Universidade Federal de Pernambuco
Centro de Informática

Pós-Graduação em Ciência da Computação

SELEÇÃO DE CARACTERÍSTICAS PARA PROBLEMAS DE CLASSIFICAÇÃO DE DOCUMENTOS

Roberto Hugo Wanderley Pinheiro

Dissertação de Mestrado

Recife
27 de julho de 2011

Universidade Federal de Pernambuco
Centro de Informática

Roberto Hugo Wanderley Pinheiro

SELEÇÃO DE CARACTERÍSTICAS PARA PROBLEMAS DE CLASSIFICAÇÃO DE DOCUMENTOS

Trabalho apresentado ao Programa de Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: *Prof. Dr. George Darmiton da Cunha Cavalcanti*
Co-orientador: *Prof. Dr. Tsang Ing Ren*

Recife
27 de julho de 2011

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

Pinheiro, Roberto Hugo Wanderley
Seleção de características para problemas de classificação de documentos / Roberto Hugo Wanderley Pinheiro - Recife: O Autor, 2011.
xi, 88 folhas : il., fig., tab.

Orientador: George Darmiton da Cunha Cavalcanti.
Dissertação (mestrado) - Universidade Federal de Pernambuco. CIn, Ciência da Computação, 2011.

Inclui bibliografia e apêndice.

1. Inteligência artificial. 2. Recuperação da informação. I. Cavalcanti, George Darmiton da Cunha (orientador). II. Título.

006.31 CDD (22. ed.) MEI2011 – 106

Dissertação de Mestrado apresentada por **Roberto Hugo Wanderley Pinheiro** à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título **"Seleção de Características para Problemas de Classificação de Documentos"**, orientada pelo Prof. **George Darmiton da Cunha Cavalcanti** e aprovada pela Banca Examinadora formada pelos professores:

Profa. Flávia de Almeida Barros
Centro de Informática / UFPE

Prof. Renato Fernandes Corrêa
Departamento de Ciência da Informação/UFPE

Prof. George Darmiton da Cunha Cavalcanti
Centro de Informática / UFPE

Visto e permitida a impressão.
Recife, 27 de julho de 2011.

Prof. Nelson Souto Rosa
Coordenador da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.

Resumo

Os sistemas de classificação de documentos servem, de modo geral, para facilitar o acesso do usuário a uma base de documentos. Esses sistemas podem ser utilizados para detectar *spams*; recomendar notícias de uma revista, artigos científicos ou produtos de uma loja virtual; refinar buscas e direcioná-las por assunto. Uma das maiores dificuldades na classificação de documentos é sua alta dimensionalidade. A abordagem *bag of words*, utilizada para extrair as características e obter os vetores que representam os documentos, gera dezenas de milhares de características. Vetores dessa dimensão demandam elevado custo computacional, além de possuir informações irrelevantes e redundantes. Técnicas de seleção de características reduzem a dimensionalidade da representação, de modo a acelerar o processamento do sistema e a facilitar a classificação. Entretanto, a seleção de características utilizada em problemas de classificação de documentos requer um parâmetro m que define quantas características serão selecionadas. Encontrar um bom valor para m é um procedimento complicado e custoso. A idéia introduzida neste trabalho visa remover a necessidade do parâmetro m e garantir que as características selecionadas cubram todos os documentos do conjunto de treinamento. Para atingir esse objetivo, o algoritmo proposto itera sobre os documentos do conjunto de treinamento e, para cada documento, escolhe a característica mais relevante. Se a característica escolhida já tiver sido selecionada, ela é ignorada, caso contrário, ela é selecionada. Deste modo, a quantidade de características é conhecida no final da execução do algoritmo, sem a necessidade de declarar um valor prévio para m. Os métodos propostos seguem essa ideia inicial com certas variações: inserção do parâmetro f para selecionar várias características por documento; utilização de informação local das classes; restrição de quais documentos serão usados no processo de seleção. Os novos algoritmos são comparados com um método clássico (*Variable Ranking*). Nos experimentos, foram usadas três bases de dados e cinco funções de avaliação de característica. Os resultados mostram que os métodos propostos conseguem melhores taxas de acerto.

Palavras-chave: Classificação de Documentos, Seleção de Características, k vizinhos mais próximos, *Naïve Bayes*, Recuperação de Informação.

Abstract

Text classification systems are used, generally, to facilitate user access to a database of documents. These systems can be used to detect *spam*; recommend a news magazine, journal articles or products of a virtual store; refine searches and direct them by subject. A major difficulty in the text classification is their high dimensionality. The bag of words approach, used to extract features and obtain the vectors representing the documents, generates tens of thousands of features. Vectors of this size require high computational cost and also has irrelevant and redundant information. Feature selection techniques reduce the dimensionality of the representation in order to accelerate the system process and facilitate the classification. However, the feature selection used in text classification problems requires a parameter m which defines how many features are selected. Finding a good value for m is a complicated and costly procedure. The idea introduced in this work aims to remove the need for the parameter m and ensure that the selected features cover all documents in the training set. To achieve this goal, the algorithm iterates over the documents' training set for each document and choose the most relevant feature. If the chosen feature has already been selected, it is ignored, otherwise it is selected. Thus, the amount of features is known in the execution of the algorithm, without the need to declare a previous value to m. The proposed methods follow this initial idea with some variations: insertion of the parameter f to select multiple features per document, use of local information in classes, restriction of which documents will be used in the selection process. The new algorithms are compared with a classical method (*Variable Ranking*). The experiments used three databases and five feature evaluation functions. The results show that the proposed methods achieve better rates.

Keywords: Text Classification, Feature Selection, k-Nearest Neighbor, Naïve Bayes, Information Retrieval.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ALOFT	*At Least One Feature*
ALfFT	*At Least f Features*
ALfFT-L	*At Least f Features - Local*
ALfFT-R	*At Least f Features - Restricted*
ALfFT-LR	*At Least f Features - Local Restricted*
BNS	*Bi-Normal Separation*
BOW	*Bag of Words*
CHI	*Chi-Squared Statistic*
CDM	*Class Discriminating Measure*
DF	*Document Frequency*
NB	*Naïve Bayes*
FEF	*Feature Evaluation Function*
FS	*Feature Selection*
IDF	*Inverse Document Frequency*
IG	*Information Gain*
kNN	*k Nearest Neighbor*
ML	*Machine Learning*
MOR	*Multi-class Odds Ratio*
OR	*Odds Ratio*
SCD	Sistema de Classificação de Documentos
SFS	*Sequential Forward Selection*
SRI	Sistema de Recuperação de Informação
SVM	*Support Vector Machine*
TC	*Text Classification*
TF	*Term Frequency*
TF-IDF	*Term Frequency - Inverse Document Frequency*
VR	*Variable Ranking*

Tabela de Símbolos

$\mathscr{C}$	Conjunto com todas as classes
c_j	Classe de índice j
$\mathscr{D}$	Conjunto com todos os documentos (treino e teste)
$\mathscr{D}_{tr}$	Conjunto com todos os documentos de treino
$\mathscr{D}_{te}$	Conjunto com todos os documentos de teste
d_i	Documento de índice i
f	Parâmetro dos métodos propostos
h	Índice dos termos
i	Índice dos documentos
j	Índice das classes
k	Quantidade de vizinhos usada no classificador kNN
M	Quantidade de documentos no *Corpus*
m	Quantidade de características selecionadas por um método
N	Quantidade total de categorias pré-definidas
n	Número de caracteres ou palavras na composição de um n-gram
V	Quantidade de termos (tamanho do vocabulário)
w	Termo qualquer
w_h	Termo de índice h sem relação direta com qualquer documento
$w_{h,i}$	Valor do termo de índice h no documento i

CAPÍTULO 1

Introdução

Informação, em diversos meios, é considerada um dos maiores bens existentes. Contudo, não basta possuir muita informação e informação de qualidade, é necessário que seja possível encontrar esta informação com rapidez e com a maior precisão possível.

Para suprir essas necessidades, na década de 1950, surgiram os Sistemas Recuperação de Informação (SRI). Sua tarefa se resume a recuperar documentos relevantes, baseando-se em uma determinada consulta inserida pelo usuário. Inicialmente os SRIs foram criados para pesquisas em bibliotecas, mas não tardou para se propagar no meio acadêmico (publicações científicas) e no meio profissional, entre jornalistas, advogados e médicos.

Com o surgimento da *Internet*, a disponibilidade de documentos passou a aumentar exponencialmente. Com tamanho crescimento, o custo computacional das buscas cresceu e o interesse aumentou neste domínio. Várias ramificações de aplicações passaram a surgir para vencer os novos problemas. Uma das ramificações é o que conhecemos hoje por Classificação de Documentos, cuja aplicação mais comum para SRIs é a indexação de documentos (LEWIS, 1992b), apesar de ser uma tarefa de mineração de texto.

1.1 A Tarefa de Classificação de Documentos

A Classificação de Documentos (TC, do inglês *Text Classification* ou *Text Categorization*) é a tarefa de atribuir a um determinado texto em linguagem natural – em algum idioma específico – uma classe de um universo finito e pré-estabelecido. Aplicações que resolvam esse problema estão tornando-se cada vez mais difundidas, seja na detecção de *spam* em *e-mails*, removendo as mensagens de conteúdo suspeito; organização de documentos em tópicos hierárquicos, para facilitar pesquisas diversas; ou até mesmo como sistemas de recomendação, indicando certos artigos de interesse de um usuário.

Apesar de ser um problema antigo – estudado desde a década de 1960 – sua importância cresceu apenas na década de 1990, pela grande demanda de aplicações. Com a difusão da *Internet*, houve um crescimento bastante acelerado na quantidade de informação digital, principalmente devido à facilidade em inserir novos documentos neste meio de comunicação.

Pela necessidade de obter essas informações de modo rápido e prático, passou a existir um interesse em desenvolver aplicações voltadas para documentos digitais.

1.2 Motivação

Uma das dificuldades mais marcantes em problemas de TC é a alta dimensionalidade do vetor de características. Tendo em vista que, normalmente, esse vetor é composto por termos (palavras) que aparecem num documento, uma quantidade de dezenas de milhares de termos é facilmente obtida numa base mediana, tornando proibitivo o uso de diversas técnicas de Aprendizagem de Máquina (ML, do inglês *Machine Learning*). Portanto, é extremamente necessária uma redução do vetor de características, desde que esta não comprometa o resultado da classificação. Também é esperado que essa redução seja automática, sem a necessidade de uma intervenção humana.

Métodos automáticos para redução da dimensão do vetor de características são conhecidos na literatura por métodos de Seleção de Características (FS, do inglês *Feature Selection*), sendo estes segmentados em dois tipos: métodos *wrapper* (KOHAVI; JOHN, 1997) e métodos de filtragem (YU; LIU, 2003; ALMUALLIM; DIETTERICH, 1991; KIRA; RENDELL, 1992). É importante salientar, porém, que grande quantidade dos estudos sobre FS são focados em aplicações de domínio não-textual (PENG; LONG; DING, 2005; WANG et al., 2007; BENSCH et al., 2005; WANG et al., 2009). Esses outros domínios possuem, em sua maioria, dimensões bastante inferiores ao problema em estudo e conseguem melhores resultados utilizando métodos *wrapper*.

Métodos *wrapper* – como Seleção Sequencial Crescente (SFS, do inglês *Sequential Forward Selection*) ou Busca Genética – realizam uma busca sobre alguns (ou todos, dependendo do método) subconjuntos de características e avalia-os usando um classificador, selecionando ao final o subconjunto que conseguir a melhor taxa de acerto. Por causa de sua alta complexidade, esses métodos se tornam impraticáveis para problemas de grande escala.

Métodos de filtragem – diferentemente dos métodos *wrapper* – selecionam uma quantidade m de características (definida pelo usuário) sem a necessidade de avaliá-las usando um classificador, realizando essa seleção com algoritmos determinísticos e métricas que avaliam cada característica individualmente. Apesar de sua seleção não ser tão precisa em comparação aos métodos *wrapper*, elas gastam menos tempo de execução, permitindo seu uso em problemas de altíssimas dimensões como TC. Entretanto, mesmo com uma seleção realizada rapidamente, os métodos de filtragem precisam otimizar o parâmetro m. Essa otimização atrasa a definição final do subconjunto de características, além de não garantir o valor ótimo do parâmetro m, salvo

nos casos da busca exaustiva que impossibilitaria o uso dos métodos de filtragem.

Os métodos *wrapper* e os métodos de filtragem podem ser utilizados em conjunto. Um método de filtragem realiza uma varredura inicial para remover o excesso de características, e depois algum método *wrapper* é ser usado numa fase de refinamento. Deste modo, o método *wrapper* terá menos possibilidades em sua busca, reduzindo o custo computacional, e o método de filtragem fará uma seleção menos agressiva, evitando grandes prejuízos na classificação.

1.3 Objetivos

Este trabalho tem por objetivo principal propor uma nova abordagem para selecionar características no domínio textual utilizando métodos de filtragem. Essa nova abordagem deve superar as dificuldades encontradas nos métodos de filtragem clássicos e, por consequência, obter melhores taxas de acerto. Como objetivos específicos, destacam-se:

- Revisar literatura no domínio da classificação de documentos para aprofundar os conhecimentos neste problema específico;
- Estudar os fundamentos de Aprendizagem de Máquina visando o conhecimento dos classificadores mais utilizados em problemas de classificação de documentos;
- Implementar classificadores clássicos;
- Estudar técnicas de seleção de características que utilizem filtragem;
- Realizar experimentos usando diversas configurações a fim de comparar o desempenho das técnicas de seleção de características propostas com técnicas clássicas encontradas na literatura;
- Analisar os resultados obtidos.

1.4 Organização da Dissertação

Neste capítulo introdutório, foi apresentada uma breve descrição do problema estudado, quais motivos tornam este tema digno de um estudo mais aprofundado e quais objetivos este trabalho visa alcançar. No Capítulo 2 é descrito a tarefa de Classificação de Documentos: seus domínios, a arquitetura de um sistema de classificação de documentos, os classificadores utilizados nos experimentos, bem como as medidas para avaliar estes classificadores. No Capítulo

3 é definido o problema de Seleção de Características, apresentando o algoritmo mais utilizado para seleção no domínio estudado e, ao final do capítulo, ainda são vistas diversas funções de avaliação de características. No Capítulo 4 são introduzidos os métodos propostos. No Capítulo 5 são relatados e discutidos os resultados obtidos nos experimentos. Finalmente, o Capítulo 6 fecha o trabalho com as conclusões realizadas e apontando as principais contribuições.

CAPÍTULO 2

Classificação de Documentos

Segundo Sebastiani (SEBASTIANI, 2002), TC é a tarefa de designar um valor booleano (*true* ou *false*) ao par $\langle d_i, c_j \rangle \in \mathscr{D} \times \mathscr{C}$, no qual $\mathscr{D}$ é o conjunto de todos os documentos e $\mathscr{C} = \{c_1, ..., c_N\}$ é o conjunto pré-definido de N categorias. Um valor *true* em $\langle d_i, c_j \rangle$ indica que o documento d_i pertence à classe c_j, enquanto um valor *false* indica que o documento d_i não pertence à classe c_j.

Por ser uma tarefa que pode ser naturalmente modelada como um problema de aprendizado supervisionado, técnicas da área de ML são costumeiramente utilizadas em TC, tais como Árvores de Decisão (LEWIS; RINGUETTE, 1994; APTE; DAMERAU; WEISS, 1998), Redes Neurais (WIENER; PEDERSEN; WEIGEND, 1995; SOUZA et al., 2009), k vizinhos mais próximos (kNN, do inglês *k-Nearest Neighbors*) (TAN, 2006; LAM; HO, 1998), *AdaBoost* (SCHAPIRE; SINGER, 2000), Lógica Fuzzy (ZADROZNY; KACPRZYK, 2006), Máquina de Vetor Suporte (SVM, do inglês *Support Vector Machine*) (GODBOLE; SARAWAGI; CHAKRABARTI, 2002; LEE; KAGEURA, 2007) e *Naïve Bayes* (NB) (MCCALLUM; NIGAM, 1998; LEWIS, 1998).

Dentre as aplicações de TC, destacam-se seu uso como etapa de pré-processamento para indexação de documentos em SRIs, detecção de *spam* em e-mails, organização de uma base de documentos em tópicos hierárquicos para buscas mais intuitivas e sistemas de recomendação de artigos científicos. Todo problema que envolva texto e onde exista uma necessidade de distribuir os documentos em categorias, é aplicação de TC.

Neste Capítulo é apresentado todo o processo para classificar um documento, bem como conceitos básicos para compreensão do processo de classificação. Na Seção 2.1 são descritos os conceitos de domínios, classes e tipos de etiquetagem. Na Seção 2.2 é apresentada de modo abrangente a Arquitetura Geral de um Sistema de Classificação de Documentos (SCD). Na Seção 2.3 inicia-se uma descrição aprofundada dos SCDs com as rotinas de pré-processamento, seguido pela Seção 2.4 que apresenta a etapa de Extração de Características. Ao final do Capítulo, estão os classificadores, na Seção 2.5, e suas medidas de avaliação são vistas na Seção 2.6. A Seção 2.7 traz as conclusões deste capítulo.

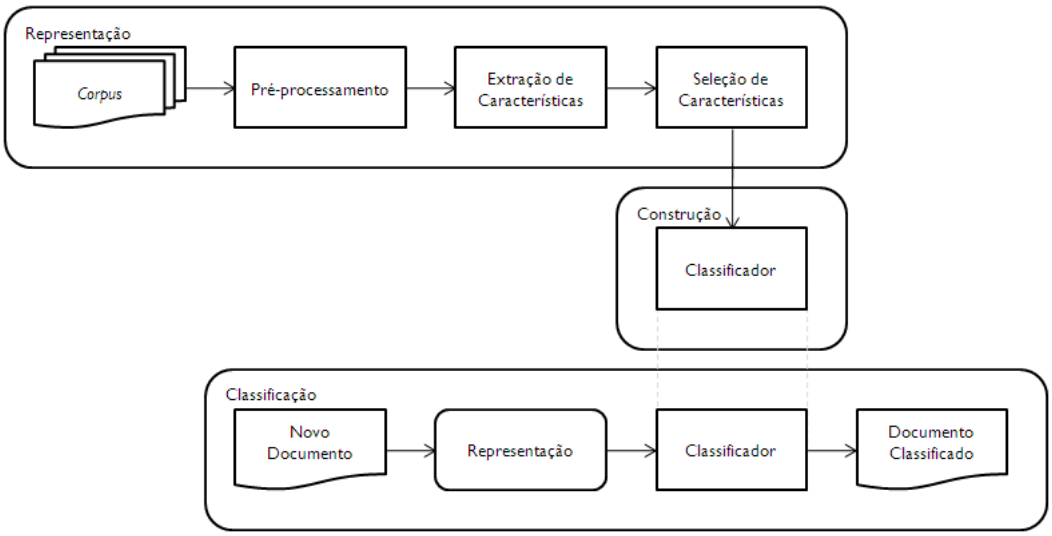

Figura 2.1 Arquitetura de um Sistema de Classificação de Documentos.

2.1 Domínios de Classificação

Domínios são restrições (ou regras) que são usadas no momento da classificação. O domínio é definido pela base de dados em estudo. Existem dois grandes domínios em TC: o binário e o multi-classe (SEBASTIANI, 2002).

A classificação binária é o caso fundamental. Sua tarefa é classificar um determinado documento em uma das duas classes existentes, normalmente chamadas de classe positiva e, a outra, classe negativa. Domínios binários são comuns e aplicáveis até mesmo para um problema multi-classe. Afinal, um problema multi-classe pode ser decomposto em x problemas binários, no qual x é o número de classes pré-definidas e cada problema binário será responsável simplesmente por afirmar se o documento pertence a uma determinada classe ou não.

A classificação multi-classe pode ser utilizada de dois modos: *single-label* (etiquetagem-única) e *multi-label* (etiquetagem-múltipla) (SEBASTIANI, 2002). Na *single-label*, um documento será associado a obrigatoriamente uma, e somente uma, das classes existentes. Já na *multi-label*, um documento pode pertencer a uma, nenhuma, várias ou até mesmo todas as classes existentes.

2.2 Arquitetura Geral

A arquitetura de um Sistema de Classificação de Documentos (SCD) pode ser dividida em, essencialmente, três módulos: um módulo responsável criação da representação dos documentos (Representação); outro módulo encarregado da construção do classificador responsável pela classificação dos documentos (Construção); e o terceiro módulo responsável pela utilização do classificador para, de fato, classificar um novo documento – não classificado – apresentado (Classificação). Detalhes podem ser observados na Figura 2.1.

A Representação é iniciada com um conjunto de documentos escritos em linguagem natural (*Corpus*). Esses documentos passam por diversas rotinas de pré-processamento, que irão tratar os documentos visando uma melhor classificação. Após esse tratamento, uma extração de características é realizada sobre os documentos, modificando sua representação original (linguagem natural) para uma sequência de números (vetor de características). Esse vetor é reduzido por uma rotina de Seleção de Características, criando a representação final dos documentos. Essa representação final é armazenada numa memória principal ou secundária, para ser usada na construção do classificador.

No módulo de Construção o classificador é criado de modo manual, utilizando um conjunto de regras estabelecidas pelo desenvolvedor, ou automático, com técnicas de aprendizado supervisionado, citadas no início deste capítulo. Os vetores de características dos documentos – obtido no módulo de Representação – podem ser utilizadas nessa construção, como no caso das Redes Neurais que a construção do classificador é um processo de treinamento.

No módulo de Classificação, um novo documento é apresentado. O documento sofrerá as operações presentes no módulo de Representação, pois é necessário que siga o mesmo padrão utilizado nos documentos do *Corpus*. O vetor de características do novo documento é apresentado ao classificador e, por fim, o documento é etiquetado. Essa etiquetagem índica que o documento pertence à uma, diversas ou nenhuma das classes, dependendo do domínio utilizado.

Todas as etapas presentes na Figura 2.1 serão descritas em detalhes no decorrer deste trabalho: Pré-processamento na Seção 2.3; a Extração de Características ocupa a Seção 2.4; a Seleção de Características é vista no Capítulo 3; e o Classificador é apresentado na Seção 2.5.

2.3 Pré-processamento

O Pré-processamento contém as primeiras operações realizadas sobre o *Corpus*, sendo executado antes da Extração de Características. Seu objetivo principal é remover informações irrelevantes presentes nos documentos. Por consequência, as modificações realizadas nos textos também irão reduzir a quantidade de informação que será extraída, sendo um importante processo, já que problemas de TC possuem muitas características.

Existem diversas rotinas de pré-processamento de texto. A utilização dessas rotinas deve ser avaliada pelo desenvolvedor, pois dependendo da base de dados certas rotinas podem trazer boas contribuições, enquanto algumas devem ser completamente evitadas. As principais rotinas são:

- **Análise Léxica**

  Converte o texto original para uma lista de palavras. Normalmente tem a tarefa adicional de remover pontuação, hífens e dígitos, além de transformar letras maiúsculas em minúsculas. Entretanto, algumas remoções podem alterar o sentido do texto ou, até mesmo, eliminar informações relevantes: datas importantes seriam perdidas com a remoção dos dígitos; palavras compostas como "estado-da-arte" não teriam sentido separadas com a remoção do hífen; casos específicos da língua, como Banco (instituição) e banco (assento), cuja transformação da letra maiúscula para minúscula resultaria em termos idênticos.

- **Stopwords**

  Palavras que são muito comuns no *Corpus* em questão não possuem um caráter discriminante. O papel das *stopwords* é justamente remover estas palavras. Essa tarefa é feita com base numa lista de palavras – associadas à língua do domínio – mais comuns e que não possuem um significado semântico relevante, como artigos, conjunções, preposições e alguns adjetivos e advérbios.

  Em alguns casos, esse processo pode ser danoso, pois a remoção de um termo pode remover a conexão existente entre outros dois, como é o caso de "redes de computadores" que com a remoção da preposição "de" perderia o significado. Entretanto, esse problema é menos prejudicial quando as palavras são tratadas individualmente, como na abordagem *bag of words* (introduzida na Seção 2.4). Portanto, *stopwords* é uma rotina essencial para diminuir a quantidade de características irrelevantes de modo rápido e eficiente.

- **Stemming**

  Para a tarefa de classificação é irrelevante manter termos de mesmo radical[1], como plurais, mudança de gênero e tempo verbal. A tarefa do *stemming* é justamente reduzir as palavras ao seu radical. Deste modo, a quantidade de informação é reduzida e aglutinada, permitindo que esses novos termos sejam características mais relevantes do que antes da aplicação da rotina *stemming*.

  Um exemplo claro disso seriam as palavras conectar, conectado, conectando, conectados. Todas elas remetem a uma mesma raiz, porém antes da aplicação da rotina *stemming* eram diversas características de pouca relevância. Reduzindo todas as palavras ao seu radical "conect" originaria uma característica de maior relevância, facilitando o trabalho do classificador.

  Vários algoritmos de *stemming* foram propostos, normalmente usando gramáticas (PORTER, 2006) ou regras condicionais (PAICE, 1990). Uma revisão geral dos algoritmos de *stemming* é feita no trabalho de Viera e Virgil (VIERA; VIRGIL, 2007). É importante salientar que, devido a construção dos radicais, prefixos e sufixos variarem para cada língua, a maioria dos algoritmos são criados para solucionar o problema numa determinada língua. Uma proposta robusta foi a utilização de *n*-grams para gerar pseudo-radicais (MAYFIELD; MCNAMEE, 2003). Com essa abordagem, é possível realizar o *stemming* independente da língua, apesar dos resultados serem inferiores com relação a alguns algoritmos específicos de uma língua.

- **Tesauro**

  É um dicionário de sinônimos de uma língua específica. Quando usado nessa etapa de pré-processamento, sua função é diminuir a quantidade de termos, substituindo todos os de mesmo significado por apenas um, reduzindo a dimensionalidade da representação do documento. O uso do Tesauro é recomendado apenas em domínios de vocabulário controlado, quando o desenvolvedor sabe quais palavras serão convertidas, para evitar substituições indesejáveis. Afinal, numa loja de brinquedos não é intessante substituir "boneca" e "jogo" por "brinquedo"[2], pois ambos devem pertencer a categorias diferentes.

  Essa rotina de pré-processamento pode ser implementada com o apoio do WordNet (MILLER, 1990; FELLBAUM, 1998), um grande Tesauro *online* com diversos sinônimos.

[1] O radical, ou raiz, é a parte do verbo ou substantivo que exprime a idéia geral da palavra, ou seja, seu significado mesmo sem o prefixo ou sufixo. É a parte invariável do vocábulo.

[2] Essas palavras são consideradas sinônimos de acordo com o Thesaurus.com.

- **Grupos Nominais**

  Grupos nominais são termos compostos, como “Inteligência Artificial”. Sua utilização é bastante útil em certos problemas, como categorizar artigos de ciência da computação nas suas diversas áreas. A utilização de termos compostos, como “Rede de Computadores”, “Aprendizagem de Máquina” e “Engenharia de Software” facilitariam o trabalho do classificador.

  Para encontrar os grupos nominais, normalmente é criada uma lista de substantivos com potencial para formação de termos compostos, ou é usado um etiquetador indicando que determinado termo é um substantivo. Quando o algoritmo encontra um substantivo desta lista em algum documento, inicia-se uma verificação de termos compostos naquela região. Uma região é definida pelas palavras próximas de um ponto de referência, neste caso um substantivo, cuja distância máxima é definida pelo desenvolvedor. Os diversos substantivos dessa região são agrupados na tentativa de obter termos compostos, levando em consideração a distância entre os substantivos, devido a presença de algum artigo ou preposição, como no caso de “Recuperação de Informação”.

- **Corretor Ortográfico**

  Uma das problemáticas quando se trabalha com texto são os erros ortográficos. A presença desses erros cria características irreais, pois a mesma palavra é escrita de várias maneiras diferentes. O problema é agravado quando os documentos são escritos por diversas pessoas e abrangem diversos temas.

  Uma alternativa para resolver este problema é utilizar um corretor ortográfico em todo o *Corpus*. Entretanto, essa solução pode gerar outros problemas com palavras fora do vocabulário da língua utilizada, como é o caso das siglas. A sigla “ONU” poderia ser modificada para “ônus”, perdendo todo o poder discriminante do termo.

2.4 Extração de Características

Os documentos em sua estrutura original (diversos caracteres compondo palavras e frases) não são bons objetos de entrada para um SCD, tanto por sua complexidade como por seu tamanho. Portanto, todos os documentos são convertidos para uma forma de representação mais compacta: um vetor de características. Essa nova representação para os documentos é criada pela extração de característica.

De acordo com Xue e Zhou (XUE; ZHOU, 2009), quando se trata de TC, a palavra "característica" possui dois significados. Um é referente à unidade que será utilizada para representar o documento (chamado de unidade da característica). O outro foca em qual valor traz melhor representatividade às características (chamado de valor da característica).

Com relação às unidades das características, podem ser citadas:

- **Palavras isoladas**

  Palavras isoladas são sequências de caracteres que iniciam em um caractere alfabético (ou alfanumérico) e são finalizados por algum símbolo terminal (espaço, ponto, vírgula) especificado pelo desenvolvedor da rotina de extração de características. Essa abordagem é conhecida como *bag of words* (BOW), sendo a mais difundida e utilizada em toda literatura de TC (SEBASTIANI, 2002). Apesar da existência de unidades de características mais complexas, os resultados obtidos pela abordagem BOW são bem estabelecidos.

- **Sentenças**

  São cadeias de palavras, como: "o cachorro atravessou a rua" ou "enviou uma carta". A definição do que é uma sentença depende do limite imposto pelas especificações do algoritmo utilizado.

  Como diversas pesquisas foram feitas sobre o uso de sentenças como unidades das características, algoritmos foram propostos e estudados visando formar sentenças mais representativas. Uma das primeiras propostas, para melhorar a composição das sentenças, foram as sentenças sintáticas (DUMAIS et al., 1998; LEWIS, 1992a; SCOTT; MATWIN, 1999), que apesar de ser mais sofisticada por usar a gramática da língua, não trouxe melhorias significativas. Outra proposta foram as sentenças estatísticas (CAROPRESO; MATWIN; SEBASTIANI, 2001; FüRNKRANZ, 1998; MLADENIC; GROBELNIK, 1998), chamadas de *n*-grams, que coleta trechos de *n* em *n* palavras. Melhorias foram reportadas, mas com muita dependência ao valor utilizado no parâmetro *n*. Uma combinação entre as sentenças sintáticas e estatísticas foi proposta (TZERAS; HARTMANN, 1993), mas foram notados diversos problemas e elevado custo computacional.

