UMA METODOLOGIA DE CATEGORIZAÇÃO AUTOMÁTICA DE TEXTOS PARA A DISTRIBUIÇÃO DOS PROJETOS DE LEI ÀS COMISSÕES PERMANENTES DA CÂMARA LEGISLATIVA DO DISTRITO FEDERAL

Liliam Ayako Matsunaga

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA CIVIL.

Aprovada por:

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

Prof. Alexandre Gonçalves Evsukoff, Dr.

Prof. Beatriz de Souza Leite Pires de Lima, D.Sc.

Prof. Myrian Christina de Aragão Costa, D.Sc.

Prof. Antonio César Ferreira Guimarães, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
JUNHO DE 2007

MATSUNAGA, LILIAM AYAKO

Uma Metodologia de Categorização Automática de Textos para a Distribuição dos Projetos de Lei às Comissões Permanentes da Câmara Legislativa do Distrito Federal [Rio de Janeiro] 2007

XV, 157 p. 29,7 cm (COPPE/UFRJ, D. Sc., Engenharia Civil, 2007)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1. Categorização Automática de Textos
2. Classificação de Textos
3. Text Categorization
4. Text Classification

I. COPPE/UFRJ II. Título (série)

Agradecimentos

À Câmara Legislativa do Distrito Federal, por ter me liberado para fazer este doutorado.

Ao meu orientador Nelson, pelo apoio e pela ajuda em todos os momentos que precisei.

Ao Prof. Alexandre Gonçalves Evsukoff, pelas sugestões feitas no exame de qualificação.

Ao meu ex-chefe Abner Pereira Dutra, pela confiança no meu trabalho, pelo apoio e pela forma compreensiva e cortês com que sempre me tratou.

Aos amigos de trabalho Ângela Maria Vilas Boas Ribeiro, João Dino F. dos Santos, Ney Barros Luz e Marisa Peroni, pela ajuda nos momentos de necessidade. Em especial, agradeço à Áurea Helena Orlandi, pela ajuda na identificação da aplicação.

A Guilherme Saad Terra e Fábio Teodoro de Souza do NTT/COPPE/UFRJ, por serem tão prestativos e pela disposição em me ajudar.

A todos os amigos e familiares que me apoiaram e torceram por mim. Em especial, agradeço: à Márcia Shizuko Matsunaga Mizuno - minha irmã -, que mesmo superocupada tirou um tempo para me ajudar; à Terezinha Xavier, por ter me apoiado nos momentos difíceis; a Paulo Malheiro da Rocha, pela ajuda quando precisei; e a Guilherme Leal, por ter me apoiado, me incentivado e me acompanhando a partir do dia em que nos re-encontramos na UFRJ, depois de quinze anos sem nos vermos desde os tempos da USP.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

UMA METODOLOGIA DE CATEGORIZAÇÃO AUTOMÁTICA DE TEXTOS PARA A DISTRIBUIÇÃO DOS PROJETOS DE LEI ÀS COMISSÕES PERMANENTES DA CÂMARA LEGISLATIVA DO DISTRITO FEDERAL

Liliam Ayako Matsunaga

Junho/2007

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

Neste trabalho é proposta uma metodologia de categorização automática de textos para a obtenção de um modelo que indique de forma automática as comissões permanentes da Câmara Legislativa do Distrito Federal que devem apreciar cada um dos projetos de lei apresentados à Casa. Usando o algoritmo *Support Vector Machines* com várias formas de atribuição de pesos aos termos, foram estudadas as abordagens por dicionário global e local juntamente com seleção de termos e aumento de peso para os termos presentes nas ementas e para os relacionados às matérias de competência das comissões permanentes.

Duas novas formas de atribuição de pesos aos termos também foram propostas neste trabalho: TF_ABSL e TF_BNS. Elas incluem no cálculo dos pesos para os termos a importância desses para a discriminação das categorias, medidas pelas métricas abs-logito (ABSL) - proposta neste trabalho - e *bi-normal separation* (BNS).

Os resultados obtidos confirmaram a viabilidade prática da proposta, com as melhores soluções produzidas pelas duas formas de atribuição de pesos propostas, combinadas com seleção de termos e aumento de peso para alguns termos.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

AN AUTOMATED TEXT CATEGORIZATION METHODOLOGY TO DISTRIBUTE THE LAW PROJECTS TO THE PERMANENT COMMITTEE AT THE FEDERAL DISTRICT LEGISLATIVE ASSEMBLY

Liliam Ayako Matsunaga

June/2007

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

This work presents an approach for text categorization and builds a computer system that automatically indicates the permanent committees at the Federal District Legislative Assembly that are adequate to examine the proposed law projects. Support vector machines algorithm was implemented with a number of term weighting methods using a global and local dictionary, along with term selection and increase of weights for important existing words and for words related to the domain of the permanent committee jurisdiction.

Two new term weighting methods were also considered in this work: TF_ABSL and TF_BNS. They include in the term weight calculations, the term importance to the discrimination of the categories, measured by abs-logit (ABSL) – proposed in this work - and bi-normal separation (BNS).

The results obtained have confirmed the performance of the proposed strategy, and the best solutions were produced by the two methods of term weighting suggested in this work, combined with term selection and increase of weight for some terms.

Sumário

Lista de siglas e nomenclaturas

CÂMARA LEGISLATIVA DO DISTRITO FEDERAL

CLDF	Câmara Legislativa do Distrito Federal
PL	Projeto de Lei

COMISSÕES PERMANENTES DA CLDF

CAF	Comissão de Assuntos Fundiários
CAS	Comissão de Assuntos Sociais
CCJ	Comissão de Constituição e Justiça
CDC	Comissão de Defesa do Consumidor
CDDHCEDP	Comissão de Defesa dos Direitos Humanos, Cidadania, Ética e Decoro Parlamentar
CDESCTMAT	Comissão de Desenvolvimento Econômico Sustentável, Ciência, Tecnologia, Meio Ambiente e Turismo
CEOF	Comissão de Economia, Orçamento e Finanças
CES	Comissão de Educação e Saúde
CSEG	Comissão de Segurança

MÉTODOS DE ATRIBUIÇÀO DE PESOS AOS TERMOS (subseção 2.3.2)

TF_ABSL	$TF\_ABSL(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij}) ABSL(t_j, c_k)$
TF_BNS	$TF\_BNS(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij}) BNS(t_j, c_k)$
TF_GR	$TF\_GR(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij}) GR(t_j, c_k)$
TF_IDF	$TF\_IDF(t_j, \mathbf{d}_i) = (1 + \log f_{ij}) \log \frac{n}{n_j}$
TF_IG	$TF\_IG(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij}) IG(t_j, c_k)$
TF_OR	$TF\_OR(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij}) OR(t_j, c_k)$
TF_QUI	$TF\_QUI(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij}) \chi^2(t_j, c_k)$

MÉTODOS DE SELEÇÃO DE TERMOS (subseção 2.4.1)

ABSL	abs-logito
BNS	*bi-normal separation*
DF	freqüência de documentos (*Document Frequency*)
GR	razão de ganho (*Gain Ratio*)
IG	ganho de informação (*Information Gain*)
OR	razão de chances (*Odds Ratio*)
QUI	qui-quadrado (χ^2)

VETORES DE TERMOS ESTUDADOS EM REPRESENTAÇÃO VETORIAL BASE (subseção 3.4.4)

(continua)

adj	utiliza apenas os adjetivos extraídos dos PLs. Termos da forma "/ADJ/...". Exemplo: /ADJ/sustentável.
na	utiliza os substantivos (n) e adjetivos (a) extraídos dos PLs, com as identificações "/NOUN/..." e "/ADJ/...". Exemplos: /ADJ/sustentável; /NOUN/empreendimento.
nastp	utiliza os substantivos (n) e adjetivos (a) extraídos dos PLs, com as identificações "/NOUN/..." e "/ADJ/..."; elimina os termos que ocorreram em apenas um PL de treinamento; elimina os termos que não acrescentam conteúdo semântico importante ao domínio considerado (*stop words* (stp)); e padroniza a escrita de alguns termos para os quais são utilizadas mais de uma forma de escrita, como detran/detran-df e art/arts/artículo/artigo.
nav	utiliza os substantivos (n), adjetivos (a) e verbos (v) extraídos dos PLs, com as identificações "/NOUN/...", "/ADJ/..." e "/VERB/...", após limpeza dos termos extraídos pela ferramenta IdeXmlClient. Exemplos: /NOUN/empreendimento; /ADJ/sustentável; /VERB/dar.
noun	utiliza apenas os substantivos (noun) extraídos dos PLs. Termos da forma "/NOUN/...". Exemplo: /NOUN/empreendimento.

VETORES DE TERMOS ESTUDADOS EM REPRESENTAÇÃO VETORIAL BASE (subseção 3.4.4)

(conclusão)

orig	utiliza os substantivos, adjetivos e verbos extraídos dos PLs, com as identificações "/NOUN/...", "/ADJ/..." e "/VERB/...", sem limpeza dos termos extraídos pela ferramenta IdeXmlClient. Exemplos: /NOUN/N°, /ADJ/defensiva", /VERB/dar. (obs.: apenas o vetor de termos "orig" utiliza os termos sem limpeza).
sna	utiliza os substantivos (n) e adjetivos (a) extraídos dos PLs, sem (s) as identificações "/NOUN/..." e "/ADJ/...". Exemplos: sustentável; empreendimento.
snastp	utiliza os substantivos (n) e adjetivos (a) extraídos dos PLs, sem (s) as identificações "/NOUN/..." e "/ADJ/..."; elimina os termos que ocorreram em apenas um PL de treinamento; elimina os termos que não acrescentam conteúdo semântico importante ao domínio considerado (*stop words* (stp)); e padroniza a escrita de alguns termos para os quais são utilizadas mais de uma forma de escrita, como detran/detran-df e art/arts/artículo/artigo. P.S.: vetor de termos base usando dicionário global.
snav	utiliza os substantivos (n), adjetivos (a) e verbos (v) extraídos dos PLs, sem (s) as identificações "/NOUN/...", "/ADJ/..." e "/VERB/...". Exemplos: empreendimento; sustentável; dar.
verb	utiliza apenas os verbos extraídos dos PLs. Termos da forma "/VERB/....". Exemplo: /VERB/dar.

VETORES DE TERMOS ESTUDADOS EM REDUÇÃO DA DIMENSIONALIDADE (subseção 3.4.5)

dfxx	utiliza os termos que ocorreram em pelo menos xx PLs de treinamento, com representação vetorial usando dicionário global (dfxx). P.S.: df02 é o vetor de termos base “snastp” usando dicionário global.
dlxx	utiliza os termos que ocorreram em pelo menos xx (xx) PLs de treinamento (representação vetorial usando dicionário local (dl)). P.S.: dl02 é o vetor de termos base “snastp” usando dicionário local.
rdxxxx	utiliza os xxxx (xxxx) termos com os maiores escores (rd) globais obtidos para os termos nas diversas categorias temáticas do problema (escores globais calculados pela função máximo e pelo correspondente método de seleção de termos que compõe o peso calculado por atribuição supervisionada de pesos).

VETORES DE TERMOS ESTUDADOS EM AUMENTO DE PESO PARA ALGUMAS PALAVRAS (subseção 3.4.6)

coxxx	aumenta em xxx% (xxx) o peso das palavras relacionadas às matérias de competência das comissões permanentes (co).
em2	dobra (2) a freqüência de ocorrência das palavras presentes nas ementas (em).

VETORES DE TERMOS ESTUDADOS NO ESTUDO FINAL COMBINANDO REDUÇÃO DE DIMENSIONALIDADE COM AUMENTO DE PESO PARA ALGUMAS PALAVRAS (subseção 3.4.7)

(continua)

df09	utiliza os termos que ocorreram em pelo menos nove (df09) PLs de treinamento (representação vetorial usando dicionário global).
dl02	vetor de termos base “snastp” usando dicionário local.
dl09	utiliza os termos que ocorreram em pelo menos nove (09) PLs de treinamento (representação vetorial usando dicionário local (dl)).
co30	aumenta em 30% (30) o peso das palavras relacionadas às matérias de competência das comissões permanentes (co).
c30e2	dobra a freqüência de ocorrência das palavras presentes nas ementas (e2) e aumenta em 30% os pesos das palavras relacionadas às matérias de competência das comissões permanentes (c30).
c30f9	utiliza apenas as palavras que ocorreram em nove ou mais PLs de treinamento (f9), e aumenta os pesos para as palavras relacionadas às matérias de competência das comissões permanentes em 30% (c30).
c30e2f9	utiliza apenas as palavras que ocorreram em nove ou mais PLs de treinamento (f9), dobra a freqüência de ocorrência das palavras presentes nas ementas (e2), e aumenta em 30% os pesos das palavras relacionadas às matérias de competência das comissões (c30).

VETORES DE TERMOS ESTUDADOS NO ESTUDO FINAL COMBINANDO REDUÇÃO DE DIMENSIONALIDADE COM AUMENTO DE PESO PARA ALGUMAS PALAVRAS (subseção 3.4.7)

(conclusão)

dgc30e2f9	utiliza dicionário global (dg), apenas palavras que ocorreram em nove ou mais PLs de treinamento (f9), dobra a freqüência de ocorrência das palavras presentes nas ementas (e2), e aumenta em 30% os pesos das palavras relacionadas às matérias de competência das comissões (c30).
dlem2f9	utiliza dicionário local (dl), apenas palavras que ocorreram em nove ou mais PLs de treinamento (f9), e dobra a freqüência de ocorrência das palavras presentes nas ementas (em2).
em2	dobra (2) a freqüência de ocorrência das palavras presentes nas ementas (em).
em2f9	utiliza apenas as palavras que ocorreram em nove ou mais PLs de treinamento (f9), e conta em dobro a freqüência de ocorrência das palavras presentes nas ementas (em2).
ementas	utiliza apenas os termos presentes nas ementas.
snastp	vetor de termos base usando dicionário global.

1 Introdução

1.1 Objetivo

No cumprimento de uma de suas funções básicas, que é a elaboração de leis, a Câmara Legislativa do Distrito Federal (CLDF) conta com as comissões permanentes, órgãos técnico-legislativos e especializados, formados por grupos de deputados, que têm entre suas atribuições apreciar e emitir parecer sobre as propostas de lei submetidas a seus exames.

