Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL'2006), Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

Identificação da Autoria de Documentos Digitais com Base em Atributos Estilométricos da Língua Portuguesa

Daniel F. Pavelec[1], Edson J. R. Justino[1], Cinthia O. De A. Freitas[1]

[1] Pontifícia Universidade Católica do Paraná, Rua Imaculada Conceição, 1155 Prado Velho, Curitiba, Brasil
Laboratório de Computação Forense e Biomentria
pavelec@ppgia.pucpr.br, {edson.justino, cinthia.freitas }@pucpr.br

Resumo. Com a popularização dos *e-mails* e o uso de outras formas de documentos digitais, é cada vez mais difícil comprovar a autoria de textos em demandas judiciais. Os procedimentos tradicionais de perícia grafoscópica não são aplicáveis nesses casos, uma vez que os textos não possuem elementos gráficos que permitam associá-los ao autor. Este artigo tem por finalidade apresentar um método para a identificação da autoria de documentos digitais, com base em atributos estilométricos do autor. A abordagem computacional baseia-se em um classificador SVM (*Support Vector Machine*) e em atributos estilométricos da língua portuguesa.

Palavras Chaves: Estilometria, SVM, Reconhecimento de Padrões.

1 Introdução

Com o crescimento vertiginoso do uso de computadores pela sociedade e com o aumento da popularidade da Internet, é cada vez mais comum o uso de ambientes digitais para a troca de correspondências, armazenamento e manipulação de documentos. Essa mudança de hábito cotidiano trouxe para o ambiente digital demandas judiciais decorrentes das relações sociais que até então, eram comuns ao ambiente convencional [1], [2]. A associação de um documento digital ao autor do mesmo, nem sempre pode ser formalizada ou caracterizada, uma vez que a origem do documento (computador, disquete, CD, impresso, etc.), nem sempre permite o elo entre o documento e o autor.

As características inseridas em um texto através do estilo literário muitas vezes são pistas únicas, pois independe do meio ou suporte em que o documento ou o texto foi aposto (físico ou digital). Em documentos que tiveram sua origem em ambientes digitais (impressos ou em meio digital), não é possível a extração de características grafoscópicas. Dessa forma, torna-se necessário a análise de fatores inerentes ao estilo literário do autor [11].

Este artigo apresenta um método para a identificação da autoria de documentos digitais, com base em atributos estilométricos do autor e está organizado em 7 (sete) seções, a saber: A primeira seção contém essa introdução; Na segunda são apresentado os preceitos da estilística forense e a identificação dos atributos

estilométricos do autor; Na terceira seção são apresentados os conceitos de identificação da autoria usando SVM; Na quarta seção é apresentada uma descrição da base utilizada; Na quinta seção é apresentado o método proposto; A sexta contém a análise dos resultados; Na sétima seção são feitas as considerações finais.

2 Estilística Forense

Estilística Forense é uma subárea da Lingüística Forense dedicada à aplicação da estilística no contexto da identificação da autoria em documentos questionados. A identificação da autoria é realizada através da análise do estilo da linguagem escrita, isto é, a estilística lingüística. A estilística explora as duas premissas de variabilidade da linguagem [10]:

- Dois escritores de uma língua não escrevem exatamente do mesmo modo;
- Um mesmo escritor não escreve do mesmo modo todo o tempo.

A estilística lingüística pode analisar o estilo de duas formas distintas, a qualitativa e quantitativa [6].

2.1 Análise Qualitativa e Quantitativa

O estudo qualitativo da escrita consiste nas formas usadas pelo autor, como e porque elas foram utilizadas. Apesar de algumas linhas de pesquisa questionarem a cientificidade da análise qualitativa, existem algumas razões importantes pelas quais a mesma deve ser considerada e contextualizada [12]:

- A Descrição qualitativa é o passo inicial da análise. A métrica depende das descrições e da categorização dos elementos lingüísticos analisados;
- As Evidências qualitativas são mais demonstráveis em uma audiência nos tribunais, pois ela é a linguagem do conteúdo apresentado;
- Resultados qualitativos mostram um senso de probabilidade estruturado (não quantitativo).

