Authorship Attribution in Latin Languages Using Stylometry

P. Varela, M. Albonico, E. Justino and J. Assis

***Abstract*—In this paper we presented a computational approach to authorship attribution in a multilingual environment, based on latin languages. Initially, we defined the databases of literary texts, written by consecrated authors of Portuguese, Spanish and French literature. Subsequently, we established a set formed by groups of stylometric characteristics, which are: morphological, flexors, syntactic and auxiliary. The main objective is to extract from the grammatical structures of the sentences, the stylometric pattern of each author. We perform experiments with author-dependent approach, using verification and identification strategies. In the classification process we use the *Support Vector Machines* – SVM, with a linear kernel.**

***Index Terms*—Authorship Attribution, Latin Languages, Stylometric.**

I. Introdução

NOS últimos anos os meios de comunicação se tornaram mais acessíveis, e com isso problemas relacionados à atribuição da autoria em documentos digitais se tornaram mais constantes, envolvendo diversos casos, principalmente no meio jurídico, tais como: a disputa da autoria de textos [1], [2] [3], detecção de plágios [4], mensagens de ameaça e difamação [5], [6], [7], e casos forenses [8], [9]. Nestes casos, a tarefa principal é descobrir se um texto foi escrito por determinado autor ou saber o autor do texto entre diversos autores. Para tanto, é necessário que amostras de textos de diferentes autores sejam coletadas e armazenadas em uma base de dados. Posteriormente, destes textos são extraídos padrões que fornecem um conjunto de características de estilo para cada autor. Por conseguinte, amostras desta base são confrontadas com a amostra de texto que está sendo questionada. Por fim, o objetivo é saber se a amostra questionada e a amostra de um determinado autor foram escritas ou não pela mesma pessoa.

Neste tipo de problema, onde a análise manual não é possível, pela quantidade de informações a ser processada, o mais usual pela literatura é a aprendizagem de máquina [3], [10], [11], [12]. A aprendizagem de máquina consiste em um método de análise de dados que automatiza o desenvolvimento de modelos analíticos por meio de algoritmos que aprendem por uma série de exemplos, com a finalidade de obter a melhor tomada de decisão [13].

P. J.Varela, Universidade Tecnológica Federal do Paraná, Francisco Beltrão, Paraná, Brazil. (e-mail:paulovarela@utfpr.edu.br).

M. Albonico, Universidade Tecnológica Federal do Paraná, Francisco Beltrão, Paraná, Brazil. (e-mail: michelalbonico@utfpr.edu.br).

E. J. R. Justino, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil. (e-mail:justino@pucpr.br).

J. L. V. Assis, Universidade Tecnológica Federal do Paraná, Francisco Beltrão, Paraná, Brazil. (e-mail:joao1999@alunos.utfpr.edu.br).

Além das técnicas de aprendizagem de máquina é necessário o trato das questões associadas à determinação da autoria. Para isso, utiliza-se a linguística computacional, que usa meios computacionais para manipulação da linguagem humana, permitindo que a atribuição da autoria de um texto possa ser feita. Neste caso, o processo consiste em classificar cada palavra das amostras de textos de acordo com seus níveis gramaticais. Diante disso, aplica-se a estilística, que tem o objetivo de avaliar e identificar as características que tornam um texto, bom ou ruim para determinar o estilo. O estilo de cada autor ou grupo de autores é definido como sendo uma maneira particular de escrever, ou seja, um conjunto de características de uma obra ou de um autor, sendo considerado um elemento variável do comportamento humano por diversos fatores, tais como: grupos sociais e conhecimento técnico. O estilo é definido por um conjunto único de padrões gramaticais que o autor aplica durante o processo de escrita, mais conhecido como estilometria [12]. Com os recursos da estilometria, que visa a aplicação do estilo linguístico aprendido no texto, é possível realizar a parametrização das características de cada autor, e assim, conseguir identificar padrões na escrita dos textos.

Para o desenvolvimento dos experimentos é necessário definir as abordagens de atribuição de autoria. Neste ponto, duas abordagens são as mais usuais na literatura: verificação e a identificação da autoria [3], [10], [14], [15], [16], [17], [18], [22], [23], [24]. Tais abordagens, são executadas por intermédio da observação de atributos linguísticos, tais como os estilísticos, apresentados pelo autor do texto. A verificação da autoria tem por objetivo verificar se o modelo criado no treinamento é robusto o suficiente para conseguir classificar corretamente amostras de textos de um mesmo autor. A estratégia da abordagem é um-contra-um. Então, o autor que se deseja comparar é conhecido, realizando o processo de verificação com os modelos deste autor. O objetivo é verificar se ele acerta ou erra. Na identificação de autoria o objetivo é confrontar toda a base de textos contra todos os autores, em busca de identificar quem é o autor da amostra que está sendo questionada. A estratégia é um-contra-todos, ou seja, confrontar o texto questionado contra todos os autores constantes no treinamento, a fim de tentar identificar o provável autor. Como o confrontamento é grande e de difícil decisão por parte do classificador, utiliza-se a análise dos autores mais bem classificados (*Top-list*).

