RICI Revista Ibero-Americana de Ciência da Informação

ISSN 1983-5213

Processamento de linguagem natural e *machine learning* na categorização de artigos científicos: um estudo em torno do "patrimônio cultural"

Ananda Fernanda de Jesus
Universidade Estadual Paulista, Programa de Pós-Graduação em Ciência da Informação, Marília, SP, Brasil
af.jesus@unesp.br

Maria Lígia Triques
Universidade Estadual de Londrina, Programa de Pós-Graduação em Ciência da Informação, Londrina, PR, Brasil
mligia.triques@uel.br

José Eduardo Santarem Segundo
Universidade Estadual Paulista, Programa de Pós-Graduação em Ciência da Informação, Marília, SP, Brasil
santarem@usp.br

Ana Cristina de Albuquerque
Universidade Estadual de Londrina, Departamento de Ciência da Informação, Londrina, PR, Brasil
albuanati@uel.br

ARTIGOS

DOI: https://doi.org/10.26512/rici.v16.n1.2023.47537

Recebido/Recibido/Received: 2022-12-09
Aceitado/Aceptado/Accepted: 2023-03-10

Resumo
Objetiva verificar o potencial de aplicação de técnicas de Processamento de Linguagem Natural (PLN) e de *Machine Learning* (ML) na categorização temática de artigos científicos sobre a temática "patrimônio cultural" a partir de duas situações em que categorias são estabelecidas *a priori* e *a posteriori*. Desenvolve-se uma pesquisa aplicada, com resultados quantitativos e qualitativos. O primeiro *corpus* é constituído de artigos científicos em português, em base temática da Ciência da Informação, selecionados e categorizados manualmente; e o segundo *corpus,* composto por artigos científicos em inglês recuperados na *Web of Science*, categorizados de forma automática por estratégias de busca e aplicação de booleanos. Ambos foram submetidos à dois procedimentos de teste de categorização (algoritmo supervisionado e não supervisionado).Os resultados demonstram que em ambas a participação do pesquisador é essencial na definição da representatividade da amostra escolhida, e que esta tem impacto direto na precisão e acurácia dos algoritmos aplicados. Destaca-se a importância do detalhamento e rigor no pré-processamento dos dados e do tamanho da amostra, contudo, ressalta-se que, no caso deste estudo, somente um volume maior de dados não garantiu que os resultados fossem representativos do ponto de vista do domínio estudado, o que alerta para que haja sempre discussões e análises multidisciplinares que permitam verificar e readequar os parâmetros da amostra.
Palavras-chave: Aprendizado de máquina. Processamento de linguagem natural. Algoritmo de rede neural. Algoritmo de clusterização hierárquica. Patrimônio cultural.

Natural language processing and machine learning in the categorization of scientific papers: a study around "cultural heritage"

Abstract

Aims to verify the potential of applying Natural Language Processing (NLP) and Machine Learning (ML) techniques in the thematic categorization of scientific articles on the theme "cultural heritage" from two situations in which categories are established a priori and later. Applied research is developed, with quantitative and qualitative results, where the first corpus consisting of scientific articles in Portuguese, on a thematic basis of Information Science, manually selected and categorized; and the second corpus, composed of scientific articles in English retrieved from the Web of Science, automatically categorized by search strategies and application of Booleans. Both were submitted to two categorization test procedures (supervised and unsupervised algorithm). The results show that in both, the participation of the researcher is essential in defining the representativeness of the chosen sample, and this has an impact on the precision and accuracy of the applied algorithms. The importance of detailing and rigor in the pre-processing of data and sample size is highlighted, however, it is emphasized that, in the case of this study, only a larger volume of data did not guarantee that the results were representative from the point of view of the domain studied, which warns that there are always multidisciplinary discussions and analyzes that allow verifying and readjusting the sample parameters.

Keywords: Machine learning. Natural language processing. Neural network algorithm. Hierarchical clustering algorithm; Cultural heritage.

Procesamiento del lenguaje natural y aprendizaje automático en la categorización de artículos científicos: un estudio en torno al "patrimonio cultural"

Resumen

Objetiva verificar el potencial de aplicar técnicas de Procesamiento del Lenguaje Natural (PNL) y Aprendizaje Automático (ML) en la categorización temática de artículos científicos sobre el tema "patrimonio cultural" a partir de dos situaciones en las que se establecen categorías a priori y posteriormente. Se desarrolla una investigación aplicada, con resultados cuantitativos y cualitativos, donde el primer corpus consiste en artículos científicos en portugués, sobre una base temática de Ciencias de la Información, seleccionados y categorizados manualmente; y el segundo corpus, compuesto por artículos científicos en inglés recuperados de la Web of Science, categorizados automáticamente por estrategias de búsqueda y aplicación de booleanos. Ambos fueron sometidos a dos procedimientos de prueba de categorización (algoritmo supervisado y no supervisado). Los resultados muestran que en ambos enfoques la participación del investigador es fundamental para definir la representatividad de la muestra elegida, y que esta tiene un impacto directo en la precisión y exactitud de los algoritmos aplicados. Se destaca la importancia del detalle y rigor en el preprocesamiento de los datos y el tamaño de la muestra, sin embargo, se destaca que, en el caso de este estudio, solo un mayor volumen de datos no garantizaba que los resultados fueran representativos desde el punto de vista de vista del dominio estudiado, lo que advierte que siempre hay discusiones y análisis multidisciplinarios que permiten verificar y reajustar los parámetros de la muestra.