  Um dos problemas no uso de sentenças é conhecido como a má vizinhança de termos (KAGEURA; UMINO, 1996), que afirma ser difícil perceber quando uma ou mais palavras devem ser unidas ou não. Um exemplo pode ser dado pelas sentenças "professor assistente" e "professor alto". A primeira é uma sentença interessante por abordar o posto de um determinado professor e consequentemente não confundir com "professor associado"

ou “professor pleno”, enquanto a segunda refere-se apenas a uma característica física do professor.

- **n-grams**

  A unidade *n*-grams possui dois significados, podendo *n* ser a quantidade de palavras em sequência para compor uma sentença – como citado no item anterior – ou uma quantidade de caracteres em sequência, como é mostrado no presente tópico. A palavra “teste”, por exemplo, em 3-grams seria representada pelas cadeias tes, est, ste.

  Em algumas situações, é falho utilizar uma representação baseada em palavras ou sentenças. Línguas como chinês ou japonês não fazem uso do espaço, não sendo possível segmentar este tipo de documento por palavras, assim como documentos que utilizem muitos símbolos, como nas linguagens de programação.

  Bons resultados foram atingidos com a abordagem *n*-grams (LODHI et al., 2002). Apesar das possibilidades de características crescerem exponencialmente com o aumento do *n*, apenas uma pequena fração é utilizada pelo conjunto de treinamento e algumas destas ainda podem ser removidas por um processo de seleção de características, devido sua baixa relevância.

Com relação aos valores das características (SALTON; BUCKLEY, 1988), podem ser citados:

- **Binário**

  Um valor binário atribui em cada característica uma representação *true* ou *false*. Esse valor é usado para indicar a presença ou ausência desta característica em um determinado documento.

  É o modelo mais simples, sendo facilmente implementado, pouco custoso em termos de memória e processamento. Entretanto, tamanha simplicidade remete a uma baixa representatividade.

- **Frequência do Termo** (TF, do inglês *Term Frequency*)

  É um modelo mais completo que o binário, pois não indica apenas a presença ou ausência da característica. O valor da frequência do termo será um inteiro representando quantas vezes a característica aparece em um dado documento. Em alguns casos este valor inteiro é normalizado. Obviamente, sua implementação é mais custosa e mais complexa que a do valor binário.

- **TF-IDF** (*Term Frequency-Inverse Document Frequency*)

  É o modelo mais popular, bastante utilizado em conjunto com a abordagem BOW e, normalmente, usado em comparação de desempenho com outros novos métodos propostos (XUE; ZHOU, 2009; WANG; DOMENICONI, 2008). Sua implementação é mais complexa que as anteriores, pois necessita de cálculos menos intuitivos.

  O valor do TF-IDF$_{h,i}$ representa a importância do termo h no documento i. Seu calculo é dado pela equação:

$$\text{TF-IDF}_{h,i} = tf_{h,i} \times idf_h, \tag{2.1}$$

  sendo o valor *term frequency* definido por:

$$tf_{h,i} = \frac{w_{h,i}}{\sum_a^V w_{a,i}}, \tag{2.2}$$

  que representa a quantidade de ocorrências do termo h no documento i, sendo o denominador um fator de normalização dado pelo somatório de todos os termos existentes no documento i ($w_{1,i}, w_{2,i}, ..., w_{V,i}$).

  Enquanto que o valor *inverse document frequency* é definido pela equação:

$$idf_h = \ln \frac{M}{|\{d : w_h \in d\}|}, \tag{2.3}$$

  cujo cálculo é realizado dividindo a quantidade de documentos no *Corpus* (M) pela quantidade de documentos no qual o termo w_h aparece.

2.5 Classificadores

Como visto no início deste capítulo, diversas técnicas de ML são utilizadas em problemas de TC. Entretanto, o foco deste trabalho não é aprofundar um estudo sobre esses classificadores. Os classificadores serão apenas uma ferramenta de comparação para avaliar o desempenho dos algoritmos de seleção de características.

Nesta seção são descritos os dois classificadores utilizados em nossos experimentos: *k-Nearest Neighbors* e *Naïve Bayes*. Esses classificadores foram escolhidos por sua simplicidade e eficiência computacional, sendo de fácil implementação e rápida execução. Adicionalmente,

ambos os métodos sofrem grande influência no desempenho dependendo da seleção de características efetuada, sendo, portanto, classificadores interessantes para nosso estudo.

2.5.1 *k-Nearest Neighbors* (*k*NN)

Considerada a técnica mais comum e popular para TC (SEBASTIANI, 2002), o classificador *k*NN foi proposto em 1967 como um algoritmo de classificação genérico (COVER; HART, 2002). Foi inicialmente aplicado em TC por Masand et al. (MASAND; LINOFF; WALTZ, 1992), e mais tarde foi recomendado como uma solução prática para classificação no domínio dos documentos (YANG, 1999). A partir de então, o classificador *k*NN passou a ser utilizado como método base para comparações e verificações de resultados (SEBASTIANI, 2002).

O procedimento do classificador *k*NN pode ser resumido como: o usuário escolhe um valor para o parâmetro de entrada k, indicando quantos vizinhos serão levados em consideração no processo de classificação. Este classificador não possui uma fase de treinamento, portanto os documentos de teste (não classificados) são apresentados de imediato ao sistema. Deste modo, para cada documento de teste são encontrados os k vizinhos – documentos de treinamento previamente classificados – mais próximos dele. Essa proximidade é baseada em alguma medida de distância ou de similaridade. O documento teste será classificado como sendo pertencente à classe majoritária entre os vizinhos.

A medida mais comum para encontrar os k vizinhos é a distância euclidiana, que é definida por:

$$\text{dist}(x,y) = \sqrt{\sum_{h=1}^{V}(x_h - y_h)^2}, \tag{2.4}$$

V denota o tamanho do vetor de características dos documentos $x = [x_1, x_2, ..., x_V]$ e $y = [y_1, y_2, ..., y_V]$. Todavia, em problemas de TC, a distância euclidiana não é considerada uma boa medida. Isso se deve ao excesso de características que possuem valor zero, resultando uma proximidade entre documentos de classes distintas. Por exemplo: considere $d_1 = \{1,0,1,0,0,0,0,0,0,0\}$ da classe A e $d_2 = \{0,0,0,0,0,0,0,1,1,0\}$ da classe B, a distância entre eles é $1,41$. Agora surge $d_3 = \{1,1,0,1,0,0,0,0,0,0\}$ da classe A, cuja distância para d_1 é $1,73$. Então, mesmo d_3 sendo da mesma classe que d_1, o resultado do cálculo das distâncias não equivale à realidade, pois $\text{dist}(d_1,d_2) < \text{dist}(d_1,d_3)$ quando o necessário para uma classificação correta seria $\text{dist}(d_1,d_2) > \text{dist}(d_1,d_3)$.

Portanto, é comum adotar a medida de similaridade do cosseno (SALTON; MCGILL, 1983) em vez da distância euclidiana. O cálculo desta medida é obtido através da equação:

$$\text{sim}(x,y) = \frac{\sum_{h=1}^{V} x_h y_h}{\sqrt{\sum_{h=1}^{V} x_h^2}\sqrt{\sum_{h=1}^{V} y_h^2}}. \tag{2.5}$$

Considerando o mesmo exemplo dos três documentos (d_1, d_2, d_3), os resultados usando a similaridade do cosseno se adéquam melhor à realidade do problema: $\text{sim}(d_1, d_2) = 0$ enquanto que $\text{sim}(d_1, d_3) = 0,41$. Pois a medida de similaridade do cosseno se concentra apenas nas semelhanças de valores diferentes de zero, evitando o excesso de valores zero presente nos problemas de TC.

Definidas as modificações feitas sobre o *k*NN para adequá-lo ao domínio, o restante do classificador segue o método convencional.

Para classificar um documento de teste d_i, o *k*NN determina sua classe como:

$$\text{label}(d_i) = \arg\max_{c_j} \sum_{d_a \in kNN(d_i)} \delta(d_a, c_j), \tag{2.6}$$

no qual $kNN(d_i)$ é uma função que retorna o conjunto dos k vizinhos mais próximos ao documento d_i. Esses documentos vizinhos são representados por d_a e são encontrados pela função de similaridade do cosseno. A função $\delta(d_a, c_j)$ é uma função binária com respeito à classificação do documento vizinho d_a com relação a classe c_j, sendo definida por:

$$\delta(d_a, c_j) = \begin{cases} 1 & d_a \in c_j \\ 0 & d_a \notin c_j \end{cases} \tag{2.7}$$

2.5.2 *Naïve Bayes*

Existem dois diferentes modelos de classificadores *Naïve Bayes* criados especificamente para problemas de TC: *Multi-Variate Bernoulli Event Model* e *Multinomial Event Model* (MCCALLUM; NIGAM, 1998). Além da diferença dos cálculos das probabilidades, a maior diferença entre os dois modelos é o valor das características utilizado em cada um. No *Multi-Variate Bernoulli Event Model* são utilizados valores binários, trabalhando apenas com a ausência e com a presença dos termos. Já no *Multinomial Event Model* usam-se as frequências dos termos, isto é, números inteiros que representam a quantidade de ocorrências de um determinado termo no documento. Este trabalho é focado no segundo modelo devido à sua melhor performance (CHEN et al., 2009).

Para classificar um documento teste d_i, o classificador Naïve Bayes determina a classe do documento por meio da seguinte equação:

$$\text{label}(d_i) = \arg\max_{c_j} \left\{ P(c_j|d_i) \right\}. \tag{2.8}$$

A regra de Bayes é definida por

$$P(c_j|d_i) = \frac{P(d_i|c_j)P(c_j)}{P(d_i)}, \tag{2.9}$$

sendo $P(d_i)$ a probabilidade *a priori*, cujo valor é igual para todas as classes, assim este termo pode ser removido da equação. Como resultado, a regra de decisão final obtida é:

$$\text{label}(d_i) = \arg\max_{c_j} \left\{ P(d_i|c_j)P(c_j) \right\}. \tag{2.10}$$

Até então as equações são idênticas nos dois modelos *Naïve Bayes*. Mas a partir deste ponto, os cálculos são exclusivos do modelo *Multinomial Event Model*.

Para resolver (2.10) é preciso computar as probabilidades. A probabilidade $P(c_j)$ é facilmente encontrada ao dividir o número de documentos existentes na classe c_j pelo número de documentos existentes em todo o corpus. A probabilidade $P(d_i|c_j)$ pode ser computada como:

$$P(d_i|c_j) = P(|d_i|)|d_i|! \prod_{h=1}^{V} \frac{P(w_h|c_j)^{n_{ih}}}{n_{ih}!}, \tag{2.11}$$

no qual $|d_i|$ é a soma de todos os valores das características do documento d_i. V é o tamanho do vetor de características e n_{ih} é o número de vezes que o termo w_h aparece no documento d_i. A probabilidade $P(w_h|c_j)$ é estimada como

$$P(w_h|c_j) = \frac{1 + N_{c_j h}}{V + N_j}, \tag{2.12}$$

no qual $N_{c_j h}$ é o número de vezes que o termo w_h aparece na classe c_j, isto é, o somatório dos pesos do termo w_h em todos os documentos cuja classe seja c_j. N_j é o número de termos na classe c_j, isto é, o somatório de todos os pesos de todos os documentos cuja classe seja c_j.

2.6 Avaliação dos Classificadores

Os Sistemas de Classificação de Documentos seguem o mesmo esquema para avaliação de desempenho utilizada pelos SRIs. Essa avaliação é experimental, pois para realizar uma análise analítica seria necessário formalizar um problema que é naturalmente informal (SEBASTIANI, 2002).

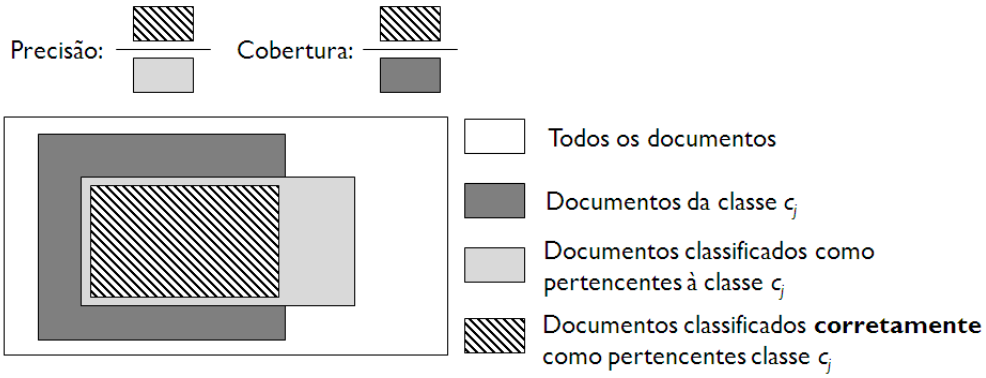

Figura 2.2 Medidas de precisão e cobertura.

Portanto, a avaliação experimental de um classificador num problema de TC é dada por medidas cujo objetivo é capturar o quão corretas são as decisões tomadas pelo classificador. As principais medidas utilizadas são a precisão e a cobertura (do inglês *recall*) (RIJSBERGEN, 1979), cujos valores são calculados para cada classe pelas equações:

$$p_j = \frac{TP_j}{TP_j + FP_j}, \tag{2.13}$$

$$r_j = \frac{TP_j}{TP_j + FN_j}, \tag{2.14}$$

Classe c_j		Resposta Real	
		Sim	Não
Resposta do Classificador	Sim	TP_j	FP_j
	Não	FN_j	TN_j

Tabela 2.1 Tabela de Contingência da classe c_j

no qual os termos de ambas as equações são esclarecidos na Tabela 2.1 e a Figura 2.2 ilustra como as medidas são obtidas. Portanto, o valor máximo da precisão ($p_j = 1$) ocorre quando todos os documentos da classe c_j, e somente estes, são classificados como pertencente à classe c_j. Enquanto que o valor máximo da cobertura ($r_j = 1$) ocorre quando todos os documentos da classe c_j são classificados como pertencente à classe c_j.

Encontrar um valor máximo de cobertura é trivial, pois ao classificar todos os documentos como pertencentes à classe c_j fará $r_j = 1$. Portanto, o valor de cobertura sozinho não é suficiente para indicar o desempenho de um SCD. Deste modo, é utilizada a *F-Measure* (ou F1), cuja função é unir os valores de cobertura e precisão numa única medida de avaliação, calculada por:

$$F1_j = \frac{2 \times p_j \times r_j}{(p_j + r_j)}. \tag{2.15}$$

Entretanto, como os SCD lidam com diversas classes, apenas os valores isolados da F1 de cada classe não mede o desempenho do sistema completo. Portanto, para avaliar o desempenho médio do sistema, são utilizadas a macro-média e a micro-média (SHANG et al., 2007; ÖZGÜR; GÜNGÖR, 2010; BEKKERMAN, 2003; XUE; ZHOU, 2009; ROGATI; YANG, 2002; CHEN et al., 2009; TAN, 2006; MANNING; RAGHAAN; SCHÜTZE, 2008). A macro-média (macro-F1) é obtida com uma média direta dos valores F1 de cada classe, seguindo a equação:

$$\text{macro-F1} = \sum_{j}^{N} \frac{F1_j}{N}, \tag{2.16}$$

e a micro-média é calculada substituindo os valores presentes na Tabela 2.1 pelos valores da tabela de contingência global (Tabela 2.2) e utilizando-os nas Equações (2.13), (2.14) e (2.15). Deste modo, serão obtidos os valores de *micro-precision*, *micro-recall* e micro-F1, sendo este último o valor da medida de avaliação do sistema.

Conjunto de Classes $\mathscr{C} = \{c_1, ..., c_N\}$		Resposta Real	
		Sim	Não
Resposta do Classificador	Sim	$TP = \sum_{j=1}^{N} TP_j$	$FP = \sum_{j=1}^{N} FP_j$
	Não	$FN = \sum_{j=1}^{N} FN_j$	$TN = \sum_{j=1}^{N} TN_j$

Tabela 2.2 Tabela de Contingência Global

Em resumo, a macro-F1 é uma média não ponderada pela quantidade de documentos presentes nas classes, enquanto a micro-F1 é uma média ponderada pela quantidade de documentos presentes nas classes. Sendo assim, um SCD é avaliado por ambas as medidas, pois apenas uma delas não é capaz de descrever o desempenho de um sistema que possua classes com diferentes quantidades de documentos. Os valores são todos no intervalo $[0, 1]$, e quanto mais próximo de 1 for este valor, melhor a classificação. No caso de problemas *single-label*, a micro-F1 tem o mesmo valor da taxa de acerto (XUE; ZHOU, 2009).

2.7 Considerações Finais

Neste capítulo foram apresentadas as seguintes etapas de um Sistema de Classificação de Documentos: Pré-processamento, que utiliza diversas rotinas para transformar os documentos originais em documentos mais simplificados, visando facilitar o trabalho das etapas subse-

quentes; Extração de Características, responsável pela geração do vetor de características; e a construção do Classificador, responsável pela classificação dos novos documentos. A descrição completa de um SCD será concluída no próximo capítulo com a apresentação da etapa de Seleção de Característica.

CAPÍTULO 3

Seleção de Características

A tarefa de redução de dimensionalidade pode ser dividida em duas grandes abordagens: extração e seleção.

A extração de características é a tarefa de gerar um novo conjunto de características $\mathscr{T}'$ baseado em combinações, transformações ou mudança de representação do conjunto original $\mathscr{T}$, de modo que $|\mathscr{T}'| \ll |\mathscr{T}|$. As características do novo conjunto $\mathscr{T}'$ são sintéticas, isto é, são unidades desconhecidas (não mais palavras, frases ou letras) geradas pelo processo de extração. Semântica latente (WEIGEND; WIENER; PEDERSEN, 1999; WIENER; PEDERSEN; WEIGEND, 1995; SCHÜTZE; HULL; PEDERSEN, 1995), *clustering* de termos (BAKER; MCCALLUM, 1998; LEWIS, 1992a; SLONIM; TISHBY, 2001; BEKKERMAN, 2003), mapeamento semântico (CORRÊA; LUDERMIR, 2007, 2006) e Análise de Componentes Principais (PCA) (SCHÜTZE; HULL; PEDERSEN, 1995; WIENER; PEDERSEN; WEIGEND, 1995) são técnicas de extração de características aplicadas em TC.

Entretanto, neste trabalho é estudado a seleção de características, cujo objetivo é gerar um novo conjunto de características $\mathscr{T}'$ que é um subconjunto de $\mathscr{T}$ (conjunto original), de modo que $|\mathscr{T}'| \ll |\mathscr{T}|$. As características do novo conjunto $\mathscr{T}'$ são características previamente existentes em $\mathscr{T}$, portanto nenhuma informação nova foi gerada, houve simplesmente uma redução do conjunto original por meio da seleção.

A seleção de características pode ser global ou local. Na seleção local, é realizada uma seleção para cada classe existente, cada seleção irá considerar uma classe como positiva e as demais como negativas. Portanto, a seleção local terá um subconjunto de características $\mathscr{T}'_i$ para cada classe c_i (LI; DONG; HUA, 2008) e estes subconjuntos são utilizados para classificações no domínio binário (Seção 2.1). Já a seleção global considera todas as classes de uma única vez, possuindo apenas um subconjunto de características $\mathscr{T}'$ (CHEN et al., 2009), que pode ser utilizado para classificações em domínios binários ou multi-classe.

O processo de seleção de características pode ser baseado em métodos de filtragem ou métodos *wrappers*. A seleção baseada em método de filtragem (YU; LIU, 2003) seleciona os termos baseado em algum algoritmo ou propriedade estatística presente nos dados. Já a seleção baseada em métodos *wrapper* (KOHAVI; JOHN, 1997) é mais complexa, pois utiliza classifi-

cadores para avaliar diversos subconjuntos de $\mathscr{T}$, de modo a escolher o subconjunto que apresentar melhor resultado dentro do conjunto de treinamento. Os métodos *wrappers* conseguem, de modo geral, melhores seleções do que os métodos de filtragem, entretanto seu custo computacional é mais elevado. Assim, os métodos de filtragem são mais utilizados em TC, devido ao tamanho do vetor de características (TANG et al., 2005).

Este trabalho é focado nos métodos de filtragem de seleção de características, por sua simplicidade e velocidade. Apesar de vários algoritmos de filtragem terem sido propostos no decorrer dos anos (como FOCUS (ALMUALLIM; DIETTERICH, 1991, 1994; ARAUZO; BENITEZ; CASTRO, 2003) e RELIEF (KIRA; RENDELL, 1992; KONONENKO, 1994)), estes não conseguiram ganhar espaço em problemas de TC, por seu elevado custo computacional ao lidar com alta dimensionalidade (KIRA; RENDELL, 1992). Na Seção 3.1 deste Capítulo será apresentada a abordagem que é constantemente utilizada na literatura (YANG; PEDERSEN, 1997; ROGATI; YANG, 2002; FORMAN, 2003; MLADENIC; GROBELNIK, 1999; XUE; ZHOU, 2009; CHEN et al., 2009), e na Seção 3.2 serão descritas algumas funções de avaliação de característica (FEF, do inglês *feature evaluation function*). A Seção 3.3 traz as conclusões deste capítulo.

3.1 Abordagem Clássica

Até o início dos anos 90, a abordagem típica para seleção de características (KIRA; RENDELL, 1992) era utilizar uma função $J(\mathscr{E}, \mathscr{D}_{tr})$. Esta função avaliaria a qualidade do subconjunto de características $\mathscr{E}$ com relação ao conjunto de treinamento $\mathscr{D}_{tr}$ (YAO, 2003). Um subconjunto $\mathscr{E}_1$ é considerado melhor que o subconjunto $\mathscr{E}_2$ se $J(\mathscr{E}_1, \mathscr{D}_{tr}) > J(\mathscr{E}_2, \mathscr{D}_{tr})$. Portanto, necessário avaliar, com a função $J(\mathscr{E}, \mathscr{D}_{tr})$, diversos subconjuntos $\mathscr{E}$ para encontrar uma boa seleção.

A abordagem modificou-se com o trabalho de Lewis e Ringuette (LEWIS; RINGUETTE, 1994), que utilizou *information gain* para selecionar as características. Mais tarde, foram usadas *mutual information* e *chi-squared statistic* (SCHÜTZE; HULL; PEDERSEN, 1995). O método passou a ser amplamente usado e métricas foram comparadas (YANG; PEDERSEN, 1997; ROGATI; YANG, 2002). Por sua simplicidade, o método sempre foi chamado pelo nome da métrica utilizada. Neste trabalho, entretanto, será usada a nomenclatura de Guyon e Elisseeff (GUYON; ELISSEEFF, 2003), reconhecendo este método pelo nome *Variable Ranking* (VR).

No Algoritmo 1 é descrito o funcionamento do método de seleção VR. Nele há um conjunto de treinamento $\mathscr{D}_{tr}$, composto por documentos $d \in \mathbb{N}^V$, V é o tamanho do vocabulário (número de características). O w_h presente na linha 3, representa a h-ésima característica do vocabulário.

Algoritmo 1 *Variable Ranking*

Input: Integer $m > 0$

Load training set $\mathscr{D}_{tr}$
for $h = 1$ to V **do** // Calcula FEF para cada termo
  $S_h = \text{FEF}(w_h)$
end for
for $i = 1$ to m **do** // Seleciona as m características de maior FEF.
  $SN = S$
  $bestscore = 0.0$
  for $h = 1$ to V **do**
    if $SN_h > bestscore$ **then**
      $bestscore = SN_h$
      $bestfeature = h$
    end if
  end for
  $SN_{bestfeature} = 0$
  $FS_i = bestfeature$
end for
execute Novo_Conjunto($\mathscr{D}_{tr}$,FS,m)

Algoritmo 2 Função Novo_Conjunto()

Input: Dataset $\mathscr{D}_{or}$, Vector FS and Integer $m > 0$

$\mathscr{D}_{nv}$ an empty dataset
for all $d \in \mathscr{D}_{or}$ **do**
  Insert document d' in $\mathscr{D}_{nv}$
  for $h = 1$ to m **do**
    $d'_h = d_{FS_h}$
  end for
end for
return $\mathscr{D}_{nv}$

O vetor *SN* é uma cópia do vetor *S* utilizado apenas para não corromper os dados do vetor original, que representa a importância de cada característica.

Apesar das 17 linhas do algoritmo, seu funcionamento é bastante simples. O usuário declara quantas características o algoritmo deverá selecionar (*m*). Uma FEF (*feature evaluation function*) (CAROPRESO; MATWIN; SEBASTIANI, 2001; YANG; PEDERSEN, 1997) será utilizada para calcular a importância de cada termo (linhas 2-4) (uma explicação detalhada sobre FEFs é feita na Seção 3.2). Então serão selecionados os *m* termos que possuírem os maiores valores na FEF (linhas 5-16). Por fim, o novo conjunto de treinamento é criado com base nas características selecionadas (linha 17). A função utilizada nesta última etapa é descrita no Algoritmo 2, cuja tarefa é simplesmente gerar o novo conjunto utilizando apenas as características selecionadas.

Outras implementações deste método modificam o parâmetro de entrada *m* por um limiar ponto flutuante. Esse limiar definirá não mais a quantidade de características, mas sim qual será o ponto de corte – com base na FEF – para que uma característica seja inserida (ou não) no novo subconjunto. Esse tipo de implementação é interessante caso o desenvolvedor conheça bem a FEF utilizada, pois, deste modo, poderia inserir um limiar capaz de selecionar apenas as características relevantes.

3.1.1 Dificuldades do método VR

Apesar do método VR (Seção 3.1) ser amplamente utilizado, essa abordagem possui uma grande dificuldade: encontrar o melhor valor para *m*. Todos os subconjuntos de características encontrados com diferentes valores de *m* devem ser testados para encontrar o melhor subconjunto. Entretanto, isso transforma o método numa busca exaustiva, sendo necessário avaliar cada subconjunto com um classificador. Deste modo, um método de filtragem passa a funcionar como um método *wrapper*, perdendo sua vantagem, que é justamente o baixo custo computacional. Portanto, para evitar a busca exaustiva, apenas alguns valores de *m* (definidos pelo usuário) são avaliados, e o melhor selecionado, sendo improvável obter o melhor subconjunto.

Outra dificuldade é utilizar as FEFs como verdade absoluta. O método faz sua seleção com base apenas nos valores da FEF atual. Então, se a FEF escolhida não for adequada, a seleção não será adequada também.

3.2 FEFs

Diversas FEFs foram propostas no decorrer dos anos e estudos comparativos foram realizados sobre elas na literatura (YANG; PEDERSEN, 1997; ROGATI; YANG, 2002; FORMAN, 2003; MLADENIC; GROBELNIK, 1999). Essas funções são baseadas em medidas estatísticas que utilizam as probabilidades de um termo w com relação aos documentos de uma classe c_j, no qual $j = 1, 2, ..., N$. São probabilidades utilizadas nestes cálculos:

- $P(w|c_j)$ a probabilidade do termo w estar presente em documentos da classe c_j;
- $P(w|\bar{c}_j)$ a probabilidade do termo w não estar presente em documentos da classe c_j;
- $P(\bar{w}|c_j)$ é a probabilidade de todos os termos – exceto o termo w – estarem presentes nos documentos da classe c_j;
- $P(\bar{w}|\bar{c}_j)$ é a probabilidade de todos os termos – exceto o termo w – não estar presente nos documentos da classe c_j;
- $P(c_j)$ e $P(w)$ são a probabilidade *a priori* da classe c_j e do termo w, respectivamente.

Todas as métricas apresentadas neste trabalho são para domínios multi-classe (Seção 2.1). Algumas delas foram, inclusive, convertidas do domínio binário para se adequar aos problemas multi-classe estudados neste trabalho, portanto o caminho inverso pode ser refeito facilmente. São métricas utilizadas em problemas de TC:

- ***Bi-Normal Separation***

  Proposta por George Forman (FORMAN, 2003, 2007), a *Bi-Normal Separation* (BNS) é uma métrica definida pela equação:

$$\mathrm{BNS}(w) = \sum_{j=1}^{C} \left| F^{-1}(P(w|c_j)) - F^{-1}(P(w|\bar{c}_j)) \right|, \tag{3.1}$$

  na qual $F^{-1}(\cdot)$ é função de probabilidade acumulativa inversa de uma distribuição normal padrão, também chamada de *z-score*. A BNS, portanto, mede a separação entre os dois limiares encontrados pelas funções.

  Para evitar o valor indefinido de $F^{-1}(0)$, Forman recomenda que o valor 0 seja substituído por $0,0005$ (FORMAN, 2003).

- ***Chi-Squared Statistics***

  Usada constantemente em problemas de TC (CAROPRESO; MATWIN; SEBASTIANI, 2001; YANG; PEDERSEN, 1997), a CHI apresentou bons resultados ao ser comparado com outras métricas clássicas (ROGATI; YANG, 2002). A *Chi-squared Statistics* (CHI) pode ser definida por:

$$\mathrm{CHI}(w) = \sum_{j=1}^{C} \frac{\left[P(w|c_j)P(\bar{w}|\bar{c_j}) - P(w|\bar{c_j})P(\bar{w}|c_j)\right]^2}{P(w)P(\bar{w})P(c_j)P(\bar{c_j})}, \tag{3.2}$$

  na qual as probabilidades seguem a mesma descrição apresentada no início desta seção.

  O valor obtido pela Equação (3.2) é referente a independência do termo w com relação às classes existentes. Portanto, quanto menor o valor retornado pela CHI, mais independente é aquele termo. Como uma boa seleção se caracteriza por encontrar termos que dependam da classe, isto é, que possuam uma forte relação com certas classes, serão selecionadas as características com elevado valor na CHI (DEBOLE; SEBASTIANI, 2003).

- ***Odds Ratio***

  A *Odds Ratios* (CAROPRESO; MATWIN; SEBASTIANI, 2001; CHEN et al., 2009; MLADENIC, 1998) é uma função de domínio binário, descrita por:

$$\mathrm{OR}(w) = \log \frac{P(w|c_j)(1-P(w|\bar{c_j}))}{P(w|\bar{c_j})(1-P(w|c_j))}. \tag{3.3}$$

  Neste trabalho será usada a *Multi-class Odds Ratio* (MOR) (CHEN et al., 2009), sua versão multi-classe, descrita por:

$$\mathrm{MOR}(w) = \sum_{j=1}^{C} \left| \log \frac{P(w|c_j)(1-P(w|\bar{c_j}))}{P(w|\bar{c_j})(1-P(w|c_j))} \right|. \tag{3.4}$$

- ***Class Discriminanting Measure***

  Proposta por Chen et al. (CHEN et al., 2009), a *Class Discriminanting Measure* (CDM) segue como:

$$\mathrm{CDM}(w) = \sum_{j=1}^{C} \left| \log \frac{P(w|c_j)}{P(w|\bar{c_j})} \right|, \tag{3.5}$$

É notável sua semelhança com a MOR, pois a CDM é sua versão simplificada (CHEN et al., 2009).

