PAULO JÚNIOR VARELA

O USO DE ATRIBUTOS ESTILOMÉTRICOS NA IDENTIFICAÇÃO DA AUTORIA DE TEXTOS

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

CURITIBA

2010

PAULO JÚNIOR VARELA

O USO DE ATRIBUTOS ESTILOMÉTRICOS NA IDENTIFICAÇÃO DA AUTORIA DE TEXTOS

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

Área de Concentração: *Computação Forense e Biometria*

Orientador: Prof. Dr. Edson José Rodrigues Justino
Co-orientador: Prof. Dr. Luis E. Soares de Oliveira

CURITIBA
2010

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

V293u
2010

Varela, Paulo Júnior
O uso de atributos estilométicos na identificação da autoria de textos / Paulo Júnior Varela ; orientador, Edson José Rodrigues Justino ; co-orientadores, Luis E. Soares de Oliveira. – 2010.
xvii, 89 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2010
Bibliografia: f. 79-84

1. Autoria - Identificação. 2. Identificação biométrica. 3. Lingüística forense. 4. Algoritmos genéticos. I. Justino, Edson José Rodrigues. II. Oliveira, Luis E. Soares. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título.

CDD 20. ed. – 005.8

Pontifícia Universidade Católica do Paraná

ATA DE DEFESA DE DISSERTAÇÃO DE MESTRADO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

DEFESA DE DISSERTAÇÃO Nº 09/2010

Aos 13 dias do mês de Setembro de 2010 realizou-se a sessão pública de Defesa da Dissertação **"O Uso de Atributos Estilométricos na Identificação da Autoria de Textos."** apresentada pelo aluno **Paulo Junior Varela** como requisito parcial para a obtenção do título de Mestre em Informática, perante uma Banca Examinadora composta pelos seguintes membros:

Prof. Dr. Edson José Rodrigues Justino PUCPR (Orientador)	*(assinatura)*	Aprovado *(aprov/reprov.)*
Prof. Dr. Jacques Facon PUCPR		Aprovado
Prof. Dr. Luiz Eduardo Soares de Oliveira UFPR		APROVADO.
Prof. Dr. Luiz Antônio Pereira Neves UFPR		APROVADO

Conforme as normas regimentais do PPGIa e da PUCPR, o trabalho apresentado foi considerado Aprovado (*aprovado/reprovado*), segundo avaliação da maioria dos membros desta Banca Examinadora. Este resultado está condicionado ao cumprimento integral das solicitações da Banca Examinadora registradas no Livro de Defesas do programa.

Prof. Dr. Mauro Sérgio Pereira Fonseca
Diretor do Programa de Pós Graduação em Informática

Programa de Pós Graduação em Informática PPGIa PUCPR

Dedico este trabalho à minha esposa Denise, meu pai Loury e minha mãe Nerci.

Agradecimentos

Agradeço primeiramente a DEUS, por me guiar, me dar forças e coragem.

A minha família por saber que o estudo é um dom precioso. A minha companheira e esposa Denise, por estar ao meu lado durante toda esta fase, sempre me dando motivação quando o cansaço tomava conta. Aos meus pais, que compreenderam a real necessidade de estar longe, apesar de a saudade doer muito, sempre me incentivaram a continuar estudando.

Ao Professor Edson Justino, meu orientador, pela oportunidade, pelo seu trabalho e dedicação, mas particularmente pela força de seu caráter e personalidade. Pelas idéias, críticas e suas preciosas sugestões.

Ao Professor Luiz Soares, pelas suas valiosas contribuições sempre bem vindas.

Ao Professor Jacques Facon, pelas dicas e motivação.

Aos professores e funcionários do PPGIA que de uma maneira ou outra contribuíram para a conclusão deste trabalho.

Aos colegas de mestrado, que juntos sofremos e apreendemos o real valor do esforço.

A CAPES pelo apoio financeiro em parte desta pesquisa.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AC	Antes de Cristo
CPLP	Comunidade dos Países de Língua Portuguesa
KB	*Kilobytes*
MRS	Minimização do Risco Estrutural
SV	*Support Vector*
SVM	*Support Vector Machine*
PPM-C	Variação do algoritmo *Prediction by Partial Matching*
RBF	*Radial Basis Function*

Resumo

A utilização do meio computacional para a resolução de casos de identificação de autoria tem crescido progressivamente em áreas como a computação, a linguística e o direito. Nos últimos anos pesquisadores tem se empenhado em estabelecer uma metodologia capaz de auxiliar na identificação de um documento textual questionado. Estas pesquisas impulsionaram o desenvolvimento de métodos computacionais para auxiliar nas tarefas de seleção e análise de características estilométricas, como também na atribuição da autoria. Entretanto, tais métodos não levam em consideração o idioma usado, o que dificulta o uso destas em países que falam a língua portuguesa. Este projeto tem por finalidade avaliar estatisticamente a importância da utilização de características da língua portuguesa para a identificação de autoria em documentos questionados. Para tal foram necessários: formação de uma base de dados de autores de língua portuguesa; seleção de características estilométricas que visam à identificação do autor; geração dos vetores de dissimilaridade; produção de modelos de aprendizado e testes; análise dos resultados obtidos em comparação com os outros métodos e características já utilizadas. Para a classificação dos textos questionados foi utilizado o classificador SVM, e para seleção das melhores características foi utilizado de algoritmos genéticos. Obtiveram-se resultados promissores, em um modelo independente do autor atingiu-se o patamar de 74,5% de reconhecimento; e em um modelo dependente do autor 80%. Ainda foram identificados quais os conjuntos de características relevantes, dependendo do assunto abordado no texto em análise.

Palavras-Chave: Estilometria, identificação de autoria, linguística forense, SVM.

Abstract

The computational solution uses to solve problems related to the authorship identification and verification has grown progressively in areas such as computing, linguistics and law. In recent years researchers have been attempting to establish a methodology that can be able to identify a questionable textual document. These studies boosted the development of computational methods to assist in the tasks of selection and analysis of style characteristics, as well as in the attribution of authorship. However, such methods do not take the language used into consideration, making the use of this approach difficult in countries that speak Portuguese. This project aims to assess statistically the importance of using the Portuguese language features for authorship identification in questionable documents. For that it was necessary: build a text database in Portuguese Language; develop a protocol for select the best features in a linguistic group to identify the authorship; create the dissimilarity protocol; produce learning and testing models; analyse the results and compare with other methods and features already used. The SVM classifier was used to classify the questioned texts and the genetic algorithms were used to select the best features. Promising results were obtained. In the authorship independent model was reached 74,5% of recognition, and in the authorship dependent model the result was 80%. It was also identified the best combinations of features, depending on the text subject.

Keywords: Stylometry, authorship identification, forensic linguistics, SVM.

Capítulo 1

Introdução

Nos últimos tempos tem sido de grande valia as pesquisas e as descobertas referentes aos estudos sobre a escrita individual manuscrita e as assinaturas, no que se refere à identificação de autoria em documentos questionados [JUS02] [BAR05]. Para tanto, são utilizadas várias técnicas computacionais para a extração e a análise de características, que asseguram a identificação da autoria de forma precisa, quando estas são submetidas a várias abordagens.

Entretanto, uma abordagem que está se sobressaindo e ganhando atenção entre os pesquisadores em relação à identificação de autoria, é o estilo literário do autor [PAV07]. Este tipo de abordagem vem tendo grande importância pelo advento da tecnologia, pois é visto uma crescente série de atividades ilícitas, como por exemplo, a utilização da internet para envio de uma mensagem anônima de ameaça, cartas de supostos suicídios e códigos maliciosos. Contudo, os computadores podem ter várias evidências desta atividade ilícita, mas como um computador pode ser acessado por várias pessoas é difícil saber quem é o autor de tal mensagem. E diante disso, um documento sem a identificação do autor, não possui amparo legal como prova de autoria, diante da justiça. As características de estilo contidas em um texto, praticamente são únicas, pois são independentes da forma que estejam armazenados (meio digital, impresso ou escrito à mão). O modo de como cada autor se expressa em um texto, chama-se estilo literário.

A estilística forense é uma sub-área da linguística forense dedicada a encontrar evidências da autoria através do estilo literário utilizado em documentos questionados. A identificação da autoria é realizada através da análise do estilo da linguagem escrita. Diante deste pressuposto surgem as características estilométricas do documento questionado,

chamada de estilometria. A estilometria visa determinar parâmetros quantitativos e estáveis de conservação e variação das características textuais, como por exemplo: o número de palavras em uma frase, o uso de palavras incomuns, variações no formato do texto e o uso de abreviações. O conjunto de características obtido definirá o estilo de cada autor. [MCM02]

1.1 Desafio

Na esfera jurídica, muitos processos estão inter-relacionados diretamente com o questionamento da autoria de documentos impressos e digitais. Podem se citar vários exemplos que podem ser encontrados nesse meio, tais como: cartas de ameaça, cartas de sequestro, e-mails, notas de resgate, bilhetes e cartas de difamação, cartas de suicídios, livros, artigos e colunas em panfletos, jornais e revistas, bem como os demais documentos que cuja autoria seja desconhecida e a análise da grafia não seja possível de aplicar para identificar o autor [PAV07].

Atualmente, a utilização de tais documentos como prova fica sujeita a análise por parte dos peritos designados pelos juízes. No entanto, este processo de análise ainda é pouco conhecido e utilizado no Brasil. O principal ponto crítico da análise destes documentos por partes dos peritos, é que os mesmos não possuem um método padrão de análise e nem mesmo ferramentas que possam auxiliar na identificação de autores de língua portuguesa. Cabe ressaltar as questões da imprecisão dos métodos linguísticos, que sofrem ainda com a influência demasiada do perito e de sua subjetividade.

O desafio é trabalhar com uma base de dados com textos de tamanho reduzido (máximo de 1200 palavras por texto), para obter resultados significativos e aplicáveis na computação, na linguística e no direito brasileiro.

1.2 Motivação

Muitas pesquisas sobre características estruturais e análise do estilo literário já foram desenvolvidas, mas em sua grande maioria em idioma inglês [MCM02] [OLS04] [CHA01] [CHA05], o que dificulta a resolução de problemas presentes na língua portuguesa. As pesquisas sobre identificação de autoria em língua portuguesa estão em plena expansão, mas ainda é embrionária [CMRB04]. Então, por si só a identificação de autoria em documentos

digitais em língua portuguesa é um dos fatores motivacionais deste estudo. Entre os outros fatores motivacionais, citam-se:

- A quantidade de elementos linguísticos associado à língua portuguesa, que ainda não foram devidamente estudados no contexto da identificação da autoria de textos;
- A ampla aplicabilidade no contexto legal que pode advir de estudos associados com o tema em questão;
- Por se tratar de um problema ainda em aberto, as contribuições de pesquisas, no tema em questão, podem ser de grande valia;
- A importância do estudo das características da língua portuguesa para identificação de autoria;

1.3 Objetivos

O objetivo geral deste trabalho é evidenciar a importância da utilização de características da língua portuguesa para a identificação de autoria em documentos textuais questionados.

Com a proposta de apresentar uma abordagem para identificação de autoria em documentos questionados, este trabalho visa:

- Criar uma base de dados com 3000 textos, sendo que estes separados por autores e por assunto (10 classes de assunto, 10 autores por classe e 30 documentos por autor);
- Propor o uso de novas características vinculadas à gramática da língua portuguesa;
- Avaliar o conjunto de características já apresentadas por Pavelec [PAV07], e adicionalmente testar o potencial de desempenho (contribuição) isolado e no conjunto das novas características;
- Realizar testes com duas abordagens de modelos: dependente e independente do autor;
- Utilizar um processo automatizado para a extração de características;
- Apresentar resultados que possam contribuir para o trabalho realizado por peritos e linguistas;

1.4 Contribuições

Nesta subseção apresentam-se contribuições deste trabalho, que são:

- Formação de uma base de dados de textos digitais para a validação de procedimentos computacionais e que sirva como suporte para trabalhos futuros;
- Análise de desempenho das características de cada classe, propostas neste trabalho;
- Desenvolvimento de uma metodologia de processos embasada cientificamente, que possa auxiliar peritos, linguistas e juízes em situações que exijam a análise de documento de autoria questionada;
- Propor novas características ainda não utilizadas (pronomes, verbos e suas conjugações);

1.5 Organização

Este trabalho está organizado em capítulos, sendo que o primeiro capítulo refere-se à introdução, que contém o desafio, a motivação, a proposta e a contribuição deste trabalho. O Capítulo 2 apresenta um estudo sobre a língua portuguesa, bem como as classificações estilométricas da mesma. No Capítulo 3 é apresentado o estado da arte relacionando os principais trabalhos em relação à identificação de autoria e a estilometria. No Capítulo 4 é detalhada a metodologia de aplicação deste trabalho. Já na seção 5 são evidenciados os resultados obtidos com a pesquisa, no que tange à base de dados utilizadas, protocolos de experimentação e de avaliação.

Capítulo 2

Fundamentação Teórica

O propósito deste capítulo é apresentar uma introdução aos assuntos necessários para o entendimento deste trabalho. Entre os assuntos abordados neste capítulo estão: a língua portuguesa, linguagem, linguística, estilística, linguística e estilística forense, estilometria. Outro assunto relacionado é a utilização de documentos questionados no âmbito jurídico, que compreende conceitos de prova, o procedimento de prova e o processo pericial. Também é apresentado o classificador SVM e suas características, bem como detalhamento de algoritmos genéticos.

2.1 A Língua Portuguesa

A língua portuguesa nasceu na península ibérica (hoje Portugal) influenciado pelo latim e pelo catalão, no século III A.C. Desde então a língua portuguesa tem se expandido, ocasionado principalmente pela colonização portuguesa em regiões da Ásia, África e América.

Hoje ao redor do mundo, cerca de 230 milhões de pessoas falam a língua portuguesa nativamente. Sendo que a língua portuguesa é a oitava língua mais falada no mundo, a terceira entre as línguas ocidentais, e a segunda língua latina.

A língua portuguesa é a língua oficial em oito países de quatro continentes, que são:

- Angola (África) – Aproximadamente 10,9 milhões de habitantes;
- Brasil (América do Sul) – Aproximadamente 185 milhões de habitantes;
- Cabo Verde (África) – Aproximadamente 415 mil habitantes;

- Guiné Bissau (África) – Aproximadamente 1,4 milhões de habitantes;
- Moçambique (África) – Aproximadamente 18,8 milhões de habitantes;
- Portugal (Europa) – Aproximadamente 10,5 milhões de habitantes;
- São Tomé e Príncipe (África) – Aproximadamente 182 mil habitantes;
- Timor Leste (Ásia) – Aproximadamente 800 mil habitantes.

O português atualmente possui dois padrões: o português europeu e africano e o português brasileiro. As suas principais diferenças estão no vocabulário, na pronúncia e na sintaxe. Com o intuito de diminuir tais diferenças a Comunidade dos Países de Língua Portuguesa – CPLP, já proveu um acordo ortográfico da língua portuguesa em 1990, e no ano de 2009 houve um novo acordo ortográfico que visa estreitar ainda mais os laços entre o português praticado na Europa e na África com o português brasileiro.

A língua portuguesa brasileira é muito rica em seu vocabulário e na sua fonética, o que proporciona uma enorme variedade linguística [NAS66], que é apresentada na seção 2.1.1.

2.1.1 Língua Portuguesa Brasileira

A colonização portuguesa no Brasil começou a partir de sua descoberta em 1500. A língua nativa nesta época era o tupi (tupinambá) que foi usado como língua geral da colônia, ao lado da língua portuguesa até 1757. A partir de 1758 a língua portuguesa de tornou o idioma oficial do Brasil.

Das línguas indígenas, a língua portuguesa brasileira herdou várias palavras, como: abacaxi, mandioca, caju. A língua portuguesa brasileira também recebeu contribuições e influências africanas ocasionado pelo fluxo de escravos. Algumas contribuições africanas são palavras como: samba, moleque, caçula. Após a independência, o português brasileiro sofreu influências de imigrantes europeus que se instalaram no centro e no sul do país. De certa forma isso explica a variedade linguística e algumas mudanças superficiais de léxico que existem entre as regiões do Brasil, que variam de acordo com o fluxo migratório das pessoas.

Existe atualmente uma proposta de classificação geográfica (Figura 2.1), baseado em diferenças de pronúncias e na cadência da fala. Segundo esta proposta, é possível distinguir dois grupos de dialetos brasileiros: o do norte que é situado na região norte e nordeste do país, e o do sul que engloba a região central e sul. Dentre esses, o do norte possui duas variedades:

amazônica e nordestina; Já o do sul, são cinco variedades: baiana, fluminense, mineira, sulista e indefinido. [NAS66]

Figura 2.1: Proposta de Classificação da Variedade Linguística Brasileira [NAS66]

Conforme [PAV07], além dos fatores geográficos, existem alguns fatores que proporcionam a variedade linguística da língua portuguesa brasileira, tais como:

- Fatores Sociais

Existem muitas diferenças entre a língua portuguesa brasileira praticada por indivíduos que tiveram acesso a educação e os indivíduos que foram privados de instrução. Neste sentido a língua torna-se uma ferramenta de dominação e discriminação social.

- Fatores Profissionais

Para exercer certas atividades profissionais é necessária a utilização de uma linguagem técnica. Tal linguagem é repleta de conceitos técnicos e específicos da área que se torna importante na comunicação entre especialistas.

- Fatores Situacionais

O ser humano tem a capacidade de adaptar-se ao meio ao qual está inserido, isso faz com que a língua utilizada seja aplicada em formas diferentes para ambientes diferentes. O

fator situacional está diretamente ligado com os resultados da análise quantitativa de atributos estilométricos.

- Fatores literários

Sendo a língua portuguesa essencialmente rica em seus escritos, quando o autor de um texto utiliza a língua, o mesmo passa a se preocupar com a estética das palavras, combinando e criando elementos linguísticos, o que gera a língua literária.

Todos estes fatores possuem formas discriminatórias, que podem ser utilizadas na identificação do estilo de autor de língua portuguesa brasileira. Para se definir o estilo de cada autor, e necessário o entendimento da linguística e suas derivações, que são apresentadas nas seções 2.2, 2.3 e 2.4.

O alfabeto da língua portuguesa é composto de 26 letras, sendo que cada uma delas tem sua forma minúscula e maiúscula, como representado na Figura 2.2.

A a	B b	C c	D d	E e	F f	G g	H h	I i	J j
K k	L l	M m	N n	O o	P p	Q q	R r	S s	T t
U u	V v	W w	X x	Y y	Z z				

Figura 2.2: Letras do Alfabeto da Língua Portuguesa

2.2 Gramática

Gramática é o estudo sistemático e descritivo do sistema interno da linguagem. É o conjunto de regras e exemplos do que se deve ou não deve fazer em uma linguagem, ou seja, regras normativas que estudam a forma, a composição e a inter-relação das palavras para o uso correto e adequado da linguagem, para se escrever e falar corretamente. [MCM02]

2.3 Linguagem

A linguagem humana é um sistema de comunicação que combina os sons com significados para produzir o que o ser humano conhece como linguagem natural. É um código que comunica o significado dos sons e dos movimentos, os gestos, a linguagem corporal e até mesmo códigos fontes de programas de computacionais.

Cada ser humano possui preferências e características tanto no se expressar através da linguagem escrita ou da escrita falada. Tais peculiaridades compõem o estilo literário do autor, independentemente de escolhidas inconscientemente (hábito) ou por opção consciente.

Alguns linguistas dividem o estudo da linguagem em certo número de áreas que são estudadas de forma independente. As divisões mais comuns podem ser observadas na Figura 2.3.

