CLASSIFICAÇÃO AUTOMÁTICA DE ARTIGOS PUBLICADOS NO ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - 2021

Automatic Classification of Articles Published in the National Meeting on Research in Information Science - 2021

Liliane Cristina Soares Sousa[1], José Eduardo Santarém Segundo [(2),] Fábio Parra Furlanete[(3)]

(1) Universidade Estadual de Londrina - UEL, Londrina/PR, lilianeli.sousa@uel.br

(2) Universidade de São Paulo – USP, Ribeirão Preto/SP, santarem@usp.br

(3) Universidade Estadual de Londrina – UEL, Londrina/PR, ffurlanete@uel.br

Resumo:

A classificação de documentos científicos, diante do número expressivo de produção e disponibilização informacional, é um desafio contemporâneo para a Ciência da Informação. Diante disso, propõe-se fazer um experimento com algoritmos de Machine Learning, para entender de que maneira essas ferramentas, podem processar dados textuais e gerar métricas de validação. A acurácia, é uma métrica de validação para entender este processo de classificação na análise de dados não estruturados, como os documentos textuais. A metodologia se baseia nas concepções de Mineração de Textos, que dispõe de ferramentas para estruturar análise de dados não estruturados ou semiestruturados. Ao aplicar diferentes algoritmos de Machine Learning na massa de textos, requer observar o comportamento destes no processamento dos dados. A pesquisa preliminar foi promissora, no que se refere, a entender como se comporta a aplicação de algoritmos de Machine Learning em dados textuais. Conclui-se que, estudos em torno da área de Data Sciense, particularmente usando algoritmos de Machine Learning podem ser aprofundados e aplicados na Ciência da Informação, com o propósito de agregar valor para diversas áreas da ciência, em particular, para os profissionais da informação.

Palavras-chave: Ciência da Informação; Machine Learning; Classificação textual; Dados não-estruturados; Algoritmo.

Abstract:

The classification of scientific documents, given the significant number of production and information availability, is a contemporary challenge for Information Science. Therefore, it is proposed to do an experiment with Machine Learning algorithms, to understand how these tools can process textual data and generate validation metrics. Accuracy is a validation metric to understand this classification process in the analysis of unstructured data, such as textual documents. The methodology is based on the concepts of Text Mining, which has tools to structure analysis of unstructured or semi-structured data. When applying different Machine Learning algorithms in the mass of texts, it is necessary to observe their behavior in the data processing. Preliminary research was promising, in terms of understanding how the application of Machine Learning algorithms in textual data behaves. It is concluded that studies around the area of Data Sciense, particularly using Machine Learning algorithms, can be deepened and applied in Information Science, with the purpose of adding value to several areas of science, in particular, for information professionals.

Keywords: Information Science; Machine Learning; Text classification; Unstructured data; Algorithm.

1. Introdução

O presente texto faz parte de uma investigação preliminar, para compreender se algoritmos de Machine Learning são capazes de obter uma boa acurácia (CA) no processo de classificação textual. Importante destacar que Machine Learning, é o aprendizado de máquina, que estabelece ferramentas de análise de dados, que permite a automação de elaboração de modelos analíticos de processamento de dados. Como observa Jordan e Mitchell (2015, p. 255, tradução nossa), “O estudo do aprendizado de máquina é importante tanto para abordar questões científicas e de engenharia fundamentais, quanto para software de computador altamente práticos que são produzidos e utilizados em diversos aplicativos.” Diante disso, deseja-se testar a

importância de estudos de Data Science e sua aplicação na Ciência da Informação, para auxiliar no processo de análise de corpus textual. Compreende-se que os estudos de Data Sciense que envolvem as informações, o processo de seleção, preparação, transformação, desenvolvimento, processamento e análise de dados, são significativos para a Ciência da Informação.

O corpus selecionado está para uma perspectiva de análise de dados não-estruturados. Sendo que, o objeto da investigação está direcionado aos trabalhos publicados no ENANCIB 2021.