Palabras clave: Aprendizaje automático. Procesamiento natural del linguaje. Algoritmo de red neuronal. Algoritmo de agrupamiento jerárquico. Patrimonio cultural.

1 Introdução

Categorizar é um ato presente em diversas atividades cotidianas, desde a criação de categorias para organização dos espaços pessoais, das agendas de trabalho, até a estruturação das cidades e dos países. A busca pela criação de categorias perpassa ainda o desenvolvimento e estabelecimento da Ciência da Informação, tendo em vista sua preocupação com a representação, "[...] origem, coleção, organização, armazenamento, recuperação, interpretação, transmissão, transformação, e utilização da informação." (BORKO, 1968, p.1). Destaca-se a necessidade de elaboração de categorias nas atividades de representação da

informação e do conhecimento, visando sua posterior recuperação, tais como a catalogação, classificação e indexação.

Categorizar também é um ato importante para o desenvolvimento de pesquisas científicas, nas quais os resultados de estudos teóricos ou aplicados precisam ser agrupados e, assim, compreendidos enquanto um conjunto que possibilita a identificação de padrões e exceções, permitindo a geração de inferências. As categorias desses resultados podem ser estabelecidas *a priori*, ou seja, os resultados são agrupados em categorias criadas antes de sua análise, ou *a posteriori*, quando as categorias são elaboradas com base em padrões identificados nos resultados obtidos.

Dado um corpus de análise, a categorização pode fornecer, portanto, uma série de informações acerca do universo estudado, uma vez que ajuda a estabelecer a definição e os limites do domínio em questão. A categorização auxilia na descrição de características de um objeto de estudo com o qual pesquisador já tem certa familiaridade - o que indicaria um estudo descritivo - assim já sendo possível estabelecer categorias *a priori* que servem para caracterizar o objeto estudado. Atua também no processo de familiarização do pesquisador com o objeto de estudo que está sendo investigado, o que resulta geralmente no estabelecimento de categorias *a posteriori*- o que indica um estudo exploratório.

O presente estudo se estrutura em torno do termo "patrimônio cultural" (no inglês: *"cultural heritage"*), que tem sido cada vez mais encontrado em publicações científicas da Ciência da Informação e áreas afins, em especial, na relação com o âmbito digital, sem que se tenham claro, no entanto, a definição, as características e os limites desse, que pode ser considerado um domínio.

Objetiva-se verificar duas possibilidades em que a categorização pode ajudar no estudo dessa temática tendo em vista o potencial de aplicação de técnicas de Processamento de Linguagem Natural (PLN) e de *Machine Learning* (ML) no processo de categorização temática de artigos científicos.

O PLN pode ser definido como "[...] um conjunto de técnicas computacionais para a análise de textos em um ou mais níveis linguísticos, com o propósito de simular o processamento humano da língua". (FERNEDA, 2003, p. 82). Já ML é pautada na construção de agentes computacionais capazes de aprender com a experiência, com base na aplicação de técnicas estatísticas, em especial, por meio de algoritmos, visando a identificação de padrões e a realização de predições. A primeira etapa do processo de ML é o treinamento, que ocorre por meio da inclusão de um *corpus* (de dados ou de recursos informacionais), que permite que o algoritmo identifique quais variáveis levam a determinado resultado. (JORDAN; MITCHELL, 2015; CONEGLIAN, 2020).

A escolha do algoritmo ou conjunto de algoritmos a ser utilizado no processo de predição é contextual, dependendo das características do *corpus* e dos objetivos da atividade a serem realizadas, destacando-se dois grandes conjuntos de algoritmos: O primeiro tipo, aprendizado supervisionado, utiliza dados para treinamento, cujo resultado é conhecido e explicitado para o algoritmo. Assim, o algoritmo conhece a solução e a partir dele e dos dados definirá quais são os aspectos que devem ser considerados para classificar algo em uma categoria. No segundo tipo, aprendizado não-supervisionado, não há um resultado ou a solução desejada previamente, sendo o treino realizado então, com padrões estatísticos nos conjuntos de dados (CONEGLIAN, 2020, p.127).

Para evidenciar e discutir tais questões, aplicou-se as técnicas de ML e PLN em duas situações distintas. A primeira se deu como uso de um *corpus* pré-selecionado e categorizado manualmente pelos pesquisadores e a segunda, com o uso de um *corpus* categorizado automaticamente, por meio da aplicação de estratégias de busca e operadores booleanos. Nesse sentido, é importante destacar que o objetivo não é comparar diretamente os dois procedimentos, tendo em vista tratar-se de amostras distintas, e sim verificar a viabilidade de cada abordagem para auxiliar no estudo de uma temática.

2 Procedimentos metodológicos

A estudo é caracterizado como aplicado, com resultados quantitativos e qualitativos, com a finalidade de verificar o potencial das técnicas de PLN e ML na categorização de resultados de um levantamento bibliográfico, tendo como recorte artigos científicos.