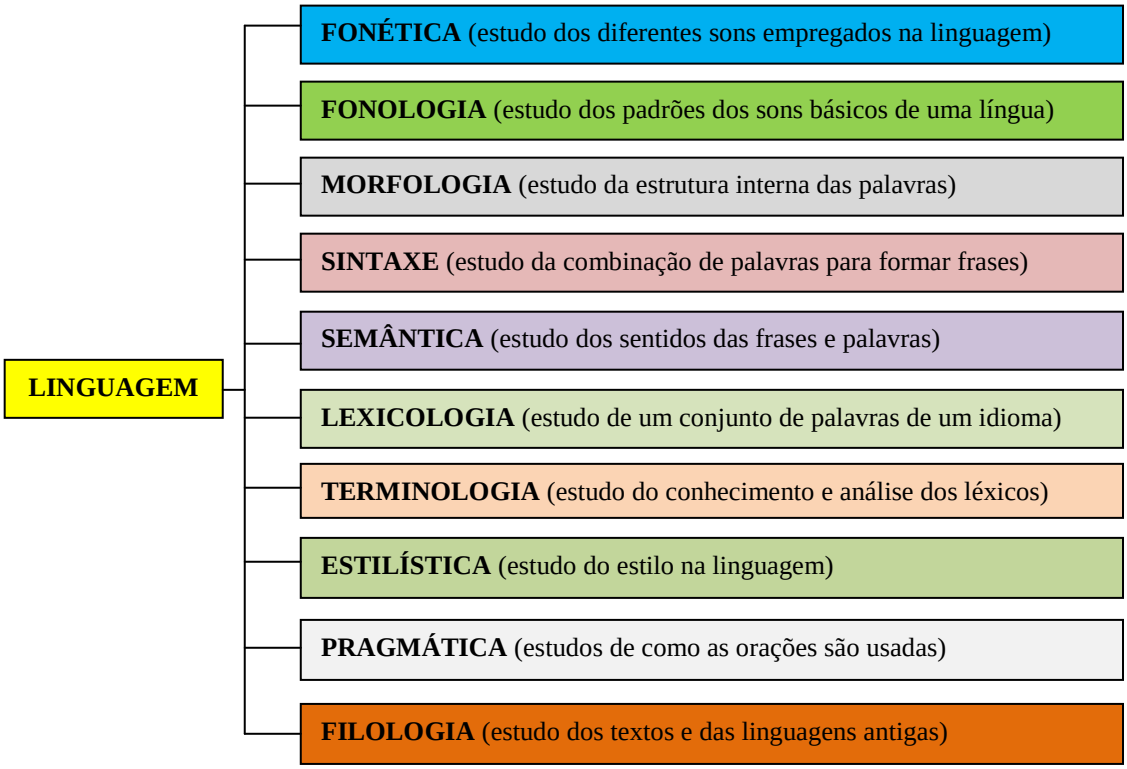

Figura 2.3 – Divisões do Estudo da Linguagem [PS06]

Na seção 2.4 é abordada a língua portuguesa e seus padrões, que são objeto de estudo deste trabalho.

2.4 Linguística

É o estudo cientifico da língua [CRY00]. Linguística é a ciência que estuda a linguagem verbal humana em seu papel teórico e prático. No aspecto teórico, estuda as características presentes em uma língua especifica ou em um grupo similar de línguas. No

campo prático, atua no conhecimento da língua objetivando a melhoria da comunicação através da língua [MCM02].

A variedade linguística deixa evidências de estilos, com os quais é possível associar indivíduos ou uma classe de indivíduos que possuam as mesmas características, existentes em documento cuja autoria seja questionada. [PAV07]

A linguística pode ser dividida em várias áreas, dentre elas algumas são de suma importância para a compreensão deste trabalho e são detalhadas nas seções 2.4.1, 2.4.2, 2.4.3, 2.4.4 e 2.4.5.

2.4.1 Linguística Estilística

A linguística estilística é a ciência que analisa o estilo individual de um escritor, ou seja, o estudo científico dos atributos estilométricos de um individuo. [MCM02]

2.4.2 Estilística

A ciência que estuda as escolhas linguísticas é chamada de estilística. Estilística é um ramo da linguística que estuda as características dos usos distintivos da língua, de acordo com as variantes (situações) [CRY00]. Tenta também estabelecer princípios capazes de explicar as opções feitas por indivíduos ou grupos sociais quando utilizam a língua.

A estilística literária trabalha com as variações próprias da literatura como gênero e do estilo de cada escritor. A quantificação dos padrões estilísticos é objeto de estudo da estiloestatistica – que geralmente se ocupa da estrutura estatística de textos literários, muitas vezes em computadores. [CRY00]

Algumas informações estilísticas são características individuais dos seres humanos enquanto autores e servem para a identificação do estilo do autor, já outras são descritas por normas e políticas (regras organizacionais) e servem para identificação do estilo de um grupo [AIR05].

2.4.2.1 Estilo

Quando o ser humano se expressa utiliza diversas normas de gramática e de uso da língua. Fica evidente que o uso da língua envolve uma grande quantidade de escolhas, entre elas a escolhas das palavras e de orações. As escolhas linguísticas são efetuadas baseando-se em características pessoais do autor (dialetos) e em restrições contextuais (formas de expressão). [AIR05]

O estilo é um fator de individualidade, sendo uma expressão distintiva de um autor, grupo, ou uma combinação destes, e pode ser definida como um conjunto de características que podem permitir a identificação dos mesmos. [AIR05]

2.4.2.2 Estilometria

A estilometria é uma área crescente dentro da estilística, que trabalha com a análise quantitativa do estilo de escrita. O principal objetivo da estilometria é encontrar informações a respeito do autor através de caracteríticas do estilo de escrita do autor que sejam mensuráveis.

Baseado em análises estatíticas de textos e o modelo estilométrico deste trabalho não foge a regra.

A estilometria vem sendo usada em três áreas principais: (i) na descrição de características estilísticas de períodos históricos; (ii) na identificação de características de estilo de escrita de um autor em particular; e (iii) na procura de conjuntos de características estilísticas associados a diferentes generos.

Esta pesquisa é baseada nas áreas (ii) e (iii) identificadas acima, e tem como aplicabilidade a área forense.

2.4.3 Linguística Forense

A linguistica forense é um ramo da linguística que estuda os diversos pontos de encontro entre a linguagem e a lei, com o fim de apontar evidências linguísticas nos processos judiciais.

As relações entre a linguística e o direito estão se aproximando com o passar dos tempos, como resultado do interesse de linguistas, juristas e pesquisadores que estão mostrando em suas respectivas áreas os resultados alcançados com suas aplicações práticas.

Existem três grandes áreas de atuação da linguística forense [GIB94], que são:

1. *A linguagem da lei*: é a linguagem dos textos legais, ou seja, a linguagem com que são escritas as leis e suas formas de interpretação;
2. *A linguagem dos processos legais*: é a linguagem como instrumento para a argumentação legal tanto nas exposições orais, como na elaboração de sentenças;
3. *Evidências linguísticas nos processos legais*: o uso, a validade e a confiabilidade de evidências linguisticas em processos judiciais, ou seja, a análise de material linguistico em diferentes níveis (fonológico-fonético, morfo-sintático, lexico-semântico, pragmático-discursivo) e seu valor probatório no desenvolvimento dos processos.

O escopo deste trabalho é atuar na terceira área com o intuito de estabelecer se uma pessoa ou um grupo de pessoas podem produzir certo tipo de linguagem, ou seja, para determinar a autoria ou não autoria por parte de um suspeito em um texto usado como evidência (cartas anônimas, notas de suicidios, ameaças).

2.4.4 Estilística Forense

A análise estilística forense busca estabelecer um dicionário de atributos estilométricos, como parâmetro estável de análise das variabilidades entre escritores distintos, tais como: a frequência de palavras incomuns; a média do tamanho das orações; o quociente de palavras diferentes em relação ao total; entre outros. Portanto, é possível afirmar que o conjunto de valores obtidos pela quantização de tais atributos definirá o estilo [MCM02].

O foco da estilistica forense é a identificação do autor em documentos cuja autoria seja questionada. Os primeiros estudos que utilizaram o conceito de estilística forense datam do século 19, onde estudiosos alemães desenvolveram métodos de identificação de autoria para fins de identificação de autoria de partes de textos bíblicos e de peças de Shakespeare.

Segundo [PAV07] na identificação de autoria existem basicamente dois modelos de análise: o modelo de identificação e o modelo de verificação.

- **Modelo de Identificação**

Com base no documento questionado é tentado identificar o autor em um conjunto de vários autores possíveis. Tal modelo de análise pode proporcionar a identificação direta do autor, porém depende do conhecimento de todos os autores de forma antecipada.

- **Modelo de Verificação**

No modelo de verificação, de posse de dois documentos de quaisquer autores, busca-se determinar se os documentos foram escritos pelo mesmo autor.

Diante dos modelos apresentados, o resultado da análise pode ser: (i) determinar a semelhança da escrita questionada; (ii) identificar um ou mais autores suspeitos; (iii) inconclusiva em virtude aos dados fornecidos para identificação ou eliminação.

Este trabalho tem como base a análise quantitativa de atributos estilometricos da língua portuguesa. Tem por finalidade obter conclusões e opiniões relacionadas tanto numa abordagem de identificação, como de verificação de autoria de um documento questionado dentro de um contexto jurídico. Na seção 2.4.5 abordar-se-á sobre atributos estilométricos.

2.4.5 Atríbutos Estilométricos

A atribuição de autoria pode ser vista com uma classificação, onde documentos de autoria conhecida são utilizadas como treinamento com o objetivo de identificar autores corretos de documentos questionados (de autoria desconhecida) baseado em modelos que foram gerados. O principal problema é não ter certeza de quais características devem ser utilizadas para fazer a classificação, ou seja, para se distinguir os autores. [PAV07]

Muitas pesquisas foram explanadas nos últimos anos, e nelas se percebe um consenso sobre quais conjuntos de características são as melhores para a atribuição de autoria. As caracteríticas estilométricas podem ser classificadas em 4 grupos (Conforme evidenciado na Figura 2.4) [ZQHC06]. A sequência destacada em negrito perfaz o caminho a ser percorrido por este trabalho (Características sintáticas utilizando palavras-funções).

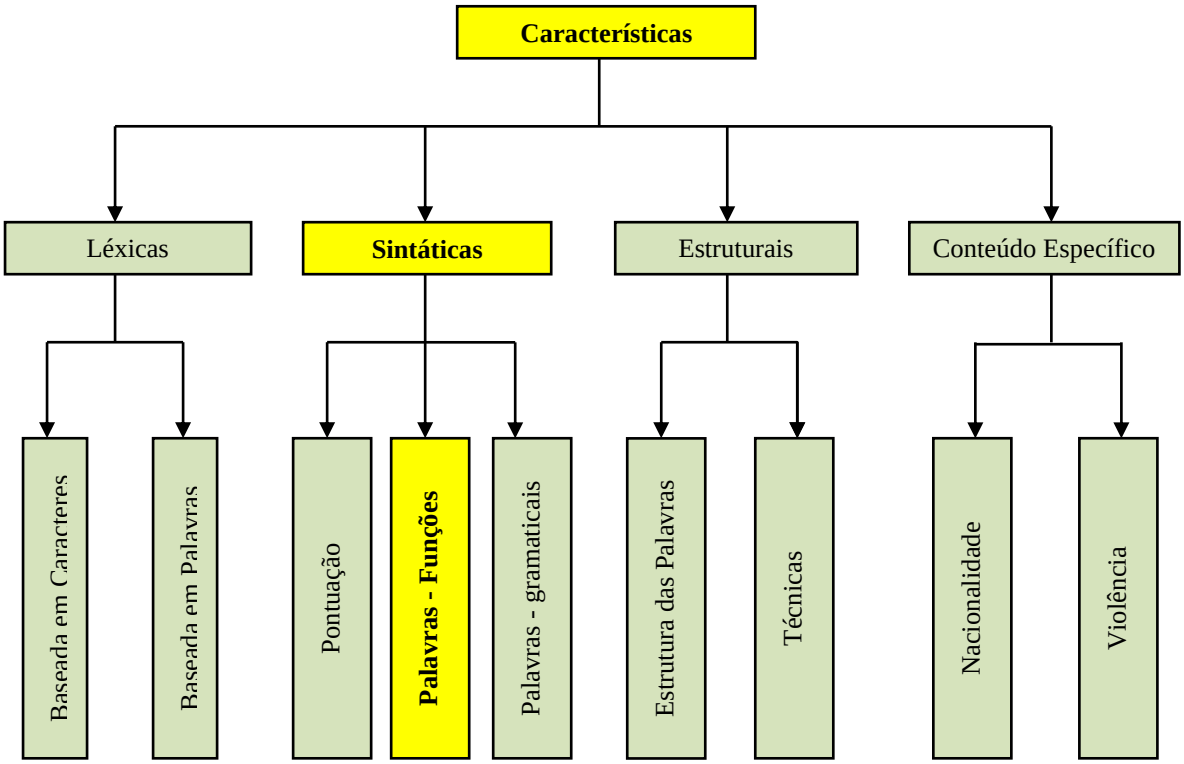

Figura 2.4: Grupos de Características de Estilo (Adaptado de [AC05])

- **Características Léxicas**

As características léxicas podem ser baseadas em caracteres ou baseadas em palavras. As características baseadas em palavras incluem o tamanho das palavras, a quantidade de palavras por frase, distribuição do tamanho das palavras, entre outras. Baseadas em caracteres, incluem o total de caracteres, caracteres por frase, caracteres por palavra e frequência das letras de forma isolada [ZQHC06].

- **Características Sintáticas**

Este grupo é formado por padrões responsáveis por formar as frases, tais como: pontuação, palavras função e palavras gramaticais. Palavras gramaticais e palavras função são palavras que indicam uma relação gramatical com outras palavras ou frases. Tais relações podem ser evidenciadas por verbos, conjunções, advérbios, pronomes, etc. Muitos trabalhos utilizaram palavras-funções como característica discriminatória e obtiveram bons resultados na criação e na identificação de um perfil de um autor [AC05] [PAV07].

Neste grupo de características (palavras-funções) enquadram-se as características utilizadas neste trabalho.

- **Características Estruturais**

O grupo de características estruturais está diretamente relacionado com a organização do texto e com a disposição das informações. Entre as características estruturais mais conhecidas, estão: a estrutura das palavras (que compreende algumas características, tais como: saudações iniciais e de encerramento, quantidade dos paragrafos, etc); e as características de ordem técnica (referentes a formatação – que podem evidenciar características importantes sobre o autor. Por exemplo, a formatação da fonte (tipos, tamanho, cor, alinhamento) [AC05] [PAV07].

- **Características de Conteúdo Específico**

As características de conteúdo específico se assemelham ao conteúdo das características léxicas, no entanto, possui um nivel de abstração e refinamento mais ampliado. Tais características são palavras relacionadas ao contexto do documento em questão. Um exemplo, pode ser que em um texto sobre qualidade de vida, utiliza-se palavras como: Nacionalidade, violência.

2.5 Classificação de textos em Classes

Quando as pessoas de comunicam de forma escrita encontra-se uma grande variação no estilo dos textos. Pode-se citar como exemplo colunas de jornais, que tem os mais variados assuntos, tais como: esporte, saúde, política e economia. Cada assunto tem formas diferentes que de certa forma são responsáveis pelas expectativas que se tem sobre o documento.

Neste trabalho os textos coletados de várias colunas de jornais brasileiros (ver mais detalhes no capítulo 4) estão classificados em 10 classes distintas entre si, com o objetivo de avaliar o impacto do tema na identificação/verificação de autoria.

2.6 Documentos Questionados e sua Aplicação no Âmbito Jurídico

Percebe-se atualmente um crescente aumento no ambiente jurídico da utilização de documentos questionados digitais, relacionado aos processos judiciais. Citam-se alguns exemplos de tais documentos: e-mails, diários eletrônicos, e-books, mensagens em meios digitais. No entanto, ainda é pequena a utilização desses documentos como prova.

Esta seção tem como objetivo descrever a prova judicial, bem como a utilização de provas digitais para aplicação em processos judiciais.

2.6.1 Conceito de Prova

As provas servem para o convencimento do juiz, e ao mesmo tempo tem o papel de justificar perante a sociedade a decisão adotada [CAL99].

Segundo [SFC04] a palavra prova próvem do latim *proba* de *probare*, que significa demonstrar, reconhecer, formar juízo de. No sentido jurídico é a demonstração que se faz pelos meios legais, da existência ou da veracidade de um ato material ou de um ato jurídico, em virtude da qual se conclui por sua existência ou se firma a certeza a respeito da existência do fato ou do ato demonstrado.

A verdade no processo deve ser sempre buscada pelo juiz, mas o legislador, embora busque pela verdade, não a coloca como um fim absoluto, em si mesmo. Ou seja, o que é suficiente, muitas vezes, para a validade e a eficácia da sentença é a verossimilhança dos fatos. [ALV97]

A pretensão de analisar a autoria de certo documento digital questionado, relaciona-se a um determinado autor com a intenção de convencer o juiz, no sentido que ele possa fazer a correta aplicação da lei. O juiz, em face do dever de solucionar o caso, utilizará as provas para formar seu convencimento, declarando o direito com a verdade encontrada (ainda que não seja a verdade real, que deve ser buscada), eis que as partes não podem restar à mercê do tempo, nem mesmo o judiciário pode omitir-se de decidir e solucionar o conflito.

Conforme o artigo 332 do Código de Processo Civil Brasileiro que todos os meios legais, bem como os moralmente legítimos, são hábeis a provar a verdade dos fatos. Assim, importante identificar a origem e conceito da prova, bem como sua finalidade, destinatário, objeto, salientando-se que os meios e tipos de prova não serão objeto de análise.

2.6.2 Procedimento de Prova

Segundo [PAV07] o procedimento de prova, ou procedimento probatório, é um espaço reservado a coleta de provas e é composto por três estágios, que são:

- **Proposição** - quando a autoria de um documento é questionada, o que requer a prova;
- **Deferimento** – é o ato em que o juiz declara a necessidade da prova;
- **Produção** – é o momento em que ocorre a efetivação para que a prova seja incorporada aos autos do processo judicial.

Muitas vezes em um processo judicial a autoria de um documento é questionada por uma parte envolvida no processo, então cabe a essa parte requer ao juiz a produção da prova de autoria ou não autoria do documento questionado. Como tal prova requer um conhecimento específico do assunto, um especialista da área de identificação de autoria é designado para proceder a prova pericial.

Existem inúmeros questionamentos sobre a possibilidade da utilização de documentos digitais como prova, devido a sua fragilidade, a possibilidade de alteração e na confiabilidade do processo de identificação da autoria.

Este trabalho visa utilizar o estilo literário para melhorar o grau de confiabilidade no processo de identificação de autoria em documentos digitais.

2.6.3 A Prova Pericial

A prova pericial é o meio de suprir a carência de conhecimentos técnicos de que se ressente o juiz para apuração dos fatos litigiosos, ou seja, o meio pelo qual no processo pessoas entendidas e sob compromisso verificam fatos interessantes à causa, transmitindo ao juiz o respectivo parecer.

De acordo com [SIL91] a perícia é feita por um perito oficial nomeado pelo juiz, que deve ter conhecimento técnico especializado sobre o assunto em questão. O perito pode ter como auxiliar o assistente técnico (conhecedor da área) que atua como auxiliar para concordar, criticar ou complementar o laudo do perito oficial.

Conforme os artigos 145/148 são direitos e deveres do perito:

- Deveres

(i) aceitar o encargo;

(ii) respeitar os prazos fixados pelo juiz para a realização da perícia;

(iii) comparecer à audiência, desde que intimado com cinco dias de antecedência;

(iv) dever de lealdade.

- Direitos

(i) escusar-se do encargo por motivo legítimo;

(ii) pedir prorrogação de prazos;

(iii) recorrer, requisitar e ter acesso às fontes de informação;

(iv) indenização pelas despesas relativas à perícia e honorários.

Segundo o Código de Processo Civil, no artigo 420 as espécies de pericias podem ser:

- **Exame**

É a inspeção por meio de perito sobre pessoas, coisas móveis ou animais para a verificação de fatos que interessam à causa.

- **Vistoria**

É a inspeção sobre bens imóveis, com os mesmos objetivos do exame.

- **Avaliação**

É a estimativa do valor, em moeda corrente, de coisas, direitos e obrigações segundo os conhecimentos técnicos do avaliador.

- **Arbitramento**

Quando se verifica o valor, quantidade ou qualidade do objeto do litígio, serviço, direito ou obrigação (espécie de avaliação, que para alguns possui autonomia).

Para a identificação de autoria de documentos digitais, o tipo de perícia utilizada é o exame que consiste na inspeção feita por um perito sobre pessoas ou coisas móveis, livros comerciais, documentos e papeis de um modo geral para a verificação de circunstâncias e fatos.

Conforme [PAV07], na análise do estilo literário se faz necessária certa quantidade de documentos do(s) autor(es) questionado(s), para fazer a análise do estilo. O perito deverá, nestes casos, solicitar ao juiz documentos que estejam em poder das partes ou em repartições públicas para que ele os requisite. Tal situação fica clara no artigo 429 do Código de Processo Civil:

> "O perito e os assistentes técnicos no desempenho de sua função, podem utilizar-se de todos os meios necessários, ouvindo testemunhas, obtendo informações, solicitando documentos que estejam em poder de parte ou em repartições públicas, bem como instruir o laudo com plantas, desenhos, fotografias e outras quaisquer peças".

Em suma, os peritos e os assistentes possuem livre acesso a recorrer a todas as informações que visem o esclarecimento dos quesitos apresentados em seu laudo.

Embora a prova técnica, científica, a perícia é uma prova como qualquer outra no que diz respeito à possibilidade de conter erros, imperfeições e até vícios que a tornem imprestável. Por isso ela está sujeita a esclarecimentos, que serão dados em audiência com a intimação do perito pelo juiz. [SIL91]

2.7 Support Vector Machine - SVM

O classificador utilizado para realização dos experimentos deste trabalho será o SVM, que foi desenvolvido por Vapnik[1] e é uma técnica de aprendizado estatístico. É utilizado neste trabalho por apresentar bons resultados no meio de identificação de autoria em textos. Também tem atraído a atenção de pesquisadores devido a sua boa capacidade de generalização e robustez diante de dados de grande dimensão.

