A Computational Approach Based on Syntactic Levels of Language in Authorship Attribution

P. J. Varela, E. J. R. Justino, F. Bortolozzi and L. E. S. Oliveira

***Abstract*— This paper aims to insert a new approach based on syntactic features of the language, which are relating to the essential terms, integrant and accessories of a sentence, such as: subject, predicate and accessories, for the resolution of cases involving the authorship attribution. To this, a database in Portuguese was collected for the experiments. The proposed approach consists of conducting experiments with various fusion methods (sum, mean, median and majority vote), to verify the best performance and method for both authorship verification and identification. To evaluate the approach we used the dependent and independent models. The results generated in authorship verification were between 89-95% accuracy, and 81-90% in the authorship identification through classifiers based on SVM − Support Vector Machines.**

***Keywords*— Authorship Attribution, Syntactic Features, Fusion Methods.**

I. INTRODUÇÃO

A TAREFA de identificação de características de estilo que sejam discriminantes em um documento de texto ou de um determinado autor sempre foi uma das áreas de interesse de pesquisa, quando se fala em atribuição de autoria. Saber se um documento foi escrito por determinado autor ou saber de quem é a autoria do texto são questionamentos interessantes e que despertam o interesse de pesquisadores e do poder judiciário. Para tanto, amostras de textos de diversos autores suspeitos devem ser coletadas e armazenadas em uma base de dados, de onde são extraídas características de estilo de cada autor. Amostras desta base são confrontadas com a amostra de texto que está sendo questionada, para ao final, saber se a amostra questionada e a amostra de um determinado suspeito foram escritas pelo mesmo autor. Como por exemplo, pode-se citar os casos dos textos questionados do *Federalist Papers* [1][2][3], do *Quintus Curtius Snodgrass* [4], das epístolas paulíneas [5], da identificação de crimes digitais [6], e na identificação de mensagens terroristas através da web [7]**Erro! Fonte de referência não encontrada.**.

Sobre o ponto de vista forense, a linguística se coloca como uma ferramenta de suma importância, no trato das questões associadas à determinação da autoria. Quer seja na língua falada ou na escrita, a linguística tem se mostrado um recurso cada vez mais necessário. No contexto da escrita, o uso intensivo dos meios digitais pela sociedade, acabou por gerar igualmente uma grande demanda no poder judiciário. Processo de toda ordem, onde a presença de textos eletrônicos são peças importantes do mesmo e como tal, ajudarão a formar a convicção do juiz. Assim, os mesmos encontram na linguística forense um recurso decisivo para a determinação da autoria.

Diferentemente da grafoscopia, a linguística forense permite que a identificação da autoria de qualquer documento possa ser feita independentemente da base de registro utilizada (papel ou formato digital). Permite também que a verificação ou a identificação da autoria de um documento questionado seja executada através da observação de atributos linguísticos, tais como os estilísticos, apresentados pelo escritor do documento. No entanto, por se tratar ainda de um campo que apresenta divergências quanto à eficácia metodológica dos métodos qualitativos e dos quantitativos, carece de aprofundamento científico [8] [9].

Independentemente das divergências, novas abordagens computacionais para a verificação e identificação da autoria de textos vêm sendo propostas pela comunidade científica de língua portuguesa, tais como: métodos de classificação baseados em compressão de dados [10] e os que utilizam os métodos quantitativos da linguística [11] [12].

Quanto as características textuais utilizadas nas abordagens, algumas classes gramaticais têm obtido resultados promissores, tais como: léxicas [13][14]; sintáticas [3][8][15][16][17][18][19]; semânticas, [15][17]; e, de conteúdo específico [7][20][21]. No entanto, as características que demonstram os melhores resultados, são as características com informações sintáticas. Todas as abordagens visam agregar subsídios ao processo de análise e identificação da autoria de documentos. A aplicação forense nesses casos passa a ser um processo natural, decorrente do uso dos métodos em ferramentas para esse fim. Entretanto, ainda existem lacunas em aberto e possibilidades de aplicação de abordagens que ainda não foram devidamente testadas e estudadas. Neste caso, por si só a atribuição de autoria em documentos digitais é um fator motivacional.

Sendo assim, a contribuição deste artigo é apresentar uma abordagem que faz uso de características sintáticas baseada em níveis estruturais da gramática da língua portuguesa que possam ser utilizados em casos que envolvam a atribuição de autoria. Utilizamos os níveis estruturais de termos essenciais (sujeito e predicado), termos integrantes (verbo, objeto direto, objeto indireto e agentes de voz ativa e passiva), e termos acessórios (adjunto adnominal e adjunto adverbial) em sequencias longas, tais como, frases (independente de verbo) e orações (dependente de verbo) [9]. Neste caso, o principal objetivo é extrair informações que os autores tendem a usar

P. J. Varela, Universidade Tecnológica Federal do Paraná (UTFPR), Francisco Beltrão, Paraná, Brasil, paulovarela@utfpr.edu.br

E. J. R. Justino, Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, Paraná, Brasil, edsonjustino@ppgia.pucpr.br

E. J. Pacheco, Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, Paraná, Brasil, edsonpacheco@ppgia.pucpr.br

L. E. S. Oliveira, Universidade Federal do Paraná (UFPR), Curitiba, Paraná, Brasil, les.oliveira@ufpr.br

inconscientemente, e que possuam padrões sintáticos semelhantes.