A pesquisa foi dividida em duas etapas principais, a de categorização utilizando artigos pré-selecionados e categorizados manualmente, como fonte para treino e teste do algoritmo (etapa 1); e a de categorização, utilizando artigos selecionados e categorizados por meio de aplicação de estratégia de busca e de operadores booleanos (etapa 2). Para cada etapa foi construído e aplicado um c*orpus*, sendo eles:

> **Corpus 1)** artigos científicos, publicados nacionalmente em língua portuguesa, em base temática da Ciência da Informação, sobre a temática 'patrimônio cultural', selecionados e categorizados manualmente;
>
> **Corpus 2)** artigos científicos publicados em inglês e recuperados na *Web of Science*, sobre a temática 'patrimônio cultural', utilizado estratégia de busca e aplicação de booleanos para categorização automática do *corpus.*

As próximas subseções detalham os procedimentos empregados em cada etapa principal da pesquisa.

2.1 Categorização utilizando corpus pré-selecionado e categorizado manualmente

Para elaboração do *Corpus* 1) recorreu-se a um levantamento bibliográfico, qualitativo e exploratório da literatura científica, delimitado à produção nacional em português, utilizando a *Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação* (BRAPCI), devido ao amplo espectro de documentos nacionais da Ciência da Informação que indexam.

A partir do uso do termo 'patrimônio cultural' na delimitação temporal de 2012 a maio 2022 (momento da coleta dos dados) como estratégia de busca, o *corpus* foi formado mediante à existência do termo em questão como descritor nas palavras-chaves das publicações e, posteriormente, identificado se havia a definição do termo em seu contexto de estudo. Ao final foram selecionados 46 artigos, cuja leitura permitiu identificar duas categorias: contexto de estudo relacionado ao meio digital (categoria A); contexto de estudo não relacionado ao meio digital (categoria B).

Esse *corpus* foi analisado tendo em vista tanto a possibilidade de sua categorização por meio de um algoritmo supervisionado, que executa suas funções utilizando as categorias estabelecidas a *priori* (A e B), como com a aplicação de um algoritmo não supervisionado, observando o potencial de sua aplicação na criação de novas categorias de análise.

Diante disso, a primeira etapa foi a de pré-processamento dos dados, que consiste em técnicas cujo objetivo é melhorar a qualidade dos dados para o posterior processamento, eliminando elementos que podem influenciar indevidamente o processo, criando resultados indesejados.

No caso deste estudo, empregou-se a divisão do conteúdo do texto em unidades menores (chamadas *tokens*), omitindo pontuação, processo conhecido como *tokenização*. Após isso, foram aplicadas opções de transformação dos dados de modo a garantir a padronização, como remoção de URLs e demais *links*, bem como uniformização em letra minúscula.

Em sequência, foram aplicados filtros, que permitem remover ou manter uma seleção de palavras, como a definição por idioma, no caso, português, dado o *corpus* de análise. Nesta fase, aplica-se principalmente o processo de identificação de *stopwords*, palavras tais como artigos e conectivos, que se repetem ao longo do texto, mas que não refletem o seu significado. Também foram excluídos os números, tais como páginas e anos, facilitando a visualização dos termos significativos.

A partir disso, gerou-se uma primeira nuvem de palavras, na qual foi possível identificar outros termos que também precisavam ser excluídos, tais como letras soltas e informações relativas ao periódico/evento do artigo, entre outros.

Para a avaliação da aplicabilidade em categorização utilizando categorias construídas, que foram rotulados manualmente *a priori*, realizou-se a etapa de teste e treino de um conjunto de

algoritmos supervisionados, sendo considerados os algoritmos Rede Neural, KNN e *Random Forest*.

Para a avaliação da aplicabilidade em categorias construídas *a posteriori* foi utilizado o processo não supervisionado, isto é, utilizando padrões estatísticos por meio de um algoritmo de clusterização. Para isso, foi retirada a *feature* que indicava categorização (*target*) dos textos selecionados.

Novamente foi repetido o pré-processamento para garantir a qualidade dos dados e, então, escolhido os parâmetros de frequência dos termos e dos documentos de forma que fosse calculada a importância de uma palavra em um documento em relação a uma coleção de documentos, e não somente as palavras de maior ocorrência total.

Posteriormente foram calculadas as métricas de distância no conjunto de dados utilizando-se as referências Euclidiana e Jaccard, resultando, assim, na aproximação entre os textos similares e, consequentemente, em sua categorização. Para a finalização do estudo, optou-se pela métrica Jaccard, pois obteve-se os melhores resultados de clusterização.

2.2 Categorização utilizando corpus criado por meio de aplicação de estratégia de busca

Para a elaboração do *Corpus* 2) utilizou-se a base de dados *Web of Science,* na qual dois grupos de amostras foram selecionados para a utilização na etapa de treino e um grupo de amostra para a etapa de teste.

As amostras utilizadas para treino foram divididas em duas categorias, construídas *a priori,* como base nos resultados obtidos na primeira etapa do estudo (categorizado manualmente) que demonstrou duas principais abordagens para o tema - relacionada ao âmbito digital (Categoria A) e não relacionada ao âmbito digital (Categoria B). Em cada uma das categorias foram identificadas as palavras mais frequentes que caracterizam cada abordagens, visíveis por meio da análise da nuvem de palavras gerada no processo.

Essas palavras foram consideradas para a elaboração da estratégia de busca da segunda etapa da pesquisa (categorizado automaticamente). O quadro 1 apresenta as palavras adotadas para construção da estratégia de busca de cada categoria.