Será apresentado nas próximas subseções um breve relato sobre o SVM duas classes e o SVM multiclasse que serão utilizados nesta pesquisa.

2.7.1 SVM Duas Classes

O classificador SVM duas classes baseia-se no principio da Minimização do Risco Estrutural (MRS), que tem por finalidade dois objetivos principais: (i) controlar o risco empírico do conjunto de treinamento; e, (ii) controlar a capacidade da função de decisão *f* usada para obtenção do valor de risco.[VAP98]

A função da decisão do SVM duas classes (linear) é dada por um vetor de peso $\vec{w}$, um bias $\vec{b}$, e um padrão de entrada $\vec{x}$ conforme mostra a equação 1 a seguir.

$$f\left(\vec{x}\right) = sign\left(\vec{w}.\vec{x} + b\right) \qquad (1)$$

Sendo um conjunto de vetores de treinamento S_l, que pertencem a duas classes separáveis $W_1(y_i = +1)$ e W_2 $(y_i = -1)$, o SVM tem por função encontrar o hiperplano de margem máxima (distância euclidiana máxima, que corresponde a maior distância de seus padrões no conjunto de treinamento – padrões estes que são chamados de vetores de suporte

[1] VAPNIK, V. **Statistical learning theory**. Wiley, N. Y., page pp. 768,1998

(SV)). Segundo o principio da MRS, somente existirá um hiperplano com margem máxima δ, que é definida como a soma das distâncias do hiperplano até o ponto mais próximo das classes. Com este limiar do classificador linear é possível obter a separação ótima do hiperplano através da equação 2, que segue.

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

A representação gráfica da classificação entre duas classes W_1 e W_2 usando hiperplanos pode ser denotada na Figura 2.5.

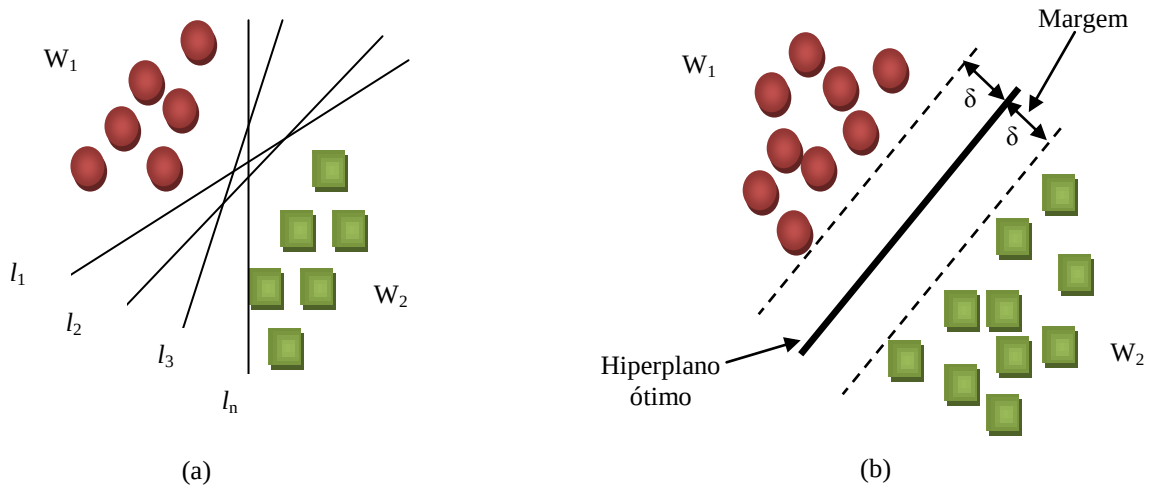

Figura 2.5: Representação de duas classes (W1 e W2) no hiperplano: (a) Hiperplanos arbitrários (li) e (b) hiperplano com separação ótima (máxima margem)

Para encontrar a superfície de decisão ótima, o algoritmo de treinamento o *SVM* tenta separar da melhor forma possível os pontos dos dados de ambas as classes. Os pontos mais próximos do limite entre as duas classes são selecionados, por serem mais importantes na solução, do que os pontos que estão mais distantes, os quais ajudam a definir a forma da melhor superfície de decisão que outros pontos.

Em um conjunto de treinamento não separáveis, o *i*-ésimo ponto de dados possui uma variável de folga ξ_i, que representa a magnitude do erro de classificação. Sendo que a função de penalidade $f'(\xi)$, representa a soma dos erros de classificação através da equação 3.

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

A solução do SVM pode ser encontrada através da minimização dos erros de treinamento de acordo com a seguinte equação (4) de minimização.

$$\min_{\overrightarrow{w},b,\overrightarrow{\xi}} = \frac{1}{2}\overrightarrow{w}.\overrightarrow{w} + C\sum_{i=1}^{n} \xi_i \qquad (4)$$

Sendo que na equação 4, C determina uma negociação entre o erro empírico e o termo de complexidade. O parâmetro C é escolhido livremente. No entanto, se um grande valor for atribuído a C, isto corresponde que existirá uma associação de uma penalidade mais alta para erros. [SJBS04]

2.7.2 SVM Multi Classes

O SVM *multiclass* utilizado nesta pesquisa é uma implementação do algoritmo multi-classe do SVM proposto por Crammer e Singer [CR01], onde um conjunto de treinamento (x_1, y_1)...(x_n, y_n) com $y_1 \in (1..k)$ encontra a solução para o problema de otimização durante o treinamento, conforme demonstrado na equação 5.

$$\begin{aligned} &\min \textstyle\sum_{i=1..k} w_i * w_i + C/n \sum_{i=1..n} \xi_i \\ &\quad s.t.\ for\ all\ y\ in\ [1..k] : [x_1 \bullet w_{yi}] >= [x_1 \bullet w_y] + 100 * \Delta(y_1, y) - \xi_1 \\ &\qquad \ldots \\ &\qquad for\ all\ y\ in\ [1..k] : [x_n \bullet w_{yn}] >= [x_n \bullet w_y] + 100 * \Delta(y_n, y) - \xi_n \end{aligned} \qquad (5)$$

Na equação 5, C é o parâmetro de regularização comum, que faz a negociação com o tamanho da margem e o erro do treinamento. Δ (y_1, y) é a função de perda que retorna 0 se y_n é igual a y e 1 caso contrário.

2.8 Algoritmos Genéticos

Algoritmos genéticos são técnicas computacionais que são aplicados em sua maior parte como mecanismo de busca e otimização de soluções em problemas complexos, inspirados na teoria da evolução natural de Darwin e na reprodução genética. [GP05]

Segundo [GON08] algoritmos genéticos são muito eficientes em busca de soluções ótimas e sub-ótimas que envolvam uma grande variedade de problemas, pois tal método não proporciona as limitações encontradas em métodos de busca tradicionais.

A menor unidade de um algoritmo genético é chamada gene. Um gene representa uma unidade de informação do domínio do problema. Um conjunto de genes forma um cromossoma, que representa uma possível solução para o problema. Neste caso um gene é uma característica e um cromossoma seria um conjunto de características.

A inicialização da população com valores aleatórios determina o processo de criação dos indivíduos para o primeiro ciclo do algoritmo. Em algoritmos genéticos vários parâmetros controlam o processo evolucionário, entre eles podem-se citar:

- Tamanho da população: é o espaço da busca;
- Taxa de Crossover: probabilidade de um indivíduo ser recombinado geneticamente com outro;
- Taxa de Mutação: probabilidade do conteúdo de cada gene do cromossoma ser alterado;
- Número de Gerações: indica a quantidade de ciclos do algoritmo genético;
- Total de Indivíduos: total de soluções a serem geradas e avaliadas pelo algoritmo genético.

O próximo passo é o cálculo da aptidão (*fitness*) de cada indivíduo da população, que é de suma importância na seleção de indivíduos usados para a reprodução, que dará origem a uma nova geração (*generation*). Verificou-se assim, que quanto maior a aptidão de um indivíduo, maior será a sua chance de ser selecionado para a reprodução. Por conseguinte a população é submetida a operações genéticas de cruzamento (*crossover*) e mutação (*mutation*).

Diferentes critérios de paradas podem ser utilizados para terminar a execução de um algoritmo genético, por exemplo: (i) após um determinado número de iterações (ciclos ou

gerações); (ii) quando a aptidão média ou do melhor individuo não melhorar mais; (iii) quando as aptidões dos indivíduos de uma população se tornarem muito parecidas. [GP05]

A estrutura geral de um algoritmo genético, pode ser representada a partir do seguinte pseudo-código:

Algoritmo Genético

```
T = 0;
Gerar População Inicial P(0);
Avaliar P(0);
Enquanto Critério de Parada não for satisfeito faça:
        T = T+1;
        Selecionar População P(T) a partir de P(T-1);
        Aplicar Operadores de Cruzamento sobre P (T);
        Aplicar Operadores de Mutação sobre P(T);
        Avaliar P(T);
Fim Enquanto
```

Já a representação da estrutura geral da função de avaliação pode ser descrita através do seguinte pseudo-código:

Função Avaliar P(T)

```
Para todo Indivíduo i da População Atual P(T) faça
        Avaliar o Indivíduo i, obtendo sua aptidão
Fim Para
```

Como neste trabalho o resultado final é obtido através da combinação do classificador SVM e de algoritmos genéticos, a seção 2.9 evidencia tal técnica.

2.9 Agrupamento de Classificadores

O agrupamento de classificadores é definido como a combinação de classificadores a fim de se obter melhores taxas de classificação. A principal função do uso de um agrupamento é que classificadores em conjunto apresentam resultados mais precisos que classificadores isolados. Subentende-se que classificadores distintos são aqueles, que diante

de um mesmo experimento, cada classificador cometa erros diferentes. Toda essa diversidade provoca que os agrupamentos apresentem uma maior precisão, se estes forem comparados aos classificadores de forma isolada. [HS90]

De acordo com [BER06] baseado nos estudos de [DIE00], o ganho de desempenho é notório. Um exemplo apresentado por [DIE00] mostra como a combinação de classificadores pode melhorar as taxas. No exemplo apresentado (Figura 2.6), existe um agrupamento de três classificadores distintos (C_1, C_2 e C_3). Se os três classificadores forem iguais, então quando $C_1(x)$ for incorreto, logo $C_2(x)$ e $C_3(x)$ também são incorretos. Porém, se os classificadores forem diferentes ou não correlatos, quando $C_1(x)$ for incorreto, C_2 e C_3 podem ser corretos. Então, pode-se utilizar o voto majoritário para classificar o exemplo *x* de forma correta.

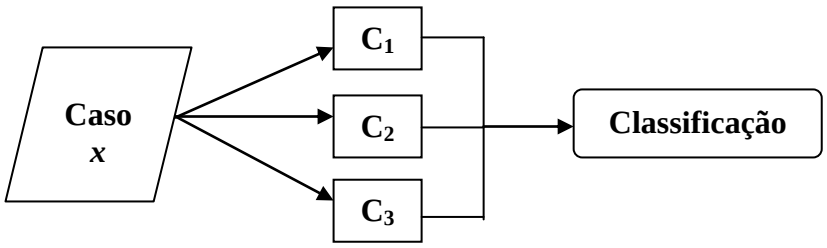

Figura 2.6: Exemplo de Combinação de Classificadores

O desenvolvimento de técnicas que combinam diversos classificadores com o objetivo de obter uma taxa de acerto cada vez melhor é uma área de pesquisa ativa e diversos estudos experimentais que avaliam a eficiência dessas técnicas têm sido produzidos. Uma condição necessária e suficiente para que um classificador formado pela combinação de diversos classificadores tenha melhor taxa de acerto que seus membros, é que os classificadores utilizados na combinação sejam diversos entre si e tenham uma taxa de acerto superior a 50%. [DIE00]

2.9.1Regras de Fusão

As regras de fusão utilizadas neste trabalho são métodos que independem dos dados, o que indica que não sofrem influencia durante a fase de aprendizagem. As regras são: Voto Majoritário, Máximo, Média e Mínimo.

Na regra do voto majoritário temos a contagem dos votos recebidos para determinada hipótese dos classificadores de forma individual. Assim, a classe com o maior número de votos é selecionada pelo consenso da maioria, conforme detalha a Equação 6.

$$\sum_{i=1}^{R} \Delta_{ji} = max_{k=1}^{m} \sum_{i=1}^{R} \Delta_{ki} \tag{6}$$

No máximo (Equação 7) selecionamos o resultado com base no valor mais elevado do conjunto de resultados e considera-se somente ele. Essa decisão se baseia no grau de certeza, quanto maior o valor mais certeza tem o SVM.

$$\begin{aligned} &atribuir\ Z \rightarrow w_j\ se \\ max_{i=1}^{R} P(w_j|x_i) &= max_{k=1}^{m} max_{i=1}^{R} P(w_k|x_i) \end{aligned} \tag{7}$$

A média escolhe o valor que mais se aproxima da média aritmética dos resultados, conforme mostra a equação 8.

$$\begin{aligned} &atribuir\ Z \rightarrow w_j\ se \\ \frac{1}{R} \sum_{i=1}^{R} P(w_j|x_i) &= max_{k=1}^{m} \frac{1}{R} P(w_k|x_i) \end{aligned} \tag{8}$$

O mínimo seria considerar o resultado pelo menor valor obtido (Equação 9).

$$\begin{aligned} &atribuir\ Z \rightarrow w_j\ se \\ min_{i=1}^{R} P(w_j|x_i) &= max_{k=1}^{m} min_{i=1}^{R} P(w_k|x_i) \end{aligned} \tag{9}$$

2.10 Considerações Finais

Neste capítulo foram apresentados os principais conceitos teóricos para a elaboração e compreensão deste trabalho. É evidenciada a importância da utilização de características de língua portuguesa para a identificação de autoria de documentos questionados em processos judiciais, que pode auxiliar linguistas e juristas na análise e na toma de decisão.

No próximo capítulo, são descritos os principais trabalhos já publicados na área de identificação de autoria de textos.

Capítulo 3

Estado da Arte

Neste capítulo são descritos os principais trabalhos já publicados na área, evidenciando os resultados já obtidos sobre identificação de autoria em textos. Ao final deste capítulo, também é apresentada às considerações finais sobre o estado da arte.

3.1 Cronologia Histórica

Os primeiros estudos sobre atribuição de autoria datam do final do século XX. A seguir são apresentadas algumas das principais contribuições efetuadas pelos pesquisadores até os tempos atuais:

Mendenhall em 1887 estudou a autoria de Bacon, por Marlowe e Shakespeare através do espectro das palavras e das curvas características que eram representações gráficas da organização do comprimento do termo e da sua frequência. Ele concluiu que a curva que se mantivesse constante conforme a curva característica do autor seria um bom método para a discriminação da autoria. [WIL75]

Já Zipf em 1932, centrou seu trabalho em frequências de palavras diferentes em um documento do autor. Ele determinou que houvesse uma relação logarítmica entre o número de palavras que aparecem exatamente r vezes em um texto e r em si. Esta expressão ficou conhecida como a Lei de Zipf. [ZIP75]

Yule em 1938 utilizou inicialmente o comprimento de frase para diferenciação de autores, mas constatou que este método não era completamente confiável. Então Yule criou uma medida baseada no método de Zipf, que era baseado na frequência das palavras. Ele

descobriu que o uso de uma palavra é probabilístico e pode ser aproximado com a distribuição de *Poisson*.[YUL38]

Williams em 1940 identificou que o *log* do número de palavras por frase dos trabalhos de Chesterston, Wells e Shaw ocorria de acordo com uma distribuição normal. [WIL40]

No ano de 1963, Monsteler e Wallace utilizaram o teorema de Bayes pela primeira vez nos problemas, ao invés de abordagens clássicas. [MW64]

Em 1967, Särndal utilizou a distribuição quantitativa de palavras para a determinação da probabilidade de erros de falsa aceitação e falsa rejeição. Baseado nestas características Särndal criou várias hipóteses arbitrárias. [SÄR67]

Holmes em 1985 efetuou uma revisão na análise do estilo literário, identificando possíveis fontes de características e técnicas de atribuição de autoria. Entre as características, citam-se: média de sílabas por palavras, tamanho de frase, frequência de palavras, riqueza de vocabulário, frequência e distribuição do tamanho das palavras, distribuição da frequência das palavras. [HOL85]

Thisted e Efron no ano de 1987 usaram conceitos de riqueza de vocabulários para a determinação de o autor de um novo poema questionado ser Shakespeare. [TE87]

Em 1996, Merriam usou palavras com comportamento gramatical para a comparação dos estilos de William Shakespeare e Christopher Marlowe. [MER96]

Ainda no ano de 1996, Foster estudou o poema "Uma Elegia Fúnebre", atribuindo a autoria a Shakespeare. Foram utilizados na pesquisa como referências os trabalhos canônicos de Shakespeare, onde foi comparado o estilo, os acidentes gramaticais do texto, a sintaxe e uso de palavras raras [FOS96A]. Foster também publicou neste mesmo ano um estudo sobre uma sátira ao então presidente dos Estados Unidos, Bill Clinton, que foi publicada sem identificação de autoria. Nesta pesquisa Foster fez a análise e a atribuição de autoria a este chamado "Primary Colors" [FOS96B].

Em 2001 Carole E. Chaski apresentou resultados empíricos divididos em três grupos de características, que foram: (1) Pontuação e estrutura da frase; (2) vocabulário, análise do conteúdo e complexidade frasal; e (3) características relacionadas a erros (por exemplo, erros gramaticais e erros de pontuação. Para comparação das medidas de um autor com outros Chaski utilizou o método estatístico X^2.[CHA01]

Yuta Tsuboi e Yuji Matsumoto no ano de 2002 fizeram um estudo sobre a identificação de autoria em japonês, através do SVM. A aplicação do trabalho foi em

identificar autores de documentos publicados em páginas da internet, a fim de investigar um grupo mais heterogêneo de documentos. Utilizaram alguns marcadores de estilo, tais como: comprimento das palavras e frases, riqueza do vocabulário e palavras reservadas de cada autor. Obtiveram resultados satisfatórios utilizando *n*-gramas e frequência de padrões sequenciais extraídos através de uma mineração técnica, chegando a sustentar resultados entre 66% e 80% de similaridades de autores. [YTYM02]

Em 2002 Smith e Kelly obtiveram resultados que demonstravam que o estilo de um autor pode variar cronologicamente com o passar dos anos. Para isso foram utilizadas características léxicas e de vocabulário, tais como: riqueza de vocabulário e frequência das palavras. A base de testes foram textos de Eurípedes, Aristophanes e Terence [JSCK02]

Em 2003, Malcolm W. Corney usou em sua dissertação de mestrado, características estilométricas para a construção de uma ferramenta para a identificação de autoria em textos de e-mails. No trabalho, foram utilizadas características léxicas (frequência de caracteres, palavras e palavras funções) e estruturais (formatação) através do classificador SVM (*Support Vetor Machine*) e os resultados atingidos foram de 85% de acerto na identificação de autoria de e-mails. [COR03]

Diederich *et. al*. em 2003, usaram o SVM para trabalhar com vetores de grande capacidade para classificar textos e palavras de jornais da Alemanha de 150 autores diferentes. Obtiveram resultados perto de 80% de acertos, utilizando a frequência de palavras. Em um segundo experimento quando foram ignorados substantivos, adjetivos e verbos os resultados foram menores (proporção de 60% de acertos). Com isso Diederich *et. al.* conseguiram descrever uma análise comparativa de seu método com os outros métodos já aplicados no mesmo problema. Evidenciaram também a importância de certas características na identificação de autoria. [DKLP03]

No ano de 2004 Gamon utilizou o SVM para identificação de autoria, analisando amostras de textos de 3 irmãs (Irmãs Brontë). Analisou características sintáticas e semânticas através da combinação de características, e atingiu resultados em torno de 85% de acerto. [GAM04]

Ainda em 2004 Van Halteren, testou características léxicas e sintáticas em separado e depois combinando-as. A base de texto utilizada para os testes é em língua holandesa (ABC-NL1). Atingiu através da combinação de características em torno de 97% de identificação. [VHA04]

Em 2005, Uzaner e Katz, fizeram o teste com dois modelos de classificação: de reconhecimento e de atribuição de autoria através da avaliação da expressão sintática inicial e final de cada estrutura da frase na categoria de verbos e nas medidas linguísticas (palavras-função, elementos sintáticos, tamanho das palavras e das frases), que foi baseado em um corpus de livros de romance. Com este experimento foi possível identificar 76% das obras literárias e 66% de elementos das expressões. [BKOU05]

Em 2005, Coutinho *et. al*. usou o algoritmo de compressão PPM-C para classificação de textos de 10 autores da literatura brasileira. Os textos escolhidos tinham entre 15 kb e 120 kb e foram testados usando ordens de Markov 4, 5 e 6. Os resultados encontrados foram bastante satisfatórios, pois atingiram uma média geral de 78% de reconhecimento dos autores dos textos.