A indicação das comissões permanentes que devem apreciar cada proposta de lei deve ser baseada nas competências estabelecidas para essas nos artigos 63 a 69-B do Regimento Interno da Câmara Legislativa do Distrito Federal, e é efetuada pela Assessoria de Plenário e Distribuição. Atualmente, essa atividade é realizada de forma manual, pela leitura da proposta de lei, e, em geral, é executada pelo próprio chefe da Assessoria.

Alguns fatores que prejudicam esse processo são: 1) o quantitativo de propostas de lei analisadas por uma única pessoa; 2) o fato da chefia da Assessoria de Plenário e Distribuição, cargo de livre provimento, da confiança da Presidência, ser, em geral, ocupada por servidores sem vínculo efetivo com a CLDF e, normalmente, sem experiência na área; 3) a mudança na composição da Mesa Diretora da Casa a cada dois anos que, muitas vezes, implica a mudança do chefe dessa Assessoria.

Esses fatores intervenientes acarretam incorreções e falta de padronização na distribuição das propostas de lei às comissões permanentes, como o não

encaminhamento de propostas de lei que tratam de matérias análogas para as mesmas comissões.

O objetivo do presente trabalho consiste em utilizar a técnica de categorização automática de textos para desenvolver um modelo que possibilite a indicação, de forma automática, das comissões permanentes que devem apreciar cada uma das propostas de lei apresentadas à Câmara Legislativa, e que produza resultados satisfatórios de modo a subsidiar ou mesmo substituir o procedimento manual atualmente realizado.

Dadas as especificidades e volume do trabalho envolvido para o desenvolvimento desta tese, abordou-se o processo de distribuição apenas com relação aos projetos de lei. Todavia, pode-se aplicar a mesma metodologia a outros tipos de proposta de lei, como projeto de lei complementar, proposta de emenda à Lei Orgânica, projeto de decreto legislativo e indicação, que são submetidas a processo de distribuição similar ao dos projetos de lei.

1.2 Categorização automática de textos

A classificação de documentos textuais é uma tarefa tipicamente realizada por humanos, especialistas no domínio de interesse, que lêem os documentos e os classificam em categorias temáticas pré-definidas.

Com o número de documentos potenciais para exame por um humano cada vez mais excedendo em muito a quantidade de documentos que uma pessoa necessita ler, técnicas como a categorização automática de textos (*text categorization*, *text classification* ou *topic spotting*) - atividade que consiste em classificar automaticamente

textos em linguagem natural[1] em categorias temáticas pré-definidas - têm testemunhado um interesse crescente nos últimos tempos. Esse interesse também se deve ao fato dessa técnica estar conseguindo atingir níveis de desempenho competitivos em relação às classificações realizadas por profissionais treinados (SEBASTIANI, 2002).

Exemplos de aplicação da técnica são: 1) classificação de notícias jornalísticas por assunto (DEBOLE & SEBASTIANI, 2003; LEWIS et al., 2004; SOUCY & MINEAU, 2005); 2) classificação automática de páginas Web na hierarquia de categorias do Yahoo! (MLADENIC, 1998); 3) filtragem de e-mail indesejado - *spam* (GÓMEZ HIDALGO, 2002); 4) filtragem de conteúdo impróprio da Internet - pornografia, violência e racismo - em ambientes educacionais (GÓMEZ HIDALGO et al., 2002); 4) envio por e-mail de notícias jornalísticas on-line personalizadas para clientes (GIRÁLDEZ et al., 2002); 5) identificação de dialetos ou língua e identificação de autor em textos literários ou artigos cuja autoria é desconhecida ou controversa (TEAHAN, 2000). Outras aplicações podem ser encontradas em Sebastiani (2002) e Gómez Hidalgo (2003).

O surgimento da técnica de categorização automática de textos remonta a pelo menos o início da década de 60, com o importante trabalho de M. E. Maron sobre classificação probabilística de textos (SEBASTIANI, 2002).

Até o final da década de 80, a abordagem dominante ao problema envolvia a construção de classificadores baseados em engenharia do conhecimento (sistema especialista), isto é, a construção manual de um conjunto de regras lógicas codificando o conhecimento de especialistas do domínio em como classificar documentos nas categorias sob consideração (SEBASTIANI, 2002).

[1]Discurso comum, linguagem utilizada habitualmente na escrita e na fala, texto livre (LANCASTER,1993, pp. 200).

Nessa abordagem, o especialista deve prever, para cada categoria temática envolvida no problema, todas as combinações de palavras cuja co-ocorrência leva um documento a ser classificado na referida categoria.

Esses classificadores baseados em sistema especialista, apesar de apresentarem níveis altos de desempenho em termos de efetividade na classificação, são computacionalmente caros e muito dependentes do domínio.

Na década de 90, com o aumento rápido da produção e disponibilidade de documentos textuais por meio eletrônico, a categorização automática de textos presenciou um maior e renovado interesse. Um novo paradigma baseado em aprendizado de máquina suplantou a abordagem anterior.

Nesse novo paradigma, usando como base um conjunto de documentos pré-classificados nas diversas categorias temáticas do domínio de interesse, um processo de indução é conduzido para identificar as características que diferenciam as categorias temáticas entre si (SEBASTIANI, 2002).

Com base nessa indução, um modelo de classificação é construído e pode ser usado para classificar um novo documento ainda não visto, comparando as características presentes neste documento com os padrões aprendidos dos documentos pertencentes às diversas categorias temáticas estudadas.

Na terminologia usada na área de aprendizado de máquina, esse problema de classificação é denominado aprendizagem supervisionada, uma vez que o processo de aprendizagem - estimação do modelo de classificação - é supervisionado por exemplos de documentos cujas categorias temáticas são conhecidas.

O tratamento textual utilizado em categorização automática de textos por aprendizado de máquina é fortemente baseado nas metodologias desenvolvidas na área de recuperação da informação, que também lida com tarefas centradas no conteúdo dos

documentos textuais e é uma área bem mais antiga e com muito mais pesquisa desenvolvida.

O presente trabalho será desenvolvimento com base na metodologia utilizada em aprendizado de máquina. Para a aplicação da técnica de categorização automática de textos sob o enfoque dessa abordagem é necessário, fundamentalmente:

- ter disponível um conjunto de documentos[2] pré-classificados nas diversas categorias temáticas de interesse;
- transformar a informação textual contida nos documentos em uma representação possível de ser tratada computacionalmente pelos modelos de classificação;
- escolher os termos mais relevantes do conjunto de documentos, isto é, os que permitem melhor discriminação entre as categorias temáticas sob estudo (etapa nem sempre realizada);
- definir pesos para os termos presentes nos documentos, de forma a obter uma representação documental que possibilite a maior discriminação possível entre as categorias temáticas consideradas;
- estimar o modelo de classificação;
- avaliar a efetividade do modelo construído.

Essas etapas serão detalhadas no capítulo seguinte.

Uma excelente visão geral sobre a técnica pode ser encontrada no *survey* elaborado por Sebastiani (2002).

[2] No contexto deste trabalho, o termo “documento” estará se referindo a objeto eletrônico de caráter estritamente textual.

1.3 Abordagem para o problema de categorização

O problema de categorização mais simples consiste na classificação binária (*single-label*), onde é possível classificar os documentos em apenas uma de duas categorias. Por exemplo, sistema de filtragem de e-mail, onde as mensagens são classificadas como importantes ou não-importantes.

No presente trabalho, no entanto, dependendo da matéria tratada, um projeto de lei (documento a ser classificado) pode ser distribuído para mais de uma comissão permanente (categoria temática). Esse tipo de classificação, onde um mesmo documento pode ser classificado em mais de uma categoria temática, é conhecido como *multilabel*.

A abordagem utilizada na maioria dos trabalhos publicados na área de categorização automática de textos e que também será adotada neste trabalho consiste em considerar o problema *multilabel* de classificação nas categorias $c_1,...,c_m$ como m problemas independentes de classificação binária em $\{c_k, \bar{c}_k\}$, para $k=1,..., m$, onde $\bar{c}_k$ refere-se ao conjunto completo de categorias sem a categoria c_k. Nesse caso, m classificadores são construídos, um por categoria, e cada problema de classificação responde à pergunta se o documento deve ou não ser classificado na referida categoria.

Devido à própria estruturação desses m problemas de classificação binária nas categorias $\{c_k, \bar{c}_k\}$, $k=1,..., m$, dependendo do número de categorias que estiverem sendo representadas em $\bar{c}_k$, a quantidade de documentos a ela pertencentes pode ultrapassar em muito a quantidade de documentos em c_k. Conjuntos de dados com essas características são conhecidos como assimétricos ou desbalanceados e são típicos em categorização automática de textos usando essa abordagem.

Na nomenclatura utilizada em classificação binária, a classe c_k (classe de interesse) é denominada classe positiva e os documentos dessa classe são denominados documentos relevantes ou exemplos positivos; a classe $\bar{c}_k$ (“demais classes”) é denominada classe negativa e os documentos dessa classe são denominados documentos irrelevantes ou exemplos negativos.

1.4 Contribuições da tese

Em geral, nos problemas de categorização automática de textos, as categorias são vistas apenas como rótulos simbólicos - esporte, cultura, política, etc. - e nenhum conhecimento adicional sobre os seus significados é incorporado ao processo de construção do classificador. No presente trabalho, no entanto, existe disponível uma descrição para as categorias temáticas - que correspondem às comissões permanentes -, uma vez que os artigos 63 a 69-B do Regimento Interno da Câmara Legislativa listam os assuntos que cada comissão deve apreciar.

Além disso, compõem um projeto de lei: uma ementa, que resume o objetivo do projeto; o corpo ou texto da lei, que encerra a matéria disciplinada; e uma justificação, onde o autor procura demonstrar a necessidade ou oportunidade do projeto usando uma série de argumentos (justificativas). Quando bem redigida, a ementa fornece uma boa indicação sobre o conteúdo do projeto de lei, devendo conter as palavras mais importantes do texto.

Dada a disponibilidade dessas informações adicionais ao problema, pretende-se investigar o efeito na efetividade da classificação de se ressaltar a

importância das palavras presentes nas ementas e das relacionadas às matérias de competência das comissões permanentes.

Também será apresentada uma métrica para seleção de termos - denominada no presente trabalho de abs-logito - sobre a qual não se tem conhecimento de sua utilização na área de categorização automática de textos. A métrica abs-logito corresponde a uma transformação da métrica razão de chances (odds ratio) - utilizada para o mesmo fim -, que corrige a assimetria apresentada por esta última métrica que privilegia a seleção dos termos mais prevalentes na categoria de interesse.

Além disso, seguindo idéia proposta por Debole & Sebastiani (2003), de incluir no cálculo dos pesos para os termos a importância desses para a discriminação das categorias, será considerada neste trabalho, para esse fim, a utilização das métricas de seleção de termos *bi-normal separation* - que apresentou excelentes resultados na pesquisa realizada por Forman (2003) – e abs-logito (proposta neste trabalho). Essas idéias também não foram vistas em trabalhos publicados.

Adicionalmente às contribuições citadas, um dos objetivos deste trabalho é colaborar com o Legislativo, mostrando a viabilidade prática da utilização de ferramentas de mineração de textos, como a categorização automática de textos. A proposta tem importância para a disseminação dessa cultura dentro do serviço público.

Este trabalho, além de ser valioso para a Câmara Legislativa do Distrito Federal, poderá ser igualmente útil para outras casas legislativas estaduais, e mesmo para a Câmara dos Deputados, que também realizam a indicação da distribuição das propostas de lei de forma manual.

1.5 Organização da tese

Os capítulos subseqüentes desta tese estão estruturados conforme descrito a seguir.

No capítulo 2, são mostrados os principais conceitos relacionados à técnica de categorização automática de textos, descrevendo em cada uma das etapas do processo as abordagens mais utilizadas.

No capítulo 3, é apresentado o estudo sobre a modelagem, por meio da técnica de categorização automática de textos, da indicação de distribuição dos projetos de lei para as comissões permanentes da Câmara Legislativa do Distrito Federal. Nesse capítulo serão detalhados a base de dados adotada, os programas utilizados, as análises realizadas e os resultados obtidos.

Por fim, no capítulo 4 são expostas as conclusões finais sobre o trabalho, apontando sugestões para trabalhos futuros.

2 Técnica de categorização automática de textos

2.1 Corpus

O recurso-chave para a construção do modelo de categorização automática de textos consiste em ter disponível um conjunto de documentos pré-classificados nas diversas categorias temáticas do domínio de interesse. Esse conjunto de documentos é denominado corpus.

Para avaliar como deverá ser o desempenho do modelo construído em documentos futuros, ainda não vistos, as seguintes abordagens são utilizadas nas análises encontradas na literatura:

a) treinamento/teste (*holdout method*): o corpus é dividido em dois subconjuntos disjuntos, denominados conjunto de treinamento e conjunto de teste. Em geral, selecionam-se aleatoriamente 2/3 dos documentos para o conjunto de treinamento e 1/3 para o conjunto de teste, onde:

- conjunto de treinamento: conjunto de documentos utilizado para construção do modelo de classificação;
- conjunto de teste: conjunto de documentos utilizado para verificar a efetividade do modelo construído, comparando, para cada documento desse conjunto, as classificações realizadas pelo especialista (classificação manual) com as classificações obtidas com base no modelo construído (classificação automática);

b) para propósitos de seleção de modelo, uma variante do método exposto no item “a)” ainda divide o conjunto de treinamento em dois subconjuntos disjuntos (um dos subconjuntos mantém o nome de conjunto de treinamento e o outro é denominado conjunto de validação). Nesse procedimento são utilizados três subconjuntos:

- conjunto de treinamento: conjunto de documentos utilizado para construção dos modelos de classificação candidatos, de acordo com parâmetros pré-especificados;
- conjunto de validação: conjunto de documentos utilizado para testar a efetividade dos modelos candidatos;
- conjunto de teste: conjunto de documentos utilizado para verificar a efetividade do “melhor” modelo entre os candidatos considerados, ou seja, do modelo com os valores de parâmetros que forneceram a maior efetividade na categorização testada sobre os documentos do conjunto de validação. O modelo aqui considerado é o “melhor” modelo, porém re-estimado, utilizando os dados dos conjuntos de treinamento e validação.