Quadro 1 - Categorias das amostras de treino

CATEGORIA A	CATEGORIA B
cultural heritage	cultural heritage
digital	social
semantic	identity
interoperability	museum
metadata	value
data	material

pattern	politics
vocabulary	institutions

Fonte: Elaborado pelos autores.

Para ambas as categorias a busca foi realizada com refinamento para Tópicos (título, palavras-chave, resumo), com filtro de "c*ategorias da Web of Science: Information Science Library Science*"; filtro de idioma para inglês, filtro de tipos de documento: Artigo de conferência or Artigo. Ainda, para ambas, foram selecionados apenas os anos pares. Essa medida foi adotada para que não houvesse a repetição de artigos do treino na amostra a ser utilizada para teste e, visse versa, sendo que a amostra para teste foi composta de artigos publicados em anos ímpares, evitando assim enviesamento do processo.

A estratégia de busca adotada para criação do *corpus* da "Categoria A" foi (TS=("cultural heritage")) AND TS=("digital" OR " semantic" OR "interoperability" OR "metadata" OR "data" OR "pattern" OR "vocabulary", enquanto que a estratégia adotada para a "categoria B" foi (TS=("cultural heritage")) AND TS=("social" OR "identity" OR "museum" OR "value" or "material" OR "politics" OR "institutions").

Realizadas as buscas nas bases de dados, foram recuperados 382 artigos na Categoria A e 267 artigos na Categoria B. Foi feita então uma seleção aleatória de 250 artigos de cada categoria, visando evitar discrepâncias entre o número de documentos utilizado para teste em cada uma das categorias.

Dando continuidade ao processo, para compor a amostra de teste foi considerada a estratégia (TS=("cultural heritage")), considerando os mesmos critérios de filtragem das amostras anteriores. Foram considerados apenas os anos ímpares, sendo selecionada uma amostra aleatória de 150 documentos.

Os artigos também foram submetidos a uma etapa de pré-processamento dos dados, com aplicação de filtros e de *stopwords*, que foram construídas para o idioma inglês. Fez-se uso também da tokenização e das opções de transformação dos dados de modo a garantir a padronização, com remoção de URLs e demais *links*.

Para o procedimento de treino e teste nessa etapa, foram considerados os mesmos algoritmos da etapa anterior, sendo eles Rede Neural, KNN e *Random Forest*.

3 Aplicação de PLN e ml em *corpus* pré-selecionado e categorizado manualmente (etapa 1)

Como resultado da primeira etapa, com o pré-processamento foi possível obter um panorama das discussões sobre patrimônio cultural, por meio da nuvem de palavras, que coloca em destaque os principais termos recorrentes no *corpus* teórico analisado.

Nesta etapa de geração da nuvem de palavras destaca-se a importância do processo de limpeza dos dados, o que fica evidente na Figura 1, em que se observa a nuvem de palavras antes e depois da remoção das *stopwords*.

Figura 1 - Nuvem de palavras antes e depois do pré-processamento

Fonte: Elaborado pelos autores.

O processo do treino/teste foi realizado levando em consideração 80% do *corpus* total (37 dos 46 artigos), os outros 9 artigos foram reservados para uma validação, realizada aleatoriamente. O procedimento de treino/teste teve o *corpus* (37 artigos) com uma aplicação de *K-fold cross validation*[1] *para 20 repetições, com divisão de 70% para treino e 30% para teste*. Esses parâmetros foram os melhores encontrados após alguns testes.

Utilizou-se a métrica de acurácia para avaliar o resultado dos algoritmos. Apesar do pequeno *corpus*, o que normalmente não favorece um algoritmo de Rede Neural, ele teve melhor desempenho com uma acurácia de 88%, já o KNN atingiu 84% e o *Random Forest* com 82%.

Para a validação, portanto, foi aplicado o algoritmo Rede Neural, levando em consideração 20% dos artigos (9 dos 46 artigos). Tais documentos já haviam sido previamente rotulados pelos pesquisadores de forma manual para permitir a checagem dos erros e acertos, mas essa rotulação não foi indicada ao algoritmo.

Após a etapa de treino e teste já descrita, a etapa de validação teve apenas 1 dos 9 artigos classificados incorretamente, baseado na classificação manual, obtendo-se uma acurácia aproximada de 89%. Compreendeu-se que acurácia do algoritmo é influenciada por diversos fatores como o detalhamento e rigor no pré-processamento e na limpeza dos dados, o tamanho e a representatividade da amostra escolhida.

[1] A validação cruzada *k-fold* destina-se a estimar a habilidade do modelo em novos dados.

Em relação ao processo não supervisionado, os resultados puderam ser verificados usando um algoritmo de clusterização hierárquica (*hierarchical clustering algorithm*) que permite a visualização dos documentos em função da aproximação ou distanciamento de seu conteúdo.

Isso se dá, pois, o algoritmo lê a métrica de distanciamento escolhida e calcula uma matriz que categoriza os documentos que são mais similares, resultando em um dendrograma. Com base nisso, o pesquisador pode combinar os itens de seu corpus a partir das métricas que lhe entregue melhores resultados.

Como o *corpus* analisado já havia passado por uma análise de conteúdo manual, resultando nas duas categorias (A e B) empregadas no primeiro procedimento (supervisionado), foi possível comparar se o segundo procedimento (não supervisionado) entregava resultados próximos ao que foi obtido manualmente.