Em correlato, objetivamos auxiliar peritos, linguistas e o poder juciário em casos que envolvam: direito autoral, mensagens de ameaças, racismo, injúria e difamação, bem como, a categorização de textos. Adicionalmente, listamos alguns dos fatores motivacionais deste artigo, que consistem em: evidenciar um grupo de características sintáticas que sejam discriminantes para atribuição de autoria, bem como identificar a melhor estratégia de fusão de classificadores a ser aplicado em casos que envolvam um documento cuja autoria seja questionada.

Para aplicação da abordagem foram utilizadas 400 características pertencentes à 8 classes sintáticas, sendo utilizadas as 50 representações mais frequentes de cada classe. Por conseguinte, foram reproduzidos alguns experimentos utilizando as abordagens propostas por [22][23][24], sob a mesma base de dados, para fins comparativos. Ao final, percebeu-se que a abordagem proposta se sobressaiu nos experimentos realizados, demonstrando que as características utilizadas podem ser discriminantes em casos que envolvam a verificação ou a identificação de autoria em documentos questionados de língua portuguesa.

Este artigo está estruturado em seções, onde: a seção 2 apresenta uma revisão da literatura e a importância do estilo e da estilística na atribuição de autoria. A seção 3 descreve os métodos utilizados para o desenvolvimento dos experimentos, tais como: base de dados, o processo de extração de características, a geração dos modelos e o processo de decisório. Os protocolos de experimentos são apresentados na seção 4. A seção 5 apresenta os resultados obtidos e, por fim, na seção 6, que consta das considerações finais.

II. ESTADO DA ARTE

Diversos pesquisadores já desenvolveram abordagens para atribuição de autoria de textos ao longo dos anos. Dentre esses trabalhos, diversas foram as línguas em que os textos foram analisados, entre elas: albanesa, alemã, árabe, chinesa, espanhola, francesa, grega, indiana, italiana, inglesa, japonesa, holandesa, russa, persa e portuguesa. Percebe-se também, que em certas línguas os trabalhos ainda são ínfimos em relação a outras línguas, tal como, a língua portuguesa em relação a língua inglesa.

Uma série de contribuições que envolvam a atribuição de autoria já foram desenvolvidas, principalmente as baseadas na ocorrência de frequência de palavras. No entanto, algumas palavras provaram ser mais discriminatórias em relação a outras, a estas palavras dá-se o nome de palavras-função [2]. Isso gerou a possibilidade do desenvolvimento de novas abordagens, focadas principalmente na análise sintática do texto, ou seja, nas regras de escritas geradas pelo autor. Tais regras são dadas pela análise da árvore sintática de cada frase, que é gerada através da análise parcial ou aprofundada de um texto (também conhecido como POS *tagger* ou *Part-of-speech*). Neste tipo de análise são observadas as rotulagens categóricas de cada palavra, ou seja, da função estrutural que cada palavra exerce na frase. Pode-se citar o exemplo da categorização em uma frase, identificando o sujeito, o predicado e seu complemento em um nível mais raso. Porém, pode-se aprofundar o nível e identificar as classes gramaticais de uma língua, tal como: verbos, advérbios, pronomes e suas subclassificações.

Um dos primeiros relatos científicos que evidenciou fazer uso de análise sintática foi desenvolvido por [27], onde um corpus sintaticamente anotado foi utilizado para investigar o potencial discriminatório das regras sintáticas em atribuição de autoria. Já em [15] o objetivo foi de categorizar um texto de acordo com seu estilo, principalmente para: verificar se um texto foi retirado de uma revista ou de um jornal; se é um texto editorial ou uma notícia; e até mesmo para identificar se foi escrito por um nativo da língua inglesa ou não, além é claro, de fazer uso em atribuição de autoria. Utilizaram como características discriminantes a frequência de um conjunto de palavras-função e a análise da estrutura sintática dos textos através de POS *tagger*. Utilizaram técnicas de aprendizagem de máquina para classificar os documentos.

Apresentando uma abordagem para a categorização de textos através de termos sintáticos para identificação de autoria e de gênero em língua grega, [14] utilizou uma ferramenta de processamento de linguagem natural – PLN - para rotular as palavras e gerar marcadores de estilo. Fizeram uso de 3 níveis para realização de seus experimentos: nível de token (análise de frase individual); análise sintática (análise do conjunto de frases); e análise de níveis dada pela ferramenta de PLN.

O uso de métodos de mineração de textos para atribuição de autoria foi utilizado inicialmente por [16], através do uso do SVM. Segundo os pesquisadores a atribuição de autoria deve ser considerada como um problema de categorização, e em contraste com outras tarefas de classificação, ainda não são claras quais as características de um texto devem ser utilizadas para identificar um autor.

Em [17], uma análise linguística profunda para fins de atribuição de autoria e análise do estilo literário foi proposta. Para testar a abordagem fez uso dos textos literários. Usou diferentes características testadas em separado, que são: frequência de palavras-função, frequência de POS *trigrams*, características sintáticas através de rotulagem das palavras e dependências semânticas das frases.

Com o intuito de conseguir identificar livros e realizar a atribuição de autoria através de características estilométricas, [28] analisaram as distribuições das estruturas sintáticas em comparação com palavras-função e outras características léxicas e estruturais. Para tanto, fizeram uso de classes sintáticas e relações semânticas dos verbos e suas estruturas, a complexidade linguística em termos de profundidade da frase, análise da sentença inicial e final da frase.

