PROCESSAMENTO DE LINGUAGEM NATURAL E MACHINE LEARNING COMO APARATO PARA A CATEGORIZAÇÃO DE ARTIGOS CIENTÍFICOS

Natural Language Processing and Machine Learning as an apparatus for the categorization of scientific papers

Ananda Fernanda de Jesus[1], Maria Lígia Triques[2], José Eduardo Santarem Segundo[3], Ana Cristina de Albuquerque[4]

(1) Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP-Marília), Av. Hygino Muzzi Filho, 737 - Bairro Mirante, Marília/SP - CEP 17.525-900, af.jesus@unesp.br.

(2) Universidade Estadual de Londrina (UEL), Rodovia Celso Garcia Cid, PR-445, Km 380 - Campus Universitário, Londrina - PR, CEP 86057-970, mligia.triques@uel.br.

(3) Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP-Marília), Av. Hygino Muzzi Filho, 737 - Bairro Mirante, Marília/SP - CEP 17.525-900, Av. Hygino Muzzi Filho, 737 - Bairro Mirante, Marília/SP - CEP 17.525-900, santarem@usp.br.

(4) Universidade Estadual de Londrina (UEL), Rodovia Celso Garcia Cid, PR-445, Km 380 - Campus Universitário, Londrina - PR, CEP 86057-970, albuanati@uel.br.

Resumo:
Este estudo 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, por meio de categorias estabelecidas *a priori* e *a posteriori*. A partir da aplicação das técnicas de ML e PLN por meio de dois algoritmos de categorização (algoritmo de rede neural e algoritmo de clusterização hierárquica) em um *corpus* documental constituído de artigos científicos brasileiros sobre a temática "patrimônio cultural", desenvolve-se uma pesquisa aplicada, com resultados quantitativos e qualitativos derivados de dois procedimentos de teste: o primeiro utilizando algoritmo supervisionado, cuja categorização foi feita *a priori*; e o segundo, utilizando algoritmo não supervisionado, com a categorização feita *a posteriori.* Os resultados demonstram que para ambos os casos há a importância do detalhamento e rigor no pré-processamento dos dados e do tamanho e da representatividade da amostra escolhida. No caso supervisionado, conclui-se que quanto mais claras forem as características específicas de cada classe estabelecida *a priori* e quanto mais representativa for a mostra treino selecionada, maiores serão as chances de acerto do algoritmo. Já no caso não supervisionado, percebe-se que o algoritmo identifica de forma satisfatória o conteúdo dos documentos, permitindo inclusive identificar mais padrões de categorização que podem ser úteis às análises dos pesquisadores.
Palavras-chave: machine Learning; processamento de linguagem natural; algoritmo de rede neural; algoritmo de clusterização hierárquica; patrimônio cultural.

Abstract: This study aims to verify the potential of applying Natural Language Processing (NLP) and Machine Learning (ML) techniques in the thematic categorization of scientific articles, through categories established a priori and a posteriori. NLP and ML techniques are applied through two categorization algorithms (neural network algorithm and hierarchical clustering algorithm) in a documentary corpus composed of Brazilian scientific articles on the theme "cultural heritage", it is develop an applied research, with quantitative and qualitative results derived from two test procedures: the first using a supervised algorithm, which categorization was made a priori; and the second, using an unsupervised algorithm, with a posteriori categorization. The results demonstrate that, for both cases, there is an importance of detail and rigor in the pre-processing of the data and the chosen sample's size and representativeness. In the supervised case, it is concluded that the clearer the specific characteristics of each class established a priori and the more representative the selected training sample is, the greater the chances of success of the algorithm. In the unsupervised case, the algorithm satisfactorily identifies the documents content, even allowing the identification of more categorization patterns that can be useful for researchers' analyses.
Keywords: machine learning; natural language processing; neural network algorithm; hierarchical clustering algorithm; cultural heritage.

1 Introdução

A elaboração de categorias 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.

2 Objetivos

O presente estudo busca verificar o potencial de aplicação de técnicas de Processamento de Linguagem Natural (PLN) e de *Machine Learning* (ML) na automação do processo de categorização temática de artigos científicos, de forma a replicar as duas principais formas manuais de categorização: *priori* e a *posteriori* por meio de um recorte experimental.

