ISSN 1678-765X doi 10.20396/rdbci.v21i00.8671396

Artigo

RDBCI
Revista Digital de Biblioteconomia e Ciência da Informação
Digital Journal of Library and Information Science

Uma proposta de arquitetura da informação aplicada ao processamento de linguagem natural: contribuições da Ciência da Informação no pré-processamento de dados para treinamento e aprendizado de redes neurais artificiais

George Hideyuki Kuroki Júnior [1] Claudio Gottschalg-Duque [2]

RESUMO

Introdução: O processamento de linguagem natural em redes neurais artificias possui lacunas passíveis de tratamento por parte da Ciência da Informação, utilizando-se de Arquitetura da Informação. Objetivo: Propor contribuições da Ciência da Informação na Organização do Conhecimento para treinamento de redes neurais artificiais utilizando Arquitetura da Informação Multimodal, posicionando-a como área do conhecimento atuante em problemas de inteligência artificial. Metodologia: Adaptando um percurso de três níveis de análise (metafísico, científico e tecnológico), verifica o atual estágio de desenvolvimento de técnicas de processamento de linguagem natural (metafísico); utiliza definições de Arquitetura da Informação Multimodal propondo um procedimento de cinco passos para delineamento, análise e transformação do espaço informacional a ser utilizado em métodos de treinamento e aprendizagem de redes neurais, complementando lacunas identificadas por autores voltados a implementações da Ciência da Computação (científico); verifica a aplicabilidade da proposta em 3 conjuntos de dados advindos de 16 áreas do conhecimento como base de avaliação (tecnológico). Resultados: Os resultados obtidos nas situações com pré-tratamento e sem pré-tratamento foram comparados observando-se potencial para desenvolvimento de um método estruturado de Arquitetura da Informação Multimodal que forneça instrumentos para a organização do pré-processamento de dados a serem utilizados como massa de teste e aprendizado em redes neurais artificiais, em particular, no processamento de linguagem natural. Conclusão: Este método posicionaria a Ciência da Informação como atuante e produtora de soluções de pré-processamento de dados, sobrepondo o papel atual de mera consumidora de soluções pré-fabricadas pela Ciência da Computação.

PALAVRAS-CHAVE

Ciência da Informação. Arquitetura de informação. Tratamento da informação. Inteligência Artificial. Processamento de linguagem natural.

Correspondência do autor

[1] Universidade de Brasília
Brasília, DF - Brasil
e-mail: kurokijr@gmail.com

[2] Universidade de Brasília
Brasília, DF - Brasil
e-mail: klaussherzog@gmail.com

Information architecture applied on natural language processing: a proposal Information Science contributions on data pre-processing for training and learning of artificial neural networks

ABSTRACT

Introduction: Natural Language Processing through artificial neural networks has gaps that can be addressed by Information Science through Information Architecture. Objective: To present Information Science contributions on Knowledge Organization applied to artificial neural networks training methods, positioning it as an active body of knowledge in artificial intelligence problems. Methodology: A three-leveled analysis path (metaphysical, scientific, and technological) is adopted to guide and ground the study. On metaphysical level, current development stage of natural language processing techniques is verified and analyzed. On scientific findings, a five-step procedure is proposed which aims to design, analyze, and prepare information spaces for artificial neural networks training and learning methods, fulfilling gaps identified by authors focused on Computer Science implementations. On technological implementation, the five-step procedure is applied to 3 datasets formed by texts from 16 scientific knowledge areas, as an evaluation basis. Results: Results obtained through pre-processed data and raw data where compared, showing great potential in developing a structured method of Multimodal Information Architecture that provide instruments able to organize data used as test and learning samples in artificial neural networks. Conclusion: This method could place Information Science as a producer of data pre-processing solutions, replacing its current role as consumer of prefabricated solutions made by Computer Science.

KEYWORDS

Information Science. Information architecture. Information treatment. Artificial Intelligence. Natural language processing.

CRediT

- Reconhecimentos: Não aplicável.
- Financiamento: Não aplicável.
- Conflitos de interesse: Os autores certificam que não têm interesse comercial ou associativo que represente um conflito de interesses em relação ao manuscrito.
- Aprovação ética: Não aplicável.
- Disponibilidade de dados e material: Os dados possuem sigilo de propriedade industrial.
- Contribuições dos autores: Conceitualização, Curadoria de dados, Análise formal, Investigação, Metodologia, Software, Visualização, Escrita – rascunho original, Escrita – revisão & edição, Redação – revisão & edição: KUROKI JÚNIOR, G.H.; Supervisão, Validação: DUQUE, C. G.

JITA: BK. Information architecture.

Artigo submetido ao sistema de similaridade

Submetido em: 03/11/2022 – **Aceito em**: 16/12/2022 – **Publicado em**: 15/01/2023

Editor: Gildenir Carolino Santos

1 INTRODUÇÃO

A crescente utilização de modelos de inteligência artificial em atividades cotidianas de classificação e tratamento de informação coloca um novo prisma de observação à questão levantada por Hjørland (2008). Segundo o autor, a Organização do Conhecimento como área de estudo teria como peças centrais a Ciência da Informação e a Biblioteconomia, todavia, sendo seriamente desafiada pela Ciência da Computação.

Ao tempo em que tal afirmativa fora realizada, uma proposição de arquitetura e implementação de redes neurais artificiais desenvolvida por Hinton, Osindero e Teh (2006) possibilitou superar um obstáculo histórico enfrentado pela Computação. Até então, a construção de redes neurais artificias padecia de falta de profundidade em suas implementações: notoriamente, o cérebro humano, base para o desenvolvimento de modelos de inteligência, possui diversas camadas de análise, o que possibilita o tratamento de problemas com maior complexidade. Com o advento da proposta em questão, o número de camadas de tratamento ultrapassou o limite de duas ou três.

O perpassar desta limitação computacional deu origem a grande variedade de implementações tecnológicas, originando inumeráveis desenhos arquiteturais de redes neurais que aplicam múltiplos algoritmos matemáticos para se obter uma medida de inteligência por meio da verificação de padrões.

Ainda que haja avanços por parte da Ciência da Computação, uma crítica feita por Hjørland ainda é passível de discussão:

> Existem muitas comunidades separadas que trabalham com diferentes tecnologias, mas muito poucas pesquisas sobre seus pressupostos básicos e méritos e lados fracos. O problema não é apenas formular uma teoria, mas descobrir suposições teóricas em diferentes práticas, formular esses pressupostos de forma tão clara quanto possível, para possibilitar a comparação das abordagens. (HJØRLAND, 2008, p.87)

Um ponto em comum a todas as iniciativas da Ciência da Computação é a sua dependência de expressiva quantidade de dados e/ou registros para obtenção de padrões a serem observados. Entretanto, a obtenção destes dados nem sempre é possível, particularmente em problemas que requerem conhecimento especializado, por exemplo, a classificação de textos técnico-científicos, altamente vinculados ao vocabulário da área em questão.

Historicamente, a Ciência da Computação se ocupou primariamente do tratamento da complexidade de um modelo de rede neural perante os dados a serem analisados conforme seu crescimento exponencial, o que Bellman (1954) denominou de *problema da dimensionalidade dos dados*. Na ausência de maiores registros, técnicas de enriquecimento de dados textuais se atem a contextos cotidianos de uso, ainda utilizando da grande gama de informação disponível em outros domínios comuns.

Posiciona-se, neste artigo, a Arquitetura da Informação Multimodal (AIM) como uma contribuição inicial da Ciência da Informação, na forma de uma contrapartida teórica de pré-processamento de dados para posterior aplicação em modelos de Inteligência Artificial (IA), mais especificamente no Processamento de Linguagem Natural (PLN), em problemas classificação de textos em domínios de conhecimento específico.

2 PROCEDIMENTOS METODOLÓGICOS

Para analisar de forma estruturada os impactos da aplicação da AIM sobre problemas de PLN, propõe-se a utilização do percurso metodológico para construção de uma Visão de Mundo (M^3) criada por Van Gigch e Moigne (1989).

Tal proposta considera a construção do conhecimento ao longo de três etapas que

guardam intima relação entre elas: um nível metafísico, anterior a formalização do objeto do conhecimento; um nível do objeto do conhecimento em si; e um nível da aplicação do conhecimento construído. Neste sentido, este artigo adaptará tal metodologia da seguinte forma:

a) No nível metafísico: identificar as questões fundamentais do atual estágio do PLN e questões fundamentais da Arquitetura da Informação Multimodal;
b) No nível do objeto do conhecimento: propor formas de aplicação da AIM em problemas de PLN;
c) No nível da aplicação do conhecimento: gerar produtos de AIM para implementação em PLN.

Percorrido os três níveis da Visão de Mundo adotada, ter-se-á um conjunto de conhecimentos, técnicas e produtos passíveis de validação e verificação de sua aderência ao problema abordado, por meio de uma comparação de resultados obtidos em simulações de redes neurais artificiais partindo de um conjunto de dados não-tratado por AIM e o mesmo conjunto de dados tratado por AIM.