Morales *et. al.* em 2006 utilizou recursos estilométricos em seu trabalho para a identificação de autores, que foi baseado em um conjunto de palavras sequenciais que combinavam com palavras funções (substantivos, verbos e adjetivos). Seu estudo foi delineado para trabalhar com documentos curtos e chegou a um patamar médio de acerto entre 60% e 80%. [MPGR06]

Em 2006 Malyutov, fez uma revisão sobre as diversas abordagens teóricas de atribuição de autoria em textos, e cita que apesar do estudo ser pioneiro e datado desde o fim do século XX, ainda há muito a se fazer na área da estilometria, tanto na teoria como nos estudos de caso, e que através destes estudos podem-se chegar a resultados mais concretos e corretos na atribuição. [MAL06]

Tufan Tas e Abdul Kadir Gorur em 2007 utilizaram 35 marcadores de estilo automáticos para definição de um grupo de autores e depois cada texto questionado era submetido a estes estilos para a identificação do autor. A base utilizada foram 20 textos diferentes para cada um dos 20 autores de diferentes jornais turcos. Com a utilização deste método, alcançou-se uma taxa de 80% de sucesso, através do classificador Naive Bayes Multimonial. [TTAKG07]

Grieve no ano de 2007 fez uma abordagem sobre 39 diferentes tipos de medições textuais (análise quantitativa), que são comunentes utilizadas em estudos de atribuição de autoria. Concluiu que os melhores resultados foram atingidos pela combinação dos melhores algoritmos (que atingiram no mínimo 75% de acerto), em uma base de dados com um elevado número de amostras de textos. [GRI07]

Em 2007, Pavelec utilizou a classificação sintática, através de 171 palavras-funções (conjunções e advérbios) para a identificação de autoria em documentos de língua portuguesa através do SVM. A base de dados utilizada foram colunas de 30 jornalistas brasileiros de diferentes jornais dos estados do Paraná e São Paulo. Os resultados médios atingidos através do experimento foi 83% de identificação de autoria nos textos questionados. [PJ007]

No ano de 2008, Stamatos apresentou um levantamento dos avanços da investigação cientifica nos últimos anos, enfatizando a identificação de autoria através de recursos computacionais. Apresentou as distintas metodologias utilizadas no processo e analisou seus pontos fortes e fracos. Identificou que é crucial para os métodos de atribuição, o mesmo ser: robusto, com textos curtos e quantidade limitada, e ser equilibrado em relação ao assunto.[STA08]

A Tabela 3.1 mostra um resumo dos principais trabalhos realizados na área de identificação de autoria nos últimos anos. Apresenta de forma resumida os autores, o ano de publicação, o grupo de característica de estilo conforme a Figura 2.6 (Ver Capítulo 2), o classificador (algoritmo) utilizado para os experimentos, a base de textos na qual foi efetuados os testes e os resultados alcançados em cada trabalho.

Tabela 3.1: Resumo dos Principais Trabalhos sobre Identificação de Autoria

Autor(es) / Ano	Características	Classificador	Bases / Textos	Resultados
Tsuboi e Matsumoto (2002)	Léxicas	SVM	Páginas Web em Japonês	66-80%
Malcom (2003)	Léxicas e Estruturais	SVM	E-mails	85%
Diederich *et al* (2003)	Sintáticas	SVM	Jornais Alemães	80%
	Frequência de Palavras			60%
Gamon (2004)	Sintáticas	SVM	Textos de 3 irmãs	75%
Uzaner e Katz (2005)	Sintáticas	-	Livros de Romance	66-76%
Coutinho *et al* (2005)	-	PPM-C	Literatura Brasileira	78%
Morales (2006)	Sintáticas	Naive Bayes	Poemas Mexicanos	60-80%
Tas e Gorur (2007)	Léxicas	Naive Bayes Multimonial	Jornais Turcos	80%
Pavelec (2007)	Sintáticas	SVM	Jornais Brasileiros	72-83%

Analisando a Tabela 3.1, percebe-se que os trabalhos sobre identificação de autoria utilizando recursos computacionais teve uma maior contribuição a partir do ano de 2002. A utilização de características léxicas e sintáticas trouxe bons resultados, independente da base de dados e do classificador. O classificador SVM foi o mais utilizado pelos pesquisadores e obteve resultados promissores na classificação de textos.

O trabalho de Pavelec [PAV07] foi pioneiro na identificação de autoria com características sintáticas utilizando o SVM, e atingiu bons resultados. Este trabalho propõe um novo grupo de características sintáticas para testar em uma nova base de dados, para a verificação da contribuição de cada grupo e no conjunto, utilizando para isso o SVM.

Na seção 3.2 são apresentadas as divisões na área de identificação de autoria.

3.2 Identificação de Autoria

No campo da literatura que se refere à identificação de autoria, o mesmo é dividido em três áreas, que são: atribuição de autoria, identificação de plágio e caracterização de autoria [COR03]. Sendo o objetivo desta pesquisa a identificação de autoria, três tipos de evidências podem ser utilizadas: externas, linguísticas e interpretativas (Figura 3.1). A evidência linguística é o foco deste trabalho, pois está focada nas palavras e padrões de palavras utilizadas em um documento. As evidências externas, por exemplo, podem ser relacionadas a traços de manuscritos. Já as evidências interpretativas estão relacionadas com o que o autor pretendia passar quando escreveu o documento. [CRA98]

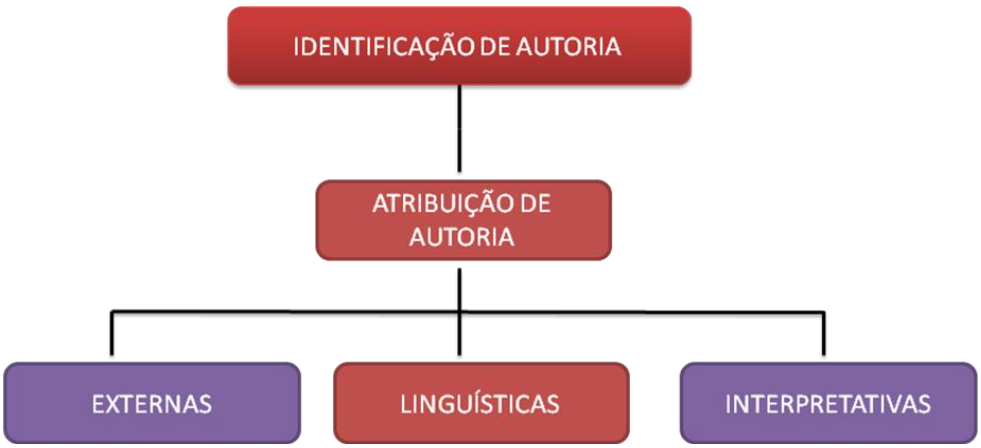

Figura 3.1: Divisão da área de Identificação de Autoria

Na seção 3.3 são apresentados alguns dos principais trabalhos que utilizaram as técnicas de aprendizado de máquina para identificação de autoria em textos.

3.3 Aprendizado de Máquina

O aprendizado de máquina na área da estilometria possui várias pesquisas, entre as mais importantes pode-se destacar: a utilização de classificadores de redes neurais para a comparação de textos de pensadores, tais como, Shakespeare, Marlowe e Fletcher aplicada por Matthews e Merriam [MM94];

O uso de rede neural com os classificadores *Naive* e *Nearest Neighbour* para a extração de *n-grams*[2] se mostrou eficaz na discriminação entre dois autores que escrevam em um estilo semelhante. Nos testes efetuados por Kjell, esta combinação de classificadores foi a mais eficaz nos resultados. [KJE94];

O uso da rede neural *Radial Basis Finction* (RBF) através de palavras-função para classificar textos de poetas holandeses (Bloem, Slauerhoff e Lucebert). Os resultados na comparação entre dois poetas foram satisfatórios, atingido um acerto médio entre 80% e 90%; Já quando os testes foram submetidos a três poetas o rendimento médio foi de 70% de identificação do poeta. [HFKV99];

Nos últimos anos ainda muitos trabalhos foram publicados com aprendizado de máquina, e cabe ressaltar que na classificação de textos, o SVM tem obtido bons resultados [DKLP03] [COR03] [PAV07].

A seguir são apresentadas as considerações finais do capítulo.

3.4 Considerações Finais

Muito se tem produzido sobre a identificação de autoria, porém não é identificado um consenso entre os pesquisadores. Constatou-se que ainda não existe um conjunto de características comuns que determinam o estilo de cada autor; por isso várias abordagens são utilizadas, reestruturadas e reutilizadas nas atuais pesquisas, em busca de metodologias e

[2] *n-gram* é a seqüência de letras de uma parte de um texto com n caracteres.

técnicas que tenham uma maior eficiência na identificação de autoria [HOL85] [MAL06] [TTAKG07] [PAV07].

O desenvolvimento tecnológico, em conjunto com a estruturação e a formação de novos pesquisadores nesta área, está levando a linha pesquisa a conquistar um espaço considerável no meio da pesquisa computacional.

No próximo capítulo é apresentado o método proposto que foi adotado para a identificação de autoria de textos.

Capítulo 4

Método Proposto

Neste capítulo são apresentadas as fases dos métodos de atribuição de autoria de documentos digitais, através da análise do estilo do autor. São abordados os seguintes assuntos: o método de identificação de autoria, a formação da base de dados, a apresentação das características utilizadas, o processo de geração dos vetores de dissimilaridade, as duas abordagens de modelos de comparação (dependente e independente de autor), o processo de classificação e a decisão final.

4.1 Método de Identificação de Autoria

O método proposto para a identificação de autoria é baseado no método apresentado por [PAV07], o qual contém as seguintes etapas:

1. Coleta e formação da base de dados;
2. Extração das características;
3. Geração dos vetores de dissimilaridade
4. Classificação – produção de modelos
5. Decisão Final

Na Figura 4.1 é possível verificar o processo esquematizado para identificação de autoria.

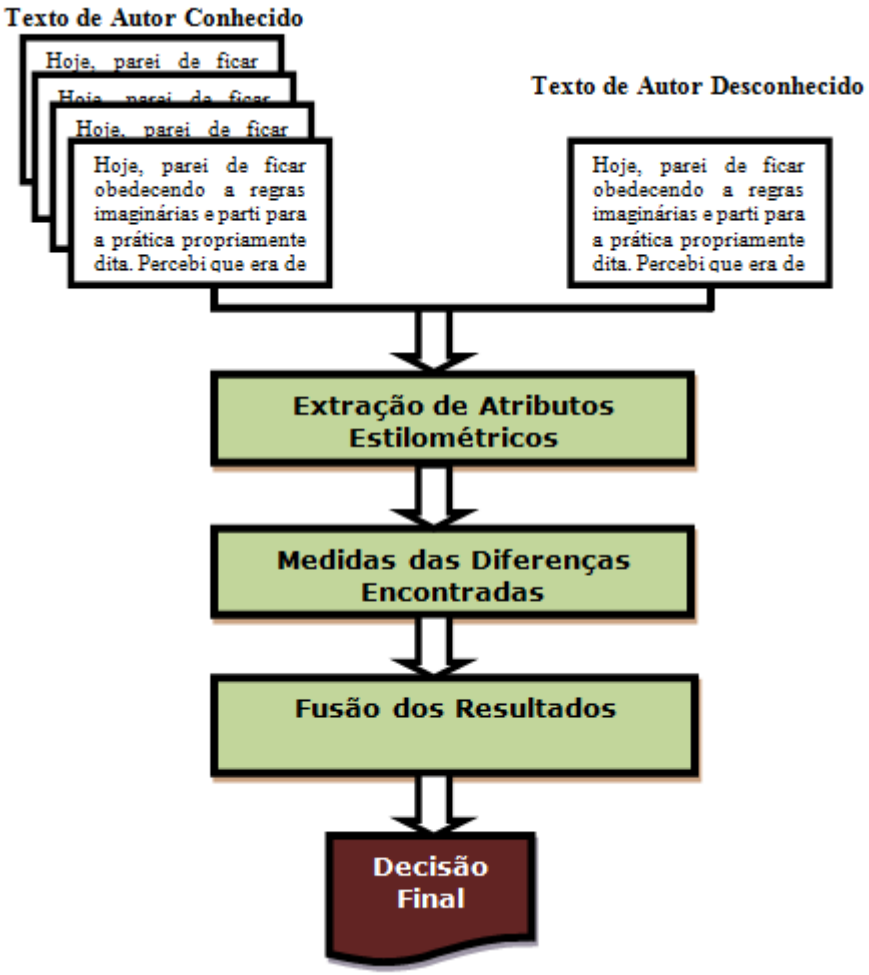

Figura 4.1: Diagrama Esquemático das Etapas Estilométricas (Adaptado de [PAV07])

Nas seções 4.2, 4.3, 4.4, 4.5, 4.6 e 4.7 são descritas as etapas do método proposto, abordando as principais características do método.

4.2 Coleta e Formação da Base de Dados

Muitas vezes, os documentos encontrados em casos que envolvem a identificação de autoria, os textos são frequentemente pequenos, se tornando de difícil análise. Linguistas forenses, diante de um documento com poucas informações, apresentam dados empíricos e não dados estatísticos (que poderiam elevar o grau de certeza na análise para identificação de autoria) [MCM02].

Para os experimentos realizados neste trabalho foram escolhidos textos pequenos, para que a pesquisa seja a mais próxima possível da realidade encontrada. Tais textos possuem entre 1KB (Kilobytes) e 9KB com no máximo 1200 *tokens*[3] e em média 378 *tokens* por texto.

Para avaliar o método proposto foram escolhidas colunas de 100 jornalistas e colunistas brasileiros de diferentes jornais. Como e-mails, cartas de sequestro, cartas de ameaça, notas de suicídios entre outros, possuem pouco conteúdo, isto é, textos pequenos. As colunas de jornais se mostram uma opção viável e que satisfazem os quesitos de tamanho e quantidade de textos desejados. Por essa razão, adotou-se as colunas de jornais como elementos das bases. Outro fator importante está no fato de se conseguir um número expressivo de amostras por autor e a possibilidade de separá-los em de temas (Apêndice A).

As colunas dos jornalistas foram obtidas através da internet, entre os principais jornais e blogs do País:

- O Povo
- Zero Hora
- Diário do Grande ABC
- A Gazeta do Povo
- A Notícia
- Jornal de Brasília
- O Extra
- O Estado do Paraná
- Paraná On-Line
- O Tempo
- Jornal de Beltrão
- O Gerente
- Folha UOL Online
- A Gazeta do Acre
- Colunistas IG

[3] Número de palavras válidas para o extrator de características

Mais informações a respeito dos jornais e colunas podem ser encontrados no Apêndice A.

A Figura 4.2 mostra um exemplo de um texto eletrônico, coletado na internet.

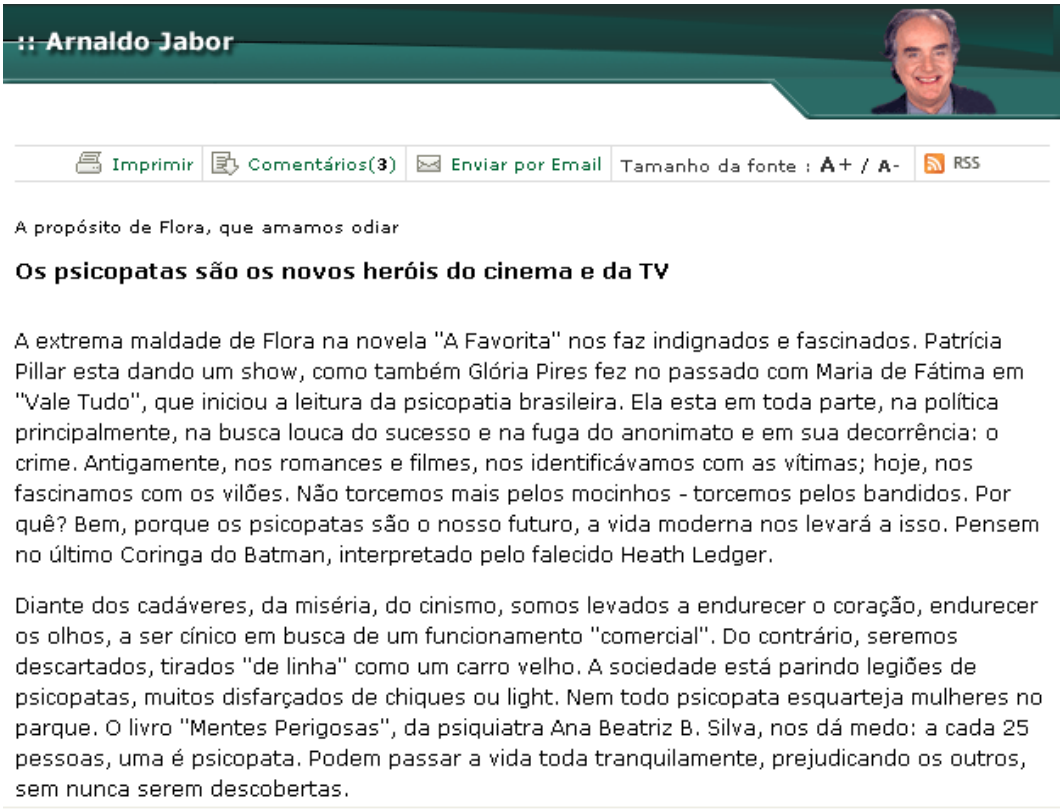

Figura 4.2: Exemplo de Coluna Eletrônica de um Jornal

A Figura 4.3 mostra um exemplo das colunas escolhidas com as informações do autor, do jornal, da classe de assuntos, o número do texto, o título da coluna, data que a coluna foi publicada, o tamanho da coluna em KB, a quantidade de *tokens* por texto, quantidade de *hapax legomena*[4] e o nível[5]. Cada autor da base de dados possui uma amostra de 30 textos.

[4] Quantidade de *tokens* que não se repetem.
[5] Quantidade de *hapax legomena* dividido pela quantidade total de *tokens*.

AUTOR			JORNAL			CLASSE
AUGUSTO MAFUZ			O ESTADO DO PARANÁ			ESPORTES
Arquivo	Texto	Data	Tamanho em Kb	Tokens	Hapax	Nível
1	Sociais	17/6/2009	2,04	352	171	0,485795455
2	Espiritos	16/6/2009	2,20	376	180	0,478723404
3	Tempo de jogo	15/6/2009	1,98	354	162	0,457627119
4	Contradições	13/6/2009	1,97	352	141	0,400568182
5	Tratamento vip	12/6/2009	1,82	300	148	0,493333333
6	Começo	10/6/2009	2,37	402	188	0,467661692
7	Ideal	9/6/2009	1,55	261	133	0,509578544
8	Gratidão	8/6/2009	2,18	383	161	0,420365535
9	Vitórias Inúteis	5/6/2009	1,81	320	137	0,428125000
10	Imponderável	3/6/2009	1,52	270	117	0,433333333
11	Lembranças	2/6/2009	2,14	370	181	0,489189189
12	De volta para o futuro	1/6/2009	1,86	326	154	0,472392638
13	Contraste	30/5/2009	1,98	339	172	0,507374631
14	Improvável	27/5/2009	1,70	296	139	0,469594595
15	Apaixonados	26/5/2009	2,85	497	211	0,424547284
16	Vergonha	25/5/2009	2,42	424	174	0,410377358
17	Imagem	23/5/2009	1,76	311	149	0,479099678

Figura 4.3: Colunas do Autor Augusto Mafuz

As colunas selecionadas foram arquivadas em formato texto com acentuação e sem hifenização (Figura 4.4). Como podem ser observados na Figura 4.3 os arquivos gerados pelas colunas são pequenos. Os autores das colunas possuem perfil profissional variado e escrevem sobre determinados assuntos, sendo que os textos foram classificados em 10 classes distintas de assunto:

- Esportes;
- Política;
- Saúde;
- Economia;
- Direito;
- Turismo;
- Tecnologia;
- Gastronomia;
- Literatura; e
- Assuntos Variados.

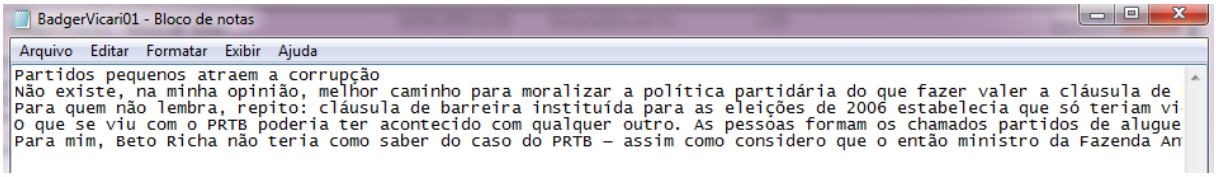

Figura 4.4: Exemplo de Armazenamento do Texto das Colunas dos Jornais

Na Tabela 4.1 é possível verificar uma amostra dos autores por classe (A relação dos autores por classe se encontra no Apêndice B deste trabalho).

