# Atribuição de autoria de comentários em português extraídos de fóruns de discussão online

**Pedro Semcovici**

**Prof. Dr. Luciano Antonio Digiampietri**

Escola de Artes, Ciências e Humanidades / Universidade de São Paulo

pedrosemcovici@usp.br, digiampietri@usp.br

## Objetivos

A atribuição de autoria é a tarefa de se identificar o autor de um item. Tipicamente aborda-se a autoria de um texto, mas o problema se aplica também a códigos-fonte, pinturas, composições etc. Ao longo das décadas, a atribuição de autoria foi utilizada para resolver diferentes problemas, desde verificar se um dado documento provavelmente não pertencia ao autor atribuído a ele, identificar os autores de textos dentro de um conjunto de pessoas que reivindicaram a autoria e até para a identificação de potenciais criminosos no contexto da linguística forense. No cenário das redes sociais online, a identificação de autoria costuma ser utilizada com dois objetivos principais: identificar se uma postagem provavelmente foi escrita pelo autor do comentário, ajudando, por exemplo, na detecção de notícias falsas ou para identificar se uma pessoa que teve a conta banida está tentando reingressar na rede social com uma nova conta, objetivando, assim, burlar o banimento. O objetivo do presente trabalho é utilizar técnicas de estilometria para a identificação da autoria em postagens em português textuais realizadas no fórum de discussão online Reddit.

## Métodos e Procedimentos

O presente estudo utiliza como corpus o conjunto de dados fornecido por Matias e Digiampietri (2023), composto por comentários extraídos dos subreddits *Brasil*, *brasillivre* e *Brasil do B*. Este conjunto de dados é composto por 1.000 comentários de 15 autores, totalizando 15.000 publicações.

Para a análise de atribuição de autoria, foram empregados três métodos distintos de classificação. O primeiro método adotou uma abordagem binária "autor x autor", com o objetivo de identificar a autoria de um determinado texto entre dois autores. O segundo método implementou uma classificação binária "autor x outros", buscando verificar se um texto é atribuível ou não a um autor específico. Por fim, o terceiro método utilizou uma classificação multiclasse para determinar a qual autor cada texto pertence.

Este estudo realizou três vertentes de experimentos para a classificação dos textos:

1. Utilização de n-gramas de *POS tags* e algoritmos não neurais.
2. Utilização dos *embeddings* BERT e algoritmos não neurais.
3. Utilização dos *embeddings* BERT com redes neurais recorrentes.

Foram empregados *postaggers* pré-treinados para rotular o corpus. Os modelos utilizados foram provenientes de portagger (Silva, E. H., 2023) e pt_core_news (Honnibal et al., 2022).

Para os experimentos 1 e 2, os algoritmos de classificação testados incluíram o *multinomial naive bayes*, regressão logística, *Support Vector Machine* (SVM), árvore de decisão, *Random Forest*, *Ada Boost* e *Gradient Boost*.

No experimento 3, foram empregados os algoritmos LSTM, GRU e SimpleRNN. Na classificação "autor x outros", observou-se um

desbalanceamento significativo das classes. Portanto, nos algoritmos neurais, foi aplicado o "*class weight*" para dar maior peso à classe minoritária. Enquanto nos demais, foi realizada uma etapa de *oversampling* utilizando RandomOverSampler.

## Resultados

Para o Experimento 1, entre os POS taggers testados, o pt_core_news_sm apresentou os melhores resultados em média para a classificação "autor x outros" e o problema multiclasse. Já para o problema "autor x autor", o pt_core_news_md foi o que obteve os melhores resultados, em média.

No Experimento 2, entre as classificações utilizando RNNs, observou-se que a vetorização com BERT Embeddings foi significativamente superior à vetorização com embedding layer e encoder. Além disso, entre as camadas recursivas testadas, a BI-LSTM destacou-se como a melhor para os problemas "autor x autor" e "autor x outros", enquanto para o problema multiclasse a BI-GRU foi a mais eficaz.

No Experimento 3, os classificadores que empregaram a média dos embeddings como características de entrada demonstraram que os embeddings do BERTimbau large foram os mais eficazes, enquanto o SVC foi identificado como o melhor classificador. Este modelo apresentou os melhores resultados de F1-score para os experimentos realizados neste estudo, conforme podem ser vistos na Tabela 1.

Tabela 1: Melhores resultados obtidos no estudo (com média de *BERT embeddings* e SVC)

| **Problema** | **F1-macro** |
|---|---|
| **autor x autor** | 89,11% |
| **autor x outros** | 76,94% |
| **multiclasse** | 55,65% |

## Conclusões

Neste trabalho de iniciação científica voluntária, foram aplicadas, a comentários em português, abordagens modernas de classificação e processamento de linguagem natural, preenchendo uma lacuna identificada na literatura que trata desse tema. Apesar de a atribuição de autoria ser uma área estudada há décadas, foi observada uma escassez de informações específicas sobre o uso das técnicas mais modernas, especialmente no contexto de postagens em língua portuguesa.

Além disso, este estudo proporciona uma comparação entre diferentes técnicas de *embeddings*, *POS tagging* e classificação para a solução do problema de atribuição de autoria de três maneiras distintas: "autor x autor", "autor x outros" e multiclasse. Essa análise contribui não apenas para o avanço do conhecimento nessa área, mas também para a aplicação prática dessas técnicas em situações reais, como a identificação de autoria em textos digitais em português.

## Referências


Matias, V.A. e Digiampietri, L.A. 2023. Authorship attribution of comments in Portuguese extracted from Reddit. Revista Brasileira de Computação Aplicada. 15, 2 (jul. 2023), 1-10.

Silva, E. H. (2023). Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo Universal Dependencies. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

Honnibal, M., Montani, I., Van Landeghem, S. and Boyd, A. (2022). spacy: Industrial-strength natural language processing in python.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12: 2825–2830.

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 18(17), 1–5.