Atribuição de Autoria utilizando Aprendizado de Métrica Profundo

Walter do Espírito Santo Souza Filho, Edson Takashi Matsubara

Laboratório de Inteligência Artificial - Faculdade de Computação
Universidade Federal de Mato Grosso do Sul - UFMS
Cidade Universitária, Av. Costa e Silva, 79070-900 - MS - Brazil

walter21souza@gmail.com, edson.matsubara@ufms.br

1 Introdução

No contexto de aprendizado de máquina, atribuição de autoria refere-se à tarefa de identificar ou atribuir a autoria de um determinado texto desconhecido a um autor específico, com base em características linguísticas, estilísticas e estruturais presentes no texto [1].

A ideia parte do pressuposto de que cada autor possui características intrínsecas de escrita, que podem ser aprendidas e discriminadas. Essa tarefa é comumente encontrada em áreas como análise de textos históricos, detecção de plágio e também de verificação de autenticidade de documentos.

Para realizar a atribuição de autoria por meio do aprendizado de máquina, é primeiro necessário um conjunto de textos conhecidos (conjunto de treinamento) atribuídos a autores específicos. Com base nesses dados rotulados, o objetivo é treinar um modelo capaz de identificar padrões ou características únicas na escrita de cada autor para predizer corretamente o autor de textos desconhecidos.

Apesar de ser uma ideia razoavelmente simples, a atribuição de autoria pode ser muito desafiadora na realidade, especialmente quando os textos disponíveis são curtos, quando existe uma sobreposição ou até mesmo manipulação no estilo de escrita feita pelo autor, e também textos que passaram por alguma revisão ou tradução. Todos esses fatores são desafios existentes nessa tarefa.

Sendo assim, o primeiro e principal passo para conseguir uma qualidade nos resultados, é justamente uma boa qualidade de dados, possuindo as características desejadas, para que se adeque da melhor forma ao modelo escolhido.

A proposta desse trabalho é justamente explorar uma nova adaptação utilizando um modelo de linguagem, porém, voltado para um contexto de Aprendizado de Métrica Profundo. Nessa abordagem, o modelo de linguagem deixa de servir como um classificador, e passa a ser usado como um backbone que irá aprender métricas baseadas nas características de interesse. Para auxiliar nesse aprendizado de métrica, utilizamos juntamente durante o processo a loss ArcFace[2]. Na sequência, tais características extraídas são utilizadas por um algoritmo que utilize métricas de distância, fornecendo como resultado, a classificação final.

2 Conjuntos de Dados

Dois grupos de datasets foram escolhidos para o caso de estudo: Evangelhos Sinóticos e Partidas de Xadrez.

2.1 Evangelhos Sinóticos

Evangelhos sinóticos é o nome dado aos evangelhos de Mateus, Lucas e Marcos, por possuírem uma grande parte de história comum entre eles, como mostrado na Figura. Para esse conjunto de dados, foi utilizado a versão da Bíblia ASV (American Standard Version).

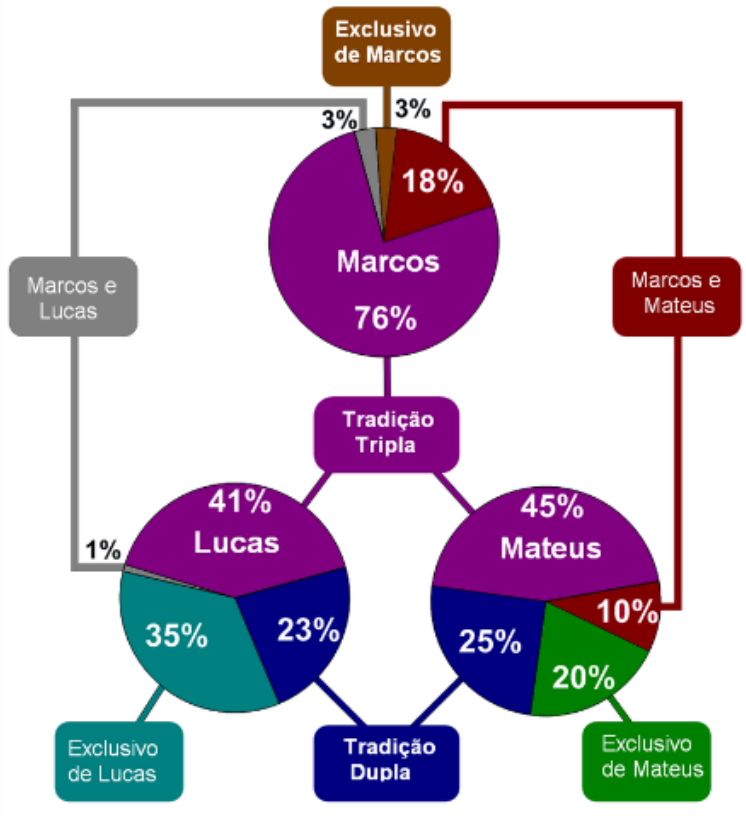

Figura 1: Relação entre os evangelhos sinóticos.

