar que papéis temáticos sejam reconhecidos ou interpretados por humanos, bem como por sistemas computacionais, como a interpretação de *Tema* e *Objeto estativo*, conforme ilustrado no Quadro 1. Nesses casos, humanos têm à disposição uma série de mecanismos (cognitivos e de conhecimento de mundo, por exemplo) que podem auxiliar na interpretação das construções linguísticas. Por conta desse aspecto e de o conhecimento nesse tipo de construto teórico não estar explícito, impõem-se mais desafios de implementar computacionalmente esse tipo de abordagem.

Ainda que os conceitos de papéis temáticos e semânticos se refiram à interpretação e representação de sentidos, a noção de papéis temáticos remonta à organização entre argumentos de um predicado na estrutura sintática, atrelando-se, por vezes, a uma perspectiva grammatical, como demonstrado em (2). Já a noção de papéis semânticos consiste normalmente em uma abordagem mais ampla, calcada no significado e na interpretação da sentença em relação ao evento descrito nela, não apenas na organização grammatical.

Gildea e Jurafsky (2002), em uma perspectiva computacional, propõem-se a trabalhar com papéis semânticos. Os autores destacam que os conjuntos de papéis temáticos mais abstratos foram propostos com o objetivo de explicar generalizações entre predicados e argumentos, apontando para uma teoria grammatical. Já os conjuntos de papéis propostos no âmbito do PLN, conforme ainda insistem, são mais específicos, enfatizando as realizações sintáticas entre predicados e argumentos.

Nesse sentido, a teoria dos papéis semânticos precisa estar formalizada de maneira tal que a relação entre argumentos e predicados possa ser explícita. Isso, a princípio, garante que os sistemas computacionais obtenham maior acurácia (ou acerto) ao identificar e/ou rotular as relações e papéis semânticos. Para tanto, o modelo *Abstract Meaning Representation* (AMR) (Banarescu *et al.*, 2013; Weischedel *et al.*, 2013) tem ganhado destaque nos estudos e aplicações em PLN, por ser um modelo que garante explicitude sobre a estrutura argumental, permitindo explicabilidade não apenas para humanos, mas também para sistemas computacionais, de como o sentido se estrutura em sentenças

de línguas naturais.

Tais aspectos são relevantes nas discussões atuais em PLN face aos *Large Language Models* (LLMs). Um dos desafios impostos por esses modelos de língua é a dificuldade de se obter explicabilidade acerca de inferências e correlações semânticas realizadas. Apesar do avanço metodológico e técnico sobre a quantidade de dados que podem ser processados em menor tempo, o aprendizado do “sentido” acontece implicitamente, sem que um humano precise supervisionar o processo, *grosso modo*. A proposta da AMR é trabalhar com a estrutura semântica de maneira explícita, permitindo maior entendimento não apenas por humanos, mas também por sistemas computacionais.

Tendo a AMR como pressuposto teórico, objetiva-se neste trabalho apresentar um algoritmo de classificação de papéis semânticos baseado em Aprendizado de Máquina (AM). Para tanto, partiu-se de um *corpus* multigênero (literário, jornalístico, opinativo e científico) em PB, em que os textos escritos já estavam pré-anotados com os papéis semânticos do modelo AMR.

Com esse intuito, este artigo está organizado em cinco seções, além desta Introdução. Na Seção 2, apresentamos com mais detalhes os pressupostos teóricos da AMR, bem como os trabalhos relacionados à tarefa de *Semantic Role Labeling* (SRL). Na Seção 3, apresentamos a metodologia empregada neste trabalho, bem como a descrição do *corpus* utilizado no desenvolvimento do algoritmo. Na Seção 4, apresentamos o resultado obtido pelos algoritmos na tarefa de classificação. Por fim, na Seção 5, tecemos algumas considerações finais.

2 Revisão da literatura

Nesta seção, serão apresentados trabalhos recuperados da literatura que dialogam direta ou indiretamente com a temática desenvolvida nesta pesquisa, sobretudo relacionados aos papéis semânticos em PLN e à aplicação de SRL.

2.1 Papéis semânticos nos estudos de PLN

Segundo Banarescu *et al.* (2013), o modelo AMR objetiva capturar e representar explicitamente aspectos do significado de uma sentença. Esse modelo teórico propõe a enumeração de argumentos na tentativa de simplificar a proposta de papéis semânticos. Weischedel *et al.* (2013) apontam que na AMR, tradicionalmente, segue-se a seguinte proposta:

Arg0 : refere-se ao argumento da ação que desempenha o papel de agente;

Arg1 : refere-se ao argumento principal que é afetado pela ação expressa pelo verbo, correspondendo, em geral, ao objeto direto;

Arg2 : refere-se a um argumento secundário ou ao objeto indireto em construções que apresentam verbos;

Arg3, Arg4 e Arg5 : são termos menos frequentes nas construções linguísticas, ocorrendo em construções verbais mais complexas, como para o verbo “dar” ou “entregar”.

A partir dessas definições, é importante ressaltar que a proposta do modelo AMR pode ou não levar em conta aspectos morfossintáticos e/ou a ordem em que os itens lexicais ocorrem nas sentenças. O Arg0, por exemplo, é definido apenas a partir de aspectos semânticos; todos os outros argumentos levam em conta, em alguma medida, informações gramaticais que consideram a ordem e/ou a classificação morfossintática. Outro ponto importante desse modelo é que itens lexicais/*tokens* que não contribuem fortemente para a construção dos sentidos em determinadas sentenças (como determinantes/artigos) não são considerados na análise, dando-se ênfase na estrutura argumental.

A título de exemplificação, tem-se a sentença apresentada em (3), que foi extraída do *corpus* AMR-PT (Lima Inácio *et al.*, 2023), que contempla sentenças extraídas da obra “O pequeno príncipe”, dentre outros gêneros textuais, atualmente.

(3) Meu desenho não representava um chapéu.

De acordo com a proposta AMR, na sentença exemplificada em (3), tem-se dois argumentos: “desenho” é o argumento principal em relação ao verbo “representar” e, por isso, recebe a classificação de Arg1; e “chapéu”, que desempenha o papel de “objeto”, recebe a classificação Arg2. Os itens lexicais “meu” e “não” indicam, respectivamente, a relação de posse e a polaridade negativa da

sentença. Tais relações também são consideradas e representadas no modelo AMR. Por fim, o item “um” não faz parte da estrutura argumental pois não exerce modificação significativa sobre “chapéu”, ainda que desempenhe função de indeterminação sobre ele.

Além disso, o modelo propõe possibilidades de representação do conhecimento semântico, aspecto extremamente importante para implementação computacional. Uma dessas representações é a *estrutura de grafos*, em que os nós representam eventos e entidades mencionados na sentença; já as arestas representam as relações entre os nós. A Figura 1 representa a estrutura argumental do exemplo em (3).

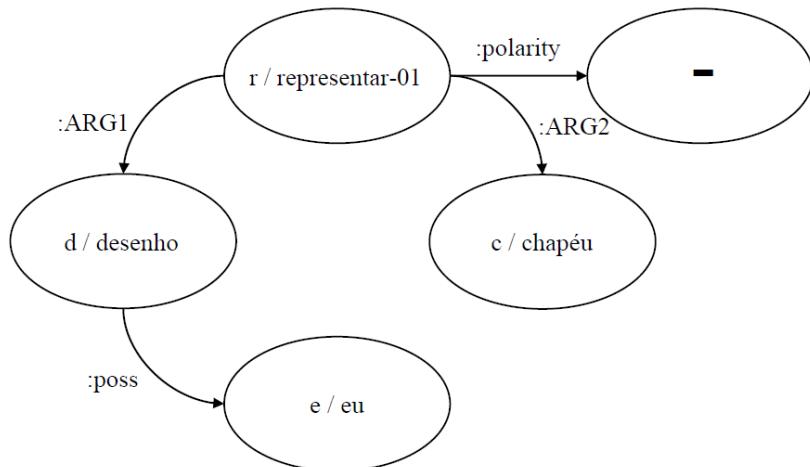


Figura 1. Exemplo de grafo AMR.

