

Atribuição autoral de textos digitais

José Eleandro Custódio

Programa de Mestrado em Sistemas de Informação - PPgSI
Universidade de São Paulo - USP

Novembro de 2018

- **Orientador:** Prof. Dr. Ivandré Paraboni
- **Semestre no curso:** 4o.
- **Qualificação:** 29/10/2018
- **Defesa:** realização planejada para 30/06/2019
- **Linha de pesquisa:** Inteligência de sistemas
- **Área de pesquisa:** Inteligência artificial
- **Área de aplicação:** Linguística computacional / Língua natural

Agenda

- 1 Introdução
- 2 Conceitos
- 3 Experimentos
- 4 Projeto de pesquisa
- 5 Referências

Contexto

- A atribuição autoral de textos digitais (AA) (do inglês, *Authorship Attribution*) visa identificar quem é o autor de um determinado texto a partir de um conjunto de autores possíveis (POTTHAST et al., 2017).
- A premissa principal da AA é que o autor deixa rastros de seu estilo, sendo que esses rastros podem ser a preferência por certas palavras, o tamanho do vocabulário, a utilização de pontuação e a repetição de certos elementos gramaticais.
- A quantificação do estilo de escrita, ou estilometria, compreende um vasto conjunto de medidas e técnicas que buscam extrair uma “biometria” textual (NEAL et al., 2017).

Aplicações da Atribuição Autoral

Aplicações da AA

Sua aplicação pode ajudar:

- em casos de escândalos de corrupção, como no caso Enron (KLIMT; YANG, 2004; CHEN et al., 2011).
- na identificação de abusos na utilização da internet (VARTAPETIANCE; GILLAM, 2012).
- na detecção de notícias falsas (PENG; CHOO; ASHMAN, 2016).
- na detecção de casos onde uma pessoa tenta se passar por outra (KOPPEL; SEIDMAN, 2018).
- na atribuição autoral de código-fonte (ALSULAMI et al., 2017)
- na detecção de pseudônimos (JUOLA, 2015)

Áreas de interesse

- Humanidades Digitais
- Análise Forense
- Linguística computacional

Métodos para atribuição autoral

Os métodos computacionais para atribuição autoral utilizam:

- Análise estatística multivariada (SAVOY, 2016; EVERT et al., 2017).
- Métodos baseados em vizinho mais próximo (KOCHER; SAVOY, 2017; KOPPEL; SEIDMAN, 2018; VARELA et al., 2016).
- Modelos de compressão (HALVANI; GRANER, 2018).
- Aprendizado de máquina com SVM (SCHWARTZ et al., 2013; STAMATATOS, 2017).
- Redes neurais recorrentes (BAGNALL, 2016).
- Redes neurais de convolução (SHRESTHA et al., 2017; SARI; STEVENSON, 2016).

Conceitos - Fatores que influenciam a AA

Canal

E-mail, jornais, livros, SMS

Textos mais ou menos formais.

Idioma

Complexidade morfológica e lexical diferentes.

Tópico

Economia, celebridades, dia-a-dia

Influencia o vocabulário.

Domínio ou Gênero do texto

Contos, romances, artigos

Influencia no rigor formal e no vocabulário.

Tamanho do texto

Métodos probabilísticos são afetados pelo número de observações formais.

Número de autores

O aumento do número de classes requer o aumento do número de classes.

Conceitos - Subtarefas da análise autoral

AA de conjunto fechado

Os textos do conjunto de teste pertencem a um dos autores candidatos presentes no cópús de treinamento.

AA de conjunto aberto

Os textos do conjunto de teste não necessariamente foram escritos por um dos autores do cópús de treinamento

K-Atribuição ou ordenação

As saídas do classificador são ordenadas pela probabilidade e são retornados os K autores mais prováveis.

Caracterização

São extraídas informações demográficas do autor do texto podem reduzir a lista de candidatos.