Nessa abordagem, os conjuntos de treinamento e validação são utilizados para avaliar o desempenho de vários modelos candidatos visando escolher o “melhor”. Porém, para evitar que o modelo assim selecionado se ajuste excessivamente bem aos dados do conjunto de validação e não consiga reproduzir o mesmo resultado em documentos diferentes desses, o desempenho de generalização do modelo é medido sobre o conjunto de teste, que é diferente do conjunto de validação.

c) validação cruzada com k subconjuntos (*k-fold cross-validation*): o corpus é dividido aleatoriamente em $k \geq 2$ subconjuntos mutuamente exclusivos, aproximadamente com o mesmo número de documentos. Um modelo de classificação é construído

usando os documentos pertencentes a (*k*-1) desses subconjuntos (conjunto de treinamento) e a efetividade do modelo é avaliada nos documentos do subconjunto não utilizado para construção do modelo (conjunto de teste). Esse procedimento é repetido *k* vezes, cada vez utilizando um subconjunto diferente para avaliação do desempenho do modelo construído, de forma que cada um dos *k* subconjuntos seja utilizado uma única vez para tal fim. A efetividade da classificação por validação cruzada é estimada como a média da medida de desempenho calculada em cada um dos *k* subconjuntos usados para avaliar a efetividade do respectivo modelo construído.

Quando o valor utilizado para *k* é igual ao número de documentos do corpus, esse método é denominado "*leave-one-out cross-validation*". Nesse caso, o modelo de classificação é construído usando (*k*-1) documentos e a efetividade do modelo é avaliada no documento não utilizado para construção do modelo.

O procedimento de validação cruzada pode ser usado simplesmente para estimar o desempenho de um modelo de categorização ou também para selecionar modelos escolhendo um entre vários candidatos - o que apresentar a melhor efetividade na categorização.

A forma de análise descrita no item "a)" - utilizada na maioria dos trabalhos publicados em categorização automática de textos - pode ser encontrada em Yang & Pedersen (1997), Joachims (1998), Debole & Sebastiani (2003), Madigan & Eyheramendy (2005) e Soucy & Mineau (2005).

A forma de análise descrita no item "b)" pode ser encontrada em Lewis et al. (2004) e Joachims (2005). Lewis et al. (2004) apresentam análises muito interessantes no corpus Reuters, volume 1, (RCV1-v2), usando três algoritmos de classificação - *Support Vector Machines* (SVMs), *k-Nearest Neighbor* (*k*-NN) e

Rocchio -, onde os melhores parâmetros para cada um desses modelos são escolhidos usando validação cruzada - divisão em 5 subconjuntos – sobre o conjunto de treinamento. Joachims (2005) utiliza a subdivisão do conjunto de treinamento em 2/3 para construção dos modelos candidatos e 1/3 para avaliação desses modelos a fim de escolher um dos parâmetros do algoritmo *Support Vector Machines*.

A forma de análise descrita no item "c)" pode ser encontrada em Schapire & Singer (2000) e Zhang & Zhou (2006), que utilizam validação cruzada com divisão do corpus em três subconjuntos.

Forman (2003) utiliza validação cruzada estratificada – divisão em quatro subconjuntos - repetida cinco vezes. Na validação cruzada estratificada, cada subconjunto contém aproximadamente as mesmas proporções de documentos que o corpus em cada uma das categorias temáticas. A repetição da validação cruzada cinco vezes gera cinco divisões distintas do corpus em quatro subconjuntos de documentos.

A metodologia adotada por Forman (2003) é explicada em Kohavi (1995), segundo o qual, com exceção do "*leave-one-out*", que sempre corresponde a um procedimento completo, a validação cruzada com *k* subconjuntos utiliza uma particular divisão dos dados em *k* subconjuntos e representa uma estimativa da validação cruzada completa com *k* subconjuntos. Ainda de acordo com Kohavi (1995), a repetição da validação cruzada várias vezes, utilizando diferentes divisões dos *k* subconjuntos, fornece uma estimativa de Monte Carlo melhor para a validação cruzada completa.

Kohavi (1995) explica que a validação cruzada completa é um método com custo muito alto e normalmente não é executada, pois exige a avaliação de $\binom{n}{n/k}$ subconjuntos, uma vez que na divisão dos *n* documentos do corpus em *k* subconjuntos,

cada subconjunto conterá *n/k* documentos, sendo que existem $\binom{n}{n/k}$ maneiras distintas dos *n* documentos serem distribuídos em *k* grupos de *n/k* documentos.

Com relação ao melhor método de estimação/avaliação de modelo para as medidas de desempenho utilizadas na área de categorização automática de textos, não se tem conhecimento da existência de nenhum estudo comparativo fornecendo indicações a respeito do assunto.

No entanto, deve-se ter em mente que um modelo de classificação é apenas tão bom quanto o conjunto de treinamento e a precisão das estimativas serão apenas tão boas quanto os dados de teste. Dessa forma, tanto os dados de treinamento como os de teste devem ser representativos dos dados que serão encontrados na prática quando o classificador for colocado em produção. Assim, para construção/avaliação do modelo de classificação, deve-se ter disponibilidade de exemplos de documentos em quantidades suficientes que permitam bem caracterizar as particularidades de cada uma das categorias temáticas sob estudo.

Em termos gerais, o método *holdout* é o método de estimação mais simples e fácil de calcular. Porém, quando os dados disponíveis são escassos, pode tornar-se difícil ainda separar uma parte desses para testar a efetividade do modelo. Além disso, com poucos dados, o desempenho do modelo pode depender da particular divisão dos dados nos conjuntos de treinamento e teste. Assim, performances bem diferentes podem ser obtidas dependendo de quais documentos são selecionados para o conjunto de treinamento e quais são selecionados para o conjunto de teste.

Nessas situações, a validação cruzada oferece uma forma de estimação melhor que o método *holdout*, pois faz melhor utilização dos dados disponíveis. A desvantagem da validação cruzada é o alto número de conjuntos de treinamento e teste gerados e a quantidade de cálculos necessários para estimação e avaliação dos modelos.

Em categorização automática de textos, a aplicação da validação cruzada apresenta algumas particularidades que devem ser observadas no momento da análise. Como será visto na seção seguinte, em categorização de textos, o conjunto de atributos (variáveis) utilizado para caracterizar os dados a serem analisados corresponde aos termos presentes em pelo menos um dos documentos que compõem o conjunto de treinamento. Assim, diferentemente das análises usuais com dados numéricos, onde o número de atributos (variáveis) é fixo, aqui o número de atributos é definido apenas após a definição de quais documentos compõem o conjunto de treinamento.

Além disso, como em cada uma das *k* rodadas da validação cruzada os documentos que formam o conjunto de treinamento mudam, o conjunto de atributos (variáveis) considerado também pode mudar dependendo dos termos presentes nesses documentos. Pelo mesmo motivo, o número de documentos que contém cada termo também pode mudar. Dado que esse número, em geral, é utilizado no cálculo dos pesos dos atributos, esses pesos também podem ser afetados tanto nos documentos de treinamento como nos de teste.

Como regra geral, independentemente da forma de estimação utilizada, para melhorar o desempenho do classificador quando esse for colocado em uso na prática, após decidir sobre o "melhor" modelo de categorização, costuma-se re-estimar o modelo escolhido utilizando todos os documentos do corpus.

2.1.1 Corpora disponíveis na Internet mais utilizados

Nesta subseção serão descritos os corpora[3] mais utilizados nos trabalhos publicados, uma vez que eles serão referenciados diversas vezes no decorrer da presente

[3] Plural de corpus.

tese. Esses corpora encontram-se disponíveis na Internet para propósitos experimentais, sendo os relativos às reportagens da agência de notícias Reuters adotados na maioria das pesquisas desenvolvidas na área de categorização automática de textos.

A versão inicial do corpus Reuters, denominada Reuters-22173, contém 21.450 notícias jornalísticas referentes ao ano de 1987. Essa versão, no entanto, possui várias ambigüidades na formatação, incorreções tipográficas e nas classificações, além de grande parte dos trabalhos publicados pela comunidade científica com base nesses dados terem utilizado subconjuntos diferentes da coleção, inviabilizando a comparação dos resultados obtidos. Essa coleção foi então revista e corrigida - tendo como principal responsável David D. Lewis - gerando a versão denominada Reuters-21578[4].

Com essa nova coleção ainda foram utilizados alguns subconjuntos de dados diferentes nos trabalhos publicados, mas muitos pesquisadores já utilizaram o mesmo subconjunto com a mesma divisão dos conjuntos de treinamento e teste. A divisão mais utilizada é denominada Reuters-21578, distribuição 1.0, "ModApte", e consiste de 12.902 reportagens, com 9.603 notícias para treinamento e 3.299 para teste. As reportagens encontram-se classificadas em 118 categorias, com o número de categorias por documento variando de zero (documento não classificado) a 16, sendo a média de categorias por documento igual a 1,08. O número de documentos por categoria varia de um a 3.964 (DEBOLE & SEBASTIANI, 2003).

Usando o corpus Reuters-21578, "ModApte", existem estudos englobando: a) todas as 118 categorias; b) as 115 categorias com pelo menos um documento de treinamento; c) as 90 categorias com pelo menos um documento de treinamento e um de teste; d) as 10 categorias com o maior número de documentos de treinamento.

[4] Disponível em http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt. Acesso em: 18 set. 2006.

Em 2000, a agência de notícias Reuters disponibilizou o corpus denominado "Reuters Corpus Volume 1" (RCV1) com reportagens em língua inglesa produzidas por seus jornalistas entre 20 de agosto de 1996 e 19 de agosto de 1997. Esse corpus corresponde ao maior corpus atualmente disponível para uso pela comunidade acadêmica, e contém 806.791 notícias distribuídas em 1.362 categorias ("*Topic*" - 126 categorias; "*Industy*" - 870 categorias; e "*Region*" - 366 categorias).

Apesar de ser superior em qualidade às coleções anteriores, a coleção RCV1 possui erros na atribuição de categorias, apresenta documentação deficiente para compreensão da atribuição das reportagens às categorias, além de descrever algumas categorias que são inconsistentes com as categorias atribuídas aos artigos (LEWIS et al., 2004).

Com base na documentação do corpus RCV1, em entrevistas com funcionários da Reuters e em análises das reportagens e categorias, Lewis et al. (2004) realizaram um trabalho de documentação e limpeza do corpus RCV1 transformando-o em uma coleção para testes. Eles denominaram essa versão corrigida de RCV1-v2 e a versão original de RCV1-v1. A coleção RCV1-v2[5] contém 804.414 notícias distribuídas em 823 categorias ("*Topic*" - 103 categorias; "*Industy*" - 354 categorias; e "*Region*" - 366 categorias), que foram separadas em 23.149 documentos para treinamento e 781.265 para teste, a fim de permitir comparações entre os experimentos a serem realizados nesse corpus. A divisão do corpus nesses conjuntos de treinamento e teste foi denominada LYRL2004 (iniciais dos seus organizadores). Após a aplicação de *stemming* e remoção de *stop words*, o vetor de treinamento possui 47.152 termos.

[5] Os dados sobre as categorias, lista de *stop words*, os termos utilizados em cada documento, os vetores utilizados para treinamento e teste e as tabelas de contingência - matriz de confusão - encontram-se disponíveis em *http://jmlr.csail.mit.edu/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm*. Acesso em: 18 de set. 2006.

Sobre a coleção Reuters, Joachims (1998) comenta que ela é conhecida pela correspondência direta entre palavras e categorias. A presença da palavra "*wheat*" em um documento, por exemplo, é um forte indicativo da categoria "*wheat*" para o documento. Leopold & Kindermann (2002) acrescentam que o tamanho dos documentos dessa coleção não é natural. A maioria dos textos da coleção tem menos de 100 palavras, enquanto que textos de notícias jornalísticas são, em média, duas vezes maiores. Soucy & Mineau (2005) reforçam que os documentos da coleção são escritos de forma sucinta usando um vocabulário limitado. Esses motivos fazem com que os valores obtidos para as medidas de desempenho na coleção Reuters sejam bastante altos.

Outra coleção também usada com certa freqüência nos trabalhos publicados, mas em proporção bem inferior à da coleção Reuters, é a coleção OHSUMED[6], que corresponde a um subconjunto da base de dados MEDLINE, a base de dados on-line da Biblioteca Nacional de Medicina dos Estados Unidos. OHSUMED consiste de 348.566 artigos de 270 revistas médicas publicadas entre 1987 e 1991. Os artigos são classificados em 14.321 categorias MeSH (*Medical Subject Headings*), que é o tesauro com vocabulário controlado da Biblioteca Nacional de Medicina dos Estados Unidos.

De acordo com Yang & Pedersen (1997), todos os 348.566 documentos da coleção OHSUMED têm títulos, mas apenas 233.445 têm *abstracts*, considerando-se como documento o título mais o *abstract*. Além disso, esclarecem Yetisgen-Yildiz & Pratt (2005) que, devido ao fato de OHSUMED incluir apenas uma pequena parte da base de dados MEDLINE, muitas categorias têm muito poucos documentos para treinamento (753 categorias têm apenas um documento), o que faz com que os

[6] Disponível em: http://www.ltg.ed.ac.uk/disp/resources/ . Acesso em: 05 dez. 2006.

pesquisadores limitem seus experimentos a um pequeno subconjunto de categorias com um mínimo de exemplos por categoria.

Yang & Pedersen (1997), por exemplo, utilizaram os documentos de 1990 para treinamento e os de 1991 para teste. Nesse experimento obtiveram 72.076 termos únicos no conjunto de treinamento, sendo o número médio de termos por documento igual a 167. Em média, 12 categorias são atribuídas a cada documento do subconjunto de dados utilizado.

Yetisgen-Yildiz & Pratt (2005) identificaram os 179.796 documentos classificados nas categorias MeSH agrupadas abaixo das categorias hierárquicas doenças ou síndromes (1.928 categorias). Elas separaram aleatoriamente 80% (143.837) desses documentos para treinamento e 20% (35.959) para teste, utilizando, no entanto, em seu experimento, apenas as 634 categorias com pelo menos 75 documentos de treinamento, considerando como documento apenas os títulos.