- ***Document Frequency***

  Apesar de ser utilizada em alguns trabalhos da literatura como um método de FS (CAROPRESO; MATWIN; SEBASTIANI, 2001; ROGATI; YANG, 2002; YANG; PEDERSEN, 1997), a função *Document Frequency* (DF) também é considerada uma etapa de pré-processamento dos documentos a serem classificados. Para tal, um limiar é definido e todos os termos com DF inferior ao limiar são removidos. Sua definição é bastante simplória e tem como função simplesmente contar quantas ocorrências o termo w teve em todo o *Corpus*.

- ***Information Gain***

  Juntamente com a *Chi-squared Statistics*, esta métrica foi reportada como uma das melhores medidas em problemas de multi-classe (YANG; PEDERSEN, 1997). Seu uso permanece frequente em trabalhos recentes (CAROPRESO; MATWIN; SEBASTIANI, 2001; CHEN et al., 2009; ROGATI; YANG, 2002).

  Sua equação é descrita de várias formas na literatura, sendo a mais comum:

$$\mathrm{IG}(w) = \sum_{j=1}^{C} P(w|c_j) \log \frac{P(w|c_j)}{P(c_j)} + \sum_{j=1}^{C} P(\bar{w}|c_j) \log \frac{P(\bar{w}|c_j)}{P(c_j)}. \tag{3.6}$$

  A Equação (3.6) terá resultado 0 quando for completamente independente da classe e crescerá monotonicamente de acordo com sua dependencia (COVER; THOMAS; MYILIBRARY, 1991).

- **Outras Métricas**

  Além das já listadas, outras FEFs são: *Accuracy* (FORMAN, 2003) e sua versão balanceada (MLADENIC; GROBELNIK, 1999), *Mutual Information* (CHANG; CHEN; LIAU, 2008; BEKKERMAN, 2003; YANG; PEDERSEN, 1997), *Gini index* e suas diversas reformulações (PARK; KWON; KWON, ; OGURA; AMANO; KONDO, 2009; SHANG et al., 2007), *Term Strength* proposta por Wilbur e Sirotkin (WILBUR; SIROTKIN, 1992) e utilizada em diversos outros trabalhos (YANG, 1995; YANG; WILBUR, 1996; YANG; PEDERSEN, 1997).

3.3 Considerações Finais

Neste capítulo foi apresentada a etapa de Seleção de Características de um Sistema de Classificação de Documentos. Nessa etapa, o vetor de características será reduzido pelo método clássico *Variable Ranking* (VR). O método VR possui problemas com otimização de parâmetros e utilização da FEF. No próximo capítulo serão apresentados novos métodos, que foram propostos para superar as dificuldades do método VR.

CAPÍTULO 4

Métodos Propostos

Como visto na Seção 3.1.1, foi constatado que o método clássico VR possui certas dificuldades. Os métodos apresentados neste capítulo foram propostos para superar essas dificuldades. Todos os métodos propostos derivam de uma idéia principal, introduzida pelo método ALOFT, portanto, é recomendada a leitura desse método, apresentado na Seção 4.1. Na Seção 4.2 é apresentada a generalização do primeiro método, na Seção 4.3 sua versão local, na Seção 4.4 sua versão restrita e na Seção 4.5 sua versão local restrita. A Seção 4.6 traz as conclusões deste capítulo.

Todos os métodos apresentados neste capítulo foram desenvolvidos para problemas *single-label* (termo explicado na Seção 2.1). Entretanto, os métodos propostos podem ser utilizados em *multi-label* sem nenhuma conversão, exceto as versões locais, cuja conversão é apresentada em detalhes na Seção 4.3.2.

4.1 ALOFT (*At Least One Feature*)

Como apresentado na Seção 3.1.1, uma das maiores dificuldades do VR é encontrar a melhor quantidade de características a ser selecionada (parâmetro m). Para solucionar este problema, foi idealizada uma abordagem automatizada, capaz de gerar um subconjunto de características $\mathscr{T}'$ sem a definição inicial do tamanho desse subconjunto.

A idéia central é varrer todos os documentos de treinamento – um por um – e selecionar uma característica valorada[1] de cada documento. Assim, ao final da seleção, cada documento possuirá uma representação de, pelo menos, uma característica valorada, daí o nome do método ser *At Least One Feature*.

Essa estratégia visa garantir alguns pontos:

- Cada um dos documentos de treinamento serão representados por pelo menos uma característica valorada;

[1] Uma característica valorada é uma característica cujo valor seja diferente de zero.

Algoritmo 3 ALOFT

1: Load training set $\mathscr{D}_{tr}$
2: **for** $h = 1$ to V **do** // Calcula FEF para cada termo
3: $S_h = \text{FEF}(w_h)$
4: **end for**
5: $m = 0$
6: FS an empty vector
7: **for all** $d_i \in \mathscr{D}_{tr}$ **do** // Seleciona, de cada documento, a característica valorada de maior FEF.
8: $bestscore = 0.0$
9: **for** $h = 1$ to V **do**
10: **if** $w_{h,i} > 0$ and $S_h > bestscore$ **then**
11: $bestscore = S_h$
12: $bestfeature = h$
13: **end if**
14: **end for**
15: **if** $bestfeature \notin FS$ **then**
16: $m = m + 1$
17: $FS_m = bestfeature$
18: **end if**
19: **end for**
20: **execute** Novo_Conjunto($\mathscr{D}_{tr}$,FS,m)

- O algoritmo não requer um limiar ou quantidade de características a serem selecionadas. Em outras palavras, ele será capaz de encontrar um subconjunto de termos sem uma busca por parâmetros adequados. A quantidade de características será descoberta no final da execução do algoritmo;
- A seleção será uma busca não exaustiva, pois as características são encontradas em cada documento sem a necessidade de revisões ou prolongamentos;
- O algoritmo é determinístico, provendo uma solução única para cada base de treinamento que lhe seja apresentada, dependendo da FEF apenas.

O método ALOFT é descrito no Algoritmo 3. Um conjunto de treinamento $\mathscr{D}_{tr}$ é carregado (linha 1) sendo composto de $d_i \in \mathbb{N}^V$ documentos, onde V é o tamanho do vocabulário (número de características). As características de um documento d_i podem ser indexadas como $w_{h,i}, h = 1,\ldots,V$. O algoritmo é dividido em três passos:

i) Linhas 2-4: calcula $S \in \mathbb{R}^V$ para cada característica com alguma FEF (descritas na Seção 3.2). O valor de S_h representa a importância da h-ésima característica (w_h) com

relação ao conjunto de características. A FEF influencia diretamente no quão efetiva será a seleção;

ii) Linhas 5-19: gera o vetor FS, que representa o novo conjunto de características. É selecionada uma característica por documento, baseado no maior valor de S dentre todas as características valoradas. Se a característica selecionada já estiver em FS, ela é ignorada e segue-se para o próximo documento. No final desta etapa, FS deverá conter m valores, e estes valores representam os índices das características selecionadas do conjunto original;

iii) Linha 20: constrói o novo conjunto de treinamento, usando a função descrita no Algoritmo 2 (Seção 3.1). Este novo subconjunto será composto de $d' \in \mathbb{N}^m$ documentos, no qual m é o número de características selecionadas pelo método. O novo conjunto de teste será gerado pela função Novo_Conjunto(T, FS, m), no qual T é o conjunto de teste original.

4.1.1 Passo-a-passo

Visando uma explicação mais detalhada do algoritmo proposto, um conjunto de treinamento hipotético (Tabela 4.1) foi criado. Este conjunto é composto de 13 documentos, cada um deles representado por 9 características com valores booleanos (presença ou ausência do termo no documento), sendo S a mesma variável presente no Algoritmo 3, mas composta por valores inteiros (em vez dos pontos flutuantes) para melhor visualização.

O primeiro passo do algoritmo é calcular os valores do vetor S. Uma FEF é utilizada para isso. Entretanto, neste exemplo é usado os valores arbitrários mostrados na última linha da Tabela 4.1.

O segundo passo é preencher o vetor FS seguindo o fluxo do Algoritmo 3. Para tal, cada documento é visitado e a melhor característica valorada, baseada no vetor S, é selecionada. O documento 1 possui w_2 e w_7 como características valoradas, então w_2 é selecionado, pois $S_2 = 7 > S_7 = 1$. O vetor FS é atualizado com o índice da característica selecionada: $FS = \{2\}$. No documento 2, w_4 é selecionado, fazendo $FS = \{2,4\}$. No documento 3, w_4 é selecionado novamente. Entretanto, o índice 4 já existe no vetor FS. Portanto, o algoritmo ignora essa seleção e passa para o próximo documento. No documento 4, w_2 seria o selecionado, mas seu índice já existe em FS, novamente a seleção é ignorada. No documento 5, w_8 é selecionado, obtendo $FS = \{2,4,8\}$. No documento 6, w_6 é selecionado, então $FS = \{2,4,8,6\}$. O restante da varredura encontrará apenas características já presentes no vetor FS.

Tabela 4.1 Conjunto de treinamento hipotético. Primeira coluna representa o índice para identificar o documento, na última coluna estão as classes de cada documento e nas colunas centrais são os pesos das características. Cada linha representa um documento, com exceção da última, que representa o vetor S.

$\mathscr{D}_{tr}$	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	C
d_1	0	1	0	0	0	0	1	0	0	A
d_2	0	0	0	1	1	0	0	1	0	A
d_3	1	0	0	1	0	0	0	1	1	A
d_4	0	1	0	0	0	0	0	0	0	A
d_5	0	0	0	0	0	0	1	1	0	A
d_6	0	0	0	0	0	1	0	0	0	B
d_7	0	0	0	1	0	1	0	0	1	B
d_8	1	0	1	1	1	0	0	0	1	B
d_9	0	0	0	0	0	1	0	0	0	B
d_{10}	0	0	0	0	0	1	1	0	0	B
d_{11}	1	0	1	1	1	0	1	0	0	B
d_{12}	0	0	0	1	1	0	0	0	1	B
d_{13}	1	0	0	1	0	1	0	0	0	B
S	11	7	4	15	10	8	1	5	13	-

Tabela 4.2 Resultado da seleção efetuada pelo ALOFT no conjunto de treinamento apresentado na Tabela 4.1.

$\mathscr{D}'_{tr}$	w_2	w_4	w_8	w_6	C
d_1	1	0	0	0	A
d_2	0	1	1	0	A
d_3	0	1	1	0	A
d_4	1	0	0	0	A
d_5	0	0	1	0	A
d_6	0	0	0	1	B
d_7	0	1	0	1	B
d_8	0	1	0	0	B
d_9	0	0	0	1	B
d_{10}	0	0	0	1	B
d_{11}	0	1	0	0	B
d_{12}	0	1	0	0	B
d_{13}	0	1	0	1	B

O passo final é gerar o novo conjunto de treinamento. O vetor FS é percorrido e um mapeamento é feito utilizando $d'_h = d_{FS_h}$ (linha 5 do Algoritmo 2). Então, $d'_1 = d_{FS_1} \therefore d'_1 = d_2$, no qual d' é o documento do novo conjunto de treinamento e d é o documento original. Segue-se o mesmo processo com: $d'_2 = d_4$, $d'_3 = d_8$ e $d'_4 = d_6$.

Na Tabela 4.2 é exibido o resultado do método utilizando a base apresentada na Tabela 4.1. É notável a presença de características apenas em classes específicas: w_1 e w_3 apenas na classe A, e w_4 apenas na classe B. Sendo, portanto, uma seleção adequada que facilita a classificação.

4.2 ALfFT (*At Least f Features*)

No método ALOFT, apenas uma característica é selecionada por documento. No (ALfFT), f características são selecionadas por documento. O processo do método é descrito no Algoritmo 4.

Este algoritmo possui quatro modificações com relação ao seu antecessor:

i) **Input**. Neste método o usuário precisa fornecer um inteiro positivo que irá representar quantas características serão selecionadas em cada documento;

ii) Linhas 8 e 21: representam o início e o fim do novo laço inserido para lidar com a seleção das f características por documento.

Essas duas são as modificações mais importantes. Entretanto, a título de implementação, alguns trechos foram inseridos para que o algoritmo funcione como esperado:

iii) Linha 19: o valor que representa a importância do termo selecionado se torna negativo no vetor FS. Com o valor negativo, a característica não será escolhida novamente;

iv) Linhas 22-24: os valores do vetor S são restaurados ao seu valor original (positivo).

Essas linhas adicionadas no algoritmo incrementam um pouco mais o tempo de processamento do método. Uma otimização seria restaurar (linhas 22-24) os valores apenas dos índices presentes no vetor FS, pois apenas eles teriam se tornados negativos na operação efetuada na linha 19. No caso de $f = 1$ o resultado obtido será o mesmo do algoritmo ALOFT, sendo o ALfFT uma generalização do método base.

Algoritmo 4 ALfFT

Input: $f \geq 1$

Load training set $\mathscr{D}_{tr}$
for $h = 1$ to V **do** // Calcula FEF para cada termo
  $S_h = \text{FEF}(w_h)$
end for
$m = 0$
FS an empty vector
for all $d_i \in \mathscr{D}_{tr}$ **do** // Seleciona, de cada documento, as f características valoradas com as maiores FEFs.
  for $n = 1$ to f **do**
    $bestscore = 0.0$
    for $h = 1$ to V **do**
      if $w_{h,i} > 0$ and $S_h > bestscore$ **then**
        $bestscore = S_h$
        $bestfeature = h$
      end if
    end for
    if $bestfeature \notin FS$ **then**
      $m = m + 1$
      $FS_m = bestfeature$
      $S_{bestfeature} = -1 \times S_{bestfeature}$
    end if
  end for
  for $h = 1$ to V **do**
    $S_h = |S_h|$
  end for
end for
execute Novo_Conjunto($\mathscr{D}_{tr}$,FS,m)

4.2.1 Exemplo de Execução

Considere a Tabela 4.1 (a mesma utilizada no exemplo do ALOF) para a execução do algoritmo ALfFT com $f = 2$. O primeiro documento a ser verificado é o d_1, cujas características valoradas são w_2 e w_7, como $f = 2$ ambas serão selecionadas e o vetor FS recebe sua primeira atualização: $FS = \{2,7\}$. No documento d_2, são características valoradas w_4, w_5 e w_8, a primeira característica a ser selecionada é o w_4, pois $S_4 > S_5 > S_8$. Como $f = 2$, a seleção continua no documento d_2 em busca da segunda melhor característica valorada: w_5, pois $S_5 > S_8$, então, $FS = \{2,7,4,5\}$. No documento d_3, a característica valorada de maior valor em S é o w_4, mas como $4 \in FS$, o algoritmo segue para verificar a segunda melhor característica w_9, pois $S_9 > S_1 > S_8$. Deste modo, ao final da varredura do terceiro documento $FS = \{2,7,4,5,9\}$.

Note que apesar de $f = 2$, não implica que todo documento deve selecionar duas novas características, portanto no caso do d_3 apenas uma característica foi novidade para o vetor FS, pois uma delas já existia previamente no vetor. No documento d_4, só existe uma característica valorada, portanto apenas ela será verificada e como já é uma característica existente em FS, nenhuma atribuição é realizada. Ao final da varredura em todos os documentos, o vetor final será $FS = \{2,7,4,5,9,8,6,1\}$.

Algoritmo 5 ALfFT-L

Input: $f \geq 1$

Load training set $\mathscr{D}_{tr}$
for $h = 1$ to V **do** // Calcula FEF para cada termo em cada classe
  for $j = 1$ to N **do**
    $S_{h,j} = \text{FEF}(w_h, c_j)$
  end for
end for
$m = 0$
FS an empty vector
for all $d_i \in \mathscr{D}_{tr}$ **do** // Seleciona, em cada documento, as f características valoradas com as maiores FEFs.
  for $n = 1$ to f **do**
    $bestscore = 0.0$
    for $h = 1$ to V **do**
      if $w_{h,i} > 0$ and $S_{h,\text{class}(d_i)} > bestscore$ **then**
        $bestscore = S_{h,\text{class}(d_i)}$
        $bestfeature = h$
      end if
    end for
    if $bestfeature \notin FS$ **then**
      $m = m + 1$
      $FS_m = bestfeature$
      $S_{bestfeature,\text{class}(d_i)} = -1 \times S_{bestfeature,\text{class}(d_i)}$
    end if
  end for
  for $h = 1$ to V **do**
    $S_{h,\text{class}(d_i)} = |S_{h,\text{class}(d_i)}|$
  end for
end for
execute Novo_Conjunto($\mathscr{D}_{tr}$,FS,m)

4.3 ALfFT-L (*At Least f Features - Local*)

Outra alternativa proposta é a abordagem local. O objetivo desta abordagem é selecionar características que melhor discriminam cada classe. O processo desta versão é descrito no Algoritmo 5.

Neste algoritmo alguns novos elementos surgem: N é o número total de classes e class(d_i) é uma função que retorna a classe de um documento d_i. O antigo vetor S agora é uma matriz, na qual $S_{h,j}$ representa a importância de um termo h com relação a uma classe j.

Diferentemente do seu antecessor (ALfFT), o ALfFT-L selecionará os termos de cada documento com base na sua classe. Na linha 13 do Algoritmo 5, é selecionada a característica que possuir o maior valor de FEF dentre, apenas, os valores da classe do documento que está sendo analisado no momento. Portanto, em cada documento serão selecionadas características especializadas da classe em questão.

4.3.1 FEFs Locais

Com a modificação do cálculo da FEF para gerar uma matriz S (linhas 2-6), também foram modificadas as FEFs para suprir essa necessidade do método. São, portanto, as FEF locais:

$$\mathrm{BNS_L}(w,c_j) = \left|F^{-1}(P(w|c_j)) - F^{-1}(P(w|\bar{c_j}))\right|, \tag{4.1}$$

$$\mathrm{CHI_L}(w,c_j) = \frac{\left[P(w|c_j)P(\bar{w}|\bar{c_j}) - P(w|\bar{c_j})P(\bar{w}|c_j)\right]^2}{P(w)P(\bar{w})P(c_j)P(\bar{c_j})}, \tag{4.2}$$

$$\mathrm{CDM_L}(w,c_j) = \left|\log\frac{P(w|c_j)}{P(w|\bar{c_j})}\right|, \tag{4.3}$$

$$\mathrm{IG_L}(w,c_j) = P(w|c_j)\log\frac{P(w|c_j)}{P(c_j)} + \sum_{j=1}^{C} P(\bar{w}|c_j)\log\frac{P(\bar{w}|c_j)}{P(c_j)}, \tag{4.4}$$

$$\mathrm{MOR_L}(w,c_j) = \left|\log\frac{P(w|c_j)(1-P(w|\bar{c_j}))}{P(w|\bar{c_j})(1-P(w|c_j))}\right|. \tag{4.5}$$

A diferença entre versões locais da FEF para versão global (Seção 3.2) reside na remoção do somatório. Deste modo, o resultado da FEF é distribuído entre cada classe, sendo atribuídas parcelas maiores nas classes onde o termo possui mais relevância e parcelas menores onde o termo possui menos relevância. Portanto, as versões locais das FEFs não modificam a essência

das FEFs, apenas dividem seu resultado entre as classes existentes para realizar uma seleção mais especifica.

4.3.2 Conversão para *multi-label*

A necessidade da versão local de lidar com as classes de cada documento traz uma dependencia de como a etiquetagem é realizada. O algoritmo apresentado na Seção 4.3 deve ser utilizado apenas em problemas *single-label*. Quando o problema é *multi-label*, isto é, cada documento pode pertencer à várias classes, o algoritmo do ALfFT-L deve ser modificado.

As modificações necessárias são nas Linhas 13 e 14 do Algoritmo 5 substituindo o termo $S_{h,\text{class}(d_i)}$ por $\text{mlabel}(S,h,\text{class}(d_i))$. Adicionalmente, é necessário que a função class retorne um vetor com as diversas classes do documento d_i. A função mlabel é descrita no Algoritmo 6, cujo objetivo é basicamente retornar a média dos valores de FEF das classes que o documento faz parte.

Algoritmo 6 Função mlabel()

Input: Vector S, Integer $h > 0$ and Vector $\text{class}(d_i)$

1: $sum = 0$
2: $total = 0$
3: **for all** $c \in \text{class}(d_i)$ **do** // Para cada classe que o documento fizer parte
4: $\quad sum = sum + S_{h,c}$
5: $\quad total = total + 1$
6: **end for**
7: **return** $sum/total$

4.4 ALfFT-R (*At Least f Features - Restricted*)

A proposta básica dos métodos introduzidos neste capítulo é seguir a idéia inicial de cada documento possuir pelo menos uma característica valorada representando-o. Entretanto, na abordagem restrita, alguns documentos serão ignorados no processo de seleção das características. Para determinar quais documentos serão ignorados durante a seleção, os documentos possuirão uma métrica para avaliar sua relevância, definida por:

$$\text{DR}(d_i) = \sum_{h=1}^{V} \big(\text{FEF}(w_h) \times \text{bin}(w_{h,i})\big), \tag{4.6}$$

no qual a função FEF (descritas na Seção 3.2) representa o valor de importância da caracterís-

tica w_h e a função bin retorna um valor binário:

$$\text{bin}(w_{h,i}) = \begin{cases} 1 & w_{h,i} \neq 0 \\ 0 & w_{h,i} = 0 \end{cases} \tag{4.7}$$

Deste modo, o valor de DR será o somatório dos valores das FEFs das características valoradas. Sendo assim, documentos que forem representados por características importantes também serão considerados importantes.

O método é descrito em detalhes no Algoritmo 7.

Três modificações foram feitas neste método com relação ao seu antecessor (ALfFT):

i) Linhas 2-4. representam o início e o fim do novo laço inserido para calcular os valores do novo vetor *DR*. Este vetor possui um valor que representa a importância de cada documento d_i;

ii) Linha 5: calcula a média do vetor *DR*. Este valor é o limiar que indica quando um documento deve ou não ser ignorado;

iii) Linha 12: condição usada para determinar se o documento será ignorado ou não.

Com essas três modificações o algoritmo irá agora selecionar uma quantidade igual ou inferior à encontrada pelo ALfFT. O objetivo da versão restrita é selecionar menos características, ignorando os documentos que tiverem menor relevância.

Apesar da necessidade de calcular o vetor *DR*, a quantidade de documentos ignorados no processo de seleção faz com que o tempo de processamento seja equivalente ao ALfFT.

4.5 ALfFT-LR (*At Least f Features - Local Restricted*)

Esse método é simplesmente a união da versão local (Seção 4.3) com a versão restrita (Seção 4.4). O método é descrito pelo Algoritmo 8. A única diferença com relação aos métodos já apresentados é o cálculo da função DR, que será definido por:

$$\text{DR}_\text{L}(d_i) = \sum_{h=1}^{V} \big(\text{FEF}(w_h, \text{class}(d_i)) \times \text{bin}(w_{h,i})\big). \tag{4.8}$$

Essa modificação é realizada, pois é necessário especificar uma classe para as funções de FEF, vide Equações (4.1), (4.2), (4.3), (4.4) e (4.5).

Para problemas *multi-label* cada documento terá diversos valores de FEF para cada classe que pertence. Portanto, para calcular o valor de DR a FEF a ser considerada é uma média entre todos os valores.

4.6 Considerações Finais

Neste capítulo foi apresentada a idéia base do método ALOFT e os métodos propostos derivados desta idéia: ALfFT, ALfFT-L, ALfFT-R e ALfFT-LR. No próximo capítulo serão apresentados os experimentos com as comparações do desempenho dos métodos propostos (apresentados neste capítulo) com relação ao método clássico (apresentado no Capítulo 3). Assim como, uma análise do tempo computacional, um estudo sobre o comportamento das FEFs.

Algoritmo 7 ALfFT-R

Input: $f \geq 1$

Load training set $\mathscr{D}_{tr}$
for all $d_i \in \mathscr{D}_{tr}$ **do** // Calcula DR para cada documento.
  $DR_i = \mathrm{DR}(d_i)$
end for
$mDR = \mathrm{mean}(DR)$;
for $h = 1$ to V **do** // Calcula FEF para cada termo
  $S_h = \mathrm{FEF}(w_h)$
end for
$m = 0$
FS an empty vector
for all $d_i \in \mathscr{D}_{tr}$ **do** // Seleciona, em cada documento, as f características valoradas com as maiores FEFs.
  if $DR_i > mDR$ **then**
    for $n = 1$ to f **do**
      $bestscore = 0.0$
      for $h = 1$ to V **do**
        if $w_{h,i} > 0$ and $S_h > bestscore$ **then**
          $bestscore = S_h$
          $bestfeature = h$
        end if
      end for
      if $bestfeature \notin FS$ **then**
        $m = m + 1$
        $FS_m = bestfeature$
        $S_{bestfeature} = -1 \times S_{bestfeature}$
      end if
    end for
    for $h = 1$ to V **do**
      $S_h = |S_h|$
    end for
  end if
end for
execute Novo_Conjunto($\mathscr{D}_{tr}$,FS,m)

Algoritmo 8 ALfFT-LR

Input: $f \geq 1$

Load training set $\mathscr{D}_{tr}$
for all $d_i \in \mathscr{D}_{tr}$ **do** // Calcula DR para cada documento.
  $DR_i = \mathrm{DR_L}(d_i)$
end for
$mDR = \mathrm{mean}(DR)$;
for $h = 1$ to V **do** // Calcula FEF para cada termo em cada classe
  for $j = 1$ to N **do**
    $S_{h,j} = \mathrm{FEF}(w_h, c_j)$
  end for
end for
$m = 0$
FS an empty vector
for all $d_i \in \mathscr{D}_{tr}$ **do** // Seleciona, em cada documento, as f características valoradas com as maiores FEFs.
  if $DR_i > mDR$ **then**
    for $n = 1$ to f **do**
      $bestscore = 0.0$
      for $h = 1$ to V **do**
        if $w_{h,i} > 0$ and $S_{h,\mathrm{class}(d_i)} > bestscore$ **then**
          $bestscore = S_{h,\mathrm{class}(d_i)}$
          $bestfeature = h$
        end if
      end for
      if $bestfeature \notin FS$ **then**
        $m = m + 1$
        $FS_m = bestfeature$
        $S_{bestfeature,\mathrm{class}(d_i)} = -1 \times S_{bestfeature,\mathrm{class}(d_i)}$
      end if
    end for
    for $h = 1$ to V **do**
      $S_{h,\mathrm{class}(d_i)} = |S_{h,\mathrm{class}(d_i)}|$
    end for
  end if
end for
execute Novo_Conjunto($\mathscr{D}_{tr}$,FS,m)

CAPÍTULO 5

Experimentos

Neste capítulo são descritas as bases de dados na Seção 5.1 e configurações dos experimentos na Seção 5.2. Cada método proposto possui uma seção para relatar, e discutir, os resultados dos experimentos: ALfFT na Seção 5.3, ALfFT-L na Seção 5.4, ALfFT-R na Seção 5.5 e ALfFT-LR na Seção 5.6. Uma análise comparativa dos métodos propostos é feita na Seção 5.7. O comportamento das FEFs é discutido na Seção 5.8. Por fim, uma análise do tempo de execução dos métodos é realizada na Seção 5.9. A Seção 5.10 traz as conclusões deste capítulo.

5.1 Bases de Dados

Para comparar o desempenho dos algoritmos apresentados neste trabalho, foram utilizadas três base de dados *single-label*. Cada base de dados possui uma diferente distribuição dos padrões entre as classes, além de variarem na quantidade de documentos total. Deste modo, as comparações entre os métodos se deu com diversos tipos de problemas dentro do domínio textual.

Seguem as descrições, em detalhes, das três bases de dados usadas:

- ***20 Newsgroup***

  O *20 Newsgroup corpus*[1] (LANG, 1995) possui 19.997 artigos obtidos pela *Usenet newsgroup collections*. Para os experimentos realizados neste trabalho, a base de dados foi usada em sua totalidade, tal qual outros trabalhos (XUE; ZHOU, 2009; BEKKERMAN, 2003; NIGAM et al., 1998). Os documentos dessa base de dados possuem um cabeçalho, mas nenhum tratamento especial foi realizado para obter metadados[2]. O vetor de características, após o pré-processamento, foi composto por 46.834 termos ($V = 46834$).

  Apesar de ser uma base com muitos documentos, sua composição é bastante uniforme. Seus 19.997 documentos são distribuídos em vinte categorias. Dezenove categorias possuem 1.000 documentos e uma possui 997 documentos.

[1] Disponível em http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes/20_newsgroups.tar.gz.
[2] Estruturas para facilitar o entendimento de certo dado. Por exemplo: autor de uma notícia.

Tabela 5.1 Distribuição dos documentos entre as classes da base *Reuters 10*.

Categoria	# Treinamento	# Teste	# Total
acq	1.777	592	2.369
corn	178	59	237
crude	434	144	578
earn	2.973	991	3.964
grain	437	145	582
interest	359	119	478
money-fx	538	179	717
ship	215	71	286
trade	365	121	486
wheat	213	70	283
Total	7.489	2.491	9.980

- **Reuters 10**

  A coleção *The Reuters-21578*[3] contém documentos coletados do *Reuters newswire* de 1987. É o *benchmark* mais utilizado para experimentos de classificação de documentos. Sua composição original é de 135 categorias. Entretanto, existem dois subconjuntos em destaque: *ModeApté* – já segmentado em treino e teste com 90 categorias e 21.578 documentos – e *Top 10* – contendo as dez categorias com maior quantidade de padrões por categoria. – Para este trabalho foi escolhido o segundo subconjunto. Ele possui 9.980 documentos. Esta mesma configuração foi utilizada em outros trabalhos (CHEN et al., 2009; CHANG; CHEN; LIAU, 2008; SHANG et al., 2007). O vetor de características, após o pré-processamento, foi composto por 10.987 termos ($V = 10987$).

  Essa base de dados possui alto desbalanceamento entre as classes, conforme exibido na Tabela 5.1. A categoria com mais padrões representa 39% de todos os documentos, enquanto que a com menos padrões possui apenas 2,3% dos documentos.

- **WebKB**

  O *WebKB corpus*[4] (CRAVEN; DIPASQUO; FREITAG, 1998) é uma coleção com 8.282 páginas da *web* de quatro faculdades americanas. O conjunto original possui sete categorias, mas nos experimentos foram utilizados apenas quatro delas: *course*, *faculty*, *project* e *student*. Este subconjunto contém 4.199 documentos e foi introduzido por Nigam et al. (NIGAM et al., 1998). Outros trabalhos também utilizaram esta mesma configuração (XUE; ZHOU, 2009; BEKKERMAN, 2003). Os documentos dessa base de dados pos-

[3]Disponível em http://dit.unitn.it/~moschitt/corpora.htm.
[4]Disponível em http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/.

Tabela 5.2 Distribuição dos documentos entre as classes da base *WebKB*.