Fonte: Adaptado de Torres Anchieta e Salgueiro Pardo (2022).

Outra possibilidade de representação é por meio de *notação lógica* (Figura 2), em que, a partir da identificação léxica do predicado, apontam-se os tipos de argumentos, seus respectivos itens lexicais e a relação semântica estabelecida entre eles. Por fim, na *notação Penman* (Figura 3), feita em formato textual, há delimitação de instâncias, que são os itens lexicais que funcionam como argumentos e/ou predicados nas sentenças, a estrutura argumental entre as instâncias e as relações que podem estabelecer entre si. Cabe destacar que as três figuras equivalem à mesma representação semântica, sendo que a Figura 1, por seu apelo gráfico, pode ser melhor interpretável por humanos, ao passo que as Figuras 2 e 3 permitem implementações computacionais por serem notações lógicas.

```

(r / representar-01
  :polarity -
  :ARG1 (d / desenho
    :poss (e / eu))
  :ARG2 (c / chapéu))

```

Figura 2. Exemplo de notação lógica no modelo AMR.

Fonte: Elaborado pelos autores.

```

instance(r, representar-01) ^
instance(d, desenho) ^
instance(e, eu) ^
instance(c, chapéu) ^
polarity(r, '-') ^
ARG1(r, d) ^
poss(d, e) ^
ARG2(r, c)

```

Figura 3. Exemplo de notação Penman no modelo AMR.

Fonte: Elaborado pelos autores.

Outro aspecto importante na proposta do modelo é a classificação dos conceitos. De acordo com Torres Anchieta e Salgueiro Pardo (2022), os conceitos AMR podem ser classificados em *concretos* (palavras em suas formas lexicalizadas, como “mulher” e “homem”), *framesets* do *Proposition Bank* – PropBank¹ – (Palmer; Gildea; Kingsbury, 2005) ou *abstratos* (que não correspondem a nenhuma

¹ PropBank pode ser definido como um recurso de sentenças anotadas com funções semânticas (Jurafsky; Martin, 2023).

unidade lexical das sentenças, como “%” e “endereço de e-mail”, por exemplo).

2.2 Trabalhos relacionados à SRL

De acordo com Hartmann, Duran e Aluísio (2017), SRL é uma tarefa em PLN que detecta eventos descritos em sentenças e os participantes desses eventos. Os eventos ocorrem nas sentenças sob formas morfológicas de verbos, nomes, adjetivos e advérbios; porém, a classe que mais é explorada na literatura é a dos verbos, pois expressam eventos, além de ser impossível construir orações completas com ausência de verbos. As demais classes de palavras não expressam eventos da mesma forma que os verbos, nem em proporção.

Ainda segundo Hartmann, Duran e Aluísio (2017), os métodos empregados na anotação automática de papéis semânticos são, na maioria, com base em AM. Nesse contexto, é necessário que haja *corpora* linguísticos anotados para que os algoritmos de AM possam ser treinados e avaliados quanto ao desempenho da tarefa a que eles foram submetidos. A literatura aponta alguns *corpora* disponíveis com anotação AMR, importantes para as tarefas de SRL, a saber: Xue *et al.* (2014) para o inglês; Vanderwende, Menezes e Quirk (2015) com uma abordagem multilíngue para Francês, Alemão, Espanhol e Japonês; Damonte e Cohen (2019) para as línguas italiana, espanhola, alemã e chinesa; Migueles-Abraira, Agerri e Diaz de Ilarraz (2018) especificamente para o espanhol; e Torres Anchíeta e Salgueiro Pardo (2018) e Lima Inácio *et al.* (2023), para o PB. Destaca-se que há também recursos com papéis semânticos desvinculados da estruturação completa proposta pela AMR, como é o caso do PropBank do inglês, citado anteriormente, e o PropBank.Br (Duran; Aluísio, 2012) e o recente PBP (*Porttinari-base Propbank*) para o português (Freitas; Salgueiro Pardo, 2024).

Em especial, o *corpus* para o PB é composto por sentenças anotadas com o modelo AMR para diferentes gêneros textuais. A primeira versão do *corpus* apresentava sentenças extraídas do romance “O pequeno príncipe”, tido como *corpus* “little prince”. Em sua segunda versão, o escopo de domínio foi ampliado para notícias extraídas do jornal Folha de S. Paulo (Duran; Aluísio, 2012), tido como o *corpus* “news”. Em sua versão mais recente², o AMR-PB apresenta textos opinativos, tido como *corpus* “opisums”, e científicos, tido como *corpus* “sci”. Ademais, todos os *corpora* apresentam sentenças anotadas, identificando os papéis semânticos seguindo a metodologia do PropBank.Br (Duran; Aluísio, 2012).

Com relação ao PropBank.Br, Duran e Aluísio (2012) destacam que o *corpus* é um repositório de proposições, alinhando-se ao que Fillmore (1968) apresenta sobre a estrutura base de uma frase, a saber: um conjunto de relações entre substantivos e verbos, sem modificadores de tempo, negação, aspecto e modo. Segundo as autoras, os verbos recebem um código que indica seu sentido com relação ao *frame* em que a sentença está associada; já os argumentos são anotados com rótulos de função numerados (de Arg0 a Arg5) e os modificadores são anotados com rótulos de função ArgMs (Modificadores de argumento).

Vale a pena ressaltar que os *corpora* delimitados aqui seguem as diretrizes da literatura quanto ao PLN. Os conjuntos de dados linguísticos acompanham os apontamentos de Gildea e Jurafsky (2002), já que representam semanticamente os papéis temáticos distanciando-se de detalhamentos analíticos, ao passo que se aproximam da objetividade e simplificação de rótulos para garantir melhor desempenho no processo de classificação automática.

Partindo de *corpora* anotados, é possível aplicar diferentes técnicas de AM. O trabalho de Hartmann (2015) realizou a análise automática de papéis semânticos utilizando AM nos *subcorpora news* e *opisums*. Aplicando diferentes abordagens (utilizando ou não árvores sintáticas treinadas), os autores obtiveram Medida-F de 87,8% para o *corpus news*, e 94,5% para o *corpus opisums*.

Hartmann, Duran e Aluísio (2012) realizaram a avaliação de dois sistemas de SRL para textos jornalísticos. O sistema desenvolvido por Fonseca e Rosa (2013) teve o propósito de anotar automaticamente os papéis semânticos para o PB, sem lançar mão de ferramentas de PLN externas ao sistema desenvolvido; os resultados dessa avaliação indicam uma Medida-F de 68%. Já o sistema de

Por conta da dificuldade de se obter um conjunto “universal” de papéis semânticos, como demonstrados nesta seção, nos PropBanks convencionou-se que as relações semânticas são numeradas (ao invés do uso de seus nomes), que é justamente a filosofia adotada pela AMR.

² Disponível em: GitHub - nilc-nlp/AMR-BP. Acesso em: 20 jul.2024.

Alva-Manchego e Rosa (2012) empregou a abordagem supervisionada de AM e obteve Medida-F de 79,6%. Destaca-se que ambos os trabalhos foram baseados no *corpus* PropBank.Br.

Ilmy e Khodra (2020) realizaram a análise de frases em indonésio utilizando a AMR com abordagem de aprendizado de máquina. Inspirado no trabalho de Zhang *et al.* (2019), o sistema desenvolvido por Ilmy e Khodra (2020) compreende três etapas: (i) previsão de pares de palavras, (ii) previsão de rótulos e (iii) construção de grafos. Em (i), utiliza-se um componente de análise de dependência para obter as conexões entre palavras. Em (ii), emprega-se um algoritmo de aprendizado supervisionado para prever os rótulos entre as conexões. Em (iii), a partir dos rótulos previstos, é criado o grafo da sentença em AMR. Avaliado com uma base de dados de frases simples coletadas de artigos e notícias, o modelo atingiu uma pontuação SMATCH³ de 0,820.

