

Atribuição autoral de textos digitais

José Eleandro Custódio

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

Novembro de 2018

Informações gerais

- **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 Trabalhos relacionados
- 4 Experimentos
- 5 Projeto de pesquisa
- 6 Referências

Tema

- 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).

Projeto - 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).
- 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.

Projeto - 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.

Projeto - Objetivo

Objetivo Geral

O objetivo geral deste trabalho é enriquecer modelos de atribuição autoral de texto digitais com conjunto fechado de autores utilizando conhecimentos dependentes e independentes de idioma combinados com técnicas de aprendizados de máquina, de modo a obter resultados superiores ao estabelecido em trabalhos anteriores.

Definições

- A atribuição autoral (AA) é uma técnica computacional que visa identificar o autor de um texto a partir de um conjunto de autores possíveis, baseando-se em padrões de estilo deixados pelos autores (POTTHAST et al., 2017).
- Do ponto de vista de aprendizado de máquina, a AA pode ser vista como um problema de classificação multi-classes (STAMATATOS, 2009).
- 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).

Apesar da similaridade com a tradicional tarefa de classificação de documentos, a AA busca elementos inconscientes da escrita que sejam independentes do conteúdo semântico do texto (KEŠELJ et al., 2003).

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, artigos, avaliações de produtos
- Influencia no rigor formal e no vocabulário.

Tamanho do texto

- Métodos probabilísticos são afetados pelo número de observações.

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 córpus de treinamento.

AA de conjunto aberto

Os textos do conjunto de teste não necessariamente foram escritos por um dos autores do córpus 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, quantidade de dígitos, tamanho das sentenças, etc.
- **Categoria sintática:** frequência da pontuação, palavras de função, frases começando com maiúscula, etc.
- **Categoria semântica:** contagem das palavras, analisadores semânticos, *word embeddings*, etc.
- **Categoria estrutura:** indentação, tamanho do parágrafo, etc.
- **Categoria específica de domínio:** palavras-chave, *tags* HTML, *emoticons*, nomes de produtos.

Conceitos - Tipos de conhecimentos usados

Outra classificação possível e simplificada dos conhecimentos utilizados na AA pode ser a utilização das famílias baseadas em palavras e caracteres.

Palavras

- As palavras mais frequentes 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.
- Capturam semântica e elementos de conexão entre sentenças.
- Diversas ferramentas são preparadas para usar a unidade *palavra*.

Conceitos - Tipos de conhecimentos usados

Caracteres

- As sequências de caracteres são considerados os modelos mais efetivos para AA (KJELL; WOODS; FRIEDER, 1994; NEAL et al., 2017).
- Os **caracteres mais frequentes (CNG)** (do inglês, *common n-grams*) (KEŠELJ et al., 2003; SAPKOTA et al., 2014).
- São independentes de idioma.
- Não precisam de *stemming* pois lidam bem com idiomas flexionais.
- Geram vetores de contagem mais densos que os vetores de palavras.
- Os n-gramas de caracteres conseguem capturar pontuação, utilização de espaços, preferências temporais, palavras de função de tamanho curto.
- O trabalho em Sapkota et al. (2015) mostra que apesar de independente de idioma nem todos os *char n-gramas* tem a mesma origem.

Modelo tradicional de representação textual - BOW

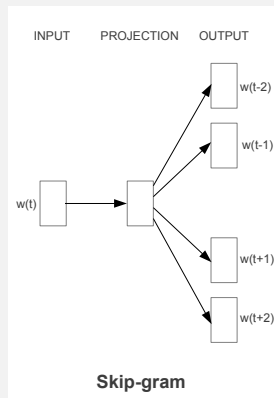
- Hipótese distribucional aplicada a documentos (TURNERY; PANTEL, 2010)
- Modelo de n-gramas
- Conhecimentos: Caracteres, Palavras, POS
- TF-IDF
- TF-IDF equações alternativas utilizadas nos experimentos:
 - $TF_{sublinear} = 1 + \log TF_{t,d} \rightarrow$ Definição I do SMART.
 - $IDF_{Suavizado}(t, D) = \log \left(\frac{D}{DF(t)} \right) + 1$

Conceitos - Modelos computacionais

Modelo de representação distribuída

- Hipótese distribucional aplicada símbolos
- Modelos neurais de língua natural (BENGIO et al., 2003)
- Aprendizado não supervisionado
- Transferência de conhecimento
- *Word embeddings*
 - Representa cada palavra, ou símbolo, em k dimensões, sendo k menor que o tamanho do vocabulário.
 - Word2Vec (MIKOLOV et al., 2013)
 - Doc2Vec (LE; MIKOLOV, 2014)
 - FastText (BOJANOWSKI et al., 2017)