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, propõe-se a aplicação das técnicas de ML e PLN em um corpus documental constituído de artigos científicos brasileiros que figurem em seus tópicos de estudo a expressão ou termo 'patrimônio cultural' como uma expressão explicitada ou definida em seu contexto de estudo.

Para refletir as principais formas manuais de categorização, o *corpus* selecionado foi submetido a algoritmos supervisionados e não supervisionados. Nesse cenário, os algoritmos supervisionados refletem a categorização *a priori*, em que se tem um conjunto de categorias conhecidas e busca-se "encaixar" novos documentos nessas categorias. Para a aplicação do algoritmo supervisionado o corpus selecionado foi previamente rotulado manualmente.

Já os algoritmos não-supervisionados foram aplicados visando refletir o processo de categorização *a posteriori*, quando os documentos observando são analisados e, então, são criadas categorias. Para essa aplicação os documentos não foram previamente rotulados, observando-se o potencial do algoritmo na criação de novas categorias de análise.

Ressalta-se que para as duas situações, objetivou-se replicar, via aplicação de

algoritmos, os procedimentos realizados de forma manual para categorização.

2 Procedimentos Metodológicos

A presente pesquisa é caracterizada como um estudo aplicado, com resultados quantitativos e qualitativos, com a finalidade de verificar o potencial de aplicação das técnicas de PLN e ML na categorização de resultados de um levantamento bibliográfico, tendo como recorte artigos científicos, publicados nacionalmente, em base temática da Ciência da Informação, a respeito da definição de 'patrimônio cultural'.

Para tanto, 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*.

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

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

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.

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.

3 Resultados

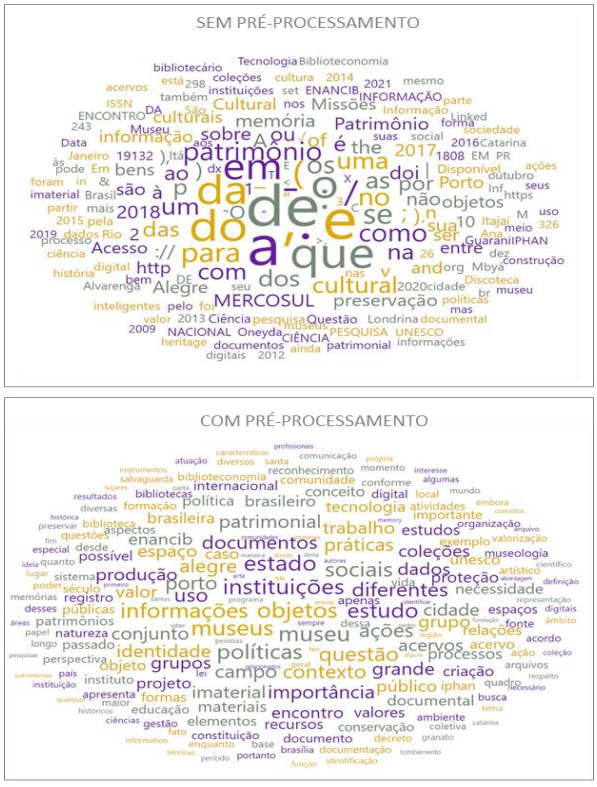
Como resultado da primeira etapa, com o pré-processamento já foi possível obter um panorama das discussões sobre patrimônio cultural, por meio da nuvem de palavras, que coloca em destaque os principais termos recorrentes no *corpus* teórico analisado.

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*.

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. Quanto mais claras forem as características específicas de cada classe estabelecida *a priori* e quanto mais representativa for a mostra treino selecionada, maiores serão as chances de acerto do algoritmo.

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



Fonte: Elaborado pelos autores.