Verificação

Verifica-se se dois documentos foram escritos pelo mesmo autor, não sendo necessário saber quem são os autores.

Demais

Agrupamento, Ligação e Quebra de estilo.

Conceitos - Tipos de conhecimentos usados

A abordagem estilométrica tradicional utiliza as seguintes fontes de conhecimento:

- **Categoria lexical:** tamanho médio das palavras, número de letras maiúsculas, tamanho das sentenças, etc.
- **Categoria sintática:** frequência da pontuação, palavras de função, etc.
- **Categoria semântica:** contagem das palavras, analisadores semânticos, *word embeddings*, etc.
- **Categoria específica de domínio:** palavras-chave, *tags* HTML, *emoticons*, etc.

Conceitos - Tipos de conhecimentos usados

Palavras

As palavras mais frequentes de um texto são independentes de domínio e utilizadas de forma inconsciente (KESTEMONT, 2014).

Palavras de função (do inglês, *function words*) compreende artigos, preposições, locuções adverbiais, e outros.

Caracteres

As sequências de caracteres são considerados os modelos mais efetivos para AA (KJELL; WOODS; FRIEDER, 1994; NEAL et al., 2017) e uma contagem simples pode produzir modelos próximos ao estado-da-arte (NEAL et al., 2017).

Os **caracteres mais frequentes (CNG)** (do inglês, *common n-grams*) (KEŠELJ et al., 2003; SAPKOTA et al., 2014).

Experimento 1: Verificação autoral

Publicação 1

CUSTÓDIO, J. E.; PARABONI, I. Similaridade de Textos aplicada à Verificação Autoral. In: 1st International Congress on Digital Humanities in Rio de Janeiro. [S.l.]: Fundação Getúlio Vargas, 2018.

Verificação autoral ou atribuição por similaridade

- Deseja-se saber se pares de documentos foram escritos pelo mesmo autor. (KOPPEL et al., 2012)
- Aplicável quando não se sabe quem são os autores.
- Modelo supervisionado por vizinho mais próximo.
 - O documento é atribuído ao vizinho mais próximo.
 - A distância pode ser usada no agrupamento autoral.
- Modelo transformado
 - Documentos são uma representação única.

Experimento 1: Verificação autoral

Extração de características

Modelo de espaço de vetores (BOW) com n-gramas de caracteres normalizados com norma L1 (TF).

Foram selecionadas os n-gramas presentes em 90% do cópulus (*Common n-grams* (KEŠELJ et al., 2003)).

Distâncias

Medidas de similaridade textual entre os documentos A e B do cópulus C:

$$\text{Cossenos}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (1)$$

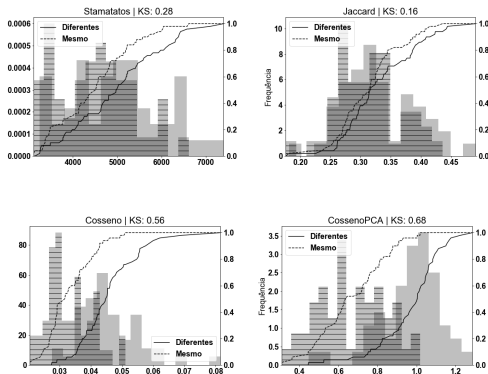
$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

$$\text{Stamatatos}(A, B, N) = \sum_i \left(\frac{2 * (A - B)}{A + B} \right)^2 * \left(\frac{2 * (A - C)}{A + C} \right)^2 \quad (3)$$

Experimento 1: Verificação autoral

Análise da capacidade de separação das medidas de similaridade aplicadas a corpus PAN-CLEF 2014 (STAMATATOS et al., 2014)

Figure 1: Diagnósticos



- Histograma para mesma autoria (tracejado).
- Histograma para autorias diferentes (liso).
- Distribuição acumuladas (linhas).
- Separação pela métrica Kolmogorov-Smirnov.
- Métricas AUC e acurácia.