Já o estudo quantitativo avalia a medida da variação na língua escrita. Trata-se de um poderoso complemento para descrição e, portanto é peça fundamental para o sucesso da análise e interpretação do estilo do autor. O foco da análise quantitativa encontra-se na determinação de quanto e com que freqüência, formas determinadas são utilizadas por um autor [12]. A medida quantitativa também apresenta problemas e limitações. Uma dessas limitações reside na escassez de ferramentas de auxílio à análise. Uma abordagem quantitativa exige métricas dos atributos estilométricos. Outra limitação encontra-se na quantidade de texto. O material a ser analisado nem sempre fornece conteúdo suficiente para permitir uma análise adequada.

Apesar das restrições apresentadas, o processo de análise quantitativa possui importantes atributos:

- Possui amparo metodológico e judicial na análise de evidências;
- Utiliza parâmetros mensuráveis.

Para um método computacional, o modelo quantitativo possui atributos que favorecem sua implementação. Em decorrência disso, foi proposto inicialmente nesse trabalho uma abordagem quantitativa.

2.2 Atributos Estilométricos

Para a estilografia forense a responsabilidade ou a consistência estilística do trabalho literário, será determinada normalmente, no momento da classificação das características usadas nos testes estatísticos [4]. As provas estilométricas buscam determinar parâmetros quantitativos e estáveis de conservação e variação das características textuais (a taxa de aparecimento de palavras incomuns, a média do tamanho das orações, o quociente de palavras diferentes em relação ao total, etc.). O conjunto de valores obtidos pela quantização de tais atributos definirá o estilo.

Existem várias classes de atributos estilométricos, tais como [6]: variações em números e símbolos; variações em abreviações; variações no formato de texto; variações em pontuação. No entanto, existem atributos estilométricos que são pertinentes à língua portuguesa e que por isso, possuem um peso maior no processo de análise da autoria de textos nessa língua. Na Tabela 1 são listados alguns atributos estilométricos relevantes.

Tabela 1. Atributos estilométricos da Língua Portuguesa [13].

Características da Língua Portuguesa	Exemplo
Porquês	por que, porque, porquê e por quê
Plural de substantivos simples e compostos	troféus, canis
Gênero de substantivos	A sentinela, o cônjuge
Figuras de estilo	eufemismos, metoníimia
Interjeição	ai! Nossa! basta!
Vícios de linguagem	cacografias, arcaísmo, neologismos, jargões
Trema	agüentar, lingüística
Crase	Ele foi à casa de Pedro
Acentuação gráfica	grátis, pêra, hífen
Numerais	cinqüenta, cincoenta, Hum, um
Iniciais maiúsculas	Brasil, Jesus
Sintaxe	Edison, Edson
Ortografia	seiscentos, sessenta
Conjunções	“ele é tal como seu pai”, “ele é que nem seu pai” “ele é tal qual seu pai”

3. Identificação da autoria utilizando SVM

Os métodos automáticos de verificação da autoria de textos baseiam-se usualmente em duas abordagens, global e pessoal [5]. A abordagem pessoal utiliza um modelo por autor, enquanto que a abordagem global faz uso de um modelo geral para todos os autores. O modelo pessoal, usualmente, exige um conjunto elevado de textos de um dado autor, para a geração de um modelo robusto. No entanto, apresenta a vantagem de modelar adequadamente os atributos estilométricos do autor. O modelo global possui a desvantagem da generalização. No entanto, possui a vantagem de necessitar um número reduzido de textos para cada autor e de não necessitar de um novo treinamento do modelo, diante da inclusão de novos autores.

No treinamento do modelo global, a classe W_1 representa a classe de textos de autores conhecidos e usados para o treinamento. A classe W_2 representa o conjunto de textos de outros autores. Na verificação, o modelo gerado é então utilizado para a comparação com os textos de autores desconhecido (Fig. 1).

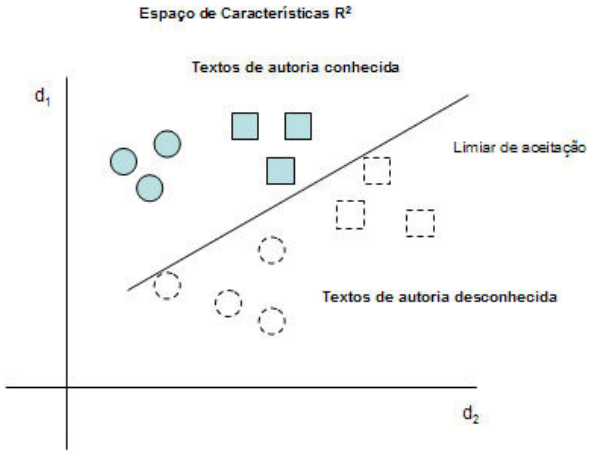


Fig. 1. Modelo global de verificação da autoria de textos.