A atribuição da autoria também é uma das principais técnicas em investigações em crimes digitais, e sendo assim, [6] relatou em seu trabalho um método computacional que faz uso de recursos estilométricos baseados em características sintáticas. Já em [22] fizeram uso de aprendizado de máquina para realizar a classificação automática de gêneros textuais em língua inglesa. No caso de [23], apresentaram um método para

atribuição de autoria baseado na frequência de *bigrams* de rótulos sintáticos através da análise parcial do texto. Em [24] resolveram explorar a autoria de obras da literatura inglesa clássica extraídos do projeto Gutenberg, utilizando palavras-função, marcadores de POS *tagger* e POS *pair* em separado e em conjunto.

III. A ABORDAGEM APLICADA

A visão geral da abordagem e dos processos aplicados neste artigo, são divididos em 4 partes, que são: formação da base de dados, processo de extração de características, geração do modelo de atribuição e processo de decisão (Fig. 1).

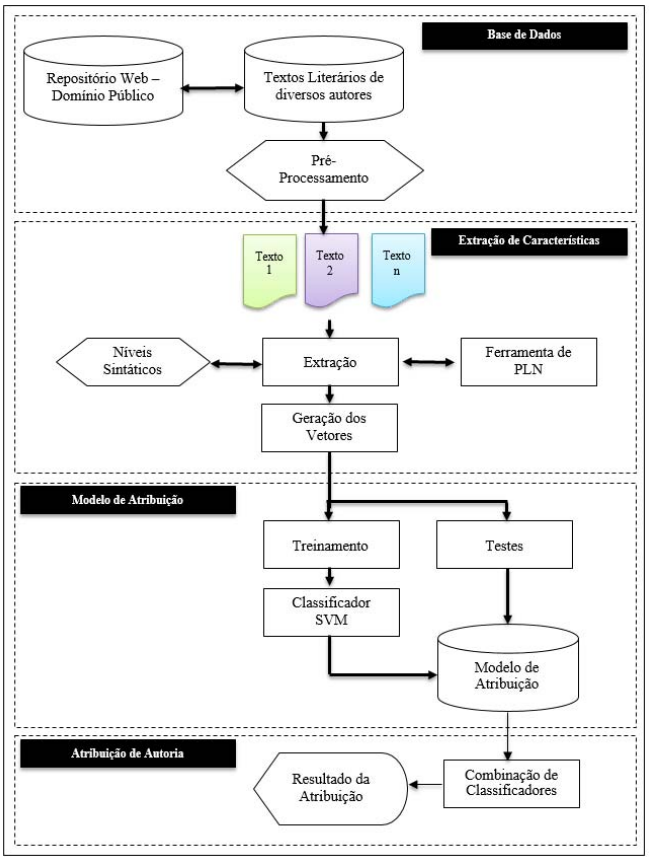

Figura 1. Visão Geral da abordagem.

A. Base de Dados

Para a aplicação da abordagem proposta neste artigo, foi coletada uma base de dados composta por 150 autores de obras literárias em língua portuguesa disponíveis em domínio público na internet. Para cada autor foram coletadas 20 amostras com textos entre 500 e 2000 palavras. Após a coleta e formação da base de dados, todas as amostras de textos coletadas passaram por um tratamento (pré-processamento) para retirar elementos que não pertenciam a obra original, tais como: cabeçalhos e número de páginas. Este processo é necessário, para que não haja interferência no texto original, o que pode influenciar nos resultados de atribuição de autoria.

B. Extração de Características

Para auxiliar no processo de extração das características sintáticas, uma biblioteca de Processamento de Linguagem Natural (PLN) [29] foi utilizada para realizar a rotulagem das palavras em língua portuguesa.

Um software foi desenvolvido para extrair os níveis de características sintáticas em separado e/ou em conjunto (Tabela I).

TABELA I. NÍVEIS SINTÁTICOS.

NÍVEIS	CARACTERÍSTICAS
Essenciais	Sujeito, predicado e verbo
Integrantes	Objeto direto, objeto indireto, voz passiva e ativa
Acessórios	Adjunto adverbial e adjunto adnominal

Quando os vetores de características de cada texto ou de cada autor são gerados, estes passam por um processo de normalização, que consiste na transformação da frequência absoluta na frequência relativa, dada pela equação 1:

$$F_r = F_i/n \quad (1)$$

Onde *Fr*, simboliza a frequência relativa, ou seja, a informação que é carregada no vetor. Já *Fi* significa a quantidade de vezes que uma certa característica aparece em uma amostra de texto, e *n* a quantidade total de palavras do texto.

Cada vetor de característica é composto por 400 características, sendo estas as 50 mais frequentes de cada classe de características.

C. Métrica de Distância

As medidas de distâncias se tornaram essenciais para diversas áreas que envolvem a computação, especialmente as que envolvem a atribuição de autoria. Saber identificar qual a melhor métrica de distância ou a mais adequada para quantificar a proximidade ou o distanciamento entre objetos, vem se tornando uma tarefa importante [30]. Sendo assim, realizamos os experimentos com a distância euclidiana, apresentada na equação 2. A métrica é aplicada para mensurar as distâncias entre textos de um mesmo autor e de autores diferentes, conforme o protocolo aplicado.

$$D_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (2)$$

D. Modelo de Atribuição

Para geração dos modelos de atribuição a tarefa de classificação envolve os conjuntos de treinamento e testes. Em cada instância do conjunto de treinamento existe um rótulo que identifica o autor e também diversos atributos, ou seja, as posições do vetor que identificam as características estilométricas. O classificador utilizado para realização dos experimentos foi o SVM na implementação LIBSVM [31], que tem por função produzir um modelo com base no conjunto de treinamento, para tentar prever a autoria de um documento questionado, que é submetido pelo conjunto de testes. O tipo de *kernel* utilizado para aplicação da abordagem e consequentemente geração do modelo foi o linear. Em

correlato, para a resolução de casos de atribuição de autoria que envolvam o processo de aprendizagem, a base de dados foi dividida em dois subgrupos: treinamento e testes.

