TÉCNICO LISBOA

Diz-me o que escreves, dir-te-ei quem és

Processamento de Língua Natural aplicado à literatura

Vanessa Alves Feliciano Baptista

Dissertação para a obtenção de Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur

Prof. João Paulo Baptista de Carvalho

Júri

Presidente: Prof. Nuno João Neves Mamede

Orientador: Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur

Vogal: Prof. Bruno Emanuel da Graça Martins

Junho 2015

Abstract

The author identification tasks of a document have long been the target of the academic community interest.

The basis of this work is a framework developed by Homem and Carvalho [1], based on top-k most frequent words for each author.

Our goal is to evaluate if the use of statistical data for each document and the top-k most frequent words, can improve the existing framework.

In the classification task of the documents it used the Weka.

In addition, we evaluated the impact of excluding stop words from the list of most frequent words.

Finally, the application of the methodology was tested in the task of identifying other author attributes, such as: sex, birth century and decade of birth.

The results suggest that the use of statistical features, together with the top-k-used words, has improved the existing framework. Furthermore, it was observed that remobing stop words of the most frequent words enhances the performance of this methodology.

Finally, it was shown that it is possible to identify the sex of the author of a document and its century of birth. But when trying to identify the decade of birth the results are clearly below.

Keywords: Authorship attribution, features, Weka, stylometrics, stop words

Resumo

As tarefas de identificação do autor de um documento são há muito tempo alvo do interesse da comunidade académica.

A base deste trabalho é uma framework desenvolvida por Homem e Carvalho [1], em que a tarefa de identificar o autor de um documento se baseia nas top-k palavras mais frequentes de cada autor.

O objetivo desta tese é avaliar se a utilização de conjunto de dados estatísticos de cada documento em conjunto com os dados relativos às top-k palavras mais frequentes, pode enriquecer a framework existente.

Na tarefa de classificação dos documentos foi utilizado o Weka.

Para além disso, avaliou-se o impacto da exclusão das Stop Words da lista de palavras mais frequentes.

Os resultados obtidos sugerem que a utilização das features estatísticas, em conjunto com as top-k palavras mais utilizadas, veio enriquecer a framework existente. Além do mais, observou-se que a exclusão de stop Words da lista de palavras mais frequentes aumenta o desempenho desta metodologia.

Finalmente, testou-se a aplicação da metodologia na tarefa de identificar outras características do autor de um documento, tais como: sexo, século de nascimento e década de nascimento.

Demonstrou-se que é possível identificar o sexo do autor de um documento e o seu século de nascimento. Mas, quando se tenta identificar a década de nascimento de um autor os resultados obtidos são francamente inferiores.

Palavras-Chave: Identificação do autor, features, Weka, stylometrics, stop words

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

ARFF	Attribute-Relation File Format
AA	Authorship Attribution
CMC	Cumulative Match Characteristic
FFS	Forward Feature Selection
nAUC	normalized Area Under Curve
SVM	Support Vector Machines

1. Introdução

1.1. Motivação e Objetivos

Historicamente, existem muitos textos cujo autor permanece desconhecido. Considere-se como exemplo os "Federalist Papers" [2] [3] [4]. "The Federalist Papers" são um conjunto de 77 ensaios políticos, escritos por Alexander Hamilton, John Jay e James Madison, que argumentam para a ratificação da Constituição dos Estados Unidos da América. Desses textos, 5 são atribuídos a John Jay, 43 a Alexander Hamilton, 14 a James Madison, 3 foram escritos em parceria e 12 desconhece-se o autor (será Alexander Hamilton ou James Madison).

Contudo, também existem documentos recentes cuja autoria é ambígua. Tal, pode dever-se a fatores como: reprodução manual de documentos, prestígio ou pressão social e/ou política [2].

A identificação do autor de um texto é uma área de interesse para a comunidade académica, contudo tem aplicações noutras áreas de interesse. Por exemplo, em Direito na ausência de outros meios de prova, pode inferir-se o autor de um documento (tal como uma carta anónima) [5].

O ponto de partida deste trabalho é um projeto desenvolvido por Homem e Carvalho [1]. Nesse projeto, os autores identificavam autores de textos jornalísticos, utilizando as palavras mais utilizadas de cada autor.

Os objetivos desta tese são:

- Estender a framework desenvolvida por Homem e Carvalho [1]
- Identificar o autor de um documento
- Identificar o sexo do autor de um documento
- Identificar o século em que nasceu o autor de um documento
- Identificar a década em que nasceu o autor de um documento

1.2. Contribuições

Com este trabalho pretende-se criar uma metodologia para identificar o autor de um documento existente, estendendo a framework existente e melhorando a sua performance. Para isso, pretende-se identificar um conjunto adicional de características que, em conjunto com as palavras mais frequentes, podem ser utilizados na tarefa de identificação de um documento.

As contribuições deste trabalho são:

1. Recolha de textos para análise (utilizando Projeto Gutenberg)
2. Análise do trabalho anterior e extensão do mesmo para análise literária

3. Levantamento do estado da arte relativo a esta temática
4. Proposta de arquitetura para análise literária e dos parâmetros a avaliar
5. Implementação do sistema
6. Avaliação do sistema
7. Testes
8. Análise dos testes efetuados

1.3. Outline

No Capítulo 2 (Trabalho Relacionado) são apresentados os conceitos teóricos em que este trabalho se baseia, alguns trabalhos relevantes e os recursos utilizados.

No Capítulo 3 (Solução Proposta) é apresentado o modelo utilizado nesta tese para caracterizar um documento, a implementação desta tese e a arquitetura a que chegámos.

No Capítulo 4 (Avaliação) será apresentada a metodologia utilizada para avaliar o sistema proposto.

No Capítulo 5 (Conclusão e Trabalho Futuro) serão descritas as conclusões obtidas e apresentaremos algumas propostas de trabalho futuro.

2. Trabalho Relacionado

A identificação do autor do texto é uma área de investigação com um longo historial, que evoluiu substancialmente nos últimos anos [1].Esta evolução deve-se sobretudo à introdução de novas técnicas de análise estatística, no século XXI [5] [3]. Para melhor ilustrar a evolução do interesse existente nesta área de investigação, basta observar que uma pesquisa por "authorship attribution" no Google obtinha 10.700 hits (em 2 de Junho de 2005) [5] e que atualmente obtém 1.670.000 hits[1]. Contudo, apesar desta evolução, esta ainda não é uma área completamente aceite, o que se pode dever sobretudo à falta de perceção da sua aplicabilidade e à complexidade matemática das técnicas utilizadas [5].

A análise de um documento pode ser feita subjetiva ou estatisticamente. Numa análise subjetiva, o documento é avaliado por peritos com base no seu estilo literário [3]. Uma vez que o estilo literário não é quantificável, com uma análise subjetiva raramente se obtém um resultado unanimemente aceite pela comunidade. Na análise estatística são utilizados métodos quantitativos, que visam estudar o comportamento linguístico de um autor, obtendo resultados objetivos [2] [3].

Esta tarefa pode ser executada sobre um texto manuscrito ou sobre um texto e formato digital. No caso de se utilizar texto manuscrito, pode analisar-se os aspetos gráficos do documento, como por exemplo, a grafia do seu autor (reconhecimento de escrita). No caso de se utilizar um texto em formato digital, não é possível analisar estas características. Como tal, é necessário avaliar a existência de um padrão linguístico entre documentos [6].

Na primeira parte deste capítulo serão introduzidos os conceitos teóricos importantes para a compreensão desta tese, assim como os recursos utilizados.

Na segunda secção, serão apresentadas diversas ferramentas que podem ser utilizadas para inferir o autor de um documento.

2.1. Conceitos Fundamentais

Na literatura aceita-se que um texto possa revelar dados acerca do seu autor, tais como: idade, género e personalidade [7]. A tarefa de inferir a autoria de um documento é uma das aplicações de *stylometrics*. Ao utilizar-se *stylometrics* para esta tarefa, estamos a assumir que cada autor possui um estilo próprio (e único) de escrita, persistente ao longo do tempo [5] [2].

Esta tarefa pode dividir-se em três tarefas distintas [8]:

- **Caracterização do autor**: Identificação do conjunto de características que definem um autor. As características do autor são inferidas utilizando um conjunto de textos previamente identificados.

[1] Pesquisa efetuada a 5 de Novembro de 2014

- **Identificação do autor**: Pretende-se determinar o autor de um texto, tendo como base um conjunto de características do texto e do autor. Esta tarefa pode ser útil quando várias pessoas reclamam a autoria de um documento, quando ninguém quer aceitar a autoria do mesmo ou quando não é possível determinar o autor de um texto.
- **Deteção de semelhança**: Avaliação do grau de semelhança entre dois textos. Neste caso, pode não ser necessário determinar os autores dos mesmos. As técnicas de deteção de semelhança podem ser úteis para detetar plágio.

Ao longo do tempo, foram apresentadas diversas características a utilizar na análise estatística de um documento. A principal dificuldade ao caracterizar um autor é definir o conjunto de características que descreve melhor o seu estilo. Bailey (1979) descreve as características desejáveis para essas variáveis: "*They should be salient, structural, frequent, and relatively immune from conscious control*" [2].

As características podem ser divididas em cinco categorias [8] [9] [10]:

- **Lexicais** - baseadas em estatísticas de palavras ou caracteres
- **Estruturais** - tendo em conta o *layout* do documento
- **Sintáticas** - tendo como base a pontuação e o tipo de palavras utilizadas
- **Específicas do conteúdo** - frases e *keywords* com particular importância para um dado domínio
- **Idiossincráticas** - anomalias existentes, tais como, erros ortográficos e gramaticais

Na tabela 1- Exemplo de características a analisar, é apresentada uma lista de diversas características que podem[2] ser analisadas [2] [1] [10] [3].

Categoria	Característica
Lexicais	# Letras
	# Palavras
	# Palavras por frase
	# Números
	# Frases
	# Sinais de pontuação
	# Reticências
	% Maiúsculas/minúsculas
	% Vogais/consoantes por palavra/frase/texto
	% Sequências de caracteres não alfanuméricos
	Média de caracteres por palavra
	Média de palavras por frase
	Frequência de cada palavra (normalizada)
	Riqueza de vocabulário (rácio de palavras distintas)
	Frequência de cada n-grama
Estruturais	# Parágrafos
	# Títulos

[2] Se considerarmos a escrita informal atual, esta característica pode ser útil para analisar a utilização de *emoticons*, como por exemplo, :-) ou ;)

	Tipo da fonte
	Tamanho da fonte
	# *Links*
	# Imagens
Sintáticas	% Nomes
	% Adjetivos
	% Verbos
	% Advérbios
	% Nomes próprios
	% Estrangeirismos
	% Abreviaturas
Específicas do conteúdo	Palavras importantes para o domínio
	Frases importantes para o domínio
Idiossincráticas	% Erros ortográficos
	% Erros gramaticais
	Outras anomalias do texto

1- Exemplo de características a analisar

Geralmente, o processo de identificação do autor de um documento pode dividir-se em três fases [11]:

1. Obtenção de um conjunto de treino
2. Extração de *features* do conjunto de treino
3. Classificação de um documento, de acordo com as features extraídas

Uma desvantagem da utilização de stylometrics para identificar o autor de um documento é a possibilidade de obter resultados enviesados, sob determinadas circunstancias. Alguns dos fatores que podem influenciar os resultados obtidos são [5] [11]:

1. Alteração do estilo do autor ao longo do tempo
2. Textos com muitas citações
3. Coautoria de um documento
4. Tipo de documento (por exemplo: livro, artigo jornalístico ou post online)

Para identificar o autor de um documento, deve-se extrair um conjunto de características que identifiquem inequivocamente o autor do mesmo.

A linguística, isto é, o estudo da linguagem humana pode dividir-se nas seguintes áreas:

- Estilística – variações de linguagem dentro de um dado contexto
- Sintaxe – disposição das palavras dentro de uma frase
- Lexicologia – estudo das palavras de um idioma
- Morfologia – estudo da estrutura interna das palavras
- Gramática – conjunto e regras estruturais que definem a composição das frases

Cada uma das áreas da linguística pode contribuir para a tarefa de identificar o autor de um documento, apresentando diferentes vantagens e desvantagens, dependendo do objetivo do trabalho.

A estilística define atributos que dependem do tema presente no documento; logo, poderá ser difícil distinguir entre autores que se debrucem sobre o mesmo tema.

A lexicologia pode ser útil para identificar a riqueza de vocabulário do autor; contudo, poderá ser difícil distinguir autores com riqueza de vocabulário semelhantes.

Os atributos morfológicos poderão ajudar a identificar diferentes formas de escrever, sobretudo em documentos informais, tais como, e-mails; mas, em textos formais essas características não estão presentes.

A sintaxe e a gramática possuem um conjunto de características que poderão ser úteis para identificar o autor de um documento, independentemente do seu tema e tipo.

2.2. Projeto Base: Authorship Identification and Author Fuzzy "Fingerprints"

Em [1], Homem e Carvalho apresentam a ferramenta de deteção de autor, por si desenvolvida.

Nesse trabalho pretende-se extrair a impressão digital de um conjunto de documentos e utilizá-la para identificar o autor de um documento distinto.

Cada autor é caracterizado pela sua impressão digital. Uma impressão digital pretende representar informação muito complexa de forma compacta (e única), da mesma forma que uma impressão digital humana identifica uma pessoa.

Em ciências da computação, as impressões digitais podem ser utilizadas para evitar a comparação/transmissão de dados. Por exemplo, para verificar se um ficheiro foi alterado bastaria comparar a sua impressão digital (caso esta seja igual, significa que o ficheiro está igual).