2.2 Partidas de Xadrez

Esse dataset foi construído a partir das partidas de xadrez de alguns jogadores. Tais partidas foram retiradas do conhecido portal de xadrez Chess.com. Uma partida pode ser interpretada por uma string contendo um código de posicionamento para cada jogada. Sendo assim, construímos essa string que representa a partida como uma sequência de jogadas, e atribuímos a ela o seu jogador (autor) correspondente.

3 Resultados

3.1 Evangelhos Sinóticos

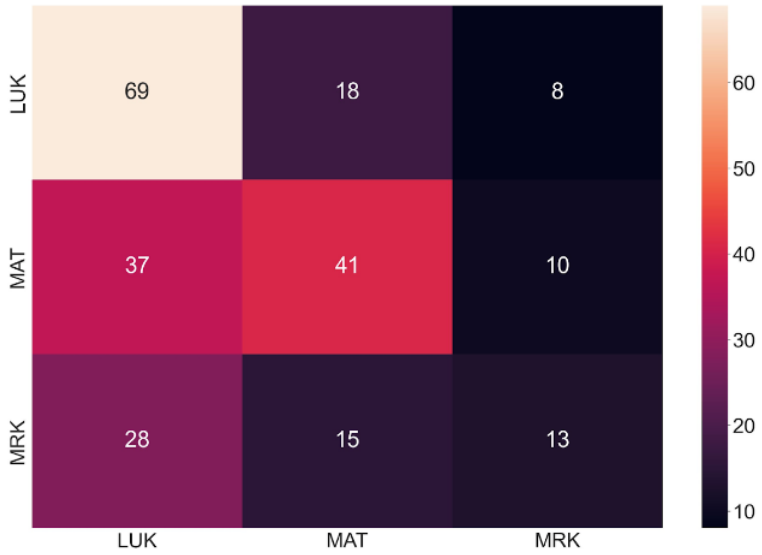

Figura 2: Matriz de Confusão para os evangelhos de Mateus, Lucas e Marcos.

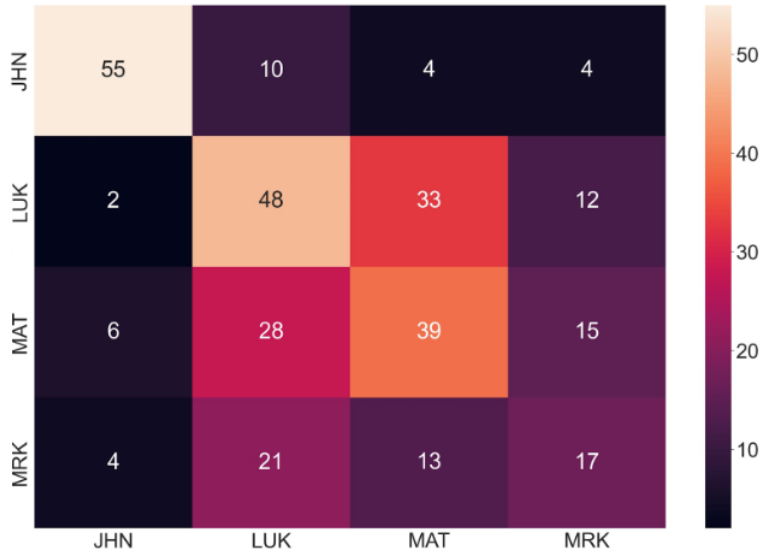

Figura 3: Matriz de Confusão para os evangelhos de Mateus, Lucas, Marcos e João.