E. Treinamento

O treinamento tem por função treinar o modelo de classificação, ou seja, aplicar uma técnica de classificação que consiga diferenciar as classes e características dos autores. No processo de treinamento e geração dos modelos são utilizadas duas abordagens distintas: independente e dependente do autor.

Modelo Independente: é utilizado o conceito de dicotomia, ou seja, existem somente duas possibilidades: autoria ou não autoria. Nesta abordagem é gerado um modelo genérico de autoria e de não autoria através da combinação de amostras de um mesmo autor (amostras positivas – autoria) e a combinação de amostras de autores distintos (amostras negativas – não autoria). Neste caso, os autores que participam do treinamento não fazem parte dos testes, tendo assim um modelo que realizará a classificação de autores nunca vistos anteriormente. Para construção deste modelo foram utilizadas 2 amostras de textos por autor, pois este modelo não necessita de um grande número de amostras por autor. Neste modelo a ênfase está na diversidade de autores e não na quantidade e variedade de amostras. Uma das vantagens do modelo independente é que para a inclusão de novos autores não é necessário a realização de um novo treinamento do modelo.

Modelo Dependente: Neste caso, para cada autor é gerado um modelo de atribuição baseado nas características de estilo sintático de escrita do autor. Este por sua vez, é baseado no conceito da policotomia, ou seja, na classificação do problema em diversos modelos. Neste modelo geralmente utiliza-se um grande número de amostras de texto por autor, pois a ênfase principal está nas características estilométricas de cada autor. Na construção da base de treinamento deste modelo também são geradas amostras verdadeiras e amostras falsas. As amostras verdadeiras são constituídas pela combinação das amostras de um mesmo autor. Já as amostras falsas são constituídas pela combinação do autor com amostras de outros autores. Ao final, o arquivo de treinamento possuirá o mesmo de número de amostras verdadeiras e falsas, ou seja, um vetor balanceado. Isso é necessário para que não haja desiquilíbrio ou tendência do modelo. No modelo dependente do autor todos os autores da base participam do treino e também dos testes. Para tanto, as amostras de textos dos autores que fazem parte do treinamento não fazem parte dos testes, porém as amostras de textos separadas para testes são combinadas com amostras de referências que podem ou não fazer parte do treinamento.

F. Testes

O grupo de testes, é usado para realizar a validação do modelo, ou seja, para verificar o poder de previsão do modelo de atribuição de autoria gerado pelo modelo de classificação. No processo de testes são utilizadas duas abordagens quanto ao processo de atribuição de autoria, que são: verificação e identificação.

Verificação: o objetivo principal é verificar se o modelo criado no processo de treinamento é robusto o suficiente para conseguir classificar corretamente amostras de textos de um mesmo autor. A estratégia utilizada neste tipo de abordagem é um-contra-um. Neste caso é sabido o autor que se deseja comparar, sendo assim realiza-se o processo de verificação com o modelo deste autor. O resultado é verificar se ele acerta ou erra.

Identificação: o objetivo é confrontar a base de dados toda com todos os autores em busca de identificar quem é o autor da amostra questionada. A estratégia utilizada nesta abordagem é um-contra-todos, ou seja, confrontar um texto questionado contra todos os modelos ou classes constantes do treinamento, afim de tentar identificar o provável autor. Neste caso, como o confrontamento é grande e de difícil decisão por parte do classificador, também realizamos a análise dos autores melhores classificados (Top 1, Top 3, Top 5 e Top10).

G. Processo de Decisão

No processo de atribuição de autoria o processo de decisão final é baseado em métodos de fusão. Neste artigo são analisados e explorados os métodos de fusão: voto majoritário, soma, média e mediana conforme a Tabela II. O intuito é verificar qual o melhor método para aplicação em textos.

TABELA II. MÉTODOS DE FUSÃO.

MÉTODO	FÓRMULA
Soma	$soma\ (x) = \max_{k=1}^{c} \sum_{i=1}^{n} P\ (w_k \mid y_i\ (x))$
Média	$média\ (x) = \frac{1}{n} \max_{k=1}^{c} \sum_{i=1}^{n} P\ (w_k \mid y_i\ (x))$
Mediana	$mediana\ (x) = \max_{k=1}^{m} \underset{i=1}{R}\ mediana\ P\ (w_k \mid x_i)$
Voto Majoritário	$voto\ majoritário\ (x) = \max_{k=1}^{c} \sum_{i=1}^{n} y_{i,k}$

IV. PROTOCOLOS DE EXPERIMENTOS

Os protocolos de experimentos foram divididos em treinamento e testes. Na fase de treinamento foram utilizadas as abordagens dependente e independente do autor. Sendo assim, nas Tabelas III e IV são apresentados os protocolos de experimentos de ambas as abordagens.

TABELA III. PROTOCOLO DO MODELO INDEPENDENTE.

PROTOCOLO	TREINAMENTO		TESTES	
	Nº DE AUTORES	%	Nº DE AUTORES	%
A	50	33%	100	67%
B	75	50%	75	50%
C	100	67%	50	33%

TABELA IV. PROTOCOLO DO MODELO DEPENDENTE.