3 *DEEP LEARNING*: APLICAÇÕES, DESENVOLVIMENTO E DESAFIOS EM PROCESSAMENTO DE LINGUAGEM NATURAL

Os ditames fundamentais para a construção de redes neurais artificias foram sedimentados ao longo das décadas de 60 a 90. Com a entrada dos anos 2.000 e a proposta de Hinton, Osindero e Teh (2006), uma nova gama de implementações passaram a se valer da profundidade de camadas de análise, dando origem ao termo *Deep Learning*.

Wason (2018) realiza levantamento sobre a utilização das descobertas realizadas por Hinton, Osindero e Teh (2006), verificando sua utilização de forma ampla em variada gama de domínios como, por exemplo, reconhecimento de voz independente da fonte sonora; redes neurais recorrentes; reconhecimento de caligrafia; redes de crença profundas; auto-encodificadores; modelagem acústica; detectores de características de classes; síntese de caligrafia; modelagem de linguagens; melhoria e desenvolvimento de modelos dentre outros. Conclui que três grandes desafios ainda perduram na maioria das aplicações de IA:

a) Volume de dados: a massa de dados necessária para se obter aprendizado satisfatório seria da natureza de dez vezes a quantidade de parâmetros (neurônios) da rede desenhada;
b) Fenômeno de *Overfitting*: quão maior o tamanho da rede, em termos de número de parâmetros, maior a probabilidade de que o aprendizado esteja superdimensionado, resultando em uma baixa capacidade de generalização (mudanças pequenas nos objetos de entrada resultam em um resultado insatisfatório);
c) Natureza frágil: redes neurais tendem a serem especializadas, de forma que ao serem treinadas em uma determinada tarefa, seu desempenho em outra tarefa é extremamente insatisfatório.

Da junção dos dois primeiros desafios citados por Wason (2018), ainda se identifica problema anteriormente mapeado por Bellman (1954), também endereçado por Arel, Rose e Kanowski (2010) denominado *problema da dimensionalidade dos dados*, onde a complexidade de aprendizado cresce de forma exponencial em detrimento ao aumento linear do número de dimensões dos dados.

Segundo Minaee (2021), as mais recentes tentativas de obtenção de resultados otimizados em PLN baseiam-se em Transformadores e Modelos Pré-Treinados — MPT. Desde as primeiras implementações de redes neurais para PLN, como Redes Convolucionais, Redes

Recorrentes e Redes LSTM (*Long Short-Term Memories*, ou Memórias Longas de Curto-Período), percebe-se a dificuldade em capturar as relações entre palavras dentro de uma frase. Com o advento de modelos baseados em Mecanismos de Atenção proposto inicialmente por Bahdanau, Cho e Bengio (2015), redes neurais passaram a tratar diversos objetos de forma agrupada. Com base neste avanço, Vaswani *et al*. (2017) propuseram uma nova arquitetura denominada Transformadores, que trouxe duas inovações relevantes: atribuição de uma pontuação de atenção que avalia a influência de uma palavra sobre outra e melhoria nos métodos de paralelização, reduzindo o tempo de treinamento. A partir de 2018 observa-se um crescimento em MPTs baseados em Transformadores, dotados de arquiteturas mais densas e pré-treinados em grandes volumes de dados textuais o que, de forma conjunta, acarreta melhor contextualização de palavras e sentenças. Qiu *et al*. (2020) realizaram um levantamento sobre os MPTs mais utilizados, classificando-os por meio de quatro categorias:

a) Tipo de representação: forma de representação do idioma, visando a identificação de regramentos linguísticos implícitos e conhecimento de senso comum que não são explícitos em dados textuais;
b) Modelo arquitetural: modo de captura dos contextos, se de forma sequenciada (palavra após palavra) ou não-sequenciada (utilizando uma estrutura sintática ou semântica pré-definida);
c) Tipo de tarefa de pré-treinamento: objetivo pretendido ao longo do treinamento. Em aprendizado supervisionado, busca-se uma função capaz de mapear pares de entrada e saída; em aprendizado não-supervisionado, busca-se obter conhecimento intrínseco a partir de dados não-classificados; em aprendizado auto-supervisionado, há a junção dos tipos anteriores, onde o método de treinamento é baseado em aprendizado supervisionado, mas a classificação dos dados é gerada de forma automática.
d) Extensões ao modelo: MPTs geralmente visam representações universais de um idioma para aplicações genéricas. Para aplicações específicas, maior enriquecimento do modelo é desejável como multi-idioma, multimodal ou específico de um domínio ou tarefa.

Qiu *et al*. (2020) também dividem os MPTs em duas gerações conforme seus objetivos. A primeira geração busca bons modelos de mapeamento de palavras, obtendo classificação hierárquica de palavras em detrimento de um modelo da linguagem. São independentes do contexto. Word2vec de Mikolov *et al*. (2013a), GloVe de Pennington, Socher e Manning (2014) assim como CBow e Continuous Skip-Gram de Mikolov *et al*. (2013b) são exemplos. A segunda geração busca produzir vetores de palavras a nível de frases, levando em consideração o contexto em que as palavras se encontram. CoVe de McCann *et al*. (2017), ELMo de Peters *et al.* (2018), OpenAI GPT de Radford *et al.* (2018) e BERT de Devlin *et al*. (2019) são exemplos.

Dada amplitude de modelos disponíveis, Minaee *et al*. (2021) propõem um procedimento de cinco passos para a escolha de uma rede neural de PLN:

a) Seleção do MPT;
b) Adaptação ao domínio do problema;
c) Inserção de camada adaptada a tarefa;
d) Ajuste de pesos a tarefa;
e) Compressão do modelo.

Decorrida a análise de mais de 150 modelos voltados a PLN utilizando mais de 40 conjuntos de dados, os autores concluem que por maior que sejam os avanços obtidos, algumas questões permanecem desafiadoras:

a) Ausência de dados para tarefas mais complexas: embora a quantidade de dados coletados ao longo dos anos seja expressiva, tarefas como perguntas e respostas com raciocínio de múltiplos passos, classificação de textos para documentos com múltiplos idiomas e classificação de texto para documentos longos;
b) Modelos de conhecimentos de senso comum: a falta de modelos com conhecimentos de senso comum limita a capacidade de análise de redes neurais como, por exemplo, responder a perguntas sobre o mundo real ou lidar com a incompletude de informações;
c) Modelos com uso eficiente de memória: a maioria dos modelos modernos requerem grande quantidade de memória, o que leva a necessidade de compressão;
d) Aprendizado com menor esforço: a maioria dos modelos de *Deep Learning* são treinados por meio de aprendizado supervisionado. Na prática, coletar e classificar dados para um novo domínio é uma tarefa complexa e desafiadora.

Os avanços das ferramentas de processamento de linguagem natural são notórios, tanto em diversidade de implementações quanto em espectro de tratativas empreendidas, entretanto, a representação de conhecimento específico (tratado em alguma medida por Minaee *et al.* (2021) como conhecimento de senso comum) ainda representa um desafio a ser melhor abordado.

4 ARQUITETURA DA INFORMAÇÃO MULTIMODAL: CONTRIBUIÇÕES PARA O DESENVOLVIMENTO DE PLN

Segundo Kuroki Júnior (2018), define-se a Arquitetura da Informação Multimodal – AIM – como a construção e distinção de Mundos Arquiteturais, por meio de suposição de Modelos Relacionais agrupados por contextos espaço-tempo de Estados de Informação correlacionados ou não.

Estaria o termo para o autor intimamente ligado a Ciência da Informação, pela sua disposição a atuar no que Hjørland (2008) referia-se ao sentido estrito de Organização do Conhecimento: descrição, indexação e classificação de documentos. Uma imposição de *Ordem* (pela arquitetura) para ambas as correntes de conceitos de *Informação,* definidas por Capurro e Hjørland (2007): uma objetiva, tratando-a como coisa (número de *bits*, por exemplo) e outra subjetiva, que dependeria da interpretação de um agente cognitivo. Em ambos os casos, citam os autores, a Ciência da Informação se voltaria aos fenômenos de relevância e interpretação como aspectos básicos do conceito de informação.

A proposta de Kuroki Júnior (2018) estende o conceito tradicional de Arquitetura da Informação por meio da adição do conceito de *Modo* dado por Kress e Van Leeuween (2001) e Kress (2009), como qualquer recurso socialmente e culturalmente moldado para se construir significados. Para os autores, qualquer *Modo*, incluindo a língua (na concepção de idioma escrito e falado e suas possibilidades) possui limitações e potencialidades.

