DOI: 10.5748/16CONTECSI/LIS-6321

A STRATEGY FOR GENDER IDENTIFICATION IN OPEN DATA REPOSITORIES USING AN ARTIFICIAL NEURAL NETWORK MODEL

Sérgio José de Sousa - Centro Federal de Educação Tecnológica de Minas Gerais - sergio7sjs@gmail.com
Monique de Oliveira Santiago - Centro Federal de Educação Tecnológica de Minas Gerais - moniqueosantiago@gmail.com
Adilson Luiz Pinto - Universidade Federal de Santa Catarina - adilson@cin.ufsc.br
Thiago Magela Rodrigues Dias - Centro Federal de Educação Tecnológica de Minas Gerais - thiago@div.cefetmg.br

Abstract. Many open datasets do not present information about gender, which makes it difficult to analyze this type of information, such as the identification of polarities and inequalities. Some works try to perform this classification using the first name and using traditional techniques like SVM, other than this, this work tries to find a relation between the characters of the full name in order to identify a possible gender for the names. This neural network model proved to be very effective reaching a 98.99% accuracy.

Keywords: Neural Network, Full name, Gender Identification, Deep Learning.

Resumo. Muitas bases de dados abertas não apresentam informações com relação ao gênero o que dificulta analises sobre esse tipo de informação, como identificação das polaridades e desigualdades. Alguns trabalhos tentam realizar essa classificação através do primeiro nome e utilizando técnicas tradicionais como SVM, diferente disso, este trabalho tenta encontrar uma relação entre os caracteres do nome completo com a finalidade de identificar um possível gênero para os nomes. O modelo de rede neural se mostrou bem efetivo chegando a atingir 98,99% de acurácia.

Palavras-chave: Rede Neural; Nome Completo; Identificação de gênero; Deep Learning.

1. INTRODUÇÃO

Diversos são os estudos que tentam identificar e relacionar o gênero a uma quantidade de produção, em especial, na área acadêmica existe um esforço em responder a questão: há diferenças entre a produtividade científica feminina e masculina? (LETA, 2014; HILL, CORBETT e ST ROSE, 2010; BLICKENSTAFF, 2005; HYDE, 2005, TUESTA et al., 2019; SANTIAGO e DIAS, 2018).

Responder esse tipo de questão é de suma importância para identificar polaridades e desigualdades, assim, servir de apoio a tomada de decisões como tornar o ambiente mais homogêneo.

Um dos problemas encontrado na maioria das bases de dados é que nem sempre a informação de gênero está disponível, o que leva a necessidade de obter essa informação por outros meios, como no trabalho realizado por Naldi (2004) onde o autor utilizou uma base de dados de primeiros nomes para auxiliar na classificação do gênero. Há outros trabalhos utilizam dados extraídos do Twitter (LIU e RUTHS, 2013; CIOT, SONDEREGGER e RUTHS, 2013) onde é gerado uma nova base de dados com informação de perfil de usuário e gênero no qual foi usado técnicas como SVM e K-Means para classificar o gênero através do primeiro nome, conseguindo uma acurácia média de 87%.

2. METODOLOGIA

Este trabalho utiliza dados dos grupos de pesquisa do censo de 2016 que estão disponíveis na Plataforma Lattes. Consiste em XMLs que contém informações dos integrantes dos grupos, como nome completo, gênero, identificação do currículo. Essas informações foram extraídas e seu quantitativo pode ser conferido na Tabela 1.

Tabela 1. Quantitativo dos dados utilizados

Tipo de Dado	Quantidade
Registros válidos	622.385
Nomes femininos	352.804
Nomes masculinos	269.581

Com as informações importantes extraídas, inicia-se o pré-processamento para o modelo que consiste em inverter a ordem dos nomes, como por exemplo: “Alcides da Silva Diniz” se torna “Diniz Silva da Alcides”. Dessa maneira, as informações mais importantes para a classificação são posicionadas no fim, dando uma indicação mais forte ao modelo, proporcionando que ela aprenda mais rapidamente e de uma maneira mais eficiente. Seguindo, os nomes são transformados em tensores cujas letras são convertidas em valores inteiros sequencialmente. No mesmo exemplo temos o tensor([7, 39, 44, 39, 56, 0, 22, 39, 42, 52, 31, 0, 34, 31, 0, 4, 42, 33, 39, 34, 35, 49]).