Segundo Lewis et al. (2004), a coleção OHSUMED tem desvantagens, como não conter o texto completo dos documentos, a linguagem médica ser de difícil compreensão para não-especialistas - vocabulário bem técnico - e sua categoria hierárquica (MeSH) ser grande e estruturalmente complexa.

Também utilizado com certa freqüência é o corpus 20 *Newsgroups*[7]. Esse corpus consiste de 20.000 mensagens de texto, denominadas artigos, que foram postadas por usuários em 20 grupos de notícias (*newsgroups*) no meio de comunicação on-line Usenet (*Unix User Network*). 1.000 artigos foram coletados em cada um dos 20 *newsgroups*. Entre os grupos de notícias do corpus, alguns discutem assuntos relacionados. Por exemplo: três grupos discutem assuntos religiosos (talk.religion.misc, alt.atheism e soc.religion.christian); três discutem assuntos políticos (talk.politics.misc,

[7] Disponível em: http://people.csail.mit.edu/jrennie/20Newsgroups/. Acesso em: 05 dez. 2006.

talk.politics.guns, talk.politics.mideast); e outros discutem temas bem distintos, como produtos para venda (misc.forsale).

2.2 Indexação automática

Textos não podem ser diretamente interpretados por um modelo de categorização automática. Assim, um procedimento que mapeia um documento em uma representação compacta de seu conteúdo necessita ser obtida. Esse processo de extração automática das palavras do texto é denominado indexação automática, e deve ser aplicado de forma uniforme em cada um dos documentos dos conjuntos de treinamento, teste e validação (quando usado).

Documentos são tipicamente *strings* de caracteres. Dessa forma, antes de iniciar o processo de indexação automática, os textos são convertidos para letra minúscula e as palavras - seqüências máximas de caracteres não-brancos - são identificadas e separadas dos sinais de pontuação. Os sinais de pontuação, dígitos e outros símbolos sem conteúdo semântico são removidos.

Na forma mais simples de mapeamento de um texto para tratamento por computador, desconsidera-se a ordem em que as palavras aparecem no texto e extraem-se apenas palavras, como abastecimento, estudante e terreno, ignorando-se locuções[8], como meio ambiente ou Região Administrativa.

Em levantamento realizado por Sebastiani (2002), sobre estudos englobando também no conjunto de palavras extraídas dos textos locuções no sentido sintático (locuções formadas de acordo com a gramática da língua) e/ou “locuções estatísticas”

[8] Conjunto de palavras que equivale a uma só.

(locuções formadas por seqüência de palavras cujo padrão de ocorrência contígua na coleção são estatisticamente significantes), os resultados obtidos não foram uniformemente encorajadores em termos de melhorias na efetividade da categorização.

Para a área de recuperação da informação, Salton & Buckley (1988) comentam que, na extensiva literatura acumulada sobre o assunto, a esmagadora evidência é que o uso criterioso de palavras é preferível à incorporação de entidades mais complexas como locuções extraídas dos próprios textos ou obtidas a partir de outras fontes de vocabulário disponível, como dicionários, tesauros, etc.

Mochitti & Basili (2004) realizaram um estudo sobre a incorporação de representações mais ricas para os documentos na área de categorização automática de textos. Usando cinco corpora - corpus Reuters-21578, divisão Apté; corpus Reuters3, preparado por Y. Yang e colaboradores; coleção ANSA; coleção OHSUMED; corpus 20 *NewsGroups* - e três algoritmos de classificação - *Rocchio, Parameterized Rocchio Classifier e Support Vector Machines* -, eles estudaram a inclusão: a) da classe gramatical das palavras (substantivos, adjetivos e verbos); b) de substantivos próprios relacionados às categorias do domínio, como “Rome”, “George Bush”, “Audi 80”, e conceitos também referentes ao domínio, como “*bond issues*” e “*beach wagon*”, que usualmente são identificados por múltiplas palavras. Para lidar com problemas de polissemia - mesma palavra com mais de um significado -, incluíram, ainda, para os substantivos, palavras que ajudam a identificar o contexto em que a palavra é empregada (*word sense disambiguation*). No caso das locuções, essas foram acrescentadas à representação dos documentos e não substituíram as palavras que as compunham, pois nesse experimento, assim como em outras pesquisas anteriores, foi observado que a substituição das palavras de uma locução pela própria locução diminui o desempenho dos classificadores usados.

A conclusão do experimento realizado por Mochitti & Basili (2004) foi que as representações lingüísticas mais sofisticadas utilizadas não foram capazes de melhorar a efetividade na classificação dos algoritmos considerados. Eles observaram que locuções e palavras de contexto são bem representadas pelas palavras do texto, uma vez que uma palavra em uma categoria sempre assume o mesmo sentido, ao passo que as categorias diferem nas palavras e não no sentido das palavras.

Por outro lado, no estudo apresentado por Yetisgen-Yildiz & Pratt (2005) com o corpus OHSUMED, porém considerando como documentos apenas os títulos, a representação dos documentos acrescentando conceitos médicos afins (locuções relacionadas) às palavras extraídas dos títulos apresentou melhora em torno de 2% no desempenho do algoritmo *support vector machines*, em relação à representação dos documentos usando apenas as palavras extraídas dos títulos.

Assim, sem resultados consistentes em termos de ganhos em efetividade na classificação apontando em direção à utilização de formas mais sofisticadas de representação dos documentos, normalmente trabalha-se com a forma mais simples, extraindo dos textos apenas as palavras e ignorando as locuções.

Além dessa simplificação, outras simplificações são, em geral, realizadas e os documentos não são representados em um computador pelo seu conjunto completo de palavras. Como na área de recuperação da informação, em categorização automática de textos também são utilizados alguns recursos para melhorar a efetividade da classificação. Esses recursos aumentam a compactação do conjunto de palavras derivadas dos textos e acabam beneficiando a sua representação em computadores, principalmente para grandes coleções de textos. É importante ressaltar, no entanto, que a utilização desses recursos, que serão listados a seguir, deve ser criteriosa, uma vez que

alguns deles aplicados de forma agressiva - redução drástica de palavras - podem prejudicar a efetividade na classificação.

Desse modo, palavras sem poder de discriminação sobre o conteúdo dos textos, como conjunções, preposições e artigos, denominadas *stop words*, normalmente são excluídas. Além dessas, podem ser excluídas as palavras que não fornecem informações a respeito do domínio considerado. Nesse último caso, especialistas do domínio devem fornecer a lista de palavras a serem desconsideradas.

Também, freqüentemente, as palavras aparecem nos documentos com muitas variações morfológicas, sendo cada uma dessas variações vistas como palavras distintas pelo computador. Na maioria dos casos, no entanto, as variações morfológicas têm interpretações semânticas similares e podem ser consideradas como equivalentes para o propósito de aplicações de recuperação da informação e também de categorização automática de textos.

Nesse sentido, outra forma bastante utilizada de compressão do conjunto de palavras para representação dos textos em computadores é a aplicação de *stemming*, que substitui as variações morfológicas das palavras por seu *stem* (conjunto de caracteres que não necessariamente corresponde ao radical da palavra). Para propósitos de recuperação da informação, em geral, não importa se os *stems* gerados são palavras genuínas ou não, desde que palavras com o mesmo significado básico sejam reduzidas ao mesmo *stem*, e palavras com significados distintos sejam mantidas separadamente.

De acordo com Lopes (2004), existem vários tipos de algoritmos de *stemming*. Uns são mais conservadores, como o *stemmer* S (remove apenas uns poucos finais de palavras), outros são mais agressivos, como o método de Lovins (remove 250 sufixos), sendo o método de Porter (remove 60 sufixos) considerado como referência no processo de *stemming* para a língua inglesa (existe uma adaptação deste algoritmo para

a língua portuguesa). Como exemplo da aplicação de um algoritno de *stemming*, considere a aplicação do algoritmo de Porter às palavras “considerar”, “considerado”, “consideração” e “considerações”, que causa a substituição dessas por seu *stem* “consider”.

Uma forma mais sofisticada de compressão do conjunto de palavras, porém pouco utilizada devido à maior complexidade para a sua obtenção, consiste na aplicação de *lemmatization* ao invés de *stemming*. O processo de *lemmatization* envolve a identificação da classe gramatical da palavra e a redução de cada palavra do corpus para seu respectivo *lemma*, que corresponde à palavra primitiva a partir da qual todas as outras a ela relacionadas derivam. O *lemma* corresponde – grosso modo - às palavras da forma como as buscamos em um dicionário da língua: as formas verbais flexionadas são substituídas pela forma infinitiva, os substantivos plurais são substituídos pelo singular, etc. Como exemplo dessa aplicação, considere as formas verbais flexionadas como “modificou”, “modificaram” e “modifica”, que são substituídas pela forma infinitiva “modificar”.

Uma vez que na aplicação do processo de *lemmatization* identifica-se, além do *lemma*, a classe gramatical da palavra, pode-se também, ao utilizar esse processo, obter uma redução ainda maior do conjunto de palavras que representam um documento, eliminando-se classes gramaticais que apresentam pouco poder de discriminação entre os documentos, como a classe dos verbos, por exemplo.

As palavras resultantes da aplicação dos recursos acima descritos são denominadas termos. Desses termos, apenas os presentes nos documentos do conjunto de treinamento serão usados para compor a representação dos documentos para fins de construção dos modelos de classificação, posto que o objetivo dos documentos do

conjunto de teste é representar documentos novos, desconhecidos, justamente para permitir avaliar o modelo construído.

Nos problemas de categorização de textos, cada termo distinto presente nos documentos de treinamento corresponde a um dos atributos usados para caracterizá-los.

2.3 Representação dos documentos

Para aplicação dos algoritmos de categorização, os documentos do corpus sob estudo devem ser representados em forma de um vetor, que é denominado vetor de atributos. Nessa modelagem, conhecida como modelo de espaço vetorial (*vector space model*), cada dimensão do vetor corresponde a um dos termos do dicionário de termos adotado, expresso como um peso que reflete a sua importância no problema de categorização considerado.

2.3.1 Dicionários de Termos

O dicionário de termos é construído a partir dos termos presentes nos documentos que formam o conjunto de treinamento, sendo que existem duas abordagens possíveis para a sua construção (APTÉ et al., 1994; Ng et al., 1997):

- Dicionário Local: um dicionário é construído por categoria temática, sendo cada dicionário composto apenas pelos termos presentes nos documentos do conjunto de treinamento pertencentes à respectiva categoria. Nessa abordagem, um mesmo documento terá representação vetorial distinta dependendo da categoria temática que o dicionário local de termos estiver representando.

- Dicionário Global: um único dicionário é construído e é composto pelos termos presentes em pelo menos um documento do conjunto de treinamento, considerando todas as categorias temáticas do problema sob estudo. Nessa situação, o mesmo documento terá a mesma representação, independentemente da categoria considerada.

Para entender as representações dos documentos por dicionário local e global, suponha um problema de classificação em três categorias - c_1, c_2 e c_3 - com cinco documentos no conjunto de treinamento e três no conjunto de teste. Considere que o interesse esteja em construir um classificador para a categoria c_1, isto é, um classificador que responda à pergunta se o documento deve ou não ser classificado em c_1. Nesse caso, as demais categorias do problema são representadas pela categoria não c_1, ou seja, $\bar{c}_1$. Os dados do problema são mostrados na Tabela 2.1.

Tabela 2.1 - Problema de classificação em três categorias com cinco documentos de treinamento e três de teste.

DOCUMENTO	CONJUNTO	TERMOS	CATEGORIA
doc_1	treinamento	term_3; term_5; term_8	c_1
doc_2	treinamento	term_2; term_3	c_1
doc_3	treinamento	term_1; term_5; term_7	c_2 ($\bar{c}_1$)
doc_4	treinamento	term_3; term_4	c_2 ($\bar{c}_1$)
doc_5	treinamento	term_2; term_4; term_5; term_6	c_3 ($\bar{c}_1$)
doc_6	teste	term_3; term_5; term_9	c_1
doc_7	teste	term_3; term_4; term_7; term_10	c_2 ($\bar{c}_1$)
doc_8	teste	term_2; term_4; term_11	c_3 ($\bar{c}_1$)

Para esse problema, considere a representação mais simples de um documento, onde a coordenada do vetor é zero, quando o correspondente termo do dicionário adotado está ausente no documento e é um, quando está presente (peso

Booleano). As matrizes de treinamento e teste para a categoria c_1, usando as abordagens por dicionário global e local, são mostradas em forma de tabela - para facilitar a visualização dos dados -, nas Tabelas 2.2 a 2.5 a seguir.

Tabela 2.2 - Representação dos documentos de treinamento usando dicionário global para a categoria c_1.

Termo / Documento	term_1	term_2	term_3	term_4	term_5	term_6	term_7	term_8
doc_1	0	0	1	0	1	0	0	1
doc_2	0	1	1	0	0	0	0	0
doc_3	1	0	0	0	1	0	1	0
doc_4	0	0	1	1	0	0	0	0
doc_5	0	1	0	1	1	1	0	0

Tabela 2.3 - Representação dos documentos de treinamento usando dicionário local para a categoria c_1.

Termo / Documento	term_2	term_3	term_5	term_8
doc_1	0	1	1	1
doc_3	1	1	0	0
doc_2	0	0	1	0
doc_4	0	1	0	0
doc_5	1	0	1	0

Tabela 2.4 - Representação dos documentos de teste usando dicionário global para a categoria c_1.

Termo / Documento	term_1	term_2	term_3	term_4	term_5	term_6	term_7	term_8
doc_6	0	0	1	0	1	0	0	0
doc_7	0	0	1	1	0	0	1	0
doc_8	0	1	0	1	0	0	0	0

Tabela 2.5 - Representação dos documentos de teste usando dicionário local para a categoria c_1.

Termo / Documento	term_2	term_3	term_5	term_8
doc_6	0	1	1	0
doc_7	0	1	0	0
doc_8	1	0	0	0

2.3.2 Pesos para os termos

Com relação à utilização dos pesos Booleanos para o problema de classificação apresentado na Tabela 2.1, Forman (2003) explica que em documentos curtos é improvável que as palavras se repitam, fazendo com que a utilização do peso Booleano seja adequada. Todavia, em documentos de uma forma geral, considerações sobre discriminação de termos sugerem que os melhores termos para a identificação do conteúdo de um documento são aqueles capazes de distinguir certos documentos dos documentos remanescentes da coleção.