Tabela 4.1: Autores da Classe Esportes

AUTOR	JORNAL
ANDRÉ RIBEIRO	DIÁRIO DO GRANDE ABC
AUGUSTO MAFUZ	O ESTADO DO PARANÁ
DIOGO OLIVIER	ZERO HORA
MARCELO SENNA	O EXTRA
MÁRCIO BERNARDES	DIÁRIO DO GRANDE ABC
SÉRGIO REDES	O POVO
TOSTÃO	A GAZETA DO POVO
VALDIR BICUDO	PARANÁ ONLINE
VICENTE DATOLLI	JORNAL DE BRASÍLIA
WIANEY CARLET	A NOTÍCIA

Na seção 4.3 serão apresentadas as características estilométricas utilizadas neste trabalho e como foi o processo de extração de tais características dos textos.

4.3 Extração de Características

A extração de características é uma das etapas mais importantes deste trabalho, pois nesta fase são escolhidas as características que poderão obter bons resultados na identificação de autoria. As características utilizadas para a identificação da autoria de textos questionados serão 150 verbos conjugados no infinitivo, particípio e gerúndio e 87 pronomes em conjunto com as 171 palavras-funções (advérbios e conjunções) já testadas por [PAV07].

Os pronomes são palavras utilizadas que representam os nomes dos seres, determinando e indicando a pessoa do discurso. Já o verbo é uma palavra essencial para exprimir uma idéia ou apresentar um enunciado.

Na identificação de autoria, o estilo literário é um conjunto de elementos que personaliza a escrita de um autor, neste caso representado por verbos e pronomes, além das conjunções e advérbios já utilizados por [PAV07].

Os pronomes a serem utilizados são especificados nas Tabelas 4.2 (Pronomes Relativos), 4.3 (Pronomes Possessivos), 4.4 (Pronomes Demonstrativos), 4.5 (Pronomes Pessoais) e 4.6 (Pronomes de Tratamento) respectivamente.

Tabela 4.2: Pronomes Relativos

quem	o qual	a qual	os quais	as quais
onde	em que	quanto	quanta	quantos
quantas	cujo	cuja	cujos	cujas

Tabela 4.3: Pronomes Possessivos

meu	minha	meus	minhas	teu
tua	teus	tuas	seu	sua
seus	suas	nosso	nossa	nossos
vosso	vossa	vossos	vossas	

Tabela 4.4: Pronomes Demonstrativos

este	esta	estes	estas	isto
esse	esses	essa	essas	isso
aquele	aquela	aqueles	aquelas	aquilo
nessa	desta	daquela		

Tabela 4.5: Pronomes Pessoais

eu	tu	ele	nós	vós
eles	me	te	se	lhe
o	a	nos	vos	lhes
os	as	mim	comigo	conosco
ti	contigo	convosco	si	consigo

Tabela 4.6: Pronomes de Tratamento

você	vocês	senhor	senhores	senhora
senhoras	senhorita	senhoritas	vossa senhoria	vossas senhorias

Os verbos utilizados foram conjugados de forma nominal: infinitivo, gerúndio e particípio, e são especificados na Tabela 4.7. Esta lista de verbos foi escolhida por serem alguns dos verbos mais comumente utilizados em textos escritos na língua portuguesa [RYA,06].

Tabela 4.7: Verbos

Infinitivo	**Gerúndio**	**Particípio**
Escrever	Escrevendo	Escrito

Falar	Falando	Falado
Jogar	Jogando	Jogado
Andar	Andando	Andado
Ver	Vendo	Visto
Ser	Sendo	Sido
Cantar	Cantando	Cantado
Pular	Pulando	Pulado
Ler	Lendo	Lido
Ter	Tendo	Tido
Achar	Achan	Achado
Colar	Colando	Colado
Estar	Estando	Estado
Dizer	Dizendo	Dito
Dar	Dando	Dado
Escolher	Escolhendo	Escolhido
Fechar	Fechando	Fechado
Entender	Entendendo	Entendido
Fazer	Fazendo	Feito
Trocar	Trocando	Trocado
Abrir	Abrindo	Aberto
Acabar	Acabando	Acabado
Declarar	Declarando	Declarado
Completar	Completando	Completado
Visitar	Visitando	Visitado
Encerrar	Encerrando	Encerrado
Comer	Comendo	Comido
Beber	Bebendo	Bebido
Pensar	Pensando	Pensado
Possuir	Possuindo	Possuído
Efetuar	Efetuando	Efetuado
Atingir	Atingindo	Atingido
Melhorar	Melhorando	Melhorado
Achar	Achando	Achado
Realizar	Realizando	Realizado
Haver	Havendo	Havido
Viver	Vivendo	Vivido
Aplicar	Aplicando	Aplicado
Gerar	Gerando	Gerado
Melhorar	Melhorando	Melhorado
Pagar	Pagando	Pagado
Distribuir	Distribuindo	Distribuído

Ligar	Ligando	Ligado
Usar	Usando	Usado
Projetar	Projetando	Projetado
Desenvolver	Desenvolvendo	Desenvolvido
Poder	Podendo	Podido
Implantar	Implantando	Implantado
Trazer	Trazendo	Trazido
Iniciar	Iniciando	Iniciado

As características utilizadas por [PAV07] e testadas neste trabalho, são 77 conjunções e 94 advérbios conforme especificam as tabelas 4.8 (Conjunções) e 4.9 (Advérbios).

Tabela 4.8 : Conjunções

Grupo	**Palavras-Funções**
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, 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

Tabela 4.9: Advérbios

Grupo	**Palavras-Funções**
Lugar	aqui, ali, aí, cá, lá, acolá, além, longe, perto, dentro, adiante, defronte, onde, acima, abaixo, atrás, em cima, de cima, ao lado, de fora, por fora
Tempo	hoje, ontem, amanhã, atualmente, sempre, nunca, jamais, cedo, tarde, antes, depois, já, agora, então, de repente, hoje em dia
Afirmação	certamente, com certeza, de certo, realmente, seguramente, sem dúvida, sim
Dúvida	porventura, provavelmente, talvez
Intensidade	ainda, apenas, de pouco, demais, mais, menos, muito, pouca, pouco, quase, tanta, tanto
Negação	absolutamente, de jeito nenhum, de modo algum, não, tampouco
Quantidade	todo, toda
Modo	assim, depressa, bem, devagar, face a face, facilmente, frente a frente, lentamente, mal, rapidamente, algo, alguém, algum, alguma, bastante, cada, certa, certo, muita, nada, nenhum, nenhuma, ninguém, outra, outrem, outro, quaisquer, qualquer, tudo.

A seleção das palavras-função (pronomes, verbos, conjunções e advérbios) utilizadas como características, propostas para a identificação de autoria neste trabalho, se deve ao fato de as mesmas poderem permitir:

- Identificar traços inconscientes do autor (não identificáveis a primeira vista);
- Tais características já foram testadas e utilizadas para identificação de autoria em outros idiomas e obtiveram bons resultados;
- A grande gama de possibilidades que a língua portuguesa oferece, por sua enorme quantidade de elementos linguísticos que podem ser discriminantes da identificação de autoria;

Seguindo o mesmo processo apresentado por [PAV07], o processo de extração de características obedece às seguintes regras:

- Não houve diferenciações entre letras maiúsculas e minúsculas;
- Espaços em branco e de finais de linha não foram considerados *tokens* válidos;
- Palavras hifenizadas, mesóclises, próclises e ênclises foram consideradas palavras únicas;
- Utilização de algoritmo de busca de características.

Na seção 4.4, são evidenciados os processos de medidas das diferenças encontradas entre os textos, e como foram gerados os vetores de dissimilaridade.

4.4 Vetores de Dissimilaridade

De posse das características extraídas todos os documentos são distribuídos em vetores, que representam o conjunto de características (fv), a identificação do documento (Q) e o número total de características (f_L), conforme a equação 6.

$$fv_Q = (f_1, f_2, \ldots, f_L) \tag{6}$$

O vetor de dissimilaridade é composto pelo módulo da diferença entre as características extraídas das colunas de acordo com os protocolos de aprendizado e testes.

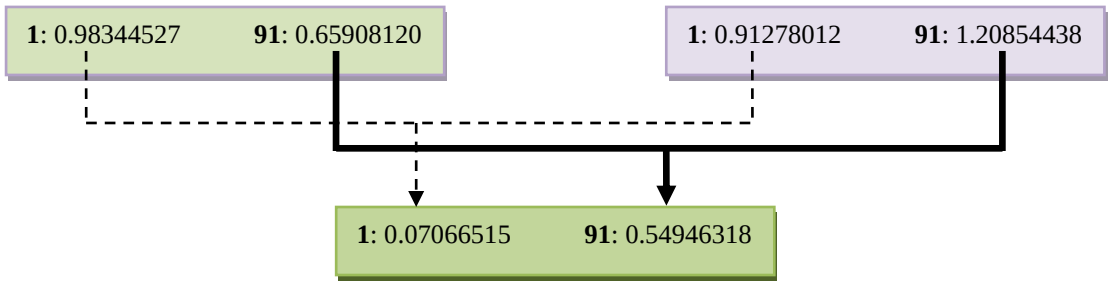

Figura 4.5: Vetor de Dissimilaridade

A expressão do exemplo acima é representada por $Z_i = \|V_i - Q_i\|$, tal que Z_i é a resultante da subtração entre os vetores de características estilométricas extraídas das colunas V_i e Q_i.

4.5 Modelos de Comparação

Os sistemas automáticos de verificação da autoria de textos baseiam-se usualmente em duas abordagens de modelos para classificação: dependente e independente do autor [JSB03]. Diante disso, este trabalho utilizará os dois modelos, que são descritos a seguir.

4.5.1 Modelo Independente do Autor

O modelo independente do autor utiliza o conceito da dicotomia, ou seja, a divisão do modelo em duas classes, sendo elas: autoria (+1) e não autoria (-1). A geração do modelo independente do autor ocorre com um conjunto de autores escolhidos aleatoriamente, combinando-se amostras de um mesmo autor e de autores diferentes. O modelo independente do autor possui a vantagem de necessitar um número pequeno de exemplares de cada autor e de não necessitar de um novo treinamento do modelo, na inclusão de novos autores. [BAR05]

No treinamento do modelo independente do autor, a classe w_1 representa a classe de amostras genuínas dos autores usados para o treinamento (autoria). A classe w_2 representa o conjunto de amostras pertencentes a autores distintos (não autoria). Na verificação, o modelo gerado é então utilizado para a comparação com a amostra desconhecida (fase de testes).

Na fase de treinamento alguns pontos importantes têm que ser considerados:

(i) Os autores utilizados para a fase de treinamento devem ser exclusivos para tal, ou seja, os autores utilizados na fase de treinamentos não podem ser utilizados na fase de testes [PAV07];

(ii) A quantidade de vetores de autoria e de não autoria devem ser as mesmas [JUS02];

(iii) Evitar o uso de sobre-treinamento [PAV07];

O modelo independente do autor em dicotomia possui vetores de autoria e de não autoria, que são gerados a partir dos vetores de dissimilaridade entre documentos de um mesmo autor, que são separados exclusivamente para os testes, e vetores de dissimilaridades gerados de documentos de autores diferentes (aleatórios). Com estes vetores de autoria e de não autoria o classificador SVM se encarregará de gerar o modelo unívoco que será utilizado na fase de testes. Foi utilizado o pacote freeware SVM^{light} [JOA02] para as etapas de aprendizado e teste com o modelo independente do autor. Com relação à configuração do classificador, foram feitos testes com *kernel* linear iniciando com 1, parâmetros –d em 3, –g, -r e -s em 1. Maior detalhamento sobre as configurações e parâmetros do SVM pode ser encontrado em [JOA02].

4.5.2 Modelo Dependente do Autor

O modelo dependente do autor é baseando no conceito da policotomia, ou seja, a classificação do problema em *n*-classes [BAR05]. Nesse modelo, cada autor representa uma classe. O modelo dependente do autor exige um conjunto elevado de amostras genuínas para sua geração, pois para cada autor será gerado um modelo específico e que descreve adequadamente as características do mesmo. Este modelo apresenta a vantagem de descrever adequadamente as variabilidades intrapessoais do autor, apresentando, porém, a desvantagem da geração de um novo modelo a cada inclusão de um novo autor [BAR05][PAV07].

Como no modelo independente do autor, no modelo dependente do autor gera-se um único modelo, porém cada vetor treinado possui a informação a qual classe pertence ($w_{1..n}$), significando que só vetores de autoria podem ser treinados.

Os vetores de autoria são resultado das possibilidades dada pela equação 7, representada a seguir.

$$[b!]A_n^d = \frac{\frac{n!}{(n-d)!}}{2} \qquad (7)$$

Onde *A* é igual ao arranjo de *d* elementos em *n* (número de documentos separados para treinamento).

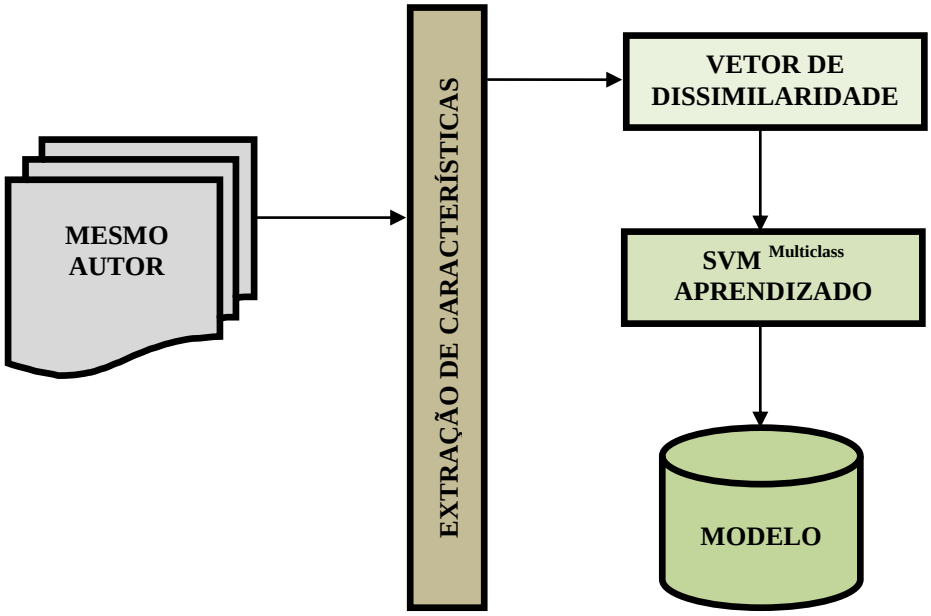

Figura 4.6 – Fluxo de operações com o SVM Multiclasse – Adaptado de [PAV07]

Com o modelo dependente por autor através do classificador SVM Multiclasse este trabalho se propõe a comparar os resultados de cada autor e cada classe, a fim de comparar com os resultados atingidos por Pavelec [PAV07].

4.6 Classificação

Acontece nesta fase a classificação dos documentos questionados, que são classificados em função do modelo gerado pelo treinamento (Figura 4.7). O processo de teste ocorre especificamente para cada um dos modelos propostos respeitando o protocolo de testes.

Seguindo o modelo apresentado por [PAV07] a classificação no modelo independente do autor consiste em testar a base de dados contra o modelo gerado (vetores de autoria e de não autoria) para a obtenção da taxa de erros de falsa rejeição (quantidade de vezes que um autor genuíno é rejeitado) e de falsa aceitação (quantidade de vezes que um não autor é aceito como se fosse o autor verdadeiro).

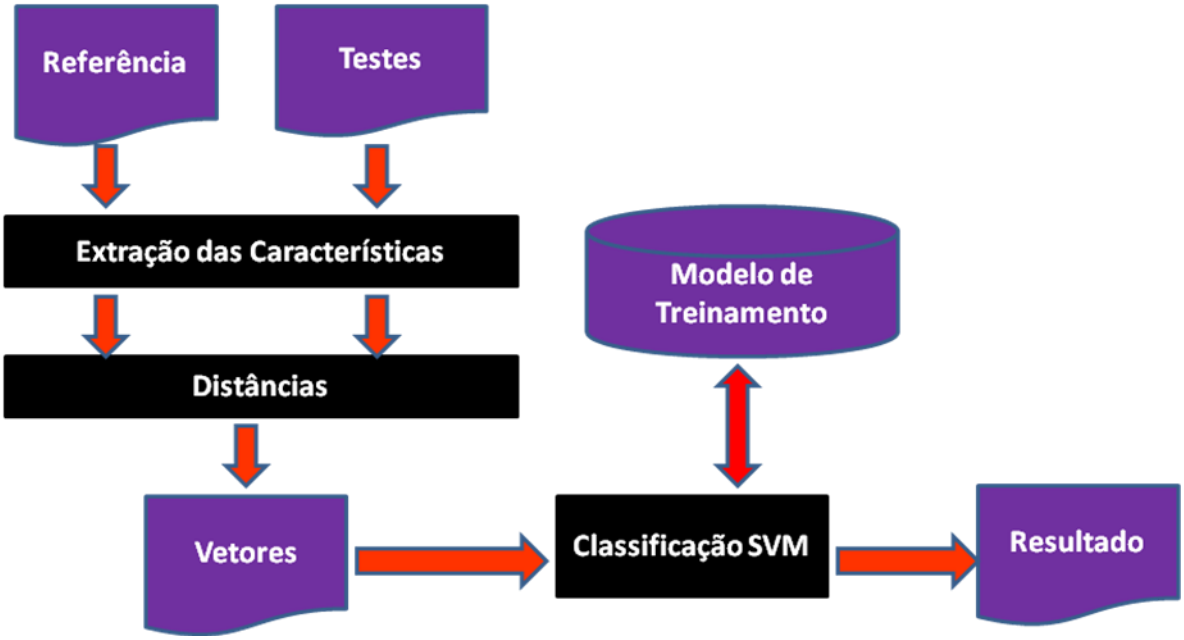

Figura 4.7: Modelo de Classificação

Os vetores de autoria são gerados através das possibilidades dada pela equação 7, onde *A* é o arranjo de *d* (exemplo: igual a 2) elementos em *n* (números de documentos separados para teste). Já para a geração dos vetores de não-autoria, os documentos utilizados no treinamento funcionam como referência, ou seja, os vetores de dissimilaridade são formados por um documento de testes e um documento de treinamento de autor diferente (aleatório). Ao final do processo, é gerado uma mesma quantidade de vetores de autoria (+1) e de não autoria (-1), que são classificados em inferência com o modelo independente do autor gerado na fase de treinamento.

No modelo dependente do autor, acontece o mesmo procedimento que no modelo independente do autor, onde os modelos gerados pelo treinamento são utilizados como referência, ou seja, os vetores de dissimilaridade são formados por um documento de teste e um documento de treinamento de mesmo autor.

4.7 Decisão Final

Nesta fase o documento é classificado para cada um dos modelos propostos. Tendo como saída um resultado de autoria ou de não-autoria.

No modelo independente do autor é aplicado um processo de voto, utilizando a combinação de classificadores que analisa o grau de confiança de cada uma das comparações para o documento. Por exemplo: Se um determinado protocolo possui 3 documentos de referência, um documento questionado (Q) terá seus vetores de dissimilaridade com cada um dos 3 documentos de referência gerando 3 vetores. Assim a classificação final para o documento questionado (Q) é gerada pelo voto destes 6 vetores (3 de autoria e 3 de não autoria). Com base nestes votos é atribuída a autoria como falsa ou verdadeira, podendo também através deste processo serem calculadas as taxas de falsa rejeição e falsa aceitação.

No modelo dependente do autor o processo de decisão é gerado pela própria saída do classificador, pois o resultado do documento classificado gera uma classe, ou seja, cria uma matriz de confusão onde se atribui ao documento questionado a classe associada.

4.8 Considerações Finais

Neste capítulo foi apresentada a metodologia para a identificação de autoria de textos. As seções abordaram os procedimentos gerais do processo identificação, como: o método de identificação de autoria, a coleta e formação da base de dados, a apresentação das características e o processo de extração, a geração dos vetores de dissimilaridade, os modelos de comparação dependente e independente de autor, bem como a classificação e a decisão final.

São apresentados dois modelos para atribuição de autoria em documentos questionados de língua portuguesa. Os modelos se diferenciam, sendo um o modelo independente do autor, e outro o modelo dependente do autor. O modelo independente tem sua contribuição através da sua generalização e utilizar apenas um pequeno número de amostras do autor. Já o modelo dependente apresenta a vantagem de descrever adequadamente as variabilidades intrapessoais do autor, sendo que para isso o modelo exige

um conjunto elevado de amostras. Temos assim, dois modelos diferentes para realização dos experimentos.