O Encontro de Pesquisa em Ciência da Informação (ENANCIB), representa o principal evento de pesquisa e de pós-graduação da área de Ciência da Informação do país. Aspira problematizar e refletir a produção científica da área, visto que, requer fomentar amplo diálogo entre os pesquisadores protagonistas nos programas de pós-graduação. O evento é direcionado para estimular troca de experiências acadêmico-científicas e pela solidificação de laços acadêmicos no âmbito nacional e internacional.

A presente investigação trará a aplicação de técnicas de Mineração de Texto, para alcançar a hipótese projetada. A Mineração de Texto, segundo Barion e Lago (2008, p. 125), é identificada como uma extensão da área de Data Mining; e de acordo com Thuraisinggham (1999, p. 210), tem por objetivo central extrair modelos e associações singulares de uma gama significativa de um banco de dados textual. Ainda na concepção de Barion e Lago (2008, p. 123), esses processos possibilitam usar conjuntos de estratégias para navegar, organizar, achar e descobrir informações em base de textos.

No dizer de Wives (2004, p. 66) é enfatizado que a maneira básica de Mineração de Texto, está relacionado na exploração e identificação de terminologias relevantes na composição de um grupo de documentos, como também instituir padrões textuais e projetar grupos temáticos de conteúdos pela periodicidade de aparecimento de termos no corpus a ser investigado. Esse processo de mineração de texto, se caracteriza como um método de suporte para pesquisadores, a fim de possibilitar um novo prisma informacional, em grande e significativa coleção de textos.

A Mineração de Texto emerge em consequência da necessidade de explorar, de maneira automática, modelos e anomalias informacionais em documentos. Segundo Aranha e Passos (2006, p. 125), a Mineração de Texto, é uma área do conhecimento multidisciplinar, que permeia diversos campos da ciência, como “Informática, Estatística, Linguística e Ciência Cognitiva”, conforme esta investigação preliminar, entende-se que a Ciência da Informação, também é um campo científico de atuação na Mineração de Texto.

Nossa questão é: Em que medida algoritmos de Machine Learning são capazes de processar métricas de validação em classificação de dados textuais?

2. Objetivos

Este artigo tem o objetivo de diagnosticar se algoritmos de Machine Learning, aplicados em dados não-estruturados, são capazes de classificar documentos textuais, pelo olhar das técnicas de classificação dos bibliotecários. Tendo como corpus de pesquisa os artigos publicados no ENANCIB 2021.

3. Procedimentos Metodológicos

A execução dessa investigação fundamenta-se na Mineração de Texto, uma vez que, estabelece procedimentos de extração de padrões em significativas quantidades de textos, tendo como atributo a linguagem natural, e geralmente, para ser utilizados com objetivo específico. Visto que, a Mineração de Texto, requer selecionar conhecimentos relevantes de dados não estruturados ou semiestruturados.

As etapas desenvolvidas nesta pesquisa, está representada logo a seguir:

Figura 1 – Etapas do processo de análise dos dados

Fonte: Os autores (2022)

O primeiro passo para atingirmos o objetivo proposto, foi compreender os processos necessários para a aplicação da Mineração de Texto, em particular, nos dados selecionados para esta investigação. Seguindo a concepção de Aranha (2007, p. 150), a Mineração de Texto, está permeada por quatro macro etapas: “coleta, pré-processamento, indexação e análise da informação”.

A posteriori, iniciamos a coleta dos dados, nos Anais do ENANCIB - 2021. Essa etapa, segundo Schiessl e Medeiros (2011, p. 100), consiste na elaboração do objeto da pesquisa, formada na Mineração de Texto, por uma base textual, no qual será trabalhada no processo de análise. Pode-se denominar esta base de textos, como Corpus; composto por um grupo de textos, que representa um conjunto de linguagens naturais. O corpus desta investigação foi extraído do Anais do ENANCIB - 2021, foram selecionados 342 artigos, separados por GTs: GT1 (19); GT2 (38); GT3 (38); GT4 (49); GT5 (42); GT6 (32); GT7 (26); GT8 (43); GT9 (15); GT10 (25); GT11 (15).

