O Uso de Dicionário de Atributos Estilométricos na Identificação de Autoria de Textos de Língua Portuguesa

Paulo Júnior Varela[1], Edson J. R. Justino[1], Luiz E. S. Oliveira[1]

[1]Pontifícia Univeridade Católica do Paraná (PUCPR),
Rua Imaculada Conceição, 1155, Curitiba, PR, Brasil.

{varela,justino,soares}@ppgia.pucpr.br

Abstract. *Currently, the growing use of digital documents such as e-mails as evidence in judicial proceedings. This practice is usually related to the determination of authorship of the text. The use of computational methods in solving the cases relating to identification of authorship of texts, has been growing. Some methods are based on structural analysis of the document, while others use a linguistic approach. This article aims to provide a method for the identification of authors of text, based on a dictionary of attributes stylometry, with focus on the characteristics of the Portuguese language.*

Resumo. *Na atualidade, vem crescendo o uso de documentos digitais, tais como os e-mails, como evidências em processo judiciais. Esta prática está, usualmente, relacionada como a determinação da autoria do texto. A utilização de métodos computacionais, na solução dos casos associados à identificação da autoria de textos, tem sido crescente. Alguns métodos se baseiam na análise estrutural do documento, enquanto outros utilizam uma abordagem linguística. Este artigo tem como objetivo apresentar um método para a identificação de autoria de texto, com base em um dicionário de atributos estilométricos, com enfoque nas características linguísticas do idioma Português.*

1. Introdução

A linguística forense dedica-se à aplicação da estilística no contexto da identificação da autoria em documentos questionados. A identificação da autoria é realizada através da análise do estilo de escrita do autor. A estilística explora as duas premissas de variabilidade da linguagem. As mesmas estabelecem que, dois escritores de uma língua não escrevem exatamente do mesmo modo e um mesmo escritor, não escreve do mesmo modo todo o tempo [Black *et al* 1990].

A linguística forense divide a análise estilística em duas categorias, a qualitativa e a quantitativa [McMenamim 2002]. A análise qualitativa consiste em avaliar as formas usadas pelo autor, como e porque elas foram utilizadas. Sua grande limitação encontra-se no processo de inferência utilizado pelos peritos. Já o estudo quantitativo avalia a medida da variação na língua escrita. Trata-se de uma ferramenta importante na caracterização da autoria [Johnstone 2000]. A medida quantitativa também apresenta problemas e limitações. Uma dessas limitações reside na escassez de ferramentas de auxílio à análise. Uma abordagem quantitativa exige a mensuração dos atributos estilométricos.

Este artigo apresenta um método para a identificação da autoria de documentos, com base em um dicionário de atributos estilométricos, com enfoque nas características linguísticas do idioma Português. A abordagem adotada é a quantitativa, tendo como objeto o desenvolvimento de métricas através de modelos computacionais.

2. Dicionário de Atributos Estilométricos

A análise estilística forense busca estabelecer um dicionário de atributos estilométricos, como parâmetro estável de análise das variabilidades entre escritores distintos, tais como: a freqüência de palavras incomuns; a média do tamanho das orações; o quociente de palavras diferentes em relação ao total; entre outros. Portanto, é possível afirmar que o conjunto de valores obtidos pela quantização de tais atributos definirá o estilo [McMenamim 2002].

Existem várias classes de atributos estilométricos, entre as quais se destacam a estrutural e a linguística. Na estrutural é possível encontrar atributos, tais como: variações em números e símbolos; variações em abreviações; variações no formato de texto; variações em pontuação; ente outros. Alguns exemplos dessa abordagem são apresentados por Chaski em [Chaski 2005]. Já na lingüística estão presentes atributos intimamente ligados ao idioma utilizado, tais como: conjugações verbais, pronomes, conjunções, entre outros [Pavelec *et al* 2008]. Portanto, um dicionário de atributos estilométricos busca atender aos diferentes pontos de viste de uma análise estilística forense, incorporando em seu contexto, diferentes classes de atributos estilométricos.

3. Base de Dados

A base de dados utilizada neste trabalho consta de colunas de jornais de 30 diferentes autores, sendo 20 textos de cada autor [Pavelec *et al* 2008]. Foram selecionados autores de maneira aleatória que escrevem sobre diferentes temas, incluindo Esportes, Economia, Vinho, Política, Sociedades, Literatura, etc. Os textos da base de dados possuem um tamanho reduzido, variando entre 3KB e 6 KB e com um limite de informações de 735 palavras (tokens).

4. Método Proposto

O método proposto se baseia nos procedimentos de análise estilística forense, Figura 1. A mesma estabelece a associação ou dissociação da autoria do texto, em relação a um provável autor, como base num conjunto de atributos estilométricos previamente estabelecido. A associação indica a existência de atributos estilométricos suficiente para garantir estatisticamente, que o texto, de autoria desconhecida, pertence ao autor avaliado. A dissociação indica que o mesmo não pertence ao autor avaliado.