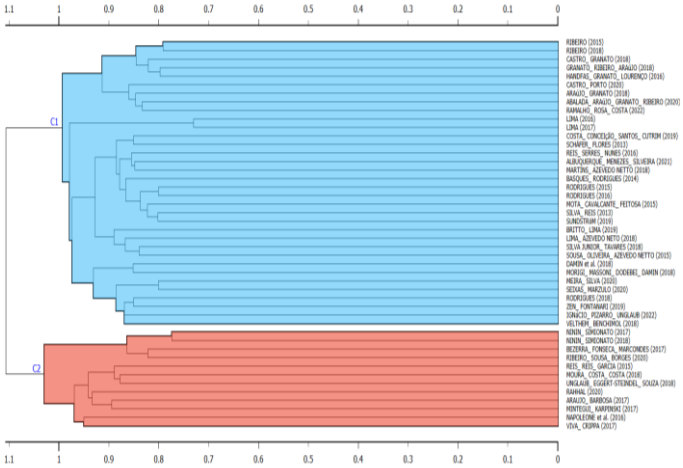
Em relação ao processo não supervisionado, os resultados puderam ser verificados usando um algoritmo de clusterização hierárquica (*hierarchical clustering algorithm*) que permite a visualização dos documentos em função da aproximação ou distanciamento de seu conteúdo.

Isso se dá, pois, o algoritmo lê a métrica de distanciamento escolhida - no caso deste estudo, o índice Jaccard - e calcula uma

matriz que categoriza os documentos que são mais similares, resultando em um dendrograma, tal como na Figura 3.

Com base nisso, o pesquisador pode combinar os itens de seu corpus a partir das métricas que lhe entregue melhores resultados.

Figura 3 - Dendograma do corpus de análise

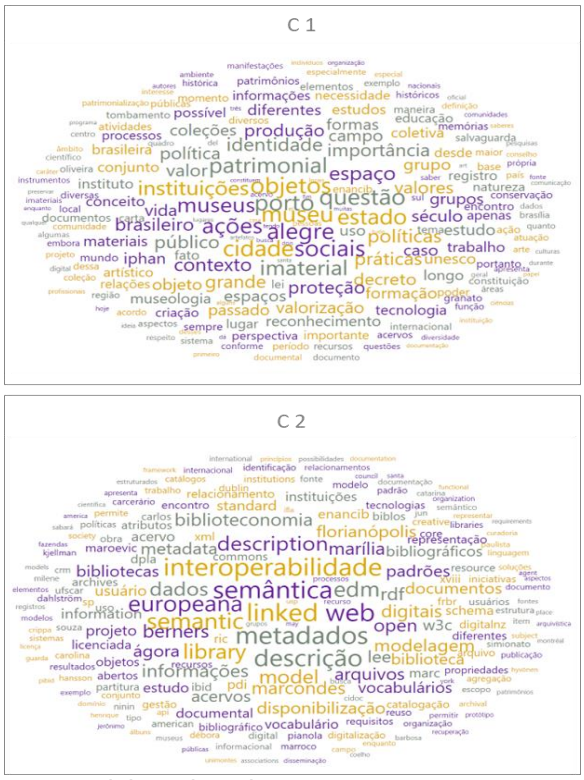


Fonte: Elaborado pelos autores.

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 3, 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 4), 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 4 - 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 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, com a inclusão de um *corpus* mais robusto.

Em relação a predição de novos documentos com base em categorias obtidas *a priori*, conclui-se que 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 novo cenário, com o algoritmo atuando como uma pré-classificação, em que o pesquisador atuaria como validador, representando uma redução significativa de trabalho manual.

Outro papel importante do pesquisador no processo será o de seleção da amostra utilizada para treino do algoritmo e estabelecimento correto das características que diferenciam os conjuntos de dados e que permitem a criação das categorias, levando em consideração a influência desses quesitos na acurácia do algoritmo. Conclui-se ainda que o potencial de contribuição desse procedimento seria ampliado em análise de grandes volumes de documentos.

Já em relação a criação de categorias *a posteriori*, aplicando técnicas de PLN e ML, conclui-se que os resultados são mais promissores, tendo em vista que com base na aplicação desse procedimento é possível identificar novos padrões que poderiam passar despercebidos pelos próprios pesquisadores. Novamente o potencial do procedimento seria ampliado em um contexto de grandes volumes de documentos cuja análise aprofundada pelo pesquisador seria um processo longo e exaustivo, e em muitas situações inviável.

Tendo em vista a ampliação do potencial dessas técnicas na análise de grandes conjuntos de documentos, como estudos futuros, pretende-se ampliar a amostra utilizada e ainda realizar teste com outros algoritmos e esgotar (saturar) o uso de seus parâmetros com busca de melhores resultados. Pretende-se ainda verificar o potencial de aplicabilidade das técnicas em outras formas de análise de resultados de pesquisa científica, como na geração de métricas, em processos como Bibliometria.

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 para a 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.