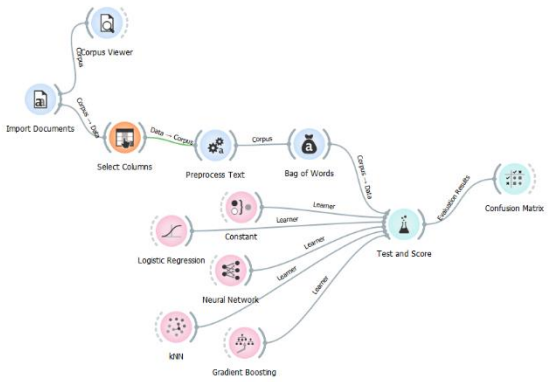
Posteriormente, na segunda etapa denominada pré-processamento, que é caracterizada por ser responsável pela construção de uma estrutura representativa dos documentos textuais. No qual pretende-se com este processamento, aprimorar a organização e a qualidade dos dados. Esta ação consiste em aplicar diversas maneiras de transformações nos documentos textuais, com a ideia de estruturar estes grupos de documentos, para poderem ser submetidos a algoritmos de mineração de texto. No dizer de Aranha e Passos (2006, 132), as estratégias utilizadas nos “tratamentos dados ao texto durante essa fase podem ser feitos tanto de forma automatizada como feitos por humanos, porém o desempenho dos sistemas automáticos é extremamente superior.” E segundo Spark-Jones e Willet (1997, p. 5), essa etapa de pré-processamento “inclui tokenização, limpeza de dados e eliminação de stopwords.” Nesta fase, foi analisado a necessidade de preparar os textos para que a observação seja eficiente. Identificou-se que a massa documental estava desbalanceada, percebendo que em cada delimitação dos grupos de trabalho, há uma quantidade diferente de artigos. Deste modo, opta-se por equilibrar a massa textual, para obter um mapeamento equitativo dos dados, sem enviesar os resultados pelo desbalanceamento das categorias. Isto posto, os dados textuais foram reorganizados, equiparando cada grupo de trabalho do ENANCIB 2021, totalizando 165 artigos, distribuindo 15 artigos por GT (GT1 / GT2 / GT3 / GT4 / GT5 / GT6 / GT7 / GT8 / GT9 / GT10 / GT11).

Ainda nesta perspectiva de pré-processamento dos dados, como terceira etapa do processo, tem-se a indexação dos dados, que objetiva uma acessibilidade rápida e eficiente destas informações, a busca por palavras. Faz parte da Mineração dos dados, onde se seleciona a tarefa que será executada, conforme a necessidade desejada.

O próximo passo refere-se à construção de uma estrutura de dados, composta pelas etapas anteriores, no qual, aplica-se algoritmos de mineração de dados, com o objetivo de extrair os conhecimentos. Por fim, é feita a análise da aplicação dos algoritmos, entrando na etapa de análise e leitura dos dados gerados.

Todos os processos embasados teoricamente, expostos anteriormente, foram seguidos para obter-se os resultados preliminares de análise. As etapas seguidas para a execução do mapeamento do processamento dos textos, estão ilustradas na Figura 2:

Figura 2 – Mapeamento do processo de execução da Mineração de Texto



Fonte: Elaborado pelos autores usando o software Orange Canvas (2022)

4. Resultados

Nesse passo, processa-se os dados textuais com objetivo de identificar a acurácia de acerto de classificação pelos algoritmos. O escopo da acurácia para o processamento e leitura dos dados, fornece-nos uma junção entre exatidão e precisão, consequentemente, oferece-nos resultados próximos do valor real das classificações. O fator selecionado para o processamento dos dados foi a validação cruzada, com número dez de dobras e 70% para treino com 30% dos dados para teste.