Experimento 1: Verificação autoral

Modelo proposto 1 - MP1

- As distâncias foram utilizadas como variáveis para o modelo.
- Aplicado a normalização minmax.
- Aplicado a regressão logística.

Modelo proposto 2 - MP2

- Os documentos conhecidos C e de autoria desconhecidas D foram unificados em um único BoW através da equação:

$$MP2(C_{ij}, D_{ij}) = \log \left(1 + \frac{(C_{ij} - D_{ij})^2}{C_{ij} + 1} \right) \quad (4)$$

- Aplicado a normalização minmax.
- Aplicado a regressão logística.

Experimento 1: Verificação autoral

Table 1: Verificação autoral - Resultados médios das métricas AUC e acurácia em 5-partições.

Modelo	PAN2014 (EE e EM)		PAN2014-SP	
	ROC	Acurácia	ROC	Acurácia
Jaccard	0,60	0,56	0,57	0,52
Cossenos	0,63	0,50	0,88	0,77
Cossenos_PCA	0,63	0,55	0,92	0,83
Keselj	0,61	0,54	0,71	0,60
Stamatatos	0,60	0,55	0,59	0,54
MP1 – Mix	0,75	0,67	0,72	0,62
MP2 – BOW	0,62	0,53	0,93	0,85

PAN2014 (EE e EM) corpus com textos em língua inglesa, PAN2014-SP textos em língua espanhola.

Experimento 2: Atribuição Autoral

Publicação 2

CUSTÓDIO, J. E.; PARABONI, I. EACH-USP Ensemble Cross-domain Authorship Attribution: Notebook for PAN at CLEF 2018. In: CAPPELLATO, L. et al. (Ed.). Working Notes Papers of the CLEF 2018 Evaluation Labs. [S.l.]: CLEF and CEUR-WS.org, 2018. (CEUR Workshop Proceedings). ISSN 1613-0073.

Atribuição por aprendizado de máquina supervisionado

- Tem-se um conjunto de documentos para os quais se sabe quem são os autores e um documento do qual deseja-se atribuir.
- O classificador extrai a “assinatura do estilo”.
- Aspectos inconscientes, como a sintaxe, são mais importantes que a semântica.
- O trabalho apresentado foi parte da participação da tarefa de AA da competição PAN-CLEF2018.

Experimento 2: Atribuição Autoral

Baseline *Bas.PAN*

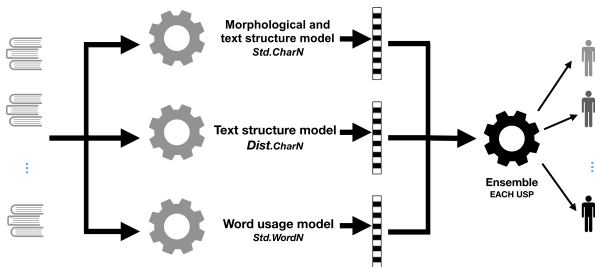
Os organizadores forneceram um sistema *baseline* pelos com as seguintes características:

- N-gramas de caracteres de tamanho fixo.
- Normalização no documento por TF.
- Sem normalização no cópús.
- Frequência mínima de 4 ocorrências.
- Classificador SVM encapsulado nas estratégias um-contra-um e um-contra-todos.
- Foi otimizado por *grid search* com validação cruzada com 5 partições, e os melhores parâmetros foram:
 - n-gramas de tamanho 4.
 - Frequência mínima de 5 documentos.
 - SVM com estratégia um-contra-todos.

Experimento 2: Atribuição Autoral

Dado nossas premissas, o sistema final PAN2018 para AA consistiu de um comitê que concatenou as fontes de informações em uma saída única.

Figure 2: Método proposto final



O sistema foi otimizado por *grid search* com validação cruzada de 5 partições.

Experimento 2: Atribuição Autoral

Premissa: O estilo de escrita de autor pode ser capturado através de diversas fontes de informação, como sintática, léxica e semântica.