A metodologia de classificação supervisionada *Support Vector Machine (SVM)* foi desenvolvido por V. Vapnik [9] e é uma nova técnica no campo teórico do aprendizado estatístico. A técnica se baseia no princípio da Minimização do Risco Estrutural (MRS). O princípio da indução do MRS possui dois objetivos. O primeiro é controlar o risco empírico no conjunto de treinamento. O segundo é controlar a capacidade da função de decisão usada para obter esse valor de risco. A Função de decisão do *SVM* linear é descrito por um vetor de peso $\vec{w}$, um limiar b e um padrão de saída $\vec{x}$ (Equação 1).

$$f(\vec{x}) = sign(\vec{w}.\vec{x} + b) \quad \textbf{(1)}$$

Dado um conjunto de vetores de treinamento S_l (Equação 2) pertencente a duas classes separáveis, W_1 ($y_i = +1$) e W_2 ($y_i = -1$), o *SVM* encontra o hiperplano com a máxima distância Euclidiana do conjunto de treinamento. De acordo com o princípio do MRS, existirá somente um hiperplano com a margem máxima δ, definida como a soma das distâncias do hiperplano até o ponto mais próximo das classes. Esse limiar do classificador linear é a separação ótima do hiperplano (Fig. 2).

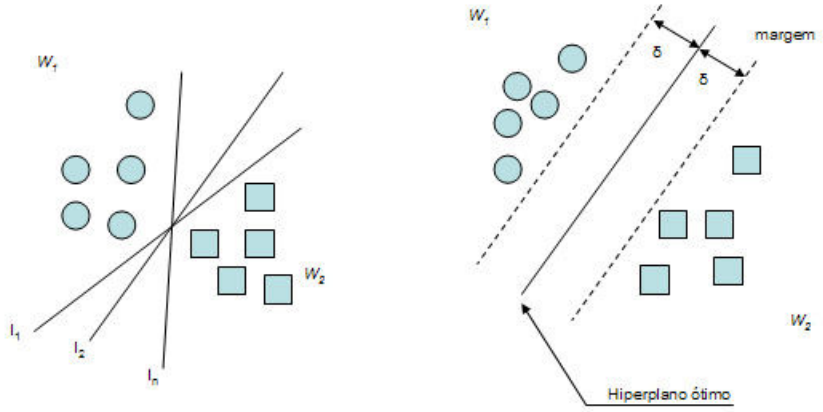
$$S_l = ((\vec{x}_1, y_1), \ldots (\vec{x}_l, y_l)),\ \vec{x}_i \in \Re^n,\ y_i \in \{-1, +1\} \quad \textbf{(2)}$$

Nos casos de conjuntos de treinamento não separáveis, o *i*-ésimo ponto possui uma variável ξ_i, que representa a magnitude do erro de classificação. A função de penalidade $f(\xi)$ representa a soma dos erros de classificação (Equação 3).

$$f(\xi) = \sum_{i=1}^{l} \xi_i \tag{3}$$

A solução do *SVM* pode ser encontrada através da minimização dos erros de treinamento (Equação 4).

$$\min_{\vec{w},b,\vec{\xi}} = \frac{1}{2}\vec{w}.\vec{w} + C\sum_{i=1}^{n} \xi_i, \tag{4}$$



(a) (b)

Fig. 2. Classificação entre duas classes W_1 e W_2 usando hiperplanos: (a) Hiperplanos arbitrários l_i e (b) hiperplano com separação ótima, máxima margem.

A literatura apresenta várias possibilidades de *kernels* para o *SVM* em aplicações envolvendo diversas áreas do conhecimento [7], [8], [10]. Nesse estudo inicial foi utilizado apenas o *kernel* linear (Equação 5).

$$K(\vec{x},\vec{y}) = (\vec{x}.\vec{y}) \tag{5}$$

4. Base de Textos Digitais

Para a formação da base de dados foram coletadas colunas de jornais disponíveis na Internet, de sites oficiais de jornais. Foram escolhidos 10 (dez) colunistas de 2 (dois)

jornais de maior circulação local da cidade de Curitiba-PR. Os dois jornais escolhidos foram a Gazeta do Povo (http://www.gazetadopovo.com.br) e a Tribuna do Paraná (http://www/parana-online.com.br). As colunas foram escolhidas por apresentarem pequenos textos para análise, possuírem conteúdo polêmico e usualmente expressarem a opinião pessoal do colunista.