Para a construção das amostragens, foram utilizados os seguintes algoritmos: KNN, executa previsões conforme as classes mais próximas, utilizou-se cinco como vizinhos mais próximos, sendo a métrica euclidiana e peso uniforme; Redes Neurais (Scikit-Learn), usando Perceptron multicamadas, que é capaz de aprender modelos não lineares e lineares, sendo parâmetro utilizado para esse algoritmo 200 como número máximo de interações; Logistic Regression, essa ferramenta representa o algoritmo de classificação de regressão logística com regularização, onde utilizou-se o tipo de regularização Ridge (L2); Gradient Boosting, algoritmo que utiliza a combinação de resultados de preditores fracos, com a intenção de elaborar modelos preditivos, os parâmetros aplicados foram de 100 para o número de árvores e taxa de aprendizagem de 0,100 e Constant, que gera uma amostra o qual na maioria das vezes prevê a maior parte para tarefas de classificação e valor médio para funções de regressão.

Posterior a execução dos testes, com os diversos algoritmos de Machine Learning, foi selecionado o algoritmo Logistic Regression. A seleção desse algoritmo, fundamenta-se, pela sua característica de processamento e visualização das relações entre os dados. Em face do exposto, elegeu-se a acurácia como métrica de validação para a presente observação. Visto que, a CA (acurácia) significa a proximidade de resultado levando em consideração o seu valor de referência real, ou seja, quanto maior o nível de acuracidade, mais perto do modelo significa o resultado encontrado.

Dessa maneira, segue-se com os resultados dos testes aplicados nos dados textuais selecionados e pré-processados. Observa-se a figura 3:

Figura 3 – Análise do algoritmo selecionado

Model	AUC	CA	F1	Precision	Recall
kNN	0.932	0.952	0.733	0.733	0.733
Neural Network	0.970	0.945	0.710	0.688	0.733
Logistic Regression	0.960	0.952	0.714	0.769	0.667
Gradient Boosting	0.927	0.939	0.643	0.692	0.600
Constant	0.500	0.909	0.000	0.000	0.000

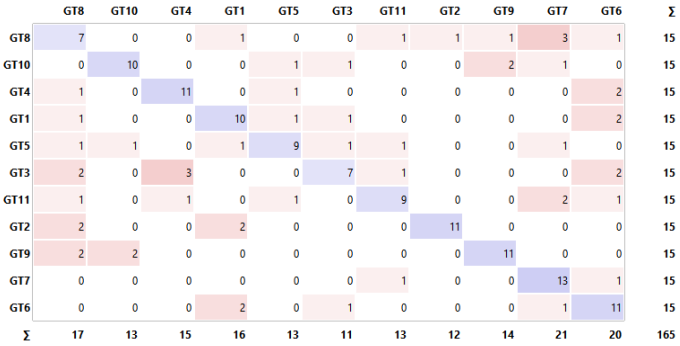
Fonte: Elaborado pelos autores usando o software Orange Canvas (2022)

Ao extrair os dados processados no teste com o algoritmo Logistic Regression, verifica-se que a CA (acurácia) extraída foi de 95%.

Na figura 4, é possível fazer a leitura que o algoritmo Logistic Regression representou na Mineração dos textos selecionados.

Para executar a leitura destes dados gerados, utilizou-se a Confusion Matrix, que apresenta o número ou proporção entre a classe prevista e a atual. Esses resultados foram consequência dos dados gerados pela avaliação do algoritmo Logistic Regression.

Figura 4 – Análise do processamento dos dados



	GT8	GT10	GT4	GT1	GT5	GT3	GT11	GT2	GT9	GT7	GT6	Σ
GT8	7	0	0	1	0	0	1	1	1	3	1	15
GT10	0	10	0	0	1	1	0	0	2	1	0	15
GT4	1	0	11	0	1	0	0	0	0	0	2	15
GT1	1	0	0	10	1	1	0	0	0	0	2	15
GT5	1	1	0	1	9	1	1	0	0	1	0	15
GT3	2	0	3	0	0	7	1	0	0	0	2	15
GT11	1	0	1	0	1	0	9	0	0	2	1	15
GT2	2	0	0	2	0	0	0	11	0	0	0	15
GT9	2	2	0	0	0	0	0	0	11	0	0	15
GT7	0	0	0	0	0	0	1	0	0	13	1	15
GT6	0	0	0	2	0	1	0	0	0	1	11	15
Σ	17	13	15	16	13	11	13	12	14	21	20	165