Neste trabalho, propõe-se uma abordagem computacional para atribuição de autoria de textos em um ambiente multilíngue baseado em línguas de origem latina. Apresenta-se um conjunto de características estilométricas baseadas nas estruturas gramaticais internas das frases, por meio de seus níveis estruturais. A ideia principal é extrair funções de cada palavra, necessárias para a formação de uma frase, tais como:

verbo, adjetivo, advérbio, sujeito e predicado. Estes elementos irão compor um padrão da escrita, delineando um perfil estilístico para cada autor.

Assim sendo, realizaram-se os experimentos utilizando textos literários em domínio público nos idiomas espanhol, francês e português. Utilizou-se a abordagem dependente do autor no treinamento, e nos testes, a verificação e identificação de autoria. Para averiguação do processo foi usado o classificador SVM (*Support Vector Machines*).

Então, este trabalho possui duas contribuições principais: Primeiro, apresentar uma abordagem computacional baseada na estilística que seja discriminante e aplicável em casos que envolvam a atribuição de autoria. E por segundo, avaliar o comportamento do conjunto de características estilométricas em textos literários em idiomas de origem latina (português, espanhol e francês).

Este artigo é estruturado da seguinte forma: A seção 1 apresenta a introdução. A parte 2 descreve os materiais e métodos. A parte 3 apresenta os experimentos e a discussão. Na parte 4, é exposto um comparativo com a literatura. E, por fim, na parte 5 as conclusões e trabalhos futuros.

II. Materiais e Métodos

Nesta seção, apresenta-se as bases de textos literários, o conjunto de características e a abordagem proposta.

A. Bases de Textos

Para realização dos experimentos, foram utilizadas três bases de dados contendo textos de autores consagrados da literatura em línguas derivadas do latim, que são as línguas: portuguesa, espanhola e francesa. Tais línguas foram escolhidas por derivarem da mesma família de origem (Indo-europeia/Itálica/Latina/Românica), entretanto, com diferentes evoluções em seus aspectos gramaticais e sintáticos. Além do que, tais línguas possuem abundantes recursos linguísticos, o que reflete diretamente no contexto dos experimentos, que é avaliar os recursos estilísticos como atributos discriminantes na atribuição de autoria.

Para todos os textos, foi realizado uma limpeza, excluindo textos que não eram dos autores, tais como: número de páginas, cabeçalhos e rodapés. Isso se faz necessário, para que não houvesse interferência no processo de aprendizado do classificador. A base de textos coletada é formada por contos, histórias, romances, fábulas, crônicas, peças teatrais e novelas, excluindo-se poemas e poesias.

As bases de textos de cada língua são compostas por 150 autores distintos e nativos no idioma, com 40 amostras de textos por autor. Cada amostra de texto possui no mínimo 1000 frases. Para avaliar o comportamento da abordagem foi variada a quantidade de informação (baixa, média, alta), ou seja, foi trabalhado com subconjuntos aninhados por quantidade de frases por amostra, sendo: baixa (< 100 frases), média (entre 100 e 500 frases) e alta (> 500 frases). Em todos os casos, as amostras foram divididas 1/2 para treino e 1/2 para testes.

B. Características Estilométricas

Para identificar os padrões de escrita de cada autor, foram utilizadas características linguísticas de estilo, provenientes dos textos. Tais características são divididas em 4 categorias, que representam os níveis estruturais de uma frase: gramaticais, modificadoras, sintáticas e auxiliares (Tabela I) [3].

TABELA I
CLASSES E ATRIBUTOS ESTILÍSTICOS

Classes	Características
Gramaticais	Substantivos, adjetivos, advérbios, verbos, preposições, conjunções e etc.
Modificadoras	Número, gênero, pessoa, tempo, modo, e etc.
Sintáticas	Sujeito, predicado, objeto direto, objeto indireto, verbo principal, verbo auxiliar e etc.
Auxiliares	Artigo, pronome e etc.

O processo de rotulagem indica que, para cada palavra, em diferentes idiomas pode haver vários rótulos (classificações), ou seja, uma palavra pode ter várias funções em uma única frase. Maiores detalhes da rotulagem das palavras podem sem consultados em [3] [20].

C. Abordagem Proposta

No treinamento e testes dos experimentos foi aplicada a abordagem dependente do autor. É baseada em um modelo para cada autor, ou seja, é embasada na policotomia [3], [15], [17], [18]. Nesta abordagem, muitas amostras por autor são necessárias, porque o objetivo principal é enfatizar as características individuais de cada autor.

Todos os autores participam das fases de treinamento e testes. Por esse motivo, o subconjunto de amostras usadas na fase de treinamento não é usado na fase de testes, mas pode ser usado como referência.