3.2 Partidas de Xadrez

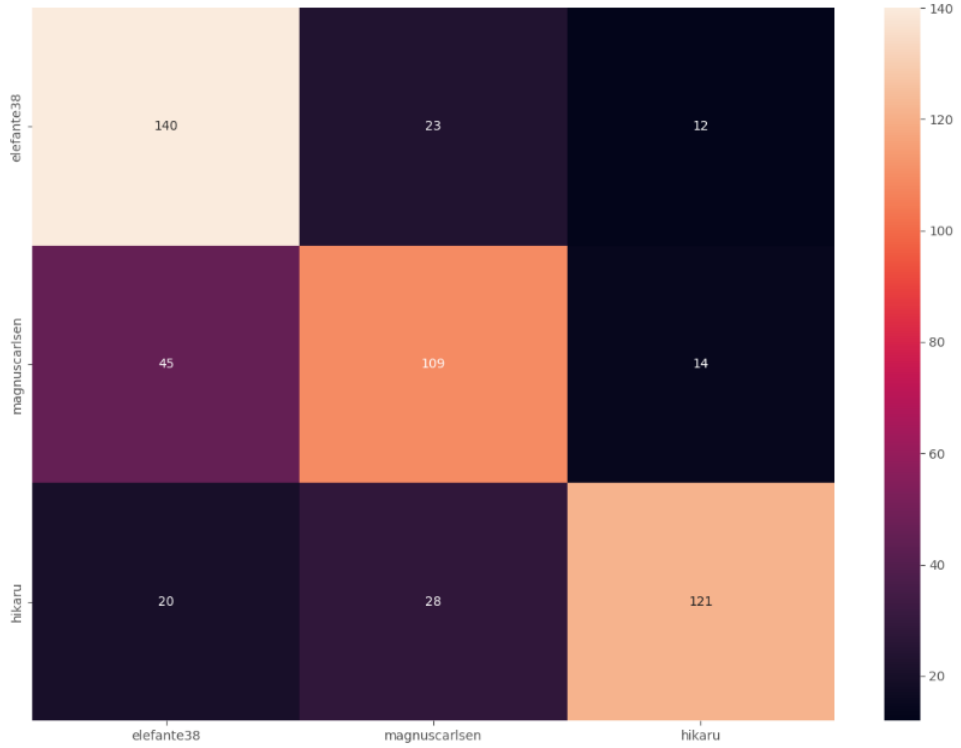

Figura 4: Matriz de Confusão para jogadores elefante38, magnuscarlsen, hikaru.