Durante o processo de análise é utilizado um conjunto n de amostras de texto de autoria conhecida (modelos de referência) K_i (i=1,2,3...,n), em comparação com a amostra de autoria desconhecida (questionada) Q. Durante o processo é observado, tendo como base o dicionário de atributos estilométricos, diferenças de medidas entre as amostras conhecidas e a desconhecida e, posteriormente, é apresentado o resultado parcial da análise. O resultado final depende da soma dos resultados obtidos nas comparações individuais dos pares (referência e questionada), Figura 1.

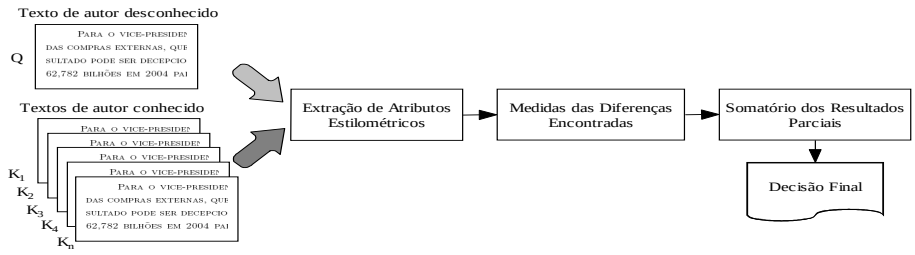

Figura 1. Diagrama esquemático do procedimento de análise estilística forense.

Do ponto de vista computacional, o modelo da Figura 1 é conhecido como modelo global, pois um único modelo é utilizando na associação ou dissociação da autoria do texto questionado. Outro modelo computacional utilizado nesses casos é o modelo por autor ou pessoal, onde deve ser construído um modelo para cada autor. Uma das desvantagens dessa última estratégia consiste na necessidade de um novo treinamento, quando da inclusão de um novo autor.

Nesse trabalho são apresentados resultados usando tanto o modelo global quanto o modelo pessoal. O objetivo é propiciar a análise do desempenho do dicionário, utilizando características linguísticas do idioma Português, sobre os dois pontos de vista, global e pessoal.

O dicionário de atributos estilométricos usado é composto por palavras-funções com 50 verbos no infinitivo, segundo [Ryan 2006] são os mais utilizados em textos no idioma Português e 91 pronomes. A Tabela 1 apresenta o dicionário de atributos estilométricos utilizados nesse trabalho.

Tabela 1. Dicionário de Atributos Estilométricos

Tipo	Palavras
Pronomes Relativos	quem, o qual,a qual, os quais, as quais,onde, em que, quanto, quanta, quantos, quantas, cujo, cuja, cujos, cujas
Pronomes Possessivos	meu, minha, meus, minhas, teu, tua, teus, tuas, seu, sua, seus, suas, nosso, nossa, nossos, vosso, vossa, vossos, vossas
Pronomes Demonstrativos	este, esta, estes, estas, isto, esse, esses, essa, essas, isso, aquele, aquela, aqueles, aquelas, aquilo,nessa, desta, daquela, cujo, cuja,,cujos, cujas
Pronomes Pessoais	eu, tu, ele, nós, vós, eles, me, te, se, lhe, o, a, nos, vos, lhes, os, as, mim, comigo, conosco, ti, contigo, convosco, si, consigo
Pronomes de Tratamento	você, vocês, senhor, senhores, senhora, senhoras, senhorita, senhoritas, vossa senhoria, vossas senhorias
Verbos	escrever, achar, abrir, efetuar, pagar, falar, colar, acabar, atingir, distribuir, jogar, estar, declarar, melhorar, ligar, andar, dizer, completar, achar, usar, ver, dar, visitar, realizar, projetar, ser, escolher, encerrar, haver, desenvolver, cantar, fechar, comer, viver, poder, pular, entender, beber, aplicar, implantar, ler, fazer, pensar, gerar, trazer, ter, trocar, possuir, melhorar, iniciar

Além desses últimos, 171 advérbios e conjunções propostos por Pavelec *et al* em [Pavelec *et al* 2008] foram incorporados ao dicionário e utilizado nos testes. Em outras palavras, o vetor de quantização dos atributos estilométricos é composto de 312 componentes, sendo que cada componente representa a quantidade de cada atributo estilométrico encontrado no texto. Para implementar o modelo global mostrado na Figura 1, os vetores de quantização são usados para calcular os vetores de dissimilaridades, através da Equação 1.

$$V_i = |K_i - Q| \quad (1)$$

4.1. Classificação

No modelo global, o processo de comparação é composto por duas fases, o treinamento e a verificação. No estágio de treinamento, as medidas de dissimilaridades V_i (i=1,2,3,..,n), são calculadas entre pares de textos. Quando dois textos pertencerem a um mesmo autor, o vetor de dissimilaridades é indicado com +1 (associação). Quando dois textos pertencerem a autores diferentes, o mesmo é indicado com -1 (dissociação). O vetor de dissimilaridades tenderá a possuir valores iguais a zero, quando as amostras pertencerem a um mesmo autor. Um SVM (*Support Vector Machine*) [Vapnik 1998] com *kernel* linear é então treinado para separar pequenas dissimilaridades entre atributos estilométricos (associação) e grandes dissimilaridades entre atributos estilométricos (dissociação).