3 Metodologia

Como esta pesquisa possui um viés linguístico dentro do PLN, a proposta dos métodos e de seus respectivos algoritmos refletirá essa perspectiva, como apontam Rodrigues, Souza e Santos (2022). Destaca-se que o ambiente utilizado para criar, treinar e testar os modelos apresentados neste trabalho foram configurados na versão 3.12.3 do Python, tendo como sistema operacional o POP!_OS 22.04. Ademais, as bibliotecas utilizadas em Python serão especificadas ao longo desta seção. Destaca-se que o código completo para desenvolvimento deste estudo, bem como sua documentação detalhada, está disponível na plataforma GitHub dos autores⁴.

A proposta de classificação de papéis semânticos neste trabalho segue a proposta metodológica de Ilmy e Khodra (2020). Os autores propõem identificar relações linguísticas utilizando uma estrutura hierárquica entre *parent* e *child*, sendo, respectivamente, as palavras de saída e de chegada em uma representação em grafo de sentenças anotadas com AMR. Após isso, constrói-se um arquivo no formato texto (com extensão txt) em que todas essas informações estão organizadas. A fim de exemplificação, apresentamos a sentença anotada em (4), bem como o Tabela 2, em que as informações linguísticas estão organizadas.

(4) (4.a) Sentença: Meyer tem duas explicações para esses resultados.

(4.b) Anotação: (h3 / ter-01 :ARG0 (p0 / pessoa :name (n1 / name :op1 "Meyer") :ARG1 (e5 / explicar-01 :ARG0 p0 :ARG1 (t8 / coisa :mod (t7 / esse) :ARG2-of (r9 / resultar-01) :quant 2) :wiki -))

Tabela 2. Mapeamento estrutural em *parent* e *child* de informações linguísticas.

Dependência sintática	<i>parent</i> <i>child</i> 5 dep	ter Pessoa (<i>person</i>) nsubj
Características ligadas a <i>parent</i>	Texto (<i>Text</i>) Lema (<i>Lemma</i>) Etiqueta morfossintática (POS)	tem ter verbo (VERB)
Características ligadas a <i>child</i>	Entidade nomeada Texto (<i>Text</i>) Lema (<i>Lemma</i>) Etiqueta morfossintática (POS)	<i>não se aplica</i> Meyer Meyer Nome próprio (PROPN)
Papel semântico	Entidade nomeada Rótulo (<i>Label</i>)	Pessoa (pers) Arg0

Fonte: Elaborado pelos autores.

³ SMATCH é uma métrica que tem como objetivo avaliar estruturas semânticas em ocorrências linguísticas (Cai; Knight, 2013).

⁴ Disponível em: <https://github.com/semcovici/AMR-BP-prediction>

No Tabela 2, observam-se características extraídas da relação apenas entre as palavras “Meyer” e “ter”. Para que seja feita a classificação do papel (a saber, Arg0), são consideradas informações morfológicas, morfossintáticas e de Entidades nomeadas tanto da palavra na posição *parent*, quanto da palavra na posição *child*. Tais informações servirão de *features* para os modelos automáticos de classificação, ao passo que o valor da aresta entre *parent* e *child* (a saber, a dependência sintática) servirá como variável *target*.

Nesta pesquisa, foram utilizados os quatro *corpora* do AMR-BP, os quais foram padronizados na extensão txt. Tal procedimento não foi necessário para o *corpus* PropBank, que estava em formato CoNLL. Em seguida, foi gerado um arquivo de extensão json para todos os *corpora*. Para extrair as informações linguísticas exemplificadas no Tabela 2, foi utilizada a biblioteca spaCy (SpaCy, 2024). Ao final, o arquivo json contém as seguintes informações:

- ID da sentença;
- Nome do *corpus* de origem;
- Informações dos nós: IDs dos nós e seus respectivos valores;
- Informações das arestas: ID das arestas e os nós que elas conectam;
- Sentença original;
- Sentença anotada em AMR;
- Alinhamentos entre itens léxicos e conceitos AMR, quando for possível criar ou estiverem disponíveis. O *corpus news* possui parte de suas sentenças com alinhamentos; no caso do PropBank, é possível criar os alinhamentos dada a formatação CoNLL do *corpus*;
- Sentença com anotação linguística, contendo texto, lema, POS, dependência e Entidades nomeadas.

Algumas sentenças dos *corpora* não apresentaram alinhamentos entre as palavras analisadas como *parent* e *child*. Nesse sentido, optou-se pela inferência utilizando o método *match* da *string* do AMR com a sentença original. Porém, houve casos em que isso não foi possível, já que a anotação semântica não apresentava uma palavra/*token* correspondente ao texto original, como “Tenho de admitir” em relação à anotação “*oblige*”. Assim, ao final, somaram-se 2.212 casos em que não foi possível inferir o alinhamento, frente a 3.760 em que foi possível, sendo estes últimos desconsiderados neste trabalho.

Ao final, obteve-se a seguinte distribuição de papéis semânticos: (i) Arg0 com 4.154 exemplares; (ii) Arg1 com 7.052; (iii) Arg2 com 1.714; (iv) Arg3 com 188; (v) Arg4 com 137; e (vi) Arg5 com 1 único exemplar. Diante dessa distribuição, optou-se por não considerar Arg5 por conta de sua baixa representatividade. Além disso, dado o desbalanceamento dos outros papéis, optou-se por realizar dois experimentos para a construção dos classificadores automáticos: Experimento 1, utilizando apenas os dados relativos a Arg0 e Arg1; Experimento 2, utilizando os dados relativos a Arg0 até Arg4.

A construção do modelo de predição de papéis semânticos deste trabalho baseou-se em seis etapas. Na primeira, realizou-se a separação dos conjuntos de treino e teste. Os dados provenientes do PropBank foram utilizados apenas para treino; já os dados restantes foram divididos em 80% para treino e 20% para teste, considerando a informação “*label*” das instâncias que, neste trabalho, são os próprios papéis semânticos.

Na segunda etapa foi necessário aplicar a vetorização em variáveis preenchidas com valor textual (ou seja, categóricas). Para tanto, foi aplicada a técnica *One-Hot Encoding* (Rodríguez *et al.*, 2018). Ainda, para as *features* relacionadas ao lema das palavras das instâncias, foi feita a vetorização utilizando o modelo Word2Vec de *Continuos Bag-of-Words* com 300 dimensões de Hartmann, Duran e Aluísio (2017).

A terceira etapa consistiu na sequência organizada de subetapas. Foi utilizado o método Pipeline do *imbalanced-learn*⁵, consistindo em:

- *Normalização*, em que as variáveis foram normalizadas utilizando *Z-score*⁶ através do método *StandardScaler*⁷ do *scikit-learn* (Pedregosa *et al.*, 2011);

⁵ Disponível em: Pipeline — Version 0.12.3. Acesso em: 20 jul.2024.

⁶ O *Z-score* mede quantos desvios padrão um valor está distante da média da distribuição. Trata-se de uma forma de transformação dos dados que torna a distribuição deles mais compreensível e comparável entre si.

⁷ Disponível em: StandardScaler – scikit-learn 1.5.1 documentation. Acesso em: 20 jul.2023.