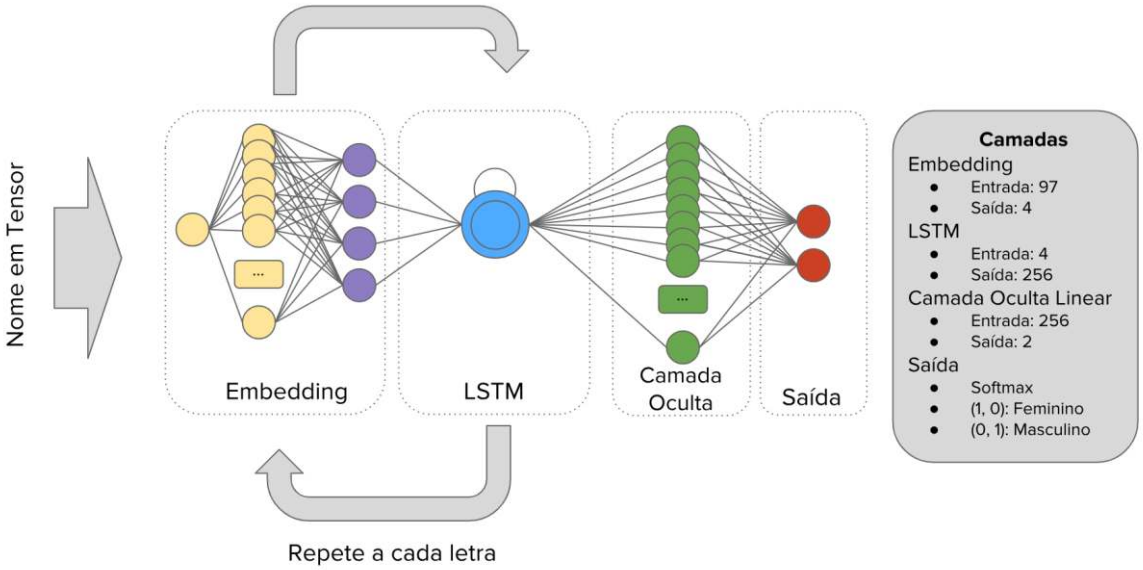

Figura 1 – Modelo de rede neural
Fonte: Próprio autor

A Figura 1 mostra o modelo proposto, a primeira camada de Embedding (ROWEIS e SAUL, 2000) com o propósito de reduzir de maneira não linear a dimensão de entrada que vai de 97 para 4. Seguindo tem-se uma camada de memória LSTM com 256 neurônios internos (GERS, SCHMIDHUBER e CUMMINS, 1999) que possui a capacidade de aprender padrões importantes e esquecer informações não são relevantes. Essas duas camadas se repetem letra a letra do nome, ao finalizar, a arquitetura segue com uma camada oculta linear com 256 neurônios ligados a uma camada de saída com 2 neurônios que classificam o dado de entrada. A função de perda utilizada foi de erro quadrático médio e a função de otimização Adam (KINGMA, 2014) com taxa de aprendizagem 0,00001 até a 7ª época, decaindo para 0,0000001 nas épocas seguintes.

Para validar o método foi utilizado a validação cruzada (GOLUB, HEATH, e WAHBA, 1979) dividindo o dado em 5 grupos, reservando um quinto para teste e os quatro quintos restante para treino. Treina-se então 5 modelos variando o conjunto de treino/teste, assim, tem-se uma medida mais confiável sobre a capacidade do modelo.

3. EXPERIMENTO

Os experimentos nos 5 modelos foram executados por 8 épocas que levou cerca de 220 horas. Os resultados médios podem ser vistos na Figura 2 onde o valor de perda é menor a cada iteração e a acurácia crescente, porem com valores mínimos nas ultimas iterações, indicando que o modelo está convergindo.

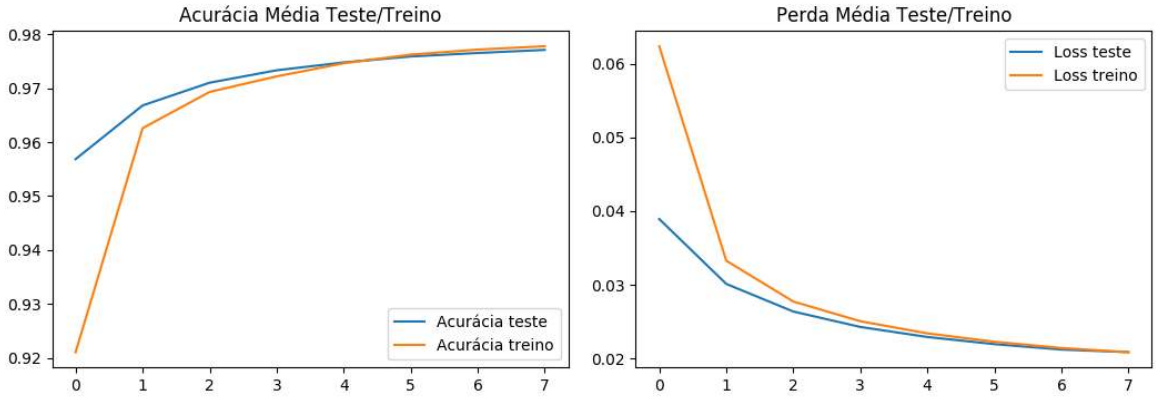

Figura 2 – Resultados da validação cruzada
Fonte: Próprio autor

Continuando o treinamento do modelo reduzindo a taxa de aprendizado é notado um aprimoramento, Figura 3, melhorando ainda mais os resultados, atingindo o melhor valor na época 16 com 99,549% de acurácia no treino e 98,995% de acurácia no teste. Avançando mais uma época o modelo começa a sofrer sobreajuste (overfitting) quando a acurácia de treino aumenta e a de teste começa a reduzir, ou seja, o modelo começou a decorar os dados de treino.