PROTOCOLO	TREINAMENTO		TESTES	
	Nº DE TEXTOS	%	Nº DE TEXTOS	%
D	5	25%	15	75%
E	8	40%	12	60%
F	10	50%	10	50%
G	12	60%	8	40%
H	15	75%	5	25%

Na Tabela III são apresentadas as quantidades de autores utilizados para treino e testes do modelo independente do autor. Para realização do treinamento foram utilizadas 2 amostras de textos por autor. Já para a fase de testes foram utilizadas amostras de textos como referências variando entre 3, 5, 7 e 9, ou seja, uma quantidade de referências ímpares para facilitar o processo de decisão dos métodos de fusão. Na Tabela IV, são apresentados 5 protocolos de experimentos do modelo dependente do autor. Foram utilizadas 20 amostras de textos de cada autor para realização dos experimentos. Nos testes também foram utilizadas referências variando entre 3, 5, 7 e 9. Para cada um dos experimentos realizados foi utilizado o método de *holdout,* que consiste em dividir aleatoriamente a base de dados em uma percentagem fixa para treinamento e testes. Para verificar a robustez dos modelos foram aplicados 3 experimentos para cada protocolo, sendo assim os resultados apresentados são a média das 3 execuções, através da taxa de acurácia.

Para fins comparativos, também foram replicadas as abordagens proposta por [22][23][24]utilizando a mesma base de dados em língua portuguesa. Para implementação completa das abordagens propostas foram utilizadas ferramentas para auxiliar no processo de extração de características disponibilizadas por [32][33].

V. RESULTADOS

H. Verificação de Autoria

A verificação de autoria consiste no processo de verificar se um determinado texto foi escrito ou não por um determinado autor. Inicialmente, foram analisados os resultados dos métodos de fusão utilizados (média, mediana, soma e voto majoritário). Na Tabela V é possível observar os resultados dos métodos de fusão nos 3 protocolos utilizados com número de referência igual a 9 (que proporcionaram os melhores resultados).

TABELA V. RESULTADOS – MÉTODOS DE FUSÃO.

PROTOCOLO	MÉTODOS DE FUSÃO – TAXA DE ACERTO EM %			
	MÉDIA	MEDIANA	SOMA	VOTO
A	84.7	77.3	83.7	89.0
B	87.6	81.3	86.2	94.2
C	89.7	82.3	86.7	95.0

De acordo com a Tabela V verifica-se que o método de fusão que obtém os melhores resultados é o voto majoritário. Observando todos os experimentos realizados com os métodos de fusão denota-se que o voto majoritário se saiu melhor que as outras regras de fusão em todos os experimentos, ou seja, em 100% dos casos. Por conseguinte, a regra de fusão que obteve os melhores resultados foi o voto majoritário.

Também foi analisado o comportamento do modelo em relação a quantidade de autores submetidos ao processo de treinamento. Percebe-se, que nos experimentos houve um acréscimo na taxa de reconhecimento conforme aumenta-se a quantidade de autores no processo de treinamento.

Por conseguinte, foi analisado o comportamento do modelo em relação a quantidade de referências utilizadas para os testes. É possível observar (Tabela VI) que os resultados obtidos em relação a quantidade de referências utilizados são estáveis e com pouca variação. Também é possível observar uma pequena tendência de aumento da taxa de acerto conforme aumenta-se a quantidade de referências. Sendo assim, também denota-se que em todos os experimentos realizados com outros métodos de fusão e quantidade de autores no modelo, a variação média fica entre -0.3 e 2.2%, preconizando uma certa estabilidade do modelo. Neste caso, os resultados apresentados são em função da estratégia do voto majoritário, que reportou as melhores taxas de acurácia.

TABELA VI. VERIFICAÇÃO DE AUTORIA.

PROTOCOLO	NÚMERO DE REFERÊNCIAS – TAXA DE ACERTO EM %			
	3	5	7	9
A	87.0	87.0	88.3	89.0
B	92.0	92.0	92.0	94.2
C	94.0	93.7	94.0	95.0

I. Identificação de Autoria

A abordagem dependente do autor preconiza que exista um modelo para cada autor da base de dados. Sendo assim, a ênfase está na diversidade de características de cada autor. Então, o principal problema do modelo dependente é tentar identificar um determinado autor entre todos os autores que estão no conjunto do experimento. Para tanto, no modelo dependente é utilizada a abordagem de identificação de autoria. A identificação consiste na chamada do processo de verificação *n* vezes, onde *n* é igual a quantidade de autores submetidos ao processo de testes. Muitas vezes, o processo de identificação pode causar confusão entre autores que possuem estilo semelhantes, para tanto, é fornecida uma lista com as amostras que são semelhantes ao documento questionado. Neste caso, os resultados que são apresentados podem ser analisados utilizando os Top 1, Top 3, Top 5 e Top 10. Isto é importante, pelo fato do perito, pesquisador ou linguista pode reduzir o escopo de seu trabalho, ou seja, o trabalho se atém somente a determinados autores e não a sua totalidade, como por exemplo: no Top 10, a o escopo total de 100 autores é reduzido para 10 autores que possuem amostras mais semelhantes com o texto questionado.

Para realização dos experimentos com a abordagem dependente foram gerados modelos individuais para cada um dos autores separados para o experimento.

Um dos experimentos realizado na identificação de autoria, foi verificar os resultados do acerto principal (Top 1) para cada um dos protocolos em função do método de fusão e do número de referências utilizados. Para tanto foram gerados 80 conjuntos de resultados, sendo 5 protocolos x 4 métodos de fusão x 4 referências. No entanto, como o conjunto de resultados é grande, somente são apresentados os melhores resultados para cada protocolo, como pode ser visto na Tabela VII.