- *Seleção de melhores features*, em que foram selecionadas as melhores features por meio do método *SelectPercentile*⁸ do *scikit-learn*;
- *Sobre-amostragem*, em que o método *Synthetic Minority Over-sampling Technique* (SMOTE), do *imbalanced-learn*⁹ (Lemaître; Nogueira; Aridas, 2017), foi utilizado para sobre-amostrar as classes minoritárias até que elas possuíssem o mesmo número de amostras da classe majoritária;
- *Classificação*, em que foi utilizado o algoritmo *Extreme Gradient Boosting* (XGBoost) (Chen; Guestrin, 2016) da biblioteca XGBoost¹⁰.

A quarta etapa consistiu no ajuste de hiperparâmetros do algoritmo XGBoost. Trata-se de uma etapa necessária para otimizar o desempenho do modelo de classificação. Para tanto, foi utilizado o método *RandomizedSearch* do *scikit-learn*¹¹, que realiza uma validação cruzada com combinações aleatórias de parâmetros pré-definidos.

A quinta etapa de avaliação foi realizada a partir das métricas clássicas de AM, a saber:

- *Precisão* (P), em que é medida a proporção de verdadeiros positivos entre os exemplos classificados como positivos pelo modelo. Para tanto, baseia-se no seguinte cálculo: $P = \frac{VP}{VP+FP}$, em que VP é a quantidade de verdadeiros positivos, e FP é o número de falsos positivos.
- *Revocação* (R), em que é medida a proporção de verdadeiros positivos entre os exemplos que são realmente positivos. Para tanto, baseia-se no seguinte cálculo: $R = \frac{VP}{VP+FN}$, em que FN é o número de falsos negativos.
- *Medida-F* (MF), que é a média harmônica entre P e R, proporcionando balanceamento entre as duas métricas. Para tanto, baseia-se no seguinte cálculo: $MF = 2 \frac{Precision \cdot Recall}{Precision + Recall}$
- *Acurácia* (A), em que é medida a proporção de exemplos corretamente classificados pelo modelo entre todos os exemplos. Para tanto, o cálculo é feito da seguinte maneira: $acurácia = \frac{VP+VN}{VP+VN+FP+FN}$, em que VN é a quantidade de falsos negativos.

Por fim, a sexta etapa consistiu na seleção de *features* mais relevantes para a classificação de papéis semânticos. Foram utilizados os métodos *feature_importance* do algoritmo XGBoost e a biblioteca SHAP (*SHapley Additive exPlanations*) (Lundberg; Lee, 2017), a qual permitiu comparar as *features* mais importantes para cada tipo de argumento.

4 Resultados e discussão

Como dito anteriormente, foram feitos dois experimentos de classificação de papéis semânticos. O primeiro envolve uma classificação binária (Arg0 e Arg1); já o segundo aborda uma classificação multiclasse com cinco rótulos (Arg0, Arg1, Arg2, Arg3 e Arg4). Na Tabela 3, apresentamos os resultados obtidos, considerando as medidas P, R, MF e A.

Tabela 3. Resultados para os dois experimentos de classificação.

CLASSES	EXPERIMENTOS							
	1º experimento (Arg0 e Arg1)				2º experimento (Arg0 a Arg4)			
	P	R	MF	A	P	R	MF	A
Arg0	0,83	0,82	0,82	0,85	0,82	0,79	0,81	0,76
Arg1	0,88	0,89	0,88	0,85	0,77	0,84	0,80	0,76
Arg2	-	-	-	-	0,40	0,32	0,36	0,76
Arg3	-	-	-	-	0,33	0,14	0,20	0,76
Arg4	-	-	-	-	1,00	0,20	0,33	0,76

Fonte: Elaborado pelos autores.

Para o primeiro experimento, os resultados mostram um desempenho equilibrado entre as duas classes. As métricas P, R e MF para Arg0 são, respectivamente, 83%, 82% e 82%, enquanto para Arg1 são 88%, 89% e 88%. Esses resultados indicam um classificador bem balanceado, sugerindo que

⁸ Disponível em: SelectPercentile – scikit-learn 1.5.1 documentation. Acesso em: 20 jul.2023.

⁹ Disponível em: SMOTE — Version 0.12.3. Acesso em: 20 jul. 2023.

¹⁰ Disponível em: Python API Reference – xgboost 2.1.0 documentation. Acesso em: 20 jul.2023.

¹¹ Disponível em: RandomizedSearchCV — scikit-learn 1.5.1 documentation. Acesso em: 20 jul.2024.

o modelo consegue distinguir bem entre essas duas classes.

Por haver um aumento de classes, já era esperado um desbalanceamento na performance do modelo entre as diferentes classes, no segundo experimento. Para as classes Arg0 e Arg1, os resultados são semelhantes aos do primeiro experimento, com MF de 81% e 80%, respectivamente. No entanto, para as outras classes (Arg2, Arg3 e Arg4), as métricas são significativamente mais baixas.

Para a classe Arg2, as medidas P, R e MF são 40%, 32% e 36%, respectivamente. A classe Arg4 apresenta uma precisão perfeita (100%), mas uma revocação extremamente baixa (20%), resultando em uma MF de apenas 33%. A classe Arg3 também apresenta baixos valores em todas as métricas, com precisão, revocação e medida F de 33%, 14% e 20%, respectivamente.

A acurácia de 76% no segundo experimento sugere que o modelo tem uma alta taxa de acertos nas classes majoritárias (Arg0 e Arg1), mas tem dificuldade para classificar corretamente as classes minoritárias (Arg2, Arg3 e Arg4), o que é evidenciado pelo baixo desempenho nestas classes. Esse desbalanceamento indica que o modelo pode tender a classificar mais exemplos nas classes majoritárias, possivelmente devido a um número insuficiente de exemplos das classes minoritárias durante o treinamento.

Esse resultado pode ser justificado não apenas do ponto de vista computacional, perpassado pelo desbalanceamento das classes, mas também de uma perspectiva linguística. No primeiro experimento, consideraram-se papéis de argumentos mais prototípicos e mais claramente associados a posições sintáticas estabelecidas como sujeito e objeto, sendo mais fáceis de se classificar. Já no segundo, observa-se uma diversidade semântica e estrutural, sendo argumentos menos prototípicos, ocasionando equívocos de classificação de predicados e adjuntos como argumentos. Nesse caso, parece pertinente pontuar que é quase impossível desassociar a estrutura argumental da sintaxe, cabendo mais discussões sobre essa correlação em estudos posteriores.

Na língua geral, os papéis Arg0, Arg1 e Arg2 (como exemplificado em (5.a), (5.b) e (5.c), respectivamente) são mais profícuos, e os falantes constroem rotineiramente sentenças que cumprem essas estruturas argumentais. Já as construções que apresentam Arg3 e Arg4 (como exemplificado em (5.d) e (5.e), respectivamente) seguem a proporção inversa. Destaca-se que os exemplos demonstrados em (5) foram extraídos do *corpus little prince*, já anotado com o modelo AMR, tendo a indicação do argumento correspondente em foco.

(5) (5.a) [Ela]_{Arg0} me perfumava.

(5.b) Perdoa-me [(eu)]_{Arg1}.

(5.c) Mas era tão [comovente!]]_{Arg2}

(5.d) [Perfur-o]_{Arg3} com suas raízes.

(5.e) Vou passear até a [vinha.]_{Arg4}

É importante ressaltar que os predicados no modelo AMR equivalem à descrição de uma ação, evento, estado ou relação e, por isso, são frequentemente associados aos verbos, ainda que seja possível associá-los a adjetivos e/ou substantivos. Além disso, frequentemente os predicados correspondem a *frames* semânticos, construindo cenas e relações que exigem certos complementos. Já os argumentos são partícipes ou entidades que se relacionam aos predicados, respondendo a perguntas como “quem”, “o que?” e “para quem?”, por exemplo, sendo enumerados em função da posição nos *frames* em que são evocados.

