15° Seminário de Iniciação Científica e Tecnológica

Reconhecimento de nomes de instituições utilizando Inteligência Artificial

João Gabriel Grandotto Viana*, Waldeyr Mendes Cordeiro da Silva**

Câmpus Formosa

**waldeyr.mendes@ifg.edu.br, ***joao.grandotto@estudantes.ifg.edu.br*

***Palavras Chave**: Inteligência Artificial; Classificação de Texto; Aprendizado de Máquina.*

Introdução

Mimetizar a capacidade humana de interpretação de textos está entre as promessas da Inteligência Artificial (IA). Já é possível classificar textos utilizando IA para identificar o estilo literário, analisar sentimentos embutidos, ou mesmo reconhecer uma mensagem como *spam* ou não. O preenchimento da afiliação por parte dos autores de publicações científicas, geralmente não segue padrões, o que causa problemas como diferentes formas de escrita, traduções, abreviações, siglas, ambiguidades e mesmo erros de digitação. Uma mesma instituição pode estar representada de diferentes formas em diferentes trabalhos acadêmicos. Por exemplo, é possível encontrar para o IFG variações tais como: IFG, Instituto Federal de Goiás, Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Federal *Institute of Goiás*, IFGO, entre outros. E, ainda que corretamente escritos, os nomes das instituições e suas siglas podem conter ambiguidades. Existe, por exemplo, um IFG (Instituto Federal de Goiás) no Brasil e um IFG *Institut für Gebirgsmechanik* na Alemanha. Há, portanto, uma demanda crescente por soluções eficientes e produtivas de análise de textos estruturados e não-estruturados. O problema de identificação de afiliações em artigos pode ser entendido como um problema de classificação de um texto e subsequente associação deste texto a uma categoria pré-estabelecida a partir do treinamento de um modelo de *Machine Learning*.

Metodologia

Foram elaboradas *strings* de busca compatíveis com as bases *Web of Science*, *Scopus* e *Scielo*. A partir das quais realizou-se uma busca automatizada nessas bases utilizando suas *APIs* com *scripts* Python. Os resultados foram tratados por também por *scripts* para limpeza, eliminação de redundâncias e identificação de ocorrência de termos com pesos relativos à sua ocorrência no conjunto de documentos: TF-IDF (RAMOS, 2003). Uma vez tratados os dados compuseram o *dataset* utilizado como entrada para o treinamento, teste e validação do modelo de *Machine Learning*. Foram explorados os algoritmos de aprendizado de máquina para classificação*: Support Vector Machines (SVM)*, *Naive Bayes (NB)* e *k- Nearest Neighbors (KNN)* em busca da alternativa que apresentasse os melhores resultados para o problema proposto sobre o conjunto de dados disponível.

Resultados e Discussão

Entre os modelos para classificação binária, que identifica se a instituição é ou não IFG, o resultado mais expressivo foi o algoritmo *LogisticRegression,* que obteve acurácia de 97%. A Figura abaixo mostra a curva ROC do modelo:

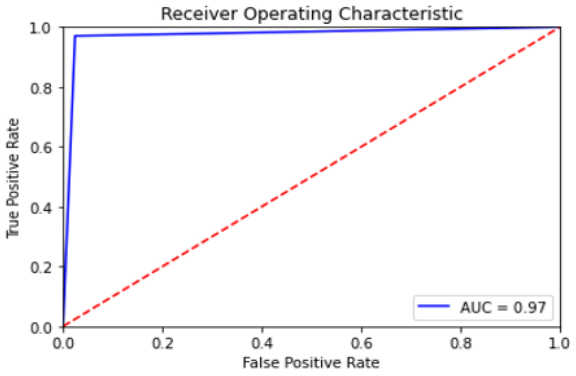

Figura 1. Curva ROC do modelo treinado para classificação binária (IFG ou não IFG) de nomes de instituições em afiliações.

Conclusões

O modelo treinado foi capaz de classificar corretamente a maior parte dos dados e pode ser adaptado para classificar dados desse tipo em sistemas já existentes, como o IFG Produz. Uma possível atualização para trabalhos futuros é expandir a classificação binária para multiclasses reconhecendo os *campi* do IFG.

Referências - RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: OF THE FIRST INSTRUCTIONAL CONFERENCE ON MACHINE LEARNING. Proceedings. . .[S.l.: s.n.], 2003. v.242, n.1, p.29–48.