Método proposto *Std.word*

Consistiu de um modelo BOW de n-gramas de palavras tradicional.

Método proposto *Std.char*

Consistiu de um modelo BOW de n-gramas de caracteres tradicional.

Método proposto *Dist.char*

Consistiu de um modelo BOW de n-gramas de caracteres onde são letras maiúscula e minúsculas sem acento são distorcidas, deixando a pontuação, espaços e letras com diacríticos.

Experimento 2: Atribuição Autoral

Table 2: Valores ótimos encontrados para PAN2018

Módulo	Parâmetros	Valores ótimos
Extração de características	Faixa n-gram	Std.charN - Início=2 Fim=5 Dist.charN - Início=2 Fim=5 Std.WordN - Início=1 Fim=3
	Freq. min. doc.	0,05
	Freq. max. doc.	1,0
	TF	Sublinear
	IDF	Suavizado
	Normalização no documento	L2
Transformação	PCA	0,99

Experimento 2: Resultados obtidos em treinamento

Table 3: F1 para PAN-CLEF 2018 AA no corpus de desenvolvimento

Problema	Língua	Autores	Bas.PAN	Std.charN	Dist.charN	Std.wordN	Comitê
01	EN	20	0,514	0,609	0,479	0,444	0,625
02	EN	5	0,626	0,535	0,333	0,577	0,673
03	FR	20	0,631	0,681	0,568	0,418	0,776
04	FR	5	0,747	0,719	0,586	0,572	0,820
05	IT	20	0,529	0,597	0,491	0,497	0,578
06	IT	5	0,614	0,623	0,595	0,520	0,663
07	PL	20	0,455	0,470	0,496	0,475	0,554
08	PL	5	0,703	0,948	0,570	0,922	0,922
09	ES	20	0,709	0,774	0,589	0,616	0,701
10	ES	5	0,593	0,778	0,802	0,588	0,830
Média			0,612	0,673	0,551	0,563	0,714

Experimento 2: Resultados obtidos em produção

Resultado geral apresentados pelos organizadores do PAN2018 (KESTEMONT et al., 2018).

Table 4: PAN-CLEF 2018 - 3 melhores equipes - por língua

Equipe	F1 Geral	EN	FR	IT	PL	ES
Custódio e Paraboni (2018)	0,685	0,744	0,668	0,676	0,482	0,856
Murauer, Tschuggnall e Specht (2018)	0,643	0,762	0,607	0,663	0,450	0,734
Halvani e Graner (2018)	0,629	0,679	0,536	0,752	0,426	0,751
PAN18-BASELINE	0,584	0,697	0,585	0,605	0,419	0,615

Table 5: PAN-CLEF 2018 - 3 melhores equipes - por língua

Equipe	Quantidade de autores			
	20	15	10	5
Custódio e Paraboni (2018)	0,648	0,676	0,739	0,677
Murauer, Tschuggnall e Specht (2018)	0,609	0,642	0,680	0,642
Halvani e Graner (2018)	0,609	0,605	0,665	0,636
PAN18-BASELINE	0,546	0,532	0,595	0,663

Lacunas e motivação

Lacunas gerais:

- A AA é um problema de pesquisa não totalmente resolvido (POTTHAST et al., 2017).
- É o tema da série de competições PAN-CLEF (KESTEMONT et al., 2018) (KOCHER; SAVOY, 2017; KOPPEL; SEIDMAN, 2018; VARELA et al., 2016).
- Estudos desta área exploram técnicas independentes de idioma e de domínio, subutilizando recursos linguístico-computacionais.

O trabalho em Custódio e Paraboni (2018) apresentou o melhor desempenho global na edição de 2018 da competição PAN-CLEF, no entanto, deixa as seguintes lacunas:


- não tirou proveito de conhecimentos dependentes de idioma como *part-of-speech* (POS).
- e modelos de representação distribuída (*word embeddings*).
- foi restrito ao domínio *Fanfic*.
- não considerou dados em português brasileiro.