Partindo desse princípio, os exemplos (5.a), (5.b) e (5.e) cobrem essa concepção; ainda que o quinto exemplo seja, na teoria linguística, um argumento da preposição, na perspectiva da AMR ele é um argumento do *frame* “passear”. Ao analisar os exemplos presentes em (5.c) e (5.d), evoca-se a necessidade de realizar revisões da anotação realizada. Em (5.c), “comovente” deveria ser considerado como predicado, mas recebeu a anotação de argumento, pois completa o sentido de estado (“ser comovente”, no caso). Já em (5.d), o argumento deveria ser apenas o objeto direto (no caso, “o”) e não o verbo “perfurar”. Esta última colocação é justificada por uma questão de tokenização do *corpus* e do modelo de segmentação morfossintática utilizado, aglutinando o “o” a “perfura”.

Ainda, em seus estudos, Cançado (2009) encontrou predicados com no máximo cinco lugares,

como o verbo *transportar* em “João transportou os livros no carro de São Paulo para a Bahia”. Nesse caso, a autora admite que o complemento locativo “no carro” não é parte obrigatória da estrutura argumental.

Nesse sentido, parece pertinente observar não apenas o quanto os classificadores propostos acertam, mas também como se distribui o equívoco de classificação no conjunto de dados. Por conta disso, foi feita uma análise sobre a matriz de confusão dos modelos criados, em que é possível avaliar a performance dos modelos em função de verdadeiros e falsos positivos e negativos. Apresentamos as matrizes de confusão para os dois experimentos nas Figuras 4 e 5.

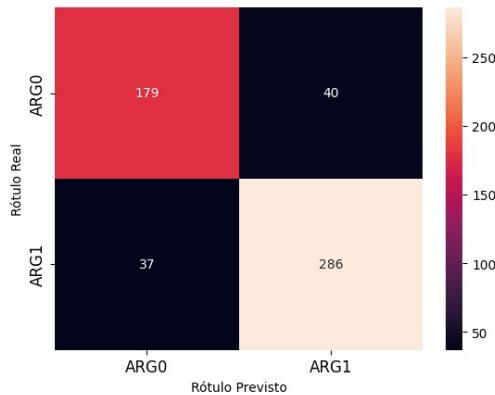


Figura 4. Matriz de confusão do primeiro experimento

Fonte: Elaborado pelos autores.

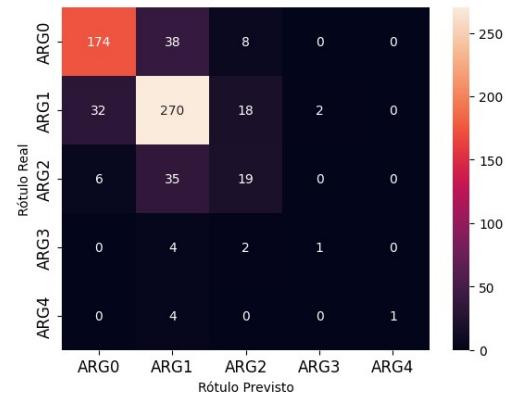


Figura 5. Matriz de confusão do segundo experimento

Fonte: Elaborado pelos autores.

As matrizes de confusão para os dois experimentos complementam as métricas de avaliação apresentadas. No primeiro experimento, o modelo mostra um bom equilíbrio entre P e R, com 179 (de 219) exemplos de Arg0 e 286 (de 323) exemplos de Arg1 corretamente classificados.

No segundo experimento, o modelo apresentou desequilíbrio em seu desempenho. As classes Arg0 e Arg1 mantêm uma alta taxa de acerto, ao passo que Arg2, Arg3 e Arg4 sofrem de baixa precisão e alta taxa de erros. Essa confusão se deve ao fato de haver um desbalanceamento entre os tipos de argumentos, conduzindo à classificação conforme a classe majoritária, a saber, Arg1, conforme já discutido anteriormente.

Ainda, foi feita uma análise sobre o desempenho do modelo considerando as classes e os *corpora* utilizados nos dois experimentos.

Tabela 4. Resultados do Experimento 1 em função dos *corpora* utilizados.

Corpora	Classe	P	R	MF	Qnt. de instâncias
<i>little prince</i>	Arg0	0,89	0,87	0,88	119
	Arg1	0,90	0,91	0,91	151
<i>news</i>	Arg0	0,74	0,75	0,74	60
	Arg1	0,87	0,86	0,87	116
<i>opisms</i>	Arg0	0,77	0,71	0,74	34
	Arg1	0,78	0,84	0,81	43
<i>sci</i>	Arg0	0,86	1,00	0,92	6
	Arg1	1,00	0,92	0,96	13

Fonte: Elaborado pelos autores.

Na Tabela 4, relativa ao Experimento 1, observa-se que, considerando a medida de avaliação MF, o modelo apresenta um melhor desempenho para classificar Arg1 do que Arg0 para todos os

corpora utilizados. Tal resultado pode ser explicado pela quantidade de instâncias analisadas em Arg1 ser maior que Arg0. Ainda sobre a distribuição das instâncias entre os *corpora*, destaca-se que, apesar do *corpus sci* apresentar menos instâncias (no total, 19), o *corpus opisums* apresentou o menor desempenho com relação a MF, a despeito da quantidade de instâncias (no total, 77). No entanto, é possível notar que há um bom equilíbrio entre P e R para todos os *corpora*, indicando resultados consistentes para a classificação do modelo.

Tabela 5. Resultados do Experimento 2 em função dos *corpora* utilizados.

Corpora	Classe	P	R	MF	Qnt. de instâncias
<i>little prince</i>	Arg0	0,85	0,89	0,87	120
	Arg1	0,81	0,82	0,81	151
	Arg2	0,22	0,19	0,21	21
	Arg3	0,50	0,20	0,29	5
	Arg4	1,00	0,25	0,40	4
<i>news</i>	Arg0	0,77	0,68	0,73	60
	Arg1	0,74	0,86	0,80	115
	Arg2	0,57	0,39	0,46	31
	Arg3	0,00	0,00	0,00	1
	Arg4	0,00	0,00	0,00	1
<i>opisums</i>	Arg0	0,77	0,59	0,67	34
	Arg1	0,69	0,84	0,76	43
	Arg2	0,29	0,29	0,29	7
	Arg3	0,00	0,00	0,00	1
	Arg4	0,00	0,00	0,00	0
<i>sci</i>	Arg0	0,86	1,00	0,92	6
	Arg1	1,00	0,85	0,92	13
	Arg2	1,00	1,00	1,00	1
	Arg3	0,00	0,00	0,00	0
	Arg4	0,00	0,00	0,00	0

Fonte: Elaborado pelos autores.

Na tabela 5 relativa ao Experimento 2, observa-se que o Arg0 obteve um melhor desempenho de classificação nos *corpora* *little prince* e *sci*, com MF de 0,87 e 0,92, respectivamente. Já Arg1 foi melhor classificado nos *corpora* *news* e *opisums*, com MF de 0,88 e 0,76, respectivamente. Destaca-se também que no *corpus news*, Arg3 e Arg4 não foram classificados, apesar de ter um único exemplo para cada uma dessas classes; o mesmo ocorre para Arg3 no *corpus opisums*. Em contrapartida, apenas uma única instância de Arg2 no *corpus sci* foi classificada corretamente pelo modelo.

Ainda com relação à Tabela 5, é possível inferir que as sentenças dos *corpora* podem apresentar estruturas argumentais distintas a depender do domínio e/ou gênero textual a que se vinculam.

Além disso, foi feito um estudo sobre as *features* mais relevantes para a classificação dos papéis semânticos utilizando o método *feature importance*. Para tanto, somou-se a importância das *features* do Experimento 1 (Figura 6) e do Experimento 2 (Figura 7).