Como é possível visualizar na Figura 2, o algoritmo gera um dendograma, ou seja, uma forma de visualização em formato de árvore com ramificações clusterizadas por similaridade. Ao marcar as duas categorias de maior nível na clusterização (C1 e C2), é possível gerar uma nuvem de palavras para cada uma (Figura 2), demonstrando que a primeira (C1) se aproxima da Categoria B identificada manualmente, isto é, referente a um contexto não vinculado ao digital. Já a segunda (C2), se aproxima da Categoria A identificada manualmente, o que corresponde ao contexto digital. Palavras como 'museus', 'objetos' e 'sociais', são as mais relevantes de C1, enquanto que 'interoperabilidade', 'metadados' e 'web' são os principais destaques da C2.

Figura 2 - Nuvem de palavras de C1 e C2

Fonte: Elaborado pelos autores.

Diante disso, verifica-se que o procedimento não supervisionado cria categorias que podem ajudar o pesquisador a decidir quais serão seus próprios parâmetros e suas próprias categorias de análise, isto é, categorias *a posteriori*, direcionando a leitura dos documentos e facilitando na compreensão do *corpus* em questão.

4 Aplicação de PLN E ML em *corpus* criado por meio de aplicação de estratégia de busca (etapa 2)

O processo do treino/teste foi realizado levando em consideração 225 artigos representativos de cada categoria (A e B). O procedimento de treino/teste teve como *corpus*, portanto, 450 artigos que submetidos a testes de acurácia, permitiu estabelecer a aplicação de *random sampling*, com parâmetros em 20 repetições, com divisão de 60% para treino e 40% para teste. Foram testados os algoritmos KNN, Random Forest e Rede Neural.

Também foram realizados testes com a parametrização dos algoritmos em busca de resultados mais representativos, entretanto, foram obtidos resultados significativamente inferiores ao da categorização apresentada na seção 3, descrita anteriormente. O algoritmo KNN obteve melhor desempenho, com uma acurácia de 47%, a rede neural obteve 43% e o *Random Forest* com 39%.

Buscando compreender os resultados de acurácia significativamente inferiores aos obtidos na etapa anterior foi realizada uma análise das categorias criadas por meio da aplicação da estratégia de busca. Essa análise foi realizada a partir da observação dos artigos e especialmente das nuvens de palavras resultantes de cada categoria.

Observou-se que embora existam particularidades em cada nuvem, essas ainda são significativamente semelhantes, em especial os termos "*digital*" e "*information*", grandes destaques em ambas, tal como é possível visualizar na Figura 3.

Figura 3 - Nuvem de palavras de A e B

Fonte: Elaborado pelos autores.

Essa semelhança entre as nuvens indica que mesmo com o uso de termos representativos de cada categoria, o emprego da estratégia de busca não foi suficiente para individualizá-las corretamente, o que teve impacto direto na acurácia dos algoritmos testados.

Com o objetivo de validar a hipótese de que os baixos níveis de acurácia obtidos estavam relacionados a pouca diferença existente entre as categorias obtidas por meio das estratégias

de busca, foi realizada a conferência manual de cada categoria, selecionando os artigos que melhor representavam cada uma. Para isso, considerou-se a aderência do tema do artigo (título; descritores; resumo) em relação à sua categoria - âmbito digital (Categoria A) e não relacionada ao âmbito digital (Categoria B).

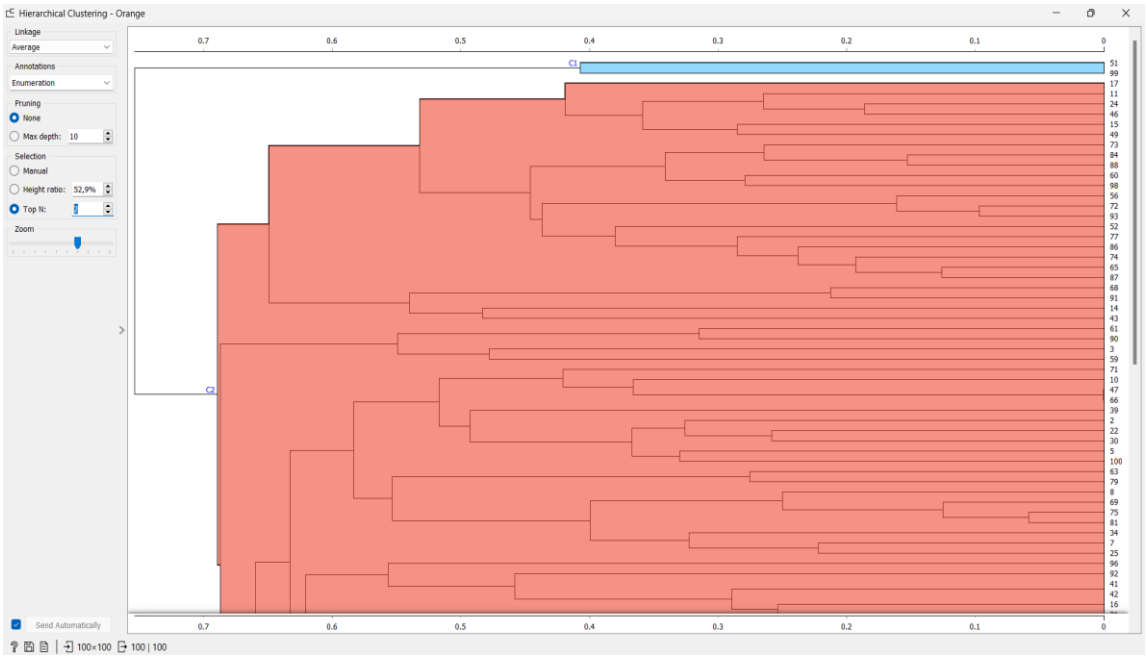
A nova amostra, proveniente da seleção manual, resultou em 50 artigos representativos de cada categoria, totalizando 100 artigos a serem utilizados para treino e teste. Nessa rodada foram testados novamente os algoritmos KNN, *Random Forest* e Rede Neural, com aplicação de *random sampling*, com parâmetros em 20 repetições, sendo 60% para treino e 40% para teste.