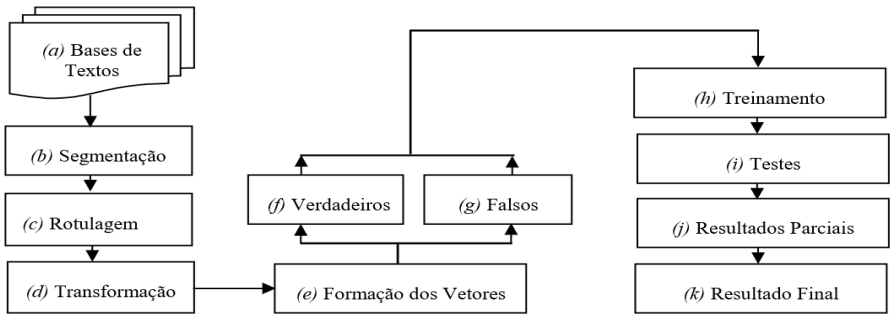

Fig. 1 Visão Geral da Abordagem.

Para fins de estabelecer relação entre as amostras de textos, foi utilizado o conceito de dissimilaridade, onde cada objeto (x) é descrito por suas diferenças em relação a um conjunto de objetos (R) [19].

Na Fig. 1, mostra-se uma visão geral da abordagem. Neste caso, o passo a passo foi adaptado da proposta de [3], onde:

(a) Todas as amostras de textos de autores (A_c), são organizadas. *(b)* Cada amostra de texto é segmentada em frases, e cada frase é processada para extração das características. *(c)* As informações de cada autor são extraídas através do processo definido em [20], que realiza a rotulagem de cada palavra em uma frase, através da identificação das funções que as palavras exercem e dos níveis estruturais da frase. *(d)* O conjunto de atributos é dividido em quatro vetores com diferentes características conforme a Tabela I. *(e)* Para cada texto T contendo F_k frases, os vetores Vt_i são criados com base na Tabela I, onde $i \in R \wedge 1 \leq i \leq 4$ e i é o número de frases. O número de palavras N_k que compõe cada frase F_k é calculado, e o número de palavras marcadas em cada classe é computado. Os vetores de Vt_i, que se referem às quatro classes da Tabela I, contém o número de vezes que cada característica aparece na frase. Os vetores Vt_i são divididos pelo número de palavras N_k na frase, para criar os vetores F_i. *(f)* Um conjunto de amostras genuínas é criado pela combinação aleatória de amostras de um mesmo autor $Z_{(+)}$. Um subconjunto de amostras é usado como referência e para treinamento, e outro subconjunto é usado para testes. *(g)* O subconjunto falso $Z_{(-)}$ é gerado pela combinação de amostras de diferentes autores. Neste caso, um vetor de características de um autor A é combinado aleatoriamente com um vetor de características de um autor B. *(h)* As amostras positivas $Z_{(+)}$ e negativas $Z_{(-)}$ geram um conjunto de treinamento T_s. É iniciado o processo de treinamento, onde são gerados os modelos de cada autor. *(i)* Um conjunto de vetores de testes Q_α, onde $\alpha \in R \wedge 1 \leq \alpha \leq \omega$ e ω é o número de autores, que é integrante de um subconjunto de vetores de características D_t, onde $t \in R \wedge 1 \leq t \leq \xi$ e ξ é o número de amostras de cada autor para os testes. O procedimento básico é calcular o vetor de dissimilaridade entre uma instância de Q_α e um subconjunto de amostras de referências R_p de um autor, escolhidas de forma aleatória. *(j)* Um conjunto de resultados parciais Pr_{ap} é obtido, pela saída de cada um dos classificadores (conforme classes da Tabela I) de forma probabilística. *(k)* É tomada a decisão final pelo somatório dos resultados parciais Pr_{ap}.

III. Resultados e Discussão

Para avaliar os experimentos, a apresentação dos resultados foi dividida em duas partes: verificação e identificação de autoria. Em todos os casos as taxas de acerto são representadas pela acurácia, que mostra o quanto o resultado do experimento é preciso. O processo de validação cruzada foi utilizado para dividir a base de dados, e para a aprendizagem e tomada de decisão foi utilizado o SVM (*Support Vector Machines*) que apresenta resultados promissores relatados pela literatura [3], [10], [17], [18], [22], e com kernel linear que apresentou os melhores resultados diante dos modelos não lineares com kernel polinomial, gaussiano e sigmoidal. Entretanto, é necessário observar que o melhor método depende da base de dados [25].

A. Verificação de Autoria

O objetivo principal de verificação de autoria é determinar se um texto foi escrito por um determinado autor, ou seja, é um problema de duas classes, onde pode-se ter somente duas respostas: autoria ou não-autoria. Então, quando tem-se um vetor de características de um texto questionado Q_α, que pertence *a priori*, a um autor desconhecido (A_d), o objetivo é determinar se o texto questionado pertence a um autor conhecido (A_c) ou não.

Na Tabela II, apresentam-se os resultados da verificação de autoria, ou seja, no processo um-contra-um. Foram analisados os resultados pela quantidade de informações (número de frases) constantes em cada amostra de texto. Neste caso, com baixa quantidade de informação ($10 \leq F_k \leq 100$), para média quantidade de informações linguísticas ($100 \leq F_k \leq 500$), e, para alta quantidade de informações ($500 \leq F_k \leq 1000$). Na Fig. 2, é possível observar a evolução dos resultados para cada língua.