No contexto da identificação do autor de um texto, uma impressão digital pretende capturar as características do mesmo. Neste caso, pretende-se que a probabilidade de dois autores distintos terem a mesma impressão digital seja baixa. Para além disso, esta deve ser robusta. Ou seja, deve conseguir identificar o autor de um texto, mesmo que este mude ligeiramente o seu estilo.

Uma impressão digital deve ter as seguintes características:

- Incluir o mínimo de características que consegue descrever o autor
- Permitir ser atualizada, caso surjam novos textos do autor
- Permitir uma rápida comparação
- Ser escalável, isto é, a sua qualidade deve manter-se quando houver mais textos do autor
- Ser flexível, ou seja, deve poder ser aplicada a novos autores

Neste método, as frequências de cada palavra são utilizadas para obter a informação que representa dado autor. Posteriormente, essa informação será utilizada para identificar o autor de outro texto.

O trabalho de Homem e Carvalho [1] utiliza o modelo *bag-of-words*, ou seja, assume-se que a ordem pela qual as palavras surgem em dado documento é irrelevante.

O primeiro passo é calcular as top-k palavras mais frequentes, para todos os textos cujo autor esteja devidamente identificado. Para calcular as top-k palavras mais frequentes, foi utilizado o algoritmo Filtered Space-Saving, por uma questão de performance. Segundo o autor, esta aproximação não deverá comprometer os resultados obtidos, já que deverá ser introduzido algum nível de aleatoriedade no modelo de identificação de autores.

As top-k palavras mais frequentes são depois utilizadas para criar uma impressão digital.

Segundo Homem e Carvalho [1], a impressão digital criada é comparada com as impressões digitais de cada autor. O documento é atribuído ao autor que tenha a impressão digital mais semelhante com a do documento em análise.

O algoritmo implementado pressupõe que cada autor é suficientemente estável para que se possa extrair e comparar features. Pressupõe também que as top-k palavras mais frequentes de cada autor possuem essa estabilidade.

Este trabalho foi aplicado a um conjunto de artigos publicados no jornal "Público". Os artigos foram processados cronologicamente e divididos em dois conjuntos, com metade dos artigos de cada autor. Um dos blocos foi utilizado para calcular as impressões digitais e o outro foi utilizado para testar o algoritmo criado.

Homem e Carvalho [1] testaram o seu modelo utilizando vários algoritmos e obtendo uma taxa de sucesso perto dos 60%.

Para este trabalho, os autores desenvolveram um classificador. Esta ferramenta, desenvolvida com recurso a Java, estava preparada para ler um conjunto de documentos a partir de uma Base de Dados SQL.

2.3. Projetos Relevantes

2.3.1. Conversationally-inspired Stylometric Features for Authorship Attribution in Instant Messaging

No trabalho apresentado em [12] o autor tenta aplicar as técnicas de Authorship Attribution (AA) para identificar os intervenientes em conversas decorridas num *chat* e propõe melhorias ao métodos tradicionais de AA que visam melhorar a sua performance para este tipo de dados.

Anteriormente, já haviam sido utilizadas features baseadas em stylometrics para categorizar o conteúdo de um *chat* e/ou o comportamento dos participantes. Contudo, havia poucas tentativas de identificar os seus participantes. Para além disso, não tinham sido consideradas as características inerentes a este tipo de dados. Neste trabalho, o autor propõe um conjunto de features que têm em consideração estas características.

Foram utilizadas 77 conversas, em que cada conversa é modelada como uma sequência de "turnos".

Oralmente, um turno corresponde a um intervalo de tempo em que um dos interlocutores fala. Paralelamente, numa conversa de *chat*, um turno corresponde a um conjunto de símbolos e/ou palavras escritas consecutivamente (isto é, sem que o seu autor tenha sido interrompido pelo interlocutor).

Estas conversas foram realizadas de forma espontânea, isto é, sem que o seu propósito inicial fosse a recolha de dados. Desta forma, evita-se o enviesamento dos dados devido à mudança de estilo de escrita dos interlocutores.

Neste trabalho, as features são extraídas por turno, contrariamente aos métodos tradicionais de AA, em que são extraídas por documento.

A introdução do conceito de "turno" permitiu criar novas features, tais como:

- Duração do turno
- Velocidade de escrita
- Número de caracteres "*return*"
- Mimetismo (rácio entre o número de palavras no turno atual e o número de palavras no turno anterior)

Neste trabalho, as features foram extraídas a partir dos turnos individuais e não a partir de toda a conversa.

Devido a questões éticas e de privacidade apenas foram utilizadas features que não envolvam o conteúdo das conversas. Assim sendo, as features utilizadas neste trabalho foram:

Feature	Intervalo
# Palavras	[0,260]
# Emoticons	[0,40]
# Emoticons por palavra	[0,1]
# Emoticons por caracter	[0,0.5]
# Pontos de exclamação	[0,12]
# Pontos de interrogação	[0,406]
# Caracteres	[0,1318]
Tamanho médio das palavras	[0,20]
# Reticências	[0,34]
# Maiúsculas	[0,94]
# Maiúsculas / # Palavras	[0,290]

Duração do turno	[0,1800] (segundos)
# Caracteres "*return*"	[1,20]
# Caracteres por segundo	[0,20] (caracteres/segundo)
# Palavras por segundo	[0,260]
Nível de mimetismo	[0,1115]

2 - Features utilizadas neste trabalho

Uma vez que o número de turnos variava entre 60 e 100, foram tidos em conta 60 turnos por pessoa. Deste modo, evita-se que os resultados fossem enviesado devido à existência de diferentes quantidades de dados.

Dos 60 turnos escolhidos por pessoa, estes foram divididos em conjunto de testes e de treino, cada um com 30 turnos.

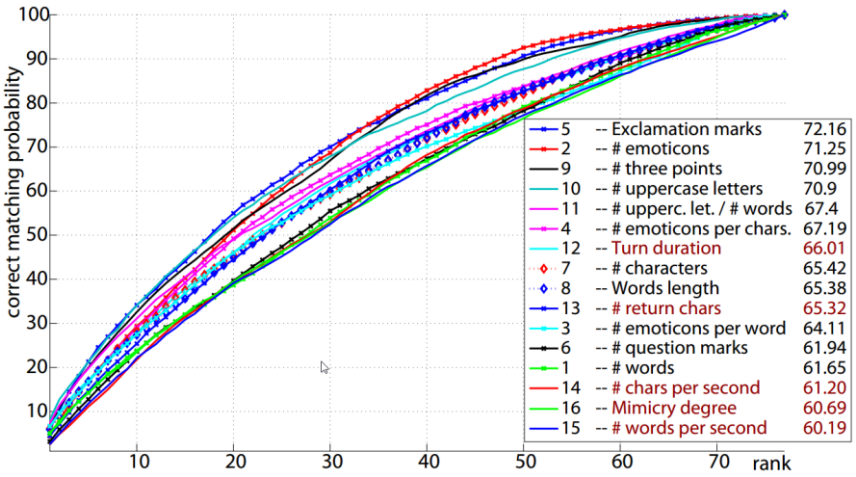
Os testes efetuados começaram por ter em conta apenas uma feature de cada vez. Neste passo, uma feature de uma pessoa identificada no conjunto de testes foi escolhida e comparada com a mesma feature para todas as pessoas no conjunto de treino, utilizando uma métrica adequada (por exemplo, a distância Euclidiana). Utilizando os resultados obtidos neste passo podemos criar uma matriz N x N com a distância calculada, em que N representa o total de pessoas.

De seguida, as distâncias para cada elemento no conjunto de testes foram ordenadas por ordem crescente e calculada a curva CMC (*Cumulative Match Characteristic*).

A curva CMC é utilizada para medir a performance de sistemas de identificação [13] e representa a probabilidade de encontrar a correspondência correta nas primeiras *n* posições do ranking. A posição 1 de uma curva CMC representa a probabilidade de obter a correspondência correta para cada pessoa.

Após calcular a curva CMC para cada feature, a área abaixo da curva (nAUC) pode ser utilizada como uma medida de performance.

Na figura 1- CMCs para as features proposta [12] podemos observar que a performance de cada feature é baixa (abaixo dos 10% para a posição 1 da curva). O número à direita representa a nAUC.



1- CMCs para as features propostas

Estas experiências serviram de base para o passo seguinte: aplicar a estratégia FFS (Forward Feature Selection).

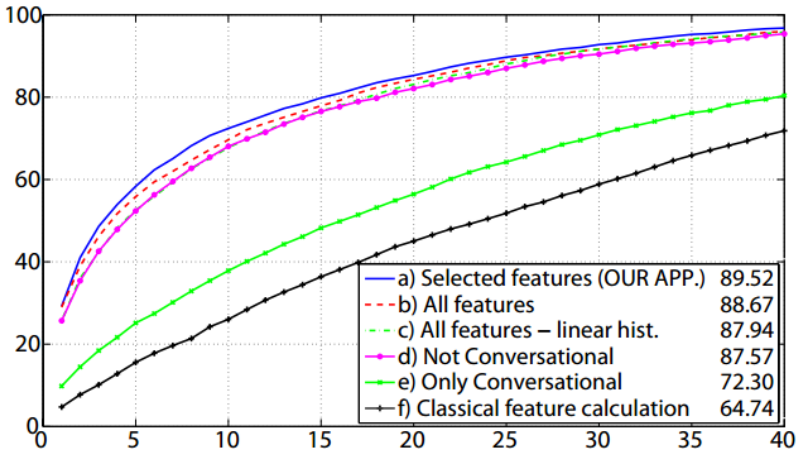
Nesta estratégia, na primeira iteração seleciona-se a feature com maior nAUC e, de seguida, seleciona-se a feature que, combinada com a primeira, dá uma maior nAUC. Aplica-se o mesmo raciocínio até que a área abaixo da curva decresça.

Deste processo resultou a seguinte lista de features:

1. #Pontos de Exclamação
2. #emoticons
3. #Reticências
4. #Maiúsculas
5. Duração do turno
6. #Caracteres return
7. Tamanho das palavras
8. #Caracteres por segundo
9. #Pontos de Interrogação
10. #Caracteres
11. Nível de mimetismo
12. #Palavras por segundo.

Observou-se ainda que algumas destas features têm em conta a natureza das conversas pelo que, apesar de individualmente terem uma performance baixa, complementam a informação comparativamente aos métodos tradicionais de AA.

Finalmente, analisando a curva CMC para estas features podemos observar que houve uma melhoria da performance, obtendo 29,2% na posição 1 da curva. A Figura 2 [12] representa os resultados obtidos para diferentes conjuntos de features.



2 - Comparação entre diferentes conjuntos de features

2.3.2. Selecting Syntactic Attributes for Authorship Attribution

Em [6] o autor começa por distinguir os desafios entre a identificação do autor de um documento em formato manuscrito ou digital. No caso de ser um documento manuscrito, podem analisar-se as características gráficas do texto. Contudo, este trabalho debruça-se sobre os documentos em formato digital, utilizando atributos sintáticos e gramaticais. A principal dificuldade da utilização de atributos sintáticos e gramáticas reside principalmente no enorme número de features disponível. Este trabalho propõe uma metodologia que pretende identificar as features mais significativas para o idioma português.

Inicialmente, parte-se de um conjunto de 408 palavras composto por atributos sintáticos variáveis (verbos e nomes) e invariáveis (conjunções e advérbios).

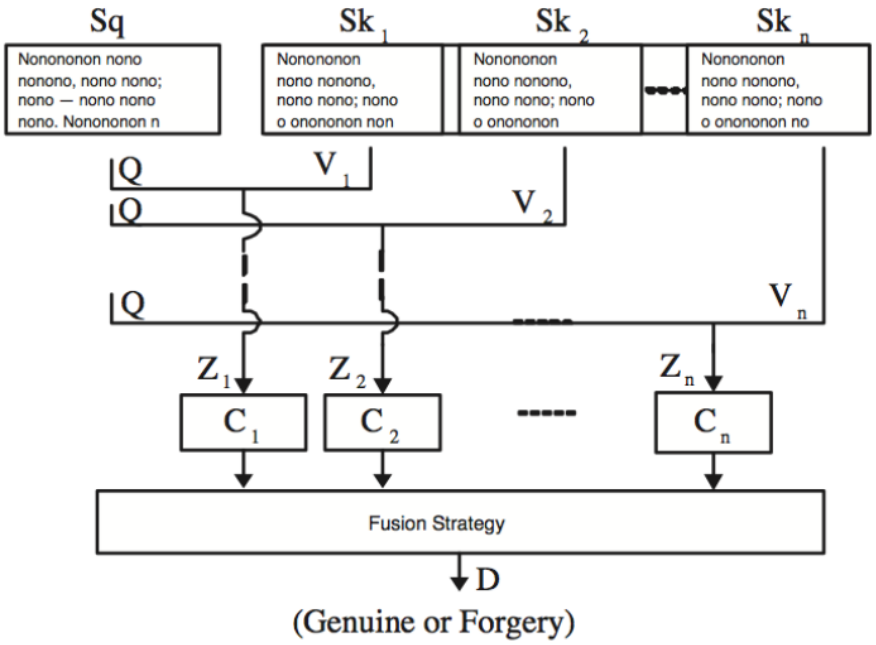
Normalmente, para identificar o autor de um documento utilizam-se modelos específicos do autor. Nestes modelos são utilizadas duas classes ω_1 e ω_2, em que ω_1 representa o nível de semelhança e ω_2 o nível de diferença. A principal desvantagem desta abordagem é a necessidade de aprender o modelo sempre que é incluído um novo autor no sistema. Para além disso, é necessário um grande número de documentos genuínos para construir um modelo fiável.