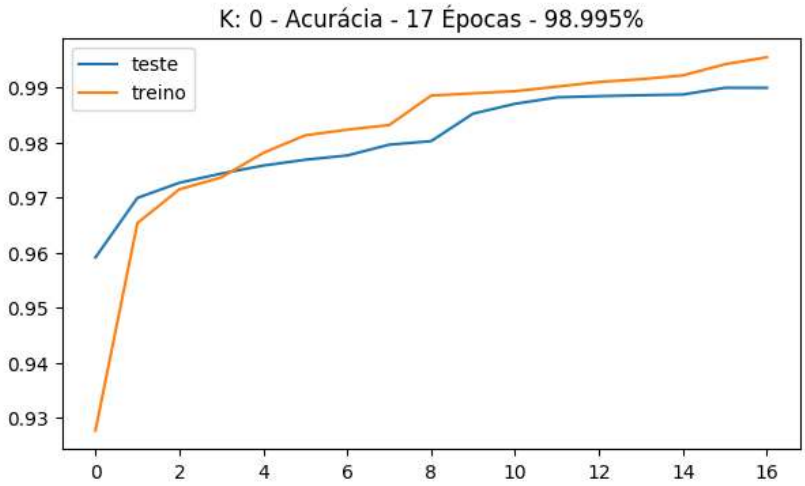

Figura 3 – Resultados com 17 épocas
Fonte: Próprio autor

4. CONCLUSÃO

O modelo baseado em redes neurais profundas proposto identificou uma relação entre as letras do nome completo de forma que conseguiu uma boa acurácia tendo em vista que a base de dados inclui alguns nomes estrangeiros que dificultam no treinamento e acurácia foi superior ao encontrado na literatura.

Trabalhos futuros incluem aplicar o modelo em nomes estrangeiros e utilizar outras informações para aprimorar o modelo, como por exemplo foto de perfil.

REFERÊNCIAS

BLICKENSTAFF, Jacob Clark. *Women and science careers: leaky pipeline or gender filter?*. Gender and education, v. 17, n. 4, p. 369-386, 2005.

CIOT, Morgane; SONDEREGGER, Morgan; RUTHS, Derek. *Gender inference of Twitter users in non-English contexts*. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. p. 1136-1145.

GERS, Felix A.; SCHMIDHUBER, Jürgen; CUMMINS, Fred. *Learning to forget: Continual prediction with LSTM*. 1999.

GOLUB, Gene H.; HEATH, Michael; WAHBA, Grace. *Generalized cross-validation as a method for choosing a good ridge parameter*. Technometrics, v. 21, n. 2, p. 215-223, 1979.

HILL, Catherine; CORBETT, Christianne; ST ROSE, Andresse. *Why so few? Women in science, technology, engineering, and mathematics*. American Association of University Women. 1111 Sixteenth Street NW, Washington, DC 20036, 2010.

HYDE, Janet Shibley. *The gender similarities hypothesis*. American psychologist, v. 60, n. 6, p. 581, 2005.

KINGMA, Diederik P.; BA, Jimmy. *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980, 2014.

LETA, Jacqueline. *Mulheres na Ciência Brasileira: desempenho inferior?*. Revista feminismos, v. 2, n. 3, 2014.

LIU, Wendy; RUTHS, Derek. *What's in a name? using first names as features for gender inference in twitter*. 2013 AAAI Spring Symposium Series. 2013.

NALDI, Fulvio et al. *Scientific and technological performance by gende*r. Handbook of quantitative science and technology research. Springer, Dordrecht, 2004. p. 299-314.

ROWEIS, Sam T.; SAUL, Lawrence K. *Nonlinear dimensionality reduction by locally linear embedding*. science, v. 290, n. 5500, p. 2323-2326, 2000.

SANTIAGO, Monique de Oliveira; DIAS, Thiago Magela Rodrigues. *Distribuição e Análise do Conjunto de Doutores Brasileiros Baseado em Gênero*. Pesquisa Brasileira em Ciência da Informação e Biblioteconomia, v. 13, n. 2, 2018.

TUESTA, Esteban Fernandez et al. *Análise da participação das mulheres na ciência: um estudo de caso da área de Ciências Exatas e da Terra no Brasil*. Em Questão, v. 25, n. 1, p. 37-62, 2019.