TABELA VII. CONJUNTO DE MELHORES RESULTADOS NO TOP 1

CONJUNTO DE MELHORES RESULTADOS.			
PROTOCOLO	TAXA DE ACERTO (%)	REFERÊN-CIAS	MÉTODO DE FUSÃO
D	81.1	9	Voto Majoritário
E	84.5	7	Voto Majoritário
F	88.6	9	Voto Majoritário
G	88.3	9	Voto Majoritário
H	90.3	5, 7, 9	Voto Majoritário

Observando a Tabela VII, pode-se verificar que conforme aumenta-se o número de amostras para a fase de treinamento (conforme protocolos) aumenta-se também a eficácia do modelo, evoluindo de 81-90% de acerto. Porém, percebe-se que o modelo atinge uma certa estabilidade entre os protocolos F e G, o que representa dizer que bons resultados são atingidos utilizando entre 50-60% das amostras para treinamento e 40-50% para realização dos testes. Também observa-se nos resultados que quanto mais referências foram utilizadas no processo de identificação, melhor foi o desempenho, tendo seus ápices entre 7 e 9 referências. Já quando observados os métodos de fusão empregados para análise dos resultados, percebe-se que em todos os protocolos de experimentos envolvendo os textos da base, o melhor método de fusão foi o voto majoritário.

Na Tabela VIII são demonstrados os resultados da abordagem proposta por protocolo de experimento, mostrando os resultados pelo *Hit List* (Top 1, Top 3, Top 5 e Top 10). A estratégia de voto majoritário também foi a estratégia de fusão escolhida para a apresentação dos resultados.

TABELA VIII. IDENTIFICAÇÃO DE AUTORIA.

PROTOCOLO	TAXA DE ACERTO EM %			
	TOP 1	TOP 3	TOP 5	TOP 10
D	78.9	88.3	92.1	97.8
E	84.6	90.6	95.4	100
F	87.9	95.4	100	100
G	88.2	94.2	99.0	100
H	90.3	96.0	100	100

Observa-se na Tabela VIII que o protocolo que possui uma maior quantidade de textos no treinamento, possui uma maior taxa de acurácia (acerto). Isso se deve ao fato de que no modelo dependente do autor, quanto mais informações se tiver sobre o autor, melhor será o modelo criado para a classificação. Um outro fator importante, é observar a taxa de acerto referente a *Hit List.* Neste caso, o acerto é dado verificando se o autor a que pertence o texto questionado está entre a lista de suspeitos (Top1 – Top10) dada pelo classificador. Diante disso, percebemos que abordagem proposta relatou resultados entre 78-90% no Top 1, entre 88-96% no Top 3, entre 92-100% no Top 5 e entre 97-100% no Top 10. Isso evidencia, que o conjunto de características utilizadas na abordagem se mostra uma opção para casos que envolvam a atribuição de autoria em língua portuguesa.

J. Resultados Comparativos

A Tabelas IX apresenta um resumo comparativo do método proposto com alguns trabalhos apresentados na literatura e reproduzidos neste artigo. Neste caso, utilizamos os conjuntos de características propostos por [22][23][24]. Então, é possível observar e estimar as contribuições dadas pelo método proposto.

Observando os resultados obtidos pelo método proposto, na verificação de autoria, percebemos que é possível notar um acréscimo médio de 28% para a abordagem 1[22], de 11% para a abordagem 2 [23] e de 5% para a abordagem 3 [24]. Isto indica, que a melhora decorreu principalmente pela adoção de um conjunto robusto de características. Isso nos leva a crer que para casos que envolvam a verificação de autoria a abordagem proposta se mostra estável.

No caso da identificação de autoria, percebe-se que o método proposto se sobressai as abordagens propostas por [22] e [23]. No entanto, na comparação com a proposta de [24] as abordagens se mostram semelhantes, porém com leve vantagem de cerca de 1-2% da abordagem que propomos. Percebe-se então que as palavras-função, regras *uni* e *bigrams* propostos por [22][23] e [24]obtiveram resultados inferiores aos níveis sintáticos que propomos neste artigo. Isso se deve ao fato da língua portuguesa possuir uma grande quantidade de fatores e complexidades linguísticas, que fazem com que as funções sintáticas se sobressaiam sobre outras características. Sendo assim, constata-se que os resultados da abordagem proposta, através das características sintáticas são promissoras tanto na verificação como na de identificação de autoria.

TABELA IX – RESULTADOS COMPARATIVO COM A LITERATURA.

Método	***Referências***	***Características***	***Taxas de Acurácia da Abordagem***	
			Dependente	***Independente***
Verificação de Autoria	Finn e Kushmerick [22]	Todas as palavras, Palavra-função e 36 rótulos unigrams	-	58-70%
	Hirst e Feiguina [23]	Regras bigrams em análise parcial	-	76-86%
	Zhao e Zobel [24]	Regras uni e bigrams em separado e em conjunto	-	82-91%
	Nossa Abordagem	**Níveis Sintáticos**	**-**	**87-95%**
Identificação de Autoria	Finn e Kushmerick [22]	Todas as palavras, Palavra-função e 36 rótulos unigrams	52-70%	-
	Hirst e Feiguina [23]	Regras bigrams em análise parcial	70-85%	-
	Zhao e Zobel [24]	Regras uni e bigrams em separado e em conjunto	77-89%	-
	Nossa Abordagem	**Níveis Sintáticos**	**79-90%**	**-**

VI. CONCLUSÕES

Foram realizados experimentos utilizando duas abordagens, que são: independente e dependente do autor. No modelo independente do autor a verificação de autoria apresentou resultados promissores, mesmo quando houve a variação da quantidade de autores que foram submetidos ao processo de treinamento e testes, apresentando taxas de acerto entre 87-94%. Quando realizada a variação da quantidade de textos de referências (3, 5, 7 e 9), os resultados apresentados também se demonstraram estáveis, variando cerca de 1,3% entre o pior e o melhor resultado (93,7-95,0%).