TABELA II
RESULTADOS GERAIS DA VERIFICAÇÃO DE AUTORIA

Língua	**Quantidade de Informação por Amostra de Texto – Resultados em Taxa de Acerto (%)**		
	Baixa	Média	Alta
Espanhola	59-80,3	80-93,9	93-97,8
Francesa	57-76,3	76-90,2	90-96,0
Portuguesa	61-81,0	81-93,6	93-98,1

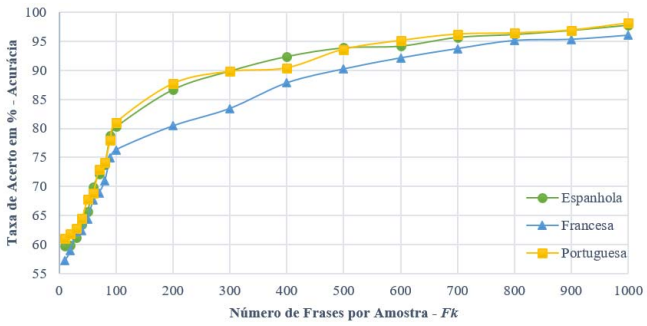

Fig. 2. Evolução dos Resultados – Verificação de Autoria.

Em língua espanhola, percebe-se que com baixa quantidade de informações, a acurácia do modelo variou entre 59-80% de acerto. Isso demonstra, que o modelo com pouca informação sintática é capaz de classificar corretamente cerca de 59% dos textos questionados quando testados com 10 frases aninhadas, cerca de ≈150 palavras. E, com cerca de 100 frases por amostra, o que equivale a um texto com média 1500 palavras, o modelo atinge em torno de 80% de acurácia. Para amostras de textos com média quantidade de informações linguísticas, percebe-se uma evolução considerável nos resultados, alcançando entre 80-93% nas taxas de acerto. Quando testadas as amostras com mais de 500 frases (alta quantidade de informações), verifica-se que os resultados foram entre 93-98%. Isso indica, que quanto mais informações linguísticas se obtém sobre o texto

questionado, e consequentemente, sobre os autores, mais eficaz será a aplicação da abordagem. Entretanto, percebe-se que até mesmo com baixa quantidade de texto, a abordagem se mostra viável e aplicável na verificação de autoria de textos escritos em língua espanhola.

Em língua francesa, verifica-se que as taxas de acerto foram levemente menores que nas outras línguas. Entretanto, evidencia-se que existe uma regularidade na evolução dos resultados, conforme se aumenta a quantidade de frases por amostra, como pode ser visto na Fig. 2. Para textos com baixa quantidade de informações, as taxas de acerto variaram entre 57-76%, preconizando uma similaridade dos resultados com a literatura e com outros idiomas testados. Quando foi testado com o número de frases variando entre ($100 \leq F_k \leq 500$), as taxas de acerto ficaram entre 76-90%. E, para amostras de textos com mais de 500 frases, a acurácia foi de 90-96%. Esses resultados indicam que as características estilométricas propostas são robustas e discriminantes, sendo passíveis de uso em casos de atribuição de autoria em língua francesa.

Para os textos de língua portuguesa, resultados promissores e aceitáveis foram atingidos. Em textos com baixa quantidade de informação, as taxas de acerto variam entre 61-81%. Com textos com média quantidade de informações a acurácia fica entre 81-93%. E, para textos com grande quantidade de informações linguísticas, as taxas de acerto oscilam entre 93-98%. Com estes resultados, pode-se dizer que a capacidade de discernimento das características estilométricas é realmente considerável no processo de atribuição de autoria de textos de língua portuguesa.

Por conseguinte, foram observadas algumas semelhanças nas taxas de acerto dos idiomas português, espanhol e francês. Esta semelhança, podem estar diretamente ligadas às estruturas sintáticas, pois as línguas portuguesa, espanhola e francesa possuem uma gramática e sintaxe muito próximas, condizendo que existe uma constância e robustez do modelo, já que existe uma baixa variabilidade dos resultados em idiomas distintos. Um outro ponto a ser levado em consideração, é que conforme incrementa-se o número de frases em cada amostra de texto (F_k) do autor, os resultados vão evoluindo consideravelmente. Isso significa, que quanto mais informações linguísticas tem-se de cada autor, melhor será a performance do modelo. Ainda que, com a limitação de frases de um texto com baixa quantidade de palavras, os resultados apresentados na verificação de autoria são condizentes com a literatura.

B. Identificação de Autoria

A identificação de autoria consiste no processo de identificar o autor desconhecido de um documento. Neste caso, identificar um autor A_c de um texto T_d, onde $c \in R \wedge 1 \leq c \leq \delta$, onde δ é o número de autores da base de textos literários. Para tanto, é maximizada a relação $F_d = max\ \{D_i\ (x,\ R_c)\}$, e com isso, se obtém o retorno estimado da probabilidade de acerto *a posteriori*, onde D_i representa o modelo treinado. Isso, indica se um texto T_d e as referências R_c pertencem ou não à um mesmo autor.