Categoria	# Treinamento	# Teste	# Total
course	698	232	930
faculty	843	281	1.124
project	378	126	504
student	1.231	410	1.641
Total	3.150	1.049	4.199

suem *tags html*[5], mas nenhum tratamento especial foi realizado para removê-las, visando verificar o impacto das *tags html* nos experimentos. O vetor de características, após o pré-processamento, foi composto por 21.324 termos ($V = 21324$).

Na Tabela 5.2 é mostrada a distribuição dos documentos entre as classes. A maior categoria possui 39% de todos os documentos e a menor 12%. Sendo uma diferença de 27 pontos percentuais, uma discrepância menor do que os 36,7 pontos percentuais de diferença da base *Reuters 10*.

5.2 Configurações dos Experimentos

Todos os experimentos, tanto os métodos propostos como o método VR, utilizaram as mesmas configurações:

- Validação cruzada estratificada com *4-folds*. Isto é, quatro conjuntos (*folds*) com documentos divididos de modo proporcional às classes. Estes *folds* são unidos para gerar diferentes conjuntos de treinamento e de teste. O conjunto de treinamento é composto por três *folds* (75%) e o de teste por um *fold* (25%). Portanto cada experimento é realizado quatro vezes utilizando diferentes combinações dos *folds*, conforme mostra a Figura 5.1. Os resultados apresentados nas tabelas são as médias dos quatro resultados seguidas pelo desvio padrão entre parênteses;

- As rotinas de pré-processamento (descritas na Seção 2.3) utilizadas foram: remoção de *stopwords* (lista no Apêndice A), *stemming* com o algoritmo *Iterated Lovins Stemmer* (LOVINS, 1968) e remoção dos termos com $DF < 2$ (descrito na Seção 3.2);

- Vetor de características foi composto seguindo a abordagem BOW e com valores baseados na Frequência do Termo. Ambos apresentados na Seção 2.4. Um exemplo da repre-

[5]Rótulos utilizados para a página ser construída pelo navegador. Por exemplo: <html>, <header>, <title>

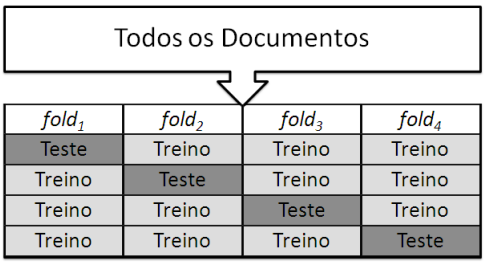

Figura 5.1 Construção do conjunto de treinamento, composto por três *folds*, e do conjunto de teste, composto por um *fold*.

sentação é exibido na Figura 5.2. Apesar do TF-IDF ter mais poder de representação e ser mais utilizado, o modelo do classificador *Naïve Bayes* usado nos experimentos requer que os valores das características sejam apenas a Frequência do Termo;

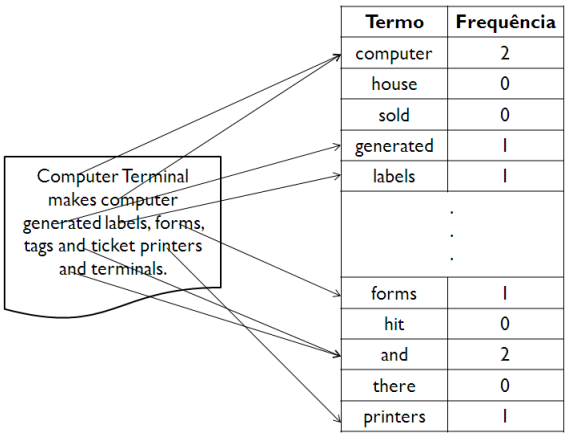

Figura 5.2 Exemplo da representação usada nos experimentos (*bag of words* e Frequência do Termo).

- Nos experimentos foram utilizadas as cinco FEFs: *Bi-Normal Separation* (BNS), *Chi-Squared Statistics* (CHI), *Class Discriminanting Measure* (CDM), *Information Gain* (IG) e *Multi-class Odds Ratio* (MOR) (apresentadas na Seção 3.2). No caso da versão local (ALfFT-L, descrito na Seção 4.3), as funções foram modificadas para a abordagem local, como descrito na Seção 4.3.1;

- Os classificadores são os mesmos apresentados nas Seções 2.5.1 e 2.5.2. No caso do classificador kNN, foi fixado o valor de $k = 5$ em todos os experimentos;

- Para avaliar o desempenho dos algoritmos foram utilizadas as micro-F1 e macro-F1 descritas na Seção 2.6. Entretanto, os valores originais destas medidas estão no intervalo $[0, 1]$, para facilitar a visualização, os valores foram multiplicados por 100;

- O parâmetro f teve cinco valores: 1, 2, 3, 4 e 5. Esse parâmetro é utilizado pelos métodos propostos, descritos no Capítulo 4. A escolha desses valores foi feita com base em experimentos preliminares, pois $f > 5$ exibiu quedas no desempenho do algoritmo;

- O método clássico VR (descrito na Seção 3.1) requer a declaração de quantas características devem ser selecionadas (parâmetro m). Porém, nos métodos propostos (descritos no Capítulo 4), a quantidade de características selecionadas é encontrada durante a execução do método, sem a necessidade de um parâmetro de entrada. Portanto, para uma comparação justa, o valor de m utilizado pelo método VR será o mesmo encontrado pelos métodos propostos.

- Foi utilizada uma máquina dedicada exclusivamente aos experimentos com um processador de 2.5 GigaHertz e uma memória RAM de 2 Gigabytes, com o sistema operacional Windows XP.

5.3 Resultados com o método proposto ALfFT

O algoritmo apresentado na Seção 4.1 (ALOFT) é uma versão simplificada do ALfFT, por isso seus resultados estão reunidos nesta mesma seção. A apresentação dos resultados é dada por diversas tabelas. Cada tabela mostra os resultados com relação a um valor do parâmetro f. Deste modo, os resultados do ALfFT são distribuídos entre cinco tabelas: na Tabela B.1 os resultados com $f = 1$, na Tabela B.2 com $f = 2$, na Tabela B.3 com $f = 3$, na Tabela B.4 com $f = 4$ e, finalmente, na Tabela B.5 com $f = 5$. Em todas as tabelas, foram destacados em negrito as melhores médias dos valores de macro-F1 e micro-F1 de cada linha para facilitar a visualização da comparação entre os métodos.

Todos os resultados das Tabelas B.1 à Tabela B.5 foram resumidos em dois gráficos:

O gráfico da Figura 5.3 contabiliza o percentual de vitórias do método proposto (ALfFT) sobre o método clássico (VR), isto é, quantas vezes o resultado do ALfFT foi marcado em negrito. Deste modo, quando o valor apresentado no gráfico é inferior à 50%, indica que o método propostos perdeu mais vezes do que ganhou.

O gráfico da Figura 5.4 exibe o somatório das diferenças entre os resultados dos métodos comparados. Este calculo é dado por:

$$\text{SUMDIF}(x,y) = \sum_{a \in \mathscr{P}} \left[(\text{macro-F1}_a^x - \text{macro-F1}_a^y) + (\text{micro-F1}_a^x - \text{micro-F1}_a^y) \right], \qquad (5.1)$$

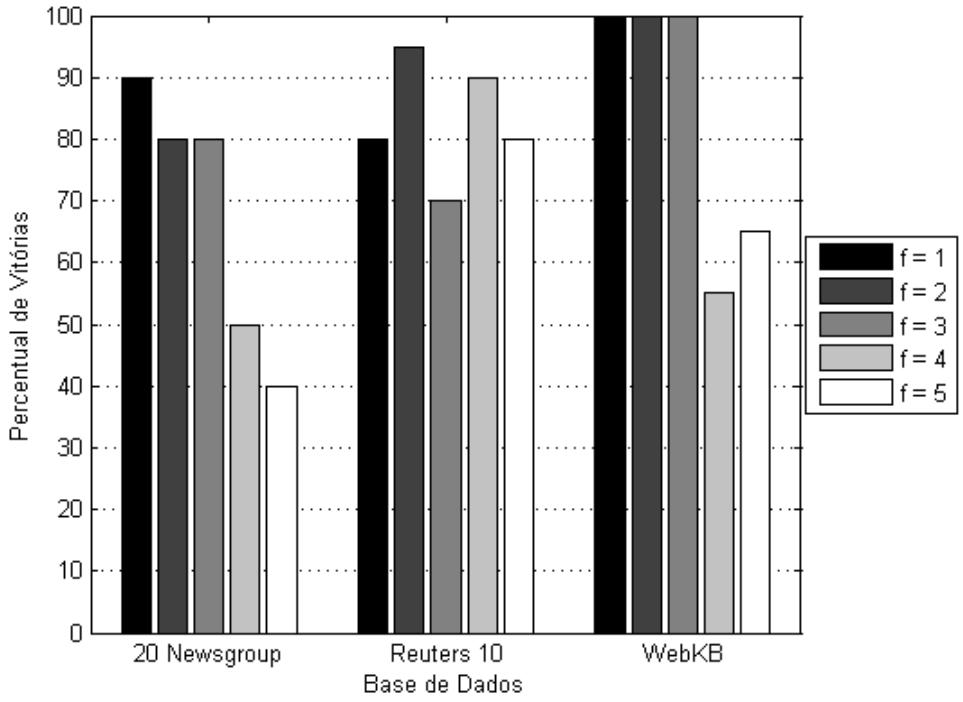

Figura 5.3 Percentual de vitórias dos resultados do ALfFT sobre o VR.

no qual x e y são dois métodos em comparação. $\mathscr{P}$ é o conjunto de todos os experimentos realizados com os dois métodos em comparação. Portanto, macro-F1$_a^x$ é o resultado médio da macro-F1 num experimento a do método x, a mesma idéia se aplica aos demais termos da equação. Então, o resultado de SUMDIF(x,y) será o somatório das diferenças de todas as médias dos experimentos realizados na comparação de dois métodos. No caso de $x =$ALfFT e $y =$VR, a primeira iteração do somatório (primeira linha da Tabela B.1), seria: $(12,49 - 12,60) + (17,60 - 17,13) = 0,47$. Um valor negativo indica que o método x foi inferior ao método y. Nos resultados o método x é sempre o método proposto e o método y é sempre o VR.

Com esses dois gráficos é possível comparar os métodos de modo quantitativo (quantas vezes venceu) e de modo qualitativo (o quão bem venceu). Ambos os gráficos mostram todas as bases de dados e todos os valores de f usados nos experimentos.

O gráfico da Figura 5.3 mostra a superioridade do ALfFT sobre o VR. O VR vence em apenas uma ocasião, com $f = 5$ na *20 Newsgroup*, e empata em outra, com $f = 4$ na mesma base de dados. Portanto, o ALfFT consegue melhor percentual de vitória em 13 das 15 situações.

No gráfico da Figura 5.4 as maiores somas das diferenças são com $f = \{1,2,3\}$. Na *20 Newsgroup* com $f = 5$ o ALfFT chega a obter um valor negativo na soma das diferenças,

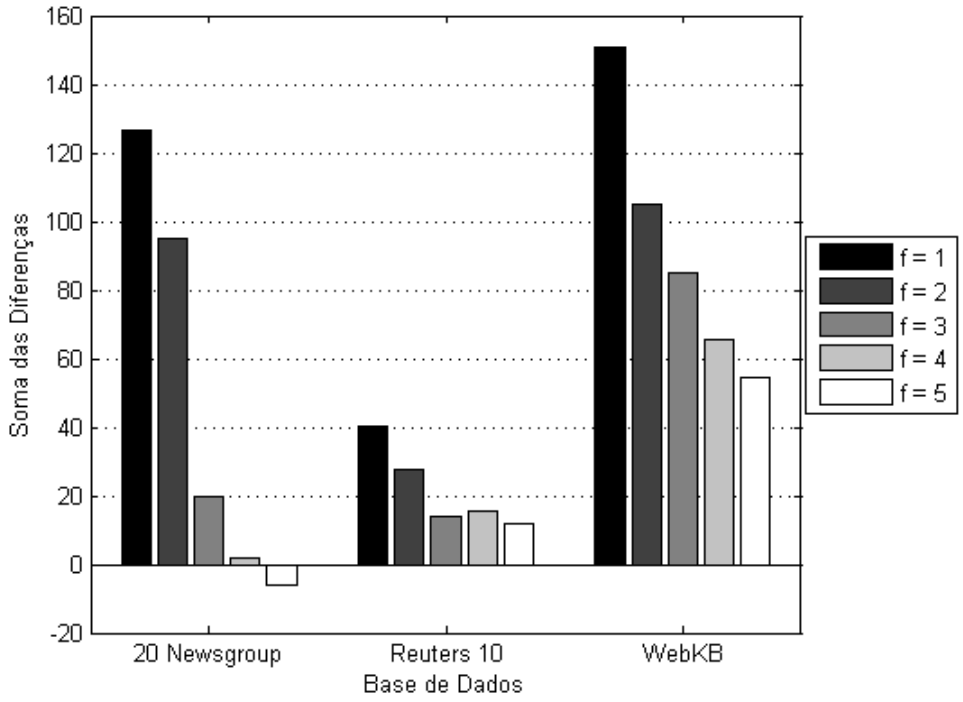

Figura 5.4 Soma das diferenças dos resultados do ALfFT sobre o VR.

indicando que o método VR conseguiu melhores resultados tanto em termos de percentual das vitórias como na soma das diferenças. Essa foi a única vitória dentre os 16 casos apresentados no gráfico, sendo uma soma das diferenças 20 vezes inferior à soma das diferenças do melhor caso do método proposto na mesma base.

Em ambos os gráficos o ALfFT lida melhor com poucas características. Com o aumento do f, mais características são selecionadas. Entretanto, o aumento na quantidade de características é feito selecionando as características valoradas do mesmo documento. Caso um documento possua, por exemplo, duas características relevantes (capaz de contribuir na classificação), uma irrelevante (não causa nenhuma interferencia na classificação) e o restante prejudicial (capaz de dificultar a classificação), com $f = 4$ uma característica prejudicial seria selecionada. Portanto, quanto maior o f, maior é a probabilidade de características irrelevantes ou prejudiciais (baixo valor de FEF) serem selecionadas.

O contrário acontece com o VR, pois na maioria dos casos, seus resultados são melhorados com o incremento da quantidade de características selecionadas. Sua metodologia baseia-se apenas na FEF, portanto ao selecionar x ou $x+1$ características, isso não trará tantos problemas, afinal, a nova característica selecionada será o maior FEF possível dentre as características restantes (diferentemente do ALfFT que ao selecionar x ou $x+1$ características não garante

que essa nova característica encontrada possua um bom valor de FEF).

Para os experimentos do ALfFT, os melhores desempenhos – de acordo com as Figuras 5.3 e 5.4 – são obtidos com $f = 1$. Todavia, uma análise sobre os dados das Tabelas B.1 e B.2 mostram que os resultados de mesma configuração (base de dados, classificador e FEF) do AL2FT são superiores aos do AL1FT em 99% dos casos.

5.4 Resultados com o método proposto ALfFT-L

Os resultados apresentados nesta seção seguem a mesma organização utilizada na Seção 5.3: cinco tabelas variando os valores de f e dois gráficos exibindo um resumo dos resultados presentes nas tabelas.

O objetivo da versão local (ALfFT-L) é selecionar características mais focadas em cada classe. Para cobrir todas as classes, o ALfFT-L, na maioria dos casos, seleciona mais características que seu antecessor (ALfFT). Entretanto, apesar do aumento na quantidade de características selecionadas o ALfFT-L não consegue superar os resultados do ALfFT. Essa afirmação pode ser comprovada ao comparar o gráfico da Figura 5.3 com o gráfico da Figura 5.5 e o gráfico da Figura 5.4 com o gráfico da Figura 5.6.

No gráfico da Figura 5.5 o VR consegue vencer o ALfFT-L em dois casos: *20 Newsgroup* com $f = 2$ e *WebKB* com $f = 5$.

A dificuldade do método proposto, com relação ao aumento do f, se torna clara no gráfico da Figura 5.4. O ALfFT-L atinge ótimas taxas com $f = 1$ – superando a versão original na *20 Newsgroup* – mas com $f > 1$ elas são reduzidas pela metade a cada incremento do f, atingindo um valor negativo com $f = 5$ na *WebKB*.

Apesar dos piores resultados com relação ao seu antecessor, o ALfFT-L possui um comportamento semelhante: resultados iniciam com elevada soma das diferenças e decaem à medida que o f aumenta.

A quantidade de características da Tabela B.1 à Tabela B.5 (experimentos do ALfFT) aumentaram em média 4,1 vezes com relação aos presentes da Tabela C.1 à Tabela C.5 (experimentos do ALfFT-L). Com este aumento, foi observado nos experimentos deste trabalho que, o VR deixou de ser beneficiado pelo aumento das características quando $m > V/10$. A partir deste ponto, características prejudiciais são selecionadas com maior frequência e dificultam o trabalho do classificador.

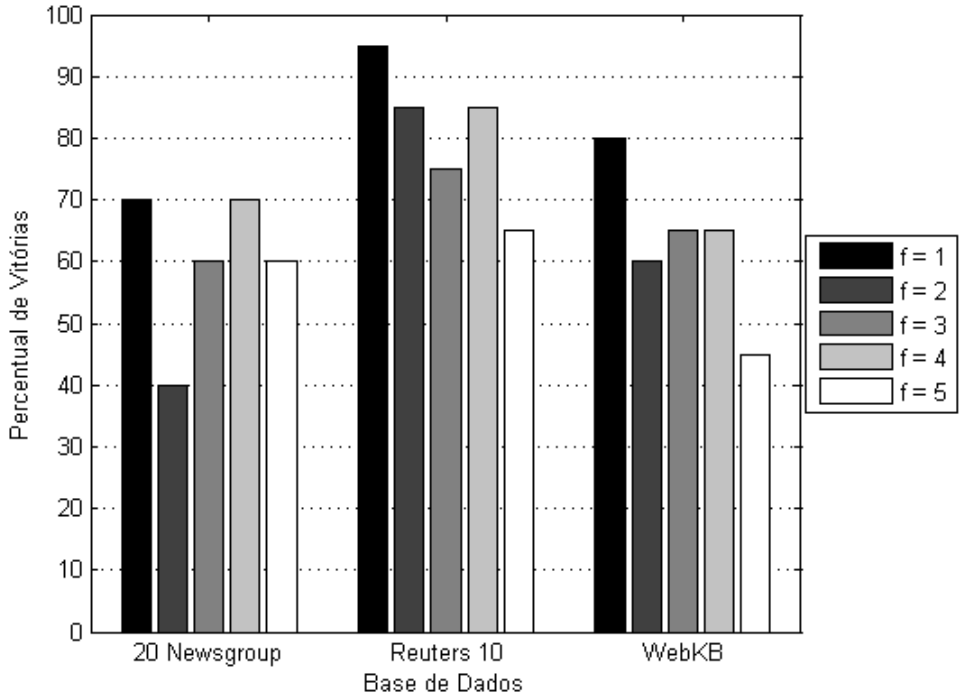

Figura 5.5 Percentual de vitórias dos resultados do ALfFT-L sobre o VR.

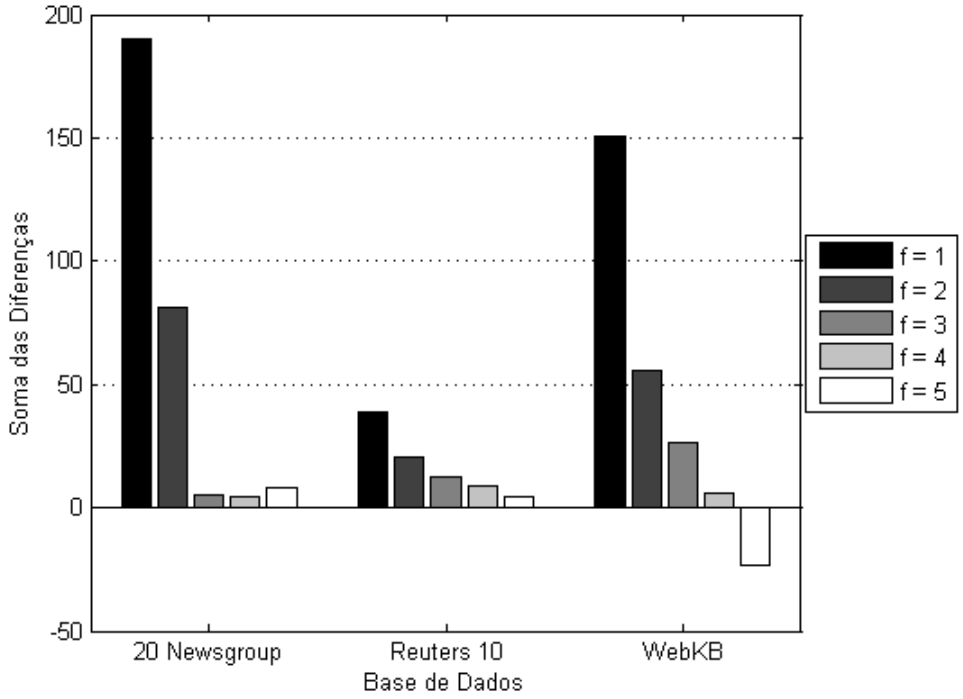

Figura 5.6 Soma das diferenças dos resultados do ALfFT-L sobre o VR.

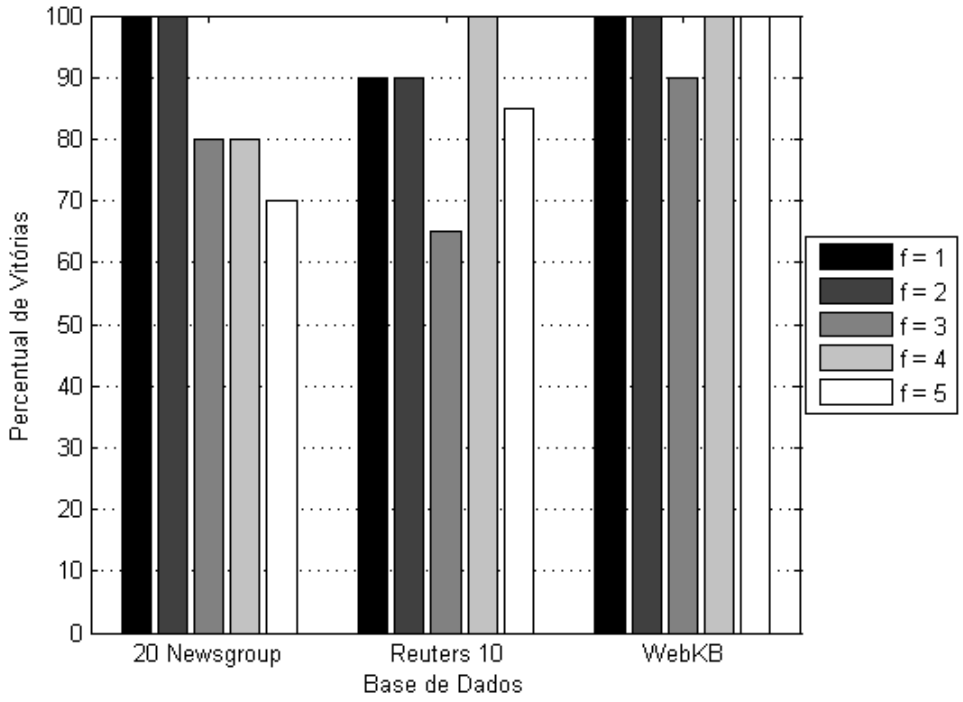

Figura 5.7 Percentual de vitórias dos resultados do ALfFT-R sobre o VR.

5.5 Resultados com o método proposto ALfFT-R

Os resultados apresentados nesta seção seguem a mesma organização utilizada na Seção 5.3.

O objetivo da versão restrita (ALfFT-R) é selecionar menos características que seu método base (ALfFT). O limiar usado para distinguir quais características selecionar é uma média, portanto, a quantidade de características selecionadas será reduzida em torno de 50%.

O método ALfFT-R é superior ao VR. No gráfico da Figura 5.7, o ALfFT-R consegue 100% de vitórias em 7 de 15 das configurações. Mesmo com $f > 3$ – situação que provoca queda de desempenho nos outros métodos propostos – a versão restrita consegue 100% de vitória em três ocasiões. Essa situação acontece por causa da restrição, imposta pelo método, que reduz a quantidade de características selecionadas. Assim, o VR sofre um retardo na evolução de suas taxas, pois requer uma maior quantidade de características para superar o método proposto.

O gráfico da Figura 5.8 mostra outra exceção: o ALfFT-R é o único método proposto que possui todas as somas das diferenças com valor valores positivos.

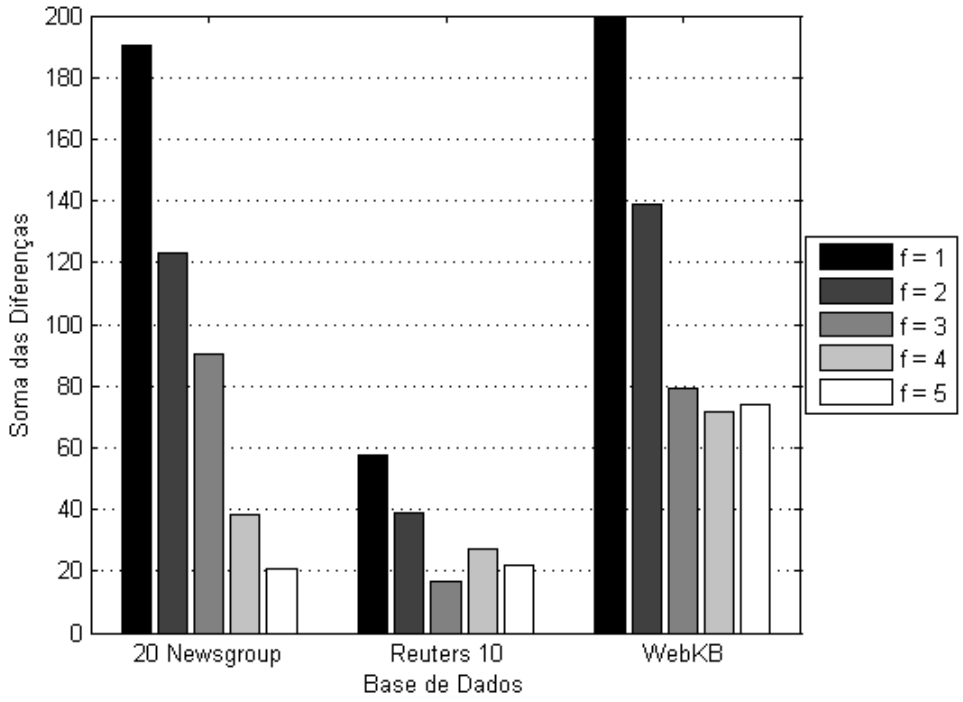

Figura 5.8 Soma das diferenças dos resultados do ALfFT-R sobre o VR.

5.6 Resultados com o método proposto ALfFT-LR

Os resultados apresentados nesta seção seguem a mesma organização utilizada na Seção 5.3.

O objetivo da versão local restrita (ALfFT-LR) é selecionar menos características que seu método base (ALfFT-L). O limiar usado para distinguir quais características selecionar é uma média, portanto, a quantidade de características selecionadas será reduzida em torno de 50%.

O método ALfFT-LR é inferior ao VR na base de dados *Reuters 10*. No gráfico da Figura 5.9, o ALfFT-LR consegue duas vitórias sobre o VR na *Reuters 10* com $f = \{1,2\}$. Enquanto que na *20 Newsgroup* consegue três vitórias com $f = \{1,4,5\}$. Na *WebKB*, a versão local restrita consegue uma uniformidade de 80% no percentual de vitórias para todos os valores de f.

O gráfico da Figura 5.10 confirma que o ALfFT-LR é inferior ao VR na base de dados *Reuters 10*, com três dos somatórios em negativo. Na *20 Newsgroup* – apesar de dois dos somatórios em negativo – com $f = \{1,2\}$, o somatório atinge valores mais de dez vezes maiores em comparação com os negativos. A *WebKB* consegue somatório positivo para todos os valores de f, atingindo a marca mínima de 60 com $f = 5$, somatório igual ao segundo melhor de seu

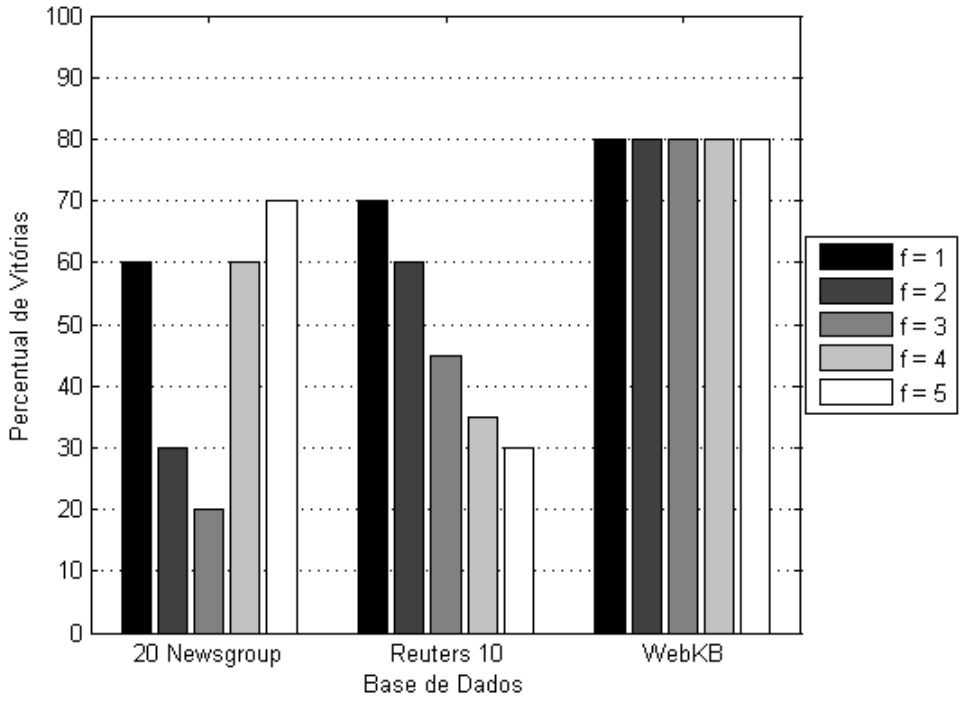

Figura 5.9 Percentual de vitórias dos resultados do ALfFT-LR sobre o VR.

antecessor (ALfFT-L) na mesma base de dados.