Modelo Word2Vec



Fonte: Mikolov et al. (2013)

Modelos computacionais baseados em distância

As medidas, ou funções, de distância devem obedecer quatro propriedades (DEZA; DEZA, 2009):

- 1 $D(A, B) \geq 0$ para todo A e B , D é positiva.
- 2 $D(A, B) = 0$ se e, somente se, $A = B$.
- 3 $D(A, B) = D(B, A)$, D é uma função simétrica.
- 4 $D(A, C) \leq D(A, B) + D(B, C)$, a desigualdade triangular.

Distância de cosseno

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

Manhattan

$$\text{Manhattan}(A, B) = \sum_i |A_i - B_i|$$

Jaccard

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

Interseção binária

MinMax, ou Jaccard generalizado

$$\text{MinMax}(A, B) = \frac{\sum_i^n \min(A_i, B_i)}{\sum_i^n \max(A_i, B_i)}$$

Modelos computacionais baseados em distância

Regra Δ de Burrows

$$\Delta_{Burrow}(A, B) = \frac{1}{N} \sum_{i=1}^N |Zscore(A_i, C_i) - Zscore(B_i, C_i)| \quad (1)$$
$$Zscore(X_i, C_i) = \frac{X_i - \mu(C_i)}{\sigma(C_i)}$$

Distância de Manhattan dos Z-score das frequências

Keselj

$$Keselj(A, B) = \sum_i^n \left(\frac{2 * (A_i - B_i)}{A_i + B_i} \right)^2$$

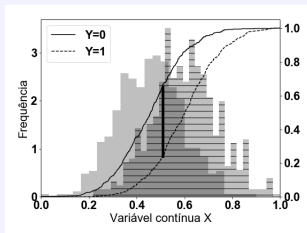
Stamatatos

$$Stamatatos(A, B, C) = \sum_i^n \left(\frac{2 * (A_i - B_i)}{A_i + B_i} \right)^2 * \left(\frac{2 * (A_i - C_i)}{A_i + C_i} \right)^2$$

Onde A e B são vetores de frequências documentos que se desejam comparar e C é o vetor de frequência do corpus.

Modelos computacionais baseados em distância

Distância Komogorov-Smirnov



Fonte: José Eleandro Custódio, 2018

Definição e propriedades

- $KS(X, Y) = \arg.\max_x \text{abs}(CDF(x|y = 0) - CDF(x|y = 1))$
- Distância entre as curvas de probabilidades acumuladas.
- Mede se a variável binária representa distribuições diferentes.

Modelos de aprendizado de máquina (AM)

Regressão logística softmax

$$\begin{array}{ccc} f_1(X) - \ln Z & = & \ln P(Y = 1|X) \\ \dots & & \dots \end{array} \quad (2)$$

$$f_c(X) - \ln Z = \ln P(Y = c|X)$$

⇓

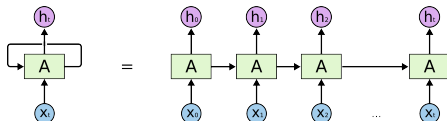
$$P(Y = c|X) = \text{softmax}(X, c) = \frac{e^{f_c(X)}}{\sum_{k=1}^K e^{f_k(X)}} \quad (3)$$

Propriedades

- Não assume a independência das variáveis.
- Possui saída probabilística.
- As probabilidades refletem o balanceamento das classes.
- Possui saída contínua.
- É um classificador linear.
- É usado em aprendizado profundo.

Modelos de aprendizado de máquina (AM)

Figure 2: Expansão de um neurônio de uma rede recorrente



Fonte: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

- Redes neurais recorrentes (RNN) (ELMAN, 1990), são uma família de redes neurais que fazem o processamento de dados sequenciais, onde a dependência temporal é importante para o resultado (GOODFELLOW; BENGIO; COURVILLE, 2016).
- Devido à recorrência, um neurônio codifica uma sequência temporal.
- O estudo em Mikolov (2012) traz aplicações das RNNs aplicadas a modelagem da língua.

Trabalhos relacionados

Foram estudados os trabalhos:

- relacionados ao histórico da competição PAN-CLEF.
- trabalhos recentes encontrados no Scopus, IEEE e ACL Anthology
- que usaram n-gramas, embeddings e variações.
- que usaram técnicas de similaridade/distância/vizinhos mais próximos.
- que usaram métodos de distância específicos para AA.
- que usaram RNN e CNN.
- que analisaram línguas europeias.