Alternativamente, pode utilizar-se um modelo independente ao autor, em que o número de modelos é independente do número de autores. Uma vez que este é um modelo global pode construir-se um sistema mais robusto, mesmo que haja poucos documentos genuínos de cada autor.

Esta abordagem classifica os documentos em duas categorias: genuíno ou falsificação, utilizando o modelo global. Para isso, utiliza um conjunto de *n* documentos devidamente identificados e compara o novo documento com cada um destes. O objetivo é identificar as discrepâncias entre cada um dos documentos e o documento a identificar.

Sendo V_i o vetor de características extraído de cada um dos documentos genuínos e Q o vetor de características extraído do novo documento, calcula-se a medida de dissimilaridade $Z_i = \|V_i - Q\|_2$,

originando *n* decisões parciais (C). A decisão final (D) depende da combinação de todas as decisões parciais, tal como se pode observar na figura 3, retirada de [6].



3 - Arquitetura da solução

Do ponto de vista prático o objetivo deste trabalho pode ser encarado como uma otimização multicritério. Para este cenário utilizaram algoritmos genéticos que, devido às suas características, têm um bom desempenho em pesquisas em espaços não lineares e pouco compreendidos. Um problema de otimização multicritério consiste num determinado número de objetivos e respetivos constrangimentos. Originalmente, este problema era resolvido com recurso a uma soma ponderada. Isto é, todos os objetivos eram agregados num único valor. Contudo, o processo de agregação dos objetivos poderá passar por uma normalização dos mesmos.

Neste trabalho foram utilizados documentos de 100 autores diferentes, distribuídos por 10 temas: Direito, Economia, Desporto, Gastronomia, Literatura, Política, Saúde, Tecnologia, Turismo e Diversos. Os documentos foram recolhidos de 15 jornais brasileiros. Por cada autor foram recolhidos 30 documentos, sendo que, em média, cada artigo tem 600 palavras. Foi ainda referido que o processo de revisão dos artigos poderá descaracterizar os mesmos.

Neste trabalho foram escolhidos 4 tipos de features: conjunções, advérbios, verbos e pronomes.

As conjunções servem para palavras e frases. Estas podem ser utilizadas de diferentes formas, sem que alterem o significado do texto. Considere-se como exemplo a frase “Ele é *tal qual* seu pai”. Facilmente podemos observar que frases semelhantes, tais como: “Ele é *tal e qual* seu pai” ou “Ele é *tal como* seu pai” apesar de estarem escritas de forma distinta encerram o mesmo significado, pelo que, a utilização de uma ou outra conjunção poderá representar o estilo de escrita do autor. As conjunções utilizadas neste trabalho podem ser consultadas na tabela 3.

Grupo	Conjunções
Coordenativas copulativas	e, nem, mas também, senão também, bem como, como também, mais ainda
Coordenativas adversativas	porém, todavia, mas, ao passo que, senão, entretanto, não obstante, apesar disso, em todo o caso, contudo, no entanto
Coordenativas conclusivas	logo, portanto, por isso, por conseguinte
Coordenativas explicativas	porquanto, que, porque
Subordinativas comparativas	tal qual, tais quais, assim como, tal e qual, tão como, tais como, mais do que, tanto como, menos do que, que nem, tanto quanto, o mesmo que, tal como, mais que
Subordinativas conformativas	consoante, segundo, conforme
Subordinativas concessivas	embora, ainda que, ainda quando, posto que, nem que, por muito que, e bem que, por menos de, dado que, mesmo que, por mais que
Subordinativas condicionais	se, caso, contanto que, salvo que, a não ser que, a menos que
Subordinativas consecutivas	de sorte que, de forma que, de maneira que, de modo que, sem que
Subordinativas finais	para que, fim de que
Subordinativas proporcionais	a proporção que, quanto menos, quanto mais, a medida que

3 - Conjunções utilizadas

Um advérbio serve para modificar um verbo, um adjetivo, uma frase ou mesmo outro advérbio e procuram responder a questões como: “como?”, “quando?”, “onde?” ou “quanto?”. Os advérbios utilizados neste trabalho podem ser consultados na tabela 4.

Grupo	Advérbios (em português)
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

Intensidade	ainda, apenas, de pouco, demais, mais, menos, muito, pouca, pouco, quase, tanta, tanto
Negação	absolutamente, de jeito nenhum, de modo algum, não tampouco
Subordinada concessiva	embora, ainda que, ainda quando, posto que, por muito que, se bem que, por menos que, nem que, dado que, mesmo que, por mais que
Quantidade	todo, toda
Modo	assim, depressa, bem, devagar, face a face, algo, facilmente, frente a frente, lentamente, mal, rapidamente, alguém, algum, alguma, bastante, cada, certa, certo, muita, nada, nenhum, nenhuma, ninguém, outra, outrem, outro, quaisquer, qualquer, tudo

4 - Advérbios utilizados

Relativamente aos verbos, o autor recorreu aos 50 verbos mais utilizados no português do Brasil nas formas: infinitivo, gerúndio e particípio passado. Os verbos utilizados estão representados na tabela 5.

Verbos
escrever, falar, jogar, andar, ver, ser, cantar, pular, ler, ter, achar, colar, estar, dizer, dar, escolher, fechar, entender, fazer, trocar, abrir, acabar, declarar, completar, visitar, encerrar, comer, beber, pensar, possuir, atingir, melhorar, achar, realizar, haver, viver, aplicar, gerar, melhorar, pagar, distribuir, ligar, usar, projetar, desenvolver, poder, implantar, trazer, iniciar, efetuar

5 - Verbos utilizados

Finalmente, foram utilizados 87 pronomes que podem ser consultados na tabela 6.

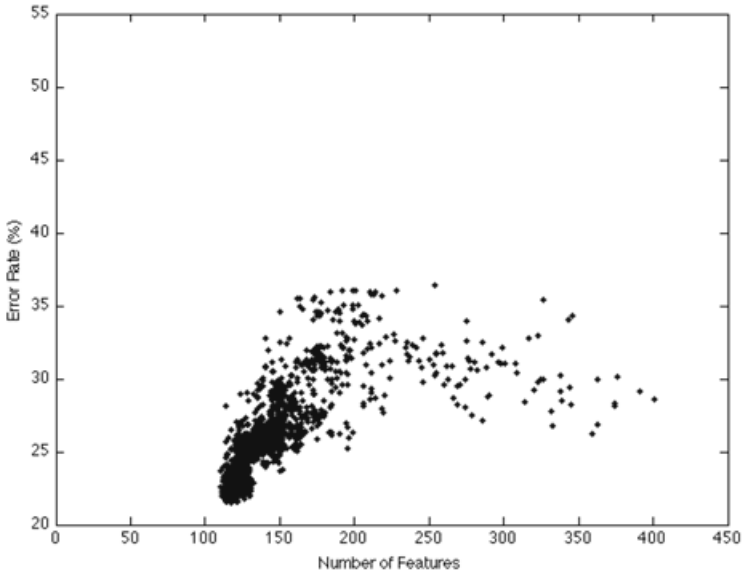
Grupo	**Pronomes**
Relativos	quem, o qual, a qual, os quais, as quais, onde, em que, quanto, quanta, quantos, quantas, cujo, cuja, cujos, cujas
Possessivos	meu, minha, meus, minhas, teu, tua, teus, tuas, seu, sua, seus, suas, nosso, nossa, nossos, nossas, vosso, vossa, vossos, vossas
Demonstrativos	este, esta, estes, estas, isto, esse, esses, essa, essas, isso, aquele, aquela, aqueles, aquelas,

	aquilo, nessa, desta, daquela, cujo, cuja, cujos, cujas
Pessoal Subjetivo	eu, tu, ele, nós, vós, eles, me, te, se, lhe, o, a, nos, vos, lhes, os, as, mim, comigo, connosco, ti, contigo, convosco, si, consigo
Pessoal Objetivo	você, vocês, senhor, senhores, senhora, senhoras, senhorita, senhoritas, vossa senhoria, vossas senhorias

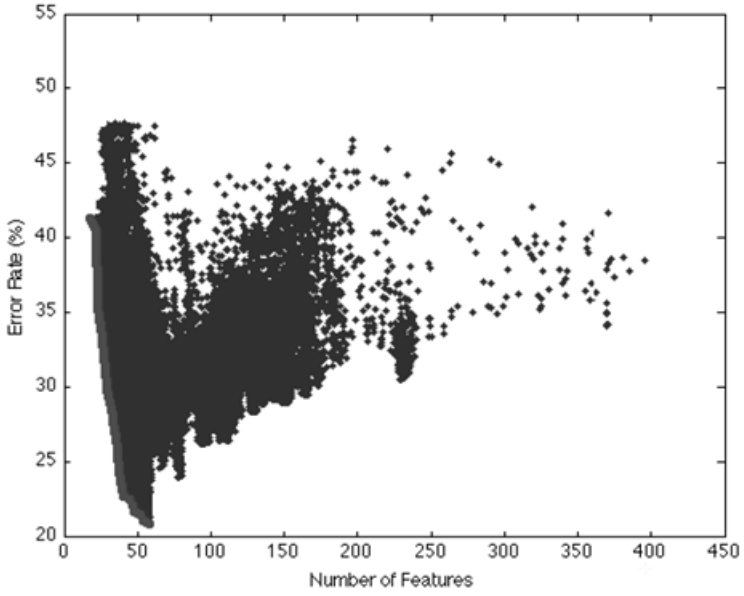
6 - Pronomes utilizados

No processo de testes, o autor utilizou um classificador SVM, testando-o com todas as features e também com os grupos de features mencionados (conjunções, advérbios, verbos e pronomes).

Além disso, comparou-se o desempenho do algoritmo utilizando uma soma ponderada e um algoritmo genético. Nestas experiências, observou-se que, no caso de uma soma ponderada, a performance é melhor com cerca de 100 features, tal como está representado na Figura 4 [6]. Paralelamente, no caso da utilização de um algoritmo genético, a performance é melhor com cerca de 50 features (ver Figura 5 [6]).



4 - Performance com uma soma ponderada



5 - Performance utilizando um algoritmo genético

De seguida, foi utilizado apenas o subconjunto de features com melhor performance. Desta forma, conseguiu-se diminuir a taxa de erro de cerca de 42% para 26%. O que provou a eficiência de escolher apenas um subconjunto de features.

Na tabela 7 pode-se observar as features selecionadas neste trabalho.

Grupo	Quantidade	Features
Advérbios	22	lá, dentro, adiante, em cima, ao lado, depois, sempre, com certeza, sem dúvida, ainda, quase, apenas, mais, todo, toda, bastante, nada, ninguém, nenhum, antes, qualquer, outro
Conjunções	11	porém, por isso, assim como, que nem, segundo, embora, portanto, tais como, contanto que, de modo que, caso
Pronomes	10	seu, sua, quem, cujo, este, esta, o, a, aquele, onde
Verbos	15	ser, ver, pular, estar, ligar, estando, efetuando, fazendo, tendo, sendo, usando, pagando, aberto, visto, usado

7 - Features utilizadas

Analisando a lista de features utilizadas pode ainda concluir-se que alguns subgrupos de features nunca são selecionados: conjunções coordenativas copulativas, conjunções coordenativas explicativas, conjunções subordinativas finais, conjunções subordinativas proporcionais, pronomes pessoais e advérbios de negação.

Devido ao elevado número de autores (100) o autor deste trabalho optou por criar uma matriz de confusão com base no tema do documento. Analisando esta matriz (tabela 8), podemos concluir que a taxa de sucesso deste algoritmo ronda os 86%, sendo que a categoria com pior performance é "Diversos", tal como seria expectável.

	a. Diversos	b. Direito	c. Economia	d. Desporto	e. Gastronomia	f. Literatura	g. Política	h. Saúde	i. Tecnologia	J. Turismo
a.	82	7	1	2	1	2	1	2		1
b.	5	84	3		1	2	2	1		
c.	3	3	84			1	4	2		
d.	2		1	86	1	1	1	7	1	
e.		1		1	87	2	1	3		3
f.	4	3	2	1		87	3			
g.	1	1	6			3	88			
h.	1		2	4	3		1	88		2
i.	3	3	3			1	1		89	1
j.	1	1	2		6	1				89

8 - Matriz de confusão por temas

2.4. Recursos Utilizados

2.4.1. Projeto Gutenberg

O Projeto Gutenberg é uma biblioteca digital, fundada em 1971 por Michael Hart. Michael Hart batizou o projeto em honra a Johannes Gutenberg, um impressor alemão que impulsionou a prensa móvel.

A maior parte dos documentos disponibilizados são textos completos de livros em domínio público.

O objetivo do projeto é tornar estes documentos livres, publicando os documentos em formatos duradouros e abertos.

Inicialmente, todos os textos eram introduzidos manualmente. Contudo, com a massificação dos digitalizadores e do *software* de reconhecimento ótico de caracteres, a digitalização dos livros tornou-se um processo facilmente exequível.

Mais tarde, foram criados projetos filiados que correspondem a organizações independentes que partilham os ideais do Projeto Gutenberg e que são autorizados a utilizar a sua marca registada.

No ano 2000 foi criada uma organização sem fins lucrativos, que gere o projeto. Nesta altura, foi também criado um projeto que fez com que a revisão dos livros fosse feita por voluntários, através da Internet, de forma distribuída. Este projeto permitiu um grande aumento do número de livros disponibilizado no Projeto Gutenberg

Um dos principais objetivos do Projeto Gutenberg é encorajar a livre reprodução e distribuição dos textos publicados.