Fonte: Elaborado pelos autores usando o software Orange Canvas (2022)

Visualiza-se na figura 4, o processamento dos dados, e indicam a seguinte leitura:

- GT1 (15 artigos), 10 artigos estão conforme respectiva classe, no entanto, existem: 1 artigo para o GT8, 1 artigo para o GT5, 1 artigo para o GT3 e 2 artigos para o GT6. Dessa maneira, de acordo com a previsão de proporção, 66,66% dos artigos do GT1 pertencem a sua classificação prévia;
- GT2 (15 artigos), 11 artigos estão conforme respectiva classe, no entanto, existem: 2 artigos para o GT8 e 2 artigos para o GT1. Sendo assim, de acordo com a previsão de proporção, 73,33% dos artigos do GT2 pertencem a sua classificação prévia;
- GT3 (15 artigos), 7 artigos estão conforme respectiva classe, no entanto, existem: 2 artigos para o GT8, 3 artigos para o GT4, 1 artigo para o GT11 e 2 artigos para o GT6. Dessa maneira, de acordo com a previsão de proporção, 46,66% dos artigos do GT3 pertencem a sua classificação prévia;
- GT4 (15 artigos), 11 artigos estão conforme a respectiva classe, no entanto, existem: 1 artigo para o GT8, 1 artigo para o GT5 e 2 artigos para o GT6. Dessa maneira, de acordo com a previsão de proporção, 73,33% dos artigos do GT4 pertencem a sua classificação prévia;
- GT5 (15 artigos), 9 artigos estão conforme a respectiva classe, no entanto, existem: 1 artigo para GT8, 1 artigo para GT10, 1 artigo para GT1, 1 artigo para GT3, 1 artigo para o GT11 e 1 artigo para GT7. Dessa maneira, de acordo com a previsão de proporção, 60% dos artigos do GT5 pertencem a sua classificação prévia;
- GT6 (15 artigos), 11 artigos estão conforme respectiva classe, no entanto, existem: 2 artigos para o GT1, 1 artigo para o GT3 e 1 artigo para o GT7. Sendo assim, de acordo com a previsão de proporção, 73,33% dos artigos do GT6 pertencem a sua classificação prévia.
- GT7 (15 artigos), 13 artigos estão conforme a respectiva classe, no entanto, existem: 1 artigo para o GT11 e 1 artigo para o GT6. Sendo assim, de acordo com a previsão de proporção, 86,66% dos artigos do GT7 pertencem a sua classificação prévia;
- GT8 (15 artigos), 7 artigos estão conforme a respectiva classe, no entanto, existem: 1 artigo para o GT1, 1 artigo para o GT11, 1 artigo para o GT2, 1 artigo para o GT9, 3 artigos para o GT7 e 1 artigo para o GT6. Dessa maneira, de acordo com a previsão de proporção, 46,66% dos artigos do GT8 pertencem a sua classificação prévia;
- GT9 (15 artigos), 11 artigos estão de acordo com a respectiva classe, no entanto, existem: 2 artigos para o GT8 e 2 artigos para GT10. Sendo assim, de acordo com a previsão de proporção, 73,33% dos artigos do GT9 pertencem a sua classificação prévia;
- GT10 (15 artigos), 10 artigos estão conforme respectiva classe, no entanto, existem: 1 artigo para o GT5, 1 artigo para o GT3, 2 artigos para o GT9 e 1 artigo para o GT7. Dessa forma, de acordo com a previsão de proporção, 66,66% dos

artigos do GT10 pertencem a sua classificação prévia;

- GT11 (15 artigos), 9 artigos estão conforme a respectiva classe, no entanto, existem: 1 artigo para o GT8, 1 artigo para o GT4, 1 artigo para o GT5, 2 artigos para o GT7 e 1 artigo para o GT6. Sendo assim, de acordo com a previsão de proporção, 60% dos artigos do GT11 pertencem a sua classificação prévia;

O processamento dos dados com o algoritmo Machine Learning, demonstrou a possibilidade de validação, no processamento de mineração de texto com os cinco algoritmos testados, inclusive, com o Logistic Regression, selecionado para a apresentação dos índices de CA da análise.