A expressão de significados e a consistência de um modelo relacional entre os diversos agrupamentos formados deve ser balizado por alguma medida de ordenamento. O problema reside em situações as quais uma mesma premissa possa ser considerada verdadeira em um contexto, mas falsa em outro e, ainda assim, ambos os contextos devem coexistir no mesmo modelo informacional. De forma ilustrativa simples e reduzida, a mesma rede neural de PLN deveria assumir que um termo específico (por exemplo, “sistema”) tem impacto positivo e negativo ao mesmo tempo. Eis a questão da multimodalidade em arquiteturas da informação: o custo de se modelar cada caso em que as todas as premissas sejam verdadeiras em todas as configurações possíveis excede os benefícios encontrados por esta extrema individualização e granularização dos problemas. É o dilema da Navalha de Ockham também adotado pela AIM de Kuroki Júnior (2018), por meio de dois princípios que guardam intima relação entre eles:

economia e relevância. “Pluralitas non est ponenda sine necessitae” (a pluralidade não deve ser posta sem necessidade, representada pela relevância) e “Frustra fit per plura quod potest fieri per pauciora” (é infrutífero fazer com mais o que se pode fazer com menos, representada pela economia).

No enfretamento da questão, Kuroki Júnior (2018) lança mão de estruturais lógicas modais, baseadas em operadores de possibilidade e necessidade conforme Carnielli e Pizzi (2008) e Portner (2009). Uma proposição é possível caso seja verdadeira em alguma configuração de um domínio. Uma proposição é necessária, caso seja verdadeira em todas as configurações de um domínio.

Atuaria a Ciência da Informação, por meio da AIM, na Organização do Conhecimento no sentido estrito de Hjørland (2008), produzindo visões ou agrupamentos de dados que possam expressar de forma mais eficaz um domínio ou um contexto de informação para facilitar o reconhecimento de padrões por meio de redes neurais. Na AIM, um mundo arquitetural é um contexto de relações entre sujeitos e objetos, ou seja, um conjunto de domínios semânticos, passíveis de serem moldados de múltiplas formas, partindo do mesmo conjunto de sujeitos e objetos.

Os itens a seguir detalham o procedimento de cinco passos propostos para se obter uma nova configuração de domínio informacional voltado a PLN, como fase anterior ao pré-processamento de dados executado no desenvolvimento de redes neurais artificiais.

4.1 Identificação de entidades de contexto

Para PLN e MPTs, um contexto pode ser visto tão somente como um grupo de textos agrupados por semelhança linguística, semântica, factual, senso comum ou qualquer outra característica. Tal assertiva não se aplica para a AIM. Um contexto só se torna um espaço arquitetural quando é considerado o ponto de vista de um sujeito de ao menos um objeto. Em contrapartida, um objeto pode ser classificado de forma diferente por múltiplos sujeitos, assim como determinada sequência de texto pode exprimir significados distintos em contextos distintos. Redes neurais de PLN visam sobrepor esta barreira por meio de volume de dados o que, conforme Minaee *et al.* (2021), é restrito para tarefas mais complexas. Neste sentido, a primeira intervenção da AIM visa definir os sujeitos e objetos de um contexto, sendo:

a) SUJEITO uma entidade dotada de capacidade de produzir e manipular informação;
b) OBJETO é uma entidade com potencial de significação, dotada de atributos que possam ser interpretados por sujeitos de forma comum;
c) Uma CORRELAÇÃO ocorre quando um sujeito transforma um objeto por meio de DEFINIÇÃO, COMPARAÇÃO, FUSÃO OU DECOMPOSIÇÃO e o produto desta operação é aceito dentro do corpo de conhecimento compartilhado pelos sujeitos que compõem o contexto.

Definidos os sujeitos que figuram em um contexto, a forma com que estes manipulam objetos determina a configuração do momento observado. Kuroki Júnior (2018) explicita estes diversos momentos denominando correlação a unidade fundamental de conexão entre sujeitos e objetos. Ainda que sujeitos distintos concordem que um conjunto de características definam um objeto, suas correlações são distintas, passíveis de diferenças intrínsecas não observáveis no momento de análise.

4.2 Identificação de correlações entre entes

Para Kuroki Junior (2018), as relações conectam instâncias de um contexto ou os próprios contextos em si, sendo uma correlação um tipo específico de relação. Uma correlação se forma entre um sujeito e um objeto em um determinado contexto. Em uma abordagem de PLN à luz da AIM, as correlações fundamentais propostas são quatro:

a) DEFINIÇÃO é uma correlação realizada por um sujeito que transforma o estado de um ente em um contexto para objeto, abrindo a possibilidade de agregar outros entes como atributos.
b) COMPARAÇÃO só é aplicável a objetos definidos por um sujeito. Qualquer nível de comparação se dá por meio de análise dos atributos assinalados a objetos distintos.
c) FUSÃO é a junção de dois objetos para a formação de um terceiro.
d) DECOMPOSIÇÃO é a operação oposta a fusão, onde um objeto dá origem a outros dois distintos.

Por meio destas operações são colhidas as impressões dos sujeitos atuantes em um contexto no tocante às características de um grupo de objetos denominadas atributos. Importante ressaltar que somente por meio da adoção de modelos lógicos modais a AIM pode tratar os diferentes *Modos* em que entes se agregam de forma diferente. Por exemplo, em um determinado *Modo Tecnologia* o ente "sistema" seria um objeto com atributos [informação, desenvolvimento, linguagem] ao passo que em um *Modo Política*, o mesmo ente "sistema" seria tão somente um atributo do objeto "governo". Para abrigar ambos os *Modos*, o contexto informacional deve ser subdividido em unidades menores e então, estas unidades terem suas relações (não correlações, que se referem a sujeitos e objetos) identificadas.

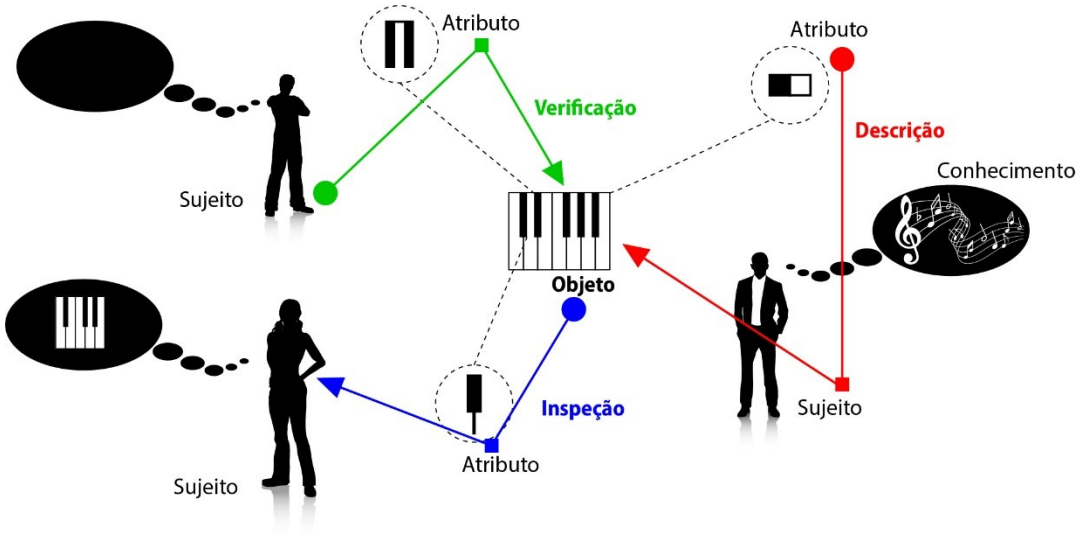
4.3 Distinção de domínios

Em uma AIM aplicada a PLN, um DOMÍNIO é um grupo de atributos de objetos que podem ser identificados de forma comum por diferentes sujeitos por meio correlações similares. Desta forma, sujeitos e objetos podem compor diversos domínios. Três formas possíveis de estabelecimento de domínios são:

a) Descrição: partindo de um conjunto de atributos potenciais, verifica-se o acolhimento semântico destes por sujeitos para então encontrar tais atributos em determinados objetos, agrupando os mesmos;
b) Inspeção: analisando um conjunto de objetos, agrupa-se os mesmos por atributos comuns e verifica-se o reconhecimento comum em um determinado grupo de sujeitos;
c) Verificação: inquirindo determinado grupo de sujeitos, identifica-se atributos percebidos de forma comum pelos indivíduos deste grupo e agrupa-se objetos que contenham estes atributos.

Uma representação gráfica destas formas pode ser observada na abaixo Figura 1 a seguir.

Figura 1. Formas de estabelecimento e distinção de domínios



Fonte: Produzido pelos autores em maio de 2022.

As subdivisões provenientes desta etapa completam o ciclo informacional da AIM: define-se qual é o escopo de "*estados de informação*" (provenientes dos itens *4.1 Identificação de entidades de contexto* e *4.2 Identificação de correlações entre entes*) "*correlacionados ou não*" (proveniente do item *4.1 Identificação de entidades de contexto4.3 Distinção de domínios*).