As colunas selecionadas, num total de 15 colunas por autor, foram arquivadas no formato TXT, com acentuação e sem hifenização. Os tamanhos dos arquivos gerados variam em média de 3 a 6 kbytes. Os autores possuem perfis profissionais variados (economistas, empresários, ex-esportistas, etc.) e tratam de assuntos específicos como: economia, política, comércio exterior, vinhos e cultura, esportes, humor e notícias gerais.

Para cada autor foram usadas 3 (três) colunas distintas para treinamento, 5 (cinco) colunas para referência e 7 (sete) colunas para teste. A base de treinamento é composta por duas classes, a de mesmo autor W_1 (30 amostras) e de autores diferentes W_2 (30 amostras), sendo a primeira (W_1) formada pela comparação, dois a dois, entre as três colunas de um mesmo autor e a última (W_2), formada entre três colunas de autores diferentes, da base selecionada para treinamento. As 5 (cinco) colunas de referência são utilizadas no processo de verificação do exemplar questionado (Fig. 3). Para os testes foram usadas as colunas de referência e teste, numa comparação dois a dois (350 amostras). Para autores diferentes foram combinadas as colunas de referência, de um autor, com as de teste de outros autores, selecionados aleatoriamente na base (350 amostras).

5. Método Proposto

O método proposto se baseia nos princípios da análise das provas estilométricas forense (Fig. 3). Os peritos classificam os textos, em relação à autoria, como associação ou dissociação [5]. 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 a prova pericial, o perito utiliza 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. O perito observa, tendo como base os atributos estilométricos, diferenças de medidas entre as amostras conhecidas e a desconhecida e, posteriormente, apresenta um resultado parcial. O resultado final depende da soma dos resultados obtidos nas comparações individuais dos pares (referência e questionada) (Fig. 3).

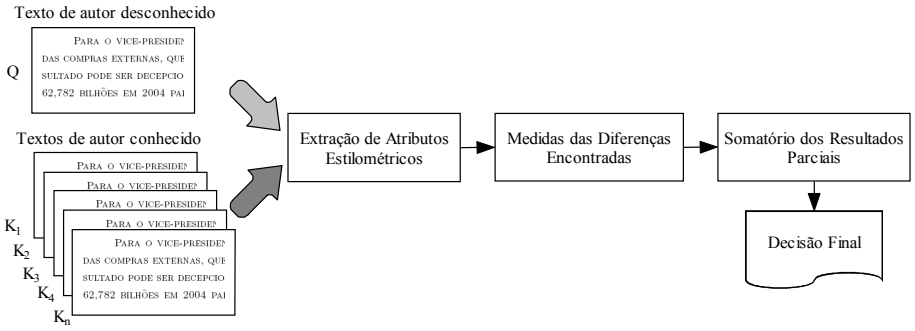


Fig. 3. Diagrama esquemático das provas estilométricas forenses.

5.1 Extração de Atributos Estilométricos

A seleção de atributos estilométricos, para o presente estudo, envolveu o nível léxico e de análise sintática. As características alvo estudadas foram as ocorrências normalizadas de conjunções. A utilização de conjunções foi escolhida por explorar traços inconscientes do autor [4] independendo do assunto tratado pelo autor. Várias conjunções podem ser trocadas por outras sem modificar o sentido e a idéia do texto. Exemplos: "Ele é tal qual seu pai.", "Ele é tal e qual seu pai.", "Ele é tal como seu pai.", "Ele é que nem seu pai.", "Ele é assim como seu pai.".

A Tabela 2 mostra um exemplo de uma coluna obtida e das informações extraídas. Foi extraída das colunas a quantidade de ocorrências de 77 conjunções da língua portuguesa, pertencentes a 12 grupos (Tabela 3).