O algoritmo KNN obteve melhor desempenho, com uma acurácia aproximada de 74%, seguido do *Random Forest* (71%), tendo o pior desempenho a Rede Neural (67%). O KNN obteve os melhores resultados de acurácia, sendo estes alcançados com a parametrização em 10 para o "número de vizinhos" e utilizando a Métrica Euclidiana e o Peso Uniforme.

Em relação a aplicação do algoritmo não supervisionado, foi aplicado um algoritmo de clusterização hierárquica (*hierarchical clustering algorithm*), utilizando a métrica de distância Euclidiana. Na aplicação desse algoritmo é possível pré-determinar o número de categorias a ser gerado, sendo os artigos agrupados por aproximação de maneira não supervisionada (sem o uso de uma amostra inicial para treino).

Ao aplicar a mesma parametrização do estudo descrito na seção 3, ou seja, indicado a criação das duas categorias, não foram obtidos resultados significativos. Apenas um único artigo foi considerado um *outlier*, sendo os demais todos agrupados em uma grande e única categoria. A figura 4 apresenta o dendrograma gerado no processo que permite a visualização das ramificações clusterizadas por similaridade.

Figura 4 - Dendrograma da nova amostra (100 artigos)

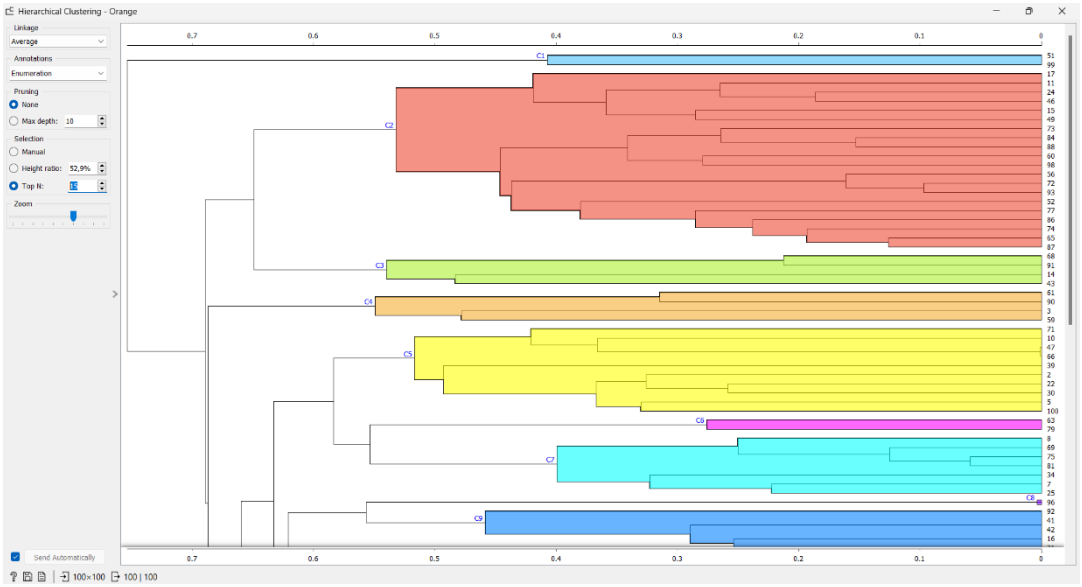


Fonte: Elaborado pelos autores.

Com a alteração do tipo de métrica de Euclidiana para *Cosine*e de quantidade de categorias a serem visualizadas, os resultados obtidos se tornaram mais significativos. A partir de 5 categorias apresentadas pelo algoritmo já foi possível analisar a nuvem de palavras gerada para cada uma delas e identificar os padrões que levaram ao agrupamento dos artigos. Após, detalhada análise, estabeleceu-se que para esse estudo, o número ideal foi de 15 categorias. A figura 5, mostra essa configuração

Figura 5 - Dendrograma da nova amostra (100 artigos) com mais de 5 categorias

Fonte: Elaborado pelos autores.



A análise da nuvem de palavras de cada categoria permite identificar quais os principais termos utilizados pelo algoritmo para a inferência de proximidade dos artigos. Um exemplo desse agrupamento pode ser observado na figura 6, na qual a nuvem de palavras da categoria 2 (C2), em que estão reunidos 20 artigos.

Figura 6 - Nuvem de palavras da categoria 2 (C2)

Fonte: Elaborado pelos autores.