4.4 Proposição de relações entre domínios

As três primeiras operações visam identificar entes, correlações e domínios de um modelo endereçando o conjunto informacional a ser tratado. As relações entre estes domínios dão o caráter arquitetural da proposta, no sentido de uma imposição de economia e ordem. Para que uma AIM impacte de alguma forma em um contexto ou até um domínio, alguma alteração neste espaço informacional deve ser realizada. Isto se dá por meio de relações entre domínios.

Para Kuroki Júnior (2018), relações são dotadas de regras que as restringem. A definição primária da AIM utiliza de lógica modal para expressar relações. Três relações básicas de manipulação de domínio são propostas para alterar este domínio ou produzir um novo:

a) Identidade: uma relação de identidade é obtida quando todos os objetos de um domínio podem ser encontrados em um outro domínio. Corresponde ao operador modal de necessidade;
b) Proximidade: uma relação de proximidade é identifica quando parte dos objetos de um domínio pode ser encontrado em um outro domínio. Corresponde ao operador modal de possibilidade;
c) Incidental: relações incidentais nem sempre são perceptíveis, com certa medida de aleatoriedade em suas incidências. A forma mais simples de defini-las seria como uma relação de segunda ordem.

Quanto a extensão das relações, o autor utiliza de estruturas modais lógicas citadas por Carnielli e Pizzi (2008):

a) Reflexiva: uma estrutura reflexiva é identificada quando uma relação proposta é

aplicável de um domínio para ele mesmo;

b) Serial: uma estrutura serial é identificada quando uma relação proposta é aplicável de um domínio para outro;
c) Simétrica: uma estrutura simétrica é identificada quando uma relação proposta é aplicável mutuamente entre dois domínios;
d) Transitiva: uma estrutura transitiva é identificada quando, supondo três domínios [A, B, C], caso A tenha a relação proposta com B, e B possui a relação proposta com C, então A possui a relação proposta com C;
e) Euclidiana: uma estrutura euclidiana é identificada quando uma relação proposta é reflexiva, simétrica e transitiva.

Da combinação entre tipo e extensão obtemos a classificação completa de uma relação. Por exemplo: as relações entre os domínios A = {1,3,4}; B = {1, 3, 5} e C = {1,2,3,4,5} seriam e identidade serial de A para C e B para C; e proximidade simétrica entre A, B e C.

4.5 Agrupamento por contextos espaço-tempo

Aplicar todo o regramento possível a um domínio ou conjunto de domínios não é o objetivo da AIM. Uma medida de economia das relações deve ser levada em consideração, do contrário, qualquer configuração tenderia a mapear a realidade objetiva da forma mais próxima possível. Para Kuroki Júnior (2018), distinções espaço-tempo podem ser identificadas por meio de estruturas deônticas, que exprimem uma lógica de obrigações e permissões. Distinguem-se estas de estruturas epistêmicas, as quais tratam de conhecimento. A principal diferença reside na impossibilidade de estruturas deônticas assumirem uma verdade imutável: tão somente consideram a possibilidade de uma ocorrência. Um exemplo simples citado por Portner (2009) seria o regramento moral "não assassinar". Ainda que este seja elencado como necessário (deve existir em todos os contextos possíveis), assassinatos ocorrem ainda assim.

Todos os regramentos listados até este ponto endereçaram questões espaciais de uma arquitetura da informação: o quão abrangente um modelo é no tocante as relações, objetos e atributos que considera. A questão temporal torna-se, de fato, um limitador para qualquer modelo estático, o que leva a necessidade de um modelo cíclico, conforme Figura 2.

Figura 2. Ciclo de construção de AIMs

Fonte: Produzido pelos autores em maio de 2022.

5 IMPLEMENTAÇÃO DE UMA ARQUITETURA DA INFORMAÇÃO MULTIMODAL

Seguindo o percurso metodológico proposto, sugere-se uma aplicação de AIM em um problema de PLN a título de exemplificação. A situação selecionada refere-se à classificação de textos. A dificuldade reside tanto na ausência de dados suficientes para aprendizado quanto na abrangência semântica destes dados. Em resumo, trata-se de uma análise de tendência positiva ou negativa de um conjunto de textos segundo uma legislação de incentivos a pesquisa, desenvolvimento e inovação. Anualmente, são submetidos mais de 10.000 textos que podem ser classificados em 16 categorias de conhecimento: Agroindústria, Alimentos, Bens de Consumo, Construção Civil, Farmacêutica, Metalurgia, Mineração, Moveleira, Outras, Papel e Celulose, Têxtil, Petroquímica, Mecânica e Transportes, Eletroeletrônica, TICs e Telecomunicações. Até o presente momento, tão somente os dados de 2014 e 2015 encontram-se e analisados e classificados como "aprovado" ou "reprovado".

5.1 Aplicação de PLN em um conjunto de dados não tratados por AIM

Os textos classificados nos anos de 2014 e 2015 foram submetidos a treinamento, validação e teste em uma rede neural de classificação de textos. Para tal tarefa fora utilizado o modelo *BERTimbau* de Souza, Nogueira e Lotufo (2020), treinado por meio do corpus *brWaC* de Filho *et al.* (2018), o qual possui 3.5 milhões de documentos e 2.68 bilhões de *tokens*. O modelo utilizado separa os dados em três partes: Treinamento, Validação e Teste. Para cada conjunto, duas variáveis são observadas. Perda (*loss*) representa a diferença entre os resultados esperados e os resultados obtidos pela máquina. Por meio da perda que se obtém os ajustes dos pesos da rede neural, o que possibilita o avanço no aprendizado ao longo do experimento. Menores valores de perda indicam melhor aprendizado da rede. Acurácia (*acc*) representa o

percentual de acertos obtidos em cada etapa do experimento. Esta variável expressa o quão assertivo é o modelo com base nos dados apresentados.

Para isolar os produtos da AIM de qualquer interferência advinda de técnicas vinculadas a ciência da computação (enriquecimento da base de dados, mudança no algoritmo de aprendizado, aumento do escopo de análise), nenhum procedimento de melhoria será aplicado tanto ao ambiente quando ao conjunto de dados originais, o que garante que qualquer resultado esteja vinculado única e exclusivamente a AIM.

Foram realizados 10 experimentos com 20 ciclos de treinamento para os dados de 2014, 2015 e 2014 e 2015 em conjunto, obtendo-se os seguintes resultados médios apresentados na Tabela 1:

Tabela 1. Média de resultados dos experimentos realizados com dados não tratados

Variável	2014	2015	2014 e 2015
Perda em treinamento	0,7087808	0,5627345	0,6463273
Acurácia em treinamento	53,55%	76,38%	63,39%
Perda em validação	0,6949488	0,5708822	0,6765008
Acurácia em validação	54,52%	74,14%	59,08%
Perda em teste	0,7416452	0,4740491	0,6142412
Acurácia em teste	54,79%	77,57%	58,22%

Fonte: Produzidos pelos autores em agosto de 2022

Observa-se sensível diferença nos resultados provenientes dos dados de 2014 e 2015, estes últimos apresentando valores mais assertivos. A diferença percentual de acurácia em teste chega a 21,78% entre os anos em separado. Ao juntar ambos os conjuntos, a acurácia pende para resultados mais próximos ao ano de 2014, representando um decréscimo em relação ao melhor resultado (de 2015) de 15.52%.

Partindo dos mesmos conjuntos de dados fornecidos, os objetivos a serem alcançados por meio do pré-tratamento de dados baseado em AIM serão:

a) Encontrar configurações de agrupamento de domínios que aumentem a acurácia do algoritmo de PLN sem que haja intervenções de cunho técnico-computacional (baseados em alterações no código-fonte);
b) Identificar os domínios que apresentem dados com maior ou menor potencial de extração de aprendizado.