2.4.2. Weka

O Weka é uma ferramenta de *Data Mining* desenvolvida pelo grupo de *Machine Learning* da Universidade de Waikato, na Nova Zelândia.

Esta ferramenta começou a ser desenvolvida em 1993, utilizando Java. Este é um *software* livre, ao abrigo da licença *GNU General Public License*.

Neste *software* são disponibilizados vários algoritmos de *Machine Learning*. Estes algoritmos podem ser aplicados a um conjunto de dados (como é o caso deste trabalho), ou integrados no *software* desenvolvido.

O Weka integra também algumas ferramentas típicas para o processo de *Data Mining*, nomeadamente: pré-processamento, classificação, regressão, *clustering*, visualização e seleção de features. A integração de todas estas ferramentas é essencial para o sucesso do Weka, já que o seu principal objetivo é ser aplicado a dados reais.

Na tabela 9 estão referidas alguns cenários reais a que se tentou responder utilizando o Weka [14].

Dados	Aplicação real
Propriedades da uva	Que fatores influenciam a qualidade do vinho?
Diabetes	Será possível produzir um bom diagnóstico clínico para a diabetes humana?
Doenças das vacas	Poderão ser previstas?

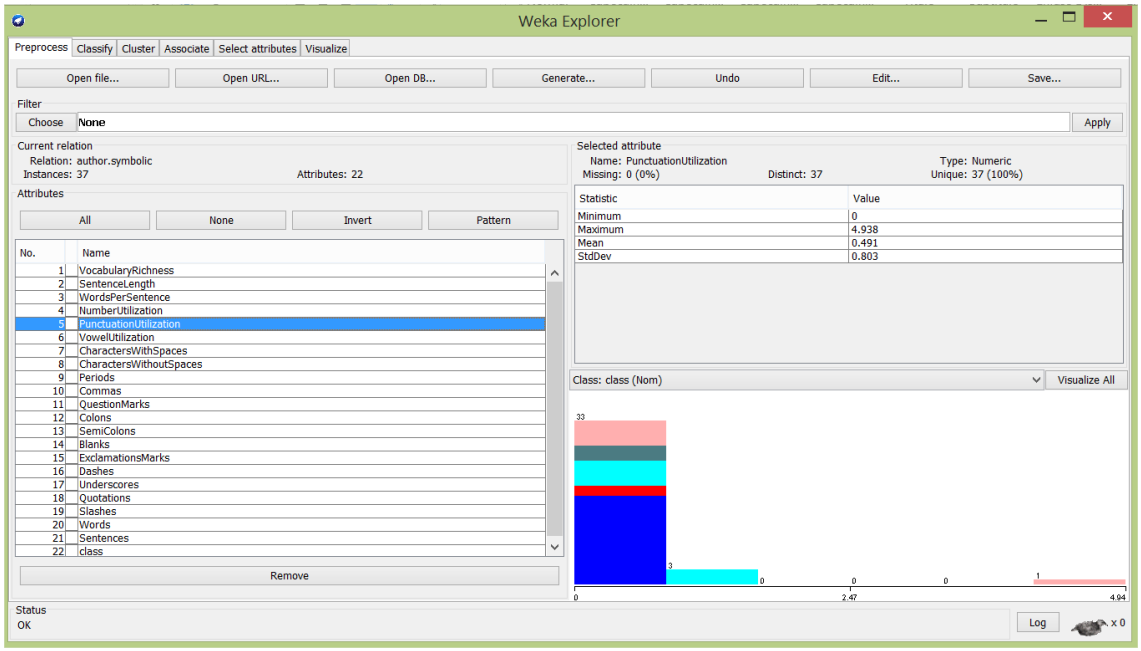
9 - Casos de estudo do Weka com dados reais

Esta ferramenta é passível de ser utilizada em *big data*.

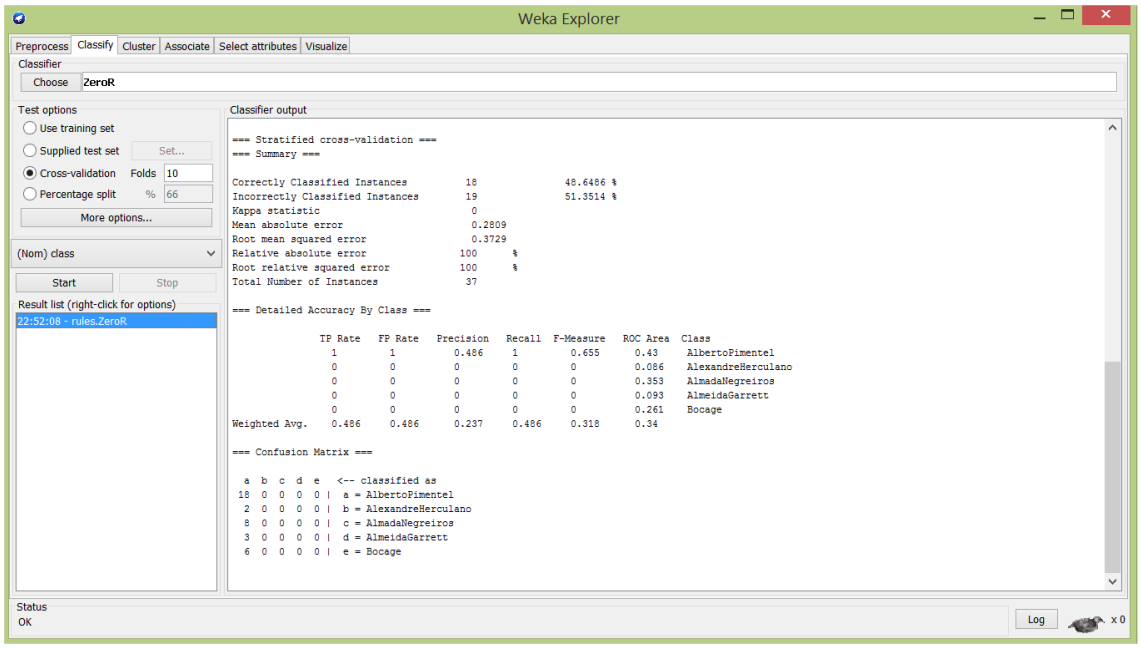
A principal *interface* do Weka é o *Explorer* que tem os seguintes componentes:

- Pré-processamento: Permite importar dados em vários formatos (a partir de uma base dados, de um ficheiro "csv" ou de um ficheiro "arff", por exemplo) e filtra-los (transformando atributos numéricos em discretos, por exemplo) já que, normalmente, os dados podem fazer parte de uma grande base de dados e poderá ser necessário reduzir a quantidade de dados antes de fazer qualquer processamento [14]. Neste painel é ainda possível visualizar os atributos dos dados importados e algumas estatísticas de cada atributo, tal como é possível visualizar na figura 6.
- Classificação: permite aplicar algoritmos de classificação e regressão aos dados. Permite ainda definir os critérios de cross-validation (se aplicável) e visualizar os resultados do processo de classificação, tal como representado na figura 7.

- Associação: permite identificar relações entre os atributos dos dados
- *Clustering*: aplicação de técnicas de *clustering*
- Seleção de atributos: disponibiliza algoritmos que identificam os atributos mais previsíveis
- Visualização: visualização de dados relativos aos atributos



6 - Painel de pré-processamento do Weka



7- Painel de Classificação do Weka

Quando os dados são importados para o Weka, são convertidos para um formato intermédio chamado ARFF (*Attribute Relation File Format*).

Um ficheiro ARFF é um ficheiro de texto ASCII que descreve uma lista de instâncias que partilham dados atributos.

Estes ficheiros são compostos por duas partes:

1. Cabeçalho: Definição da relação e dos atributos a analisar. A relação define o conceito a aprender. Os atributos são definidos pelo seu nome e tipo de dados. Os atributos podem ser do tipo numérico, nominal, texto ou data. No caso de um atributo ser do tipo nominal, deve definir-se também a lista de valores possíveis.
2. Dados: Listagem das instâncias analisar, com os valores dos atributos definidos (separados por vírgulas).

Na figura 8 pode ver-se um exemplo de um ficheiro no formato ARFF.

```
@relation author.symbolic
@ATTRIBUTE NumberUtilization  NUMERIC
@ATTRIBUTE VowelUtilization  NUMERIC
@ATTRIBUTE class        {FranciscoJorgedeAbreu,JoséMartinianodeAlencar}
@DATA
0,37,FranciscoJorgedeAbreu
0,38,FranciscoJorgedeAbreu
0,37,JoséMartinianodeAlencar
0,38,JoséMartinianodeAlencar
0,37,JoséMartinianodeAlencar
```

8- Exemplo de um ficheiro no formato ARFF

2.5. Conclusão

Uma vez que o nosso objetivo é construi um sistema de AA, é essencial compreender alguns conceitos subjacentes a este tema.

Para isso, neste capítulo, começámos por introduzir o conceito de stylometrics, que será fundamental para definir as características que pretendemos extrair de modo a identificar cada documento.

De seguida, apresentámos o trabalho desenvolvido por Homem e Carvalho em [1], que serviu de base para este trabalho e cuja framework criada se pretende estender.

Posteriormente, foram apresentados alguns trabalhos no âmbito de AA.

Finalmente, foram apresentados os recursos utilizados no decorrer deste trabalho: o Projeto Gutenberg e o Weka. O Projeto Gutenberg foi a fonte de recolha de todos os documentos utilizados neste trabalho. O Weka foi a ferramenta de Data Mining utilizada para avaliar os resultados obtidos.

3. Solução Proposta

Neste capítulo será apresentada a solução proposta nesta tese.

Para isso, começaremos por apresentar a estrutura de dados utilizada para caracterizar um documento. De seguida, será apresentada a implementação desta solução.

3.1. Caracterização de um documento

Além das top-k palavras mais utilizadas, decidimos utilizar um conjunto de features para identificar cada documento.

As features utilizadas neste trabalho foram:

- **Riqueza de vocabulário** - Utilização de palavras variadas, isto é, rácio entre o número de palavras distintas e o número total de palavras
- **Tamanho das frases** - Número de caracteres por frase
- **Nº médio de palavras por frase**
- **Utilização de algarismos** - Percentagem de caracteres utilizados que correspondem a algarismos
- **Utilização de pontuação** - Percentagem de caracteres utilizados que correspondem a sinais de pontuação
- **Utilização de vogais** - Percentagem de caracteres utilizados que correspondem a vogais
- **Número de caracteres (incluindo espaços em branco)**
- **Número de caracteres (sem espaços em branco)**
- **Nº pontos finais**
- **Nº de vírgulas**
- **Nº de pontos de interrogação**
- **Nº de dois pontos**
- **Nº de ponto e vírgula**
- **Nº de pontos de exclamação**
- **Nº de hífens**
- **Nº de underscores**
- **Nº de aspas**
- **Nº de barras**

Além destas features, forma ainda tidas em conta as top-10, top-25 e top-50 palavras mais utilizadas de cada documento.

Uma vez que a ferramenta de Data Mining utilizada será o Weka, era necessário criar um formato vetorial para identificar os documentos, que conciliassem estas duas informações.

Para isso, criou-se um vetor cujas 18 entradas iniciais continham a informação relativa aos valores calculados para as features estatísticas.

As restantes posições do vetor contêm informação relativa às top-k palavras mais utilizadas. Isto é, esta é uma lista das *n* top-k palavras mais frequentes distintas, tendo em conta todos os documentos do corpora. Assim, a posição w_x contém 1 se essa palavra faz parte das top-k palavras mais frequentes para esse documento, e 0 caso contrário.

A estrutura criada encontra-se representada na figura 9.

9 - Representação vetorial das características de um documento

3.2. Implementação

3.2.1. Recolha de Corpora

Os corpora utilizados foram recolhidos do Projeto Gutenberg (descrito com mais detalhe na secção 2.4.1). Reúne 233 textos de 50 autores portugueses diferentes, tal como se mostra na tabela que se segue.