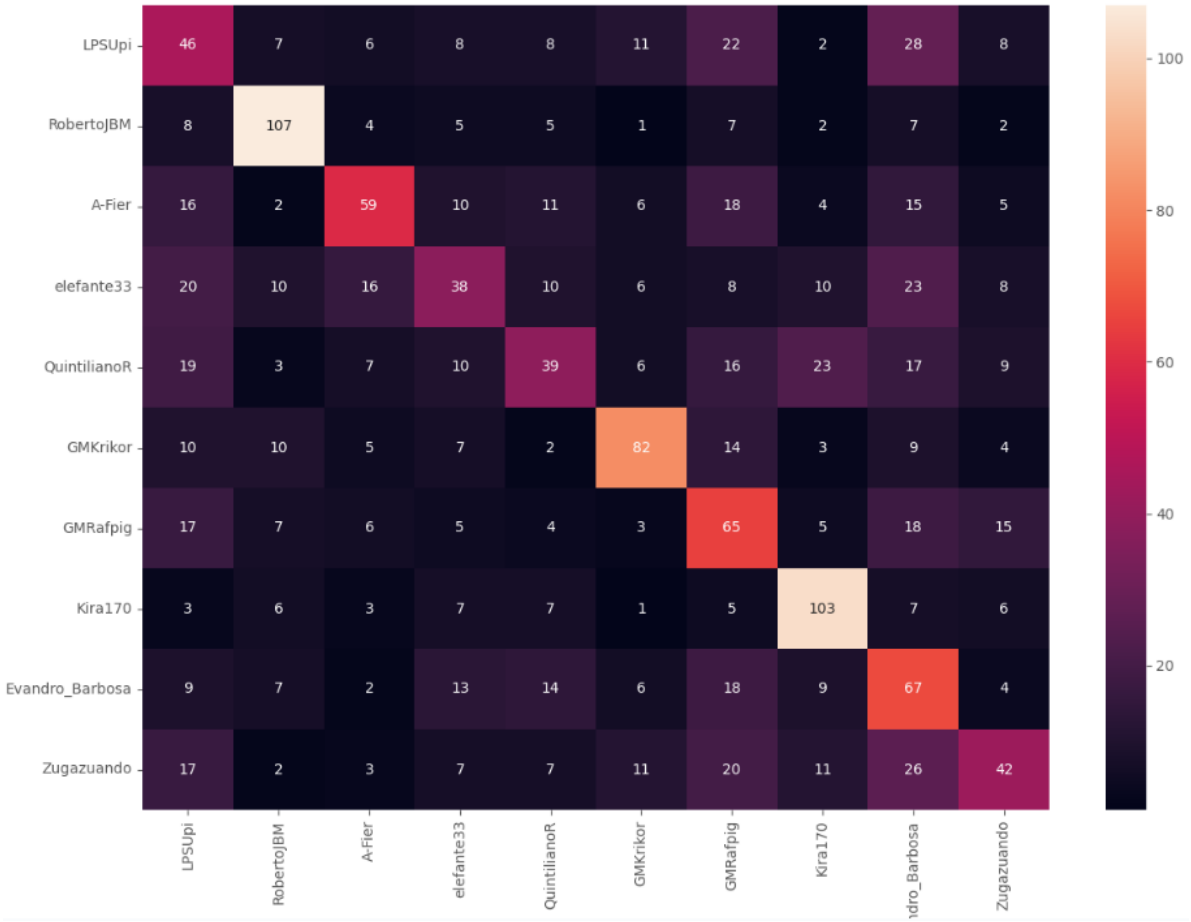

Figura 5: Matriz de Confusão para 10 jogadores Grão Mestre brasileiros.

4 Discussão dos Resultados

4.1 Evangelhos Sinóticos

Considerando os dados fornecidos pela matriz de confusão na Figura 2, e a relação entre os evangelhos sinóticos na Figura 1, é possível perceber que os resultados obtidos estão de certa forma de acordo com o esperado.

Pela Figura 1, Lucas possui uma boa porcentagem de exclusividade. Isso é refletido pela pontuação alta na matriz de confusão. Além disso, também possui uma baixa relação com Marcos e, novamente, isso é observável pela pontuação baixa na matriz de confusão.

De modo semelhante, uma análise semelhante pode ser refletida para o caso de Mateus, que possui uma relação menor com Marcos do que com Lucas. Novamente, isso é refletido na matriz de confusão. Entretanto, apesar da porcentagem de relação entre Mateus e Lucas ser maior do que a que é exclusiva de Mateus, a matriz de confusão mostra um resultado contrário. Isso pode ser explicado pela parte correspondente à Tradição Tripla, que é comum a todos os 3 evangelhos, criando uma distribuição um pouco diferente do esperado.

Já para o caso de Marcos, os resultados se aparentam mais inconclusivos. Isso se deve, novamente, ao fato da Tradição Tripla compor uma boa parte da totalidade desse evangelho. Dessa forma, é difícil verificar as relações criadas entre evangelhos escolhidos aos pares.

4.2 Partidas de Xadrez

Na Figura 4, é possível observar uma boa confusão entre os jogadores elefante38 e magnuscarlsen. Isso pode indicar que ambos os jogadores possuem um estilo de jogo mais similar, do que quando comparado ao jogador hikaru. Isso na verdade pode ser um resultado esperado, visto que magnuscarlsen é o atual campeão mundial. Sendo assim, o resultado poderia indicar que o elefante38 estudaria e usaria as técnicas e jogadas do atual campeão.

Já na Figura 5, uma outra ideia que podemos extrair dos resultados é a de previsão na mudança do ranking. Os jogadores estão em ordem ordenada de pontuação, sendo LPSupi o jogador com mais pontos, e Zugazuando com a menor quantidade de pontos entre os 10 jogadores. Por exemplo, o jogador QuintilianoR possui mais jogadas sendo confundidas com jogadores inferiores do que superiores. Isso mostraria uma tendência de movimentação no ranking para esse jogador em específico.

5 Conclusão

Este trabalho mostra que uma abordagem no âmbito do Aprendizado de Métrica Profundo para a Atribuição de Autoria é possível. Mesmo tendo alguns resultados inconclusivos, foi possível retirar algumas relações entre resultados esperados e obitidos. Além disso, é importante ressaltar que a tarefa de Atribuição de Autoria enfrenta desafios significativos, especialmente quando lida com textos curtos, estilo de escrita variável ou manipulações de tradução. A qualidade dos resultados está intrinsecamente ligada à qualidade e representatividade dos conjuntos de dados utilizados para treinamento, ressaltando a importância da disponibilidade de dados diversificados e abrangentes.

Referências

[1] Mohamed Amine Boukhaled e Jean-Gabriel Ganascia. "8 - Stylistic Features Based on Sequential Rule Mining for Authorship Attribution". Em: *Cognitive Approach to Natural Language Processing*. Ed. por Bernadette Sharp, Florence Sèdes e Wiesław Lubaszewski. Elsevier, 2017, pp. 159–175. ISBN: 978-1-78548-253-3. DOI: https://doi.org/10.1016/B978-1-78548-253-3.50008-1. URL: https://www.sciencedirect.com/science/article/pii/B9781785482533500081.

[2] Jiankang Deng et al. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”. Em: vol. 44. 10. Institute of Electrical e Electronics Engineers (IEEE), out. de 2022, pp. 5962–5979. DOI: 10.1109/tpami.2021.3087709. URL: https://doi.org/10.1109%2Ftpami.2021.3087709.