Na Figura 6, observa-se que os *embeddings* dos lemas, tanto de *parent* quanto de *child*, são particularmente significativos para o modelo de classificação. No entanto, destaca-se que *embeddings* são as *features* mais numerosas após a seleção. Portanto, embora tenham uma grande soma de importância, individualmente, sua relevância não é tão alta. Já na Figura 7, além dos *embeddings* dos lemas, *child_tag*, *child_pos* e *dep* também se destacam. Nesse sentido, é possível inferir que, no modelo criado, características de ordens morfossintáticas e sintáticas são relevantes para a indicação

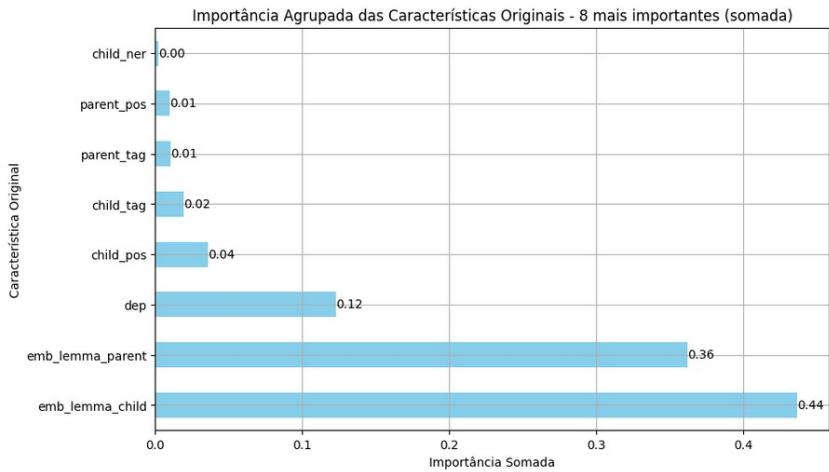


Figura 6. *Features* mais relevantes para o Experimento 1.

Fonte: Elaborado pelos autores.

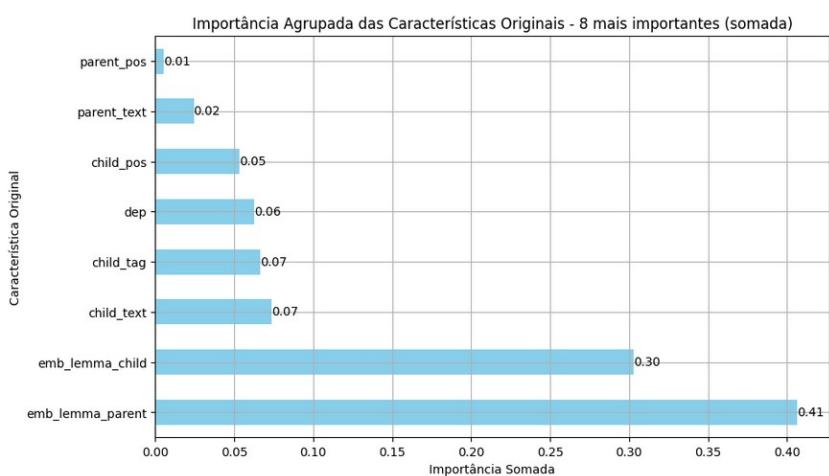


Figura 7. *Features* mais relevantes para o Experimento 2.

Fonte: Elaborado pelos autores.

dos papéis semânticos. Ressalta-se que este tipo de observação é relevante em análises futuras, indicando quais *features* são mais relevantes em tarefas como a que foi desenvolvida neste trabalho, além de contrapor o modelo teórico AMR que, em tese, exclui esses níveis de conhecimento linguístico na representação do significado e de sua estrutura argumental.

Por fim, foi empregado o método de *shap values* para uma análise mais profunda. Nesse método, conseguimos analisar como cada *feature* performa em função dos papéis semânticos observados. É importante destacar que tal método se difere do *feature_importance* por considerar a interação entre as *features* e a sua influência nas predições específicas. Vale ressaltar que utilizamos apenas os *shap values* para o conjunto de treino, tendo em vista que o objetivo da análise é entender quais padrões o modelo entendeu no seu treinamento como sendo mais importantes. Para tanto, somaram-se os *shap values* absolutos para cada uma das *features* originais e tirou-se a média de todas as instâncias agrupadas pelo argumento para o Experimento 1 (Figura 8) e o Experimento 2 (Figura 9).

Na Figura 8, evidencia-se que *emb_lemma_parent* tem discretamente maior relevância na classificação automática para Arg1 do que Arg2, o que é inversamente proporcional para *dep* e *emb_lemma_child*. Quanto a *child_tag*, *child_ner* e *child_pos*, a relevância é bem baixa nesse cenário para as duas classes analisadas; ao passo que *parent_pos* e *parent_tag* não apresentaram qualquer relevância para nenhum dos dois papéis semânticos. Quando o cenário de classificação é ampliado para os Args 0 a 4, o classificador recorre a algumas *features* diferentes daquelas utilizadas no primeiro experimento. Na Figura 9, em especial, tem-se que *emb_lemma_parent* e *emb_lemma_child* foram as *features*

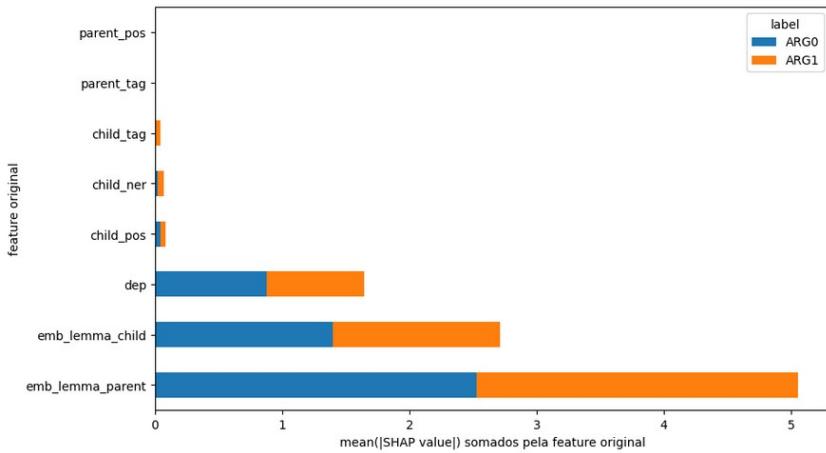


Figura 8. Features mais relevantes para o Experimento 1 em função dos argumentos utilizando *shap values*.

Fonte: Elaborado pelos autores.

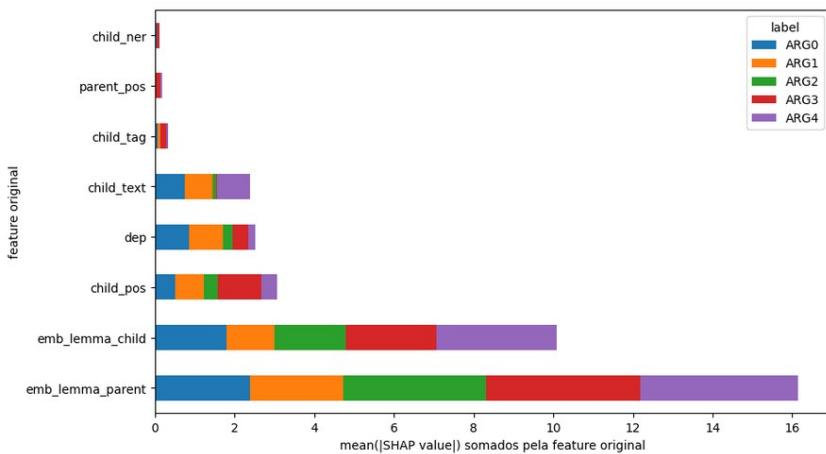


Figura 9. Features mais relevantes para o Experimento 2 em função dos argumentos utilizando *shap values*.