Seguindo esse raciocínio, os termos mais importantes devem apresentar alta freqüência de ocorrência no documento - refletindo sua importância em relação aos demais termos utilizados no mesmo documento (peso local) -, porém, devem ocorrer em poucos documentos, pois um termo que ocorre em todos ou grande parte dos documentos da coleção não possui poder de discriminação (peso global). Uma medida razoável da importância de um termo pode então ser obtida calculando, por exemplo, o produto da freqüência do termo no documento pelo inverso da freqüência de documentos em que o termo ocorre (SALTON & BUCKLEY, 1988).

Baseada nessa idéia, a medida mais utilizada para atribuir peso aos termos, tanto na área de recuperação da informação como em categorização automática de

textos é denominada *TF_IDF* (T*erm Frequency Inverse Document Frequency*), que é calculada como segue.

Sejam:

- f_{ij}, a freqüência do termo t_j no documento $\mathbf{d}_i$;
- *n*, o número de documentos do conjunto de treinamento;
- n_j, o número de documentos do conjunto de treinamento em que o termo t_j ocorre.

$$TF\_IDF(t_j, \mathbf{d}_i) = f_{ij} \log \frac{n}{n_j}, \qquad (2.1)$$

onde o peso local f_{ij} é denominado $TF(t_j, \mathbf{d}_i)$ e o peso global $\log \frac{n}{n_j}$ é denominado $IDF(t_j)$, $i = 1, \cdots, n$, $j = 1, \cdots, p$.

Existem muitas variantes da fórmula TF_IDF. Em estudo realizado por Tantrum (2003), aplicando o classificador *k-nearest neighbor* (*k*-NN) em 1.131 notícias das agências Reuters e CNN classificadas em 25 tópicos pré-selecionados, as representações para os pesos dos termos que apresentaram as maiores efetividades (menor taxa de erro) na classificação foram:

- $\log f_{ij} \log \frac{n}{n_j}$
- $\sqrt{f_{ij}} \log \frac{n}{n_j}$

Tantrum (2003) explica que a substituição de f_{ij} por $\sqrt{f_{ij}}$ ou por $\log f_{ij}$ é adotada como forma de reduzir a influência de altas contagens, uma vez que a diferença entre um termo que ocorre dez vezes em um documento contra um que ocorre onze

vezes não é tão significativa quanto a diferença entre um termo que ocorre uma vez e um outro que não ocorre no documento.

Debole & Sebastiani (2003) e Lewis el al. (2004) utilizaram em seus estudos a variação de TF_IDF dada por

$$TF\_IDF(t_j, \mathbf{d}_i) = \begin{cases} (1+\log f_{ij}) \log \dfrac{n}{n_j} & \text{se} \quad f_{ij} > 0 \\ 0 & \text{caso contrário,} \end{cases} \qquad (2.2)$$

onde a soma do valor um ao logaritmo da freqüência do termo no documento em (2.2) evita atribuir peso zero aos termos que ocorrem apenas uma vez no documento.

Debole & Sebastiani (2003) ponderam que a forma de atribuição de pesos aos termos TF_IDF, vinda da área de recuperação da informação (IR), não representa uma escolha ótima para atribuir pesos aos termos na área de categorização automática de textos. Eles esclarecem que nos contextos padrões de IR, como não há disponibilidade de documentos cuja relevância ou irrelevância para uma consulta q seja conhecida, a utilização do termo *IDF* na composição TF_IDF é razoável. Nessa situação *IDF* codifica a intuição bastante plausível de que se um termo, presente em uma consulta q, também estiver presente em muitos documentos da base de dados, ele não é suficientemente útil para discriminar os documentos relevantes para q dos irrelevantes.

Em categorização automática de textos, ao contrário do que ocorre em IR, existe disponível o conhecimento sobre a pertinência dos documentos às categorias, uma vez que o pré-requisito para construção do modelo de categorização consiste em ter disponível um conjunto de documentos pré-classificados nas diversas categorias da área de interesse.

Assim, visando melhor discriminar os documentos pertencentes às diversas categorias temáticas, Debole & Sebastiani (2003) criaram o conceito de atribuição supervisionada de pesos (*Supervised Term Weighting* (STW)), no qual a informação sobre a pertinência dos documentos de treinamento às categorias é incorporada ao processo de atribuição de pesos aos termos. Essa incorporação é realizada a partir da utilização dos escores $f(t_j, c_k)$, que medem a capacidade do termo t_j discriminar as categorias c_k, $k = 1, \cdots, m$, e que são calculados por meio de um dos métodos de seleção de termos que serão vistos na próxima seção.

A utilização de STW permite atribuir maior peso aos termos com melhor poder de discriminação das categorias, segundo os princípios em que se baseiam os métodos de seleção de termos a serem apresentados.

No caso de adoção do dicionário local, o método de atribuição supervisionada de pesos substitui, na fórmula $TF\_IDF(t_j, \mathbf{d}_i)$, o termo $IDF(t_j)$ por $f(t_j, c_k)$. Considerando que o valor de $IDF(t_j)$ a ser substituído é o da expressão (2.2), a medida de atribuição supervisionada de pesos é calculada por:

$$TF\_F(t_j, \mathbf{d}_i, c_k) = \begin{cases} (1 + \log f_{ij}) f(t_j, c_k) & \text{se} \quad f_{ij} > 0 \\ 0 & \text{caso contrário} \end{cases} \qquad (2.3)$$

Na adoção do dicionário global, para acessar a "importância" do termo t_j na discriminação das categorias é necessário obter uma medida global que resume os escores calculados para o termo t_j nas diversas categorias temáticas c_k, $k = 1, \cdots, m$.

Para definir tal medida, sejam:

- n_{c_k} : número de documentos do conjunto de treinamento classificados na categoria c_k ;
- n : número de documentos do conjunto de treinamento.

Uma das medidas globais apresentadas em (2.4) a (2.6) a seguir pode ser usada:

$$f_{soma}(t_j) = \sum_{k=1}^{m} f(t_j, c_k) \tag{2.4}$$

$$f_{somapond}(t_j) = \sum_{k=1}^{m} \frac{n_{c_k}}{n} f(t_j, c_k) \tag{2.5}$$

$$f_{\max}(t_j) = \max_{k=1,\cdots,m} f(t_j, c_k) \tag{2.6}$$

Entre essas medidas globais, a $f_{\max}(t_j)$ é a que tem produzido os melhores resultados em efetividade na categorização (YANG & PEDERSEN, 1997; DEBOLE & SEBASTIANI, 2003; LEWIS ET AL, 2004). De acordo com Debole & Sebastiani (2003), a razão para tal fato é que essa medida prefere termos que são bons separadores mesmo que seja em uma única categoria a termos que são apenas separadores justos em muitas categorias.

Substituindo em (2.2) o termo $IDF(t_j)$ por uma das medidas globais apresentadas em (2.4) a (2.6), obtém-se uma medida de atribuição supervisionada de pesos com dicionário global. Usando, por exemplo, a fórmula (2.6), essa medida é calculada como:

$$TF\_F(t_j, \mathbf{d}_i, c_1, \cdots, c_m) = \begin{cases} (1+\log f_{ij})\, f_{\max}(t_j) & \text{se} \quad f_{ij} > 0 \\ 0 & \text{caso contrário} \end{cases} \qquad (2.7)$$

Independentemente da medida de atribuição de pesos adotada, em geral, costuma-se normalizar os vetores correspondentes a cada documento de forma a terem cumprimento unitário $\left(L_x = \sqrt{\mathbf{x}^t \mathbf{x}} = 1 \right)$. Essa normalização é utilizada para permitir comparar em bases iguais documentos longos e curtos. Com essa normalização os pesos individuais dos termos nos documentos caem no intervalo [0,1] e dependem de alguma forma dos pesos dos outros termos no mesmo vetor.

Considerando, por exemplo, a fórmula *TF_IDF* dada em (2.2), a fórmula normalizada fica:

$$x_{ij} = \frac{(1+\log f_{ij}) \log \dfrac{n}{n_j}}{\sqrt{\displaystyle\sum_{j=1}^{p} \left((1+\log f_{ij}) \log \frac{n}{n_j} \right)^2}}.$$

Em notação vetorial, o *i*-ésimo documento é representado como um vetor de pesos

$$\mathbf{x}_i^t = \left[x_{i1}, \quad \cdots \quad , x_{ip} \right],$$

onde p corresponde ao número de termos do dicionário de termos considerado e $0 \leq x_{ij} \leq 1$ é o valor que representa o quanto o j-ésimo termo contribui para a semântica do documento $\mathbf{x}_i$.

A matriz de treinamento de documento por termo é dada por:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

2.3.3 Comentários

- Como o papel do conjunto de teste é representar documentos novos, desconhecidos, os documentos desse conjunto não são utilizados no cálculo de $IDF(t_j)$ ou $f(t_j, c_k)$.
- Utiliza-se como valor de $IDF(t_j)$ para o termo t_j dos documentos do conjunto de teste, o valor de $IDF(t_j)$ calculado com base nos documentos de treinamento; da mesma forma, utiliza-se o valor de $f(t_j, c_k)$ (de uma das funções dadas em (2.4) a (2.6)) calculadas a partir dos documentos de treinamento, no cálculo de $TF\_F(t_j, \mathbf{d}_i, c_k)$ ($TF\_F(t_j, \mathbf{d}_i, c_1, \cdots, c_m)$) para o termo t_j dos documentos do conjunto de teste.
- Apté et al. (1994), utilizando um classificador baseado em regras de decisão (sistema especialista) no corpus Reuters-22173, compararam as abordagens por dicionário global e local, verificando melhor efetividade na categorização usando dicionário local.

- A utilização do dicionário local em comparação ao dicionário global também apresentou melhor efetividade na categorização no estudo de Ng et al. (1997) com o corpus Reuters-22173, porém usando o algoritmo perceptron.
- No estudo conduzido por Debole & Sebastiani (2003), aplicando o algoritmo *support vector machines* (SVMs) no corpus Reuters-21578, "ModApté", os resultados por dicionário global apresentaram melhor efetividade na categorização do que por dicionário local. Eles compararam a utilização dos pesos calculados por *TF_IDF* - expressão (2.2) - com os calculados por STW usando dicionário local (expressão (2.3)) e dicionário global (expressão (2.7)). Nas expressões (2.3) e (2.7), foram usadas as funções de seleção de termos qui-quadrado, ganho de informação e razão de ganho. Os melhores resultados foram obtidos com *TF_IDF* e STW usando razão de ganho e qui-quadrado com dicionário global.

2.4 Redução da dimensionalidade

Uma propriedade dos problemas de categorização automática de textos é a alta dimensionalidade do espaço de atributos, que corresponde ao número de termos distintos presentes nos documentos do conjunto de treinamento, e que frequentemente excede em muito o número de documentos desse conjunto. Esse valor pode chegar a milhares de termos mesmo para uma coleção de documentos de tamanho moderado.

No corpus Reuters-21578, distribuição 1.0, "ModApte", por exemplo, considerando as 90 categorias com pelo menos um documento de treinamento e um de teste foram gerados 27.658 termos - após remoção de *stop words* e aplicação de *stemming* - para um conjunto de treinamento com 9.603 documentos (JOACHIMS,

2005). No corpus Reuters volume 1, versão RCV1-v2, foram gerados 47.152 termos - após remoção de *stop words* e aplicação de *stemming* - para um conjunto de treinamento com 23.149 documentos (LEWIS et al., 2004).

Sebastiani (2002) e Joachims (1998) comentam que o número elevado de atributos é problemático para muitos dos algoritmos de categorização. Sebastiani (2002) ainda acrescenta que a redução da dimensionalidade do espaço de atributos tende a reduzir o problema de *overfitting*[9].

Também esclarece Forman (2003), que atributos bem escolhidos podem melhorar a efetividade na classificação, fato confirmado, por exemplo, nos experimentos realizados por Yang & Pedersen (1997), Mladenic & Grobeonik (1999) e Forman (2003).

Desse modo, com o intuito de melhorar a efetividade na classificação, a eficiência computacional ou ambas as coisas, muitas vezes é executada a redução de dimensionalidade nos problemas de categorização automática de textos. Nessa área, a redução da dimensionalidade do espaço de atributos é feita de duas formas:

- Redução da dimensionalidade por seleção de termos: os termos escolhidos formam um subconjunto dos *p* termos originais;
- Redução da dimensionalidade por extração de termos: os termos escolhidos são termos sintéticos, criados a partir de combinações ou transformações dos termos originais, não formando, portanto, um subconjunto dos *p* termos originais.

As duas formas são apresentadas nas subseções seguintes.

[9] Situação onde os classificadores se ajustam excessivamente bem aos dados de treinamento, e assim tendem a ser extremamente bons para classificar esses dados nos quais foram treinados, mas são notavelmente piores na classificação de dados diferentes desses.

2.4.1 Redução da dimensionalidade por seleção de termos

Nesta subseção são descritos os métodos mais utilizados para seleção de termos. Os métodos apresentados são baseados na abordagem por filtro, onde os termos são selecionados independentemente do algoritmo de categorização que os utilizará. Essa abordagem é adotada por ser a que apresenta maior eficiência computacional com os problemas típicos de categorização automática de textos.

Com exceção do método baseado na freqüência de documentos, os demais métodos empregam princípios que se baseiam na intuição de que os melhores termos para discriminar a categoria c_k são aqueles cuja distribuição de freqüências nas categorias c_k e $\bar{c}_k$ se apresenta da forma mais diferente possível.

Para medir o poder de discriminação de um termo t_j, $j = 1, \cdots, p$, com relação à categoria c_k, $k = 1, \cdots, m$, um escore $f(t_j, c_k)$ é calculado usando um desses métodos de seleção de termos - exceto freqüência de documentos. Os maiores escores são atribuídos aos termos com maior poder de discriminação das categorias.

Considerando dicionário local, a seleção de termos é feita com base nos escores calculados para os termos representados no dicionário. Com dicionário global, como uma única representação é obtida para todas as categorias temáticas, os termos são selecionados com base em seus escores globais (expressões (2.4) a (2.6)).