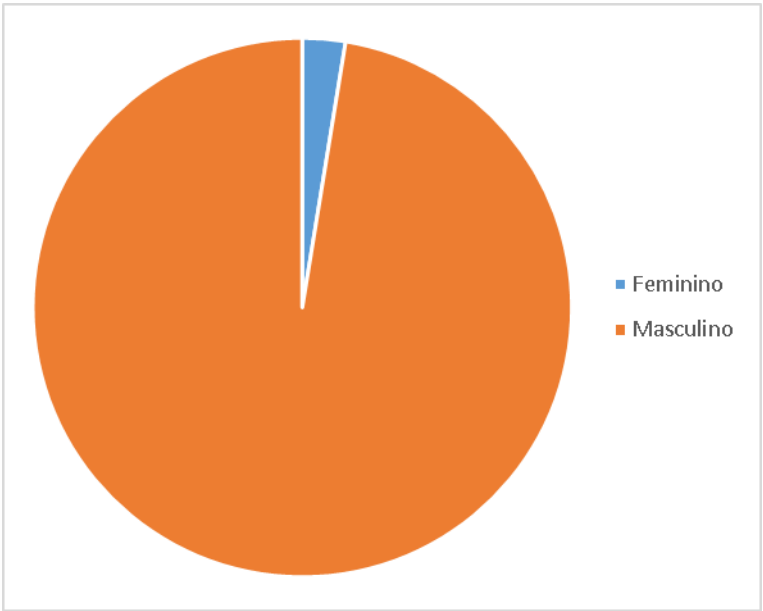
Autor	Nº de Documentos	Nº médio de caracteres/documento	Nº médio de palavras/documento
Abílio Manuel Guerra Junqueiro	5	75897	12897
Adolfo Coelho	2	121559	19129
Alberto Leal Barradas Monteiro Braga	2	88739	15420
Alberto Pimentel	16	143932	24406
Alexandre Herculano	14	315536	55977
Ana de Castro Osório	2	198280	33804
Antero de Quental	10	48038	8067
António Augusto Teixeira de Vasconcelos	2	53343	9059
António Duarte Gomes Leal	4	69060	12176
Antonio Feliciano de Castilho	3	204893	35127
António Pereira Nobre	2	102231	18489
Augusto Gil	2	30225	5223
Camilo Castelo Branco	43	170240	29951
Carlos Testa	4	92509	15223
Eça de Queirós	9	582911	132204
Fernandes Costa	2	83438	13472
Florbela de Alma da Conceição Espanca	1	18874	3451
Francisco Jorge de Abreu	2	311606	50525
Gil Vicente	2	10264	1828
Gonçalo Anes Bandarra	2	37477	6535
Henrique Ernesto de Almeida Coutinho	2	9951	1678
Jaime de Magalhães Lima	11	143909	24634
João Augusto Marques Gomes	3	69393	11823
João Manuel Pereira Silva	2	206571	33432
João Marques de Carvalho	3	107898	17404
Joaquim Carlos Paiva de Andrada	2	86888	15018
José Agostinho de Macedo	3	60668	10379
José da Silva Mendes Leal	2	128381	21478
José Daniel Rodrigues da Costa	3	43614	7613
José Martiniano de Alencar	3	113704	19749
José Sobral de Almada Negreiros	6	25010	4486
Júlio Dinis	3	824104	141541
Luciano Cordeiro	3	82074	13120
Luís de Camões	2	416361	74512
Luiz Augusto Rebello da Silva	3	250542	68348
Manoel Caldas Cordeiro	2	37185	6156

Manuel Maria Barbosa du Bocage	5	19474	3321
Manuel Pinheiro Chagas	4	156502	27333
Maria Amália Vaz de Carvalho	3	303638	60453
Nicolau Tolentino de Almeida	2	84543	14651
Raimundo António de Bulhão Pato	2	64083	11280
Raul Germano Brandão	2	347551	88733
Rui de Pina	7	148940	32229
Sebastião de Magalhães Lima	8	117402	19076
Teixeira Bastos	3	66505	10605
Teixeira de Pascoais	4	19412	3452
Teófilo Braga	4	184989	31000
Vicente de Almeida de Eça	2	68148	11463
Visconde de João Batista da Silva Leitão de Almeida Garrett	3	186479	47617
Wenceslau José de Sousa de Morais	2	140996	23386

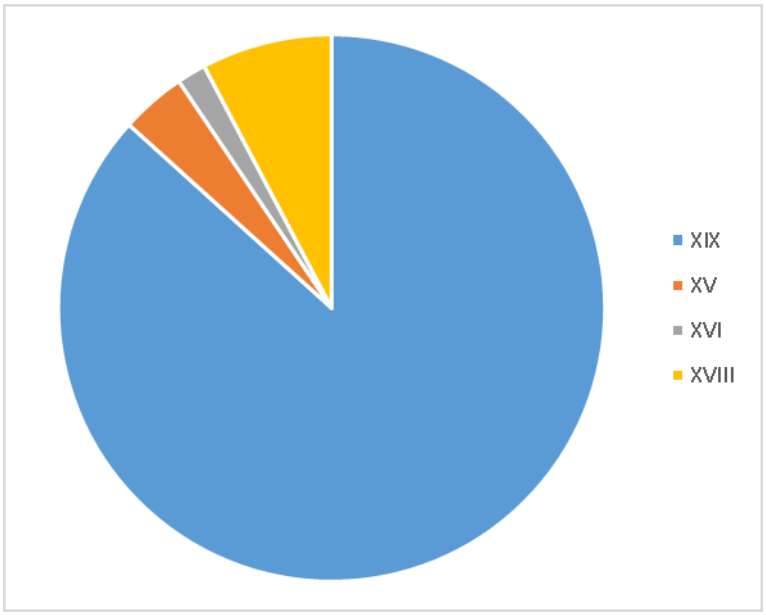
Foram também recolhidos dados que caracterizam os autores dos documentos recolhidos, nomeadamente:

- Sexo
- Século de nascimento do autor
- Década de nascimento do autor

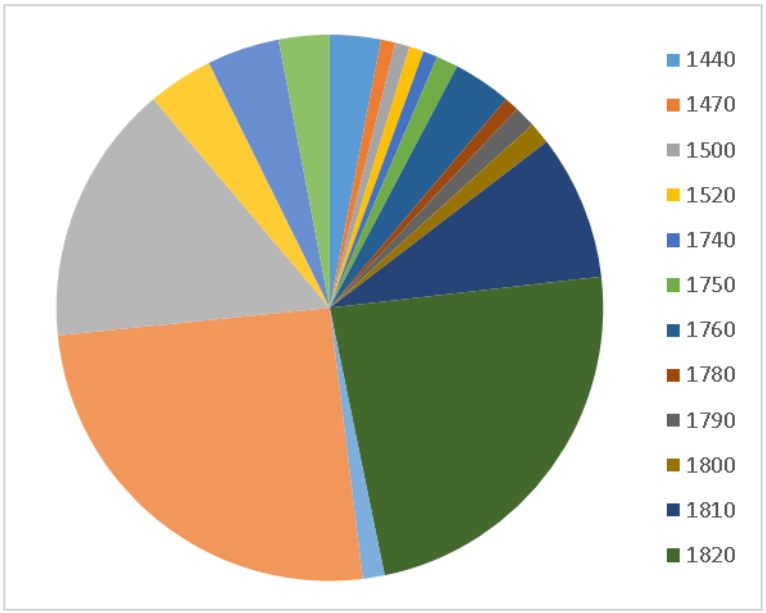
Quanto a estas características, os corpora distribuem-se como demonstrado nas figuras 10, 11 e 12.



10- Distribuição de livros segundo o sexo do autor



11 - Distribuição de livros segundo o século de nascimento do autor



12 - Distribuição de livros segundo a década de nascimento do autor

Nos 233 documentos analisados existem 7131908 palavras, que correspondem a 170276 palavras distintas.

As 5 palavras mais frequentes destes documentos são: “a” (260986 ocorrências), “de” (256343 ocorrências), “e” (236781 ocorrências), “que” (236245 ocorrências) e “o” (228450 ocorrências).

Na tabela 10 apresentam-se os 5 documentos mais longos, isto é, com maior número de caracteres.

Documento	Autor	Nº Caracteres
Os Maias	Eça de Queirós	1290356
A Morgadinha dos Cannaviaes	Júlio Dinis	891951
O crime do padre Amaro, scenas da vida devota	Eça de Queirós	842007
Os fidalgos da Casa Mourisca	Júlio Dinis	840921
Uma família inglesa - Scenas da vida do Porto	Júlio Dinis	739440

10 - Documentos mais longos

3.2.2. Tratamento de dados

De forma a possibilitar uma análise automática dos documentos recolhidos, verificou-se que era necessário submetê-los a um pré-processamento.

Para começar, analisando os documentos, demos conta da existência de um "cabeçalho" e um "rodapé" *standard* criado pelo Projeto Gutenberg. Para que estes não enviesassem as métricas que pretendíamos calcular, optámos por removê-los.

De seguida, normalizou-se o *encoding* utilizado, para que todos os documentos tivessem o mesmo *encoding*.

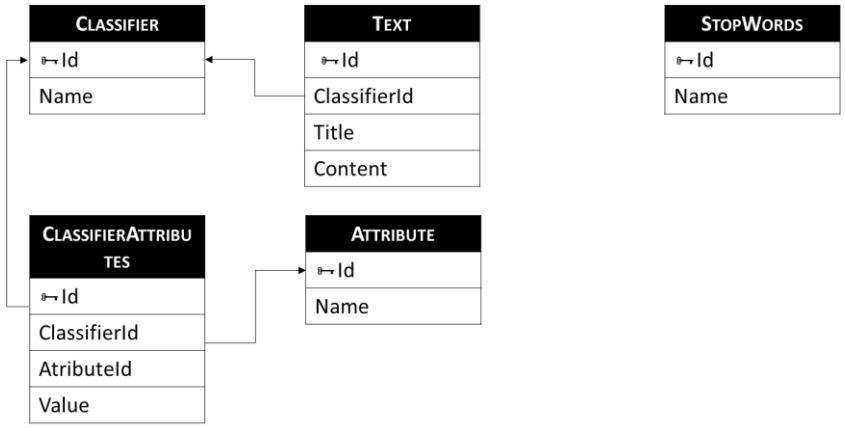
Finalmente, nos testes preliminares observou-se que a distribuição desigual do número de documentos por autor causava um enviesamento dos resultados obtidos, já que grande parte dos documentos eram (erradamente) classificados como sendo de Camilo Castelo Branco. Para resolver este problema, optou-se por extrair 10 excertos aleatórios, de cada autor. Cada excerto extraído contem 1000 palavras.

3.2.3. Criação da baseline

O primeiro passo na implementação desta tese foi adaptar a ferramenta existente (desenvolvida por Homem e Carvalho [1]) para identificação o autor de livros e não de textos jornalísticos.

3.2.4. Arquitetura

Como suporte de dados foi criada uma Base de Dados SQL, com a arquitetura apresentada na figura 13.



13 - Arquitetura da Base de Dados criada

Por exemplo, na tabela "Attribute" pode-se definir que cada autor possui como atributos o seu sexo, o século em que nasceu e ainda a sua década de nascimento. A tabela "Attribute" contém um exemplo do que poderá ser o conteúdo da tabela "Attribute".

Id	Name
95	FriendlyName
96	DecOfBirth
97	SecOfBirth
98	Sex

11 – Exemplo de conteúdo da tabela "Attribute"

A tabela "ClassifierAttributes" pretende mapear os atributos de cada autor que tenham sido definidos. A tabela contém um exemplo do que poderá ser o conteúdo da tabela "ClassifierAttributes".

Id	ClassifierId	AttributeId	Value
3816	1477	95	Francisco Jorge de Abreu
3817	1477	96	1870
3818	1477	97	XIX
3819	1477	98	M
3820	1478	95	José Martiniano de Alencar
3821	1478	96	1820
3822	1478	97	XIX
3823	1478	98	M

12 - Exemplo de conteúdo da tabela "ClassifierAttributes"

Analisando, estas tabelas podemos observar todos as características de um dado classificador (isto é, autor). Por exemplo, para o classificador com o "Id" 1477, sabemos que o seu Nome é "Francisco Jorge de Abreu", nasceu no século XIX, na década de 1870 e é do sexo Masculino.

Mais do que isso, esta arquitetura permite que facilmente sejam adicionados novos atributos a analisar (inserindo um novo atributo na tabela "Atributo" e efetuando o respetivo mapeamento na tabela "ClassifierAttribute"), sem que haja necessidade de alterar a Base de Dados.

À arquitetura foi ainda adicionada uma tabela com uma lista de Stop Words. Esta tabela foi populada com a lista de Stop Words recolhida da internet[3].

Além disso, a Base de Dados possui algumas tabelas auxiliares onde é guardada a informação para calcular as features definidas.

Para o carregamento da informação na Base de Dados e o seu tratamento foi desenvolvido um conjunto de procedimentos que calculam as features de cada documento e inserem a informação nas tabelas respetivas.

Finalmente, foi criado um procedimento que exporta a informação relativa a cada feature disponível na Base de Dados para um formato ARFF (formato suportado pelo Weka).

3.3. Conclusão

Ao longo deste capítulo foi apresentada a solução proposta nesta tese.

Em primeiro lugar, foi descrito o modelo de caracterização de documentos utilizado.

De seguida, foi apresentada a implementação efetuada, passando pelo processo de recolha de corpora, o seu tratamento e o arquitetura da Base de Dados SQL utilizada como suporte de dados.

[3] https://code.google.com/p/stop-words

4. Avaliação

Neste capítulo será apresentada a metodologia utilizada para avaliar a solução proposta, bem como os resultados obtidos com a mesma.

O objetivo deste trabalho é avaliar a eficácia da utilização de features simples em conjunto com as top-k palavras mais frequentes, na tarefa de identificar o autor de um documento.

Para isso, pretende-se avaliar se é possível enriquecer a baseline, com a adição de features simples.

4.1. Metodologia

Como ferramenta de análise de dados foi utilizado o Weka (ver secção 2.4.2). Para tal, foram extraídas da Base de Dados as características associadas a cada documento e, posteriormente, foram inseridas em ficheiros ARFF.

Assim sendo, foram extraídos 4 ficheiros ARFF distintos, apenas com as features de stylometric e os que além destas features continham informação relativas as top-10, top-25 e top-50 palavras mais usadas.

Uma vez que as palavras mais frequentes de todos os corpora recolhidos correspondem maioritariamente a Stop Words, optou-se por extrair os ficheiros que incluíam informação relativa às palavras mais utilizadas, excluindo as Stop Words.

Posteriormente, os ficheiros ARFF foram analisandos utilizando os classificadores do Weka. Neste processo de classificação, optou-se sempre por utilizar cross-validation com 10 *folds*.

Cada um dos ficheiros ARFF foi analisado utilizando 14 dos algoritmos disponíveis no Weka:

1. BayesNet
2. ComplementNaiveBayes
3. NaiveBayes
4. IB1
5. LWL
6. Bagging
7. RandomSubSpace
8. HyperPipes
9. PART
10. FT
11. J48
12. RandomForest
13. RandomTree
14. SimpleCart

Finalmente, utilizando o cenário que obteve melhores resultados, comparou-se com uma implementação do trabalho de Homem e Carvalho [1] adaptado ou Weka. Ou seja, depois de determinar o cenário com melhores resultados, avaliou-se o mesmo sem recurso às features introduzidas nesta tese.