Fonte: Elaborado pelos autores.

com maior relevância na classificação dos papéis semânticos, sobretudo entre as classes Arg2, Arg3 e Arg4; ao passo que *child_text*, que no experimento anterior não havia sido utilizada, apresenta maior impacto para Arg0, Arg1 e Arg4.

5 Considerações finais

Ao longo deste trabalho, como demonstrado, utilizaram-se métodos automáticos de anotação de papéis semânticos em textos da língua portuguesa a partir de um *corpus* multigênero anotado com o modelo AMR. É importante ressaltar que a proposta AMR objetiva explicitar o conhecimento semântico em construções linguísticas visando a implementação computacional. Porém, sua base teórica tem nascedouro e correlação direta com teorias linguísticas, como apontado neste trabalho.

Nesse sentido, destaca-se a importância de implementar computacionalmente uma proposta que modele o sentido de maneira explícita. Esta demonstra ser uma alternativa bastante interessante frente às propostas neurais atuais da Inteligência Artificial, em que há dificuldade de se obter explicabilidade das correlações semânticas, que podem produzir equívocos e “alucinações”.

Os resultados deste trabalho evidenciam uma melhor performance do modelo com relação à distinção entre Arg0 e Arg1. Um dos fatos que justifica esse apontamento é a grande quantidade de casos no *corpus*, refletindo a realidade para além do conjunto de dados. Entretanto, o modelo enfrenta desafios significativos quando a tarefa de classificação se estende a um conjunto maior e mais

diversificado de argumentos, como o acréscimo de Arg2, Arg3 e Arg4.

Se, por um lado, as dificuldades enfrentadas pelo modelo desenvolvido neste trabalho salientam os desafios impostos pela própria língua, por outro indicam a necessidade de técnicas adicionais para lidar com o desbalanceamento entre os argumentos, como demonstrado. Nesse sentido, em trabalhos futuros, caberá o uso de estratégias de reamostragem (real ou artificial) e/ou ajuste de hiperparâmetros que penalizem mais severamente equívocos na classificação em classes com menor quantidade de exemplares.

Após observar as métricas de avaliação, é possível constatar que, a despeito dessas limitações, os resultados do modelo de classificação desenvolvido neste trabalho são compatíveis com o cenário atual de PLN em PB. Tais constatações também vão ao encontro das reflexões linguísticas tecidas e demonstradas ao longo da discussão. Nesse sentido, é possível destacar que o trabalho foi bem-sucedido.

Quanto aos trabalhos futuros, destaca-se ainda que os resultados são condizentes com a quantidade de dados que se tinha à disposição na época para compor etapas de treinamento e teste do modelo de classificação. Caso a quantidade de instâncias analisadas fosse maior, sobretudo nas classes Arg3, Arg4 e Arg5, é possível que o desempenho do modelo pudesse apresentar melhores resultados de classificação. Assim, caberá aumentar a quantidade de exemplos das classes minoritárias e refazer as etapas de treinamento e teste, avaliando o modelo novamente.

Além dessa estratégia, uma outra que pode ser adotada em trabalhos futuros está relacionada à padronização dos dados. É possível que utilizar um alinhador AMR para inferir os alinhamentos das instâncias aprimore o desempenho do classificador. Porém, em PB, há escassez de ferramentas com esse propósito, quando comparado ao cenário de pesquisas em língua portuguesa.

Para o leitor interessado, mais detalhes sobre este trabalho podem ser encontrados no portal web do projeto POeTiSA¹².

6 Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - <http://c4ai.inova.usp.br/>), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Além disso, contou com o apoio financeiro do Edital PRPPG-UFBA 010/2024 – Programa de Apoio a Jovens Professores(as)/Pesquisadores(as) 2024 – e do Edital FAPESB/CNPq 004/2023 – Programa Primeiros Projetos.

Referências

ALVA-MANCHEGO, Fernando Emilio; ROSA, João Luís G. Semantic Role Labeling for Brazilian Portuguese: A Benchmark. In: PAVÓN, Juan; DUQUE-MÉNDEZ, Néstor D.; FUENTES-FERNÁNDEZ, Rubén (ed.). *Advances in Artificial Intelligence–IBERAMIA 2012: 13th Ibero-American Conference on AI*. Cartagena de Indias, Colombia: Springer Berlin Heidelberg, 2012. p. 481–490. DOI: 10.1007/978-3-642-34654-5_49.

BANARESCU, Laura et al. Abstract Meaning Representation for Sembanking. In: PROCEEDINGS of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 178–186.

CAI, Shu; KNIGHT, Kevin. Smatch: An Evaluation Metric for Semantic Feature Structures. In: SCHUETZE, Hinrich; FUNG, Pascale; POESIO, Massimo (ed.). *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 748–752.

CAMACHO, R.G. Estrutura Argumental e Funções Semânticas. *Alfa*, São Paulo, v. 43, p. 145–170, 1999.

¹² <https://sites.google.com/icmc.usp.br/poetisa>

- CANÇADO, Márcia. Argumentos: Complementos e Adjuntos. *ALFA: Revista de Linguística*, v. 53, n. 1, p. 35–59, 2009.
- CANÇADO, Márcia. Verbos Psicológicos: Uma Classe Relevante Gramaticalmente? *Veredas-Revista de Estudos Linguísticos*, v. 16, n. 2, p. 1–18, 2012.
- CANÇADO, Márcia; AMARAL, Luana. *Introdução à Semântica Lexical: Papéis Temáticos, Aspecto Lexical e Decomposição de Predicados*. [S. l.]: Editora Vozes Limitada, 2017.
- CANÇADO, Márcia; GONÇALVES, Anabela. Lexical Semantics: Verb Classes and Alternations. In: WETZELS, Leo; COSTA, João; MENUZZI, Sergio (ed.). *The Handbook of Portuguese Linguistics*. [S. l.: s. n.], 2016. p. 374–391. DOI: 10.1002/9781118791844.ch20.
- CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe; PAGANO, Adriana. O que é PLN? In: CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe (ed.). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. [S. l.]: Bpln, 2023. ISBN 978-65-00-80693-9. Disponível em: <https://brasileraspln.com/livro-pln/1a-edicao/parte1/cap1/cap1.html>.
- CHAFE, W.L. *Meaning and the Structure of Language*. Chicago, USA: University of Chicago Press, 1970.
- CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: KRISHNAPURAM, Balaji; SHAH, Mohak; SMOLA, Alexander J.; AGGARWAL, Charu; SHEN, Dou; RASTOGI, Rajeev (ed.). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA: Acm, 2016. p. 785–794. DOI: 10.1145/2939672. Disponível em: <http://doi.acm.org/10.1145/2939672>.
- DAMONTE, Marco; COHEN, Shay B. Structural Neural Encoders for AMR-to-Text Generation. In: BURSTEIN, Jill; DORAN, Christy; SOLORIO, Thamar (ed.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota: Association for Computational Linguistics (ACL), 2019.
- DURAN, Magali Sanches; ALUÍSIO, Sandra Maria. PropBank-Br: A Brazilian Treebank Annotated with Semantic Role Labels. In: CALZOLARI, Nicoletta; CHOUKRI, Khalid; DECLERCK, Thierry; DOĞAN, Mehmet Uğur; MAEGAARD, Bente; MARIANI, Joseph; MORENO, Asuncion; ODIJK, Jan; PIPERIDIS, Stelios (ed.). *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. Istanbul, Turkey: European Language Resources Association, 2012. p. 1862–1867.
- FILLMORE, C.J. Lexical Entries for Verbs. *Foundations of Language*, v. 4, p. 373–393, 1968.
- FONSECA, Erick R.; ROSA, João Luís Garcia. A Two-Step Convolutional Neural Network Approach for Semantic Role Labeling. In: THE 2013 International Joint Conference on Neural Networks. Dallas, USA: Ieee, 2013. p. 1–7.
- FREITAS, Cláudia; SALGUEIRO PARDO, Thiago Alexandre. PropBank e Anotação de Papéis Semânticos para a Língua Portuguesa: O que Há de Novo? In: ANAIS do 15º Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL). Belém: Sociedade Brasileira de Computação, 2024. p. 118–128. DOI: 10.5753/stil.2024.245377. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/31123>.
- GERALDI, J.W.; ILARI, R. *Semântica*. São Paulo: Ática, 1987. v. 3.
- GILDEA, Daniel; JURAFSKY, Daniel. Automatic Labeling of Semantic Roles. *Computational Linguistics*, v. 28, n. 3, p. 245–288, 2002.
- HALLIDAY, M. Some Notes on 'Deep' Grammar. *Journal of Linguistics*, v. 2, n. 1, p. 57–67, 1966.
- HARTMANN, Nathan Siegle. *Anotação Automática de Papéis Semânticos de Textos Jornalísticos e de Opinião sobre Árvores Sintáticas Não Revisadas*. 2015. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brasil. DOI: 10.11606/D.55.2015.tde-27112015-140053.
- HARTMANN, Nathan Siegle; DURAN, Magali Sanches; ALUÍSIO, Sandra Maria. Automatic Semantic Role Labeling on Non-Revised Syntactic Trees of Journalistic Texts. In: SILVA, João; RIBEIRO, Ricardo;