Nos experimentos da abordagem dependente do autor, através da identificação de autoria, verificaram-se resultados inferiores ao da verificação de autoria. Houve uma variação de resultados de taxa de acerto entre 81-90% em virtude da quantidade de autores que foram submetidos ao processo de treinamento e testes. Nesta abordagem, também foram analisados os resultados proporcionados pelo *hit list* (Top3, Top 5 e Top 10), onde se denotou que no Top 3 os resultados variando entre 88-96% dependendo do protocolo; entre 92-100% no Top 5; e entre 97-100% na análise do Top 10.

Também foram realizados experimentos comparativos com outras abordagens. Em ambos os experimentos foram aplicados os mesmos protocolos, e a abordagem proposta neste artigo se sobressaiu, gerando taxas de acerto maiores. Sendo assim, conclui-se que as características estilométricas utilizadas neste trabalho são promissoras e já relatam bons resultados no comparativo com outras abordagens.

Sendo assim, a contribuição deste artigo foi apresentar uma abordagem consistente e promissora que faz uso de um grupo de características sintáticas baseada em níveis estruturais da gramática da língua portuguesa que possam ser utilizados em casos que envolvam a atribuição de autoria (verificação e identificação).

Como inserções e trabalhos futuros, citam-se: a inserção de novos grupos de características sintáticas para realização de experimentos em separado e em conjunto, tais como: características morfológicas e sintagmas; a inserção de pesos conforme o conteúdo de cada frase; e a aplicação de algoritmos genéticos para seleção das melhores características e otimização do modelo.

AGRADECIMENTOS

Agradecimentos à Pontifícia Universidade Católica do Paraná pela possibilidade de realização deste trabalho através do fomento do projeto.

REFERÊNCIAS

[1] F. Mosteller and D. L. Wallace, “Inference and disputed authorship: The Federalist”. Reading, Addison-Wesley, 1964.

[2] D. I. Holmes and R. S. Forsyth, “The 'Federalist' Revisited: New Directions in Authorship Attribution”. *Literary & Linguistic Computing*, 10(2), 111-127, 1995.

[3] C. Martindale and D. Mckenzie, “On the utility of content analysis in author attribution: The “Federalist””. *Computer and Humanities*, 29, 259-270, 1995.

[4] C. S. Brinegar, “Mark Twain and the Quintus Curtius Snodgrass Letters: A statistical test of authorship”. *Journal of the American Statistical Associations*, 58, 85-96, 1963.

[5] A. Q. Morton, “The authorship of Greek prose”. *Journal of the Royal Statistical Society*, 128, 169-233, 1965.

[6] C. E. Chaski, “Who’s at the keyboard? - authorship attribution in digital evidence investigations”. *International Journal of Digital Evidence*, 2005.

[7] A. Abbasi and H. Chen, “Applying authorship analysis to extremist group web forum messages”. *IEEE Intelligent Systems*. 20, 67-75, 2005.

[8] E. Stamatatos, “A survey of modern authorship attribution methods”. *Journal of the American Society for Information Science and Technology*. 60, 538-556, 2009.

[9] G. R. McMenamim, “Forensic Linguistics – Advances in Forensic Stylistics”. CRC Press, New York, 2002.

[10] B. Coutinho. J. Macedo, A. Rique and L. V. Batista, “Atribuição de autoria usando PPM”. *In*: *III Workshop em Tecnologia da Informação e da Linguagem Humana*, v. 1, 2208-2217, São Leopoldo, 2005.

[11] D. F. Pavelec, E. J. R. Justino, L. V. Batista and L. E. S. Oliveira “ Author Identification using Writer-Dependent and Writer-Independent Strategies”. *Proceedings of the 23th Annual ACM Symposium in Applied Computing*, 2008. v. 1. p. 414-418.

[12] P. J. Varela, E. J. R. Justino, and L. E. S. Oliveira, “Selecting syntactic attributes for authorship attribution”, *Proceedings of the International Joint Conference on Neural Networks*, pp. 167-172, 2011.

[13] D. I. Holmes, “The Evolution of Stylometry in Humanities Scholarship”. *Literary and Linguistic Computing*. 13, 111-117, 1998.

[14] E. Stamatos, G. Kokkinakis and N. Fakotakis, “Automatic Text Categorization in Terms of Genre and Author”. *Computational Linguistics*. 26, 471-495, 2001.

[15] S. Argamon, M. Koppel, J. Fine and A. Shimoni, “Gender, genre, and writing style in formal written texts”. *Text*. 23, 321–346, 2003.

[16] J. Diederich, J. Kindermann, E. Leopold and G. Paass, “Authorship attribution with support vector machines”. *Applied Intelligence*, 2003.

[17] M. Gamon, “Linguistic correlates of style: Authorship classification with deep linguistic analysis features”. *In Proceedings of the 20th International Conference on Computational Linguistics* (pp. 611–617), 2004.

[18] J. Savoy, “Authorship attribution based on a probabilistic topic model”. *Information Processing and Management*. 49, 341-354, 2012.