Para selecionar os “melhores” termos, independentemente da representação usada para os documentos, o primeiro passo é ordenar os escores calculados em forma decrescente. Partindo da dimensão completa *p*, reduz-se o número de termos selecionados, considerando vários pontos de corte, até no máximo o ponto de corte onde nenhum documento é eliminado (YANG & PEDERSEN,1997). Em cada ponto de

corte, o modelo de categorização é estimado e avaliado com relação à sua efetividade na categorização. O objetivo é selecionar os termos que quando usados em conjunto fornecem a maior efetividade possível na categorização.

Os métodos são apresentados a seguir, lembrando que todos os cálculos são efetuados usando apenas os documentos do conjunto de treinamento.

2.4.1.1 Freqüência de documentos (DF)

O método de seleção de termos pela freqüência de documento (d*ocument frequency* (DF)) seleciona os termos com base no número de documentos em que o termo está presente. Esse método é considerado uma abordagem *ad hoc* para melhorar a eficiência dos algoritmos de categorização, não um critério baseado em princípios para selecionar termos preditivos (YANG & PEDERSEN,1997).

Yang & Pedersen (1997) consideram a utilização deste método para a remoção de termos que ocorrem em poucos documentos do conjunto de treinamento, esclarecendo que a suposição básica do método é que termos raros ou não são informativos para a predição das categorias ou não influenciam no desempenho global dos algoritmos de categorização.

De acordo com Forman (2003), termos raros podem ser eliminados uma vez que são improváveis de estarem presentes para auxiliar em classificações futuras.

Yang & Pedersen (1997) esclarecem que esse método não é tipicamente usado para remoção agressiva de termos devido à suposição amplamente aceita na área de recuperação da informação de que termos que ocorrem em um número baixo de documentos são relativamente informativos e, portanto, não devem ser removidos agressivamente.

Sebastiani (2002) comenta que vários pesquisadores removem todos os termos que ocorrem em no máximo r documentos de treinamento (valores usuais de r variando entre 1 a 3) como a única forma de redução de termos ou como a forma usada antes da aplicação de métodos mais sofisticados de redução de termos.

Schweighofer et al (2001) e Forman (2003) consideram a utilização deste método também para a remoção de termos que ocorrem em muitos documentos do conjunto de treinamento. Eles explicam que embora uma lista de *stop words* elaborada manualmente permita a exclusão de palavras específicas utilizadas com freqüência, a remoção das palavras presentes em muitos documentos da coleção (por exemplo, mais de 50%), representa uma forma mais automática para alcançar o mesmo propósito.

2.4.1.2 Qui-quadrado (QUI)

Considerando a distribuição conjunta das freqüências de documentos segundo o termo t_j e a categoria temática c_k, a estatística χ^2 (qui-quadrado (QUI)) mede o grau de afastamento entre a distribuição de freqüências observadas e a de freqüências esperadas sob a hipótese de independência entre o termo e a categoria. Quanto maior o valor de χ^2 maior a associação existente entre o termo e a categoria. A medida qui-quadrado é dada por:

$$\chi^2(t_j, c_k) = \frac{n(AD - BC)^2}{(A+C)(B+D)(A+B)(C+D)}. \qquad (2.8)$$

Para entender o raciocínio envolvido no cálculo da medida, considere as Tabelas 2.6 e 2.7 a seguir.

Tabela 2.6 - Distribuição conjunta das freqüências observadas de documentos segundo o termo t_j e a categoria c_k.

termo \ categoria	c_k	$\bar{c}_k$	Total
t_j	A	B	(A+B)
$\bar{t}_j$	C	D	(C+D)
Total	(A+C)	(B+D)	n

Tabela 2.7 - Distribuição conjunta das freqüências esperadas segundo o termo t_j e a categoria c_k, sob a hipótese de que um documento conter ou não o termo t_j independe da categoria temática do documento ser ou não c_k.

Termo \ Categoria	c_k	$\bar{c}_k$	Total
t_j	$\frac{(A+B)(A+C)}{n}$	$\frac{(A+B)(B+D)}{n}$	(A+B)
$\bar{t}_j$	$\frac{(C+D)(A+C)}{n}$	$\frac{(C+D)(B+D)}{n}$	(C+D)
Total	(A+C)	(B+D)	n

onde:

- A: número de documentos da categoria c_k que contém o termo t_j;
- B: número de documentos que não são da categoria c_k e que contém o termo t_j;
- C: número de documentos da categoria c_k que não contém o termo t_j;
- D: número de documentos que não são da categoria c_k e que não contém o termo t_j;
- n: número total de documentos de treinamento.

A Tabela 2.7 é obtida a partir da Tabela 2.6, de acordo com a explicação que segue.

Da Tabela 2.6, pode-se verificar que a proporção de documentos com o termo t_j é $\frac{(A+B)}{n}$ (estimativa da probabilidade do termo t_j estar presente em um documento) e a proporção de documentos da categoria c_k é $\frac{(A+C)}{n}$ (estimativa da probabilidade de um documento pertencer à categoria c_k).

Se a presença do termo t_j em um documento fosse independente[10] da categoria temática do documento ser c_k, dos n documentos do conjunto de treinamento esperaríamos observar a proporção de $\frac{(A+B)}{n}\frac{(A+C)}{n}$ documentos da categoria c_k com o termo t_j, ou seja, $n\frac{(A+B)}{n}\frac{(A+C)}{n}$ documentos (freqüência esperada de documentos sob a hipótese de o documento conter o termo t_j independer da categoria do documento ser c_k).

De forma análoga, se a ausência do termo t_j em um documento fosse independente da categoria temática do documento ser c_k, esperaríamos $n\frac{(C+D)}{n}\frac{(A+C)}{n}$ documentos da categoria c_k sem o termo t_j (freqüência esperada de documentos sob a hipótese de o documento não conter o termo t_j independer da categoria do documento ser c_k).

O mesmo raciocínio segue para a obtenção das freqüências esperadas das demais caselas da Tabela 2.7.

[10] Se dois eventos R e S são independentes, então a probabilidade de ocorrência simultânea de R e S pode ser escrita como o produto das probabilidades de ocorrência de R e de S, ou seja, $P(R \cap S) = P(R)P(S)$.

A medida qui-quadrado é calculada como:

$$\chi^2(t_j, c_k) = \frac{\left(A - \frac{(A+B)(A+C)}{n}\right)^2}{\frac{(A+B)(A+C)}{n}} + \frac{\left(B - \frac{(A+B)(B+D)}{n}\right)^2}{\frac{(A+B)(B+D)}{n}} + \frac{\left(C - \frac{(C+D)(A+C)}{n}\right)^2}{\frac{(C+D)(A+C)}{n}} + \frac{\left(D - \frac{(C+D)(B+D)}{n}\right)^2}{\frac{(C+D)(B+D)}{n}} \quad (2.9)$$

Cada parcela da soma na expressão (2.9) é do tipo $\frac{(O-E)^2}{E}$, onde "O" corresponde à freqüência observada na casela considerada e "E" refere-se à correspondente freqüência esperada sob a hipótese de independência entre o termo e a categoria. Se as freqüências observadas são próximas das esperadas, naturalmente essas diferenças serão pequenas, e conseqüentemente o valor da estatística qui-quadrado também será pequeno.

A estatística qui-quadrado assume o valor zero quando a presença ou ausência do termo t_j ocorre de forma independente da categoria temática ser ou não c_k. Por outro lado, quanto maior a diferença entre as freqüências observadas e esperadas, maior será o valor da estatística e tanto maior será a probabilidade de que a presença ou ausência do termo t_j esteja relacionada com a categoria temática ser ou não c_k. A expressão (2.8) é resultante da expressão (2.9), após simplificações.

2.4.1.3 Ganho de informação (IG)

O ganho de informação (*information gain* (IG)) é uma medida da teoria da informação usada na área de categorização de textos para medir o número de bits de informação obtidos para a predição de uma categoria c_k a partir do conhecimento de que um termo t_j está presente ou ausente em um documento. Quanto maior o ganho de informação mais informativo é o termo para predição da categoria.

O ganho de informação para a categoria c_k com relação ao termo t_j é definido como:

$$IG(t_j, c_k) = H(c_k) - H(c_k / t_j), \qquad (2.10)$$

onde:

- a expressão[11]:

$$H(c_k) = -P(c_k)\log_2 P(c_k) - P(\bar{c}_k)\log_2 P(\bar{c}_k) \qquad (2.11)$$

  representa a entropia para a categoria c_k, e corresponde à quantidade de informação necessária, em média, para classificar um documento em uma de duas categorias temáticas: c_k ou não c_k ($\bar{c}_k$). Quanto maior a entropia, ou incerteza, com relação à categoria c_k, mais informação é necessária para classificar um documento nela;

[11] A função logarítmica na base 2 é usada, pois a informação está codificada em bits.

- a expressão:

$$H(c_k / t_j) = P(t_j)\left[-P(c_k / t_j)\log_2 P(c_k / t_j) - P(\bar{c}_k / t_j)\log_2 P(\bar{c}_k / t_j)\right] + P(\bar{t}_j)\left[-P(c_k / \bar{t}_j)\log_2 P(c_k / \bar{t}_j) - P(\bar{c}_k / \bar{t}_j)\log_2 P(\bar{c}_k / \bar{t}_j)\right] \quad (2.12)$$

representa a entropia condicional da categoria c_k dado o termo t_j, e corresponde à quantidade de informação necessária, em média, para classificar um documento nas categorias temáticas c_k ou não c_k ($\bar{c}_k$), quando é conhecido que o termo t_j está presente ou ausente em um documento. Quanto menor $H(c_k / t_j)$, menos informação é necessária, em média, para classificar um documento nas categorias c_k ou não c_k ($\bar{c}_k$), a partir do conhecimento de que o termo t_j está presente ou não no documento.

Assim, o ganho de informação calcula a redução na entropia, o ganho em informação, resultante do conhecimento de que um determinado termo t_j está presente ou ausente em um documento.

Nas expressões (2.11) e (2.12):

- $P(c_k)$ $\left(P(\bar{c}_k)\right)$ é a probabilidade de um documento pertencer (não pertencer) à categoria c_k;
- $P(t_j)$ $\left(P(\bar{t}_j)\right)$ é a probabilidade do termo t_j estar presente (ausente) em um documento;
- $P(c_k / t_j)$ $\left(P(\bar{c}_k / t_j)\right)$ é a probabilidade da categoria temática de um documento ser c_k (não ser c_k) dado que o termo t_j está presente no documento;

- $P(c_k / \bar{t}_j)$ $\left(P(\bar{c}_k / \bar{t}_j)\right)$ é a probabilidade da categoria temática de um documento ser c_k (não ser c_k) dado que o termo t_j não está presente no documento.

As probabilidades necessárias para o cálculo de $IG(t_j, c_k)$ são estimadas com base nos dados relativos às freqüências observadas dispostas na Tabela 2.6. Portanto:

- $\hat{P}(c_k) = \frac{(A+C)}{n}$; $\hat{P}(\bar{c}_k) = 1 - \hat{P}(c_k)$
- $\hat{P}(t_j) = \frac{(A+B)}{n}$; $\hat{P}(\bar{t}_j) = 1 - \hat{P}(t_j)$
- $\hat{P}(c_k / t_j) = \frac{A}{(A+B)}$; $\hat{P}(\bar{c}_k / t_j) = 1 - \hat{P}(c_k / t_j)$
- $\hat{P}(c_k / \bar{t}_j) = \frac{C}{(C+D)}$; $\hat{P}(\bar{c}_k / \bar{t}_j) = 1 - \hat{P}(c_k / \bar{t}_j)$.

2.4.1.4 Razão de ganho (GR)

A medida razão de ganho (gain ratio (GR)), na área de categorização automática de textos, é definida como a razão entre o ganho de informação da categoria c_k com relação ao termo t_j (expressão (2.10)) e a entropia da categoria c_k (expressão (2.11)), isto é,

$$GR(t_j, c_k) = \frac{IG((t_j, c_k)}{H(c_k)}. \qquad (2.13)$$

Debole & Sebastiani (2003) esclarecem que o ganho de informação (expressão (2.10)) aumenta não apenas com o grau de dependência entre o termo e a

categoria, mas também com as suas entropias ($H(c_k)$ e $H(t_j)$). Assim, a divisão do ganho de informação pela entropia da categoria temática permite comparar em bases iguais os escores do termo t_j obtidos nas diversas categorias temáticas, pois enquanto

$0 \leq IG(t_j, c_k) \leq \min\{H(t_j), H(c_k)\}$,

$0 \leq GR(t_j, c_k) \leq 1$.

Quanto maior a razão de ganho mais informativo é o termo para predição da categoria.

2.4.1.5 Razão de chances (OR)

A medida razão de chances (*Odds Ratio* (OR)) representa a chance de um documento da categoria c_k conter o termo t_j sobre a chance de um documento que não é da categoria c_k - é da categoria $\bar{c}_k$ - conter o termo t_j. Quanto maior a razão de chances mais chance tem um documento da categoria c_k de conter o termo t_j do que um documento da categoria $\bar{c}_k$. A razão de chances é definida como:

$$OR(t_j, c_k) = \frac{P(t_j / c_k) \,/\, (1 - P(t_j / c_k))}{P(t_j / \bar{c}_k) \,/\, (1 - P(t_j / \bar{c}_k))}, \qquad (2.14)$$

onde:

- $P(t_j / c_k)$ é a probabilidade de um documento da categoria c_k conter o termo t_j; $(1 - P(t_j / c_k))$ é a probabilidade de um documento da categoria c_k não conter o termo t_j ;
- $P(t_j / \bar{c}_k)$ é a probabilidade de um documento que não é da categoria c_k conter o termo t_j; $(1 - P(t_j / \bar{c}_k))$ é a probabilidade de um documento que não é da categoria c_k não conter o termo t_j.

A razão $\frac{P(t_j / c_k)}{1 - P(t_j / c_k)}$ é definida como a chance de um documento da categoria c_k conter o termo t_j. Quando $P(t_j / c_k) > 0.5$, a chance de um documento da categoria c_k conter o termo t_j é maior que a chance de não conter. De forma análoga, a chance de um documento que não é da categoria c_k - é da categoria $\bar{c}_k$ - conter o termo t_j é definida como $\frac{P(t_j / \bar{c}_k)}{1 - P(t_j / \bar{c}_k)}$. Quando $P(t_j / \bar{c}_k) > 0.5$, a chance de um documento que não é da categoria c_k - é da categoria $\bar{c}_k$ - conter o termo t_j é maior que a chance de não conter.