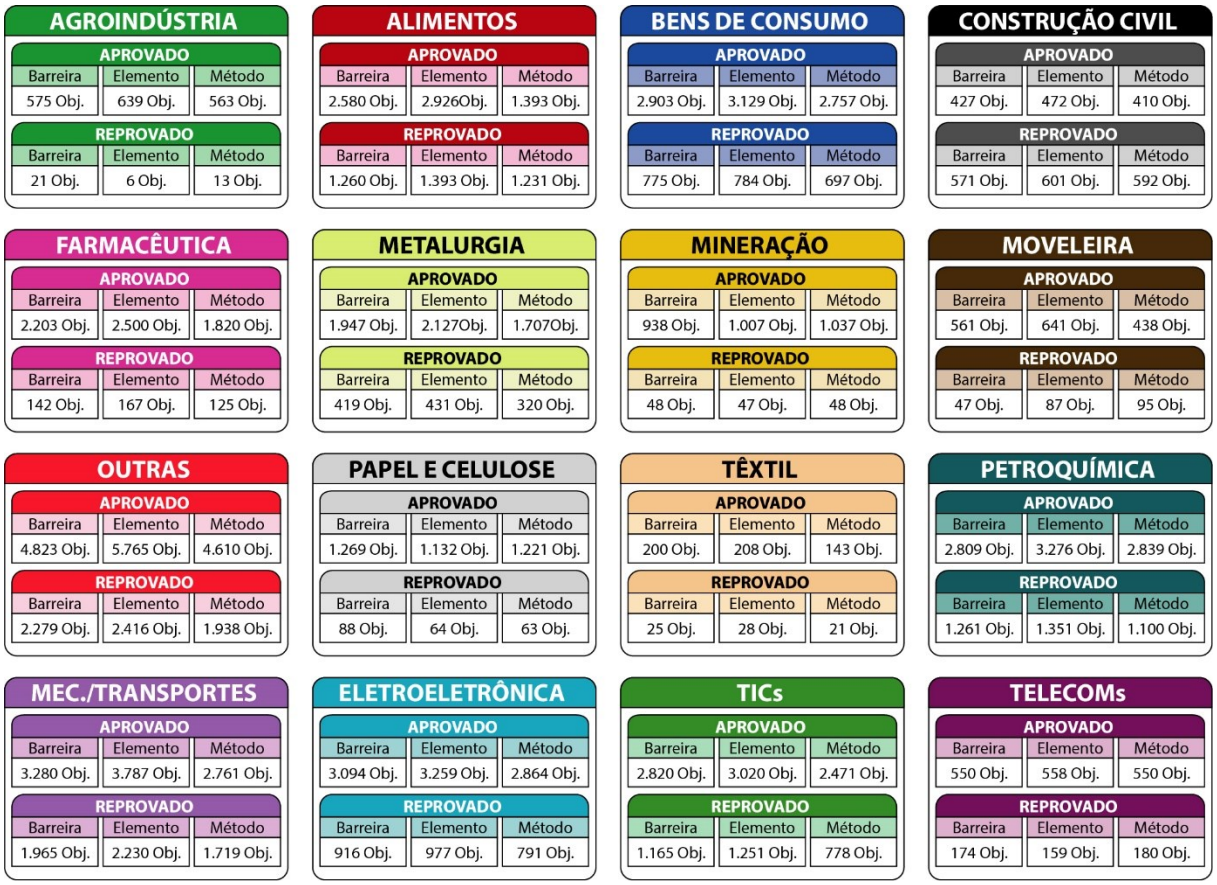
5.1 Passo 1: identificar entidades de contexto

O primeiro passo para transformar o ambiente informacional em questão é a identificação de entidades de cada contexto original. Os sujeitos ativos na configuração inicial analisam textos submetidos em 16 áreas do conhecimento. Como a classificação destes é dada por meio de vários indivíduos (pessoas naturais), aplicando-se a AIM de Kuroki Júnior (2018), o conjunto de conhecimentos expressos em cada área pode ser considerado um sujeito, obtendo-se, por conseguinte, 16 sujeitos.

De forma reflexa, o corpus de objetos também é definido por esta distinção de sujeitos, dado que há um acordo semântico entre as pessoas que analisam os textos em cada área (são

especialistas. A diferença reside no fato de que cada área de conhecimento possui duas partições de valor binário — Aprovado ou Reprovado — dotado de 3 agrupamentos semânticos — Elemento Inovador, Barreira Tecnológica e Método — totalizando 96 contextos semânticos. Neste sentido, dado que objetos são expressos por meio de atributos, tão somente substantivos são elegíveis como entidades, dada sua capacidade de absorção de atributos por meio de outros termos semânticos que os modifica. A Figura 3 demonstra os quantitativos obtidos por contexto para o ano de 2015.

Figura 3. Objetos identificados por contexto — Ano-base 2015



Fonte: Produzido pelos autores em agosto de 2022.

5.2 Passo 2: identificar correlações entre entes

A segunda fase para produção de uma AIM é a identificação de correlações entre os sujeitos e objetos do domínio. Neste sentido, uma técnica denominada Frequência Inversa em Documentos (FID), originariamente proposta por Jones (1973) fora utilizada. Trata-se de uma medida logarítmica de relevância de um termo perante um conjunto de documentos: quão menor a incidência de determinada palavra em um texto, maior é a probabilidade de sua relevância. A seleção de entidades do modelo deve garantir a manutenção a relação de relevância do ente potencial perante o contexto original não tratado. Neste sentido, 5 etapas de análise são propostas:

a) Cálculo do FID de cada ente perante cada um dos 96 domínios semânticos;
b) Obtenção da média FID de cada ente considerando a soma dos valores dos 96 domínios semânticos;
c) Seleção dos entes cuja média FID (calculado no passo anterior) seja maior que o

desvio-padrão considerando todas as médias FID;
d) Identificação de objetos por meio de Definição, Comparação, Fusão e Decomposição.

Para o ano de 2015 foram identificadas 21.142 entidades potenciais. Ao aplicar o sequenciamento dos passos "a", "b" e "c", este número decai para 513. Dentre as entidades potenciais, foram identificados os conjuntos semânticos [método, metodologia], [fabricação, produção], [necessário, necessidade], [produtivo, produtividade], [fim, final, resultado], [sistema, software], os quais apresentam potencial similaridade. Os atributos de tais pares foram analisados por meio de **comparação**, a fim de verificar a necessidade de **definição** de dois termos ou de **fusão** em apenas um termo. Os resultados das relações potenciais em questão foram:

- Conjunto semântico [método, metodologia]: percentual de semelhança entre atributos de 1,69%. Relação de **definição**;
- Conjunto semântico [fabricação, produção]: percentual de semelhança entre atributos de 1,85%. Relação de **definição**;
- Conjunto semântico [necessário, necessidade]: percentual de semelhança entre atributos de 5,26%. Relação de **definição**;
- Conjunto semântico [produtivo, produtividade]: percentual de semelhança entre atributos de 8,47%. Relação de **definição**;
- Conjunto semântico [fim, final, resultado]: percentual de semelhança entre atributos de 4,81%. Relação de **definição**;
- Conjunto semântico [sistema, software]: percentual de semelhança entre atributos de 1,96%. Relação de **definição**;

Destarte, as 513 entidades potenciais obtidas ao longo das quatro etapas de seleção são reconhecidas e correlacionadas como objetos do domínio.

5.3 Passo 3: distinguir domínios

Identificados os 16 sujeitos e 513 objetos atuantes no domínio original, passa-se a alteração da configuração deste espaço informacional por meio de descrição, inspeção ou verificação. Dado que o percurso para se obter esta configuração iniciou-se da análise de um conjunto de textos por pessoas naturais, o procedimento de **verificação** torna-se a escolha mais assertiva para a distinção dos domínios. O procedimento é dotado de 3 passos:

a) Inquirir um grupo de sujeitos;
b) Identificar atributos comuns;
c) Agrupamentos de objetos que possuam tais atributos.

A primeira etapa fora realizada de forma prévia a aplicação da AIM, quando da análise dos textos por pessoas naturais, ou seja, fora realizada ao se obter o conjunto de dados original classificado por área de conhecimento e aprovação/reprovação de cada texto individualmente. A segunda etapa fora realizada no item Identificação de correlações entre entes, onde foram obtidos os 513 objetos reconhecidos pelos 16 sujeitos do contexto inicial. Para a realização da terceira etapa, quatro procedimentos foram realizados:

a) Cálculo da relevância dos objetos para cada um dos 16 sujeitos: cada área possui duas classificações de mérito (aprovado ou reprovado) para três contextos

semânticos (Elemento Inovador, Barreira Tecnológica e Método), totalizando seis parâmetros de análise. Soma-se os valores FID de cada objeto nos seis parâmetros de análise, obtendo-se o valor de relevância do objeto para cada um dos 16 sujeitos. Este valor representa o quão relevante é cada objeto para os sujeitos;

b) Índice de adesão do sujeito ao ambiente: dotado do valor de relevância dos objetos, a soma destes valores representa o quão aderente está o escopo de conhecimento do sujeito ao contexto analisado;
c) Obtenção do índice de dispersão do contexto informacional: por meio do cálculo do desvio-padrão dos índices de adesão calculados no procedimento anterior verifica-se o quão uniforme é o ambiente informacional.
d) Concepção de domínios com base no índice de dispersão do ambiente informacional: quão maior o índice de dispersão, maior a quantidade de agrupamentos, observando-se a necessidade de compensação entre os índices de adesão dos sujeitos ao ambiente.

O índice de dispersão calculado com base no passo "c" para o ano de 2015 fora de 562,38, o que divide o espectro de valores da Tabela 7 em 4 faixas:

- 0 a 562,38: composto pelos sujeitos Metalurgia, Farmacêutica, Papel e Celulose, Mineração, Moveleira, Construção Civil, Agroindústria, Telecomunicações e Têxtil;
- 562,39 a 1.124,76: composto pelos sujeitos Petroquímica, Bens de Consumo, TICs, Alimentos e Eletroeletrônica;
- 1.124,77 a 1.687,14: composto pelo sujeito Mecânica e Transporte;
- 1.687,15 a 2.249,52: composto pelo sujeito Outros.