Adicionalmente a identificação de autoria fornece uma lista de amostras de textos que são mais semelhantes à amostra questionada, ou seja, uma *Top-list*. Esta lista possui a função de fornecer maiores subsídios para tomada de decisão em ambientes complexos. Uma amostra será considerada correta se pelo menos uma ocorrência for listada entre as listas *Top-1*, *Top-5* e *Top-10*, por exemplo. Apesar que o resultado almejado seja próximo de 100%, ou seja, estar no *Top-1*, muitas vezes pode se tomar decisão com base em listas de *Top-5* e *Top-10*.

Na Tabela III, são apresentados os resultados da identificação de autoria para os idiomas. Inicialmente, foi analisado o *Top-List*, para baixa (B), média (M) e alta (A) quantidade de informações linguísticas em cada amostra, considerando as melhores taxas de acerto.

Em língua espanhola observa-se que o classificador retornou o autor correto no topo da lista (*Top-1*) em 72,3% dos casos para amostras de textos com até 100 frases. Quando aumenta-se a quantidade de texto para até 500 frases por amostra, a taxa de acerto ampliou cerca de 10%, chegando a 82,9%. Nos testes com mais de 500 frases por amostra, o melhor resultado no *Top-1* foi de 90,1%. Isso demonstra um resultado promissor, já que o classificador tem que escolher um texto entre 150 possíveis. Quando se amplia a análise da *Top-list*, consequentemente aumenta-se o resultado, pois o autor do texto questionado pode não ter sido elencado como primeiro da lista, entretanto, estar listado na sequência. Neste caso, em uma análise do *Top-5*, percebe-se um incremento nas taxas de acerto de 4,6%, 3,8% e 2,1%, respectivamente, para textos com amostras de baixa, média e alta quantidade de informação, em comparação com o *Top-1*. E, quando se expande para o Top-10, verifica-se que textos com até 100 frases por amostra, atingem 80,4% de acerto; com média quantidade, a acurácia chega a 90,9%; e, com grande quantidade os resultados chegam à 95,7%.

Para os experimentos em língua francesa, no *Top-1* atingiu-se 70,4%, 79,6% e 86,4% de acurácia em textos com baixa, média e alta quantidade de informação, respectivamente. No *Top-5*, tais resultados tiveram incrementos de 3,3%, 4,8% e 3,5%, nesta ordem para as amostras de textos relatadas anteriormente, e vistas na Tabela III. Na análise *Top-10*, em língua francesa, a abordagem retornou precisão de 79,9% para amostras de textos com até 100 frases, 89,1% para até 500 frases, e 93,1% para mais de 500 frases. É possível constatar, que apesar de ter resultados inferiores à língua espanhola, existe uma uniformidade nas taxas de acerto nos experimentos com a língua francesa, preconizando que as características estilométricas e a abordagem proposta podem ser aplicadas em casos que envolvam a identificação de autoria.

Na língua portuguesa a abordagem demonstrou os melhores resultados gerais entre as línguas testadas. Com isso, atribui-se esta leve superioridade ao fato da língua portuguesa possuir em sua gramática, uma maior quantidade de elementos sintáticos, na comparação com as línguas espanhola e francesa. Em todo caso, quando analisado o *Top-1*, percebe-se que a acurácia foi de 73,9%, 81,7% e 90,3% quando testadas amostras com pequena, média e grande quantidade de informações linguísticas. Para o *Top-5*, estes resultados tiveram um aumento de 3%, 7, 9% e 4%, respectivamente. E, quando realizada uma análise observando os 10 textos mais bem classificados, verificam-se taxas de precisão de 81,2% para textos com até 100 frases

TABELA III
MELHORES RESULTADOS DA ATRIBUIÇÃO DE AUTORIA POR QUANTIDADE DE INFORMAÇÃO E TOP-LIST

Língua	Top-1			Top-5			Top-10		
	B	M	A	B	M	A	B	M	A
Espanhola	72,3	82,9	90,1	76,9	86,7	92,2	80,4	90,9	95,7
Francesa	70,4	79,6	86,4	73,7	84,4	89,9	79,9	89,1	93,1
Portuguesa	73,9	81,7	90,3	76,9	89,6	94,3	81,2	92,3	96,2

por amostra, 92,3% para até 500 frases, e 96,2% para até 1000 frases. Denota-se então, que a abordagem se mostra robusta e estável, pois apresenta um comportamento regular e contínuo nas taxas de acerto.

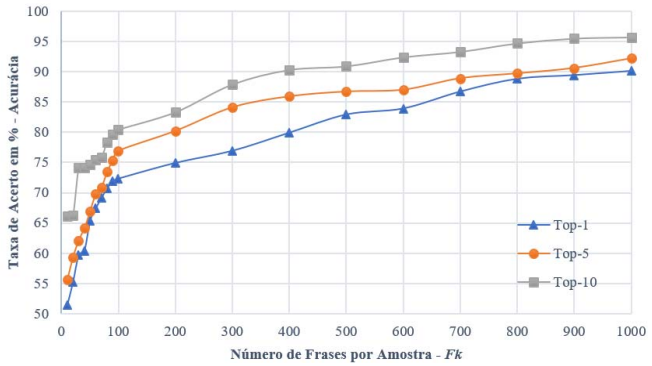

Fig. 3. Resultados Língua Espanhola – Identificação de Autoria.