Usualmente, em uma análise estilística forense, se faz uso de um conjunto de amostras de textos de origem conhecida. Cada amostra conhecida pertencente ao conjunto de referência (4 a 10 amostras). As mesmas são comparadas com a amostra de autoria desconhecida ou questionada. Nesse experimento foram utilizadas 5 amostras de referência para cada autor. O método proposto classifica as saídas através de um somatório dos resultados, Figura 1.

Já no modelo por autor, os vetores de dissimilaridades são utilizados para treinar modelos específicos para cada autor. Nesse caso, foram 20 SVMs, um para cada autor da base, treinados usando um protocolo um-contra-todos [Vapnik 1998].

5. Resultados

No modelo global proposto, 10 autores foram separados para gerar o modelo de treinamento, sendo 5 os exemplares de textos de cada autor. No modelo de testes 20 autores são separados e neste caso, 15 exemplares de textos são utilizados para cada um. Outros 5 exemplares, para os mesmos 20, foram reservados para as referências, sendo estes escolhidos de forma aleatória.

Como citado anteriormente, o modelo por autor é baseado no conceito da policotomia, ou seja, a classificação do problema em n-classes. Nesse modelo, cada autor representa uma classe. A quantidade de exemplares utilizada para cada autor foi a mesma do global.

Os testes seguiram os mesmos protocolos propostos por Pavelec *et al* [Pavelec *et al* 2008], a fim de que comparações de resultados pudessem ser feitas. Os resultados

obtidos são apresentados na Tabela 2. Nas duas abordagens utilizadas foi possível observar um ganho de 5% nas taxas de erro, o que demonstra a importância da inclusão dos verbos e dos pronomes, na classe de características linguísticas, no dicionário de atributos estilométricos. O destaque ficou para os verbos, uma vez que os mesmo foram utilizados apenas na forma infinitiva.

Tabela 2. Resultados

Atributos Estilométricos	Modelo Global	Modelo por Autor
Advérbios e Conjunções [5]	72.5%	83.2%
Dicionário Proposto	76.5%	87.0%

5. Conclusão e Trabalhos Futuros

O objetivo principal desse artigo foi apresentar um método computacional para a análise estilística forense, na identificação da autoria de textos. Para esse propósito, foi criado um dicionário de tributos estilométricos, usando a classe de características linguísticas do idioma Português, composta por advérbios, conjunções, verbos no infinitivo e pronomes. Os modelos propostos mostraram-se robustos em textos com menos de 1000 palavras. O destaque ficou para o modelo por autor ou pessoal, onde a especialização do mesmo permitiu um ganho de 10,5% em relação ao modelo global. Ambos os modelos mostraram um ganho significativo com a inclusão de novos atributos, em relação ao trabalho anterior.

Como proposta para trabalhos futuros encontra-se os testes de cada característica linguística, individualmente, para avaliar o impacto da mesma no conjunto do dicionário. Uma segunda proposta seria incorporar a classe de características estruturais ao dicionário, a fim de avaliar as contribuições dessa classe de atributos no conjunto do dicionário.

Referencias

Black, H. C., Nolan, J.R., Nolan-Haley, J.M. (1990) “Black’s Law Dictionary”. West Publishing, 6 edition, St. Paul, p. 1810.

Johnstone, B. (2000) “Qualitative Methods in Socioliguistics”. Oxford University Press, New York, p. 450.

McMenamin, Gerald R. (2002) “Forensic Linguistics - Advances in Forensic Stylistics”. CRC Press, p. 330.

Chaski C. E. (2005) “Who is At The Keyboard. Authorship Attribution in Digital Evidence Investigations”. International Journal of Digital Evidence, 4,1.

Pavelec, D., Oliveira, L. S., Justino, E., Batista, L. V.(2008) “Using Conjunctions and Adverbs for Author Identification”. Journal of Universal Computer Science, 14,18, p. 2967-2981.

Vapnik, V.(1998) “Statistical learning theory”. Wiley, N. Y. p. 768.

Ryan, M. A. (2006) “Conjugação dos Verbos em Português – Prático e Eficiente”. 17ª Ed. Ática, São Paulo, p. 176.