Table 1: Trabalhos selecionados

Estudo	Idioma	Tarefa	Conhecimento	Método
Sapkota et al. (2015)	EN	A	<i>C</i>	SVM
Stamatatos (2017)	EN	A,V	<i>C, W</i>	SVM
Schwartz et al. (2013)	EN	A	<i>C, W</i>	SVM
Rocha et al. (2017)	EN	A,V	<i>C, W, P</i>	SVM, RF e SCAP
Evert et al. (2017)	EN	C	<i>W</i>	Clusterização
Varela et al. (2016)	PT-BR	A,V	<i>P</i>	SVM
Posadas-Durán et al. (2017)	EN	A	D2V de <i>W</i>	Softmax e SVM
Rhodes (2015)	EN	A	W2V	CNN-Softmax
Shrestha et al. (2017)	EN	A	<i>C One-hot</i>	Softmax
Bagnall (2016)	PAN2015	C	<i>C One-hot</i>	RNN-Softmax

José Eleandro Custódio, 2018

Trabalhos relacionados - Considerações

- Os métodos de similaridade representam uma ferramenta importante para AA.
- As **redes de convolução** apresentaram resultados equivalentes aos *baselines*, entretanto, apresentaram custo computacional maior.
- As **redes recorrentes** apresentaram desempenho superiores ao *baseline*, no entanto, apresentaram custo computacional elevado e precisou de dados adicionais.
- **Embeddings pré-treinados** obtiveram desempenho equivalente ao *baseline*.
- As técnicas de aprendizado profundo possuem aplicações limitadas na AA porque nem sempre é possível tem um volume de dados expressivo de forma a garantir a estabilidade dos métodos.
- Os modelos baseados em caracteres representaram desempenho consistentes.
- Os modelos que usam distorção se demonstraram promissores.

Experimentos

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:

$$\text{Cossenos}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (4)$$

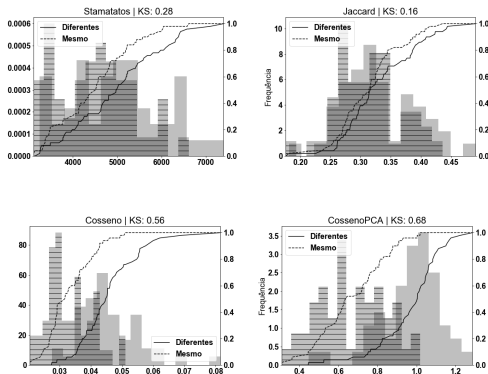
$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (5)$$

$$\text{Stamatatos}(A, B, N) = \sum_i \left(\frac{2 * (A - B)}{A + B} \right)^2 * \left(\frac{2 * (A - C)}{A + C} \right)^2 \quad (6)$$

Experimento 1: Verificação autoral

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

Figure 3: 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:

$$MP2(C_{ij}, D_{ij}) = \log \left(1 + \frac{(C_{ij} - D_{ij})^2}{C_{ij} + 1} \right) \quad (7)$$

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

Experimento 1: Verificação autoral

Table 2: 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*

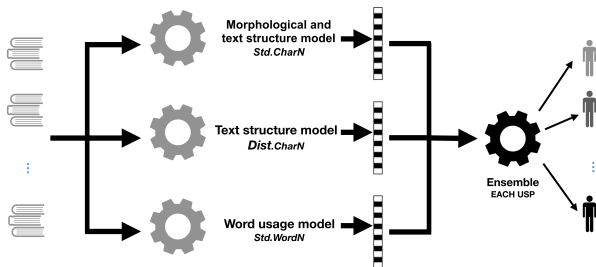
Os organizadores forneceram um sistema *baseline* pelos com as seguintes características:

- N-gramas de caracteres de tamanho fixo.
- Normalização no documento por TF.
- Sem normalização no *cópus*.
- 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 4: 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 - Modelo de distorção textual em caracteres

O modelo *Dist.charN* utiliza distorção textual e foi inspirado em Stamatatos (2017). O objetivo da distorção é mascarar conteúdos que estariam se comportando como ruído para o modelo. Stamatatos (2017) utiliza distorções para mascarar palavras menos frequentes, que estariam relacionadas ao tópico. O modelo *Dist.charN* atua sobre carácter, filtrando a utilização de espaços, pontuação e caracteres com diacríticos, de forma a mascarar caracteres comuns, ou seja, letras minúsculas e maiúsculas. A tabela 3 ilustra a aplicação desse método.