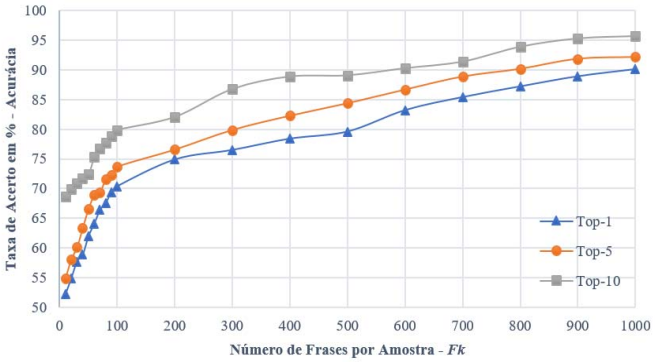

Fig. 4. Resultados Língua Francesa – Identificação de Autoria.

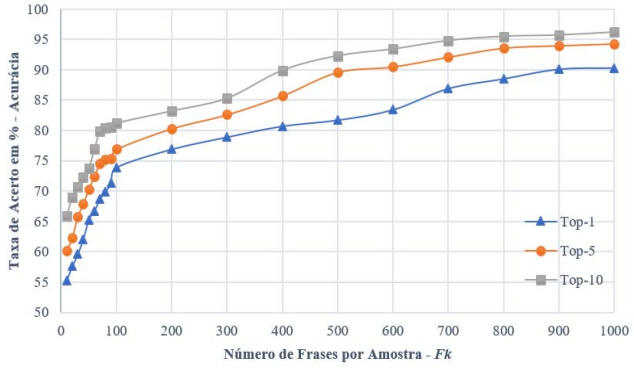

Fig. 5. Resultados Língua Portuguesa – Identificação de Autoria.

Nas Fig. 3, 4 e 5 são mostradas as evoluções em termos de acurácia para os experimentos nos três idiomas, baseado no *Top-1*, *Top-5* e *Top-10*. Nestes casos, fica evidenciado em cada gráfico, as curvas de precisão do modelo, conforme a variação da quantidade de frases por amostra, variando de no mínimo 10 e no máximo 1000 frases.

Com os experimentos realizados neste trabalho, percebe-se que a abordagem proposta se mostra robusta e eficiente em textos literários das línguas espanhola, francesa e portuguesa. Quando foi aplicado o método em amostras de textos com no máximo 100 frases, as taxas de acerto variaram entre 70-74% no *Top-1*. Isso representa um bom indicador e coerente com a literatura [15][18][21], pois foi trabalhado com a identificação de autoria em um ambiente complexo. Para uma quantidade média de frases por amostra (entre 100 e 500), os resultados para o *Top-1* varia entre 79-83%, indicando que conforme se aumenta a quantidade de informação linguística das amostras de cada autor, o classificador consegue tomar melhores decisões. E, para textos com grande quantidade de informação, as taxas de acerto variaram entre 86-90% para o *Top-1*. Em se tratando de identificação de autoria, verifica-se um comportamento promissor das características estilométricas utilizadas, tanto em textos pequenos como em textos grandes. Por conseguinte, observa-se que a diferença média entre as taxas de acerto dos idiomas, não ultrapassa os 4%, indicando a consistência [2] [3] [6] [12] da abordagem proposta.

IV. Comparativo com a Literatura

Na Tabela IV, é efetuada uma comparação da nossa abordagem com alguns dos principais trabalhos em atribuição de autoria de textos [3] [15] [21] [22]. A comparação não é exata, pois os protocolos ou bases de dados não são as mesmas. Entretanto, é possível estimar as contribuições efetuadas pelo método apresentado neste trabalho.

Em língua espanhola a abordagem proposta se mostrou mais eficiente cerca de 25% na verificação de autoria, em comparação com o trabalho desenvolvido por [21] que utilizou textos de diversos tamanhos e variados atributos de estilo. No comparativo com [3], que fez uso da mesma base de textos, percebe-se que a abordagem proposta se sobressaiu em 2,8% na verificação de autoria e 9,7% na identificação de autoria. Entretanto, em ambas as comparações as características estilométricas são distintas.

Para língua francesa, a abordagem proposta obteve resultados semelhantes ao trabalho apresentado por [15], que usou textos de romances e lemas como atributos de estilo. Em

relação ao trabalho de [3], a abordagem proposta foi 2% menor na verificação de autoria e, 2.1% maior na identificação de autoria. Com isso, pode-se dizer que a abordagem proposta, é aplicável em língua francesa e condizente com a literatura.

TABELA IV
COMPARAÇÃO COM A LITERATURA

Autor	Idioma	Verificação	Identificação
Pavelec [22]	Português	-	75-83%
Halvani [21]	Espanhol	72%	-
Savoy [15]	Francês	-	70-100%
Varela [3]	Espanhol	95%	86%
	Francês	98%	91%
	Português	98%	93%
Abordagem Proposta	**Espanhol**	**97,8%**	**95,7%**
	Francês	**96%**	**93,1%**
	Português	**98,1%**	**96,2%**

Para textos escritos em língua portuguesa, a abordagem se mostra praticamente igual aos resultados apresentados por [3] na verificação de autoria, e 3,2% maior na identificação de autoria. Contudo, a abordagem proposta neste trabalho atua com uma maior quantidade de informação textual do que em [3]. Quando comparado aos resultados apresentados por [22], que utilizou palavras-função como atributos discriminantes, percebe-se que a abordagem proposta é cerca de 13% maior.