Assim, quando:

- $OR(t_j, c_k)$ > 1, a chance de um documento da categoria c_k conter o termo t_j é maior que a chance de um documento que não é da categoria c_k conter.
- $OR(t_j, c_k)$ = 1, a chance de um documento da categoria c_k conter o termo t_j é a mesma de um documento que não é da categoria c_k conter.

- $OR(t_j, c_k) < 1$, a chance de um documento da categoria c_k conter o termo t_j é menor que a chance de um documento que não é da categoria c_k conter.

Suponha que $OR(t_j, c_k)$=2. Esse valor indica que os documentos da categoria c_k têm duas vezes mais chance de conter o termo t_j do que os documentos que não são da categoria c_k (são da categoria $\bar{c}_k$).

As probabilidades necessárias para o cálculo de $OR(t_j, c_k)$ são estimadas com base nos dados relativos às freqüências observadas dispostas na Tabela 2.6. Desse modo:

- $\hat{P}(t_j / c_k) = \frac{A}{(A+C)}$; $(1 - \hat{P}(t_j / c_k)) = \frac{C}{(A+C)}$

- $\hat{P}(t_j / \bar{c}_k) = \frac{B}{(B+D)}$; $(1 - \hat{P}(t_j / \bar{c}_k)) = \frac{D}{(B+D)}$.

Substituindo as quantidades acima na expressão (2.14) e somando 0.5 a cada uma das freqüências observadas da Tabela 2.6, a razão de chances é estimada como:

$$\hat{OR}(t_j, c_k) = \frac{(A+0.5)(D+0.5)}{(B+0.5)(C+0.5)}. \qquad (2.15)$$

Na expressão (2.15), a constante 0.5 – valor mais utilizado - é somada a cada uma das freqüências observadas da Tabela 2.6 a fim de evitar problemas numéricos na estimação de $OR(t_j, c_k)$ (AGRESTI, 1988). Eyheramendy & Madigan (2005) utilizaram a constante 0.1 em seu estudo.

2.4.1.6 Abs-logito (ABSL)

A medida abs-logito (ABSL) é definida como o valor absoluto do logaritmo natural da razão de chances. Quanto maior o valor assumido pela medida, mais diferentes são as chances de documentos das categorias c_k e $\bar{c}_k$ conterem o termo t_j.

$$ABSL(t_j, c_k) = \left|\ln(OR(t_j, c_k))\right| \qquad (2.16)$$

Da descrição da medida razão de chances, pode-se verificar que quando a chance de um documento da categoria c_k conter o termo t_j é maior que a chance de um documento da categoria $\bar{c}_k$ conter, $OR(t_j, c_k)$ varia de 1.001 até infinito. Porém, quando a chance de um documento da categoria $\bar{c}_k$ conter o termo t_j é maior do que a chance de um documento da categoria c_k conter, $OR(t_j, c_k)$ varia apenas de 0 a 0.999 (considerando três casas decimais).

Essa assimetria representa um inconveniente para a utilização da medida razão de chances como medida de força de associação entre atributos. Porém, o problema pode ser contornado, aplicando a transformação logarítmica à razão de chances, obtendo assim uma medida de associação simétrica, que é denominada logito (utilizada em estatística).

Como o interesse em seleção de termos está nos termos cuja distribuição de freqüências nas categorias c_k e $\bar{c}_k$ se apresenta da forma mais diferente possível, não importando se o termo é mais prevalente na categoria c_k ou na $\bar{c}_k$, a medida adequada para essa consideração é a utilização do valor absoluto do logito (logaritmo da razão de chances), conforme dado em (2.16).

2.4.1.7 *Bi-normal separation* (BNS)

A medida Bi-Normal Separation(BNS) é estimada como:

$$BNS(t_j, c_k) = \left| \Phi^{-1}\left(\frac{A}{A+C}\right) - \Phi^{-1}\left(\frac{B}{B+D}\right) \right|, \qquad (2.17)$$

onde Φ é a função de distribuição da normal padrão (normal com média zero e variância 1) e Φ^{-1} é a sua inversa. Isto é:

$$\Phi(x) = p_1 = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

e

$\Phi^{-1}(p_1) = x$, ou seja, x é tal que $P(X \leq x) = p_1$.

Quanto maior o valor assumido pela medida, maior a indicação de diferença entre as prevalências do termo t_j nas categorias c_k e $\bar{c}_k$.

Como pode ser verificado na Tabela 2.6, $A/_{A+C}$ corresponde à proporção de documentos da categoria c_k com o termo t_j (prevalência do termo t_j na categoria c_k) e $B/_{B+D}$ corresponde à proporção de documentos da categoria $\bar{c}_k$ ("outras categorias temáticas") com o termo t_j (prevalência do termo t_j na categoria $\bar{c}_k$).

Considere a representação gráfica da distribuição normal padrão na figura a seguir.

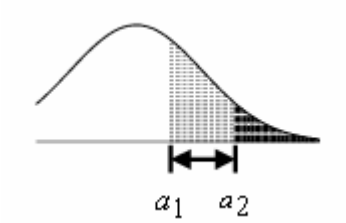


Figura 2.1 - Distribuição normal padrão na situação onde $a_1 < a_2$.

Na Figura 2.1, seja a_1 tal que $P(X > a_1) = A/A+C$ (área hachurada à direita de a_1) e a_2 tal que $P(X > a_2) = B/B+D$ (área hachurada à direita de a_2). A medida $BNS(t_k, c_i)$ compara os valores a_1 e a_2 calculados considerando a distribuição normal padrão. Quanto maior a diferença entre esses dois valores maior a indicação de diferença entre a prevalência do termo t_j nas categorias c_k e $\bar{c}_k$ ("outras categorias temáticas").

Para evitar problemas computacionais no cálculo da métrica, $\Phi^{-1}(0)$ (que é igual a $-\infty$) é substituído por $\Phi^{-1}(0.00001)$. Forman (2003) substitui o valor zero por 0.0005.

2.4.1.8 Comentários

Estudos específicos sobre métricas para seleção de termos podem ser encontrados em Yang & Pedersen (1997), Mladenic & Grobelnik (1999) e Forman (2003).

Os métodos de seleção de termos comparados por Yang & Pederson (1997) foram: freqüência de documentos, ganho de informação, informação mútua, qui-quadrado e força do termo; os comparados por Mladenic & Grobelnik (1999) foram:

ganho de informação, entropia cruzada, informação mútua, freqüência de documentos, peso de evidência, razão de chances, razão de chances ponderada, logaritmo da razão de probabilidades, diferença esperada de probabilidades, razão de chances condicional e seleção aleatória; os comparados por Forman (2003) foram: acurácia esperada, acurácia esperada balanceada, *bi-normal separation*, qui-quadrado, freqüência de documentos, F_1, ganho de informação, numerador da razão de chances, razão de chances, potência, razão de probabilidades e seleção aleatória de termos.

No estudo de Yang & Pedersen (1997), aplicando os algoritmos de classificação *k*-NN (*k nearest neighbor*) e LSSF (*Linear Least Squares Fit Mapping*) - um método de regressão - nos corpora Reuters-22173 e OHSUMED, usando pesos calculados por *TF_IDF*, as métricas que apresentaram os melhores resultados foram: freqüência de documentos, qui-quadrado e ganho de informação. Para os diversos pontos de corte considerados para o número de termos selecionados, a medida qui-quadrado - escore global calculado pelo máximo (expressão (2.6)) - apresentou desempenho geral um pouco melhor que as outras duas medidas citadas. No entanto, para uma redução agressiva de termos - baixo número de termos considerados -, o ganho de informação apresentou os melhores resultados.

Mladenic & Grobelnik (1999) aplicaram o algoritmo Naïve Bayes Multinomial a cinco áreas dispostas em categorias hierárquicas no diretório Yahoo!. Os melhores resultados foram obtidos com a métrica razão de chances e suas variantes, e os piores com as métricas ganho de informação e seleção aleatória de termos.

Forman (2003), usando peso Booleano, aplicou o algoritmo *support vector machines* a 229 problemas de classificação em duas classes, criados com textos originados dos corpora Reuters-21578, TREC, OHSUMED, WebACE e CORA. Os melhores resultados gerais foram obtidos pela métrica *bi-normal separation*,

principalmente para distribuições com alta assimetria entre as classes[12]. Considerando um número reduzido de termos selecionados, assim como no estudo realizado por Yang & Pedersen (1997), o ganho de informação, em geral, apresentou os melhores resultados.

Em seu estudo, Yang & Pedersen (1997) observaram, nos dois corpora utilizados, alta correlação entre os escores calculados pelas métricas qui-quadrado, ganho de informação e freqüência de documentos. Forman (2003) também observou correlação positiva nos resultados obtidos pelas métricas qui-quadrado e ganho de informação.

Nos estudos comentados, todos os termos distintos presentes em pelo menos um documento de treinamento foram considerados para seleção. Porém, apenas no estudo de Yang & Pedersen (1997) os termos foram selecionados definindo um escore global que resume a importância do termo na discriminação geral das categorias. Neste caso, definindo um ponto de corte para o número de termos selecionados, os mesmos termos foram selecionados em todos os problemas de classificação binária considerados (categoria de interesse x demais categorias). Nos outros estudos, em cada problema de classificação na categoria de interesse, apenas os escores obtidos com relação a essa categoria foram considerados para seleção, isto é, a seleção de termos foi considerada separadamente em cada problema de classificação binária.

A diferença entre o dicionário usado por Mladenic & Grobelnik (1999) e por Forman (2003) e o dicionário local considerado neste trabalho é que neste, apenas os termos presentes nos documentos de treinamento pertencentes à categoria de interesse compõem o dicionário e naquele, os termos presentes em pelo menos um documento de treinamento compõem o dicionário. Todavia, em ambos os casos, a

[12] Em seu estudo, Forman (2003) considerou como altamente assimétricos os problemas com um documento na classe de interesse para pelo menos 67 na outra classe.

seleção de termos é feita separadamente em cada categoria, diferentemente do dicionário global onde os termos são selecionados definindo um escore global que resume a importância do termo na discriminação geral das categorias.

2.4.2 Redução da dimensionalidade por extração de termos

Na redução da dimensionalidade por extração de termos, a técnica, em geral, utilizada é denominada indexação semântica latente (*Latent Semantic Indexing* (LSI)). Por meio dessa técnica, a redução do espaço de atributos - redução do número *p* de termos usados - é obtida usando uma aproximação $\tilde{\mathbf{X}}$ (matriz $n \, x \, k$) para a matriz **X** (matriz $n \, x \, \mathrm{p}$) de documentos por termos fatorada por decomposição em valores singulares (*Singular Value Decomposition* (SVD)).

A partir da aplicação dessa técnica um conjunto *k* (*k*<*p*) de fatores (novos termos) não-correlacionados é derivado. Esses fatores podem ser vistos como "conceitos artificiais" que representam grupos de termos similares.

O propósito em indexação semântica latente é tentar modelar explicitamente as relações entre os termos em um espaço de dimensão menor do que a original, onde inter-relações úteis entre os termos, caso existam, possam tornar-se mais evidentes.

Considere, por exemplo, as palavras "carro" e "automóvel". Como essas palavras ocorrem juntas com muitas outras palavras, como motor, modelo, veículo, chassi, fabricante, etc., elas devem compartilhar pelo menos um "conceito sintético" em comum. Assim, documentos contendo as palavras carro e automóvel deverão ter representações similares nesse novo espaço de menor dimensão. Nesse novo espaço, documentos com o termo motorista também deverão compartilhar uma similaridade

com esses documentos, porém em menor extensão. Já os documentos contendo a palavra elefante deverão ser bem dissimilares.

Tantrum (2003) mostra em seu estudo, que a indexação semântica latente é uma ferramenta similar a bastante utilizada forma de redução de dimensionalidade denominada análise de componentes principais (*Principal Componente Analysis* (PCA)).

Em indexação semântica latente, a redução do espaço de atributos é obtida a partir da decomposição em valores singulares (SVD) da matriz **X** de documento por termo, enquanto que em análise de componentes principais ela é obtida a partir da decomposição em valores singulares da matriz $\mathbf{X}^*$, matriz **X** de documento por termo, porém centrada na média das colunas ($\mathbf{X}^* = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathbf{t}} \right)\mathbf{X}$).

2.4.2.1 Comentários

- Na análise de componentes principais, o interesse é identificar algumas poucas combinações de termos - denominadas de “conceitos artificiais” neste trabalho - que possam explicam a maior parte da variabilidade dos dados. Nadler & Smith[13] alertam que as combinações de termos obtidas por componentes principais podem não ser as mesmas que melhor discriminam as categorias, posto que as componentes principais não são derivadas visando discriminar categorias e sim visando explicar a variabilidade dos dados. Dessa forma, a utilização dessa técnica em reconhecimento de padrões pode apresentar resultados pobres.

[13]Disponível em: http://www.sandia.gov/imrl/XVisionScience/Xlimit.htm. Acesso em 05 dez. 2006.

- A redução da dimensionalidade por extração de termos é pouco utilizada na área de categorização automática de textos. Exemplos de aplicação da técnica podem ser encontrados em Schütze et al. (1995) e Wiener et al. (1995).

2.5 Classificadores

Como descrito na seção 1.3, neste trabalho o problema *multilabel* de classificação nas categorias $c_1,...,c_m$ será abordado considerando m problemas independentes de classificação binária em $\{c_k, \bar{c}_k\}$, para $k=1,..., m$, onde $\bar{c}_k$ refere-se ao conjunto completo de categorias sem a categoria c_k.

O objetivo de um classificador - função de classificação, função de decisão - é fazer o computador reproduzir, da forma mais próxima possível, o comportamento de um sistema de classificação já existente, criado por humanos, de modo a classificar novos documentos com o menor erro possível. Para isso, o classificador é treinado em um conjunto de exemplos pré-classificados por esse sistema (conjunto de treinamento) nas diversas categorias temáticas do domínio considerado.