Table 3: Exemplo de distorção de texto aplicado o 1o. documento do 9o. problema da base de treinamento.

Texto original	Texto transformado
-¿Y cómo sabes que no lo ama?	-¿* *ó** ***** *** ** ** ***?
-Inglaterra se preguntó a su	-***** ** *****ó * **
vez si habría un muñeco del	*** ** *****í* ** **ñ*** **
esposo también.	***** *****é*.

Experimento 2: Atribuição Autoral

Table 4: 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 5: 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 na PAN2018

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

Table 6: 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 7: 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

Experimento - Características mais importantes

Table 8: Características textuais mais relevantes para *Std.charN*

Candidatos				
01	02	03	04	05
_as_l	_Sti	_sub	_joi	_day,
_'	_" Can	_suc	_gh	_dev
_prec	_" Ca	_l_fi	_er	_dete
_l'd	_" Be	_succ	_glow	_plac
_" Are	_K	_subs	_ls	_mut
_Re	_but_	_l_f	_sta	_must
_smel	_Ofte	_" T	_gor	_Dro
_leak	_posi	_a_t	_sorr	_day_
_is_s	_For	_" St	_eat_	_she_
_spu	_Ri	_a_sw	_lf_t	_chi

Extraído do subconjunto 02 com textos em inglês e com 5 autores.

Experimento - Anexo: Características mais relevantes

Table 9: Características textuais mais relevantes para *Dist.charN*

Candidatos				
01	02	03	04	05
*_**	_*_*	"*'	*_~	*_~*
**	_**_ ("*_**	*_~	'*_*
*'	_**_*	!)*_*	*_~*	"_
).	*!	*!!	'**	*_~
),_	__*'	*'*_*	'*****	*_~
~	*!*	**_*'	"_**'	'*.
*_~	*_**	**_*	_É***	_**
'**	_~	**_*'	_**'	_~
!),	_~*	_**!	_**..	_~

Extraído do subconjunto 02 com textos em inglês e com 5 autores.

Experimento - Características mais relevantes

Table 10: Características textuais mais relevantes em *Std.wordN*

Candidatos				
01	02	03	04	05
about_what	against_his	an_odd	although	and_pulled_him
and_practically	and_it_was	and_then_he	an_eye	and_pulling
any_of	and_so	acknowledged	and_said	across_his
any_more	and_already	and_he_had	and_takes	across_the
and_nearly	and_steve	are_your	and_just	and_all
and_pulled	and_say	again_to	ancient	against_her
agree	accent	and_tell	amount_of	among
all_tony	and_wet	and_forth	always	about_what_to
ah	apparently	are_just	and_grinned	acting
and_wet_and	after	and_grabbing	about_the	about_their

Extraído do subconjunto 02 com textos em inglês e com 5 autores.

Projeto de pesquisa

Projeto de pesquisa

Projeto - 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.

Projeto - Objetivo

Objetivo Geral

O objetivo geral deste trabalho é enriquecer modelos de atribuição autoral de texto digitais com conjunto fechado de autores utilizando conhecimentos dependentes e independentes de idioma combinados com técnicas de aprendizados de máquina, de modo a obter resultados superiores ao estabelecido em trabalhos anteriores.

Projeto - Conjunto de dados e Avaliação

Avaliação das hipóteses

- Comparação com *baselines* pertinentes, como o modelo apresentado em Custódio e Paraboni (2018).
- Serão utilizadas as medidas tradicionais de AM, como *medida F*, acurácia, auROC e outros.
- Espera-se que o resultado médio seja superior ao dos modelos de *baseline*.

Table 11: Córpus para avaliação dos métodos de AA

Córpus	No. Autores	Idioma	Domínio/Gênero
PAN-CLEF2014	-	EN, ES, DU, GR	NV, AR, RV, ES
PAN-CLEF2018	20	EN, ES, FR, IT, PL	NV
RCV1	50	EN	AR
Nus-SMS	116	EN	SMS
b5-post	1.019	PT-Br	Facebook
BlogSet-BR	4.331	PT-Br	AR

Projeto - Escopo e limitações

Este projeto de pesquisa se limita

- ao estudo das técnicas de distorção textual
- ao estudo das técnicas de anotações linguísticas
- ao estudo das técnicas de representação distribuída
- e utilizará métodos de aprendizado de máquina.
- aos idiomas considerados primordialmente são inglês e português brasileiro.