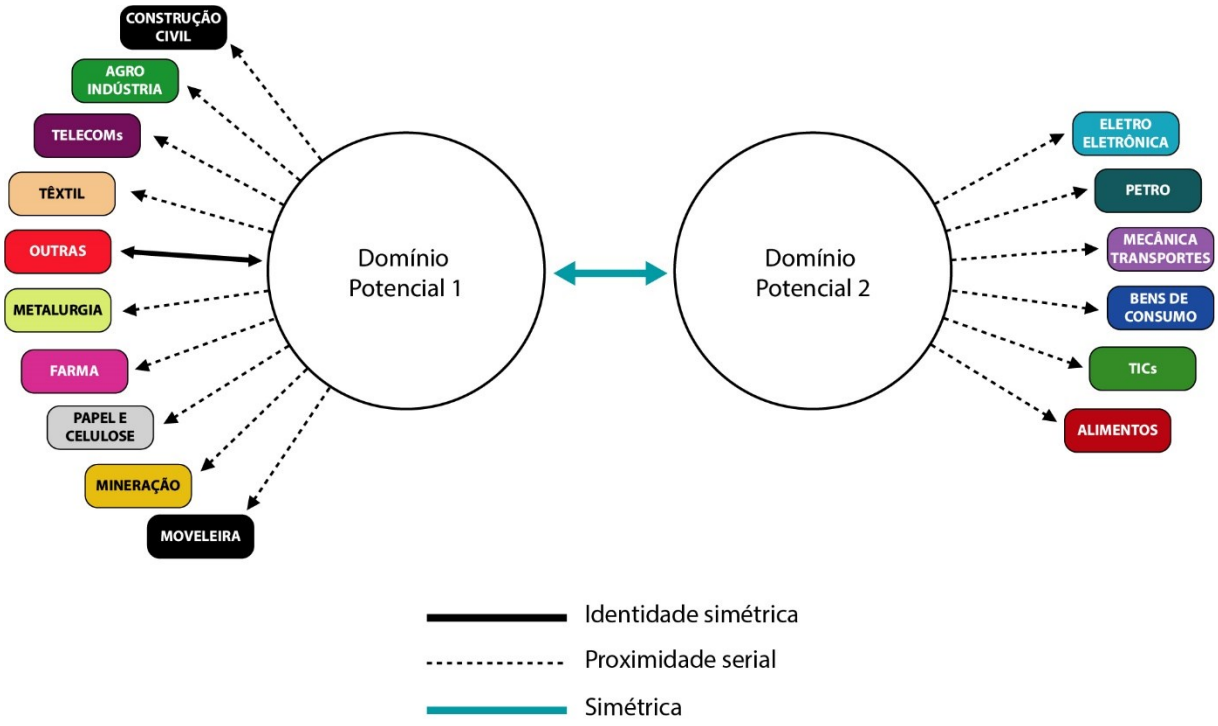
O menor nível de distinção/agregamento possível no referido contexto informacional, defeso considerar a totalidade dos 16 sujeitos, é a divisão em dois domínios. Tal divisão deve considerar um equilíbrio no índice de adesão do sujeito ao contexto informacional. Neste sentido, os agrupamentos [1, 4] e [2, 3] apresentam-se como os mais equilibrados, dando origem a:

- Domínio potencial 1, composto pelos sujeitos Metalurgia, Farmacêutica, Papel e Celulose, Mineração, Moveleira, Construção Civil, Agroindústria, Telecomunicações, Têxtil e Outros;
- Domínio potencial 2, composto pelos sujeitos Petroquímica, Bens de Consumo, TICs, Alimentos, Eletroeletrônica e Mecânica e Transporte.

5.4 Passo 4: identificar relações entre domínios

Dotados de dois domínios potenciais encontrados na etapa anterior, passa-se ao estabelecimento de relações entre as áreas de conhecimento e estes domínios, bem como entre os domínios em si. Neste sentido, a Figura 4 demonstra as relações de identidade e proximidade que originaram os domínios potenciais, bem como a extensão das relações entre estes domínios.

Figura 4. Relações entre áreas de conhecimento e domínios potenciais — Ano-base 2015



Fonte: Produzido pelos autores em agosto de 2022.

Observa-se que, em sua formação, tão somente o Domínio Potencial 1 possui uma relação de Identidade Simétrica, dado que a área de conhecimento "Outros" é a única que possui todos os objetos presentes no domínio. Todas as relações identificadas para formação dos domínios potenciais 1 e 2 são reflexivas, uma vez que tal operação parte da identificação de objetos comuns o que, necessariamente, requer a verificação de existência deste objeto no próprio domínio para só então proceder a verificação de existência do referido objeto em outro domínio.

No tocante às relações entre os domínios potenciais, verifica-se uma única relação simétrica [1,2], dado que todos os objetos podem ser encontrados em qualquer configuração possível de ambos os domínios, o que demonstra que ambos coexistem de forma independente sendo micro organizações do contexto informacional original.

5.5 Passo 5: agrupamento por contextos espaço-tempo

Conforme descrito no item *4.2 Identificação de correlações entre entes*, foram utilizados os dados do ano-base de 2015 para se conceber a distribuição de domínios obtida no item *4.3 Distinção de domínios*. De forma a verificar extensão temporal da alteração da arquitetura proposta ao longo dos anos, realizou-se o ciclo da AIM exposto na figura 3, juntamente com os procedimentos descritos ao longo dos itens 4.1 a 4.4 para o ano-base de 2014, obtendo-se configuração distinta de domínios.

Para o passo de identificação de correlações entre entes, o número de entidades potenciais passa a ser 480 em 2014, em detrimento das 513 obtidas em 2015. O índice de dispersão do contexto informacional para o ano de 2014 fora de 798,84. Tal alteração resultou em uma agregação de sujeitos ligeiramente distinta do ano de 2015:

- 0 a 798,84: composto pelos sujeitos Metalurgia, Farmacêutica, Papel e Celulose, Mineração, Moveleira, Construção Civil, Agroindústria, Telecomunicações e Têxtil;
- 798,85 a 1.597,68: composto pelos sujeitos Petroquímica, Bens de Consumo, TICs, Alimentos e Eletroeletrônica;
- 2.396,53 a 3.195,37: composto pelos sujeitos Mecânica e Transporte e Outros;

As três mudanças mais significativas são: a separação dos sujeitos Mecânica e Transporte e Outros em duas faixas distintas; reclassificação do sujeito Tecnologia da Informação e Comunicações para faixa abaixo do índice de dispersão do contexto; e a reordenação das faixas de agregação. Apesar das mudanças serem aparentemente desprezíveis, há que se considerar o equilíbrio entre os índices de adesão dos sujeitos. Neste sentido, são propostos 3 domínios potenciais para o ano de 2014:

- Domínio potencial 3, composto pelo sujeito Mecânica e Transporte e de parte dos sujeitos que compõem a primeira faixa de agregação do contexto original para o ano de 2014, a saber: Agroindústria, Moveleira, Papel e Celulose, Farmacêutica e TICs;
- Domínio potencial 4, composto pelo sujeito Outros e a parte restante dos sujeitos que compõem a primeira faixa de agregação do contexto original para o ano de 2014, a saber: Têxtil, Telecomunicações, Construção Civil, Mineração e Metalurgia;
- Domínio potencial 5, composto pela totalidade dos sujeitos que compõem a segunda faixa de agregação, a saber: Química e Petroquímica, Bens de Consumo, Eletroeletrônica e Alimentos.

Verifica-se a alta sensibilidade do problema a separação espaço-temporal: uma AIM utilizada em um ano não pode ser tomada, de início, como aplicável a um novo contexto temporal. Confirma-se tal premissa quando se procede a análise dos dados de 2014 e 2015 em conjunto. O número de entidades potenciais identificados é de 1.192. O índice de dispersão do contexto informacional elevou para 10.243,65, criando 3 domínios diferentes dos identificados anteriormente:

- Domínio potencial 6, composto pelos sujeitos Mecânica e Transporte, Telecomunicações, Construção Civil, Papel e Celulose, Farmacêutica e Metalurgia;
- Domínio potencial 7, composto pelos sujeitos Outros, Têxtil, Agroindústria, Moveleira, Mineração e Bens de Consumo;
- Domínio potencial 8, composto pelos sujeitos Química e Petroquímica, Alimentos, TICs e Eletroeletrônica.

6 APLICAÇÃO DE PLN COM DADOS PRÉ-TRATADOS POR AIM

Identificada a impossibilidade de se proceder a produção de um modelo preditivo para o problema selecionado com base na seleção indistinta de dados e, dotado dos produtos de AIM obtidos ao longo dos passos de identificação de entidades de contextos até o agrupamento por contextos espaço-tempo, proceder-se-á a validação do modelo obtido. Para tal intento, os dados de 2014 e 2015 foram divididos e concatenados conforme os domínios potenciais construídos e treinados por 10 vezes, mantendo-se as condições de treinamento descritas no item 5.1

Aplicação de PLN em um conjunto de dados não tratados por AIM. Os resultados obtidos são apresentados na Tabela 2.