Hipóteses

Este trabalho considera as seguintes hipóteses:


- H1: O uso de modelos independentes de idioma do tipo de distorção textual permite filtrar aspectos específicos do texto, e a combinação de diversos tipos de distorção pode aumentar o desempenho de sistemas de AA.
- H2: O uso de modelos dependentes de idioma do tipo *part-of-speech* extraídos por anotadores baseados em aprendizado profundo pode aumentar o desempenho de sistemas de AA.
- H3: O uso de modelos dependentes de idioma do tipo representação distribuída (*embeddings*) pode aumentar o desempenho de sistemas de AA.


Estas hipóteses serão testadas comparando-se os modelos propostos com sistemas de *baseline* que se façam pertinentes como o próprio modelo apresentado em Custódio e Paraboni (2018). A avaliação será feita por meio de medidas tradicionais em aprendizado de máquina, como a medida F, acurácia e outros. Espera-se que o resultado médio seja superior ao dos modelos de *baseline* de acordo com as métricas estipuladas.

 ALSULAMI, B. et al. Source code authorship attribution using long short-term memory based networks. In: *Computer Security - ESORICS 2017 - 22nd European Symposium on Research in Computer Security, Oslo, Norway, September 11-15, 2017, Proceedings, Part I*. [s.n.], 2017. p. 65–82. Disponível em: https://doi.org/10.1007/978-3-319-66402-6_6.

 BAGNALL, D. Authorship clustering using multi-headed recurrent neural networks. In: Cappellato L. Ferro N., M. C. B. K. (Ed.). *CEUR Workshop Proceedings*. [S.l.]: CEUR-WS, 2016. v. 1609, p. 791–804. ISSN 16130073.


Referências II


 CHEN, X. et al. Authorship similarity detection from email messages. In: *Machine Learning and Data Mining in Pattern Recognition - 7th International Conference, MLDM 2011, New York, NY, USA, August 30 - September 3, 2011. Proceedings.* [S.l.: s.n.], 2011. p. 375–386.


 CUSTÓDIO, J. E.; PARABONI, I. EACH-USP Ensemble Cross-domain Authorship Attribution: Notebook for PAN at CLEF 2018. In: CAPPELLATO, L. et al. (Ed.). *Working Notes Papers of the CLEF 2018 Evaluation Labs.* [S.l.]: CLEF and CEUR-WS.org, 2018. (CEUR Workshop Proceedings). ISSN 1613-0073.

 EVERT, S. et al. Understanding and explaining delta measures for authorship attribution. *Digital Scholarship in the Humanities*, v. 32, n. suppl_2, p. ii4–ii16, 2017.


Referências III


 HALVANI, O.; GRANER, L. Cross-Domain Authorship Attribution Based on Compression: Notebook for PAN at CLEF 2018. In: CAPPELLATO, L. et al. (Ed.). *Working Notes Papers of the CLEF 2018 Evaluation Labs*. [S.l.]: CLEF and CEUR-WS.org, 2018. (CEUR Workshop Proceedings). ISSN 1613-0073.


 JUOLA, P. The rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanitie*, v. 30, n. Suppl-1, p. i100–i113, 2015.

 KEŠELJ, V. et al. N-Gram-Based Author Profiles for Authorship Attribution. In: *Proceedings of the conference pacific association for computational linguistics (PACLING)*. [S.l.: s.n.], 2003. v. 3, p. 255–264.


Referências IV


 KESTEMONT, M. Function Words in Authorship Attribution From Black Magic to Theory ? *3rd Workshop on Computational Linguistics for Literature (CLfL 2014)*, n. January 2014, p. 59–66, 2014.


 KESTEMONT, M. et al. Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In: CAPPELLATO, L. et al. (Ed.). *Working Notes Papers of the CLEF 2018 Evaluation Labs*. [S.l.]: CLEF and CEUR-WS.org, 2018. (CEUR Workshop Proceedings). ISSN 1613-0073.