Finalmente, utilizando os algoritmos que tiveram melhor desempenho nos testes, avaliou-se se este modelo seria aplicável noutro contexto que não o de identificar o autor de um documento.

Para isso, extraíram-se novos ficheiros ARFF em que os documentos estavam classificados segundo as novas características, em vez de estarem identificados por autor. As características analisadas foram:

- Sexo do autor
- Século de nascimento do autor
- Década de nascimento do autor

4.2. Experiências

4.2.1. Ferramenta desenvolvida por Homem e Carvalho

Utilizando a baseline criada (ver secção 0), isto é a adaptação da ferramenta existente para identificar o autor de livros, obteve-se uma taxa de sucesso na ordem dos 60%.

4.2.2. Features + Algoritmos

Nesta fase foram considerado os seguintes cenários:

1. Apenas features de *stylometric*
2. Features de *stylometric* e as 10 palavras mais utilizadas em cada documento (com Stop Words)
3. Features de *stylometric* e as 10 palavras mais utilizadas em cada documento (sem Stop Words)
4. Features de *stylometric* e as 25 palavras mais utilizadas em cada documento (com Stop Words)
5. Features de *stylometric* e as 25 palavras mais utilizadas em cada documento (sem Stop Words)
6. Features de *stylometric* e as 50 palavras mais utilizadas em cada documento (com Stop Words)
7. Features de *stylometric* e as 50 palavras mais utilizadas em cada documento (sem Stop Words)

Os resultados desta fase de testes podem ser analisados na tabela 13.

Algoritmo	Apenas stylometrics	Stylometrics + Top 10 (com Stop Words)	Stylometrics + Top 10 (sem Stop Words)	Stylometrics + Top 25 (com Stop Words)	Stylometrics + Top 25 (sem Stop Words)	Stylometrics + Top 50 (com Stop Words)	Stylometrics + Top 50 (sem Stop Words)
Bayes Net	19,2	19,2	21,8	21,2	22	23,8	22,2
Complement Naive Bayes	4	13,8	44,2	35	59,4	54,2	67,8
Naive Bayes	41	50,6	65,2	64,2	71,8	71	79
IB1	47,6	35,4	60,4	48,2	64,6	57,4	47,8
LWL	22,2	23,2	40,4	30,8	46,2	34,4	51,8
Bagging	40,4	41,4	43,2	43,6	43,4	42,8	42,4
Random Sub Space	40,2	42,8	42	43,4	42,6	45,2	45
Hyper Pipes	25,2	30,2	54,8	49,2	68	67	77,2
PART	31,8	28,6	40,4	33	35	38,4	43,2
FT	46,2	51,2	57,2	54,6	59,2	58,2	61,4
J48	35	32	41	32,4	40,4	38,6	42
Random Forest	42,4	39,4	47	37,4	41,6	35,8	37,4
Random Tree	30,6	28,6	28,6	21,4	25,2	19,4	19,8

13 - Resultados iniciais

Consolidando os resultados obtidos, de acordo com o algoritmo utilizado, concluímos que os algoritmos que apresentam melhores resultados são (ver tabela 14):

- Complement Naive Bayes
- Naive Bayes
- IB1
- Random Sub Space
- Hyper Pipes
- FT

	Bayes Net	Complement Naive Bayes	Naive Bayes	IB1	LWL	Bagging	Random Sub Space	Hyper Pipes	PART	FT	J48	Random Forest	Random Tree	Simple Cart
Média	21,34	39,77	63,26	51,63	35,57	42,46	43,03	53,09	35,77	55,43	37,34	40,14	24,80	36,03
Mínimo	19,20	4,00	41,00	35,40	22,20	40,40	40,20	25,20	28,60	46,20	32,00	35,80	19,40	31,00
Máximo	23,80	67,80	79,00	64,60	51,80	43,60	45,20	77,20	43,20	61,40	42,00	47,00	30,60	42,00
Desvio Padrão	1,27	19,00	9,98	7,86	9,05	0,91	1,29	15,62	4,20	4,08	3,61	3,02	3,94	3,40

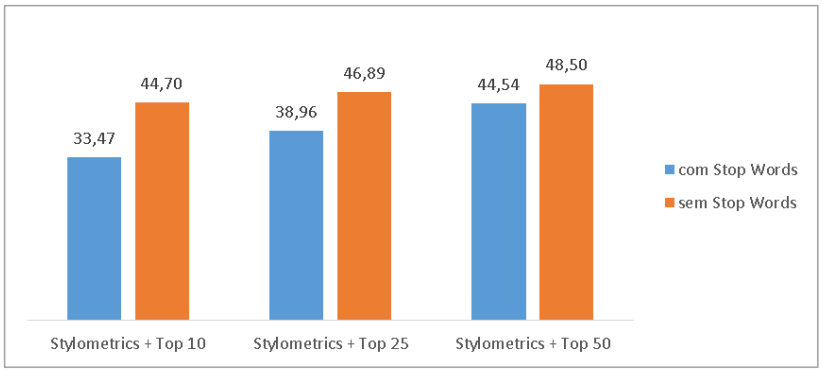
14 - Resultados por Algoritmo

Analisando os resultados com base na técnica utilizada (apenas features stylometrics ou utilizando também as palavras mais utilizadas), podemos observar que os piores resultados são obtidos quando não se tem em consideração as palavras mais utilizadas (ver tabela 15).

	Média	Mínimo	Máximo	Desvio Padrão
Apenas stylometrics	32,77	4,00	47,60	9,09
Stylometrics + Top 10	33,47	13,80	51,20	8,57
Stylometrics + Top 25	38,96	21,20	64,20	9,92
Stylometrics + Top 50	44,54	19,40	71,00	12,25

15 - Resultados por Técnica

De seguida, decidiu-se averiguar se, excluindo as Stop Words consideradas da lista de palavras mais frequentes, se poderia melhorar estes resultados. Utilizando esta técnica, obteve-se uma melhoria média de 7,7%, sendo que esta melhoria se verificou em todos os cenários testados (10 palavras mais utilizadas, 25 palavras mais utlizadas e 50 palavras mais utilizadas), tal como ilustrado na figura 14.

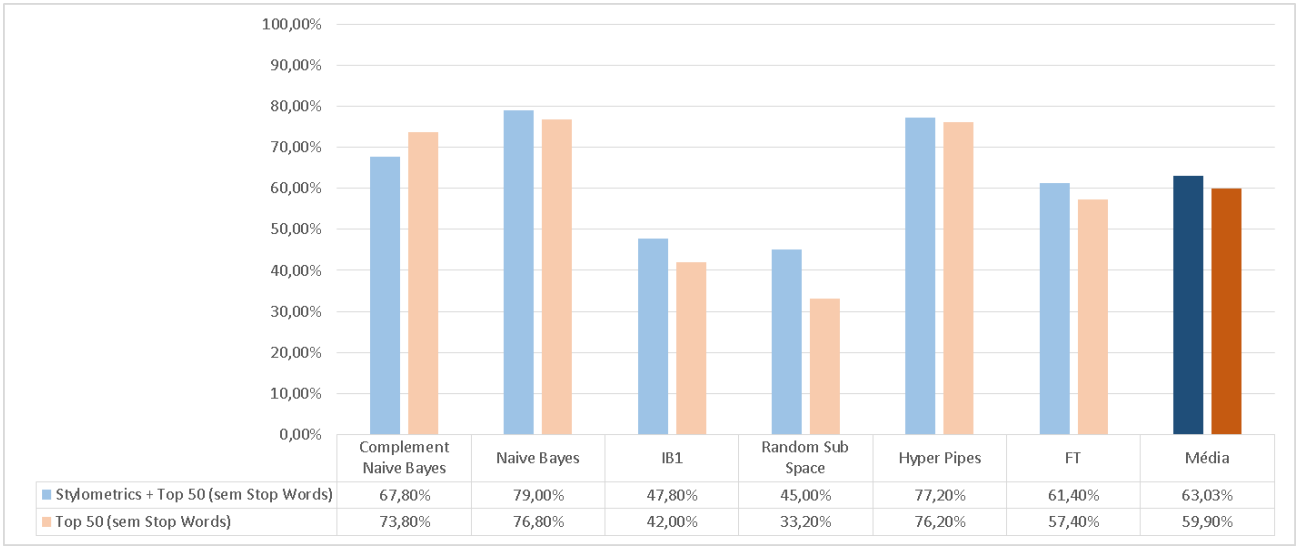


14- Remoção de Stop Words

4.2.1. Importância da utilização de features

De forma a avaliar se a introdução das features havia melhorado o modelo inicial, decidiu-se utilizar a versão que obteve melhores resultados (top-50 palavras mais utilizadas) e comparar a sua performance com e sem as features introduzidas neste trabalho.

Tal como se pode observar na figura 15, houve uma melhoria em praticamente todos os algoritmos considerados (a exceção é o algoritmo "Complement Naive Bayes"), que se traduziu numa melhoria média de 3,13%, provando que estas features vieram enriquecer o modelo previamente.



	Complement Naive Bayes	Naive Bayes	IB1	Random Sub Space	Hyper Pipes	FT	Média
Stylometrics + Top 50 (sem Stop Words)	67,80%	79,00%	47,80%	45,00%	77,20%	61,40%	63,03%
Top 50 (sem Stop Words)	73,80%	76,80%	42,00%	33,20%	76,20%	57,40%	59,90%

15 - Utilização de features

Um último teste efetuado foi a avaliação desta metodologia aplicada a outros cenários. Para isso, utilizou-se a informação acerca de cada autor que havia sido previamente recolhida e inserida na base de dados.

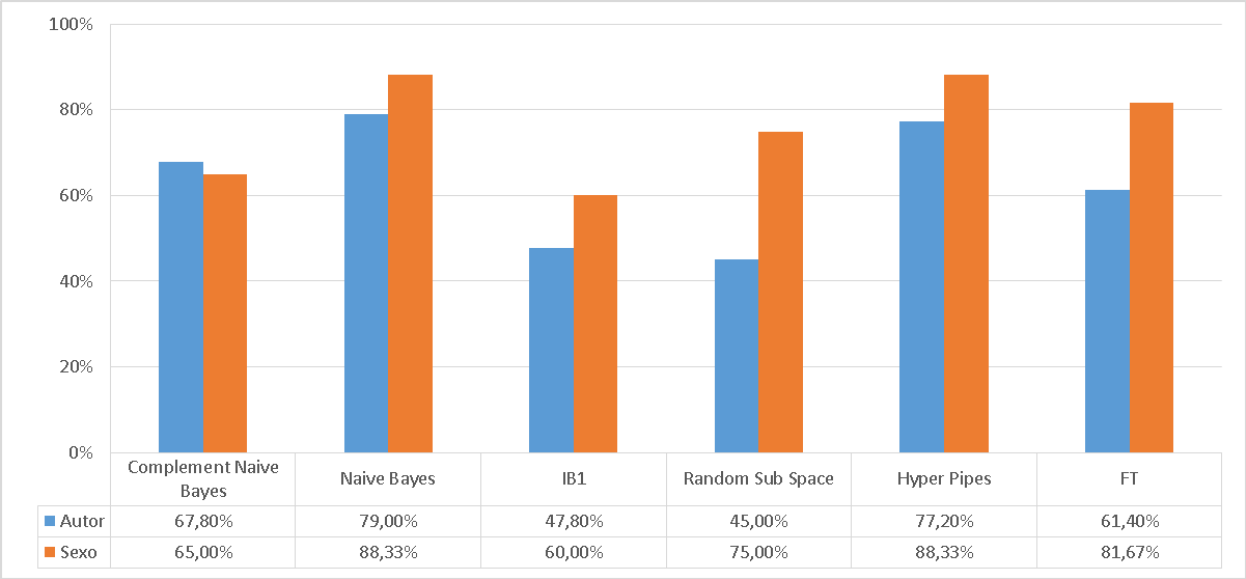
4.2.2. Identificação do sexo do autor

Depois de determinar que os melhores resultados são obtidos utilizando as top-50 palavras mais utilizadas e as features introduzidas nesta tese, decidimos averiguar se este modelo poderá ser utilizado para inferir o sexo do autor de um documento.

Para isso, foi gerado um ficheiro ARFF em que cada documento está classificado de acordo com o sexo do seu autor (em vez de estar classificado por autor).

Mais uma vez, foi necessário fazer uma seleção dos documentos a utilizar, de forma a haver uma distribuição equitativa por sexo. Para isso, foram escolhidos aleatoriamente apenas 30 excertos de cada sexo.

Para avaliar os resultados desta experiência comparou-se com os resultados já obtidos na tarefa de identificar o autor de um documento, tal como ilustrado na figura 16.



	Complement Naive Bayes	Naive Bayes	IB1	Random Sub Space	Hyper Pipes	FT
Autor	67,80%	79,00%	47,80%	45,00%	77,20%	61,40%
Sexo	65,00%	88,33%	60,00%	75,00%	88,33%	81,67%

16 - Identificação do sexo do autor

Como se pode observar, houve uma melhoria em praticamente todos os algoritmos analisados (a exceção é o algoritmo "Complement Naive Bayes").

Quando se tenta identificar um autor a taxa de sucesso média é de 63.03% e nesta experiência a taxa de sucesso média é de 76.39%, ou seja, houve uma melhoria de 13.36%.

Assim, podemos concluir que este modelo pode ser aplicado para identificar o sexo do autor de um documento.