No próximo capítulo são apresentados os experimentos realizados e a análise dos resultados obtidos nesta pesquisa.

Capítulo 5

Experimentos e Análise dos Resultados

Neste capítulo são evidenciados os experimentos e a análise dos resultados obtidos com este trabalho. São expostos os resultados nos métodos: independente e dependente do autor e também comparados com os resultados de Pavelec [PAV07], bem como os resultados de contribuição de cada grupo de características para a identificação de autoria em textos de língua portuguesa.

5.1 Ambiente de Software e Hardware

Para a realização dos experimentos foram utilizadas as seguintes plataformas de hardware e software (Tabela 5.1 e 5.2).

Tabela 5.1 – Ambiente de Hardware

HARDWARE	
Processador	Intel Pentium Dual Core 2.1 Ghz
Memória RAM	3 GB

Os experimentos foram realizados com as especificações de software detalhadas na Tabela 5.2.

Tabela 5.2 – Ambiente de Software

SOFTWARE	
Sistema Operacional	Ubuntu 10.04
Linguagem/Ferramenta do Algoritmo de Extração de Características	Java / BlueJ
Classificador - Modelo Independente do Autor	SVM *Light*
Classificador - Modelo Dependente do Autor	SVM *Multiclass*

5.2 Modelo Independente do Autor

O modelo independente do autor apresenta o conceito de dicotomia, ou seja, apenas dois grupos de vetores de características: de autoria (+1) e de não autoria (-1). Este modelo se comporta de maneira generalista, pois através das características geradas o classificador automaticamente criará os modelos do processo de aprendizagem.

No modelo independente do autor foram criados dois ambientes para realização dos experimentos, sendo que no primeiro são utilizados 29400 vetores para os testes, dentre os quais 14700 de autoria e 14700 de não autoria. No segundo ambiente são efetuados os testes por classes de assuntos, onde foram postos em testes 4200 vetores, sendo 2100 de autoria e 2100 de não autoria para cada uma das 10 classes de assuntos testadas.

5.2.1 Protocolo de Experimentos - Modelo Independente do Autor

A base utilizada para os experimentos no modelo independente do autor foram textos de colunas de jornais e blogs disponibilizados na internet. A base de textos construída para a realização dos experimentos possui 100 autores, dividido em 10 classes de assuntos, sendo que cada autor possui uma amostra de 30 textos, totalizando 3000 documentos. Com isso a base foi dividida para o aprendizado e os testes no modelo, que ficaram assim compostos (Tabela 5.3):

Tabela 5.3 – Divisão da base de Dados para o Modelo Independente do Autor

Modelo	Quantidade		
	Autores	**Documentos por Autor**	**Referência**
Aprendizado	30	7	-
Testes	70	30	7

Os protocolos de aprendizado e de testes do modelo independente do autor se encontram de forma detalhada nas próximas subseções.

5.2.2 Protocolo de Aprendizado

No modelo independente do autor foram criados dois ambientes. O primeiro ambiente foi utilizado para testar as regras de fusão, a fim de selecionar a regra que obtivesse o melhor resultado para utilização nos testes seguintes (Ambiente 1). Em um segundo ambiente, a regra de fusão que obteve o melhor resultado no primeiro ambiente foi utilizada em um experimento por classe de assunto (Ambiente 2) para verificar qual dos grupos de características obtinha o melhor resultado.

Para uma melhor confiabilidade do modelo nenhum dos autores participantes da fase de aprendizado participa da fase de testes, pressupondo que o classificador fará todo o processo de autenticação com autores nunca vistos anteriormente.

De posse dos 7 documentos de cada autor escolhidos para a fase de aprendizado, são gerados dois conjuntos de vetores de dissimilaridade: (i) de autoria, e (ii) de não-autoria. Como no modelo independente do autor é necessário que o modelo de aprendizagem seja balanceado deve-se ter a mesma quantidade de vetores de dissimilaridade de autoria e de não-autoria. Os vetores de dissimilaridade de autoria são gerados entre documentos de um mesmo autor. Neste caso o número de vetores será igual a 630, pois a equação 8 mostra os detalhes do cálculo da análise combinatória, através de uma combinação simples multiplicado pelo número de autores.

$$A_7^2 = \frac{\frac{7!}{(7-2)!}}{2} \qquad (10)$$

Aonde se chega ao resultado do arranjo que é 21, que depois é multiplicado pela quantidade de autores (neste caso, 30), totalizando 630 vetores de autoria (+1). Para se gerar

os mesmos 630 vetores de não autoria (-1), cada um dos 7 documentos gerou um vetor em comparação com um documento de outro autor que estão na relação separada exclusivamente para a fase de aprendizagem.

Diante dos cálculos apresentados e de posse dos vetores de dissimilaridade de autoria e de não-autoria, os vetores foram submetidos ao aprendizado com o SVM, onde se gerou um modelo independente do autor para posterior aplicação na fase de testes. Na Figura 5.1, detalha-se o esquema, onde:

- **TrainOK.txt**: representa os vetores de autoria;
- **TrainFalse.txt**: representa os vetores de não-autoria;
- **Train.txt**: Representa a junção dos vetores de autoria e não-autoria em um só documento;
- **Svm_learn.exe**: representa o classificador SVM^{Light}, para realização do treinamento do modelo;
- **Model.dat**: é modelo de treinamento gerado pelo SVM^{Light} em função da entrada do documento train.txt que contém os vetores de autoria e não-autoria concatenados.

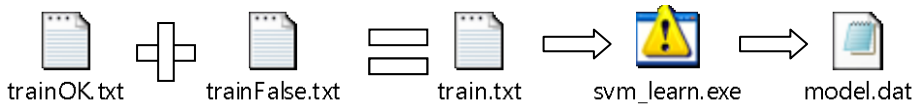

Figura 5.1 – Representação do Processo de Treinamento

5.2.3 Protocolo de Testes – Ambiente 1 (Seleção da Melhor Regra de Fusão)

Da mesma forma que no protocolo de aprendizado, no protocolo de testes é necessário a geração de vetores de autoria e de não-autoria. O processo consiste em testar contra o modelo independente do autor gerado, vetores de autoria e de não-autoria, a fim de se obter as taxas de erros de falsa rejeição (FR) e falsa aceitação (FA).

Para geração dos vetores de autoria foram utilizados os 30 documentos de cada autor separado para os testes. Cada um dos 30 documentos gerou um vetor de autoria utilizando 7 amostras de seus próprios documentos como referência. Diante disso, efetuou-se o seguinte cálculo: 70 autores * 30 documentos * 7 documentos de referência, totalizando 14700 vetores (Representado pela Equação 11).

$$Nv = Qa \times Nd \times Dr \quad (11)$$

Onde,

N_v = Número de vetores

Q_a = Número de autores

N_d = Número de documentos por autor

D_r = Número de documentos de referência

Para gerar os vetores de não-autoria foi aplicado o mesmo cálculo, mas com 7 amostras de autores diferentes que foram escolhidos aleatoriamente.

Para uma melhor confiabilidade nos resultados o processo foi executado 3 vezes. Sendo uma vez com uma base *x* de autores no aprendizado e outra base *yz* como teste. Uma segunda execução com a base *y* no aprendizado e base *xz* para teste. E a última, com a base *z* como aprendizado e a base *xy* para teste. A base de autores escolhidas para o aprendizado foi efetuada de forma aleatória e sem repetição. O protocolo de testes é mais detalhado na Tabela 5.4.

Tabela 5.4 - Protocolo de Testes – Modelo Independente do Autor

Modelo Independente do Autor				
Processo	**Autores**	**Documentos**	**Vetores de Autoria**	**Vetores de Não Autoria**
Aprendizado	1-30	1-7	630	630
Testes	31-100	1-30	14700	14700
Voto Majoritário Simples			14700	14700

Após a definição do protocolo de testes (no qual 29400 vetores são postos para verificação da autoria), os vetores de autoria e não-autoria são postos em testes para a classificação.

5.2.4 Resultados - Ambiente 1 (Seleção da Melhor Regra de Fusão)

Os resultados proporcionados pelos testes, no modelo independente do autor são mostrados na Tabela 5.6. Observa-se que os resultados produzidos foram com base no

protocolo de aprendizado e testes, onde foi utilizado o conjunto de características proposto por este trabalho. Houve nesta fase a seleção das melhores características, que resultaram nos melhores resultados, que foi efetuada através de algoritmo genético, que utilizou dos parâmetros de configuração conforme detalha a Tabela 5.5.

Tabela 5.5 – Parâmetros do Algoritmo Genético para escolha das Melhores Características

Numero de Gerações	1000
Tamanho da População	50
Tamanho do Vetor de Características	408
Estratégia de Seleção	*Roulette*
Mínimo de Características a serem eliminadas	10
Máximo de Características a serem eliminadas	380
Tamanho da Base de Testes	29400

No processo de escolha do melhor conjunto de características, foram testadas 1000 gerações, onde, em cada geração o tamanho da população foi estipulada em 50 combinações diferentes. Observa-se que no mínimo a melhor solução teria 10 características discriminantes e no máximo 380.

Foram utilizadas diferentes regras de fusão para combinar a saída do classificador. As regras do máximo, mínimo e média e voto majoritário foram avaliadas nos experimentos. O melhor resultado foi produzido pela regra do voto majoritário.

Tabela 5.6: Resultados dos Testes – Regras de Fusão

TAXAS DE ACERTO			
MÁXIMO	**MÉDIA**	**MÍNIMO**	**VOTO MAJORITÁRIO**
69,7%	71,4%	72,8%	74,3%

Coma base neste teste, todos os experimentos a seguir foram baseados na regra do voto majoritário.

5.2.5 Protocolo de Testes – Ambiente 2

Neste protocolo de testes, foram utilizados como referência os 10 autores da classe de assuntos variados por serem mais generalistas quanto a amostras de seus textos. Foram utilizados 7 amostras de cada autor de forma aleatória, escolhida entre os 30 textos disponíveis. Na fase de teste, todos os 10 autores e suas 30 amostras de textos foram confrontados com 7 documentos de referência, gerando assim 2100 vetores de autoria e 2100 vetores de não autoria.

5.2.6 Resultados por Classe - Ambiente 2

Os resultados foram concatenados em grupos de características e também pelo grupo geral que contém todas as características.

Para fins de identificação de autoria, os melhores resultados foram proporcionados pelo uso da regra do voto majoritário, e são estes os resultados que serão utilizados neste trabalho. A seguir (Tabela 5.7) são mostrados os resultados (em %) pelas suas respectivas classes de assuntos.

Tabela 5.7 – Resultados dos Testes

Grupo de Características	**Classes de Assuntos**									
	Direito	**Economia**	**Esportes**	**Gastronomia**	**Literatura**	**Politica**	**Saúde**	**Tecnologia**	**Turismo**	**Variados**
Advérbios	61,3	68,4	67,5	60,6	63,4	64,3	75,0	77,4	71,8	75,2
Conjunções	60,6	61,2	60,0	66,1	60,0	75,8	60,4	61,7	74,7	70,5
Pronomes	68,2	70,3	71,6	71,7	65,5	64,6	65,6	68,4	66,5	74,4
Verbo	63,9	67,6	64,0	65,1	65,6	64,5	64,8	61,1	62,8	64,8
Todas	63,8	65,6	69,7	69,4	62,9	77,1	67,5	63,2	63,1	74,5

Na classe direito, os melhores resultados foram proporcionados pelo uso dos pronomes, com acerto em 68,2% dos casos. Na base de testes verificou-se que 82 formas de pronomes apareceram das 87 possíveis, o que representa 94% de efetividade. Como o

algoritmo genético seleciona as melhores características, as quais de certa forma se tornam discriminantes para o processo de identificação de autoria, nesta classe a seleção foi de 41 características do universo de 87, representando o uso de 47% do total de características propostas.

No assunto economia o grupo de característica que se sobressaiu foram os pronomes, com 70,3% de acerto. Sobre a sua efetividade a mesma foi de 94%, ou seja, 82 de 87 características apareceram na base de testes. Neste caso, o melhor resultado apresentou a exclusão de 51 características e o uso de 36 pronomes (41%).

Para a classe esportes, o melhor resultado foi atingido pelos pronomes, com um acerto de 71,6%. Sua efetividade foi de 94%. Na seleção das características discriminantes, foram excluídas 48, e utilizadas 39 pronomes que representa 45% do total de características.

Sobre o assunto gastronomia, o grupo de características que obteve o melhor resultado foram os pronomes, com um acerto médio de 71,7%. Das 87 características pertencentes ao grupo de pronomes, 82 apareceram nas amostras de textos, caracterizando 94% de efetividade. Na seleção genética, 38 pronomes foram selecionados como discriminantes, gerando uma taxa de utilização de 44% do total de características propostas.

Na classe de assunto literatura, dois grupos se destacaram sendo verbos com 65,6% e pronomes com 65,5% de acerto. Observa-se que os verbos tiveram uma ligeira vantagem de 0,1%, então a observação e análise foi efetuada em cima deste grupo de características. A efetividade do grupo de verbos de 81%, pois 122 de 150 características surgiram nos textos da base testada. Foram selecionadas 47 características discriminantes no processo de seleção genética, o que representa o uso de 31% do total de características propostas.

O grupo de características que obteve o melhor resultado na classe de assunto política foram as conjunções, que atingiram um acerto de 75,8%. Sua efetividade foi de 78%, onde 60 das 77 características propostas apareceram nas amostras dos textos. O algoritmo genético selecionou 37 características e excluiu as outras 40, representando que o uso de 48% das características atingiu o melhor resultado.

Para a classe de assunto saúde, o melhor resultado foi obtido pelo grupo de advérbios, o qual atingiu um acerto de 75%. No que se refere a sua efetividade, a mesma foi de 97%, pois 91 das 94 características surgiram nas amostras testadas. Na seleção genética, 41 características foram selecionadas como discriminantes, o que representa 44% do uso de sua totalidade de características.

Na classe de tecnologia, os advérbios obtiveram o melhor desempenho entre os grupos, perfazendo 77,4% de acerto. Em respeito a sua efetividade, 91 características foram encontradas do total de 94, o que representa 97%. Foram selecionadas como melhores características, o total de 39, que representa 41% do universo de características do grupo de tecnologia.

Para o grupo de assunto de turismo as melhores características selecionadas foram as conjunções, que tiveram um acerto de 74,7%. A efetividade das características foi de 77%, ou seja, 59 das 77 características do grupo de conjunções surgiram nos textos. De 77 características, 35 foram selecionadas pelo algoritmo genético, o que corresponde a 45% da sua totalidade.

Na classe assuntos variados o treinamento foi efetuado com os autores aleatórios de diversas classes, seguindo o mesmo protocolo apresentado anteriormente.

Para o grupo de assuntos variados, o melhor grupo de características selecionado foram os advérbios, que atingiram um resultado de 75,2%. A efetividade foi de 98%, pois 92 das 94 características propostas surgiram nos textos. Na seleção genética, o algoritmo selecionou 37 características como sendo discriminantes, que representa 39% do total de características propostas.

Na Tabela 5.8 são apresentadas as características que apresentaram os melhores resultados por classe de assunto, conforme mostrado na Tabela 5.7. Tais características foram selecionadas através de seleção genética.

Tabela 5.8 – Características Selecionadas pelo Melhor Grupo

CLASSE	MELHOR GRUPO DE CARACTERÍSTICAS	CARACTERÍSTICAS SELECIONADAS
Direito	Pronomes	A qual, a, aquele, aqueles, as, cuja, cujo, desta, ele, eles, em que, essa, essas, esse, esses, esta, estas, este, estes, eu, isso, isto, lhe, lhes, me, meu, minha, nessa, nos, nossa, nosso, o, onde, os, quanto, quem, se, seu, seus, sua, suas.
Economia	Pronomes	A, aquele, aqueles, as, cuja, desta, ele, eles, em que, essa, essas, esse, esses, esta, este, estes, eu, isso, me, minha, nessa, nos, nós, nossa, nosso, o, onde, os, quanto, quem, se, seu, seus, sua, suas, você.
		A, aquela, aquele, aqueles, as, desta, ele, eles, em que, essa, essas, esse, esses, esta, este, eu, isso, lhe,

Esportes	Pronomes	me, meu, minha, nessa, nos, nós, nossa, nosso, nossos, o, onde, os, quanto, quem, se, seu, seus, si, sua, suas, você.
Gastronomia	Pronomes	A, as, aquela, aquelas, aquele, aqueles, ele, eles, em que, essa, esse, esses, esta, este, estes, eu, isso, isto, me, meu, meus, mim, minha, minhas, nos, nós, nosso, o, onde, os, quanto, quem, se, seu, seus, sua, suas, você.
Literatura	Verbos	Aberto, abrir, acabar, achando, achar, andar, beber, comer, completar, dado, dando, dar, dito, dizendo, dizer, entender, escolher, escrevendo, escrever, escrito, estado, estar, falando, falar, fazendo, fazer, feito, haver, jogar, lendo, ler, lido, ligar, pensando, pensar, poder, sendo, ser, sido, tendo, ter, tido, usar, vendo, ver, visto, viver.
Politica	Conjunções	Ainda que, assim como, bem como, caso, como também, como, conforme, contudo, desde que, e, embora, entretanto, já que, logo, mais do que, mais que, mas ainda, mas também, mas, mesmo que, não ser que, nem que, nem, para que, por isso, porém, porque, portanto, quanto mais, que nem, que, se, se bem que, segundo, sem que, tal como, todavia.
Saúde	Advérbios	Acima, agora, aí, ainda, além, algo, alguém, algum, alguma, antes, apenas, aqui, assim, atrás, atualmente, bastante, bem, cada, demais, dentro, depois, então, hoje, já, longe, mais, menos, muita, muito, não, nenhum, nunca, onde, outra, outro, pouco, qualquer, quase, sempre, tanto, todo.
Tecnologia	Advérbios	Abaixo, acima, agora, aí, ainda, além, algo, alguém, algum, alguma, antes, apenas, aqui, assim, atrás, bem, cada, dentro, depois, então, hoje, já, lá, longe, mais, menos, muito, não, nunca, onde, outro, pouco, qualquer, sempre, talvez, tanto, toda, todo, tudo.
Turismo	Conjunções	à medida que, ainda que, assim como, bem como, caso, como, conforme, desde que, e, entretanto, embora, já que, logo, mais do que, mais que, mas ainda, mas também, mas, menos do que, mesmo que, nem, para que, por isso, porque, portanto, porém, quanto mais, que, que nem, se, senão, segundo, uma vez que, visto como, visto que.
Assuntos Variados	Advérbios	Abaixo, acima, agora, aí, ainda, além, algo, ali, antes, apenas, aqui, assim, bem, cada, dentro, demais, depois, então, hoje, já, mais, menos, muito, nada, não, ninguém, nunca, onde, pouco, qualquer, quase, sempre, sim, tanto, tarde, todo, tudo.

Na seção 5.2.7 são apresentados os resultados finais do modelo independente do autor de forma concatenada.

5.2.7 Resultados Finais Concatenados

Para 40% das classes de assuntos o melhor grupo de características para a identificação de autoria foram os pronomes que se sobressaíram quando o assunto é relacionado a direito, economia, esportes e gastronomia. Em 30% dos casos os advérbios são as melhores características (saúde, tecnologia, e assuntos variados). Quando se refere ao assunto de política e turismo, o grupo de características que tem um melhor resultado são as conjunções. E, na classe de literatura os verbos são os melhores identificadores. (Ver Tabela 5.9)

Os resultados apresentados na Tabela 5.9, foram obtidos através da melhor seleção de características selecionadas pelo algoritmo genético.

Tabela 5.9: Resultados Concatenados por Classe de Assuntos

CLASSE DE ASSUNTO	MELHOR GRUPO DE CARACTERÍSTICAS	TAXA DE ACERTO
ASSUNTOS VARIADOS	ADVÉRBIOS	**75,2%**
DIREITO	PRONOMES	**68,2%**
ECONOMIA	PRONOMES	**70,3%**
ESPORTES	PRONOMES	**71,6%**
GASTRONOMIA	PRONOMES	**71,7%**
LITERATURA	VERBOS	**65,6%**
POLITICA	CONJUNÇÕES	**75,8%**
SAÚDE	ADVÉRBIOS	**75,0%**
TECNOLOGIA	ADVÉRBIOS	**77,4%**
TURISMO	CONJUNÇÕES	**74,7%**

Fazendo uma análise mais minuciosa dos grupos de características em função da classe de assunto, verificou-se que as classes assuntos variados, saúde e tecnologia possuem os advérbios como características discriminantes. O advérbio é a palavra que acompanha ou

modifica o verbo, dando a ideia de tempo, modo e lugar por exemplo. Nas classes de assunto tecnologia e saúde, por exemplo, os advérbios são muito utilizados como comparações de igualdade, superioridade ou inferioridade e também para dar uma visão mais analítica do assunto, tais como: a causa, a finalidade e o meio. Já na classe assuntos variados, os textos são muitos heterogêneos o que dificulta a análise, no entanto, prevalece a utilização de comparações e da visão analítica.