A mineração de textos indicou, nesta perspectiva, uma estratégia importante, cuja intenção, de acordo com Feldman e Hirsh, citados por Wives (2004, p. 80), significa "constituir-se em um meio efetivo de recuperação, filtragem, manipulação e resumo do conhecimento contido em grandes volumes de informações textuais, para apresentá-lo em forma de gráficos, listas ou tabelas para o consumo de suas informações."

5. Considerações Finais

Esta pesquisa preliminar, identificou que analisar informações em documentos não estruturados significa um desafio. Mesmo com os avanços no âmbito da tecnologia informacional, salienta-se a necessidade de trilhar caminhos de preparação dos documentos textuais, a fim de qualificar os dados, para obter resultados no processamento desses dados.

Diante dos resultados obtidos, a partir do processamento dos dados textuais, demonstrou-se a viabilidade de aprofundamento na investigação. A possibilidade de entender quais os elementos foram levados em consideração na classificação dos textos feitas pelo algoritmo.

É interessante observar que ao aumentar o volume de dados a ser testado, pode-se conseguir identificar elementos interessantes de análise, como a GTs de trabalho que podem ter conteúdos mais aderentes entre eles, ou ainda grupos de trabalho que tenham termos mais significativos que os colocam em situação de exclusividade de tema em relação a outros GTs.

A mineração de textos, a partir das concepções de Data Science, permite a construção de ferramentas de extração e uso de dados, que têm a fortalecer as pesquisas na Ciência da Informação.

6. Referências

ARANHA, Christian; PASSOS, Emmanuel. A Tecnologia de Mineração de Textos. **Revista Eletrônica de Sistemas de Informação**, [S.l.], v. 5, n. 2, ago. 2006. ISSN 1677-3071. Disponível em: http://periodicosibepes.org.br/index.php/reinfo/article/view/171>. Acesso em: 12 set. 2022. doi:https://doi.org/10.21529/RESI.2006.0502001.

ARANHA, Christian N. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português**: Sob o Enfoque da Inteligência Computacional. 2007. 144f. Tese (Doutorado) - Programa de Pós-graduação em Engenharia Elétrica, Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, 2007. Disponível em: https://www.maxwell.vrac.puc-rio.br/10081/10081_4.PDF.Acesso em 03 jun. 2022.

BARION, E. C.N. **Mineração de textos**. 2008. Disponível em: https://revista.pgsskroton.com.br/index.php/rcext/article/view/2372/2276. Acesso em: 24 nov. 2018.

INGERSOLL, Grant S.; MORTON, Thomas S.; FARRIS, Andrew L. 2013. **Taming Text: How to find, organize and manipulate it.** Shelter Island, NY (USA): Manning Publications Co., 2013. 298p.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255-260, 2015. Disponível em:

<https://cs.uwaterloo.ca/~y328yu/mycourses/480-2018/readings/JordanMitchell.pdf>. . Acesso em: 20 nov. 2022.

SPARK-JONES, K; WILLET, P. (1997). **Readings in Information Retrieval**. Morgan Kaufmann. 1997.

SCHIESSL, M.; MEDEIROS, M. Descoberta de Conhecimento em Texto aplicada a um Sistema de Atendimento ao Consumidor. **Revista Ibero-Americana de Ciência da Informação**, v. 4, n. 2, p. 94-111, 2011.

THURAISINGHAM, Bhavani M. Data mining: technologies, techniques, tools, and trends. Boca Raton:
Editora CRC Press, 1999. 288p.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos**. Tese (Doutorado em Computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.