O antecessor da versão local restrita: ALfFT-L, seleciona uma quantidade de características superior aos demais métodos. A aplicação de uma restrição nesta seleção, visando frear a quantidade de características parece uma boa alternativa. Entretanto, os resultados mostraram que por ser uma versão local, ao ignorar certos documentos muita informação relevante da classe daquele documento é perdida. Então, com exceção da *WebKB*, sua versão restrita removeu informação preciosa da seleção, prejudicando alguns de seus resultados.

5.7 Comparação dos Métodos Propostos

Na Tabela 5.3 são apresentados as melhores médias dos resultados dos métodos propostos. Os resultados são bastante próximos uns dos outros, com uma diferença máxima de 2,82. Portanto, dependendo das configurações escolhidas (classificador, FEF e f), todos os métodos podem atingir boas taxas.

Para uma maior percepção do desempenho dos métodos, serão utilizados os mesmos gráficos dos experimentos que comparam os métodos propostos contra o método clássico. No

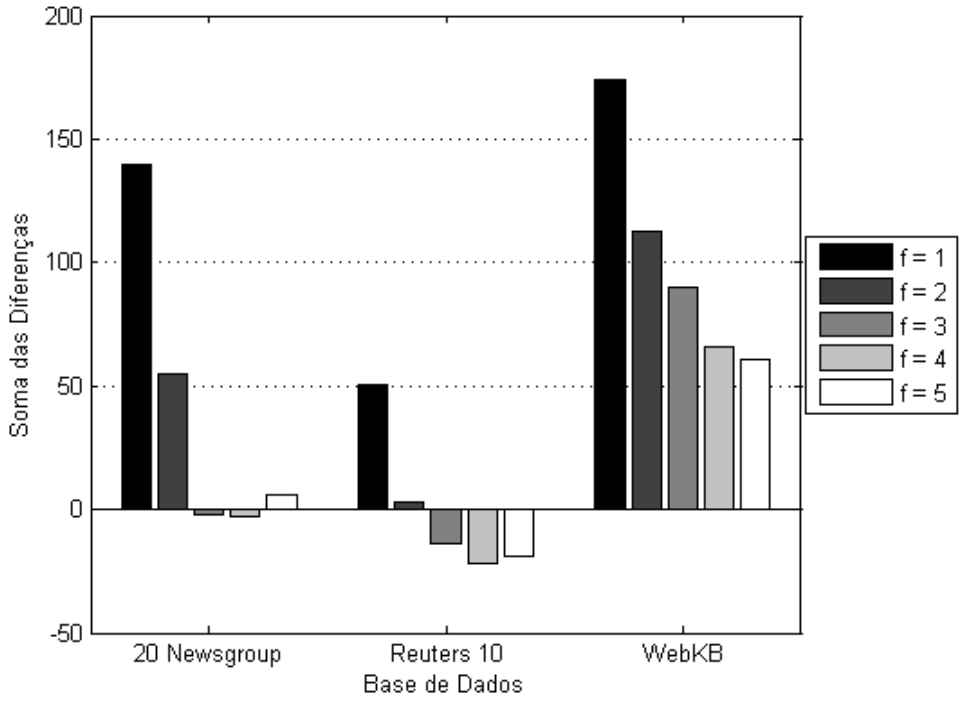

Figura 5.10 Soma das diferenças dos resultados do ALfFT-LR sobre o VR.

Tabela 5.3 Tabela com os melhores resultados de macro-F1 e micro-F1 dos métodos propostos.

Base de Dados	ALfFT		ALfFT-L		ALfFT-R		ALfFT-LR	
	macro-F1	micro-F1	macro-F1	micro-F1	macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	94,63	94,61	93,67	93,67	94,41	94,40	94,45	94,48
Reuters 10	64,12	82,43	64,77	82,15	61,95	80,63	64,12	82,40
WebKB	86,39	87,31	86,79	87,38	83,23	85,12	85,78	86,48

gráficos exibidos pela Figura 5.11 e Figura 5.12 é visto claramente o melhor desempenho do ALfFT-R, seguido pelo ALfFT como o segundo melhor. Com os outros dois métodos as posições misturam-se entre as bases de dados: o ALfFT-L é superior ao ALfFT-LR nas bases de dados *20 Newsgroup* e *Reuters 10* tanto com relação ao percentual de vitórias quanto a soma das diferenças. Portanto, pode ser concluído que o ALfFT-L segue na terceira posição por vencer do ALfFT-LR em duas das três bases.

Uma ordenação dos métodos, de acordo com seus respectivos desempenhos, seria: ALfFT-R $>$ ALfFT $>$ ALfFT-L $>$ ALfFT-LR $>$ VR. No caso da versão local (ALfFT-LR), apesar de ter perdido em algumas ocasiões contra o VR (duas bases de dados na Figura 5.11 e uma base de dados na Figura 5.12), essas derrotas foram pequenas (5% na Figura 5.11) e menos de 10 no somatório da Figura 5.12) enquanto suas vitórias sobre o VR foram grandes (30% na

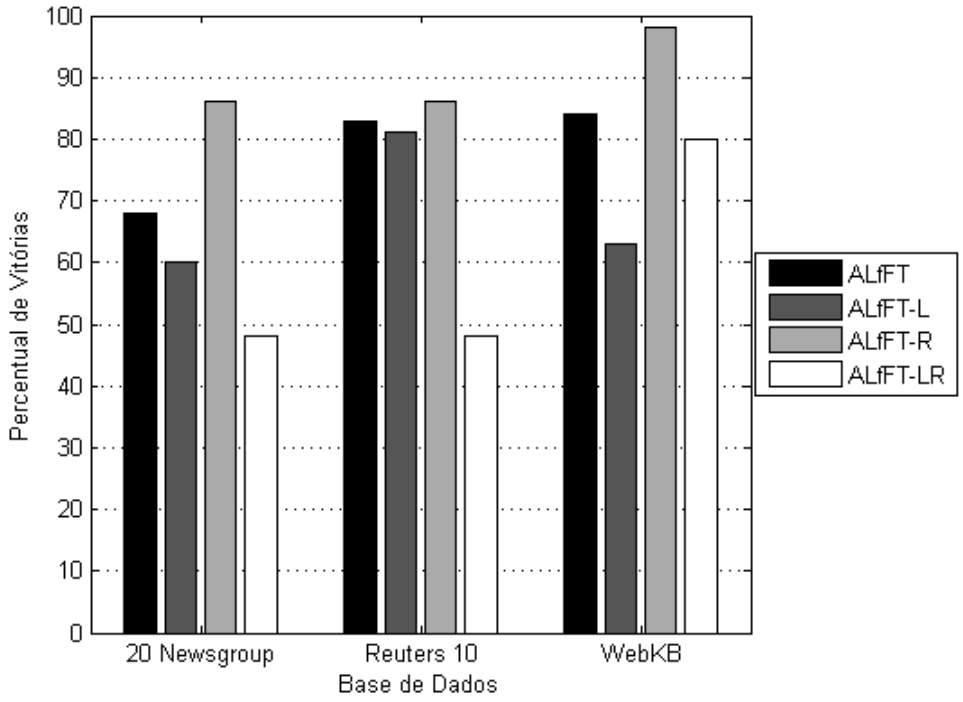

Figura 5.11 Percentual de vitórias dos resultados dos métodos propostos sobre o VR.

Figura 5.11 e média de 340 no somatório da Figura 5.12).

5.8 Comportamento das FEFs

Nesta seção serão discutidos os comportamentos das FEFs e para tal serão utilizados os resultados do método proposto ALfFT presentes nas Tabelas B.1, B.2, B.3, B.4 e B.5. Os resultados foram divididos entre bases de dados e inseridos nos gráficos da Figura 5.13.

Oito resultados foram removidos dos gráficos das Figuras 5.13(a) e 5.13(b): quatro da BNS ($m = 18$) e quatro da IG ($m = 10$). Com esses resultados no gráfico, todos os outros resultados ficavam dentro de um pequeno intervalo, impossibilitando a visualização nos gráficos da *20 Newsgroup*. Com os resultados que permaneceram, são percebidos alguns comportamentos: a BNS consegue o melhor resultado em cinco dos gráficos com uma quantidade intermediária de características em relação às demais FEFs (mais que a CHI e IG, menos que a CDM); a CHI é a FEF que atinge melhores resultados com poucas características; CDM consegue boas taxas, mas sempre seleciona mais características que as outras FEFs; IG possui um comportamento semelhante à CHI, mas seus resultados são, na maioria dos casos, inferiores; MOR seleciona

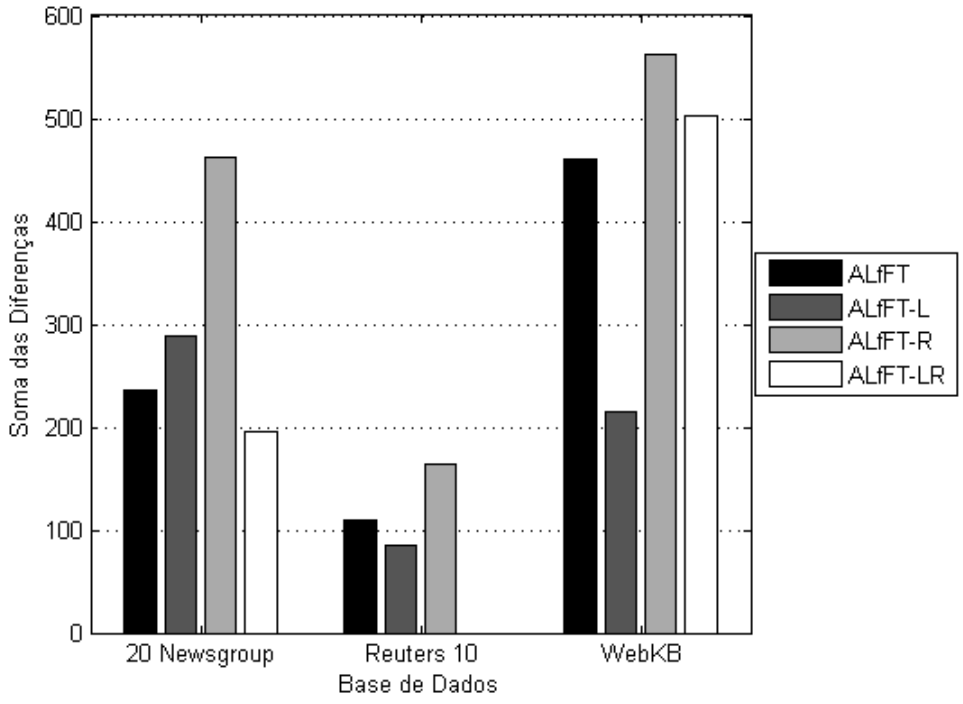

Figura 5.12 Soma das diferenças dos resultados dos métodos propostos sobre o VR.

quantidades semelhantes à BNS, mas seus resultados são inferiores.

Uma ordenação de desempenho das FEFs não é uma tarefa trivial, pois a escolha da FEF dependende do objetivo desejado na seleção. Se a prioridade for selecionar menos características, as FEFs CHI e IG são mais recomendadas. Se a prioridade for obter melhores taxas de acerto, a BNS é mais recomendada. Se o objetivo for para ter a garantia de uma boa seleção, independente da quantidade de características selecionadas talvez a CDM fosse a melhor opção, mas se a quantidade de características importar, seria melhor a MOR.

5.9 Análise do Tempo de Execução

Os métodos propostos são, aparentemente, mais custosos que o método VR. Afinal, o método VR realiza apenas uma varredura sobre o vetor S (vetor com os valores da FEF), enquanto os métodos propostos percorrem todos os documentos e verifica dentre as características valoradas qual delas possui o maior valor na FEF. Entretanto, o método VR necessita da declaração de quantas características deve selecionar (parâmetro m). Deste modo, seu tempo total de execução é o tempo de execução do método VR (tempo gasto para selecionar m ca-

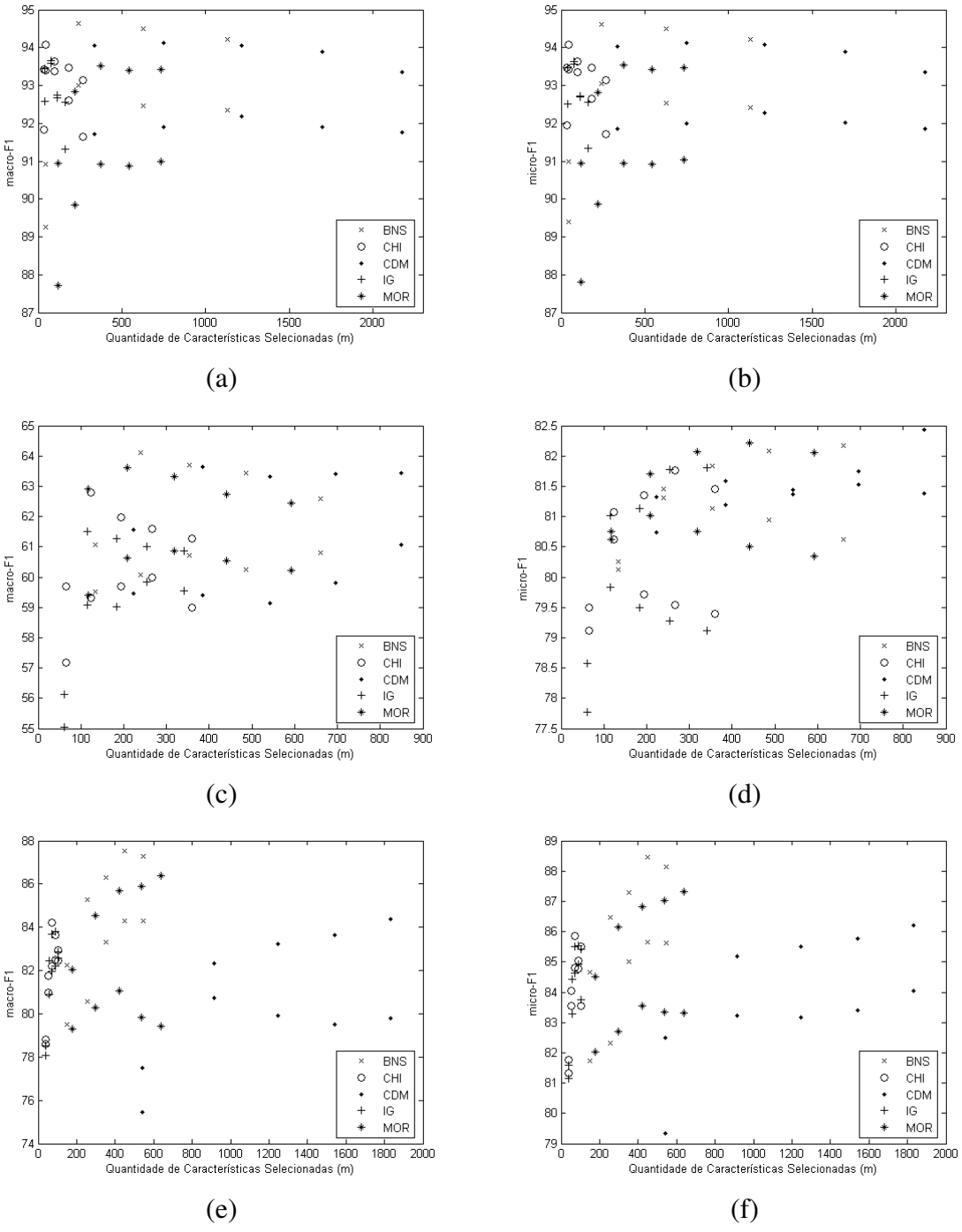

Figura 5.13 Resultados do AL*f*FT: (a) macro-F1 na *20 Newsgroup*; (b) micro-F1 na *20 Newsgroup*; (c) macro-F1 na *Reuters 10*; (d) macro-F1 na *Reuters 10*; (e) macro-F1 na *WebKB*; (f) micro-F1 na *WebKB*.

racterísticas) somado com o tempo de execução do classificador (tempo gasto para verificar a qualidade da seleção). O processo pode ser repetido diversas vezes, mudando o valor de m, caso um bom subconjunto não tiver sido encontrado.

Tabela 5.4 Tabela de tempo de execução dos métodos na *WebKB*. A primeira coluna indica a quantidade de características (m), as demais são os tempos em milissegundos de cada programa quando executado com m características.

m	Tempo com ALfFT	Tempo com VR	Tempo no NB
162	15	4	62
193	16	12	78
674	32	47	281
1563	47	110	656
1852	62	125	766

Na Tabela 5.4 são apresentados os custos computacionais de algumas execuções na base de dados *WebKB*. Com poucas características, o método ALfFT é mais custoso que o método VR, mas a medida que a quantidade de características aumenta o método VR demanda mais tempo de execução. Essa mudança ocorre devido à estrutura de dados otimizada implementada no ALfFT, que não aloca os documentos no formato de matriz, mas sim numa lista encadeada que armazena apenas as características valoradas. Como o vetor de características original (gerado pela abordagem BOW) normalmente supera a quantidade de documentos (no caso da *WebKB* a razão é de 5 documentos para cada característica) uma varredura no vetor S e a quantidade de características valoradas é pequena em cada documento, a varredura se torna rápida. Na Tabela 5.5 estão os tempos de algumas execuções na base de dados *20 Newsgroup* e o método VR consegue ser mais rápido que o ALfFT, pois no caso dessa base a razão é de apenas 2,3 documentos para cada característica.

Tabela 5.5 Tabela de tempo de execução dos métodos na *20 Newsgroup*. A primeira coluna indica a quantidade de características (m), as demais são os tempos em milissegundos de cada programa quando executado com m características.

m	Tempo com ALfFT	Tempo com VR	Tempo no NB
112	31	15	1125
219	47	31	2187
365	78	42	3625
555	109	75	5516
745	156	109	7406

É importante salientar que o algoritmo de ordenação utilizado é o *Insertion sort*, provavelmente, utilizando um algoritmo de ordenação mais veloz, o método VR executaria mais rápido.

Entretanto, não é relevante o quão rápido o VR é, pois esse método necessita de um classificador para avaliar o subconjunto selecionado e, como evidenciado nas Tabelas 5.4 e 5.5, o tempo gasto pelo classificador é de, no mínimo, 30 vezes superior ao utilizado pelo método ALfFT. Sendo assim, o fato de verificar somente um subconjunto – fato irreal, pois são normalmente vistos vários subconjuntos até atingir uma boa seleção – com um classificador, transforma o VR num método mais lento que o ALfFT.

As mesmas conclusões podem ser feitas com relação aos outros métodos propostos, pois a diferença conceitual dos métodos não traz consigo uma diferença significativa no tempo de execução do algoritmo.

5.10 Considerações Finais

Neste capítulo foram apresentados os experimentos. Nesses experimentos os métodos propostos (ALfFT, ALfFT-L, ALfFT-R e ALfFT-LR) foram comparados com o método clássico VR, com o intuito de conhecer o comportamento de cada método com base numa abordagem bastante utilizada na literatura. Adicionalmente, também foram apresentadas análises do tempo computacional e das FEFs. No próximo capítulo será finalizado o trabalho com as conclusões, principais contribuições e trabalhos futuros.

CAPÍTULO 6

Conclusões

Classificação de documentos é um tema bastante diverso e amplo, bem como a necessidade de seleção de características intrínseca deste domínio. Devido a vasta existência de algoritmos complexos e custosos, neste trabalho é proposta uma nova abordagem para selecionar características com um algoritmo simples, intuitivo, de fácil implementação e rápido processamento.

Quatro métodos foram criados a partir da abordagem proposta: ALfFT, a versão básica do algoritmo; ALfFT-L, a versão local; ALfFT-R, a versão restrita; e ALfFT-LR, a versão local restrita. Os experimentos, realizados em três bases de dados, mostram que os métodos propostos são superiores ao método clássico (VR). Essa superação é atingida tanto em termos de macro-F1 e micro-F1 (medidas de avaliação dos SCDs), como pela remoção do parâmetro m. Apesar da existência de um novo parâmetro (f) nos métodos propostos, cuja função é similar ao m (definir quantas características serão selecionadas), o valor de f tem uma escala útil de três valores: $f = \{1,2,3\}$, pois $f > 3$ traz uma queda no desempenho; enquanto m pode variar entre centenas ou milhares de valores. Com a diminuição da variedade de valores do parâmetro de entrada, os testes para verificar a qualidade do subconjunto de características selecionadas serão em menor quantidade. Deste modo, os métodos propostos são mais velozes e mais eficientes que o VR.

Neste Capítulo serão descritos as contribuições deste trabalho e os trabalhos futuros.

6.1 Contribuições

Destacam-se como contribuições deste trabalho:

- Estudo, em detalhes, de todas as etapas de um SCD. Possibilitando leitura introdutória do assunto para futuras pesquisas;
- Nova abordagem de seleção de características para problemas de classificação de documentos;
- Quatro métodos propostos baseados na nova abordagem. Os métodos propostos solu-

cionam os problemas presente na abordagem clássica e, por consequência, atingem melhores taxas de acerto;

- Avaliação experimental do desempenho de cinco métodos de seleção de características (quatro métodos propostos e um método clássico) em três bases de dados, utilizando cinco FEFs;
- Análise do comportamento de cinco FEFs, dentre elas três clássicas e duas recentes.

6.2 Trabalhos Futuros

Algumas possibilidades para trabalhos futuros:

- Verificar o desempenho dos métodos em problemas *multi-label*;
- Realizar testes de hipóteses para investigar com certeza o desempenho dos métodos;
- Analisar a complexidade dos algoritmos para verificação real do tempo computacional;
- Buscar a melhor configuração dos métodos propostos testando-o com diferentes classificadores, FEFs e rotinas de pré-processamento;
- Realizar testes em outras bases de dados de documentos;
- Verificar o desempenho dos métodos propostos de seleção em domínios não-textuais;
- Comparar a seleção do método proposto com outros métodos de seleção existentes na literatura (tanto métodos de filtragem como métodos *wrappers*);

APÊNDICE A

Lista de Stopwords

a, about, above, across, after, afterwards, again, against, all, almost, alone, along, already, also, although, always, am, among, amongst, amoungst, amount, an, and, another, any, anyhow, anyone, anything, anyway, anywhere, are, around, as, at, back, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, below, beside, besides, between, beyond, bill, both, bottom, but, by, call, can, cannot, cant, co, computer, con, could, couldnt, cry, de, describe, detail, do, done, down, due, during, each, eg, eight, either, eleven, else, elsewhere, empty, enough, etc, even, ever, every, everyone, everything, everywhere, except, few, fifteen, fify, fill, find, fire, first, five, for, former, formerly, forty, found, four, from, front, full, further, get, give, go, had, has, hasnt, have, he, hence, her, here, hereafter, hereby, herein, hereupon, hers, herse', him, himse', his, how, however, hundred, i, ie, if, in, inc, indeed, interest, into, is, it, its, itse', keep, last, latter, latterly, least, less, ltd, made, many, may, me, meanwhile, might, mill, mine, more, moreover, most, mostly, move, much, must, my, myse', name, namely, neither, never, nevertheless, next, nine, no, nobody, none, noone, nor, not, nothing, now, nowhere, of, off, often, on, once, one, only, onto, or, other, others, otherwise, our, ours, ourselves, out, over, own, part, per, perhaps, please, put, rather, re, same, see, seem, seemed, seeming, seems, serious, several, she, should, show, side, since, sincere, six, sixty, so, some, somehow, someone, something, sometime, sometimes, somewhere, still, such, system, take, ten, than, that, the, their, them, themselves, then, thence, there, thereafter, thereby, therefore, therein, thereupon, these, they, thick, thin, third, this, those, though, three, through, throughout, thru, thus, to, together, too, top, toward, towards, twelve, twenty, two, un, under, until, up, upon, us, very, via, was, we, well, were, what, whatever, when, whence, whenever, where, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, which, while, whither, who, whoever, whole, whom, whose, why, will, with, within, without, would, yet, you, your, yours, yourself, yourselves.

APÊNDICE B

Tabelas de Resultados do método ALfFT

Tabela B.1 Comparação dos métodos ALfFT com $f = 1$ contra VR.

Base de Dados	Classificador	FEF	m	AL1FT (ALOFT)		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	18 (0,82)	**68,57** (3,23)	**70,47** (2,77)	46,05 (1,98)	53,33 (3,33)
		CHI	32 (0,00)	**93,43** (0,26)	**93,47** (0,27)	92,37 (2,21)	92,61 (1,81)
		CDM	338 (7,26)	**94,05** (0,32)	**94,04** (0,31)	86,94 (0,43)	87,35 (0,37)
		IG	10 (1,00)	**48,05** (0,31)	**53,96** (0,36)	47,91 (3,16	53,49 (3,40)
		MOR	121 (21,44)	**90,93** (1,26)	**90,94** (1,22)	87,32 (0,83)	87,47 (0,83)
	NB	BNS	18 (0,82)	**67,32** (3,98)	**69,68** (3,12)	43,06 (0,20)	53,47 (0,14)
		CHI	32 (0,00)	**91,82** (0,33)	**91,94** (0,33)	90,39 (2,96)	90,85 (2,31)
		CDM	338 (7,26)	**91,72** (0,29)	**91,85** (0,28)	83,44 (0,37)	84,37 (0,29)
		IG	10 (1,00)	49,90 (0,57)	54,70 (0,62)	**51,91** (2,36	**55,68** (2,14)
		MOR	121 (21,44)	**87,71** (0,95)	**87,80** (0,88)	83,83 (0,54)	84,08 (0,60)
Reuters 10	*k*-NN	BNS	134 (9,63)	**59,51** (1,61)	**80,13** (0,88)	55,09 (1,44)	76,78 (0,90)
		CHI	65 (3,74)	57,18 (2,03)	**79,49** (0,71)	**57,56** (1,89)	76,83 (1,18)
		CDM	223 (6,98)	**59,46** (1,64)	**80,74** (1,07)	55,65 (2,17)	78,14 (0,85)
		IG	61 (1,41)	55,05 (1,48)	**78,57** (0,40)	**56,76** (1,64	77,71 (0,93)
		MOR	118 (3,87)	**59,40** (1,86)	**80,76** (0,49)	56,49 (1,65)	78,10 (0,67)
	NB	BNS	134 (9,63)	**61,08** (1,47)	**80,26** (0,61)	57,51 (0,50)	77,15 (0,28)
		CHI	65 (3,74)	59,70 (2,41)	**79,12** (0,99)	**60,92** (1,72)	77,21 (0,81)
		CDM	223 (6,98)	**61,56** (1,24)	**81,33** (0,65)	54,58 (0,50)	78,32 (0,06)
		IG	61 (1,41)	56,13 (1,93)	**77,77** (1,11)	**59,77** (1,31)	77,69 (0,91)
		MOR	118 (3,87)	**62,90** (1,14)	**80,62** (0,50)	60,61 (1,06)	77,86 (0,71)
WebKB	*k*-NN	BNS	148 (5,00)	**79,50** (1,36)	**81,73** (0,99)	69,14 (1,42)	72,78 (2,09)
		CHI	38 (1,00)	**78,79** (1,80)	**81,33** (1,17)	75,00 (1,13)	78,61 (0,83)
		CDM	543 (14,36)	**75,45** (2,27)	**79,33** (2,82)	55,13 (1,43)	62,35 (0,89)
		IG	38 (1,00)	**78,52** (1,42)	**81,14** (0,71)	75,13 (0,63	78,83 (0,48)
		MOR	176 (21,43)	**79,32** (0,89)	**82,04** (0,71)	71,71 (1,49)	74,97 (1,77)
	NB	BNS	148 (5,00)	**82,26** (1,15)	**84,66** (1,37)	74,46 (1,84)	80,68 (0,88)
		CHI	38 (1,00)	**78,60** (0,80)	**81,76** (0,83)	73,16 (0,86)	78,42 (1,11)
		CDM	543 (14,36)	**77,50** (2,05)	**82,50** (1,66)	59,33 (1,61)	70,50 (0,82)
		IG	38 (1,00)	**78,07** (1,56)	**81,59** (1,08)	74,09 (0,62)	79,21 (0,75)
		MOR	176 (21,43)	**82,04** (1,09)	**84,52** (1,18)	76,15 (1,58)	80,04 (0,60)

Tabela B.2 Comparação dos métodos ALfFT com $f = 2$ contra VR.

Base de Dados	Classificador	FEF	m	AL2FT		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	44 (3,74)	**90,91** (1,74)	**90,98** (1,62)	69,66 (4,18)	72,05 (4,14)
		CHI	44 (1,15)	**94,08** (0,12)	**94,08** (0,12)	94,03 (0,32)	94,03 (0,32)
		CDM	752 (8,06)	**94,13** (0,10)	**94,13** (0,11)	92,68 (0,44)	92,71 (0,43)
		IG	40 (0,00)	93,44 (0,22)	93,46 (0,19)	**93,98** (0,33	**94,01** (0,31)
		MOR	222 (9,78)	**92,83** (0,31)	**92,82** (0,31)	90,21 (0,60)	90,28 (0,62)
	NB	BNS	44 (3,74)	**89,26** (1,92)	**89,40** (1,85)	68,07 (2,12)	71,92 (1,72)
		CHI	44 (1,15)	**93,40** (0,39)	**93,42** (0,38)	93,10 (0,31)	93,14 (0,31)
		CDM	752 (8,06)	**91,91** (0,39)	**91,99** (0,42)	90,04 (0,24)	90,16 (0,25)
		IG	40 (0,00)	92,59 (0,32)	92,51 (0,33)	**93,24** (0,41	**93,22** (0,42)
		MOR	222 (9,78)	**89,85** (0,45)	**89,87** (0,45)	86,73 (0,66)	86,89 (0,60)
Reuters 10	*k*-NN	BNS	240 (6,45)	**60,07** (2,18)	**81,45** (1,07)	59,06 (0,80)	80,04 (0,53)
		CHI	123 (4,69)	**59,32** (1,66)	**81,08** (0,62)	59,13 (1,62)	80,12 (0,84)
		CDM	384 (10,72)	**59,41** (1,52)	**81,19** (0,97)	57,81 (1,39)	79,62 (0,48)
		IG	115 (3,16)	**59,07** (1,53)	**81,02** (0,56)	58,35 (1,87)	80,39 (1,19)
		MOR	209 (9,35)	**60,64** (2,07)	**81,70** (0,73)	58,57 (1,78)	79,60 (0,77)
	NB	BNS	240 (6,45)	**64,12** (0,42)	**81,31** (0,56)	61,94 (1,31)	79,68 (0,61)
		CHI	123 (4,69)	62,79 (0,63)	**80,62** (0,29)	**63,21** (0,71)	79,36 (0,72)
		CDM	384 (10,72)	**63,64** (1,53)	**81,58** (0,52)	59,14 (1,07)	80,07 (0,53)
		IG	115 (3,16)	**61,50** (0,79)	**79,83** (0,56)	61,47 (0,50)	79,12 (0,77)
		MOR	209 (9,35)	**63,61** (1,52)	**81,02** (0,71)	61,81 (1,32)	78,74 (0,62)
WebKB	*k*-NN	BNS	254 (5,77)	**80,58** (1,43)	**82,33** (0,92)	77,76 (0,93)	80,35 (0,83)
		CHI	55 (1,91)	**80,97** (0,98)	**83,54** (0,94)	78,32 (0,56)	81,09 (0,73)
		CDM	914 (17,58)	**80,74** (1,06)	**83,23** (0,34)	60,67 (1,61)	66,75 (0,63)
		IG	56 (1,41)	**80,88** (0,65)	**83,28** (0,91)	79,60 (1,78	81,81 (1,50)
		MOR	298 (30,79)	**80,30** (0,84)	**82,71** (1,01)	77,45 (1,51)	79,99 (0,91)
	NB	BNS	254 (5,77)	**85,29** (0,37)	**86,47** (0,57)	81,58 (1,35)	84,19 (1,47)
		CHI	55 (1,91)	**81,76** (0,54)	**84,05** (0,50)	79,32 (0,76)	82,16 (1,06)
		CDM	914 (17,58)	**82,33** (1,51)	**85,19** (1,25)	65,55 (1,80)	74,23 (0,76)
		IG	56 (1,41)	**82,46** (0,94)	**84,43** (0,60)	79,92 (1,14	82,47 (1,34)
		MOR	298 (30,79)	**84,56** (1,52)	**86,14** (0,95)	80,27 (0,81)	82,97 (1,18)

Tabela B.3 Comparação dos métodos ALfFT com $f = 3$ contra VR.