[19] M. Ebrahimpour, M. Putnins, T. J. Berryman, M. J. Allison, A. Ng and B. W. H. Abbot, “Automated Authorship Attribution using advanced signal classification techniques”. *PLOS One*, Vol. 8., 2013.

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

[21] R. Zheng, J. Li, Z. Huang and H. Chen, “A framework for authorship analysis of online messages: Writing-style features and techniques”. *Journal of American Society for Information, Science and Technology*. 57, 378–393, 2006.

[22] A. Finn and N. Kushmerick, “Learning to classify documents according to genre”. *Journal of the American Society for Information Science and Technology*. 57, 1506-1518, 2006.

[23] G. Hirst and O. Feiguina, “Bigrams of syntactic labels for authorship discrimination of short texts”. *Literary and Linguistic Computing*, 22, 405–417, 2007.

[24] Y. Zhao and J. Zobel, “Searching with style: Authorship attribution in classic literature”. *In Proceedings of the 30th Australasian Conference on Computer Science*. 62, 59–68, 2007.

[25] B. Belak, S. Belak and A. R.Pesa, “Stylometry – definition and development”. *Annals of DAAAM & Procedings*.85-94, 2008.

[26] N. Besnier, “The linguistic relationships of spoken and written Nukulaelae registers”. *Language*. 64, 707-736, 1988.

[27] H. Baayen, H. Van Halteren and F. J. Tweedie, “Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution”. *Literary and Linguistic Computing*, 11, 121–131, 1996.

[28] O. Uzuner and B. Katz, “A comparative study of language models for book and author recognition”. *Springer Lecture Notes in Computer Science*, 3651, 969–980, 2005.

[29] J. J. Almeida and A. Simões, “Jspellando nas morfolimpíadas: Sobre a participação do Jspell nas morfolimpíadas”. *In Diana Santos, editor, Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa.* IST Press, 2007.

[30] M. M. Deza and E. Deza, “Encyclopedia of Distances”. Springer, 2009.

[31] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines". *ACM Transactions on Intelligent Systems and Technology*, 2:27, 1-27, 2011.

[32] A. Branco and J. Silva, "Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, Paris, ELRA, ISBN 2-9517408-1-6, pp.507-510, 2004.

[33] P. Juola, J. Sofko and P. Brennan, "A prototype for Authorship Attribution Studies*". Literary and Linguistic Computing*. 21, 169-178, 2006.

Paulo Júnior Varela é graduado em Sistemas para Internet pela Faculdade da Fronteira (FAF), Barracão, Paraná, Brasil, em 2006. Obteve o título de mestre em Informática pela Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, Paraná, Brasil, em 2010. Atualmente é professor da Universidade Tecnológica Federal do Paraná (UTFPR) e cursa doutorado em Informático pela Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, Paraná, Brasil. Suas pesquisas se concentram na área de atribuição de autoria e reconhecimento de padrões.

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 e perito Ad hoc do Tribunal de Justiça do Estado do Paraná. Tem experiência na área de Ciência da Computação, com ênfase em Desenvolvimento de Sistema Computacionais para área Forense, tais como: Análise Digital de Documentos Antigos e Contemporâneos, Análise Computacional da Escrita Manuscrita e Assinatura, e Aplicações Computacionais para a Linguística (Estilística Forense).

Flávio Bortolozzi é graduado em Matemática em 1976 e em Engenharia Civil em 1981 pela Pontifícia Universidade Católica do Paraná. Doutorado em Engenharia de Computação pela Université de Technologie de Compiègne - França em 1991. Pesquisador 2 do CNPq. Aposentado pela UTFPr. Pró-Reitor Pesquisa, Pós-Graduação e Extensão e professor do CESUMAR. Foi Diretor do Centro de Ciencias Exatas, Tecnológicas e Agrárias do CESUMAR. Pesquisador colaborador da Pontifícia Universidade Católica do Paraná. Diretor da BDF Consultoria Científica e Educacional. Consultor: do Ministério da Educação-INEP; do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq, etc... Pesquisador do Instituto de Ciências Exatas, Tecnológicas e Inovação - ICETI. Colaborador da Association For Computing Machinery - ACM. Colaborador do Institute of Electrical and Electronics Engineers, Inc.-IEEE. Tem experiência na área de Computação, com ênfase em Análise e Reconhecimento de Imagens e Visão Computacional, atuando com pesquisas nas área de Document Analysis, HMM, Pattern Recognition, Handwritten e Bank Checks. Coordenador e professor do Mestrado em Gestão do Conhecimento nas Organizações no UniCesumar. Professor do Mestrado em Promoção da Saúde no UniCesumar. Em 2015, professor visitante Sênior no PPGIA/PUCPR.

Luiz Eduardo Soares de Oliveira concluiu o doutorado em Engenharia (Philosophiae Doctor Ph. D) - Université du Quebec, École de Technologie Superiéure em 2003. Atualmente é Professor do Departamento de Informática da Universidade Federal do Paraná. Publicou 36 artigos em periódicos especializados e mais de 90 trabalhos em anais de eventos. Em 2003 recebeu o prêmio de excelência pela melhor tese de doutorado da École de Technologie Superiéure. Atua principalmente nos seguintes temas: reconhecimento de padrões, aprendizagem de máquina e visão computacional. Atualmente ocupa o cargo de vice-coordenador do Programa de Pós-Graduação em Informática (PPGInf) da UFPR. Desde 2006 é bolsista Produtividade do CNPq.