Não serão considerados

- modelos computacionais baseados em grafos, redes complexas e modelos de compressão.

Projeto - Contribuições

Contribuições

Este trabalho pretende avançar a fronteira de conhecimento sobre o problema de AA usando modelos e recursos computacionais e linguísticos dependentes e independentes de idioma.


Ao estudar os recursos linguísticos espera-se avançar o conhecimento da relação entre a linguagem e os fatores que determinam a autoria. Em especial, pretendemos avançar os estudos para o idioma português brasileiro.

Projeto - Atividades

- 1 **Revisão bibliográfica** Concluído.
- 2 **Participação na PAN-CLEF 2018** Concluído.
- 3 **Preparação dos dados** Concluído.
- 4 **Modelos independentes de idioma** Estudo dos tipo de distorção textual, construção dos modelos computacionais e refinamentos específicos.
- 5 **Modelos baseados em anotações** Estudo dos pacotes para anotações POS, como NLTK¹ e Spacy, construção dos modelos computacionais e refinamentos específicos.
- 6 **Modelos baseados em *embeddings*** Preparação de bases de dados de *embeddings*, estudo de *embeddings* específicos para AA, construção de modelos computacionais e refinamentos.
- 7 **Refinamentos**
- 8 **Avaliação**
- 9 **Redação da dissertação**
- 10 **Divulgação**

Table 12: Cronograma

	2018							2019						
Atividades	1-6	7	8	9	10	11	12	1	2	3	4	5	6	
01. Revisão bibliográfica	x	x	x	x	x									
02. Participação PAN-CLEF2018	x													
03. Preparação dos dados	x	x				x								
04. Modelos independentes de idioma						x	x	x						
05. Modelos baseados em anotações		x				x	x	x						
06. Modelos baseados em <i>embeddings</i>		x						x	x					
07. Refinamentos								x	x	x				
08. Avaliação final										x				
09. Redação da dissertação										x	x	x		
10. Divulgação												x	x	


 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.


 BENGIO, Y. et al. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, v. 3, p. 1137–1155, 2003.

Referências II


 BOJANOWSKI, P. et al. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, v. 5, p. 135–146, 2017. ISSN 2307-387X.

 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.


 DEZA, M. M.; DEZA, E. *Encyclopedia of distances*. [S.l.]: Springer, 2009. 1–583 p.

Referências III


 ELMAN, J. L. Finding structure in time. *Cognitive science*, Wiley Online Library, v. 14, n. 2, p. 179–211, 1990.


 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.


 GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

 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.


Referências IV

 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.

 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.


Referências V


 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.


 KIM, Y. Convolutional Neural Networks for Sentence Classification. In: Alessandro Moschitti and Bo Pang and Walter Daelemans (Ed.). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*,. [S.l.: s.n.], 2014. p. 1746–1751.


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

Referências VI


 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 VII


 LE, Q. V.; MIKOLOV, T. Distributed representations of sentences and documents. In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. [S.l.: s.n.], 2014. p. 1188–1196.


 LECUN, Y. et al. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, v. 1, n. 4, p. 541–551, 1989. Disponível em: <https://doi.org/10.1162/neco.1989.1.4.541>.


 MIKOLOV, T. *Statistical language models based on neural networks*. Tese (Doutorado) — Brno University of Technology, 2012. Disponível em: <http://www.fit.vutbr.cz/~imikolov/rnnlm/thesis.pdf>.

 MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.


Referências VIII


 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 IX


 POSADAS-DURÁN, J.-P. et al. Applications of the distributed document representation in the authorship attribution task for small corpora. *Soft Computing*, Springer Verlag, v. 21, n. 3, p. 627–639, feb 2017. ISSN 14327643.


 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.

 RHODES, D. Author Attribution with CNN's. *Stanford University - CS224D Projects*, p. 1–8, 2015.

 ROCHA, A. et al. Authorship Attribution for Social Media Forensics. *IEEE Transactions on Information Forensics and Security*, v. 12, n. 1, p. 5–33, 2017. ISSN 15566013.


Referências X


 SAPKOTA, U. et al. Not all character n-grams are created equal: A study in authorship attribution. In: *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA*. [S.l.: s.n.], 2015. p. 93–102.

 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.

Referências XI


 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.


 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. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, v. 60, n. 3, p. 538–556, 2009. ISSN 15322882.


Referências XII

 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.

 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.

 TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, abs/1003.1141, 2010.

 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.