Nos grupos de assuntos, direito, economia, esportes e gastronomia o melhor grupo de características foram os pronomes. O pronome é a palavra que tem por função acompanhar ou substituir um nome. Nas classes de assuntos citadas acima, percebe-se a uniformidade do tratamento, as demonstrações e relações ocorridas, e a ações reflexivas intrínsecas nestes tipos de textos.

Na classe de politica e turismo, por exemplo, o grupo de característica que teve destaque foi às conjunções, pois as mesmas tem função de estabelecer uma relação entre as frases. Nestas classes de textos podem-se identificar diversas ocorrências, por exemplo, na classe política há o uso frequente de conjunções que expressam oposição, comparações, alternativas e explicações. Na classe turismo, as comparações, as alternativas, as condições e as proporções eximem um perfil da classe.

Para a classe de literatura o grupo que se destacou foi os verbos, que tem por função principal indicar ações. Nos textos desta classe é muito comum o uso de verbos, pois na sua grande maioria são textos narrativos de determinados fatos – que expressa à ação, que representam verbos. Pode-se avaliar também, que os autores de textos de literatura expressam-se mais através de orações completas (que possuem verbo em seu contexto), e não de simples frases. Geralmente por serem indivíduos com um conhecimento maior das estruturas gramaticais da língua portuguesa, se utilizam de linguagens mais complexas e fazem do uso do verbo uma constante em seus textos.

5.3 Modelo Dependente do Autor

No modelo dependente do autor, foi verificado como se comportam os mesmos vetores com a técnica de classificação por autor. Nesta técnica cada texto questionado é classificado de acordo com o autor que mais se assemelha, onde cada autor representa uma classe.

5.3.1 Protocolo de Experimentos

Nos experimentos com este modelo, todos os 100 autores da base de dados são testados, onde são separados 7 documentos de cada autor para a fase de aprendizagem e os outros 23 documentos restantes são utilizados na fase de testes. Cada um dos 23 documentos dos 100 autores utilizam como referência os 7 documentos utilizados no aprendizado, que são escolhidos de forma aleatório e sem repetição.

5.3.2 Protocolo de Aprendizado

Para o modelo dependente do autor duas baterias de testes foram geradas. A primeira efetuou os testes com toda a amplitude de autores. Na segunda, somente foram testados os autores que pertenciam à classe em teste (este foi repetido 10 vezes, para atingir todas as classes propostas).

Na técnica de classificação dependente do autor, somente é gerado um conjunto de vetores de autoria, que representa a sua própria classe (diferente do modelo independente do autor que gera vetores de autoria e de não autoria). Este conjunto de vetores é gerado entre os 7 documentos de aprendizado do mesmo autor. A quantidade total de vetores neste caso será de 2100, sendo 21 vetores por autor e 210 vetores por classe de assunto, que é calculado pela equação 8.

A Figura 5.2 representa um exemplo da combinação dos vetores de autoria, utilizado na fase de aprendizado.

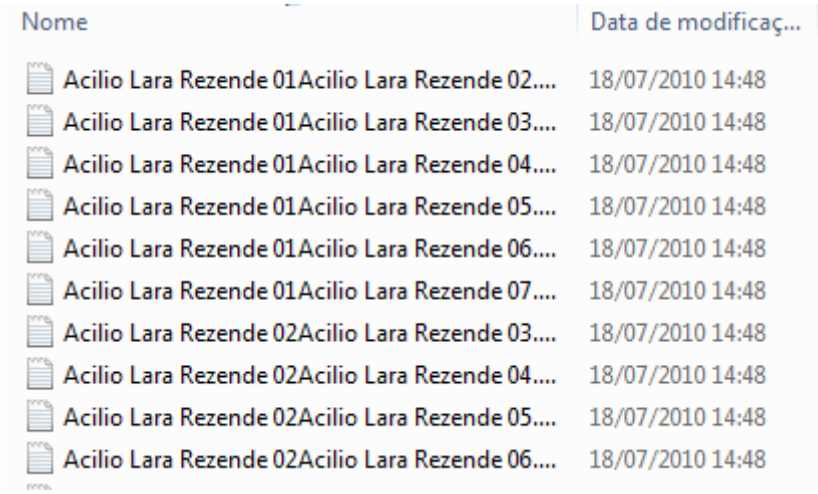

Figura 5.2: Vetores de Autoria gerados no modelo multiclasse

Com os vetores de autoria devidamente gerados, os mesmos foram submetidos ao aprendizado através do SVM ^{Multiclass}, que gerou um modelo único por autor.

5.3.3 Protocolo de Testes

Para os testes no modelo dependente do autor, cada um dos 23 documentos dos 100 autores geram 7 vetores, que é o produto da comparação com os 7 documentos do mesmo autor que foi separado para o aprendizado, que nos testes são utilizados como referência. Sendo assim, 16100 vetores são testados contra o modelo, que é dado efetuando o cálculo: 100 autores * 23 documentos * 7 documentos de referência, também representado pela equação 9.

Neste trabalho é utilizado o voto majoritário simples, pois foi a regra de fusão que obteve os melhores resultados na abordagem independente do autor e também é utilizada na abordagem dependente do autor. Neste caso o documento que for colocado em teste, através da obtenção das 7 respostas que o documento questionado foi submetido, e em razão aos 7 documentos de referência, tem a sua classificação final. Por exemplo, a atribuição da classe a um documento questionado *a* é dado através dos 7 votos deste documento, como é representado na Figura 5.3, onde cada circulo secundário representa um voto e o circulo central representa a saída final, que é dada através do voto majoritário simples (maioria dos votos), que define a atribuição de uma determinada classe ao documento questionado.

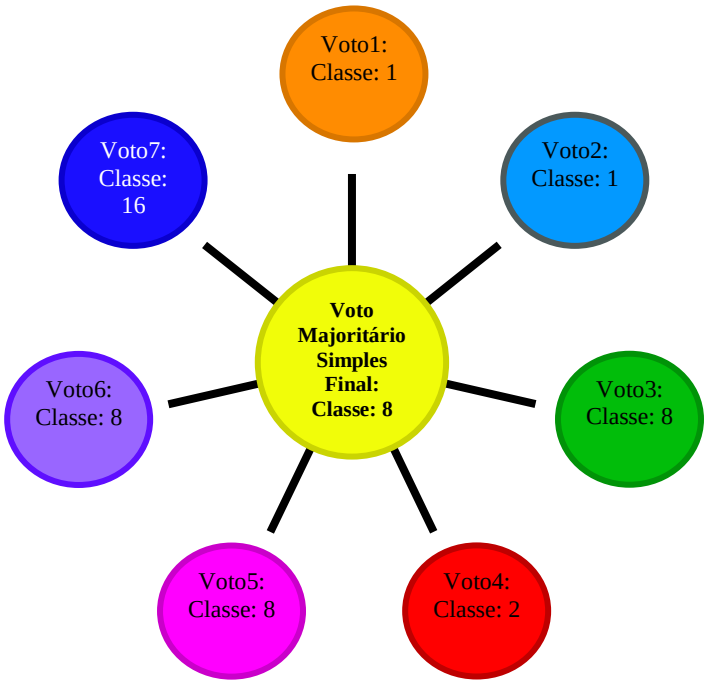

Figura 5.3: Representação do Processo de Voto Majoritário Simples

Os protocolos de testes são listados nas Tabelas 5.10 e 5.11. Na Tabela 5.10 é evidenciado o protocolo com toda a base de dados, ou seja, os testes com todo o universo de autores; Já na Tabela 5.11 é representado o protocolo por classe de assunto, onde foram testados somente os documentos que pertenciam àquela determinada classe. Este protocolo foi repetido 10 vezes, ou seja, uma vez para cada classe.

Tabela 5.10: Protocolo de Testes – Base Geral

Processo	**Autores**	**Documentos**	**Vetores de Autoria**
Aprendizado	1-100	1-7	2100
Referência	1-100	1-7	2100
Testes	1-100	8-30	16100
Voto Majoritário Simples			16100

Tabela 5.11: Protocolo de Testes – Base por Classe de Assunto

Processo	Autores	Documentos	Vetores de Autoria
Aprendizado	1-10	1-7	210
Referência	1-10	1-7	210
Testes	1-10	8-30	1610
Voto Majoritário Simples			1610

Para cada documento questionado é atribuído uma determinada classe, que gera posteriormente uma matriz de confusão, de onde serão obtidos os resultados do modelo dependente do autor.

Nas Figuras 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12 e 5.13 são apresentados exemplos de matrizes de confusão para os testes realizados somente com os autores e textos de cada classe, com base no protocolo especificado na Tabela 5.11. Neste caso, os votos dos vetores foram calculados tendo como universo, somente os autores que pertenciam àquela determinada classe, excluindo assim o restante dos autores.

ASSUNTOS VARIADOS											
AUTORES	Fátima Oliveira	Gilberto Dimenstein	Gilda de Castro	Grace Passô	Luis F. Sapori	Marcelo Rossi	Oswaldo Braga	Sebastião Nunes	Silvana Mascagna	Trigueirinho	TOTAL DE TEXTOS
Fátima Oliveira	18			1	1			2		1	23
Gilberto Dimenstein	2	17			2			1	1		23
Gilda de Castro	1		17	2			1		2		23
Grace Passô	2	1	2	16					2		23
Luis F. Sapori		3			17		2			1	23
Marcelo Rossi	1					20			1	1	23
Oswaldo Braga			2	2			18			1	23
Sebastião Nunes	1				2			18	1	1	23
Silvana Mascagna		1	1			1			20		23
Trigueirinho		1			2	1	1			18	23

Figura 5.4: Matriz de Confusão – Classe Assuntos Variados

DIREITO											
AUTORES	Boleslau Sliviany	Carlos Z. Junior	Fábio Tokars	Fernando Cesar Faria	Frederico Vasconcelos	Igor F. Rodrigues	Jorge Alberto Araújo	Maria inês Dolci	Oscar Ivan Prux	Rene Ariel Dotti	TOTAL DE TEXTOS
Boleslau Sliviany	19	1				1		2			23
Carlos Z. Junior		16		3		1		1		2	23
Fábio Tokars			19		3				1		23
Fernando Cesar Faria		2		17			2			2	23
Frederico Vasconcelo	2				18	1				2	23
Igor F. Rodrigues				2	2	17				2	23
Jorge Alberto Araújo			1	2	2		16			2	23
Maria inês Dolci							4	18	1		23
Oscar Ivan Prux						2		1	19	1	23
Rene Ariel Dotti						3			2	18	23

Figura 5.5: Matriz de Confusão – Classe Direito

ECONOMIA											
AUTORES	Ana C. Cavalcante	Antonio Pietrobelli	Benedicto Dutra	Claudio Gradilone	Fernando Canzian	Guilherme Barros	Karlon Aredes	Luis Nassif	Valdo Cruz	Vinivius T. Freitas	TOTAL DE TEXTOS
Ana C. Cavalcante	19		2		2						23
Antonio Pietrobelli		17			3			1	2		23
Benedicto Dutra	2		18	1					1	1	23
Claudio Gradilone			3	17			1	2			23
Fernando Canzian	1				18	3	1				23
Guilherme Barros			2			19	1			1	23
Karlon Aredes						1	19		2	1	23
Luis Nassif		1			3			17		2	23
Valdo Cruz			2				1		18	2	23
Vinivius T. Freitas	2							2	1	18	23

Figura 5.6: Matriz de Confusão – Classe Economia

AUTORES	ESPORTES										TOTAL DE TEXTOS
	André Ribeiro	Augusto Mafuz	Diogo Olivier	Marcelo Senna	Marcio Bernardes	Sergio Redes	Tostão	Valdir Bicudo	Vicente Datolli	Wianey Carlet	
André Ribeiro	17					4			2		23
Augusto Mafuz		19				2				2	23
Diogo Olivier		1	18	2					2		23
Marcelo Senna	1		2	18				2			23
Marcio Bernardes			2		18				3		23
Sergio Redes			1			17	3			2	23
Tostão		1		2		2	18				23
Valdir Bicudo			3					16		4	23
Vicente Datolli		3			2				18		23
Wianey Carlet						2		2		19	23

Figura 5.7: Matriz de Confusão – Classe Esporte

AUTORES	GASTRONOMIA										TOTAL DE TEXTOS
	Alessandra Blanco	Andrea Kaufmann	Carlos Bertolazzi	Cilmara Castilho	Marcia Daskal	Martha Stewart	Neide Rigo	Nigella Lawson	Ricardo Castilho	Tatiana Damberg	
Alessandra Blanco	18						3		2		23
Andrea Kaufmann		17		2		2		2			23
Carlos Bertolazzi		2	19					1		1	23
Cilmara Castilho	1			17	2		3				23
Marcia Daskal					18			4			22
Martha Stewart		1			1	19		2			23
Neide Rigo				2			17			3	22
Nigella Lawson						2	1	20			23
Ricardo Castilho	1							3	19		23
Tatiana Damberg		1				2			2	18	23

Figura 5.8: Matriz de Confusão – Classe Gastronomia

LITERATURA											
AUTORES	Arnaldo jabor	Cecilia Giannetti	Fernando Monteiro	Laura Medioli	Luiz Bras	Manoel Lobato	Marcelo Coelho	Nelson de Oliveira	Paulo Coelho	Sergio Rodrigues	TOTAL DE TEXTOS
Arnaldo jabor	20				2					1	23
Cecilia Giannetti	2	18				1		2			23
Fernando Monteiro	2		20			1					23
Laura Medioli		3		18				2			23
Luiz Bras		1			18		2		2		23
Manoel Lobato	1		3			17			2		23
Marcelo Coelho			2				18	2		1	23
Nelson de Oliveira	1			1	3			17	1		23
Paulo Coelho				2					21		23
Sergio Rodrigues		2				1				20	23

Figura 5.9: Matriz de Confusão – Classe Literatura

POLITICA											
AUTORES	Acílio L. Rezende	Badger Vicari	Carla Kreeft	Carlos Brickmann	Cluadio Humberto	Claudio Schamis	Fabio Campana	Fabio Campos	Margrit Schimidt	Vittorio Medioli	TOTAL DE TEXTOS
Acílio L. Rezende	18				3				2		23
Badger Vicari	1	19		2				1			23
Carla Kreeft			21						2		23
Carlos Brickmann		1		16		3	2			1	23
Cluadio Humberto			3	2	16					2	23
Claudio Schamis		2				16		2	3		23
Fabio Campana					2		20		1		23
Fabio Campos	2			2		1		18			23
Margrit Schimidt					3		3		17		23
Vittorio Medioli			2							21	23

Figura 5.10: Matriz de Confusão – Classe Política

SAÚDE										
AUTORES	Claudio Lima	Drauzio Varela	Fabio C. dos Santos	Fernanda Aranda	Flavio Settanni	John Cook Lane	Leandro Perché	Leo Kahn	Liliane Ferrari	Loir Carlos da Costa	TOTAL DE TEXTOS
Claudio Lima	20					3					23
Drauzio Varela		18		2				3			23
Fabio C. dos Santo	1	1	19						2		23
Fernanda Aranda				19	2					2	23
Flavio Settanni		2			18	2		1			23
John Cook Lane						20	1			2	23
Leandro Perché		1			3		17	2			23
Leo Kahn				1				21		1	23
Liliane Ferrari							1	1	21		23
Loir Carlos da Costa			1	1					2	19	23

Figura 5.11: Matriz de Confusão – Classe Saúde

TECNOLOGIA										
AUTORES	Alexandre Magalhães	Cezar Taurion	Denny Roger	Eduardo Tude	Ewandro Schenkel	Fernando Birmann	Julio Preus	Marcelo Coutinho	Marcelo Minutti	Patricia Peck	TOTAL DE TEXTOS
Alexandre Magalh	21				1				1		23
Cezar Taurion	1	20		2							23
Denny Roger			20				3				23
Eduardo Tude			2	19					2		23
Ewandro Schenkel	2				18	1	2				23
Fernando Birmann	2			2		19					23
Julio Preus			2		1		17	3			23
Marcelo Coutinho		2					1	19		1	23
Marcelo Minutti		1			4				16	2	23
Patricia Peck			2							21	23

Figura 5.12: Matriz de Confusão – Classe Tecnologia

TURISMO											
AUTORES	Adriano Gambarini	Carlos Sarli	Fabio Zanini	Ivonildo Lavor	José Pinto	Lucia Malla	Raul Lores	Roberto Couto	Roberto Linsker	Rodrigo Baleia	TOTAL DE TEXTOS
Adriano Gambarini	18				3				2		23
Carlos Sarli	2	20				1					23
Fabio Zanini		2	21								23
Ivonildo Lavor				20			2			1	23
José Pinto	2		2		17				2		23
Lucia Malla			1	3		17				2	23
Raul Lores				1	1		18			3	23
Roberto Couto			2		1			20			23
Roberto Linsker								1	20	2	23
Rodrigo Baleia				2						21	23

Figura 5.13: Matriz de Confusão – Classe Turismo

Na Figura 5.14 é mostrada a confusão inter-classes gerada pelos testes, onde é possível observar a relação das classes e suas confusões. Maiores detalhes podem ser vistos na Tabela 5.14.

CONFUSÃO INTER CLASSES											
AUTORES	Assuntos Variados	Direito	Economia	Esportes	Gastronomia	Literatura	Política	Saúde	Tecnologia	Turismo	TOTAL DE TEXTOS
Assuntos Variados	201	10	4	1	1	5	4	2	1	1	230
Direito	7	200	6	1	3	1	8	4			230
Economia	5	5	198	3	2	2	8	4	1	2	230
Esportes	2		2	201	4	4	3	10	3	1	230
Gastronomia	1	4	3	1	206	2	3	5	3	2	230
Literatura	8	2	3	3		206	6	2			230
Política	4	6	11	2	4		201		2		230
Saúde	1	2	3	4	2	2	4	208	2	2	230
Tecnologia	3	4	3	1	1	2	1	1	213	1	230
Turismo		4	1		8	3	1		2	211	230

Figura 5.14: Matriz de Confusão Inter Classes

5.3.4 Resultados

Os resultados de acordo com os protocolos de testes são evidenciados nas Tabelas 5.12, 5.13, 5.14 e 5.15 respectivamente. Na Tabela 5.12 são expostos os resultados da taxa de acerto encontrada, referente à classificação com o uso de todas as características, conforme o protocolo especificado na Tabela 5.9, onde foram realizados duas baterias de testes com o mesmo protocolo: sendo que na primeira foram utilizadas todas as características propostas (Geral); e na segunda, somente com as características selecionadas pelo algoritmo genético (Seleção).

Tabela 5.12: Taxa de Acertos – Modelo Dependente do Autor

CLASSE	GERAL	SELEÇÃO
Assuntos Variados	70,7%	**72,2%**
Direito	72,2%	**74,4%**
Economia	64,8%	**69,1%**
Esportes	68,3%	**69,6%**
Gastronomia	**75,7%**	73,5%
Literatura	72,2%	**76,1%**
Política	68,7%	**75,7%**
Saúde	72,2%	**74,4%**
Tecnologia	73,9%	**78,7%**
Turismo	78,3%	**81,7%**
ACERTO MÉDIO	**71,7%**	**74,5%**

De posse das taxas de acerto, foi possível verificar qual dos dois experimentos gerou os melhores resultados.

Foi possível observar a confusão gerada entre os autores e entre as classes. Verificou-se por exemplo, o quantitativo de votos dentro da classe correta e fora da classe, ou seja, se um texto questionado foi atribuído a um autor de uma outra classe de assunto a qual ele não pertence. Diante disso, expõem-se na Tabela 5.13 os resultados proporcionados pelos testes.

Tabela 5.13: Quantitativo de Votos dentro e fora de cada classe

CLASSES	VOTOS DENTRO DA CLASSE	
	Geral	**Seleção**
Assuntos Variados	84,8%	**87,4%**
Direito	85,7%	**86,7%**
Economia	83,5%	**86,1%**
Esportes	**89,1%**	87,4%
Gastronomia	**93,9%**	89,6%
Literatura	86,5%	**89,6%**
Política	86,1%	**87,4%**
Saúde	**90,9%**	90,4%
Tecnologia	**93,9%**	92,6%
Turismo	**92,6%**	91,7%
ACERTO MÉDIO	**88,7%**	**88,9%**

Observou-se que dois experimentos (Geral e Seleção) os resultados foram muito semelhantes na média final, porém em algumas classes, determinado grupo de características tiveram uma leve vantagem (resultado em negrito) que proporcionou melhores resultados quanto à classificação dos textos nas classes

Analisando os resultados, observou-se que determinados textos pertencentes a uma classe de assuntos, tiveram a classificação de seus textos confundidos com autores de outras classes. Por exemplo, os textos da classe politica tiveram uma maior confusão com os autores da classe assuntos variados, o que representou 5,65% do total, e nenhuma confusão com a classe esportes (Tabela 5.14).