Base de Dados	Classificador	FEF	m	AL3FT		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	242 (3,16)	**94,63** (0,29)	**94,61** (0,29)	90,94 (1,84)	90,97 (1,85)
		CHI	96 (1,83)	93,64 (0,45)	93,63 (0,45)	**94,45** (0,35)	**94,43** (0,34)
		CDM	1218 (10,91)	**94,06** (0,18)	**94,07** (0,17)	93,38 (0,33)	93,40 (0,33)
		IG	75 (1,53)	**93,58** (0,29)	**93,56** (0,30)	94,08 (0,18	94,07 (0,18)
		MOR	371 (11,17)	**93,51** (0,40)	**93,53** (0,39)	93,11 (0,10)	93,11 (0,08)
	NB	BNS	242 (3,16)	**93,00** (0,20)	**93,05** (0,19)	89,04 (1,81)	89,18 (1,75)
		CHI	96 (1,83)	93,37 (0,33)	93,36 (0,32)	**93,60** (0,35)	**93,60** (0,34)
		CDM	1218 (10,91)	**92,19** (0,27)	**92,28** (0,28)	90,74 (0,34)	90,84 (0,37)
		IG	75 (1,53)	**93,65** (0,29)	**93,63** (0,28)	93,25 (0,36	93,25 (0,34)
		MOR	371 (11,17)	**90,91** (0,63)	**90,95** (0,64)	89,88 (0,40)	89,92 (0,40)
Reuters 10	*k*-NN	BNS	353 (4,47)	**60,71** (2,15)	**81,84** (1,10)	59,81 (0,74)	80,46 (0,56)
		CHI	194 (6,24)	59,69 (1,24)	81,35 (0,24)	**60,47** (1,24)	**81,53** (0,48)
		CDM	542 (6,68)	59,13 (1,17)	**81,36** (0,68)	**59,17** (1,99)	80,64 (1,08)
		IG	184 (9,64)	59,01 (1,42)	**81,14** (0,42)	**59,26** (1,38)	81,01 (0,62)
		MOR	318 (7,07)	**60,85** (1,50)	**82,07** (0,57)	58,74 (1,58)	80,38 (0,67)
	NB	BNS	353 (4,47)	**63,69** (1,18)	**81,13** (0,78)	62,46 (0,89)	80,02 (0,35)
		CHI	194 (6,24)	61,96 (1,22)	**79,71** (0,73)	**63,06** (0,77)	79,47 (0,82)
		CDM	542 (6,68)	**63,31** (0,66)	**81,44** (0,32)	60,08 (0,94)	80,36 (0,27)
		IG	184 (9,64)	61,27 (1,31)	**79,49** (0,74)	**61,35** (1,18)	79,05 (0,54)
		MOR	318 (7,07)	**63,31** (1,69)	**80,76** (0,79)	62,51 (0,86)	79,37 (0,54)
WebKB	*k*-NN	BNS	352 (10,30)	**83,31** (1,67)	**85,02** (1,01)	76,74 (1,21)	79,07 (1,01)
		CHI	72 (2,71)	**82,20** (1,05)	**84,81** (1,21)	81,10 (0,94)	83,35 (0,83)
		CDM	1247 (16,58)	**79,92** (0,72)	**83,16** (0,63)	65,18 (1,00)	69,73 (0,36)
		IG	71 (2,08)	**81,95** (1,01)	**84,64** (1,17)	81,64 (1,18)	83,85 (0,58)
		MOR	421 (35,38)	**81,07** (0,98)	**83,54** (1,05)	78,75 (0,87)	81,45 (0,54)
	NB	BNS	352 (10,30)	**86,29** (0,45)	**87,28** (0,72)	82,74 (1,16)	84,76 (1,23)
		CHI	72 (2,71)	**84,20** (0,46)	**85,85** (0,61)	82,73 (0,77)	84,07 (0,50)
		CDM	1247 (16,58)	**83,23** (0,59)	**85,52** (0,77)	70,75 (1,62)	77,54 (0,81)
		IG	71 (2,08)	**83,68** (0,70)	**85,52** (0,54)	83,16 (0,83	84,76 (0,53)
		MOR	421 (35,38)	**85,67** (1,10)	**86,81** (0,92)	82,57 (0,63)	84,69 (0,72)

Tabela B.4 Comparação dos métodos ALfFT com $f = 4$ contra VR.

Base de Dados	Classificador	FEF	m	AL4FT		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	626 (4,83)	**94,49** (0,24)	**94,49** (0,25)	93,26 (0,40)	93,28 (0,40)
		CHI	183 (7,68)	93,46 (0,57)	93,46 (0,56)	**94,43** (0,23)	**94,43** (0,23)
		CDM	1696 (6,40)	93,89 (0,28)	93,90 (0,28)	**94,09** (0,14)	**94,09** (0,14)
		IG	116 (2,08)	92,75 (0,11)	92,72 (0,12)	**93,83** (0,32)	**93,83** (0,32)
		MOR	544 (12,25)	**93,40** (0,23)	**93,43** (0,23)	93,01 (0,22)	93,01 (0,21)
	NB	BNS	626 (4,83)	**92,47** (0,16)	**92,53** (0,16)	90,69 (0,37)	90,79 (0,38)
		CHI	183 (7,68)	92,60 (0,29)	92,65 (0,27)	**93,25** (0,44)	**93,27** (0,44)
		CDM	1696 (6,40)	**91,90** (0,25)	**92,01** (0,28)	91,56 (0,36)	91,64 (0,38)
		IG	116 (2,08)	92,67 (0,31)	92,70 (0,31)	**93,34** (0,19)	**93,35** (0,18)
		MOR	544 (12,25)	**90,87** (0,31)	**90,92** (0,32)	90,15 (0,49)	90,18 (0,50)
Reuters 10	*k*-NN	BNS	487 (7,35)	60,25 (1,82)	**82,08** (0,58)	**60,36** (1,73)	80,98 (0,94)
		CHI	266 (5,07)	**59,99** (1,35)	**81,76** (0,46)	59,96 (1,08)	81,30 (0,63)
		CDM	696 (3,56)	**59,81** (1,67)	**81,74** (0,93)	58,64 (1,35)	80,73 (1,12)
		IG	255 (7,98)	**59,85** (1,59)	**81,78** (0,59)	59,28 (1,23)	81,32 (0,53)
		MOR	440 (9,56)	**60,54** (1,40)	**82,21** (0,59)	58,96 (1,57)	80,45 (0,78)
	NB	BNS	487 (7,35)	**63,44** (0,79)	**80,95** (0,51)	62,78 (1,08)	80,29 (0,44)
		CHI	266 (5,07)	61,59 (1,02)	**79,54** (0,51)	**62,14** (0,60)	79,11 (0,47)
		CDM	696 (3,56)	**63,40** (1,04)	**81,53** (0,70)	60,76 (0,98)	80,80 (0,33)
		IG	255 (7,98)	**61,02** (1,39)	**79,28** (0,73)	60,82 (0,50)	78,68 (0,46)
		MOR	440 (9,56)	**62,72** (1,85)	**80,50** (1,10)	61,95 (0,40)	79,08 (0,24)
WebKB	*k*-NN	BNS	451 (28,34)	**84,29** (1,21)	**85,64** (1,15)	78,07 (2,15)	80,50 (1,48)
		CHI	90 (2,77)	82,50 (1,58)	84,78 (1,05)	**82,60** (0,32)	**84,93** (0,25)
		CDM	1540 (30,63)	**79,50** (0,85)	**83,40** (0,48)	67,20 (1,52)	71,49 (1,15)
		IG	88 (3,74)	82,21 (1,13)	84,92 (0,90)	**82,97** (0,84)	**85,09** (0,68)
		MOR	536 (26,66)	79,82 (0,79)	**83,35** (1,15)	**80,27** (0,52)	82,64 (0,93)
	NB	BNS	451 (28,34)	**87,53** (0,55)	**88,47** (0,82)	82,95 (1,20)	84,95 (1,16)
		CHI	90 (2,77)	83,65 (1,26)	85,04 (0,79)	**84,01** (0,76)	**85,17** (0,91)
		CDM	1540 (30,63)	**83,65** (0,22)	**85,78** (0,60)	71,81 (1,44)	78,23 (0,98)
		IG	88 (3,74)	83,81 (1,69)	85,54 (1,14)	**84,47** (0,92)	**85,59** (0,76)
		MOR	536 (26,66)	**85,90** (0,66)	**87,02** (0,54)	83,32 (0,69)	85,00 (0,65)

Tabela B.5 Comparação dos métodos ALfFT com $f = 5$ contra VR.

Base de Dados	Classificador	FEF	m	AL5FT		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	1133 (11,11)	94,22 (0,12)	94,22 (0,12)	**94,22** (0,30)	**94,22** (0,30)
		CHI	267 (5,26)	93,13 (0,27)	93,14 (0,28)	**94,22** (0,25)	**94,22** (0,24)
		CDM	2175 (12,19)	93,35 (0,16)	93,36 (0,18)	**93,93** (0,34)	**93,94** (0,35)
		IG	161 (3,74)	92,56 (0,24)	92,56 (0,26)	**93,54** (0,34)	**93,55** (0,34)
		MOR	732 (10,82)	**93,43** (0,32)	**93,46** (0,30)	92,92 (0,13)	92,94 (0,14)
	NB	BNS	1133 (11,11)	**92,35** (0,17)	**92,42** (0,18)	92,00 (0,54)	92,08 (0,57)
		CHI	267 (5,26)	91,65 (0,49)	91,72 (0,53)	**92,80** (0,29)	**92,82** (0,30)
		CDM	2175 (12,19)	**91,77** (0,26)	**91,85** (0,28)	91,56 (0,46)	91,64 (0,48)
		IG	161 (3,74)	91,32 (0,53)	91,35 (0,52)	**92,59** (0,18)	**92,67** (0,19)
		MOR	732 (10,82)	**91,00** (0,26)	**91,04** (0,25)	90,16 (0,49)	90,18 (0,49)
Reuters 10	*k*-NN	BNS	660 (8,98)	**60,79** (0,67)	**82,17** (0,21)	60,20 (1,56)	81,22 (0,70)
		CHI	359 (8,50)	58,99 (0,51)	**81,45** (0,12)	**59,74** (0,71)	81,30 (0,22)
		CDM	850 (6,06)	**61,06** (1,41)	**82,43** (0,73)	59,35 (1,19)	81,16 (0,98)
		IG	342 (6,68)	59,54 (1,20)	**81,80** (0,31)	**59,72** (1,35)	81,59 (0,70)
		MOR	592 (10,42)	**60,22** (1,18)	**82,05** (0,35)	59,33 (2,27)	80,53 (1,17)
	NB	BNS	660 (8,98)	**62,58** (1,40)	**80,62** (0,67)	62,20 (0,67)	80,13 (0,33)
		CHI	359 (8,50)	61,28 (1,45)	**79,39** (0,81)	**61,77** (0,96)	79,33 (0,58)
		CDM	850 (6,06)	**63,43** (0,95)	**81,38** (0,43)	60,98 (0,55)	80,83 (0,19)
		IG	342 (6,68)	60,85 (1,73)	**79,12** (0,89)	**61,16** (0,80)	78,99 (0,54)
		MOR	592 (10,42)	**62,45** (1,39)	**80,34** (0,88)	61,65 (0,85)	79,11 (0,50)
WebKB	*k*-NN	BNS	544 (33,14)	**84,30** (0,82)	**85,62** (0,83)	78,96 (0,89)	81,38 (0,50)
		CHI	105 (1,41)	82,94 (1,55)	**85,50** (1,04)	**83,03** (1,68)	85,14 (1,39)
		CDM	1833 (31,38)	**79,81** (0,73)	**84,04** (0,87)	67,58 (1,67)	71,87 (0,61)
		IG	105 (2,00)	82,86 (1,76)	85,42 (1,22)	**83,78** (1,17)	**85,67** (0,75)
		MOR	637 (29,41)	79,44 (1,35)	**83,31** (1,44)	**79,44** (0,92)	82,09 (0,92)
	NB	BNS	544 (33,14)	**87,30** (0,53)	**88,14** (0,62)	84,28 (0,69)	86,04 (0,96)
		CHI	105 (1,41)	**82,45** (2,34)	83,54 (1,79)	84,42 (1,45)	**85,31** (1,27)
		CDM	1833 (31,38)	**84,38** (1,33)	**86,21** (1,38)	71,88 (1,91)	78,52 (1,04)
		IG	105 (2,00)	82,58 (2,15)	83,76 (1,66)	**85,72** (0,90)	**86,47** (0,91)
		MOR	637 (29,41)	**86,39** (1,30)	**87,31** (1,09)	84,01 (0,60)	85,31 (0,61)

APÊNDICE C

Tabelas de Resultados do método ALfFT-L

Tabela C.1 Comparação dos métodos ALfFT-L com $f = 1$ contra VR.

Base de Dados	Classificador	FEF	m	AL1FT-L		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	22 (0,82)	**90,09** (0,32)	**90,29** (0,31)	61,97 (3,02)	66,04 (3,93)
		CHI	24 (1,00)	**90,86** (0,27)	**90,98** (0,28)	75,84 (2,40)	78,05 (2,43)
		CDM	2890 (15,20)	87,75 (0,25)	87,76 (0,26)	**93,66** (0,32)	**93,68** (0,32)
		IG	23 (1,41)	**90,29** (0,42)	**90,48** (0,39)	77,13 (1,53)	78,71 (1,48)
		MOR	2940 (96,65)	89,50 (1,11)	89,56 (1,09)	**91,63** (0,28)	**91,64** (0,29)
	NB	BNS	22 (0,82)	**90,39** (0,50)	**90,50** (0,48)	61,97 (2,89)	67,67 (1,87)
		CHI	24 (1,00)	**91,05** (0,42)	**91,13** (0,40)	75,47 (0,86)	77,80 (0,95)
		CDM	2890 (15,20)	88,89 (0,21)	89,00 (0,22)	**91,40** (0,36)	**91,49** (0,38)
		IG	23 (1,41)	**90,35** (0,50)	**90,46** (0,46)	77,44 (0,66)	78,74 (0,50)
		MOR	2940 (96,65)	**89,85** (0,91)	**89,92** (0,88)	89,24 (0,24)	89,34 (0,26)
Reuters 10	*k*-NN	BNS	136 (4,40)	**61,20** (1,18)	**81,37** (0,68)	55,26 (1,69)	77,06 (1,16)
		CHI	107 (3,16)	**59,71** (1,72)	**81,05** (0,82)	58,96 (1,91)	79,36 (1,02)
		CDM	1061 (18,10)	**60,13** (1,64)	**81,90** (0,74)	58,93 (1,29)	80,83 (0,74)
		IG	101 (1,83)	**59,87** (0,80)	**80,97** (0,47)	57,96 (2,11)	79,87 (0,99)
		MOR	459 (28,15)	**60,23** (1,34)	**81,22** (0,35)	58,87 (1,63)	80,45 (0,94)
	NB	BNS	136 (4,40)	**64,08** (1,47)	**81,49** (0,82)	57,75 (0,17)	77,26 (0,42)
		CHI	107 (3,16)	**63,66** (1,12)	**80,97** (0,60)	63,10 (0,76)	78,88 (0,91)
		CDM	1061 (18,10)	**61,81** (1,16)	80,53 (0,79)	61,59 (0,93)	**80,99** (0,56)
		IG	101 (1,83)	**62,97** (1,69)	**80,68** (0,95)	62,08 (1,11)	79,13 (1,05)
		MOR	459 (28,15)	**62,60** (1,50)	**81,52** (0,61)	61,79 (0,84)	79,13 (0,33)
WebKB	*k*-NN	BNS	219 (19,38)	**80,72** (2,80)	**82,38** (2,29)	76,86 (1,24)	80,04 (1,08)
		CHI	59 (2,00)	**82,14** (0,50)	**84,04** (0,63)	78,52 (1,78)	81,24 (1,23)
		CDM	1057 (26,51)	**85,00** (1,35)	**87,33** (0,73)	63,10 (2,21)	68,64 (1,19)
		IG	56 (2,58)	**82,09** (0,77)	**83,78** (0,76)	79,69 (1,09)	81,90 (0,98)
		MOR	509 (84,03)	77,35 (1,42)	80,81 (0,99)	**79,90** (1,51)	**82,28** (1,54)
	NB	BNS	219 (19,38)	**84,37** (2,04)	**85,88** (1,96)	80,36 (2,08)	83,64 (1,67)
		CHI	59 (2,00)	**83,03** (1,19)	**84,40** (1,13)	79,94 (1,47)	82,28 (1,09)
		CDM	1057 (26,51)	**86,79** (1,10)	**87,38** (1,25)	67,39 (1,26)	75,73 (0,39)
		IG	56 (2,58)	**82,82** (1,18)	**84,31** (1,17)	79,98 (1,06)	82,38 (1,08)
		MOR	509 (84,03)	80,83 (1,71)	83,76 (1,51)	**83,36** (1,39)	**84,93** (1,39)

Tabela C.2 Comparação dos métodos ALfFT-L com $f = 2$ contra VR.

Base de Dados	Classificador	FEF	m	AL2FT-L		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	43 (5,23)	**94,00** (0,26)	**94,01** (0,24)	71,11 (4,47)	73,27 (4,58)
		CHI	81 (1,83)	93,59 (0,48)	93,57 (0,48)	**94,37** (0,39)	**94,37** (0,37)
		CDM	3774 (14,25)	85,99 (0,48)	85,97 (0,48)	**93,06** (0,15)	**93,08** (0,16)
		IG	61 (1,15)	93,39 (0,55)	93,37 (0,56)	**94,28** (0,35)	**94,28** (0,34)
		MOR	5736 (152,74)	**92,20** (0,96)	**92,22** (0,94)	89,22 (0,66)	89,24 (0,67)
	NB	BNS	43 (5,23)	**93,54** (0,12)	**93,57** (0,13)	69,23 (0,31)	72,86 (0,32)
		CHI	81 (1,83)	93,32 (0,59)	93,33 (0,59)	**93,60** (0,21)	**93,58** (0,19)
		CDM	3774 (14,25)	87,02 (0,25)	87,13 (0,26)	**91,02** (0,34)	**91,12** (0,35)
		IG	61 (1,15)	93,50 (0,84)	93,47 (0,84)	**93,61** (0,40)	**93,59** (0,38)
		MOR	5736 (152,74)	**92,14** (0,64)	**92,21** (0,63)	85,74 (0,45)	85,93 (0,40)
Reuters 10	*k*-NN	BNS	322 (11,14)	**60,55** (1,50)	**81,75** (0,71)	59,61 (0,99)	80,40 (0,66)
		CHI	237 (10,47)	**60,37** (1,48)	**81,75** (0,70)	60,25 (1,35)	81,34 (0,81)
		CDM	1595 (9,54)	59,85 (1,79)	**81,89** (0,76)	**59,98** (1,87)	81,85 (0,79)
		IG	195 (6,03)	**59,92** (1,87)	**81,59** (0,74)	59,22 (1,06)	81,11 (0,43)
		MOR	1099 (31,79)	**61,12** (0,47)	**82,07** (0,24)	60,22 (1,04)	81,82 (0,35)
	NB	BNS	322 (11,14)	**64,77** (0,76)	**81,93** (0,45)	62,41 (0,76)	80,08 (0,21)
		CHI	237 (10,47)	**64,73** (1,26)	**81,31** (0,75)	62,52 (1,33)	79,16 (0,66)
		CDM	1595 (9,54)	61,68 (0,91)	80,69 (0,68)	**61,89** (1,09)	**81,22** (0,45)
		IG	195 (6,03)	**64,26** (1,31)	**81,16** (0,71)	61,51 (1,06)	79,12 (0,66)
		MOR	1099 (31,79)	**61,92** (0,58)	**81,25** (0,38)	61,03 (0,76)	79,50 (0,52)
WebKB	*k*-NN	BNS	405 (45,93)	**82,76** (1,45)	**85,16** (0,85)	77,52 (1,97)	79,88 (1,92)
		CHI	99 (5,45)	83,16 (0,69)	**85,33** (0,22)	**83,35** (1,12)	85,33 (1,12)
		CDM	1799 (32,34)	**83,09** (0,74)	**86,17** (0,84)	67,56 (1,20)	71,97 (0,32)
		IG	92 (6,06)	**83,54** (1,44)	**85,38** (1,07)	83,04 (1,25)	85,04 (1,19)
		MOR	972 (91,39)	78,18 (2,63)	82,49 (1,35)	**80,10** (0,37)	**83,16** (0,85)
	NB	BNS	405 (45,93)	**86,21** (0,82)	**87,40** (0,84)	82,52 (1,24)	84,54 (1,35)
		CHI	99 (5,45)	82,27 (1,25)	83,31 (1,09)	**84,68** (1,11)	**85,49** (1,18)
		CDM	1799 (32,34)	**86,21** (0,99)	**86,64** (1,20)	71,77 (1,75)	78,38 (1,13)
		IG	92 (6,06)	81,43 (1,38)	82,64 (1,22)	**84,91** (0,58)	**85,97** (0,32)
		MOR	972 (91,39)	84,55 (0,44)	**86,31** (0,60)	**85,22** (1,06)	86,21 (0,98)

Tabela C.3 Comparação dos métodos ALfFT-L com $f = 3$ contra VR.

Base de Dados	Classificador	FEF	m	AL3FT-L		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	548 (17,15)	**93,67** (0,16)	**93,67** (0,14)	93,33 (0,69)	93,35 (0,69)
		CHI	578 (24,63)	92,89 (0,30)	92,88 (0,30)	**93,20** (0,23)	**93,20** (0,25)
		CDM	4290 (23,81)	84,55 (0,71)	84,50 (0,70)	**92,96** (0,15)	**92,98** (0,16)
		IG	327 (4,40)	92,18 (0,47)	92,16 (0,46)	**92,82** (0,20)	**92,82** (0,21)
		MOR	8428 (237,05)	**92,62** (0,25)	**92,65** (0,24)	88,45 (0,42)	88,47 (0,42)
	NB	BNS	548 (17,15)	**93,29** (0,30)	**93,29** (0,32)	90,35 (0,42)	90,47 (0,41)
		CHI	578 (24,63)	**92,57** (0,09)	**92,60** (0,08)	91,82 (0,22)	91,92 (0,22)
		CDM	4290 (23,81)	86,03 (0,36)	86,18 (0,36)	**91,02** (0,23)	**91,10** (0,26)
		IG	327 (4,40)	**92,07** (0,38)	**92,13** (0,36)	91,02 (0,35)	91,12 (0,38)
		MOR	8428 (237,05)	**91,86** (0,27)	**91,94** (0,28)	83,94 (0,32)	84,18 (0,32)
Reuters 10	*k*-NN	BNS	581 (17,08)	**61,08** (1,12)	**82,14** (0,61)	60,66 (1,45)	81,13 (0,65)
		CHI	378 (6,35)	**60,45** (1,37)	**81,94** (0,70)	59,76 (0,61)	81,31 (0,23)
		CDM	1960 (3,87)	60,05 (1,43)	**81,98** (0,48)	**60,23** (1,51)	81,92 (0,59)
		IG	304 (4,28)	**60,61** (0,76)	**81,99** (0,43)	59,41 (1,03)	81,45 (0,38)
		MOR	1705 (25,27)	60,44 (1,02)	81,94 (0,49)	**60,64** (0,74)	**81,98** (0,41)
	NB	BNS	581 (17,08)	**63,54** (0,99)	**80,93** (0,54)	62,69 (0,76)	80,27 (0,44)
		CHI	378 (6,35)	**63,03** (0,76)	**80,40** (0,30)	61,95 (1,59)	79,43 (0,89)
		CDM	1960 (3,87)	61,36 (0,71)	80,66 (0,59)	**61,60** (0,80)	**81,26** (0,47)
		IG	304 (4,28)	**62,69** (0,86)	**80,24** (0,43)	61,06 (1,00)	78,87 (0,62)
		MOR	1705 (25,27)	**61,65** (0,55)	**81,19** (0,39)	60,71 (0,83)	79,76 (0,59)
WebKB	*k*-NN	BNS	575 (42,88)	**84,13** (0,37)	**86,02** (0,49)	78,97 (0,83)	81,47 (0,16)
		CHI	143 (6,16)	**84,40** (1,24)	**86,02** (0,95)	83,88 (0,27)	85,88 (0,51)
		CDM	2423 (23,71)	**81,36** (0,37)	**84,95** (0,92)	73,19 (1,73)	76,71 (2,13)
		IG	128 (7,07)	83,60 (1,19)	85,59 (0,80)	**83,99** (0,77)	**85,76** (0,28)
		MOR	1390 (98,34)	78,85 (1,71)	**83,45** (0,84)	**78,95** (1,79)	83,24 (2,04)
	NB	BNS	575 (42,88)	**86,85** (0,68)	**87,71** (0,43)	84,03 (0,92)	85,86 (1,14)
		CHI	143 (6,16)	80,17 (1,98)	81,38 (1,72)	**84,03** (1,61)	**84,81** (1,22)
		CDM	2423 (23,71)	**85,54** (1,14)	**86,12** (1,23)	77,12 (2,47)	82,07 (1,92)
		IG	128 (7,07)	79,64 (2,48)	80,90 (2,09)	**85,67** (0,93)	**86,14** (0,57)
		MOR	1390 (98,34)	**85,37** (0,50)	**86,71** (0,61)	85,18 (0,93)	86,26 (0,88)

Tabela C.4 Comparação dos métodos ALfFT-L com $f = 4$ contra VR.

Base de Dados	Classificador	FEF	m	AL4FT-L		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	1744 (27,49)	93,24 (0,13)	93,22 (0,13)	**94,26** (0,32)	**94,26** (0,32)
		CHI	1260 (22,87)	**92,32** (0,56)	**92,32** (0,57)	91,70 (0,38)	91,72 (0,39)
		CDM	4652 (22,43)	84,18 (0,57)	84,13 (0,56)	**92,83** (0,32)	**92,85** (0,34)
		IG	777 (7,53)	**91,72** (0,23)	**91,72** (0,25)	90,26 (0,31)	90,29 (0,33)
		MOR	10875 (184,58)	**91,82** (0,35)	**91,85** (0,36)	88,47 (0,15)	88,47 (0,14)
	NB	BNS	1744 (27,49)	**92,98** (0,07)	**93,00** (0,08)	91,93 (0,48)	92,00 (0,51)
		CHI	1260 (22,87)	**92,15** (0,14)	**92,22** (0,11)	91,47 (0,29)	91,60 (0,33)
		CDM	4652 (22,43)	85,52 (0,52)	85,68 (0,52)	**90,71** (0,42)	**90,80** (0,44)
		IG	777 (7,53)	**90,55** (0,43)	**90,67** (0,46)	89,86 (0,23)	90,02 (0,25)
		MOR	10875 (184,58)	**91,34** (0,39)	**91,44** (0,39)	81,96 (0,67)	82,24 (0,60)
Reuters 10	*k*-NN	BNS	876 (17,75)	**60,19** (1,44)	**81,87** (0,61)	59,92 (1,56)	81,06 (0,73)
		CHI	528 (5,72)	**60,46** (1,67)	**82,02** (0,79)	59,85 (1,02)	81,69 (0,49)
		CDM	2251 (9,56)	60,23 (1,04)	**82,10** (0,34)	**60,53** (1,23)	82,04 (0,52)
		IG	424 (1,83)	**59,77** (1,49)	**81,71** (0,70)	59,40 (0,63)	81,50 (0,35)
		MOR	2329 (24,96)	**60,50** (1,23)	**81,98** (0,44)	60,47 (1,15)	81,96 (0,33)
	NB	BNS	876 (17,75)	**62,98** (1,04)	**80,80** (0,67)	61,69 (0,66)	79,82 (0,46)
		CHI	528 (5,72)	**63,06** (1,39)	**80,56** (0,96)	62,54 (1,20)	79,92 (0,59)
		CDM	2251 (9,56)	61,16 (0,84)	80,68 (0,65)	**61,29** (0,94)	**81,00** (0,58)
		IG	424 (1,83)	**61,97** (0,56)	**80,03** (0,47)	61,35 (0,91)	79,23 (0,66)
		MOR	2329 (24,96)	**61,26** (0,95)	**81,06** (0,44)	60,25 (0,91)	79,78 (0,46)
WebKB	*k*-NN	BNS	734 (42,62)	**84,70** (1,13)	**86,97** (0,91)	79,87 (0,65)	82,16 (0,56)
		CHI	179 (4,43)	**83,75** (1,05)	85,69 (0,67)	83,63 (0,62)	**85,74** (0,50)
		CDM	2937 (27,19)	**80,25** (0,93)	**84,09** (1,15)	75,82 (1,73)	80,07 (1,40)
		IG	165 (5,07)	83,89 (0,92)	85,90 (0,92)	**85,02** (0,76)	**86,83** (0,59)
		MOR	1776 (117,66)	**80,89** (2,05)	**84,69** (0,70)	77,30 (2,00)	82,73 (1,88)
	NB	BNS	734 (42,62)	**86,96** (0,32)	**87,76** (0,28)	85,25 (1,09)	86,61 (1,38)
		CHI	179 (4,43)	74,66 (3,40)	76,09 (3,28)	**82,79** (1,24)	**83,66** (0,97)
		CDM	2937 (27,19)	**84,57** (1,59)	**85,45** (1,40)	79,01 (1,53)	83,43 (1,11)
		IG	165 (5,07)	75,63 (1,68)	76,71 (1,56)	**83,51** (1,35)	**84,26** (0,92)
		MOR	1776 (117,66)	**85,78** (0,90)	**86,98** (0,60)	85,25 (0,93)	86,43 (0,82)

Tabela C.5 Comparação dos métodos ALfFT-L com $f = 5$ contra VR.