Como as conjunções não são necessariamente formações de uma única palavra, o processo de extração das características seguiu os seguintes critérios:

- Separador de Palavras: (Espaço em branco), final de linha e caracteres não considerados *tokens* (palavras). Isto é, cada início de linha inicia-se uma nova palavra (sem hifenização);
- Palavras hifenizadas, mesóclise, próclise e ênclise foram consideradas separadamente. Ex: A frase "eu vou dar-te um pula-pula e também dar-te-ei um beijo, meu amor!" possui 16 *tokens* e 12 Hapax (palavras não repetidas);
- Pontuações não foram consideradas *tokens*;
- Não houve diferenciação entre letras maiúsculas e minúsculas;
- Não foram considerados algarismos como *tokens*;
- Caracteres especiais não foram considerados *tokens*.

Tabela 2. Extração de informações do texto (coluna).

Coluna "AEB prevê superávit de US$ 42 bilhões, na balança - 28/08/05"				
Autor	Tamanho	*Tokens*	Hapax Legomena	Nível Lingüístico
Antonio Pietrobelli	5,34Kbytes	800	403	0,503750000

Tabela 3 Conjunções da lingua portugessa [13].

Conjunções	Descrição
Coordenativas Aditivas	e, nem, mas também, mas ainda, senão também, bem como, como também
Coordenativas Adversativas	porém, todavia, mas, entretanto, contudo, senão, no entanto, ao passo que, não obstante, apesar disso, em todo caso
Coordenativas Conclusivas	logo, portanto, por conseguinte, por isso
Coordenativas Explicativas	porquanto, que, porque
Subordinativas Causais	como, visto que, visto como, já que, uma vez que, desde que
Subordinativas Comparativas	tal qual, tais quais, assim como, tal e qual, tal como, tão como, tais como, mais do que, tanto como, mais que, menos do que, menos que, que nem, tanto quanto, o mesmo que
Subordinativas Conformativas	consoante, segundo, conforme
Subordinativas Concessivas	embora, ainda que, mesmo que, ainda quando, posto que, por muito que, por mais que, se bem que, por menos que, nem que, dado que
Subordinativas Condicionais	se, caso, contanto que, salvo que, a não ser que, a menos que
Subordinativas Consecutivas	de sorte que, de forma que, de maneira que, de modo que, sem que
Subordinativas Finais	para que, fim de que
Subordinativas Proporcionais	à proporção que, à medida que, quanto menos, quanto mais

5.2 Medida das Distâncias entre Atributos Estilométricos

A base de dados foi convertida em vetores de características de tamanho *L*, [5]. Os vetores de característica *f* são extraídos dos textos (colunas) K_i *(i=1,2,3...n)* e *Q* (Equações 6 e 7).

$$f_{K\,i_{(i=1,2,\ldots,n)}} = (f_1\ , f_2\ , \cdots f_L\ ) \tag{6}$$

$$f_Q = (f_1\ , f_2\ , \cdots f_L\ ) \tag{7}$$

A medida das diferenças *D* é calculada através do vetor das distâncias Euclidianas D_i *(i=1,2,3,..,n)* entre os textos. O vetor *D* é calculado tanto para o conjunto de treinamento com para o de testes (Equação 8).

$$D_{\,i_{(i=1,2,\ldots,n)}} = \sqrt{(f_{Ki} - f_Q)^2} \tag{8}$$

5.3. Comparação

O processo de comparação é composto por duas fases, o treinamento e a verificação. No estágio de treinamento, as medidas das distâncias entre os atributos D_i *(i=1,2,3,..,n)*, são calculadas entre pares de textos. Quando dois textos (colunas) pertencerem a um mesmo autor, o vetor de característica é indicado com 1 (associação). Quando dois textos pertencerem a autores diferentes, o vetor de característica é indicado com -1 (dissociação). A distância entre dois textos é considerada pequena, quando as amostras pertencerem a um mesmo autor. O *SVM* é treinado então, para separar pequenas distâncias entre atributos estilométricos (associação) e grandes distâncias entre atributos estilométricos (dissociação).

No estágio de verificação, o *SVM* possui duas saídas. A primeira é composta pelos textos pertencentes a um mesmo autor W_1. A segunda é composta por textos pertencentes a autores distintos W_2.

5.4. Decisão

Usualmente, em uma prova estilométrica, o perito faz uso de um conjunto de amostras de textos de origem conhecida. Cada amostra conhecida pertencente ao conjunto de referência (4 a 10 amostras) é comparada com a amostra de autoria desconhecida ou questionada. Nesse experimento foram utilizadas 5 amostras de referência para cada autor.