Tabela 5.14: Maiores e Menores Confusões entre Classes

CLASSES	MAIOR CONFUSÃO	MENOR CONFUSÃO
Assuntos Variados	Direito	Gastronomia e Turismo
Direito	Assuntos Variados	Esportes
Economia	Política	Turismo
Esportes	Saúde	Literatura

Gastronomia	Turismo	Literatura e Tecnologia
Literatura	Assuntos Variados	Gastronomia
Política	Literatura	Tecnologia
Saúde	Esportes	Assuntos Variados
Tecnologia	Literatura	Assuntos Variados
Turismo	Direito	Assuntos Variados

As maiores confusões das classes ocorrem entre assuntos que são correlatos e podem utilizar de algumas características que sejam iguais em ambas as classes. Exemplos podem ser vistos na Tabela 5.14, onde a classe de assuntos esportes teve uma maior confusão de textos com a classe saúde, pois tratam intrinsecamente de assuntos que podem ter as mesmas características, tais como: habilidades motoras, atividades físicas, mentais e ligadas a saúde humana. Já a classe economia teve a sua maior confusão com textos da classe política, até porque economia e política estão intimamente ligadas pelo fato de serem assuntos que estão relacionados à administração, negócios e a organização.

Para os experimentos realizados de acordo com o protocolo de testes da Tabela 5.11, foi utilizado o conjunto de características que obteve a melhor taxa de acerto médio apresentado na Tabela 5.12. Na Tabela 5.15 é possível observar os resultados obtidos em cada classe.

Tabela 5.15: Resultados – Modelo Dependente do Autor por Classe

CLASSE	**TAXA DE ACERTO**
Assuntos Variados	77,8%
Direito	77,0%
Economia	78,3%
Esportes	77,4%
Gastronomia	79,1%
Literatura	81,3%
Política	79,1%
Saúde	83,4%
Tecnologia	82,6%
Turismo	83,5%
ACERTO MÉDIO	**80,0%**

5.4 Comparações entre o Modelo Proposto e o Trabalho de Pavelec

Foram realizados testes com o grupo de características proposto por este trabalho, em relação ao trabalho já efetuado por [PAV07]. Nestes experimentos, a base de textos e os protocolos de aprendizagem e testes foram os mesmos utilizados pelo autor em seu trabalho. Diante disso a Tabela 5.16, apresentam os resultados comparativos entre as duas propostas e as duas abordagens apresentados neste trabalho.

Tabela 5.16: Comparativo entre Trabalhos

ATRIBUTOS ESTILOMÉTRICOS	MODELO INDEPENDENTE	MODELO DEPENDENTE
Advérbios e Conjunções [PAV07]	72,5%	83,2%
Modelo Proposto	76,5%	87,0%

Nas duas abordagens utilizadas foi possível observar um ganho de 5% nas taxas de erro, o que demonstra a importância da inclusão dos verbos e dos pronomes, na classe de características linguísticas, no dicionário de atributos estilométricos.

A seleção de características foi relevante na melhoria dos resultados, além de reduzir significativamente o número de atributos utilizados no vetor de características.

Conclusão

A identificação de autoria em documentos questionados de língua portuguesa, através de elementos estilométricos não é uma atividade simples e trivial no processo da análise pericial. Verifica-se que a maioria dos peritos não possui uma metodologia padrão de análise e nem mesmo ferramentas que possam auxiliar na identificação de autores de língua portuguesa. Cabe ressaltar as questões da imprecisão dos métodos linguísticos, que sofrem ainda com a influência demasiada do perito e de sua subjetividade.

O foco deste trabalho é apresentar duas abordagens para atribuição de autoria em documentos questionados de língua portuguesa. As abordagens apresentadas se diferenciam, sendo um o modelo independente do autor, e outro o modelo dependente do autor. Com os experimentos realizados por este trabalho e com base na proposta, relevantes conclusões e contribuições foram observadas, através da criação da base de dados, da análise das abordagens, dos métodos e comparativos mostrados principalmente no Capítulo 5, que são:

- A criação de uma nova base de dados com 3000 textos de 100 autores diferentes, que poderá ser utilizada para realização de outros estudos e experimentos;
- O uso da seleção das melhores características que foram selecionadas através do algoritmo genético, resultou na maioria dos testes em melhores resultados em ambas as abordagens, sendo estes em comparação com o total de características propostas;
- A abordagem dependente do autor apresentou melhores resultados, sendo que nos protocolos de testes efetuados o mesmo se mostrou mais robusto em relação a abordagem independente do autor;
- Aumento da taxa de reconhecimento da base de dados proposta por [PAV07], comparada com o trabalho pioneiro realizado pelo autor, elevando a taxa de acerto médio nos duas abordagens propostas;

- Que cada classe de assunto tem um grupo de características que se sobressai para a identificação. Em alguns casos, os testes revelaram que o uso de um grupo de características (pronomes, advérbios, conjunções e verbos) apresenta melhores resultados quando testadas isoladamente do que em conjunto;
- Existe uma maior confusão entre as classes que são de assuntos correlatos, tais como saúde e esporte, política e economia;
- As abordagens se apresentaram estáveis para o número de amostras e tamanho dos textos.

Destacam-se como possibilidades de melhorias em pesquisas e trabalhos futuros os itens que segue:

- A inclusão de novos grupos de características estilométricas da língua portuguesa;
- Incorporar a classe de características estruturais ao dicionário a fim de avaliar as contribuições dessa classe de atributos no conjunto;
- Criação de um software para verificação de autoria com a escolha de características de língua portuguesa;
- Criação de um corpus de palavras da língua portuguesa para a análise automática da autoria de textos.

Portanto, os objetivos propostos por este trabalho foram cumpridos, pois: foi criada uma nova base de dados para realização de experimentos; foram incluídos nos experimentos dois novos grupos de características da língua portuguesa (verbos e pronomes); a avaliação do conjunto de características e de forma isolada foi efetuada conforme mostram os resultados no Capitulo 5; testes foram realizados utilizando duas abordagens diferentes (dependente e independente do autor); para extração das características do textos da base, foi utilizado um processo automatizado; e, os resultados apresentados por este trabalho podem contribuir com o trabalho efetuado por peritos e linguistas na identificação de autoria.

Referências Bibliográficas

[AC05] ABBASI, A. CHEN, H. **Applying authorship analysis to extremist group web forum messages.** IEEE Intelligent Systems, 20(5):67–75, 2005.

[AIR05] AIRES, R. V. X. **Uso de marcadores estilísticos para a busca na Web em português**. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação –ICMC-USP. São Carlos, 2005.

[ALV97] ALVIM, A. **Manual de Direito Processual Civil**. Volume 2. 6ª ed, São Paulo, RT, 1997, p. 437.

[BAR05] BARANOSKI, F. L. **Verificação de Autoria em Documentos Manuscritos usando SVM**. Dissertação de Mestrado. PUC-PR: Curitiba, 2005

[BER06] BERNARDINI, F. C. **Combinação de classificadores simbólicos utilizando medidas de regras de conhecimento e algoritmos genéticos**. PhD thesis, Instituto de Ciências Matemáticas e de Computação (ICMC). 2006.

[BKOU05] UZUNER, O. KATZ, B. **A comparative study of language models for book and author recognition.** IJCNLP 2005, LNAI 3651, pp. 969–980, 2005.

[CAL99] CALHAU, L. B. **O direito a prova, as provas ilícitas e as novas tecnologias**. Jus Navigandi. Minas Gerais, 1999.

[CMRB05] COUTINHO, B.C. MACEDO, J.L.M. RIQUE JUNIOR, A. BATISTA, L.V. **Atribuição de Autoria usando PPM**. XXV Congresso da Sociedade Brasileira de Computação. III TIL, 2005, pp. 2208-2217.

[COR03] CORNEY, M. W. **Analysing e-mail text authorship for forensic purposes**. Queensland University of Tecnology. Queensland: 2003

[CHA01] CHASKI, Carole E. **Empirical evaluations of language-based author identification techniques.** The International Journal of Speech, Language and Law: Forensic Liguistics, 8(1), 2001.

[CHA05] CHASKI, Carole E. **Who's at the keyboard? - autorship attribution in digital evidence investigations.** International Journal of Digital Evidence, 4(1), 2005. Spring 2005

[CRA98] CRAIN, C. **The bard's fingerprints**. Língua Franca, (4):29–39, 1998.

[CRY00] CRYSTAL, D. **Dicionário de Linguística e Fonética**. 8ª Ed. Rio de Janeiro: Ed. Jorge Zahar, 2000.

[CS01] CRAMMER, K. SINGER, Y. **On the Algorithmic Implementation of Multi-class SVMs**, JMLR, 2001.

[DIE00] DIETTERICH, T. G. **Ensemble methods in machine learning**. Lecture Notes in Computer Science, 1857:1–15, 2000.

[DKLP03] DIEDERICH, J. KINDERMANN, J. LEOPOLD, E. PAASS, G. **Authorship attribution with support vector machines**. Applied Intelligence, (1), 2003.

[FOS96A] FOSTER, D. **A funeral elegy: William Shakespeare's "best-speaking witnesses"**. Publications of the Modern Language Association of America, (111(5)):1080, 1996.

[FOS96B] FOSTER, D. **Primary culprit: An analysis of a novel of politics - who is anonymous?** New York, 26 February, 1996.

[GAM04] GAMON, M. **Linguistic correlates of style: authorship classification with deep linguistic analysis features**. Readmond: 2004

[GIB94] GIBBONS, J. **Language and the Law**. Londres: Longman, 1994.

[GON08] GONÇALVES, D. B. **Agrupamento de Classificadores na Verificação de Assinaturas *Off-Line***. Dissertação de Mestrado. Curitiba, 2008.

[GP05] GOLDSCHMIDT, R. PASSOS, E. **Data Mining – Um Guia Prático**. Rio de Janeiro: Elsevier, 2005.

[GRI07] GRIEVE, J. **Quantitative Authorship Attribution: An Evaluation of Techniques**. Literary an Linguistic Computing, Vol. 22. Nº 3. 251-270, 2007

[HOL85] HOLMES, D. I. **The analysis of literary style — a review.** J. R. Statist. Soc. A. 148, (Part 4):328–341, 1985.

[HFKV99] HOORN, J. FRANK, S. KOWALCZYK, W. VAN DER HAM, F. **Neural network identification of poets using letter sequences.** Literary and Linguistic Computing, (14(3)):311–338, 1999.

[HS90] HANSEN, L. K. SALAMON, P. **Neural network ensembles**. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(10):993–1001, 1990.

[JOA02] JOACHIMS, T**. Optimizing search engines using clickthrough data**. ACM Conference on Knowledge Discovery and Mining (KDD), pages 1–10p, 2002.

[JSB03] JUSTINO, E. J. R. SABOURIN, R, BOTOLOZZI, F. **A Autenticação de Manuscritos Aplicada à Análise Forense de Documentos**. In: TIL - 1º. Workshop em Tecnologia da Informação e Linguagem Humana, 2003, São Carlos. TIL - 1º. Workshop em Tecnologia da Informação e Linguagem Humana, 2003. v. 1. p. 102-106.

[JSCK02] SMITH, J. A. KELLY, C. **Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works.** Computers and the Humanities, (36):411–430, 2002.

[JUS02] JUSTINO, E. J. R. **Análise de Documentos Questionados**. Tese de Doutorado. PUC-PR: Curitiba, 2002.

[KJE94] KJELL, B. **Authorship determination using letter pair frequencies with neural network classifiers**. Literary and Linguistic Computing, (9(2)):119–124, 1994.

[MAL06] MALYUTOV, M.B. **Authorship attribution of texts: a review**. Information Transfer and Combinatorics, LNCS 4123, pp. 362–380, 2006.

[MCM02] MCMENAMIN, Gerard R. **Forensic Linguistics - Advances in Forensic Stylistics**. CRC Press, Florida-USA, 1a edition, 2002.

[MPGR06] MORALES, R. M. C et al. **Authorship attribution using word sequences**: CIARP 2006, LNCS 4225, pp. 844 – 853, 2006.

[MW64] MOSTELLER, F. WALLACE, D. L. **Inference and disputed authorship: The federalist.** Addison-Wesley, Reading, Massachusetts, 1964.

[NAS66] NASCENTES, A. **Tesouro da Fraseologia Brasileira**. 2 ed. Rio de Janeiro: Freitas e Barbosa, 1966.

[OLS04] OLSSON, John. **Forensic Linguistics - An Introducion to Language, Crime and Law**. Continuum, New York-NY, 1a edition, 2004.

[PAV07] PAVELEC, Daniel F. **Identificação da Autoria de Documentos: Análise Estilométrica da Língua Portuguesa usando SVM**. Dissertação de Mestrado, PUC-PR, Curitiba, 2007

[PIR03] PIRES, C. **O Surgimento da Escrita – A Escrita Hieroglífica no Egito**. Revista Temas. Porto, p. 28-32, 2003.

[PJO07] PAVELEC, Daniel F; JUSTINO, E. J. R. ; OLIVEIRA, Luiz E. S de . **Author Identication using Stylometric Features. Inteligencia Artificial**, v. 11, p. 59-66, 2007.

[PJBO08] PAVELEC, Daniel F; JUSTINO, E. J. R. ; BATISTA, Leonardo V.; OLIVEIRA, Luiz E. S. de . **Author Identification using Writer-Dependent and Writer-Independent Strategies.** In: 23th Annual ACM Symposium in Applied Computing (SAC2008), 2008, Fortaleza. Proceedings of the 23th Annual ACM Symposium in Applied Computing, 2008. v. 1. p. 414-418.

[PS06] PAVEAU, A. M. SARFATI, E. **As Grandes Teorias da Linguística: da gramática comparada à pragmática.** Editora Claraluz, 2006.

[RYA06] RYAN, Maria A. **Conjugação dos Verbos em Português – Prático e Eficiente**. 17ª Ed. Ática: São Paulo, 2006.

[SÄR67] SÄRNDAL, C. E. **On deciding cases of disputed authorship. Applied Statistics**, (16):251–268, 1967.

[SIL91] SILVA, C. A. **Ônus e qualidade da prova cível**. Aide: Rio de Janeiro, 1991.

[SIL00] SILVA, C. B. R. **História da Comunicação**. Departamento de Engenharia de informática. Universidade de Coimbra – Portugal, 2001.

[SFC04] SILVA, P. FILHO, N. S. CARVALHO, G. **Vocabulário Jurídico**. 24ª Edição. Forense: Rio de Janeiro, 2004.

[SJBS04] SANTOS, C. R. JUSTINO, E. J. R. BORTOLOZZI, F. and SABOURIN, R. **An offline signature verification method based on the questioned document expert's approach and a neural network classifier**. The Ninth International Workshop on Frontiers in Handwriting Recognition, pages 10–14p, 2004. Tokyo.

[STA08] STAMATOS, E. **A Survey of Modern Authorship Attribution Methods**. University of the Aegean: Greece, 2008

[TE87] THISTED, R. EFRON, B. **Did shakespeare write a newly-discovered poem?** Biometrika, (74(3)):445–455, 1987.

[TTAKG07] TAS, T. GORUR, A. K. **Author Identification for Turkish Texts**. Journal of Arts and Sciences Say›: 7, May›s 2007, pp. 151-160, 2007

[VAP98] VAPNIK, V. **Statistical learning theory**. Wiley, N. Y., page pp. 768,1998.

[VHA04] VAN HALTEREN, H. Linguistic Profiling for Author Recognition and Verification.

[WIL40] WILLIAMS C. B. **A note on the statistical analysis of sentence-length as a criterion of literary style**. Biometrika, 3/4(31):356–361, 1940.

[WIL75] WILLIAMS C. B. **Mendenhall's studies of word-length distribution in the works of shakespeare and bacon**. Biometrika, 1(62):207–212, 1975.

[YTYM02] TSUBOI, Y. MATSUMOTO, Y. **Authorship identification for heterogeneous documents.** Nara (Japan): 2002.

[ZIP75] ZIPF, G. K. **Selected studies of the principle of relative frequency in language**. Harvard University Press, 1975. Cambridge, MA.

[ZQHC06] ZHENG, R. QIN, Y. HUANG, Z and CHEN, H. **A framework for authorship analysis of online messages: Writing-style features and techniques**. Journal of the American Society for Information Science and Technology, (57(3)):378–393, 2006.

Apêndice A

Tabela de Autores da Base de Dados

Autor	Fonte	Assunto/Classe
André Ribeiro	Diário do Grande ABC	Esportes
Augusto Mafuz	O Estado do Paraná	
Diogo Olivier	Zero Hora	
Marcelo Senna	O Extra	
Marcio Bernardes	Diário do Grande ABC	
Sérgio Redes	O Povo	
Tostão	Gazeta do Povo	
Valdir Bicudo	A Gazeta do Paraná	
Vicente Datolli	Jornal de Brasilia	
Wianey Carlet	A Notícia	
Acílio Lara Rezende	O Tempo	Política
Badger Vicari	Jornal de Beltrão	
Carla Kreeft	O Tempo	
Carlos Brickmann	Diário do Grande ABC	
Claudio Humberto	A Gazeta do Acre	
Claudio Schamis	A Gazeta do Paraná	

Fábio Campana	A Gazeta do Paraná	
Fábio Campos	O Povo	
Margrit Schimidt	Jornal de Brasília	
Vittorio Medioli	O Tempo	
Boleslau Sliviany	O Estado do Paraná	Direito
Carlos Zamith Junior	Diário de um Juiz	
Fabio Tokars	O Estado do Paraná	
Fernando Cesar Faria	JusBrasil	
Frederico Vasconcelos	Folha UOL	
Igor Fonseca Rodrigues	Pensando Direito	
Jorge Alberto Araújo	Folha UOL	
Maria Inês Dolci	Folha UOL	
Oscar Ivan Prux	O Estado do Paraná	
Rene Ariel Dotti	JusBrasil	
Ana Cristina Cavalcante	Blogln	Economia
Antonio Pietrobelli	O Estado do Paraná	
Benedicto Dutra	O Gerente	
Claudio Gradilone	IG	
Fernando Canzian	Folha UOL	
Guilherme Barros	IG	
Karlon Aredes	O Tempo	
Luis Nassif	IG	
Valdo Cruz	Folha UOL	
Vinicius Torres Freitas	Fazenda	
Alessandra Blanco	IG	Gastronomia
Andrea Kaufmann	IG	
Carlos Bertolazzi	IG/Cuccina	
Cilmara Castilho	Gazeta do Povo	
Marcia Daskal	IG	
Martha Stewart	IG	
Neide Rigo	O Estadão	

Nigella Lawson	Nigela / IG	
Ricardo Castilho	Blog do Castilho	
Tatiana Damberg	Folha UOL	
Arnaldo Jabor	O Tempo	Literatura
Cecilia Giannetti	Folha UOL	
Fernando Monteiro	Folha UOL	
Laura Mediolli	O Tempo	
Luiz Bras	O Estado do Paraná	
Manoel Lobato	O Tempo	
Marcelo Coelho	Folha UOL	
Nelson de Oliveira	Folha UOL	
Paulo Coelho	Diário do Grande ABC	
Sergio Rodrigues	IG	
Claudio Lima	O Povo	Saúde
Dráuzio Varela	Folha UOL	
Fábio Cesar dos Santos	Diário do Grande ABC	
Fernanda Aranda	IG	
Flávio Settanni	IG	
John Cook Lane	A Hora do Povo	
Leandro Perché	Folha UOL	
Léo Kahn	Diário do Grande ABC	
Liliane Ferrari	LilianeFerrari	
Loir Carlos da Costa	Jornal de Beltrão	
Alexandre Magalhães	Folha UOL	Tecnologia
Cezar Taurion	ComputerWorld	
Denny Roger	Folha UOL	
Eduardo Tude	Folha UOL	
Ewandro Schenkel	Gazeta do Povo	
Fernando Birman	Fbirmam	
Julio Preuss	Folha UOL	
Marcelo Coutinho	Folha UOL	

Marcelo Minutti	IG	
Patricia Peck	Folha UOL	
Adriano Gambarini	National Geographic	Turismo
Carlos Sarli	Revista Trip	
Fábio Zanini	Folha UOL	
Ivonildo Lavor	O Povo	
José Pinto	O Povo	
Lucia Malla	Interney	
Raul Lores	Folha UOL	
Roberto Couto	Gazeta do Povo	
Roberto Linsker	National Geographic	
Rodrigo Baleia	National Geographic	
Fátima Oliveira	O Tempo	Assuntos Variados
Gilberto Dimenstein	Folha UOL	
Gilda de Castro	O Tempo	
Grace Passô	O Tempo	
Luiz Flávio Sapori	O Globo	
Marcelo Rossi	O Tempo	
Oswaldo Braga	O Tempo	
Sebastião Nunes	O Tempo	
Silvana Mascagna	O Tempo	
Trigueirinho	O Tempo	

Apêndice B

Distribuição das Fontes de Dados por Região