Base de Dados	Classificador	FEF	m	AL5FT-L		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	3231 (19,92)	92,66 (0,25)	92,65 (0,27)	**94,22** (0,32)	**94,23** (0,32)
		CHI	1981 (29,43)	**91,23** (0,50)	**91,23** (0,51)	89,19 (0,27)	89,20 (0,26)
		CDM	4933 (35,07)	84,02 (0,28)	83,97 (0,28)	**92,58** (0,32)	**92,60** (0,34)
		IG	1163 (24,86)	**90,28** (0,58)	**90,28** (0,58)	87,45 (0,44)	87,47 (0,45)
		MOR	13000 (95,56)	**90,85** (0,33)	**90,88** (0,33)	88,02 (0,30)	88,01 (0,30)
	NB	BNS	3231 (19,92)	**92,79** (0,31)	**92,85** (0,34)	91,80 (0,45)	91,87 (0,48)
		CHI	1981 (29,43)	**92,21** (0,23)	**92,29** (0,26)	91,31 (0,28)	91,42 (0,31)
		CDM	4933 (35,07)	85,05 (0,47)	85,22 (0,46)	**90,50** (0,41)	**90,59** (0,43)
		IG	1163 (24,86)	90,40 (0,50)	90,55 (0,52)	**90,47** (0,33)	**90,57** (0,32)
		MOR	13000 (95,56)	**90,46** (0,36)	**90,57** (0,37)	80,13 (0,39)	80,46 (0,31)
Reuters 10	*k*-NN	BNS	1224 (27,36)	60,20 (1,48)	**82,01** (0,74)	**60,86** (0,78)	81,79 (0,49)
		CHI	700 (8,04)	60,35 (2,12)	**82,12** (0,69)	**60,84** (0,84)	81,92 (0,38)
		CDM	2552 (6,58)	60,18 (1,10)	**82,15** (0,27)	**60,56** (1,12)	82,14 (0,67)
		IG	565 (5,03)	60,08 (1,58)	**81,99** (0,67)	**60,33** (1,61)	81,76 (0,80)
		MOR	2905 (11,99)	**60,72** (1,42)	**82,13** (0,62)	60,53 (1,03)	82,00 (0,40)
	NB	BNS	1224 (27,36)	**62,33** (1,11)	**80,85** (0,72)	61,21 (0,88)	79,76 (0,55)
		CHI	700 (8,04)	**63,01** (1,29)	**80,60** (0,90)	62,60 (1,06)	80,16 (0,63)
		CDM	2552 (6,58)	60,83 (0,59)	80,65 (0,52)	**60,93** (1,04)	**80,86** (0,63)
		IG	565 (5,03)	62,14 (0,82)	**80,08** (0,67)	**62,28** (1,25)	79,95 (0,81)
		MOR	2905 (11,99)	**61,14** (0,74)	**81,31** (0,33)	59,93 (0,80)	79,90 (0,56)
WebKB	*k*-NN	BNS	918 (46,07)	**83,86** (1,24)	**86,62** (0,93)	80,51 (0,58)	82,54 (0,94)
		CHI	213 (6,71)	83,53 (0,86)	85,76 (1,14)	**83,74** (1,11)	**85,88** (1,07)
		CDM	3411 (34,97)	**80,11** (1,36)	**83,73** (1,38)	75,42 (1,58)	80,73 (0,73)
		IG	193 (3,37)	83,47 (1,09)	85,71 (0,97)	**84,49** (0,69)	**86,50** (0,76)
		MOR	2167 (132,91)	**82,07** (0,81)	**85,76** (0,34)	75,93 (2,61)	81,90 (2,00)
	NB	BNS	918 (46,07)	85,91 (0,76)	86,92 (0,67)	**86,00** (0,67)	**87,09** (0,79)
		CHI	213 (6,71)	70,19 (2,41)	71,64 (2,32)	**82,11** (1,55)	**82,95** (1,48)
		CDM	3411 (34,97)	**83,90** (1,14)	84,78 (1,01)	82,12 (1,41)	**84,85** (1,09)
		IG	193 (3,37)	69,33 (0,66)	70,61 (0,78)	**82,56** (0,50)	**83,45** (0,74)
		MOR	2167 (132,91)	**85,85** (0,95)	**86,83** (0,59)	84,82 (1,20)	86,07 (1,04)

APÊNDICE D

Tabelas de Resultados do método ALfFT-R

Tabela D.1 Comparação dos métodos ALfFT-R com $f = 1$ contra VR.

Base de Dados	Classificador	FEF	m	AL1FT-R		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	16 (1,29)	**68,49** (3,22)	**70,41** (2,76)	44,23 (2,38)	51,80 (3,51)
		CHI	29 (0,58)	**93,14** (0,37)	**93,18** (0,36)	88,00 (1,94)	88,90 (1,85)
		CDM	168 (6,78)	**91,36** (0,39)	**91,40** (0,37)	75,53 (1,24)	76,61 (1,28)
		IG	10 (0,00)	**47,78** (0,15)	**53,79** (0,23)	43,66 (1,26)	50,13 (1,71)
		OR	78 (6,76)	**87,03** (0,94)	**87,17** (0,92)	84,44 (0,22)	84,71 (0,24)
	NB	BNS	16 (1,29)	**67,26** (3,98)	**69,66** (3,11)	42,72 (1,08)	52,57 (1,84)
		CHI	29 (0,58)	**91,44** (0,22)	**91,56** (0,24)	85,54 (0,34)	87,00 (0,24)
		CDM	168 (6,78)	**89,46** (0,49)	**89,52** (0,48)	72,62 (1,37)	73,87 (1,70)
		IG	10 (0,00)	**49,59** (0,32)	**54,60** (0,44)	48,52 (0,49)	53,04 (0,52)
		OR	78 (6,76)	**84,47** (0,79)	**84,75** (0,69)	80,68 (0,52)	81,26 (0,55)
Reuters 10	*k*-NN	BNS	60 (2,08)	**53,79** (2,38)	**75,68** (1,00)	48,67 (1,95)	73,39 (1,65)
		CHI	11 (0,82)	**35,56** (3,13)	**64,74** (0,87)	30,89 (1,60)	63,03 (0,77)
		CDM	123 (3,37)	**57,40** (1,86)	**78,47** (0,47)	50,80 (3,18)	75,20 (1,30)
		IG	9 (0,00)	**23,13** (1,56)	57,70 (4,51)	22,45 (1,08)	**57,84** (4,60)
		OR	52 (5,42)	**53,78** (1,65)	**75,56** (1,53)	53,08 (2,94)	75,05 (2,06)
	NB	BNS	60 (2,08)	**52,45** (0,63)	**76,11** (0,91)	46,43 (3,32)	72,79 (0,92)
		CHI	11 (0,82)	**35,00** (3,07)	**64,50** (0,86)	28,51 (0,96)	63,17 (0,69)
		CDM	123 (3,37)	**54,89** (0,69)	**78,46** (0,18)	47,76 (2,53)	74,47 (1,13)
		IG	9 (0,00)	**19,65** (0,32)	58,14 (0,42)	19,47 (0,23)	**58,17** (0,46)
		OR	52 (5,42)	**54,92** (1,93)	**75,54** (1,27)	52,41 (1,41)	74,00 (1,33)
WebKB	*k*-NN	BNS	56 (6,27)	**69,35** (1,19)	**72,99** (1,64)	60,67 (3,91)	70,11 (6,06)
		CHI	14 (1,00)	**64,83** (0,88)	**68,99** (0,97)	53,92 (4,59)	58,92 (4,61)
		CDM	244 (6,16)	**61,44** (4,56)	**67,64** (4,99)	47,32 (0,97)	56,85 (1,80)
		IG	13 (0,82)	**64,41** (1,00)	**68,51** (1,10)	54,24 (2,93)	59,87 (2,43)
		OR	94 (6,08)	**73,54** (1,95)	**76,54** (1,98)	53,94 (5,97)	59,15 (6,02)
	NB	BNS	56 (6,27)	**73,78** (1,32)	**79,54** (0,91)	63,16 (2,10)	74,61 (1,35)
		CHI	14 (1,00)	**65,77** (1,56)	**73,64** (0,72)	56,81 (3,17)	68,33 (1,42)
		CDM	244 (6,16)	**66,63** (2,83)	**74,24** (2,80)	54,46 (1,93)	67,21 (1,27)
		IG	13 (0,82)	**65,33** (2,03)	**73,44** (0,85)	56,58 (4,95)	67,90 (2,00)
		OR	94 (6,08)	**74,64** (1,05)	**79,50** (0,51)	61,72 (5,63)	70,11 (4,54)

Tabela D.2 Comparação dos métodos ALfFT-R com $f = 2$ contra VR.

Base de Dados	Classificador	FEF	m	AL2FT-R		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	24 (1,91)	**83,85** (1,92)	**84,54** (1,76)	65,42 (6,08)	68,62 (6,02)
		CHI	36 (0,82)	**94,01** (0,30)	**94,02** (0,29)	93,97 (0,30)	93,99 (0,29)
		CDM	336 (6,86)	**92,71** (0,63)	**92,73** (0,62)	86,83 (0,47)	87,25 (0,40)
		IG	29 (1,29)	**92,56** (0,12)	**92,56** (0,11)	85,35 (2,12)	86,04 (1,88)
		OR	121 (16,59)	**88,83** (0,61)	**88,95** (0,64)	87,73 (0,35)	87,85 (0,31)
	NB	BNS	24 (1,91)	**83,42** (2,67)	**83,91** (2,27)	65,03 (2,28)	69,66 (1,87)
		CHI	36 (0,82)	**92,92** (0,24)	**92,96** (0,24)	92,87 (0,22)	92,94 (0,22)
		CDM	336 (6,86)	**90,58** (0,62)	**90,69** (0,62)	83,40 (0,44)	84,36 (0,37)
		IG	29 (1,29)	**91,39** (0,21)	**91,38** (0,22)	84,46 (1,67)	85,47 (1,53)
		OR	121 (16,59)	**85,47** (0,17)	**85,67** (0,21)	84,19 (0,88)	84,40 (0,92)
Reuters 10	*k*-NN	BNS	79 (3,92)	**53,69** (1,25)	**75,92** (0,82)	51,90 (1,20)	75,21 (1,50)
		CHI	20 (1,00)	52,63 (3,27)	71,81 (1,61)	**53,22** (3,15)	**72,42** (1,73)
		CDM	193 (5,00)	**58,62** (1,56)	**79,36** (0,79)	54,55 (1,95)	77,29 (0,82)
		IG	10 (1,00)	**30,81** (3,39)	**63,35** (1,75)	25,13 (1,44)	58,49 (5,15)
		OR	59 (6,45)	**55,81** (1,05)	**77,07** (1,03)	54,85 (1,27)	76,21 (1,15)
	NB	BNS	79 (3,92)	**54,40** (1,13)	**75,86** (1,02)	53,08 (0,43)	75,45 (0,78)
		CHI	20 (1,00)	**49,60** (1,03)	**71,01** (0,74)	48,71 (1,06)	70,68 (0,34)
		CDM	193 (5,00)	**58,26** (0,64)	**79,45** (0,28)	53,20 (0,93)	77,49 (0,27)
		IG	10 (1,00)	**26,45** (0,49)	**62,04** (0,56)	20,85 (1,03)	59,12 (1,05)
		OR	59 (6,45)	**56,07** (1,89)	**76,26** (0,70)	55,88 (1,48)	75,64 (0,60)
WebKB	*k*-NN	BNS	111 (4,08)	**77,05** (1,74)	**79,64** (1,25)	67,47 (1,76)	71,75 (3,26)
		CHI	21 (1,15)	**71,36** (0,74)	**75,40** (0,58)	65,86 (0,71)	69,59 (0,65)
		CDM	440 (14,75)	**68,89** (1,75)	**74,16** (2,36)	52,75 (2,91)	60,58 (1,79)
		IG	21 (1,41)	**73,02** (1,76)	**76,88** (1,92)	66,39 (0,89)	69,99 (1,11)
		OR	144 (11,27)	**75,96** (2,35)	**78,60** (2,08)	68,67 (1,73)	72,76 (1,08)
	NB	BNS	111 (4,08)	**78,23** (0,77)	**82,04** (1,05)	71,31 (1,81)	78,35 (1,27)
		CHI	21 (1,15)	**69,19** (1,05)	**76,38** (1,03)	64,60 (1,88)	73,85 (1,09)
		CDM	440 (14,75)	**72,55** (0,99)	**78,90** (0,79)	58,34 (2,10)	69,93 (0,97)
		IG	21 (1,41)	**69,55** (0,63)	**77,12** (1,09)	67,60 (1,72)	75,23 (1,42)
		OR	144 (11,27)	**78,51** (1,63)	**82,02** (1,54)	73,21 (2,23)	78,45 (1,44)

Tabela D.3 Comparação dos métodos ALfFT-R com $f = 3$ contra VR.

Base de Dados	Classificador	FEF	m	AL3FT-R		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	79 (0,82)	**94,20** (0,30)	**94,21** (0,31)	78,52 (3,44)	79,88 (3,29)
		CHI	48 (1,15)	**94,16** (0,21)	**94,15** (0,20)	94,10 (0,21)	94,10 (0,22)
		CDM	499 (6,48)	**93,69** (0,50)	**93,70** (0,48)	88,25 (0,57)	88,56 (0,55)
		IG	39 (1,00)	94,01 (0,30)	94,04 (0,29)	**94,05** (0,33)	**94,07** (0,31)
		OR	169 (8,02)	**90,76** (1,04)	**90,83** (1,02)	89,32 (0,53)	89,40 (0,53)
	NB	BNS	79 (0,82)	**92,97** (0,33)	**93,04** (0,34)	77,12 (0,71)	78,58 (0,51)
		CHI	48 (1,15)	93,10 (0,31)	93,11 (0,31)	**93,13** (0,33)	**93,15** (0,33)
		CDM	499 (6,48)	**91,48** (0,57)	**91,58** (0,59)	85,08 (0,45)	85,74 (0,45)
		IG	39 (1,00)	**93,49** (0,41)	**93,47** (0,41)	93,08 (0,29)	93,07 (0,28)
		OR	169 (8,02)	**87,44** (0,54)	**87,56** (0,48)	85,51 (0,63)	85,72 (0,61)
Reuters 10	*k*-NN	BNS	93 (6,73)	**54,67** (0,31)	**76,40** (0,40)	54,06 (1,42)	76,21 (0,74)
		CHI	25 (1,41)	**54,52** (2,20)	73,29 (1,20)	54,20 (2,62)	**73,72** (1,46)
		CDM	266 (6,66)	**58,18** (2,13)	**80,00** (1,04)	55,70 (2,28)	78,29 (0,91)
		IG	20 (1,00)	49,36 (1,62)	72,06 (1,06)	**49,42** (2,20)	**72,08** (1,22)
		OR	72 (2,45)	**57,13** (2,00)	**77,79** (1,03)	56,60 (1,15)	77,46 (0,58)
	NB	BNS	93 (6,73)	**57,36** (1,98)	**77,08** (1,20)	54,98 (1,47)	76,11 (0,82)
		CHI	25 (1,41)	52,83 (1,16)	73,53 (0,92)	**53,06** (0,96)	**73,65** (0,41)
		CDM	266 (6,66)	**60,11** (1,32)	**79,86** (0,46)	55,37 (0,53)	78,74 (0,42)
		IG	20 (1,00)	46,24 (1,12)	71,66 (1,08)	**46,26** (1,24)	**71,79** (0,70)
		OR	72 (2,45)	**59,05** (0,59)	**77,47** (0,64)	57,56 (0,41)	76,40 (0,24)
WebKB	*k*-NN	BNS	173 (6,73)	**76,83** (1,44)	**79,62** (1,42)	73,83 (0,84)	77,31 (0,69)
		CHI	30 (2,24)	**72,76** (0,48)	**76,85** (0,62)	71,24 (1,22)	75,73 (1,45)
		CDM	618 (18,37)	**72,22** (1,55)	**77,14** (1,46)	56,44 (0,98)	63,87 (1,01)
		IG	30 (2,65)	**74,61** (1,67)	**78,62** (1,04)	73,19 (1,97)	76,88 (1,71)
		OR	200 (11,28)	**77,16** (1,61)	**79,88** (1,46)	72,94 (1,70)	75,99 (1,70)
	NB	BNS	173 (6,73)	**80,40** (1,41)	**83,19** (1,58)	77,81 (1,18)	82,26 (1,04)
		CHI	30 (2,24)	70,64 (1,85)	77,56 (1,21)	**70,97** (1,29)	**77,59** (1,42)
		CDM	618 (18,37)	**75,48** (1,38)	**80,90** (1,16)	61,30 (0,63)	71,71 (0,44)
		IG	30 (2,65)	**71,72** (1,13)	**78,52** (1,16)	71,61 (0,49)	78,14 (0,38)
		OR	200 (11,28)	**79,55** (1,56)	**82,50** (1,59)	77,27 (2,13)	80,85 (1,60)

Tabela D.4 Comparação dos métodos ALfFT-R com $f = 4$ contra VR.

Base de Dados	Classificador	FEF	m	AL4FT-R		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	168 (5,57)	**94,39** (0,44)	**94,38** (0,44)	89,13 (0,49)	89,16 (0,48)
		CHI	59 (2,45)	**94,41** (0,33)	**94,40** (0,33)	94,21 (0,25)	94,21 (0,25)
		CDM	667 (9,06)	**94,14** (0,20)	**94,13** (0,18)	92,39 (0,22)	92,43 (0,21)
		IG	45 (1,00)	94,00 (0,20)	93,99 (0,19)	**94,04** (0,31)	**94,03** (0,29)
		OR	220 (6,43)	**91,97** (1,11)	**92,00** (1,06)	90,15 (0,34)	90,21 (0,35)
	NB	BNS	168 (5,57)	**92,76** (0,22)	**92,83** (0,20)	86,95 (0,32)	87,19 (0,32)
		CHI	59 (2,45)	**93,79** (0,39)	**93,78** (0,38)	93,30 (0,39)	93,32 (0,38)
		CDM	667 (9,06)	**91,87** (0,56)	**91,95** (0,57)	89,80 (0,23)	89,91 (0,23)
		IG	45 (1,00)	93,25 (0,29)	93,23 (0,30)	**93,36** (0,26)	**93,35** (0,26)
		OR	220 (6,43)	**88,81** (0,83)	**88,86** (0,76)	86,67 (0,56)	86,82 (0,53)
Reuters 10	*k*-NN	BNS	112 (5,72)	**58,09** (1,77)	**78,99** (0,96)	54,14 (1,46)	76,18 (0,92)
		CHI	35 (1,15)	**55,85** (1,06)	**75,03** (1,14)	55,61 (1,91)	74,65 (1,14)
		CDM	337 (7,79)	**58,74** (0,93)	**80,40** (0,38)	57,08 (0,95)	79,21 (0,48)
		IG	27 (1,00)	**51,34** (1,85)	**73,68** (0,89)	50,68 (1,57)	73,38 (0,85)
		OR	87 (6,24)	**57,49** (2,37)	**78,56** (1,11)	56,58 (2,30)	77,64 (1,18)
	NB	BNS	112 (5,72)	**60,03** (1,21)	**78,32** (0,89)	55,69 (0,96)	76,22 (0,82)
		CHI	35 (1,15)	**57,00** (1,50)	**75,42** (0,77)	56,55 (1,01)	75,15 (0,75)
		CDM	337 (7,79)	**61,21** (1,87)	**80,34** (0,73)	57,98 (1,00)	79,79 (0,47)
		IG	27 (1,00)	**49,93** (1,85)	**73,99** (1,26)	48,74 (0,89)	73,22 (0,78)
		OR	87 (6,24)	**59,64** (1,35)	**78,24** (0,90)	59,20 (0,81)	77,29 (0,41)
WebKB	*k*-NN	BNS	231 (11,03)	**78,01** (0,53)	**80,71** (0,57)	76,84 (1,97)	79,93 (1,80)
		CHI	35 (2,71)	**75,86** (2,37)	**78,85** (1,68)	74,51 (1,83)	78,28 (1,35)
		CDM	777 (20,84)	**73,77** (0,87)	**77,90** (1,03)	60,45 (1,64)	66,90 (1,41)
		IG	35 (3,11)	**76,54** (1,40)	**79,23** (1,08)	74,78 (0,68)	78,66 (0,55)
		OR	257 (18,81)	**78,10** (0,93)	**80,61** (0,75)	74,44 (1,66)	77,38 (1,48)
	NB	BNS	231 (11,03)	**82,35** (0,95)	**84,59** (1,25)	80,61 (1,58)	83,83 (1,36)
		CHI	35 (2,71)	**74,27** (3,50)	**79,21** (2,11)	72,83 (1,07)	78,49 (1,05)
		CDM	777 (20,84)	**77,57** (1,31)	**82,40** (1,16)	64,36 (1,98)	73,64 (0,88)
		IG	35 (3,11)	**75,44** (3,23)	**79,85** (1,81)	73,10 (1,46)	78,71 (0,60)
		OR	257 (18,81)	**81,46** (1,20)	**83,69** (1,23)	78,83 (1,18)	81,99 (1,41)

Tabela D.5 Comparação dos métodos ALfFT-R com $f = 5$ contra VR.

Base de Dados	Classificador	FEF	m	AL5FT-R		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	284 (5,39)	**94,36** (0,31)	**94,35** (0,31)	92,54 (0,46)	92,56 (0,47)
		CHI	73 (1,29)	94,29 (0,28)	94,29 (0,28)	**94,52** (0,41)	**94,50** (0,41)
		CDM	845 (9,09)	**94,18** (0,30)	**94,18** (0,31)	92,62 (0,52)	92,65 (0,51)
		IG	58 (1,00)	94,00 (0,44)	93,99 (0,43)	**94,28** (0,28)	**94,28** (0,28)
		OR	268 (7,81)	**92,90** (0,33)	**92,89** (0,32)	91,03 (0,34)	91,10 (0,34)
	NB	BNS	284 (5,39)	**92,46** (0,25)	**92,54** (0,26)	90,51 (0,35)	90,60 (0,33)
		CHI	73 (1,29)	93,54 (0,37)	93,54 (0,35)	**93,61** (0,40)	**93,60** (0,38)
		CDM	845 (9,09)	**91,72** (0,40)	**91,81** (0,41)	90,14 (0,49)	90,25 (0,49)
		IG	58 (1,00)	**93,60** (0,31)	**93,57** (0,30)	93,58 (0,33)	93,56 (0,32)
		OR	268 (7,81)	**89,74** (0,32)	**89,75** (0,34)	87,53 (0,22)	87,66 (0,19)
Reuters 10	*k*-NN	BNS	141 (3,56)	**59,76** (0,67)	**80,31** (0,58)	55,58 (1,75)	77,15 (1,13)
		CHI	47 (1,73)	**56,72** (1,99)	75,81 (1,29)	56,63 (0,80)	**75,82** (0,58)
		CDM	409 (7,46)	**58,69** (1,48)	**80,63** (0,82)	57,76 (1,56)	79,64 (0,64)
		IG	37 (2,65)	53,48 (2,28)	**75,33** (1,63)	**53,58** (2,21)	75,03 (1,07)
		OR	109 (6,16)	**58,19** (1,29)	**78,89** (0,78)	56,38 (0,69)	78,13 (0,23)
	NB	BNS	141 (3,56)	**61,95** (1,92)	**79,23** (1,13)	58,04 (0,42)	77,42 (0,53)
		CHI	47 (1,73)	**59,28** (1,25)	**76,47** (0,78)	58,79 (1,77)	76,10 (0,93)
		CDM	409 (7,46)	**60,93** (1,51)	**80,19** (0,47)	59,24 (1,07)	80,16 (0,27)
		IG	37 (2,65)	53,69 (1,36)	**75,75** (0,98)	**53,87** (1,79)	75,28 (1,03)
		OR	109 (6,16)	**60,92** (1,40)	**78,49** (0,99)	60,30 (1,48)	77,85 (0,69)
WebKB	*k*-NN	BNS	284 (15,51)	**79,83** (0,64)	**82,19** (0,38)	77,72 (1,31)	80,47 (0,65)
		CHI	40 (2,77)	**76,56** (1,59)	**79,26** (1,24)	74,60 (0,69)	78,45 (0,66)
		CDM	936 (32,68)	**74,16** (1,64)	**78,02** (1,13)	60,96 (1,52)	66,87 (0,59)
		IG	40 (1,83)	**76,85** (1,17)	**79,59** (1,16)	74,88 (0,65)	78,57 (0,70)
		OR	313 (22,66)	**78,99** (1,46)	**81,26** (0,82)	77,55 (1,47)	80,40 (0,55)
	NB	BNS	284 (15,51)	**83,82** (1,02)	**85,62** (1,36)	81,65 (1,20)	84,14 (1,41)
		CHI	40 (2,77)	**77,11** (1,11)	**80,69** (0,76)	73,63 (1,31)	78,66 (1,27)
		CDM	936 (32,68)	**77,96** (1,76)	**82,52** (1,03)	65,60 (1,72)	74,37 (0,61)
		IG	40 (1,83)	**76,58** (0,96)	**80,38** (0,66)	74,34 (1,33)	79,42 (0,85)
		OR	313 (22,66)	**83,23** (0,97)	**85,12** (0,76)	80,53 (0,79)	83,10 (1,17)

APÊNDICE E

Tabelas de Resultados do método ALfFT-LR

Tabela E.1 Comparação dos métodos ALfFT-LR com $f = 1$ contra VR.

Base de Dados	Classificador	FEF	m	AL1FT-LR		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	21 (1,00)	**90,00** (0,40)	**90,24** (0,39)	55,46 (9,43)	61,03 (8,33)
		CHI	21 (1,41)	**82,11** (2,80)	**84,59** (2,39)	71,82 (4,76)	74,86 (4,38)
		CDM	1646 (13,75)	91,81 (0,31)	91,84 (0,31)	**94,17** (0,15)	**94,17** (0,15)
		IG	22 (2,00)	**86,83** (5,40)	**87,76** (4,18)	76,64 (1,92)	78,29 (1,82)
		OR	880 (33,00)	75,81 (2,37)	75,92 (2,33)	**93,22** (0,16)	**93,24** (0,15)
	NB	BNS	21 (1,00)	**90,34** (0,50)	**90,46** (0,48)	53,85 (9,04)	60,65 (7,52)
		CHI	21 (1,41)	**82,16** (2,85)	**83,92** (2,42)	73,34 (2,95)	75,53 (2,74)
		CDM	1646 (13,75)	90,89 (0,31)	90,98 (0,31)	**91,57** (0,30)	**91,66** (0,34)
		IG	22 (2,00)	**86,86** (5,58)	**87,57** (4,57)	77,10 (1,21)	78,48 (0,94)
		OR	880 (33,00)	77,96 (2,14)	77,76 (2,12)	**90,55** (0,35)	**90,58** (0,33)
Reuters 10	*k*-NN	BNS	64 (1,73)	**58,17** (1,36)	**76,70** (0,77)	50,27 (2,85)	74,01 (2,27)
		CHI	18 (0,82)	50,30 (6,31)	**71,34** (4,39)	**50,61** (4,01)	70,59 (2,65)
		CDM	635 (14,99)	**59,78** (1,36)	**81,81** (0,68)	58,57 (1,50)	80,56 (1,07)
		IG	11 (1,29)	**31,82** (4,85)	**62,15** (5,76)	28,30 (5,06)	60,09 (6,26)
		OR	287 (22,41)	59,03 (0,96)	78,38 (0,31)	**59,15** (1,65)	**80,41** (0,69)
	NB	BNS	64 (1,73)	**60,90** (0,65)	**78,24** (0,30)	50,13 (3,25)	74,17 (1,08)
		CHI	18 (0,82)	**51,90** (1,61)	**73,53** (1,34)	44,21 (2,20)	69,15 (0,66)
		CDM	635 (14,99)	**62,44** (1,41)	80,81 (0,73)	60,83 (1,26)	**80,86** (0,45)
		IG	11 (1,29)	**29,47** (3,88)	**62,73** (2,08)	23,39 (4,26)	60,73 (2,40)
		OR	287 (22,41)	60,26 (0,93)	79,08 (0,81)	**62,75** (1,08)	**79,46** (0,65)
WebKB	*k*-NN	BNS	91 (6,98)	**79,97** (1,07)	**81,40** (1,05)	66,04 (3,50)	70,50 (4,53)
		CHI	19 (0,58)	**74,40** (1,36)	**76,49** (1,43)	65,81 (0,63)	69,54 (0,31)
		CDM	461 (14,66)	**78,44** (0,51)	**81,28** (0,91)	53,08 (2,79)	60,90 (1,72)
		IG	18 (1,73)	**74,57** (1,35)	**77,31** (1,42)	65,74 (1,19)	69,58 (1,13)
		OR	214 (45,54)	71,14 (3,94)	75,64 (4,97)	**73,56** (0,94)	**76,64** (1,54)
	NB	BNS	91 (6,98)	**81,44** (1,99)	**83,78** (1,87)	69,67 (2,27)	77,43 (1,45)
		CHI	19 (0,58)	**74,51** (2,34)	**78,83** (1,47)	66,00 (0,91)	74,14 (0,59)
		CDM	461 (14,66)	**81,21** (0,94)	**83,16** (1,15)	58,63 (1,94)	70,16 (0,87)
		IG	18 (1,73)	**72,71** (0,91)	**78,14** (0,87)	66,57 (0,60)	74,47 (0,26)
		OR	214 (45,54)	74,84 (3,75)	80,21 (2,34)	**76,26** (3,19)	**80,43** (2,57)

Tabela E.2 Comparação dos métodos ALfFT-LR com $f = 2$ contra VR.