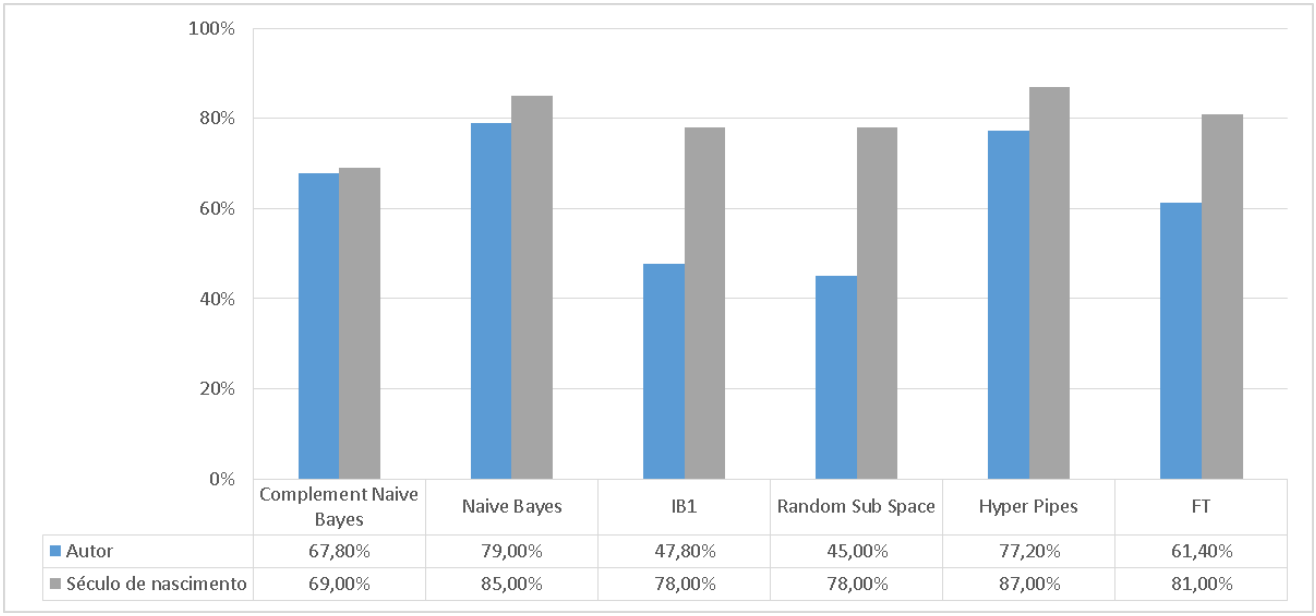
4.2.3. Identificação do século de nascimento do autor

À semelhança descrita em 4.2.2, foi utilizado o modelo criado para identificar o século de nascimento do autor de um documento.

Para isso, foi gerado um ficheiro ARFF cujos documentos estão classificados segundo o século de nascimento do autor.

Tal como na experiência anterior, apenas foram extraídos 30 excertos de cada século.

À semelhança da experiência anterior, comparou-se com os resultados já obtidos na tarefa de identificar o autor de um documento, tal como ilustrado na figura 17.



	Complement Naive Bayes	Naive Bayes	IB1	Random Sub Space	Hyper Pipes	FT
Autor	67,80%	79,00%	47,80%	45,00%	77,20%	61,40%
Século de nascimento	69,00%	85,00%	78,00%	78,00%	87,00%	81,00%

17- Identificação do século de nascimento do autor

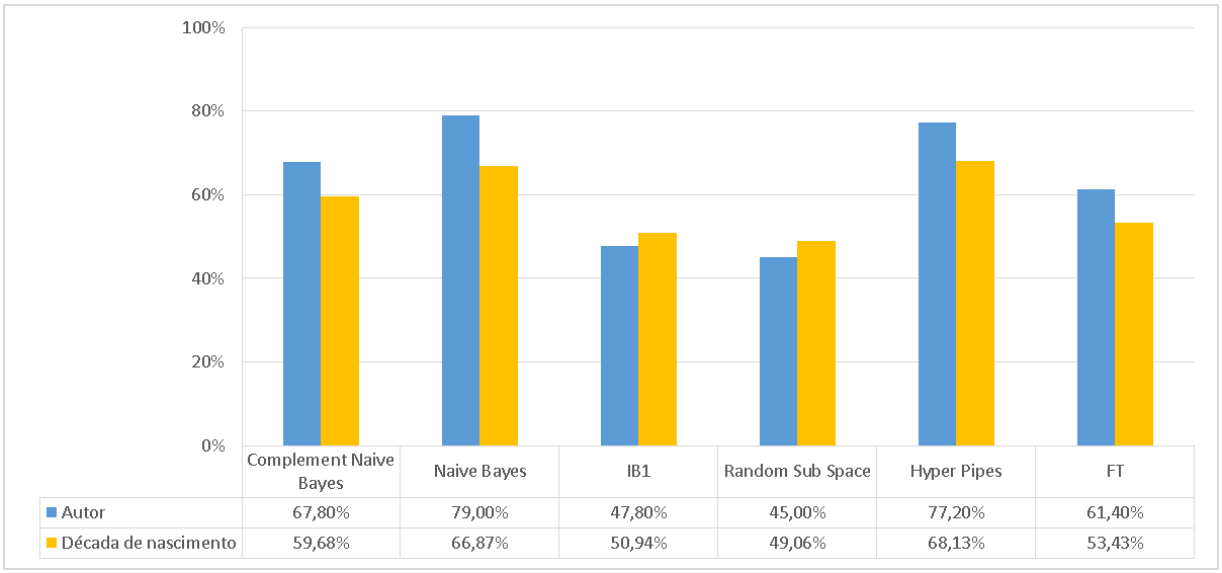
Neste caso, observou-se que houve uma melhoria geral dos resultados obtidos. Passando a taxa de sucesso média (no caso de identificação do autor) de 63.03% para 79.67%, havendo uma melhoria de 16.63%

Mais uma vez podemos concluir que este modelo pode ser aplicado para identificar o século de nascimento do autor de um documento.

4.2.4. Identificação da década de nascimento do autor

À semelhança das experiências anteriores, o modelo criado foi utilizado para identificar a década de nascimento do autor de um documento, sendo para isso gerado o ficheiro ARFF correspondente, utilizando apenas 30 excertos de cada década.

Mais uma vez, comparou-se com os resultados já obtidos na tarefa de identificar o autor de um documento, tal como representado na figura 18.



	Complement Naive Bayes	Naive Bayes	IB1	Random Sub Space	Hyper Pipes	FT
Autor	67,80%	79,00%	47,80%	45,00%	77,20%	61,40%
Década de nascimento	59,68%	66,87%	50,94%	49,06%	68,13%	53,43%

18 - Identificação da década de nascimento do autor

Neste caso, ao contrário do que havíamos verificado nas experiências anteriores, houve um decréscimo da performance do algoritmo. Sendo que a taxa de sucesso média passou de 63.03% para 58.02%.

Como poemos observar, os resultados obtidos nesta experiência são inferiores aos observados anteriormente.

Analisando a Matriz de Confusão (na figura 19) desta experiência (utilizando o Algoritmo Naive Bayes) podemos observar que em grande parte dos casos os documentos são classificados como sendo de uma época próxima da correta. Por exemplo, analisando os documentos da década de 1850, podemos observar que apesar de 16 deles serem corretamente classificados, 4 foram classificados como 1860 e outros 4 como 1840, o que sugere que o estilo de escrita não varia tão rapidamente para que se possa identificar a década em que nasceu o autor de um documento.

1440	1470	1500	1520	1740	1750	1760	1780	1790	1800	1810	1820	1830	1840	1850	1860	1870	1890	
9	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	**1440**
0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1470**
0	0	7	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	**1500**
0	0	0	5	0	0	0	0	0	0	0	1	0	0	2	0	2	0	**1520**
0	0	0	0	8	0	2	0	0	0	0	0	0	0	0	0	0	0	**1740**
0	0	0	0	0	6	2	0	0	0	0	0	0	1	0	0	1	0	**1750**
0	0	0	0	0	0	18	0	0	0	0	0	0	1	0	0	1	0	**1760**
0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	1	0	**1780**
0	0	0	0	0	0	0	0	5	0	0	1	0	3	0	1	0	0	**1790**
0	0	0	0	0	0	0	0	0	3	0	4	0	0	2	1	0	0	**1800**
0	0	0	0	0	0	0	0	0	0	24	1	0	1	3	0	1	0	**1810**
0	0	0	0	0	0	0	0	0	0	2	21	0	3	2	1	1	0	**1820**
0	0	0	0	0	0	0	0	0	0	0	0	6	1	1	2	0	0	**1830**
0	0	0	0	0	0	0	0	0	0	2	5	0	16	4	3	0	0	**1840**
0	0	0	0	0	0	0	0	0	0	2	3	0	4	16	4	1	0	**1850**
0	0	0	0	0	0	0	0	0	0	0	1	1	3	2	20	3	0	**1860**
0	0	0	0	0	1	0	0	0	0	1	1	0	3	4	0	20	0	**1870**
0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	5	0	11	**1890**

19 - Matriz de confusão

4.2.5. Avaliação da performance individual das features

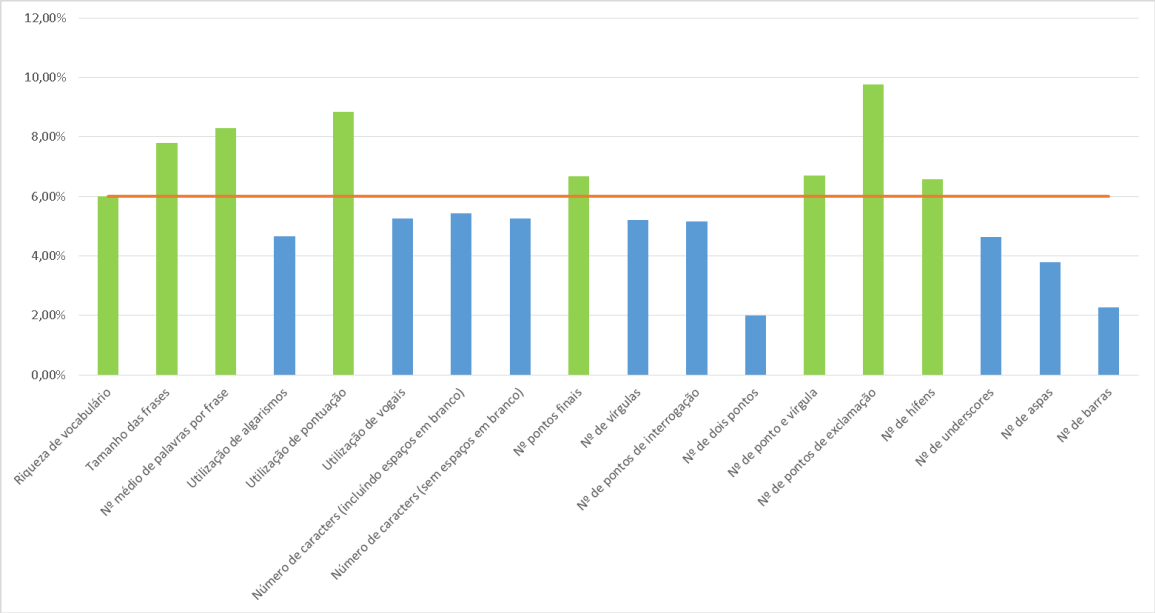
Foi também avaliado se selecionando apenas algumas das features referidas os resultados seriam melhores.

Para isso, começou-se por avaliar a performance individual de cada feature, apresentada na tabela 16.

	Complement Naive Bayes	Naive Bayes	IB1	Random Sub Space	Hyper Pipes	FT	Média
Riqueza de vocabulário	2,00%	9,20%	6,80%	7,40%	3,60%	7,00%	6,00%
Tamanho das frases	2,00%	10,80%	5,80%	13,00%	3,40%	11,80%	7,80%
Nº médio de palavras por frase	2,00%	11,00%	8,40%	12,20%	3,60%	12,60%	8,30%
Utilização de algarismos	2,00%	4,00%	7,80%	7,40%	3,00%	3,80%	4,67%
Utilização de pontuação	2,00%	12,00%	8,80%	13,00%	4,40%	12,80%	8,83%
Utilização de vogais	2,00%	8,00%	2,80%	8,60%	2,20%	8,00%	5,27%
Número de caracters (incluíndo espaços em branco)	2,00%	6,00%	7,40%	5,60%	4,80%	6,80%	5,43%
Número de caracters (sem espaços em branco)	2,00%	6,20%	6,40%	6,00%	4,80%	6,20%	5,27%
Nº pontos finais	2,00%	9,80%	6,60%	9,60%	3,00%	9,00%	6,67%
Nº de vírgulas	2,00%	6,80%	5,80%	5,40%	5,20%	6,00%	5,20%
Nº de pontos de interrogação	2,00%	5,40%	7,60%	8,20%	2,60%	5,20%	5,17%
Nº de dois pontos	2,00%	2,00%	2,00%	2,00%	2,00%	2,00%	2,00%
Nº de ponto e vírgula	2,00%	6,00%	10,40%	10,40%	3,00%	8,40%	6,70%
Nº de pontos de exclamação	2,00%	10,60%	13,00%	14,40%	6,80%	11,80%	9,77%
Nº de hífens	2,00%	9,00%	6,60%	7,00%	5,60%	9,20%	6,57%
Nº de underscores	2,00%	4,60%	5,80%	5,80%	3,00%	6,60%	4,63%
Nº de aspas	2,00%	4,00%	4,00%	4,40%	3,80%	4,60%	3,80%
Nº de barras	2,00%	2,60%	2,60%	2,00%	2,40%	2,00%	2,27%

16 - Performance individual das features

De seguida, tentou-se refazer a tarefa de identificar o autor de um documento apenas utilizando as features cuja performance individual é maior do que 6% (ver figura 20).