Tabela 2. Média de resultados dos experimentos realizados com dados tratados por AIM

Domínio potencial	Perda em treinamento	Acurácia em treinamento	Perda em validação	Acurácia em validação	Perda em teste	Acurácia em teste
Domínio Potencial 1 (2015)	0,5296761	78,43%	0,5286451	80,19%	0,4946408	84,88%
Domínio Potencial 2 (2015)	0,5505137	75,77%	0,5717295	72,78%	0,5767502	72,65%
Domínio Potencial 3 (2014)	0,7006512	55,88%	0,6701859	58,00%	0,6577451	58,70%
Domínio Potencial 4 (2014)	0,7183891	55,41%	0,7043763	54,85%	0,6313299	54,98%
Domínio Potencial 5 (2014)	0,7111632	51,85%	0,6945764	52,65%	0,7277233	52,80%
Domínio Potencial 6 (2014 e 2015)	0,6629338	63,30%	0,6571880	63,40%	0,5833146	63,94%
Domínio Potencial 7 (2014 e 2015)	0,6799421	59,75%	0,6634957	56,19%	0,6887856	55,15%
Domínio Potencial 8 (2014 e 2015)	0,6265331	67,11%	0,6602471	63,38%	0,6573988	61,58%

Fonte: Produzidos pelos autores em outubro de 2022

7 DISCUSSÃO DE RESULTADOS

Verifica-se variação nos valores de perda e acurácia após o tratamento do conjunto informacional original e sua separação em domínios de relevância. Alguns domínios apresentam melhora na acurácia de predição, outros apresentam uma piora na acurácia da predição.

O ano de 2015, utilizado como base de explanação dos procedimentos propostos no item 5, teve seu conjunto de dados divididos em 2 domínios potenciais. Os resultados iniciais se apresentaram como os mais assertivos do contexto não-tratado. A Tabela 3 apresenta a comparação entre as médias valores de perda e acurácia para o conjunto de dados do ano.

Tabela 3. Média de resultados dos experimentos realizados com dados não tratados

Variável	2015	2015 – Domínio 1	2015 – Domínio 2
Perda em treinamento	0,5627345	0,5296761	0,5505137
Acurácia em treinamento	76,86%	78,43%	75,77%
Perda em validação	0,5708822	0,5286451	0,5717295
Acurácia em validação	74,14%	80,19%	72,78%

Perda em teste	0,4740491	0,4946408	0,5767502
Acurácia em teste	77,57%	84,88%	72,65%

Fonte: Produzidos pelos autores em Agosto de 2022

Percebe-se que o domínio potencial 1 apresentou um ganho de 7,31% de acurácia em teste em oposição a perda de 4,92% assinalada para o domínio potencial 2. O potencial de aprendizado segue as mesmas tendências em ambos os domínios, apontando que há uma melhora no desempenho da rede de PLN quando utilizado o subconjunto de dados do domínio 1 e uma piora para o domínio 2. Partindo do mesmo conjunto de dados, o pré-tratamento de AIM identificou subdivisões que possuem maior e menor capacidade de extração de aprendizado, demonstrado por meio da variação de acurácia e perda nos dois conjuntos.

Seguindo na validação da AIM, o quesito temporal fora endereçado por meio da execução de experimentos baseados no pré-tratamento de dados provenientes do ano de 2014, bem como da junção dos dados de 2014 e 2015. A Tabela 4 apresenta a comparação de resultados para o ano de 2014.

Tabela 4. Comparação de resultados – Ano-base 2014

Variável	2014	2014 – Domínio 3	2014 – Domínio 4	2014 – Domínio 5
Perda em treinamento	0,7087808	0,7006512	0,7183891	0,7111632
Acurácia em treinamento	53,55%	55,88%	55,41%	51,85%
Perda em validação	0,6949488	0,6701859	0,7043763	0,6945764
Acurácia em validação	54,52%	58,00%	54,85%	52,65%
Perda em teste	0,7416452	0,6577451	0,6313299	0,7277233
Acurácia em teste	54,79%	58,70%	54,98%	52,80%

Fonte: Produzidos pelos autores em outubro de 2022

Ao longo do procedimento de construção da AIM para o ano de 2014, observa-se uma redução na quantidade de entidades potenciais em comparação ao ano de 2015 (513 para 480) e um aumento no índice de dispersão do contexto informacional (de 562,38 para 798,84). Tais números levam as seguintes considerações que norteiam a análise:

a) Os sujeitos que atuaram no contexto informacional de 2014 reconheceram menos entidades como objetos relevantes, com uma grande variação em seus índices de adesão ao ambiente, ou seja, há sujeitos que tem uma alta aderência ao contexto (os objetos que ele reconhece constam, em sua maioria, no contexto informacional relevante), e outros que possuem uma baixa aderência ao contexto (seus objetos reconhecidos, em sua maioria, não constam no contexto informacional relevante).
b) O contexto informacional relevante a ser tratado fora mais disperso, necessitando de mais subdivisões do contexto original, passando de 2 domínios para 3.

O domínio 3 apresentou melhora de 4,01% nos níveis de acurácia em teste, um ganho menor do que o registrado para o domínio 1 do ano de 2015. O domínio 4 manteve-se praticamente inalterado em relação ao contexto original, apresentando a discreta melhora de 0,19% de acurácia em teste. O domínio 5, por sua vez, apresentou queda de 1.99% nos níveis de acurácia em teste.

Observa-se o reflexo da alta dispersão de dados e a baixa adesão dos sujeitos ao contexto relevante: o conjunto de dados com a melhor predisposição a aprendizado se torna menor e, ainda assim, com um ganho pouco expressivo.

Outra situação analisada nos experimentos fora a junção dos dados de 2014 e 2015. A comparação entre os resultados sem pré-tratamento e com pré-tratamento se encontram na Tabela 5.

Tabela 5. Comparação de resultados – Anos-base 2014 e 2015 em conjunto

Variável	2014/2015	2014/2015 – Domínio 6	2014/2015 – Domínio 7	2014/2015 – Domínio 8
Perda em treinamento	0,6463273	0,6629338	0,6799421	0,6265331
Acurácia em treinamento	63,39%	63,30%	59,75%	67,11%
Perda em validação	0,6765008	0,6571880	0,6634957	0,6602471
Acurácia em validação	59,08%	63,40%	56,19%	63,38%
Perda em teste	0,6142412	0,5833146	0,6887856	0,6573988
Acurácia em teste	58,22%	63,94%	55,15%	61,58%

Fonte: Produzidos pelos autores em outubro de 2022

Conforme apresentado no item 5.5 Passo 5: agrupamento por contextos espaço-tempo, o número de entidades potenciais identificadas fora de 1.192, entretanto, o índice de dispersão do contexto informacional se elevou para 10.243,65. Novamente, temos um descompasso entre o quão aderente é o conhecimento dos sujeitos ao contexto informacional relevante. O ganho de acurácia em teste fora observado nos domínios 6 (5.72%) e 8 (3.36%) ao passo que para o domínio 7 fora observado um decréscimo de 3,07%.

Dos 8 domínios propostos, tomando como parâmetro de análise os resultados de acurácia em teste, 4 (quatro) apresentaram ganho, 3 (três) apresentaram perda e 1 (um) manteve os patamares anteriores, com pequeno acréscimo. É possível, partindo desta análise, proceder a identificação dos conjuntos de dados que possuem maior e menor potencial de extração de aprendizado.

Tabela 6. Análise de potencial de aprendizado por área do conhecimento

Área do conhecimento	2014	2015	2014/2015	Potencial
Agroindústria	1	1	-1	1
Alimentos	-1	-1	1	-1
Bens de Consumo	-1	-1	-1	-3
Construção Civil	0	1	1	2
Eletroeletrônica	-1	-1	1	-1
Farmacêutica	1	1	1	3
Mecânica e Transporte	1	-1	1	1
Metalúrgica	0	1	1	2
Mineração	0	1	-1	0
Movelaria	1	1	-1	1

Papel e Celulose	1	1	1	3
Química e Petroquímica	-1	-1	1	-1
TICs	1	-1	1	1
Telecomunicações	0	1	1	2
Têxtil	0	1	-1	0
Outros	0	1	-1	0

Fonte: Produzidos pelos autores em outubro de 2022

8 CONCLUSÃO

Por meio deste artigo, visou-se posicionar a Ciência da Informação como parte integrante do processo de construção de inteligências artificiais, figurando como disciplina anterior à formalização de algoritmos de redes neurais. O pré-tratamento de dados fornecido por meio de AIM pode contribuir no aumento da acurácia de predições realizando tão somente um rearranjo dos dados fornecidos, ou seja, impondo um senso de organização dinâmica conforme o espaço-tempo tratado.

Na seção 3 identificou-se que o atual estágio de desenvolvimento do PLN fornece uma diversa gama de implementações algorítmicas, entretanto, as técnicas de treinamento mais utilizadas (como o aprendizado supervisionado) ainda requerem grande volume de dados classificados e melhorias em modelos de conhecimentos específicos ou de senso comum (voltados a perguntas sobre o mundo real) e com incompletude de informações.