 KJELL, B.; WOODS, W. A.; FRIEDER, O. Discrimination of authorship using visualization. *Inf. Process. Manage.*, v. 30, n. 1, p. 141–150, 1994.

Referências V


 KLIMT, B.; YANG, Y. The enron corpus: A new dataset for email classification research. In: BOULICAUT, J. et al. (Ed.). *Machine Learning: ECML 2004, 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004, Proceedings*. [S.l.]: Springer, 2004. (Lecture Notes in Computer Science, v. 3201), p. 217–226.

 KOCHER, M.; SAVOY, J. A simple and efficient algorithm for authorship verification. *Journal of the Association for Information Science and Technology*, v. 68, n. 1, p. 259–269, 2017.


 KOPPEL, M. et al. The "Fundamental Problem" of Authorship Attribution. *English Studies*, v. 93, n. 3, p. 284–291, 2012. ISSN 0013838X.

 KOPPEL, M.; SEIDMAN, S. Detecting pseudepigraphic texts using novel similarity measures. *Digital Scholarship in the Humanities*, v. 33, n. 1, p. 72–81, 2018.


Referências VI


 MURAUER, B.; TSCHUGGNALL, M.; SPECHT, G. Dynamic Parameter Search for Cross-Domain Authorship Attribution: Notebook for PAN at CLEF 2018. In: CAPPELLATO, L. et al. (Ed.). *Working Notes Papers of the CLEF 2018 Evaluation Labs*. [S.l.]: CLEF and CEUR-WS.org, 2018. (CEUR Workshop Proceedings). ISSN 1613-0073.

 NEAL, T. J. et al. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, v. 50, n. 6, p. 86:1–86:36, 2017.


 PENG, J.; CHOO, K. kwang R.; ASHMAN, H. Astroturfing detection in social media: Using binary n-gram analysis for authorship attribution. In: *2016 IEEE Trustcom/BigDataSE/ISPA*. [S.l.: s.n.], 2016. p. 121–128. ISBN 9781509032051.

Referências VII


 POTTHAST, M. et al. Overview of PAN'17: Author identification, author profiling, and author obfuscation. *Lecture Notes in Computer Science*, v. 10456 LNCS, p. 275–290, 2017. ISSN 16113349.

 SAPKOTA, U. et al. Cross-Topic Authorship Attribution: Will Out-Of-Topic Data Help? In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. [S.l.: s.n.], 2014. p. 1228–1237. ISBN 9781941643266.


 SARI, Y.; STEVENSON, M. Exploring Word Embeddings and Character N -Grams for Author Clustering Notebook for PAN at CLEF 2016. *CEUR Workshop Proceedings*, 2016. ISSN 16130073.

 SAVOY, J. Estimating the probability of an authorship attribution. *Journal of the Association for Information Science and Technology*, v. 67, n. 6, p. 1462–1472, 2016. ISSN 23301643.


Referências VIII


 SCHWARTZ, R. et al. Authorship Attribution of Micro-Messages. In: *Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2013. p. 1880–1891. ISBN 9781937284978.


 SHRESTHA, P. et al. Convolutional Neural Networks for Authorship Attribution of Short Texts. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. [S.l.]: Association for Computational Linguistics (ACL), 2017. v. 2, p. 669–674. ISBN 9781510838604.

 STAMATATOS, E. Authorship attribution using text distortion. *Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, v. 1, 2017.

Referências IX

 STAMATATOS, E. et al. Overview of the author identification task at PAN 2014. In: *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*. [S.l.: s.n.], 2014. p. 877–897.

 VARELA, P. J. et al. A computational approach based on syntactic levels of language in authorship attribution. *IEEE Latin America Transactions*, v. 14, n. 1, p. 259–266, 2016. ISSN 15480992.

 VARTAPETIANCE, A.; GILLAM, L. Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification. In: *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. [S.l.: s.n.], 2012.