Pelo contraste visual nota-se o destaque dado às bibliotecas nesses artigos, combinados com os conceitos “público” e "digital". Diante isso, pode-se inferir que o que individualiza esses artigos dos demais (*corpus* de 100 artigos no total), e ao mesmo tempo os aproxima como uma categoria (C2), é o foco dados às bibliotecas, com destaque para bibliotecas públicas e digitais. Além disso, o fato de 20 artigos serem considerados semelhantes nesses parâmetros, demonstra a importância do assunto para o tema geral.

Embora não seja possível apresentar nessa seção a nuvem de palavras criada para todas as 15 categorias estabelecidas, o quadro 2 apresenta alguns dos termos de destaque visual em cada categoria e o número de artigos reunidos, o que permite delinear seus respectivos assuntos e recorrência.

Quadro 2- Categorias estabelecidas automaticamente

CATEGORIA	N° DE ARTIGOS	TERMOS INDIVIDUALIZADORES MAIS EVIDENTES
C1	2	ciência, artigo, estudo
C2	20	biblioteca, pública, digital
C3	4	coleção, objetos, instituição
C4	4	história, cultural, projetos
C5	10	informação, usuários, grupo
C6	2	arquivo, público, nacional, informação
C7	7	patrimônio, digital, cultural

C8	1	não identificado
C9	3	conteúdo, usuário, digital
C10	16	dados, gestão, recursos, patrimônio cultural, digital
C11	4	dados abertos, web, bases
C12	4	modelo, construção, cultural, tempo
C13	3	universidade, cultural, histórico, material
C14	1	não identificado
C15	5	conhecimento, comunidade

Fonte: Elaborado pelos autores.

Nesse nível de categorização apenas dois artigos foram considerados *outliers*, isto é, que não se enquadram em nenhuma categoria. Para esses dois casos não foi possível gerar e visualizar a nuvem de palavras, bem como os termos utilizados por falta de dados.

Fica evidente, portanto, que esse tipo de análise tem o potencial de contribuir com o processo de categorização em pesquisas científicas, a exemplo desse estudo, no qual as 15 categorias geradas poderiam ser empregadas como uma pré-categorização, ou seja, utilizadas como categorias construídas *a priori*, que permitiriam um olhar mais direcionado para o *corpus* a ser estudado, facilitando seu entendimento. Essa análise também permite a identificação dos principais temas que circundam a discussão sobre o patrimônio cultural, viabilizando a identificação de novos padrões ou baixos níveis de discussão de determinados assuntos correlatos.

5 Síntese e discussão dos resultados

Tendo em vista a comparação das abordagens empregadas em cada etapa, é possível sintetizar os procedimentos seguidos tal como no quadro 3.

Quadro 3–Síntese das abordagens empregadas em cada etapa

ETAPA	CORPUS	AMOSTRA	ALGORITMO	CATEGORIZAÇÃO	FINALIDADE
1	*corpus* 1 (categorização manual)	37 artigos pré-categorizados em dois grupos (A e B)	supervisionado	*a priori* (categorias estabelecidas previamente)	auxiliar a descrição de características do domínio/objeto estudado
		9 artigos (porcentagem do corpus total A + B)	não supervisionado	*a posteriori* (criação de novas categorias)	auxiliar no conhecimento do domínio/objeto estudado

2	*corpus* 2 (categorização automática a partir de estratégias de busca)	A1: 450 artigos pré-categorizados em dois grupos (A e B) A2: 100 artigos pré-categorizados em dois grupos (A e B) *anos pares	supervisionado	*a priori* (categorias estabelecidas previamente)	auxiliar a descrição de características do domínio/objeto estudado
		150 artigos selecionados sobre o tema *anos ímpares	não supervisionado	*a posteriori* (criação de novas categorias)	auxiliar no conhecimento do domínio/objeto estudado

Fonte: Elaborado pelos autores.

Diante dessa síntese, destaca-se que para ambas as etapas, quanto mais claras forem as características específicas de cada categoria estabelecida *a priori* e quanto mais representativa for a mostra de treino selecionada para criação de novas categorias *a posteriori*, maiores serão as chances de acerto do algoritmo para reconhecê-las.

Na primeira etapa do estudo, como o universo da pesquisa já era conhecido, e categorias já haviam sido estabelecidas por processo manual e a amostra delimitada, o algoritmo reconheceu mais satisfatoriamente as categorias e foi capaz de gerar novas complementares de modo a auxiliar na descrição de características do domínio/objeto estudado.

Na segunda etapa, que teve por base as palavras-chave da etapa 1, o intuito foi o de automatizar a criação prévia de categorias, de forma a ter pelo menos duas delas tal como na etapa 1. Contudo, o resultado não foi tão satisfatório, apresentando baixos níveis de acurácia e uma nova categorização foi necessária (nova amostragem - A2), sendo essa criada de forma manual.

Mais de um fator pode ser considerado para explicar esse resultado não satisfatório, mas reconhece-se que a qualidade das estratégias de busca tem grande peso e propõe-se no futuro aprimorar essa variável. Contudo, para esse estudo, optou-se por testar a hipótese de que os resultados não satisfatórios obtidos estavam relacionados a pouca diferença existente entre as categorias obtidas por meio das estratégias de busca. Com a nova amostra, o algoritmo já teve melhor desempenho no reconhecimento das categorias, apesar de não conseguir separá-las em duas grandes tal como feito manualmente. Mesmo assim, a criação de novas categorias complementares revela-se um importante aliado do pesquisador no reconhecimento do domínio/objeto estudado, ajudando a conhecê-lo.