Base de Dados	Classificador	FEF	m	AL2FT-LR		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	36 (2,77)	**93,87** (0,29)	**93,89** (0,30)	69,60 (4,21)	72,09 (4,26)
		CHI	50 (2,08)	93,26 (0,64)	93,28 (0,63)	**94,07** (0,26)	**94,07** (0,26)
		CDM	2498 (9,13)	89,85 (0,47)	89,86 (0,47)	**93,88** (0,22)	**93,89** (0,22)
		IG	37 (2,08)	93,70 (0,51)	93,71 (0,53)	**94,01** (0,29)	**94,01** (0,27)
		OR	1745 (37,03)	85,50 (2,50)	85,58 (2,50)	**92,72** (0,31)	**92,74** (0,32)
	NB	BNS	36 (2,77)	**93,51** (0,23)	**93,56** (0,24)	68,30 (2,27)	72,17 (1,86)
		CHI	50 (2,08)	92,10 (0,76)	92,28 (0,64)	**93,08** (0,37)	**93,12** (0,37)
		CDM	2498 (9,13)	89,67 (0,23)	89,75 (0,21)	**91,63** (0,38)	**91,73** (0,42)
		IG	37 (2,08)	**93,21** (0,67)	**93,25** (0,66)	92,43 (0,67)	92,46 (0,63)
		OR	1745 (37,03)	85,54 (1,99)	85,69 (2,02)	**89,99** (0,22)	**90,08** (0,20)
Reuters 10	*k*-NN	BNS	152 (8,54)	**59,34** (1,30)	**78,69** (0,88)	55,70 (1,72)	77,39 (1,16)
		CHI	39 (1,63)	**57,29** (1,18)	**75,86** (1,03)	55,58 (1,46)	74,91 (1,18)
		CDM	992 (13,44)	**60,55** (1,32)	**82,35** (0,73)	58,95 (1,83)	80,87 (0,81)
		IG	20 (2,16)	42,39 (4,50)	69,18 (3,25)	**48,19** (4,44)	**71,25** (2,07)
		OR	688 (29,27)	**59,80** (0,85)	80,28 (0,60)	58,98 (2,09)	**80,53** (1,30)
	NB	BNS	152 (8,54)	**62,96** (1,31)	**78,87** (0,89)	59,15 (0,70)	77,94 (0,39)
		CHI	39 (1,63)	57,19 (0,55)	**76,77** (0,89)	**57,23** (1,94)	75,50 (0,96)
		CDM	992 (13,44)	**62,18** (0,92)	80,82 (0,63)	61,40 (0,92)	**80,98** (0,49)
		IG	20 (2,16)	40,22 (5,04)	67,88 (3,27)	**44,58** (3,64)	**71,04** (1,79)
		OR	688 (29,27)	60,92 (0,65)	**80,05** (0,34)	**61,34** (0,86)	79,07 (0,54)
WebKB	*k*-NN	BNS	179 (17,18)	**81,19** (1,30)	**82,66** (0,71)	74,75 (1,43)	78,40 (1,50)
		CHI	30 (1,63)	**77,52** (1,36)	**79,61** (1,63)	71,44 (1,93)	75,90 (1,27)
		CDM	809 (26,77)	**79,82** (1,09)	**83,59** (0,19)	59,93 (1,71)	66,42 (1,01)
		IG	29 (2,71)	**77,61** (2,04)	**79,95** (1,98)	71,63 (2,93)	76,16 (2,34)
		OR	449 (54,22)	76,01 (1,99)	79,49 (2,20)	**79,75** (1,17)	**82,28** (1,40)
	NB	BNS	179 (17,18)	**83,73** (1,08)	**85,36** (1,53)	78,56 (1,59)	82,62 (1,36)
		CHI	30 (1,63)	**77,98** (1,08)	**80,59** (0,88)	70,68 (1,50)	77,40 (1,46)
		CDM	809 (26,77)	**84,13** (0,36)	**85,24** (0,66)	64,51 (2,01)	73,71 (1,01)
		IG	29 (2,71)	**77,31** (2,39)	**80,52** (1,71)	71,15 (1,37)	77,80 (0,93)
		OR	449 (54,22)	77,99 (0,93)	81,85 (0,88)	**82,43** (1,28)	**84,26** (1,21)

Tabela E.3 Comparação dos métodos ALfFT-LR com $f = 3$ contra VR.

Base de Dados	Classificador	FEF	m	AL3FT-LR		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	142 (4,58)	**94,45** (0,24)	**94,48** (0,22)	87,94 (1,29)	88,08 (1,22)
		CHI	82 (3,16)	94,07 (0,37)	94,08 (0,37)	**94,43** (0,50)	**94,43** (0,48)
		CDM	3063 (17,11)	88,30 (0,76)	88,29 (0,75)	**93,39** (0,38)	**93,40** (0,38)
		IG	57 (2,52)	94,00 (0,69)	94,01 (0,69)	**94,31** (0,28)	**94,29** (0,27)
		OR	2643 (61,56)	87,90 (2,19)	87,99 (2,19)	**92,02** (0,51)	**92,04** (0,53)
	NB	BNS	142 (4,58)	**93,98** (0,42)	**93,96** (0,42)	86,47 (0,44)	86,73 (0,39)
		CHI	82 (3,16)	92,57 (0,46)	92,65 (0,45)	**93,57** (0,28)	**93,57** (0,26)
		CDM	3063 (17,11)	88,71 (0,23)	88,82 (0,26)	**91,29** (0,30)	**91,39** (0,33)
		IG	57 (2,52)	93,14 (1,04)	93,19 (1,01)	**93,58** (0,34)	**93,56** (0,32)
		OR	2643 (61,56)	88,27 (1,44)	88,39 (1,41)	**89,32** (0,24)	**89,42** (0,26)
Reuters 10	*k*-NN	BNS	264 (9,18)	**60,18** (1,62)	**80,69** (0,71)	59,61 (0,76)	80,46 (0,41)
		CHI	61 (1,53)	57,24 (1,73)	**77,57** (1,08)	**57,33** (1,63)	76,70 (0,99)
		CDM	1235 (10,38)	**60,65** (1,63)	**82,40** (0,63)	59,97 (2,22)	81,64 (1,02)
		IG	40 (2,52)	49,27 (2,36)	73,88 (0,76)	**53,69** (1,29)	**75,23** (1,15)
		OR	1085 (21,70)	59,54 (1,65)	80,53 (0,71)	**60,46** (1,06)	**81,88** (0,37)
	NB	BNS	264 (9,18)	**64,12** (0,49)	**80,42** (0,77)	62,39 (1,21)	79,84 (0,61)
		CHI	61 (1,53)	58,81 (1,52)	**78,19** (0,76)	**60,68** (1,66)	77,21 (0,78)
		CDM	1235 (10,38)	61,92 (0,84)	80,73 (0,58)	**62,44** (1,46)	**81,37** (0,69)
		IG	40 (2,52)	47,92 (3,86)	73,33 (1,85)	**54,77** (1,01)	**75,51** (0,41)
		OR	1085 (21,70)	59,95 (0,77)	**80,10** (0,71)	**60,81** (1,04)	79,30 (0,58)
WebKB	*k*-NN	BNS	268 (20,62)	**82,32** (0,48)	**83,81** (0,61)	77,59 (1,03)	80,26 (0,75)
		CHI	43 (2,24)	**79,65** (1,33)	**81,59** (1,22)	74,90 (1,05)	78,79 (0,95)
		CDM	1113 (21,32)	**80,54** (1,79)	**84,90** (1,39)	63,93 (1,72)	69,28 (1,03)
		IG	41 (1,91)	**78,58** (0,92)	**80,88** (0,90)	75,04 (0,73)	78,97 (0,61)
		OR	667 (69,99)	76,53 (2,17)	80,07 (1,68)	**79,68** (0,48)	**82,33** (0,88)
	NB	BNS	268 (20,62)	**84,38** (1,10)	**85,83** (1,44)	81,58 (1,30)	84,14 (1,37)
		CHI	43 (2,24)	**80,24** (1,09)	**82,12** (1,07)	73,86 (1,24)	78,90 (1,29)
		CDM	1113 (21,32)	**85,16** (0,66)	**86,04** (0,94)	69,20 (2,50)	76,85 (1,09)
		IG	41 (1,91)	**80,01** (1,70)	**82,09** (1,46)	74,05 (0,91)	79,14 (0,89)
		OR	667 (69,99)	80,30 (0,84)	83,50 (0,67)	**84,42** (1,61)	**85,64** (1,50)

Tabela E.4 Comparação dos métodos ALfFT-LR com $f = 4$ contra VR.

Base de Dados	Classificador	FEF	m	AL4FT-LR		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	421 (8,43)	**94,45** (0,23)	**94,44** (0,23)	92,77 (0,41)	92,79 (0,40)
		CHI	155 (3,37)	**94,31** (0,35)	**94,30** (0,34)	94,23 (0,19)	94,23 (0,19)
		CDM	3469 (9,31)	87,40 (0,71)	87,41 (0,72)	**93,30** (0,19)	**93,31** (0,19)
		IG	137 (3,87)	**94,20** (0,72)	**94,19** (0,71)	93,57 (0,22)	93,57 (0,24)
		OR	3648 (87,86)	**90,62** (1,72)	**90,66** (1,71)	90,34 (0,49)	90,36 (0,49)
	NB	BNS	421 (8,43)	**93,74** (0,31)	**93,75** (0,30)	90,40 (0,87)	90,50 (0,85)
		CHI	155 (3,37)	92,66 (0,50)	92,71 (0,49)	**93,44** (0,21)	**93,46** (0,20)
		CDM	3469 (9,31)	87,77 (0,21)	87,88 (0,23)	**91,17** (0,35)	**91,28** (0,38)
		IG	137 (3,87)	92,88 (0,84)	92,93 (0,81)	**93,00** (0,51)	**93,05** (0,51)
		OR	3648 (87,86)	**90,94** (0,99)	**91,01** (0,97)	88,24 (0,31)	88,35 (0,33)
Reuters 10	*k*-NN	BNS	392 (17,22)	59,63 (1,80)	**80,86** (0,88)	**59,97** (1,48)	80,64 (0,68)
		CHI	95 (2,45)	58,29 (2,44)	79,11 (1,18)	**58,88** (1,50)	**79,31** (0,63)
		CDM	1421 (13,39)	**60,55** (1,16)	**82,19** (0,48)	60,48 (2,21)	81,91 (0,95)
		IG	63 (5,10)	51,63 (2,92)	76,18 (1,62)	**56,60** (1,90)	**77,72** (0,68)
		OR	1460 (25,90)	60,21 (1,62)	81,04 (0,89)	**60,59** (1,48)	**81,82** (0,79)
	NB	BNS	392 (17,22)	**63,58** (0,91)	**80,49** (0,71)	62,84 (0,55)	80,08 (0,18)
		CHI	95 (2,45)	61,20 (1,75)	**79,08** (1,20)	**62,93** (0,38)	78,98 (0,67)
		CDM	1421 (13,39)	61,82 (0,88)	80,66 (0,50)	**62,19** (1,55)	**81,40** (0,81)
		IG	63 (5,10)	51,37 (4,40)	75,43 (2,40)	**59,92** (1,03)	**77,82** (0,77)
		OR	1460 (25,90)	59,71 (1,63)	**80,09** (0,80)	**60,97** (1,16)	79,77 (0,78)
WebKB	*k*-NN	BNS	364 (25,84)	**82,86** (0,49)	**84,47** (0,69)	77,75 (1,79)	80,38 (1,28)
		CHI	55 (3,00)	**80,64** (0,87)	**82,54** (0,96)	78,33 (1,03)	81,04 (1,02)
		CDM	1372 (35,91)	**79,88** (1,58)	**84,76** (1,32)	66,62 (1,60)	70,94 (0,78)
		IG	51 (2,38)	**79,87** (2,67)	**82,09** (1,89)	78,08 (1,46)	80,61 (1,41)
		OR	864 (93,76)	78,33 (2,01)	81,64 (1,19)	**80,47** (0,76)	**83,50** (0,66)
	NB	BNS	364 (25,84)	**85,26** (0,23)	**86,47** (0,73)	82,22 (1,46)	84,38 (1,40)
		CHI	55 (3,00)	**81,79** (0,88)	**83,14** (0,90)	79,84 (1,76)	82,21 (1,26)
		CDM	1372 (35,91)	**85,78** (0,86)	**86,48** (1,04)	72,62 (1,29)	78,68 (1,01)
		IG	51 (2,38)	**80,75** (1,28)	**83,00** (1,02)	79,07 (0,70)	81,97 (0,87)
		OR	864 (93,76)	81,72 (1,05)	84,28 (0,48)	**84,92** (1,34)	**85,95** (1,26)

Tabela E.5 Comparação dos métodos ALfFT-LR com $f = 5$ contra VR.

Base de Dados	Classificador	FEF	m	AL5FT-LR		VR	
				macro-F1	micro-F1	macro-F1	micro-F1
20 Newsgroup	*k*-NN	BNS	804 (9,18)	**94,45** (0,14)	**94,44** (0,15)	93,38 (0,35)	93,40 (0,36)
		CHI	288 (2,71)	**94,28** (0,58)	**94,28** (0,57)	94,12 (0,25)	94,12 (0,25)
		CDM	3789 (18,38)	86,86 (0,50)	86,85 (0,49)	**93,08** (0,16)	**93,09** (0,18)
		IG	226 (2,45)	**94,11** (0,86)	**94,11** (0,84)	93,09 (0,36)	93,09 (0,36)
		OR	4719 (113,26)	**91,53** (0,57)	**91,58** (0,58)	89,39 (0,60)	89,40 (0,61)
	NB	BNS	804 (9,18)	**93,83** (0,34)	**93,85** (0,34)	90,97 (0,51)	91,07 (0,53)
		CHI	288 (2,71)	92,40 (0,35)	92,46 (0,35)	**92,72** (0,33)	**92,75** (0,35)
		CDM	3789 (18,38)	86,90 (0,13)	87,01 (0,13)	**91,03** (0,32)	**91,13** (0,34)
		IG	226 (2,45)	**92,77** (0,84)	**92,82** (0,80)	91,29 (0,32)	91,39 (0,36)
		OR	4719 (113,26)	**91,47** (0,40)	**91,53** (0,40)	86,63 (0,26)	86,76 (0,26)
Reuters 10	*k*-NN	BNS	546 (24,10)	60,27 (1,53)	**81,47** (0,91)	**60,52** (1,60)	81,02 (0,83)
		CHI	134 (2,38)	58,96 (1,09)	80,19 (0,75)	**59,21** (1,24)	**80,45** (0,65)
		CDM	1562 (18,87)	59,81 (1,57)	**81,86** (0,63)	**60,16** (2,02)	81,81 (0,77)
		IG	91 (5,00)	53,75 (1,57)	77,83 (0,70)	**57,95** (1,92)	**79,64** (0,92)
		OR	1826 (36,64)	59,77 (1,55)	81,23 (0,76)	**60,60** (0,77)	**82,08** (0,51)
	NB	BNS	546 (24,10)	**63,29** (1,05)	**80,74** (0,78)	62,99 (0,94)	80,38 (0,37)
		CHI	134 (2,38)	62,59 (0,41)	**79,51** (0,56)	**63,00** (1,04)	79,30 (0,87)
		CDM	1562 (18,87)	61,85 (0,80)	80,68 (0,49)	**61,92** (1,23)	**81,30** (0,61)
		IG	91 (5,00)	54,14 (1,49)	76,29 (1,14)	**61,80** (1,41)	**78,90** (1,02)
		OR	1826 (36,64)	59,80 (1,41)	**80,24** (0,56)	**60,51** (0,75)	79,82 (0,61)
WebKB	*k*-NN	BNS	475 (36,95)	**82,52** (0,64)	**84,52** (0,49)	78,33 (1,48)	80,66 (1,31)
		CHI	65 (1,15)	**81,25** (0,68)	**83,31** (0,83)	80,42 (0,32)	82,57 (0,21)
		CDM	1611 (35,60)	**80,07** (1,69)	**85,07** (1,10)	67,08 (1,48)	71,33 (1,00)
		IG	59 (2,45)	**80,08** (1,08)	**82,43** (1,25)	80,00 (2,36)	82,30 (1,83)
		OR	1077 (97,12)	79,28 (2,15)	82,04 (1,59)	**80,65** (0,80)	**83,90** (1,81)
	NB	BNS	475 (36,95)	**85,95** (0,43)	**87,00** (0,40)	83,01 (1,32)	85,07 (1,41)
		CHI	65 (1,15)	**82,54** (0,67)	**83,90** (0,69)	81,55 (1,28)	83,28 (1,20)
		CDM	1611 (35,60)	**85,42** (0,46)	**86,16** (0,78)	71,74 (1,68)	78,26 (1,07)
		IG	59 (2,45)	**81,80** (1,39)	**83,66** (0,98)	80,43 (0,55)	82,76 (0,80)
		OR	1077 (97,12)	83,44 (0,78)	85,35 (0,86)	**85,52** (1,38)	**86,35** (1,26)

Referências Bibliográficas

ALMUALLIM, H.; DIETTERICH, T. Learning with many irrelevant features. *Proceedings of the National Conference of Artificial Intelligence*, Oregon State University Corvallis, OR, USA, p. 547–552, 1991.

ALMUALLIM, H.; DIETTERICH, T. Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, Elsevier, v. 69, n. 1-2, p. 279–305, 1994. ISSN 0004-3702.

APTE, C.; DAMERAU, F.; WEISS, S. Text mining with decision trees and decision rules. In: *Proceedings of Conference on Automated Learning and Discovery, Workshop 6: Learning from Text and the Web*. [S.l.]: Carnegie-Mellon University, 1998.

ARAUZO, A.; BENITEZ, J.; CASTRO, J. C-focus: a continuous extension of focus. *Advances in Soft Computing: Engineering Design and Manufacturing*, Springer Verlag, p. 225–232, 2003.

BAKER, L.; MCCALLUM, A. Distributional clustering of words for text classification. In: ACM. *Proceedings of Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR)*. [S.l.], 1998. p. 96–103. ISBN 1581130155.

BEKKERMAN, R. *Distributional clustering of words for text categorization*. Dissertação (Mestrado) — Israel Institute of Technology, 2003.

BENSCH, M. et al. Feature selection for high-dimensional industrial data. In: *Proceedings of European Symposium on Artificial Neural Networks (ESANN)*. [S.l.: s.n.], 2005. p. 375–380.

CAROPRESO, M.; MATWIN, S.; SEBASTIANI, F. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *Text Databases and Document Management: Theory and Practice*, IGI Publishing, p. 78–102, 2001.

CHANG, Y.; CHEN, S.; LIAU, C. Multilabel text categorization based on a new linear classifier learning method and a category-sensitive refinement method. *Expert Systems with Applications*, Elsevier, v. 34, n. 3, p. 1948–1953, 2008. ISSN 0957-4174.

CHEN, J. et al. Feature selection for text classification with naive bayes. *Expert Systems with Applications*, Elsevier, v. 36, n. 3, p. 5432–5435, 2009.

CORRÊA, R.; LUDERMIR, T. Improving self-organization of document collections by semantic mapping. *Neurocomputing*, Elsevier, v. 70, n. 1-3, p. 62–69, 2006.

CORRÊA, R.; LUDERMIR, T. Dimensionality reduction of very large document collections by semantic mapping. In: *Proceedings of Workshop on Self-Organizing Maps (WSOM)*. [S.l.: s.n.], 2007.

COVER, T.; HART, P. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, IEEE, v. 13, n. 1, p. 21–27, 2002. ISSN 0018-9448.

COVER, T.; THOMAS, J.; MYILIBRARY. *Elements of information theory*. [S.l.]: Wiley Online Library, 1991.

CRAVEN, M.; DIPASQUO, D.; FREITAG, D. *Learning to extract symbolic knowledge from the World Wide Web*. [S.l.]: AAAI Press, 1998.

DEBOLE, F.; SEBASTIANI, F. Supervised term weighting for automated text categorization. In: ACM. *Proceedings of the 2003 ACM symposium on Applied computing*. [S.l.], 2003. p. 784–788.

DUMAIS, S. et al. Inductive learning algorithms and representations for text categorization. In: ACM. *Proceedings of International Conference on Information and Knowledge Management (CIKM)*. [S.l.], 1998. p. 148–155. ISBN 1581130619.

FELLBAUM, C. *WordNet: An Electronic Lexical Database*. [S.l.]: The MIT Press, 1998.

FORMAN, G. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, JMLR. org, v. 3, p. 1289–1305, 2003.

FORMAN, G. Feature selection for text classification. In: *Computational Methods of Feature Selection*. [S.l.]: Chapman and Hall/CRC Press, 2007.

FüRNKRANZ, J. A study using n-gram features for text categorization. *Austrian Research Institute for Artificial Intelligence Technical Report OEFAI-TR-98-30 Schottengasse*, v. 3, 1998.

GODBOLE, S.; SARAWAGI, S.; CHAKRABARTI, S. Scaling multi-class support vector machines using inter-class confusion. In: ACM. *Proceedings of International Conference on Knowledge Discovery and Data mining (ACM SIGKDD)*. [S.l.], 2002. p. 513–518.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, JMLR. org, v. 3, p. 1157–1182, 2003. ISSN 1532-4435.

KAGEURA, K.; UMINO, B. Methods of automatic term recognition: A review. *Terminology*, John Benjamins Publishing Company, v. 3, n. 2, p. 259–289, 1996. ISSN 0929-9971.

KIRA, K.; RENDELL, L. The feature selection problem: Traditional methods and a new algorithm. In: JOHN WILEY & SONS LTD. *Proceedings of the National Conference on Artificial Intelligence*. [S.l.], 1992. p. 129–129.

KOHAVI, R.; JOHN, G. Wrappers for feature subset selection. *Artificial intelligence*, Elsevier, v. 97, n. 1-2, p. 273–324, 1997.

KONONENKO, I. Estimating attributes: Analysis and extensions of relief. In: SPRINGER. *Proceedings of European Conference on Machine Learning (ECML)*. [S.l.], 1994. p. 171–182.

LAM, W.; HO, C. Using a generalized instance set for automatic text categorization. In: ACM. *Proceedings of Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR)*. [S.l.], 1998. p. 81–89.

LANG, K. Newsweeder: Learning to filter netnews. In: *Proceedings of International Machine Learning Conference (ICML)*. [S.l.]: Morgan Kaufmann, 1995. p. 331–339.

LEE, K.; KAGEURA, K. Virtual relevant documents in text categorization with support vector machines. *Information Processing & Management*, Elsevier, v. 43, n. 4, p. 902–913, 2007.

LEWIS, D. An evaluation of phrasal and clustered representations on a text categorization task. In: ACM. *Proceedings of Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR)*. [S.l.], 1992. p. 37–50. ISBN 0897915232.

LEWIS, D. *Representation and learning in information retrieval*. Tese (Doutorado) — University of Massachusetts, 1992.

LEWIS, D. Naive (bayes) at forty: The independence assumption in information retrieval. *Proceedings of European Conference on Machine Learning (ECML)*, Springer, p. 4–15, 1998.

LEWIS, D.; RINGUETTE, M. A comparison of two learning algorithms for text categorization. In: *Proceedings of Symposium on Document Analysis and Information Retrieval (SDAIR)*. [S.l.: s.n.], 1994. v. 33, p. 81–93.

LI, Y.; DONG, M.; HUA, J. Localized feature selection for clustering. *Pattern Recognition Letters*, Elsevier, v. 29, n. 1, p. 10–18, 2008. ISSN 0167-8655.

LODHI, H. et al. Text classification using string kernels. *The Journal of Machine Learning Research*, JMLR. org, v. 2, p. 419–444, 2002. ISSN 1532-4435.

LOVINS, J. *Development of a Stemming Algorithm*. [S.l.], 1968.

MANNING, C.; RAGHAAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. [S.l.]: Cambridge University Press, 2008.

MASAND, B.; LINOFF, G.; WALTZ, D. Classifying news stories using memory based reasoning. In: ACM. *Proceedings of Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR)*. [S.l.], 1992. p. 59–65. ISBN 0897915232.

MAYFIELD, J.; MCNAMEE, P. Single n-gram stemming. In: ACM. *Proceedings of Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR)*. [S.l.], 2003. p. 415–416. ISBN 1581136463.

MCCALLUM, A.; NIGAM, K. A comparison of event models for naive bayes text classification. In: AAAI PRESS. *Proceedings of Workshop on Learning for Text Categorization*. [S.l.], 1998. p. 41–48.

MILLER, G. Wordnet: an on-line lexical database. *International Journal of Lexicography*, v. 4, n. 3, 1990.

MLADENIC, D. Feature subset selection in text-learning. *Proceedings of European Conference on Machine Learning (ECML)*, Springer, p. 95–100, 1998.

MLADENIC, D.; GROBELNIK, M. Word sequences as features in text-learning. In: *Proceedings of Electrotechnical and Computer Science Conference (ERK)*. [S.l.: s.n.], 1998. p. 145–148.

MLADENIC, D.; GROBELNIK, M. Feature selection for unbalanced class distribution and naive bayes. In: MORGAN KAUFMANN PUBLISHERS INC. *Proceedings of International Conference on Machine Learning (ICML)*. [S.l.], 1999. p. 258–267.

NIGAM, K. et al. Learning to classify text from labeled and unlabeled documents. In: *Proceedings of the National Conference on Artificial Intelligence*. [S.l.]: AAAI Press, 1998. p. 792–799.

OGURA, H.; AMANO, H.; KONDO, M. Feature selection with a measure of deviations from poisson in text categorization. *Expert Systems with Applications*, Elsevier, v. 36, n. 3, p. 6826–6832, 2009. ISSN 0957-4174.

ÖZGÜR, L.; GÜNGÖR, T. Text classification with the support of pruned dependency patterns. *Pattern Recognition Letters*, Elsevier, v. 31, n. 12, p. 1598–1607, 2010. ISSN 0167-8655.

PAICE, C. Another stemmer. In: ACM. *ACM Sigir Forum*. [S.l.], 1990. v. 24, n. 3, p. 56–61. ISSN 0163-5840.

PARK, H.; KWON, S.; KWON, H. Complete gini-index text (git) feature-selection algorithm for text classification. In: IEEE. *Proceedings of International Conference on Software Engineering and Data Mining (SEDM)*. [S.l.]. p. 366–371.

PENG, H.; LONG, F.; DING, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, p. 1226–1238, 2005.

PORTER, M. An algorithm for suffix stripping. *Program: electronic library and information systems*, Emerald Group Publishing Limited, v. 40, n. 3, p. 211–218, 2006. ISSN 0033-0337.

RIJSBERGEN, V. *Information Retrieval*. [S.l.]: London: Butterworth, 1979.

ROGATI, M.; YANG, Y. High-performing feature selection for text classification. In: ACM. *Proceedings of International Conference on Information and Knowledge Management (CIKM)*. [S.l.], 2002. p. 659–661.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information processing & management*, Elsevier, v. 24, n. 5, p. 513–523, 1988. ISSN 0306-4573.

SALTON, G.; MCGILL, M. Introduction to modern information retrieval. *New York*, 1983.

SCHAPIRE, R.; SINGER, Y. Boostexter: A boosting-based system for text categorization. *Machine learning*, Springer, v. 39, n. 2, p. 135–168, 2000. ISSN 0885-6125.

SCHÜTZE, H.; HULL, D.; PEDERSEN, J. A comparison of classifiers and document representations for the routing problem. In: ACM. *Proceedings of Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR)*. [S.l.], 1995. p. 229–237. ISBN 0897917146.

SCOTT, S.; MATWIN, S. Feature engineering for text classification. In: *Proceedings of International Conference on Machine Learning (ICML)*. [S.l.: s.n.], 1999. p. 379–388.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, ACM, v. 34, n. 1, p. 1–47, 2002. ISSN 0360-0300.

SHANG, W. et al. A novel feature selection algorithm for text categorization. *Expert Systems with Applications*, Elsevier, v. 33, n. 1, p. 1–5, 2007. ISSN 0957-4174.

SLONIM, N.; TISHBY, N. The power of word clusters for text classification. In: *Proceedings of European Colloquium on Information Retrieval Research (ECIR)*. [S.l.: s.n.], 2001.

SOUZA, A. D. et al. Automated multi-label text categorization with vg-ram weightless neural networks. *Neurocomputing*, Elsevier, v. 72, n. 10-12, p. 2209–2217, 2009.

TAN, S. An effective refinement strategy for knn text classifier. *Expert Systems with Applications*, Elsevier, v. 30, n. 2, p. 290–298, 2006.

TANG, B. et al. Comparing and combining dimension reduction techniques for efficient text clustering. *Feature Selection for Data Mining*, p. 17–26, 2005.

TZERAS, K.; HARTMANN, S. Automatic indexing based on bayesian inference networks. In: ACM. *Proceedings of Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR)*. [S.l.], 1993. p. 22–35. ISBN 0897916050.

VIERA, A.; VIRGIL, J. Uma revisão dos algoritmos de radicalização em língua portuguesa. *Information Research*, v. 12, n. 3, 2007.

WANG, P.; DOMENICONI, C. Building semantic kernels for text classification using wikipedia. In: ACM. *Proceedings of International Conference on Knowledge Discovery and Data mining (ACM SIGKDD)*. [S.l.], 2008. p. 713–721.

WANG, X. et al. Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters*, Elsevier, v. 28, n. 4, p. 459–471, 2007.

WANG, Y. et al. Feature selection using tabu search with long-term memories and probabilistic neural networks. *Pattern Recognition Letters*, Elsevier, v. 30, n. 7, p. 661–670, 2009.

WEIGEND, A.; WIENER, E.; PEDERSEN, J. Exploiting hierarchy in text categorization. *Information Retrieval*, Springer, v. 1, n. 3, p. 193–216, 1999. ISSN 1386-4564.

WIENER, E.; PEDERSEN, J.; WEIGEND, A. A neural network approach to topic spotting. In: LAS VEGAS, NV, USA: UNIV. OF NEVADA. *Proceedings of Annual Symposium on Document Analysis and Information Retrieval (SDAIR)*. [S.l.], 1995. v. 332, p. 317–332.

WILBUR, W.; SIROTKIN, K. The automatic identification of stop words. *Journal of information science*, Sage Publications, v. 18, n. 1, p. 45–55, 1992. ISSN 0165-5515.

XUE, X.; ZHOU, Z. Distributional features for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, v. 21, n. 3, p. 428–442, 2009.

YANG, Y. Noise reduction in a statistical approach to text categorization. In: ACM. *Proceedings of Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR)*. [S.l.], 1995. p. 256–263. ISBN 0897917146.

YANG, Y. An evaluation of statistical approaches to text categorization. *Information retrieval*, Springer, v. 1, n. 1, p. 69–90, 1999. ISSN 1386-4564.

YANG, Y.; PEDERSEN, J. A comparative study on feature selection in text categorization. In: *Proceedings of International Conference on Machine Learning (ICML)*. [S.l.]: Morgan Kaufmann Publishers Inc., 1997. p. 412–420.

YANG, Y.; WILBUR, J. Using corpus statistics to remove redundant words in text categorization. *Journal of the American Society for Information Science*, v. 47, n. 5, p. 357–369, 1996. ISSN 0002-8231.

YAO, Y. Information-theoretic measures for knowledge discovery and data mining. *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, p. 115–136, 2003.

YU, L.; LIU, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: *Proceedings of International Conference on Machine Learning (ICML)*. [S.l.: s.n.], 2003. v. 20, n. 2, p. 856–863.

ZADROZNY, S.; KACPRZYK, J. Computing with words for text processing: an approach to the text categorization. *Information Sciences*, Elsevier, v. 176, n. 4, p. 415–437, 2006. ISSN 0020-0255.