Com o objetivo de gerar a decisão final, o método proposto classifica as saídas através de um somatório dos resultados. Esse último estágio representa a decisão final do perito (Fig. 3).

6. Resultados

A Tabela 4 mostra os resultados obtidos usando SVM com *kernel* linear. A taxa de acerto de 75,1% demonstra a capacidade discriminatória das conjunções com atributos estilométricos, apesar de apresentar uma taxa de falsa aceitação elevada 34.2%. Outro aspecto relevante encontra-se no tamanho reduzido dos textos, em média 10 vezes menores do que os utilizados por Coutinho [14], no qual obteve uma média de 78% de acerto. Mesmo usando uma abordagem global para a classificação, o método proposto se mostra robusto.

Tabela 4. Resultados obtidos usando SVM e *kerne*l linear.

Decisão Final	Falsa Rejeição (Erro Tipo I) (%)	Falsa Aceitação (Erro Tipo II) (%)	Erro Médio (%)
SVM linear	**15,7%**	**34,2%**	**24,9%**

7. Conclusão e Trabalhos Futuros

O objetivo principal desse artigo foi apresentar um método para identificação de autoria de textos em documentos digitais, tendo como base os princípios da Estilística Forense. Para esse propósito, foram utilizadas apenas duas classes (associação e dissociação). O modelo mostrou-se robusto na redução do número de exemplares por autor, durante o treinamento do modelo e na eliminação da necessidade de um novo treinamento. Os resultados apresentados demonstram a potencialidade do uso do método em procedimentos de prova pericial.

Como proposta para trabalhos futuros inclui-se a comparação com o modelo de compactação proposto por Coutinho [14] e a inclusão de outros atributos estilométricos, permitindo ao modelo absorver mais adequadamente as variabilidades estilísticas do autor, reduzindo assim a taxa de falsa aceitação e falsa rejeição.

Referências

1. Silva A. de: Onus e Qualidade da Prova Cível. Aide, Rio de Janeiro-RJ, (1991).
2. Moreira, J. C. B.: Temas de Direito Processual, Saraiva, 5a edition, (1988), pp. 304
3. Cha, S. H.: Use of the Distance Measures in Handwriting Analysis. Doctor Theses. State University of New York at Buffalo, EUA, (2001), p. 208.
4. Izquierdo, A. J.: Lo falso auténtico: cosas en personas, Borrador preliminar del texto preparado como contribución al seminario Los soportes materiales de la identidad CEIC-IKI / UPV-EHU Leioa, UNED, Madrid, (2004), pp.39.
5. Santos, C. R., Justino, E. J. R., Bortolozzi, F. Sabourin, R.: An Off-Line Signature Verification Method based on the Questioned Document Expert's Approach and a Neural Network Classifier, In: The Ninth International Workshop on Frontiers in Handwriting Recognition, Tokyo, (2004), 10-14p.
6. Olsson, J.: Forensic Linguistics - An Introducion to Language, Crime and Law. Continuum, New York-NY, 1a edition, (2004), pp. 269
7. Joachims, T., Optimizing Search Engines Using Clickthrough Data, ACM Conference on Knowledge Discovery and Mining (KDD), (2002), 1-10p.
8. Burges, C. J. C.: A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery 2, (1998), 121-167p.
9. Vapnik, V. :Statistical Learning Theory, Wiley, N. Y.(1998), pp. 768.
10. Black, H. C., Nolan, J.R., Nolan-Haley, J.M.: Black's Law Dictionary, West Publishing, 6 edition, St. Paul, (1990), pp. 1810
11. Morris, R N.: Forensic Handwriting Identification Fundamental Concepts and Principles, Academic Press, London-UK, (2000), pp. 238
12. Johnstone, B.: Qualitative Methods in Socioliguistics. Oxford University Press, New York, (2000).
13. Ulisses Infante. Curso de Gram´atica Aplicada aos Textos. Editora Scipione, São Paulo-SP, 6a edition, (2001), pp. 512.
14. Coutinho B. C, J. Macêdo L. M., Rique Júnior A., and Batista L. V.. Atribui¸cão de autoria usando ppm. XXV Congresso da Sociedade Brasileira de Computaçãao - Unisinos - São Leopoldo/RS, Julho (2004). 2208 – 2217p.