Com estas informações, constata-se que a abordagem proposta pode ser aplicada em casos de atribuição de autoria (verificação e identificação), pois ficou denotada que as características estilométricas em conjunto com a abordagem dependente do autor atingem taxas de precisão acima de 90%. Em correlato, consegue-se observar que a abordagem é robusta e aplicável em idiomas de origem latina (português, espanhol e francês).

IV. Conclusão

Este trabalho teve por finalidade apresentar uma abordagem computacional para atribuição de autoria em textos de língua portuguesa, espanhola e francesa. Trabalhou-se com a abordagem dependente do autor, e com as estratégias de verificação e identificação de autoria. Como elementos discriminadores foram usadas características estilométricas, pertencentes à quatro classes linguísticas. Tais atributos, alimentaram os vetores de dissimilaridade, que posteriormente foram utilizados para gerar os modelos de treinamento e testes através do classificador SVM. Em língua espanhola, foram alcançadas taxas de acerto máxima de 97,8% para verificação de autoria, e de 95,7% para identificação de autoria. Para língua francesa a acurácia foi de 96% em seu auge para a verificação de autoria e de 93,1% para a identificação de autoria. E, em língua portuguesa, o modelo alcançou em seu ápice, taxa de acerto de 98,1% para verificação de autoria, e de 96,2% para identificação de autoria.

Enfim, constatou-se que o uso de características estilométricas baseadas em atributos morfológicos e sintáticos para o reconhecimento de padrões de escrita em língua espanhola, francesa e portuguesa é viável, pois aponta resultados promissores. Percebeu-se que quanto mais informações textuais obtiver dos autores, melhores serão as decisões tomadas pelo classificador. Como um todo, a abordagem se mostrou estável e robusta perante os experimentos. Pretende-se ampliar e avaliar a abordagem com outros tipos de textos, e inserir atributos semânticos, como trabalhos futuros. Por conseguinte, testar a abordagem em outras línguas de origem latina, tais como: italiana e catalã.

Referências

[1] D. I. Holmes, "The evolution of stylometry in humanities scholarship", In: *Literary and Linguistic Computing*, vol. 13, nº 3, p.111-117, 1998.

[2] A. Neme, J.R.G. Pulido, A. Munoz, S. Hernandez, and T. Dey, "Stylistics analysis and authorship attribution algorithms based on self-organizing maps", In: *Neurocomputing*, vol. 147, p. 147-159, 2015.

[3] P. J. Varela, E. J. R. Justino, F. Bortolozzi and M. Albonico, "A Computational Approach for Authorship Attribution on Multiple Languages", In: 2018 *International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, July 2018.

[4] S. Burrows, A. L. Uitdenboger, and A. Turpin, "Comparing techniques for authorship attribution of source code", *Software: Practice and Experience*, vol. 44 (1), p.1-32, 2014.

[5] O. De Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics", *ACM SIGMOD*, vol. 30, 4, p. 55–64, 2001.

[6] A. Abbasi and H. Chen, "Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace", *ACM Trans. Inf. Syst*., 26(2):7:1–7:29, 2008.

[7] R. Zheng, J. Li, H. Chen, and Z. Huang, "A Framework for Authorship Identification of Online Messages: Writing-style Features and Classification Techniques", *J. Am. Soc. Inf. Sci. Technol*., 57(3):378– 393, 2006.

[8] G. R. McMenamin, "Forensic Linguistics – Advances in Forensic Stylistics", *CRC Press*, New York, 2002.

[9] C. E. Chaski, "Who's at the keyboard? - authorship attribution in digital evidence investigations", In: *International Journal of Digital Evidence*, vol. 4(1), Spring, 2005.

[10] S. Argamon, M. Koppel, J. Fine and A. Shimoni, "Gender, genre, and writing style in formal written texts", Text, 23(3), 321–346, 2003.

[11] P. Juola, "Authorship attribution for electronic documents", In M. Olivier & S. Shenoi (Eds.), *Advances in digital forensics II* (pp. 119–130). Boston: Springer, 2006.

[12] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods", *J. Am. Soc. Inf. Sci. Technol*., 60(3):538–556, March 2009.

[13] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, C.C, "Machine Learning", *Neural and Statistical Classification*, 1994.

[14] A. Finn and N. Kushmerick, "Learning to classify documents according to genre", *Journal of the American Society for Information Science and Technology*, 57(11):1506–1518, September 2006.

[15] J. Savoy, "Who Wrote this Novel? Authorship Attribution across Three Languages", Semantic Scholar, 2011.

[16] J. Diederich, J. R. G. Kindermann, E. Leopold, and G. Paass, "Authorship Attribution with Support Vector Machines", *Applied Intelligence*, 19(1-2):109–123, May 2003.

