CLASSIFICAÇÃO AUTOMATIZADA DE TEXTOS CIENTÍFICOS EM ÁREAS DE CONHECIMENTO DO ENSINO SUPERIOR

MACIEL, T. V. [1], SILVA, R.R.[1]

[1] Instituto Federal Sul-Rio-Grandense (IFSUL) – Bagé – RS – Brasill

thalesmaciel@ifsul.edu.br, profrodrigorosadasilva@gmail.com

RESUMO

Este artigo documenta uma metodologia baseada em mineração de texto para a classificação automatizada de trabalhos científicos sobre uma determinada organização das áreas do conhecimento científico. Na abordagem proposta, as palavras-chave que compunham os metadados dos textos foram utilizadas na etapa de treinamento do algoritmo, enquanto os respectivos resumos foram utilizados nos testes. Os resultados obtidos mostraram potencial de subsequentes aplicação no mundo real.

Palavras-chave: Mineração de texto, classificador, ensino superior.

1 INTRODUÇÃO

No meio acadêmico, é comum a ocorrência de eventos científicos e periódicos multidisciplinares. Neles, trabalhos de diversas áreas de conhecimento devem passar pela avaliação do respectivo mérito científico antes de serem aprovados para publicação nos anais do evento ou periódico.

Neste contexto, comumente são designados agentes humanos como avaliadores para cada trabalho submetido. Para prover maior eficiência na avaliação, esta designação ocorre através da associação manual entre as áreas de conhecimento informadas por cada avaliador como de seu interesse e aquela informada pelos autores dos trabalhos a qual a obra esteja supostamente inserida.

Ocorre que, muitas vezes, a informação da área de conhecimento dos trabalhos submetidos é incompleta, incorreta ou ausente, o que causa transtornos quando da avaliação pelos avaliadores originalmente designados ou, até mesmo, da própria designação de avaliadores para tais submissões.

O presente trabalho teve a problemática definida na forma de “como automatizar a atribuição de áreas de conhecimento para textos científicos?”. A hipótese estudada é a de que as palavras-chave cadastradas como metadados dos textos podem ser utilizadas nesta automatização, especificamente em tarefas de mineração de texto. Assim, é objetivada a caracterização de um método constante

de viabilidade prática no que se refere ao consumo de recursos computacionais, tempo de execução e âmbito aceitável para a ocorrência de erros de classificação para predição da área de conhecimento referente a textos científicos. Após esta introdução, a seção 2 apresenta os materiais e métodos utilizados no estudo. A seção 3 expõe os resultados obtidos, sendo o estudo concluído na seção 4.

2 METODOLOGIA (MATERIAL E MÉTODOS)

No Brasil, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) publicou, através da portaria no 09/2008, a formalização das áreas de conhecimento do ensino superior. Esta formalização é proposta sob a estrutura de árvores de áreas de conhecimento, cujos níveis representam grandes áreas (topo), áreas (dentro de cada grande área), subáreas (dentro de cada área) e especialidades (dentro de cada subárea), buscando uma maneira ágil e objetiva de agregar informações (CAPES, 2008).

Frank, Hall e Witten (2016) definem a mineração de texto como o processo de análise de texto automatizada que visa a extração de informação útil para fins específicos. Estes autores classificam a mineração de texto como uma especialização da mineração de dados, diferenciando-se desta no tocante a extração de informação implícita e desconhecida. Considera-se que, no caso de texto, a informação é explícita no próprio texto – apenas não é originalmente estruturada com vistas em práticas de análise automatizada. Em comum, a mineração de dados e mineração de texto mantém o critério de que a informação a ser extraída deve ser útil.

O conjunto de dados disponibilizado para análise foi composto por 997 instâncias de trabalhos científicos, todas constituídas por dois atributos: um representando o conjunto de quatro palavras-chave distintas que são comumente requisitos para submissão de trabalhos científicos e utilizadas para fins de indexação de obras e outro representando a respectiva classificação na árvore de áreas de conhecimento proposta pela CAPES (2008).

Com vistas em evitar problemáticas de qualidade dos dados e desvio de conceito (Eiwell e Polikar, 2011), foram admitidas instâncias de trabalhos cuja área de conhecimento fora descrita da forma mais específica possível, apenas. Por exemplo, em uma suposta árvore de áreas de conhecimento composta pela área de Computação tendo as sub-áreas de Teoria da Computação, Hardware e Software, não seriam admitidas classificações em Computação, apenas, mas sim em

quaisquer uma de suas especificações, que são os níveis mais detalhados dentro da respectiva abrangência.

Para fins da realização dos experimentos com mineração de texto sobre o conjunto de dados descrito, foi utilizado o Waikato Environment for Knowledge Analysis (WEKA), ou ambiente para análise de conhecimento desenvolvido pela Universidade de Waikato, localizada na Nova Zelândia (Hall et al. 2009).

O WEKA é uma coleção de algoritmos que podem ser utilizados em atividades de mineração de dados diversas, como classificação, regressão, associação, clustering e mineração de texto, além de pré-processamento de conjunto de dados e visualização de resultados através de sua interface gráfica, linhas de comando, ou a própria application programming interface (API), escrita em linguagem Java (Frank, Hall e Witten, 2016).