QUARESMA, Paulo; ADAMI, André; BRANCO, António (ed.). *Computational Processing of the Portuguese Language*. Cham: Springer International Publishing, 2017. p. 202–212. ISBN 978-3-319-41552-9. DOI: 10.1007/978-3-319-41552-9_20.

ILMY, Adylan Roaffa; KHODRA, Masayu Leylia. Parsing Indonesian Sentence into Abstract Meaning Representation using Machine Learning Approach. *In: 2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*. [S. l.: s. n.], 2020. p. 1–6. DOI: 10.1109/icaicta49861.2020.9429051.

JACKENDOFF, Ray. Toward an Explanatory Semantic Representation. *Linguistic Inquiry*, The MIT Press, Cambridge, USA, v. 7, n. 1, p. 89–150, 1976. Disponível em: <http://www.jstor.org/stable/4177913>.

JURAFSKY, D.; MARTIN, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3. ed. [S. l.: s. n.], 2023.

LEMAÎTRE, Guillaume; NOGUEIRA, Fernando; ARIDAS, Christos K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, v. 18, p. 1–5, 2017.

LIMA INÁCIO, Marcio; SOBREVILLA CABEZUDO, Marco Antonio; RAMISCH, Renata; DI FELIPPO, Ariani; SALGUEIRO PARDO, Thiago Alexandre. The AMR-PT Corpus and the Semantic Annotation of Challenging Sentences from Journalistic and Opinion Texts. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, v. 39, e202339355159, 2023. DOI: 10.1590/1678-460x202339355159. Disponível em: <https://revistas.pucsp.br/index.php/delta/article/view/55159>.

LUNDBERG, Scott M.; LEE, Su-In. A Unified Approach to Interpreting Model Predictions. *In: PROCEEDINGS of the 31st Conference on Neural Information Processing Systems*. Long beach, California, USA: Curran Associates, 2017. v. 30, p. 4768–4777.

MIGUELES-ABRAIRA, Noelia; AGERRI, Rodrigo; DIAZ DE ILARAZA, Arantza. Annotating Abstract Meaning Representations for Spanish. *In: CALZOLARI, Nicoletta; CHOUKRI, Khalid; CIERI CHRISTOPHER ANDDECLERCK, Thierry; GOGGI, Sara; HASIDA, Koiti; ISAHARA, Hitoshi; MAEGAARD, Bente; MARIANI, Joseph; MAZO, Hélène; MORENO, Asuncion; ODIJK, Jan; PIPERIDIS, Stelios; TOKUNAGA, Takenobu (ed.). Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. Miyazaki, Japan: European Language Resources Association, 2018. p. 3074–3078.

PALMER, Martha; GILDEA, Daniel; KINGSBURY, Paul. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, v. 31, n. 1, p. 71–106, 2005.

RODRIGUES, Roana; SOUZA, Jackson Wilke da Cruz; SANTOS, Roney Lira de Sales. Descrição Linguística e Aprendizado de Máquina: Análise de Verbos Locativos do Espanhol. *Cadernos de Estudos Linguísticos*, v. 64, n. 00, e022038, 2022. DOI: 10.20396/cel.v64i00.8666995.

SPACY. *Industrial-Strength Natural Language Processing*. [S. l.: s. n.], 2024. <https://spacy.io>. Acesso em: 20 jul. 2024.

TORRES ANCHIÉTA, Rafael; SALGUEIRO PARDO, Thiago Alexandre. Towards AMR-BR: A Sembank for Brazilian Portuguese Language. *In: CALZOLARI, Nicoletta; CHOUKRI, Khalid; CIERI, Christopher; DECLERCK, Thierry; GOGGI, Sara; HASIDA, Koiti; ISAHARA, Hitoshi; MAEGAARD, Bente; MARIANI, Joseph; MAZO, Hélène; MORENO, Asuncion; ODIJK, Jan; PIPERIDIS, Stelios; TOKUNAGA, Takenobu (ed.). Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. p. 974–979.

TORRES ANCHIÉTA, Rafael; SALGUEIRO PARDO, Thiago Alexandre. Análise Semântica com Base em AMR para o Português. *Linguamática*, v. 14, n. 1, p. 33–48, 2022. DOI: 10.21814/lm.14.1.358. Disponível em: <https://linguamatica.com/index.php/linguamatica/article/view/358>.

VANDERWENDE, Lucy; MENEZES, Arul; QUIRK, Chris. An AMR Parser for English, French, German, Spanish and Japanese and a New AMR-annotated Corpus. *In: GERBER, Matt; HAVASI, Catherine; LACATUSU, Finley (ed.). Proceedings of the 2015 Conference of the North American Chapter of the*

Association for Computational Linguistics: Demonstrations. Denver, Colorado, USA: Association for Computational Linguistics, 2015. p. 26–30. DOI: 10.3115/v1/N15-3006.

WEISCHEDEL, Ralph *et al.* *OntoNotes Release 5.0 LDC2013T19*. Philadelphia, USA: Linguistic Data Consortium, 2013.

XUE, Nianwen; BOJAR, Ondřej; HAJIČ, Jan; PALMER, Martha; UREŠOVÁ, Zdeňka; ZHANG, Xiuhong. Not an Interlingua, But Close: Comparison of English AMRs to Chinese and Czech. *In:* CALZOLARI, Nicoletta; CHOUKRI, Khalid; DECLERCK, Thierry; LOFTSSON, Hrafn; MAEGAARD, Bente; MARIANI, Joseph; MORENO, Asuncion; ODIJK, Jan; PIPERIDIS, Stelios (ed.). *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland: European Language Resources Association, 2014. p. 1765–1772.

ZHANG, Sheng; MA, Xutai; DUH, Kevin; VAN DURME, Benjamin. AMR Parsing as Sequence-to-Graph Transduction. *In:* KORHONEN, Anna; TRAUM, David; MÀRQUEZ, Lluís (ed.). *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 80–94. DOI: 10.18653/v1/P19-1009. Disponível em: <https://aclanthology.org/P19-1009/>.

Contribuições dos autores

Jackson Wilke da Cruz Souza: Conceituação, Curadoria de dados, Análise formal, Metodologia, Recursos, Programas, Supervisão, Validação, Escrita – rascunho original, Escrita – revisão e edição; **Pedro Semcovici:** Curadoria de dados, Programas, Escrita – rascunho original, Escrita – revisão e edição; **Thiago Alexandre Salgueiro Pardo:** Conceituação, Recursos, Supervisão, Escrita – rascunho original, Escrita – revisão e edição.