6 Considerações finais

Entende-se que o estudo realizado colabora na compreensão do processo de análise de trabalhos científicos, sem ter aqui o objetivo de determinar ou obter bons resultados. Considera-se também que não houve análises aprofundadas com experimentações de parâmetros até sua saturação nos algoritmos utilizados, o que se considera para um posterior aprofundamento da pesquisa.

Em relação a primeira etapa proposta para esse estudo - aplicação de PLN e ML em corpus pré-selecionado e categorizado manualmente -pode-se concluir que na predição de novos documentos com base em categorias obtidas *a priori* (categorização manual)ainda não é possível excluir a participação do pesquisador no processo de categorização. Entretanto, o número de acertos faz com que a aplicação do processo seja relevante e possibilite imaginar um cenário em que o algoritmo atuaria na pré-classificação e o pesquisador validação, possibilitando uma redução significativa de trabalho manual.

Outro papel importante do pesquisador no processo seria o de seleção da amostra utilizada para treino do algoritmo e estabelecimento correto das características que diferenciam os conjuntos de dados e permitem a criação das categorias, levando em consideração a influência desses quesitos na acurácia do algoritmo.

Já em relação a criação de categorias *a posteriori*, aplicando técnicas de PLN e ML em procedimento não supervisionado, conclui-se que os resultados são promissores, tendo em vista que com base na aplicação desse procedimento é possível identificar novos padrões que poderiam passar despercebidos pelos próprios pesquisadores.

Em relação a segunda etapa proposta para esse estudo - aplicação de PLN e ML em *corpus* criado por meio de aplicação de estratégia de busca - observou-se que mesmo quando as estratégias de busca são elaboradas de forma fundamentada em uma análise exploratória, isto é, conhecendo o universo da pesquisa, e na visualização de nuvem de palavras, refletindo a representatividade dos termos, ainda ocorre a recuperação de artigos que não são inteiramente representativos considerando as categorias estabelecidas (A e B), e que essa limitação tem impacto nos resultados do processo de ML.

Para obter resultados mais representativos foi necessário uma pré-categorização manual dos artigos utilizados com algoritmos supervisionados na fase de treino e teste. Esse fato reforça a importância da participação do pesquisador como validador do *corpus* utilizado.

Seguindo o mesmo procedimento aplicado na etapa 1, também foi verificado o potencial de análise dos estudos por meio da aplicação de algoritmo não supervisionado de clusterização hierárquica. Nesse caso, os resultados também se mostram promissores, sendo possível identificar um conjunto de 15 categorias que caracterizam o universo estudado e ajudam o

pesquisador a entender seu *corpus* de análise. Essas categorias revelam além das temáticas mais recorrentes, assuntos correlatos e tendencias de pesquisa.

Diante disso, reconhece-se a importância da representatividade dos termos escolhidos na estratégia de buscas empregadas na seleção da amostra. Logo, como estudos futuros pretende-se acrescentar ao estudo base, na fase de aplicação de estratégia de busca, uma etapa de refinamento manual, baseada na seleção e recategorização manual do recorte utilizado para treino e teste, verificando assim o impacto da inclusão dessas etapas na acurácia do algoritmo empregado.

Cabe ressaltar que a etapa 2 foi realizada partindo da hipótese de que um volume mais substancial de documentos ampliaria o potencial do procedimento, uma vez que mais dados seriam fornecidos para que algoritmo pudesse ser treinado. Contudo, no caso desse estudo, destaca-se que apesar da quantidade da amostra ser essencial para gerar resultados satisfatórios, somente isso não garante que estes sejam representativos do ponto de vista do domínio estudado, cabendo sempre discussões e análises multidisciplinares que permitam verificar e readequar os parâmetros de seleção e recorte da amostra.

Com base nas discussões apresentadas evidencia-se que as técnicas de Processamento de Linguagem Natural e de *Machine Learning* são promissoras para os processos de categorização de recursos informacionais, podendo contribuir assim com a redução do tempo despendido por profissionais especializados, incluindo os profissionais da informação no que tange às atividades de catalogação, classificação e indexação, bem como na categorização de resultados de pesquisas científicas. Considera-se, portanto, que as técnicas podem contribuir tanto na pré-categorização de novos recursos - quando as categorias desejadas já forem definidas -, como para a elaboração de novas categorias, permitindo assim identificação de padrões que poderiam passar despercebidos pelos pesquisadores ou evidenciando padrões já conhecidos.

Referências

BORKO, H. Information science: what is it? **American Documentation**, Washington, v. 19, n. 1, p. 3-5, Jan. 1968.

CONEGLIAN, C. S. **Recuperação da Informação com abordagem semântica utilizando Linguagem Natural**: a Inteligência Artificial na Ciência da Informação. 2020. 194 f. Tese (Doutorado) - Curso de Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista, Marília, 2020. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/193051/coneglian_cs_dr_mar.pdf?sequence=3&isAllowed=y. Acesso em: 08 set. 2022.

FERNEDA, E. **Recuperação de informação**: análise sobre a contribuição da ciência da computação para a ciência da informação. 2003. 137 f. Tese (Doutorado) - Curso de Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista, Marília, 2003.

Disponível em: https://teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/fr.php. Acesso em: 08 set. 2022.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255-260, 2015. Disponível em: https://www.science.org/doi/abs/10.1126/science.aaa8415. Acesso em: 08 set. 2022.