A experimentação central, descrita neste estudo, foi desempenhada com a utilização do pseudo-classificador para classificação filtrada, cuja implementação é disponibilizada no WEKA sob nome FilteredClassifier (Hall et al. 2009). Ele possibilita a aplicação de técnicas de filtragem ou seleção de dados em conjunto com a execução de um algoritmo de classificação ou regressão, estendendo o aprendizado realizado por tal algoritmo ao produto do melhoramento na qualidade dos dados, porém, sem alterar o estado do conjunto de dados original.

Outrossim, a atividade de classificação filtrada foi parametrizada para realizar a conversão do atributo referente ao conjunto de palavras-chave dos trabalhos em um novo conjunto de atributos, então representando a ocorrência ou não de cada palavra-chave contida no conjunto de dados original. Esta atividade de filtragem, por sua vez, foi configurada, a partir de seu comportamento padrão, para formatar as palavras-chave em letras minúsculas exclusivamente e ignorar a ocorrência de palavras contidas em uma lista auxiliar, onde estavam contidas expressões julgadas de forma a priori como irrelevantes ao contexto de classificação de textos científicos.

O algoritmo efetivamente responsável pelo processamento na mineração de dados desempenhada foi o Naive Bayes Multinominal (Mccallum e Nigam, 1998), projetado com vistas no aprendizado de máquina para classificação de texto. Este algoritmo é implementado no WEKA sob nome MultinomialNaiveBayes (Hall et al. 2009), onde não há parametrização ao seu comportamento padrão.

3 RESULTADOS E DISCUSSÃO

A experimentação com a metodologia descrita na seção 2 resultou na obtenção de um modelo para classificação de texto baseado na ocorrência ou não ocorrência das palavras-chave encontradas em cada instância em relação ao conjunto das palavras-chave encontradas na totalidade do conjunto de dados analisado.

Em outras palavras, para cada instância processada no treinamento do algoritmo, foi sintetizada uma matriz que representou a totalidade de palavras-chave encontradas no conjunto de dados, com a indicação da presença ou ausência de cada uma delas dentre as palavras-chave da instância. O atributo referente à área de conhecimento da instância, alvo da atividade de classificação, foi mantido sem alterações.

O método obteve 69% de acurácia nos testes de classificação de texto realizados sobre o mesmo conjunto de dados que fora utilizado para treinamento do algoritmo. Em função da existência de classes representadas por apenas uma instância, da totalidade do conjunto de dados, foi inviabilizada a realização de testes sob outros métodos, como a divisão do conjunto de dados em porções distintas para treinamento e testes ou estratégias de validação cruzada com ou sem estratificação.

Embora a acurácia apresentada pelos testes de classificação tenha sido pouco inferior à 70%, esta não foi a única métrica utilizada na avaliação de sucesso na construção do modelo inferido, sendo considerado, ainda, o coeficiente de kappa com valor de 0,67. Valores do coeficiente de kappa neste âmbito indicam evidências substanciais de que os resultados obtidos não foram apenas casuais, apresentando, de fato, relevância estatística para o domínio de aplicação (Vieira e Joanne, 2005).

Ademais, entende-se que a avaliação de padrões e modelos descobertos por mineração de dados são, muitas vezes, subjetivados em função da aplicabilidade prática no domínio de negócio ou aplicação ao qual pertencem (Maciel et al. 2015).

4 CONCLUSÃO

A partir dos resultados obtidos nos experimentos, foi possível verificar a eficácia da metodologia descrita. Foi possibilitada a automatização da atribuição das áreas de conhecimento de trabalhos científicos com base em uma aplicação do aprendizado de máquina, na forma de mineração de texto. Outrossim, foi julgado como satisfeito, o objetivo desta pesquisa.

Trabalhos futuros envolvem o tratamento do desbalanceamento entre as classes do conjunto de dados original, de forma que haja o nivelamento proporcional sobre a quantidade instâncias de trabalhos por categoria. Também é proposto que seja estudado um método para obter proveito da estrutura de árvore sob a qual estão organizadas as áreas de conhecimento no Brasil. Entende-se que tais medidas podem resultar no melhoramento da acurácia da metodologia apresentada e podem servir como propostas de extensão da mesma no futuro.

REFERÊNCIAS

Andrew Mccallum and Kamal Nigam (1998) A Comparison of Event Models for Naive Bayes Text Classification. In: AAAI-98 Workshop on 'Learning for Text Categorization'.

Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (2008) “Portaria no 09/2008”, http://reality.sgi.com/employees/jam_sb/mocap/MoCapWP_v2.0.html, December.

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Elwell, Ryan, and Robi Polikar. "Incremental learning of concept drift in nonstationary environments." IEEE Transactions on Neural Networks 22.10 (2011): 1517-1531.

Maciel, T., Seus, V., Machado, K. and Borges, E. (2015). Mineração de dados em triagem de risco de saúde. Revista Brasileira de Computação Aplicada, 7(2), 26-40.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.

Viera, A. and Joanne M. "Understanding interobserver agreement: the kappa statistic." Fam Med 37.5 (2005): 360-363.