[17] P. J. Varela, E. Justino, and L. S. Oliveira, "Selecting syntactic attributes for authorship attribution", The 2011 International Joint Conference on Neural Networks, pages 167–172, July 2011.

[18] P. J. Varela, E. J. R. Justino, F. Bortolozzi, L. E. S. Oliveira, "A computational approach based on syntactic levels of language in authorship attribution", *IEEE Latin America Transactions* 14 (1), 259-266, 2016.

[19] E. Pekalska and R. P. W. Duin, "Dissimilarity Representations Allow for Building Good Classifiers", *Pattern Recognition. Letters*, 23(8):943–956, June 2002.

[20] E. Bick, "Visual interactive syntax learning", Available at http://beta.visl.sdu.dk/.

[21] O. Halvani, C. Winter, and A. Pflug, "Authorship Verification for Different Languages, Genres and Topics", *Digital Investigation*, 16(S):S33–S43, March 2016.

[22] D. Pavelec, E. J. R. Justino, L. E. S. Oliveira, "Compression and stylometry for author identification", In: 2009 *International Joint Conference on Neural Networks*, IJCNN. p. 2445-2450, 2009.

[23] H. Ramnial, S. Panchoo, S. Pudaruth, "Authorship attribution using stylometry and machine learning techniques" In: *Intelligent Systems Technologies and Applications*, Springer, Cham, p. 113-125, 2016.

[24] I. Markov, J. Baptista, O. Pichardo-Lagunas, "Authorship attribution in portuguese using character n-grams", *Acta Polytechnica Hungarica*, 14.3: 59-78, 2017.

[25] R.M. Silva, T.A. Almeida, A. Yamaki, "MDLText aplicado na Filtragem Automática de SPIM e SMS Spam*", iSys - Revista Brasileira de Sistemas de Informação*, v. 11, n. 1, p. 103-132, may 2018.

Paulo Júnior Varela é Doutor em Informática pela Pontifícia Universidade Católica do Paraná - PUCPR. Professor Adjunto da Universidade Tecnológica Federal do Paraná (UTFPR) campus de Francisco Beltrão, atuando nas áreas de: reconhecimento de padrões, aprendizagem de máquina e linguística computacional. Atualmente é Chefe do Departamento de Apoio a Projetos Tecnológicos - DEPET, prospectando em Propriedade Intelectual, Inovação e Empreendedorismo. É assessor e diretor de Relações Empresariais e Comunitárias substituto (DIREC).

Michel Albonico é doutor pela Escola de Minas, Nantes, França. Fez parte do grupo de pesquisa AtlanMod desde Abril de 2014 até Agosto de 2017. Professor Adjunto da Universidade Tecnológica Federal do Paraná (UTFPR), campus de Francisco Beltrão, Paraná. Mestre em Informática pela Universidade Federal do Paraná (UFPR), tendo feito parte do grupo de pesquisa de Teste de Sistemas de Larga Escala. Pós-Graduado em Administração de Sistemas de Informação pela Universidade Federal de Lavras (UFLA), Lavras, Minas Gerais. Possui graduação em Sistemas de Informação pela Universidade do Oeste de Santa Catarina (UNOESC). Tem experiência na área de Sistemas de Informação, com ênfase em redes de computadores, sistemas distribuídos, banco de dados, desenvolvimento web e software livre. Atuou como coordenador e docente na Faculdade da Fronteira (FAF) e União de Ensino do Sudoeste do Paraná (Unisep), e como professor na Faculdade Iguaçu e Universidade do Oeste de Santa Catarina (UNOESC).

Edson José Rodrigues Justino possui graduação em Engenharia Industrial Elétrica pela Universidade Tecnológica Federal do Paraná (1985), mestrado em Engenharia Elétrica e Informática Industrial pela Universidade Tecnológica Federal do Paraná (1991) e doutorado em Informática Aplicada pela Pontifícia Universidade Católica do Paraná (2001). Atualmente é professor titular da Pontifícia Universidade Católica do Paraná. Possui experiência na área de Ciência da Computação, com ênfase em Processamento Digital de Imagens e Reconhecimento de Padrões. É líder do grupo de Pesquisa em Visão, Imagens e Robótica do Programa de Pós-Graduação em Informática da PUCPR. É coordenador do Centro de Pesquisa e Inovação em Dispositivos e Imagens Médicas da PUCPR. Coordenou vários projetos de pesquisa apoiados pelo CNPq entre os quais ADQ-AC Análise de Documentos Questionados Auxiliada por Computador, Grafoscopia e Análise de Estilo Literários em Documentos Questionados, Restauração de Documentos Auxiliada por Computador. Peer Reviewer of: Pattern Recognition, IEEE Transactions on Pattern Analysis and Machine, Intelligence, Journal of Forensic Document Examination, Neurocomputing (Amsterdam).

João Lucas Varela de Assis é acadêmico do Curso de Licenciatura em Informática da Universidade Tecnológica Federal do Paraná – UTFPR. É bolsista de Iniciação Científica e atua nas áreas de programação, reconhecimento de padrões e aprendizagem de máquina.