20 - Features com performance superior a 6%

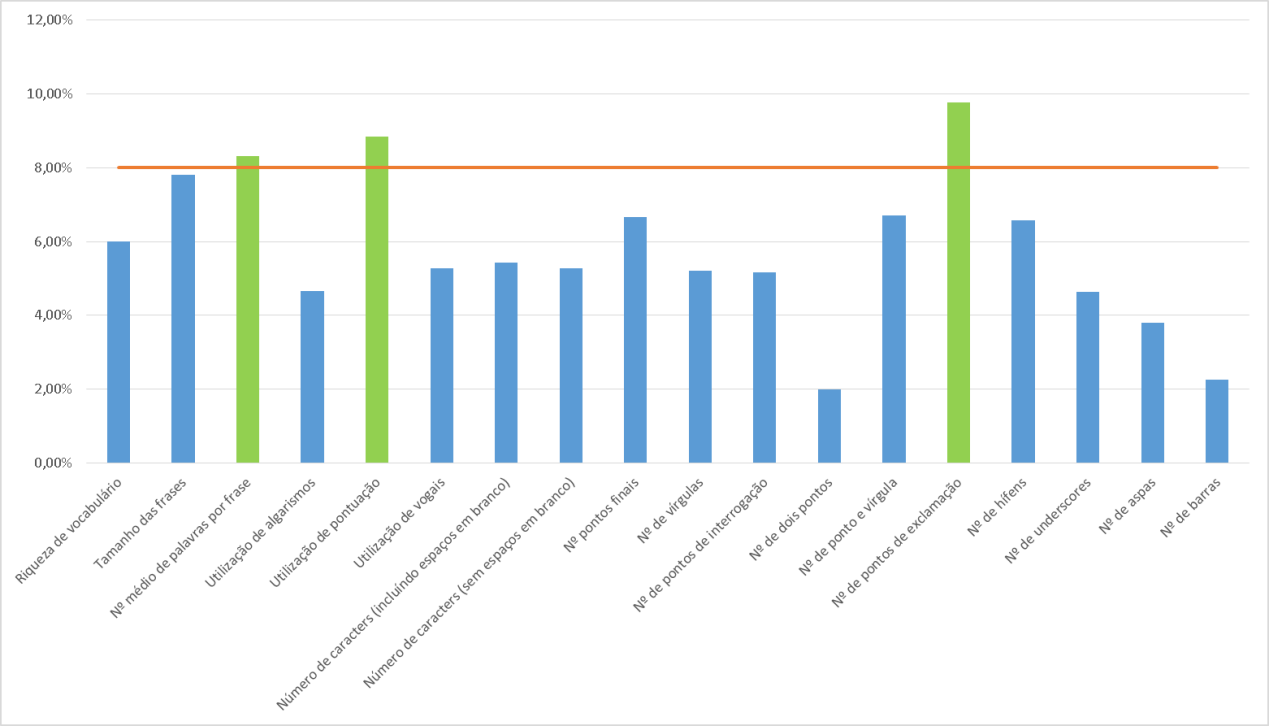
Neste caso não se consegue obter resultados superiores àqueles que já haviam sido obtidos, tal como se pode observar na tabela 17.

	Complement Naive Bayes	Naive Bayes	IB1	Random Sub Space	Hyper Pipes	FT	Média
Autor	67,80%	79,00%	47,80%	45,00%	77,20%	61,40%	63,03%
Features com performance > 6%	72,20%	78,40%	43,00%	42,40%	78,40%	62,00%	62,73%
Diferença	4,40%	-0,60%	-4,80%	-2,60%	1,20%	0,60%	-0,30%

17- Utilização de features com performance superior a 6%

De seguida, averiguou-se também se se restringisse mais as features selecionadas os resultados seriam melhores.

Para isso, utilizaram-se apenas as features com performance individual superior a 8% (ver figura 21).



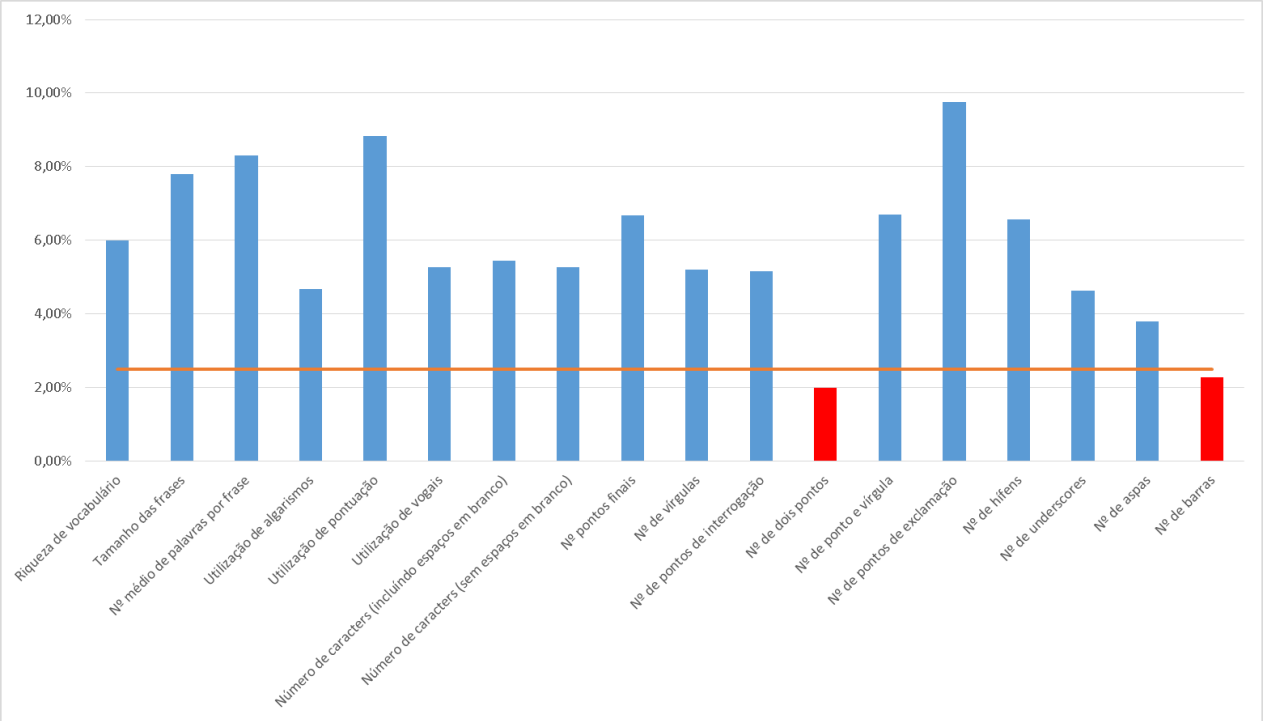
21 - Features com performance superior a 8%

Mais uma vez, os resultados obtidos foram inferiores aqueles que haviam sido obtidos com todas as features, como está representado na tabela 18.

	Complement Naive Bayes	Naive Bayes	IB1	Random Sub Space	Hyper Pipes	FT	Média
Autor	67,80%	79,00%	47,80%	45,00%	77,20%	61,40%	63,03%
Features com performance > 8%	74,00%	78,20%	43,00%	36,60%	78,40%	57,60%	61,30%
Diferença	6,20%	-0,80%	-4,80%	-8,40%	1,20%	-3,80%	-1,73%

18- Utilização de features com performance superior a 8%

Uma vez que selecionando apenas as features com melhor performance os resultados não melhoraram, decidiu-se então remover apenas as features com performance baixa (menor ou igual a 2%), de modo a verificar se estas introduziam entropia no modelo. As features utilizadas nesta experiência estão representadas na figura 22.



22- Features com performance superior a 2,5%

Também neste caso os resultados obtidos foram inferiores aos que haviam sido obtidos utilizando todas as features referidas, tal como se pode observar na tabela 19.

	Complement Naive Bayes	Naive Bayes	IB1	Random Sub Space	Hyper Pipes	FT	Média
Autor	67,80%	79,00%	47,80%	45,00%	77,20%	61,40%	63,03%
Features com performance > 2,5%	67,80%	78,60%	47,80%	43,40%	77,20%	61,40%	62,70%
Diferença	0,00%	-0,40%	0,00%	-1,60%	0,00%	0,00%	-0,33%

19 - Utilização de features com performance superior a 2,5%

Podemos então concluir que, apesar de algumas das features individualmente terem uma performance muito baixa de alguma forma contribuem para a construção de um modelo mais sólido.

4.3. Conclusão

Ao longo deste capítulo foram apresentados os testes efetuados para avaliar este trabalho.

Inicialmente, foi descrita a metodologia de avaliação utilizada.

Finalmente, foram apresentadas as várias experiências realizadas e as conclusões retiradas de cada uma.

5. Conclusão e Trabalho Futuro

Neste capítulo serão discutidos os principais resultados obtidos com este trabalho e de que forma contribuem para a criação de uma metodologia de AA.

Para além disso, serão apresentadas algumas questões às quais este trabalho poderá servir de base.

5.1. Contribuições

Nesta dissertação o nosso objetivo era demonstrar que é possível enriquecer a framework previamente existente, adicionando novas features, que correspondem a métricas estatísticas extraídas de cada documento.

De forma a provar que melhorámos a framework existente, foi recriado o trabalho existente de forma a utilizar o Weka como classificador e assim obter resultados passíveis de analisar.

Com a integração das features no vetor de caracterização de cada documento foi possível obter uma melhoria de 3,13% relativamente ao modelo anterior, demonstrando assim a utilidade da inserção das mesmas.

Para além disso, demonstrámos que se ao construir a lista de palavras mais frequentes excluirmos as Stop Words obtemos resultados melhores. Esta melhoria é de cerca de 7,7%.

Finalmente, analisámos a aplicação desta metodologia para identificar outras características do seu autor, tais como: o sexo, o século e a sua década de nascimento. Nestes testes concluímos que apesar de ser possível identificar o sexo e o século de nascimento de um autor é muito mais difícil identificar a sua década de nascimento.

5.2. Trabalho Futuro

Na nossa opinião, este trabalho poderá ser continuado em várias vertentes de forma a avaliar o potencial desta metodologia:

- Avaliar a aplicação desta metodologia com outros tipos de texto. Por exemplo: texto jornalístico, poesia, …
- Avaliar se será possível identificar o autor de um texto traduzido. Isto é, tendo vários textos de um autor (eventualmente traduzidos por diferentes tradutores), continuará a ser possível identificar o autor do documento?

- Identificar o tradutor de um documento. Ou seja, tendo vários textos classificados segundo o seu tradutor (sendo estes de diferentes autores), será possível identificar o tradutor de um novo documento?

6. Referências

[1] N. Homem e J. P. Carvalho, "Authorship Identification and Author Fuzzy "fingerprints"," *Fuzzy Information Processing Society (NAFIPS), 2011 Annual Meeting of the North American,* 2011.

[2] R. M. Dabagh, "Authorship Attribution and Statistical Text Analysis," *Metodoloski zvezki,* pp. 149-163, 2007.

[3] S. R. Pillay e T. Solorio, "Authorship Attribution of Web Forum Posts," *eCrime Researchers Summit (eCrime),* 2010.

[4] E. Stamatatos, N. Fakotakis e G. Kokkinakis , "Automatic Authorship Attribution," *Conference of the European Chapter of the Association for Computacional Linguistics,* pp. 158-164, 1999.

[5] P. Juola e J. Sofko, "A Prototype for Authorship Attribution Studies," *Lit Linguist Computing,* Junho 2006.

[6] P. Varela, E. Justino e L. Oliveira, "Selecting syntactic attributes for authorship attribution," *The 2011 International Joint Conference on Neural Networks,* 2011.

[7] R. Layton, P. Watters e R. Dazeley, "Authorship Attribution for Twitter in 140 Characters or Less," *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second,* Julho 2010.

[8] O. Aslanturk, E. Sezer, H. Sever e V. Raghavan, "Application of Cascading Rough Set-Based Classifiers on Authorship Attribution Application of Cascading Rough Set-Based Classifiers on Authorship Attribution," *2010 IEEE International Conference on Granular Computing,* 2010.

[9] N. Ali, M. Hindi e R. V. Yampolskiy, "Evaluation of authorship attribution software on a Chat bot corpus," *2011 XXIII International Symposium on Information, Communication and Automation Technologies,* 2011 Outubro 2011.

[10] H. El-Fiqi, E. Petraki e H. Abbass, "A computational linguistic approach for the identification of translator stylometry using Arabic-English text," *2011 IEEE International Conference on Fuzzy Systems,* 2011.

[11] I. N. Bozkurt, Ö. Bağlıoğlu e E. Uyar, “Authorship Attribution - Performance of various features and classification methods,” *22nd IEEE International Symposium on Computer and Information Sciences,* pp. 1-5, 2007.

[12] M. Cristani, G. Roffo e C. Segalin, “Conversationally-inspired Stylometric Features for Authorship Attribution in Instant Messaging,” *Proceedings of the 20th ACM international conference on Multimedia,* 2012.

[13] R. Bolle, J. Connell, S. Pankanti, N. Ratha e A. Senior, “The Relation Between the ROC Curve and the CMC,” *Proc. 4th IEEE Work. Automat. Identification Adv. Technol., pp.15 -20,* 2006.

[14] G. Holmes, A. Donkin e I. H. Witten, “WEKA: A Machine Learning Workbench,” *Proc Second Australia and New Zealand Conference on Intelligent Information Systems,* 1994.