Na seção 4 apresentou-se a AIM e seu tratamento de *Modos* de expressão de significados, seguindo Kress e Van Leeuween (2001) e Kress (2009); por meio de estruturas lógicas modais, conforme Carnielli e Pizzi (2008) e Portner (2009), Da junção das duas correntes de pensamento, torna-se possível o tratamento de semânticas distintas em um mesmo contexto informacional, problema este muito comum em tarefas de PLN. O enfretamento da questão por parte da AIM baseia-se, dentre outros princípios, em economia e relevância para fornecer a melhor configuração informacional possível. Utiliza-se de um procedimento de 5 passos para identificação de sujeitos e suas correlações com objetos, bem como os domínios os quais sujeitos e objetos pertencem e as relações entre estes domínios.

Na seção 5 o procedimento de construção de produtos de AIM é aplicado a um problema real de classificação de textos advindos de 16 áreas de conhecimento. Oito subdomínios foram concebidos sem qualquer alteração no quantitativo de dados original. Por meio de um algoritmo amplamente utilizado de PLN para a língua portuguesa do Brasil, os resultados obtidos a partir de dados tratados por AIM foram comparados aos obtidos sem tal tratamento.

Ainda que os valores observados tenham sido numericamente discretos sob o ponto de vista de acurácia de predição, percebe-se potencial de melhoria em grande parte dos domínios distinguidos. Considerando que nenhum procedimento de enriquecimento de dados ou aprimoramento do modelo linguístico fora realizado, é plausível a conclusão de que a AIM, por si só, indicou o melhor agrupamento de dados possível em cada momento temporal partindo-se tão somente dos registros apresentados inicialmente.

Por fim, observa-se que neste artigo, a escolha pela técnica FID proposta inicialmente por Jones (1973) para obtenção de correlações entre sujeitos e objetos no item 5.2 Passo 2: identificar correlações entre entes, não vincula a AIM a utilização da mesma, podendo ser substituída por qualquer outra técnica que forneça uma medida de relevância de objetos para cada sujeito. A investigação de outros métodos de obtenção do referido nível de relevância é encorajada.

REFERÊNCIAS

AREL, I; ROSE, D. C.; KARNOWSKI, T. P. Deep machine learning-a new frontier in artificial intelligence research [research frontier]. **IEEE computational intelligence magazine**, *[S.l.]* v. 5, n. 4, p. 13-18, 2010. DOI: 10.1109/MCI.2010.938364. Acesso em: 9 de jan. 2023.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 3, 2015, San Diego, CA. **Anals** [...]. San Diego, CA, 2015. DOI: 10.48550/arXiv.1409.0473. Acesso em: 9 de jan. 2023.

BELLMAN, R. The theory of dynamic programming. **Bulletin of the American Mathematical Society**, Providence, RI, v. 60, n. 6, p. 503-515, 1954. DOI: 10.1090/S0002-9904-1954-09848-8. Acesso em: 9 de jan. 2023.

CAPURRO, R.; HJORLAND, B. O conceito de informação. **Perspectivas em ciência da informação**, Belo Horizionte, v. 12, n. 1, p. 148–207, 2007. DOI: 10.1590/S1413-99362007000100012. Acesso em: 9 de jan. 2023.

CARNIELLI, W.; PIZZI, C. **Modalities and multimodalities**. Springer Science & Business Media, 2008. 304p.

JONES, K. S. Index term weighting. **Information storage and retrieval**, Cambridge, UK, v. 9, n. 11, p. 619-633, 1973. DOI: 10.1016/0020-0271(73)90043-0. Acesso em: 9 de jan. 2023.

HJØRLAND, B. What is knowledge organization (ko)? Knowledge organization. **International journal devoted to concept theory, classification, indexing and knowledge representation**, *[S.l.]*, ERGON-Verlag GmbH, 2008. Disponível em: http://bit.ly/3vQG7Ry. Acesso em: 9 de jan. 2023.

HINTON, G. E.; OSINDERO, S.; TEH, Y. A fast learning algorithm for deep belief nets. **Neural computation**, Boston, MA, v. 18, n. 7, p. 1527-1554, 2006. DOI: 10.1162/neco.2006.18.7.1527. Acesso em: 9 de jan. 2023.

KRESS, G. What is mode? *In*: Jewitt, C. (ed.). **The Routledge Handbook of Multimodal Analysis**. London, UK, Routledge, 2009. 340 p.

KRESS, G.; VAN LEEUWEN, T. **Multimodal discourse**: The modes and media of contemporary communication. London: Hodder Arnold Publication, 2001. 142 p.

KUROKI JÚNIOR, G. H. **Sobre uma arquitetura da informação multimodal**: reflexões sobre uma proposta epistemológica. Dissertação (Mestrado) — Universidade de Brasília, 2018. DOI: 10.26512/2018.02.D.31920. Acesso em: 9 de jan. 2023.

DEVLIN, J.D. *et al.* Pre-training of deep bidirectional transformers for language understanding**.** *In*: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, Minneapolis, MI. **Proceedings** […]. Minneapolis, MI, 2019. DOI: 10.48550/arXiv.1810.04805. Acesso em: 9 de jan. 2023.

MCCANN, B. *et al*. Learned in translation: Contextualized word vectors. *In:* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 30, Long Beach, CA, **Proceedings** […]. Boston, MA, 2017. DOI: 10.48550/arXiv.1708.00107. Acesso em: 9 de jan. 2023.

MINAEE, S. *et al*. Deep learning-based text classification: a comprehensive review. **ACM Computing Surveys (CSUR)**, *[S.l.]*, v. 54, n. 3, p. 1-40, 2021. DOI: 10.1145/3439726. Acessado em: 9 de jan. 2023.

MIKOLOV, T. *et al*. Efficient estimation of word representations in vector space. *[S.l.]*, **arXiv preprint**, arXiv:1301.3781. DOI: 10.48550/arXiv.1301.3781, 2013a. Acessado em: 9 de jan. 2023.

MIKOLOV, T. *et al.* Distributed representations of words and phrases and their compositionality. *In:* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NIPS 2013), Lake Tahoe, NV, **Proceedings** […]. Boston, MA, 2013b. v. 26. DOI: 10.48550/arXiv.1310.4546. Acesso em: 9 de jan. 2023.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. *In:* CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP), Doha, Qatar. **Proceedings** […]. DOI: 10.3115/v1/D14-1162, 2014. Acesso em: 9 de jan. 2023.

PETERS, M. E. *et al*. Deep contextualized word representations. *In:* NAACL. ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 1 (long papers), New Orleans, Louisiana, 2018. **Proceedings** […]. DOI: 10.18653/v1/N18-1202 2018. Acesso em: 9 de Jan. 2023.

PORTNER, P. **Modality**. London: Oxford University Press, 2009. 320 p.

QIU, X. *et al*. Pre-trained models for natural language processing: A survey. **Science China Technological Sciences**, *[S.l.],* v. 63, n. 10, p. 1872-1897, 2020. DOI: 10.1007/s11431-020-1647-3. Acesso em: 9 de jan. 2023.

RADFORD, A. *et al*. Improving language understanding by generative pre-training. **OpenAI**, *[S.l.]*. v. 4, n. 19, 2018. Disponível em: http://bit.ly/3Xhzao1. Acesso em: 9 de Jan. 2023.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. *In:* CERRI, R., PRATI, R.C. (ed.) **Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science**, *[S.l.],* v. 12319. Springer, Cham. p. 403-417. DOI: 10.1007/978-3-030-61377-8_28. Acesso em: 9 de jan. 2023.

VAN GIGCH, J. P.; MOIGNE, J. L. A paradigmatic approach to the discipline of information systems. **Behavioral Science**, *[S.l.],* v. 34, n. 2, p. 128-147, 1989. DOI: 10.1002/bs.3830340203. Acesso em: 9 de jan. 2023.

VASWANI, A. *et al*. Attention is all you need. **Advances in neural information processing systems**, Long Beach, CA, v. 30, p.1-15, 2017. DOI: 10.48550/arXiv.1706.03762. Disponível em: https://arXiv:1706.03762. Acesso em: 9 de jan. 2023.

WAGNER FILHO, J. A. *et al*. The brWaC corpus: a new open resource for Brazilian Portuguese. *In:* Proceedings of the eleventh international conference on language resources

and evaluation (LREC 2018), Miyazaki, Japan. **Proceedings** […]. Miyazaki, Japan: ELRA, 2018. Disponível em: http://bit.ly/3CBRzUR. Acesso em: 9 de Jan. 2023.

WASON, R. Deep learning: Evolution and expansion. **Cognitive Systems Research**, *[S.l.],* v. 52, p. 701-708, 2018. DOI: 10.1016/j.cogsys.2018.08.023. Acesso em: 9 de Jan. 2023.