Formalmente, o problema de classificação em duas classes pode ser colocado do seguinte modo: dado um conjunto de treinamento $T_r = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$, o objetivo é produzir um classificador $f : \Re^p \rightarrow \{-1, 1\}$, que mapeia um documento $\mathbf{x} \in \Re^p$ à sua categoria temática $y \in \{-1, 1\}$, e que seja capaz de fazer predições acuradas sobre a categoria temática de um documento futuro, ainda não visto, quando colocado em produção na prática.

Nesse problema:

- $\mathbf{x}_i^{\mathrm{t}} = \left[x_{i1}, \cdots, x_{ip}\right] \in \Re^p$ é denominado vetor de atributos (vetor de entradas, vetor de termos) e corresponde à representação do *i*-ésimo documento como um vetor de pesos $0 \leq x_{ij} \leq 1$ que expressam o quanto o *j*-ésimo termo, $j = 1, \cdots, p$, contribui para a semântica do documento $\mathbf{x}_i$;
- p corresponde ao número de termos do dicionário de termos adotado ou ao número de termos selecionados por um dos métodos de seleção de termos, conforme o caso;
- $y_i \in \{-1, 1\}$ é denominado rótulo da classe (saída) do *i*-ésimo documento com relação à categoria temática de interesse, e é tal que:

$$y_i = \begin{cases} 1, & \text{se o documento } \mathbf{x}_i \text{ pertence à categoria temática de interesse} \\ -1, & \text{caso contrário} \end{cases},$$

$i = 1, \cdots, n$.

Várias idéias foram propostas para solucionar o problema de classificação nas diversas áreas do conhecimento. Na área de categorização automática de textos, o algoritmo que se consolidou como o melhor em efetividade na categorização é o algoritmo *support vector machines* (JOACHIMS, 1998; YANG & LIU, 1999; FORMAN, 2003, SEBASTIANI 2003, LEWIS et al., 2004; MINEAU & SOUCY, 2005, OZGÜR et al., 2005).

Esse método foi originalmente proposto por Vapnik, e introduzido na área de categorização de textos por Joachims (JOACHIMS, 1998). Este será o algoritmo utilizado neste trabalho. O conceito em que se fundamenta o método é apresentado na subseção seguinte.

2.5.1 Algoritmo Support Vector Machines (SVMs)

O algoritmo SVMs é baseado no princípio da minimização estrutural do risco, cuja idéia é encontrar uma função de classificação – classificador - para a qual se pode garantir o menor risco esperado (JOACHIMS, 1998).

O risco esperado de uma função de classificação é uma medida de quão boa é a função para predizer corretamente a classe y de um documento de teste ***x*** (vetor de termos) ainda não visto (futuro) e aleatoriamente selecionado.

A motivação por trás da minimização estrutural do risco é manter um compromisso entre a complexidade da função de classificação e o erro na classificação dos documentos de treinamento. Uma função de classificação simples demais poderá não aproximar bem o comportamento do sistema de classificação real e assim produzir erro alto mesmo na classificação dos documentos em que foi treinada. Por outro lado, uma função de classificação muito complexa pode se ajustar excessivamente bem aos documentos de treinamento – *overfitting* -, produzindo baixo erro na classificação desses documentos, sem conseguir, no entanto, reproduzir o mesmo desempenho com documentos diferentes desses.

Tendo em vista esse compromisso, as SVMs buscam na classe dos hiperplanos a função de classificação mais simples em termos de complexidade, porém com o menor risco empírico (menor erro de classificação no conjunto de treinamento).

Para entender a idéia envolvida no algoritmo SVMs, considere um problema de classificação simples em um espaço bidimensional com dados linearmente separáveis[14].

[14] Vide definição em http://en.wikipedia.org/wiki/Linearly_separable. Acesso em: 15 abr. 2007.

Em um espaço linearmente separável, a superfície de decisão[15] é um hiperplano[16], que no caso bidimensional corresponde a uma reta. Desse modo, pode-se traçar uma reta

$$f(\mathbf{x}) = w_1 x_1 + w_2 x_2 + b\,, \qquad (2.18)$$

de forma que todos os documentos com $y_i = -1$ estejam de um lado da reta e tenham $f(\mathbf{x}_i) < 0$ e todos os documentos com $y_i = 1$ estejam do outro lado da reta e tenham $f(\mathbf{x}_i) > 0$.

Para classificar novos documentos, o vetor $\mathbf{w}$ e a constante b em (2.18) são estimados usando os documentos do conjunto de treinamento, e um novo documento $\mathbf{x}$ é classificado conforme o rótulo y recebido de acordo com a função sinal:

$$y = sign(f(\mathbf{x})) = \begin{cases} 1 & \text{se } f(\mathbf{x}) > 0 \\ -1 & \text{caso contrário}. \end{cases}$$

No entanto, tipicamente existem infinitas retas que podem separar esses documentos corretamente. Para entender como o algoritmo SVMs escolhe a "melhor" reta, considere o exemplo de classificação dos documentos pertencentes às categorias preta e branca mostrado nas Figuras 2.2 e 2.3 a seguir.

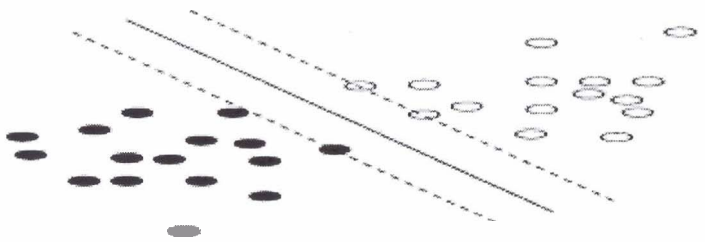


Figura 2.2 - Reta de decisão (reta sólida) com a margem (distância entre as retas paralelas tracejadas) máxima.
Fonte: página 44, Yang & Liu (1999).

[15] Vide definição em http://en.wikipedia.org/wiki/Decision_surface. Acesso em: 15 abr. 2007.
[16] Vide definição em http://en.wikipedia.org/wiki/Hyperplane. Acesso em: 15 abr. 2007.

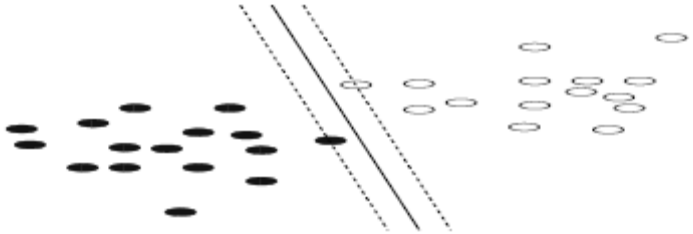


Figura 2.3 - Reta de decisão (reta sólida) com margem não-máxima.
Fonte: página 44, Yang & Liu (1999).

Nas Figuras 2.2 e 2.3, as retas sólidas constituem duas possíveis soluções para o problema de classificação apresentado: as duas separam corretamente os documentos pertencentes à categoria preta dos pertencentes à categoria branca. As retas tracejadas, paralelas às retas sólidas, mostram o quanto se pode mover cada superfície de decisão (reta sólida) sem causar um erro na classificação dos documentos. A distância entre cada par de retas tracejadas é denominada "margem". O algoritmo *support vector machines* busca encontrar o hiperplano (no caso bidimensional, reta) que separa os documentos das duas categorias com a máxima margem.

Nas Figuras 2.2 e 2.3, os pontos de treinamento que se encontram sobre as retas tracejadas paralelas e cuja remoção poderia alterar a solução encontrada são denominados vetores de suporte (*support vetors*). Os vetores de suporte são os elementos críticos do conjunto de treinamento. Eles são os elementos mais próximos do limite de decisão. Se todos os outros pontos de treinamento fossem removidos ou movidos sem cruzar as retas tracejadas e o treinamento fosse repetido, a mesma reta de decisão seria obtida (YANG & LIU, 1999).

Em sua forma geral, para lidar com casos linearmente separáveis e não linearmente separáveis no espaço p-dimensional, o cálculo do hiperplano ótimo buscado pelo algoritmo SVMs é colocado como um problema de otimização com restrições e é solucionado usando técnicas de programação quadrática. Isso leva à função de decisão

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b\,, \qquad (2.19)$$

onde $k(.,.)$ é uma função denominada *kernel*, e o sinal de $f(\mathbf{x})$ determina o rótulo y da classe do documento $\mathbf{x}$ a ser classificado.

Encontrar o hiperplano ótimo equivale a encontrar os $\alpha_i s$ maiores que zero em (2.19). Os vetores de entrada $\mathbf{x}_i$ cujos $\alpha_i s$ são maiores que zero correspondem aos vetores de suporte do hiperplano ótimo. Normalmente, o número de pontos de treinamento retidos como vetores de suporte é pequeno, fazendo com que o classificador gerado seja compacto **(**AHUJA et al., 2001**).**

No caso linearmente separável, a função kernel é simplesmente o produto escalar entre um documento de teste $\mathbf{x}$ e um documento de treinamento $\mathbf{x}_i$, calculado no espaço de entradas original. No caso não linearmente separável, usando uma transformação não-linear obtida por meio de funções kernel, como as polinomias ou funções de base radial, os vetores de entrada $\mathbf{x}_i$, $i = 1, \cdots, n$, são mapeados para um espaço de dimensão mais elevada, denominado espaço de características, onde o hiperplano de separação com máxina margem é construído. A motivação por trás desse mapeamento é que é mais provável encontrar um hiperplano no espaço de características mais elevado.

Para lidar com casos onde não é possível separar todos os dados de treinamento por meio de um hiperplano sem cometer erros, mesmo em um espaço de características de dimensão mais elevada, variáveis de folga (*slack variables*) são introduzidas para produzir hiperplanos ótimos com margens flexíveis (*soft margin optimal hyperplane*).

2.5.1.1Comentários

- Na utilização do algoritmo SVMs, a seleção de termos, em geral, não é necessária uma vez que as SVMs tendem a ser bastante robustas ao problema de *overfitting* e são escaláveis a dimensões consideráveis (JOACHIMS, 1998; SEBASTIANI, 2002). Porém, estudos usando seleção de termos com SVMs apresentaram efetividades na categorização superiores às obtidas com a dimensão completa (FORMAN, 2003; ÖZGÜR et al., 2005; TANG & LIU, 2005);
- Leopold & Kindermann (2002), em estudo utilizando três corpora (Reuters 21578, "ModApte"; notícias do jornal diário alemão "Frankfurter Rundschau"; e notícias do jornal diário alemão "Die Tageszeitung") com várias formas de representação para os pesos dos termos observaram poucas diferenças na performance do algoritmo SVMs usando kernel linear, kernel polinomial de vários graus e kernel função de base radial (*radial basis function* – RBF) com várias variâncias. O maior impacto observado no desempenho do algoritmo foi devido à forma escolhida para representação dos pesos dos termos.
- Lewis et al. (2004) comentam não terem usado *kernels* com funções não-lineares, por esses não terem mostrado vantagens significativas em estudos anteriores na área de categorização de textos. Nesse trabalho Lewis et al. (2004) também apresentam

um estudo para modelar o classificador SVMs atribuindo pesos diferenciados para os erros dos exemplos positivos e negativos no treinamento, visando compensar o desbalanceamento entre as classes. Porém, esses resultados não se apresentaram melhores do que os utilizando pesos iguais para os erros cometidos pelos exemplos das duas classes.

- Özgür et al. (2005) também comentam não terem usado *kernels* com funções não-lineares, pois em estudo piloto realizado, o algoritmo SVMs usando kernel linear apresentou consistentemente a melhor performance quando comparado com SVMs com kernel usando polinômios de vários graus e com kernel usando função de base radial com várias variâncias. Comentário similar é feito por Lan et al. (2006).

- A implementação do algoritmo SVMs mais utilizada pela comunidade científica é o SVM^{light} [17] de Joachims. Outras implementações citadas são os *softwares* livres WEKA[18] e LIBSVM[19].

- Em geral, os *softwares* são utilizados com seus parâmetros *default*. No caso do SVM^{light} isso significa, entre outras especificações de parâmetros, que se utiliza kernel linear e mesmo peso para os erros cometidos por exemplos positivos e negativos no treinamento.

2.5.2 Algoritmo k-nearest neighbors (k-NN)

Entre os algoritmos comparados com as SVMs estão *k nearest neighbors* (*k*-NN), LLSF (*liner least squares fit*) -método de regressão linear-, regressão logística,

[17] Disponível em: http://svmlight.joachims.org/. Acesso em: 21 jan. 2007.
[18] Disponível em: http://www.cs.waikato.ac.nz/ml/weka/. Acesso em: 24 jan. 2007.
[19] Disponível em: http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Acesso em: 24 jan. 2007.

redes neurais (*neural network*), Naïve Bayes, Rocchio e árvore de decisão. Dentre esses, o *k*-NN é o mais usado e apresenta efetividade competitiva ao algoritmo SVMs, motivo pelo qual será apresentado a seguir.

Um ponto negativo do *k*-NN, no entanto, é o alto custo computacional incorrido no momento da classificação, que o impede de ser visto como uma alternativa às SVMs em aplicações práticas (LEOPOLD & KINDERMANN, 2002). Por essa razão o algoritmo não será utilizado no presente trabalho.

Diferentemente de algoritmos como as SVMs que requerem a construção de *m* classificadores binários em $\{c_k, \bar{c}_k\}$, para *k=1,..., m*, para decidir se o documento deve ou não ser classificado na respectiva categoria, o *k*-NN classifica os documentos nas *m* categorias usando um único procedimento. A seguir é apresentada a versão do algoritmo descrita em Lewis et al. (2004).

Algoritmo *k*-NN: Para decidir se um documento de teste pertence a uma determinada categoria temática, o algoritmo *k*-NN calcula a similaridade - de acordo com alguma medida de similaridade - entre esse documento de teste e todos os documentos do conjunto de treinamento, e então seleciona os *k* documentos de treinamento mais similares a ele (os *k* vizinhos mais próximos).

Para cada categoria temática, as medidas de similaridade entre o documento de teste e os documentos de treinamento dos vizinhos mais próximos a ela pertencentes são somadas, produzindo um escore para a categoria com relação ao documento de teste (uma evidência favorecendo ou desfavorecendo a pertinência do documento de teste à respectiva categoria temática).

O documento de teste é classificado em uma categoria temática se o escore para a categoria com relação ao documento for maior ou igual ao ponto de corte pré-

estabelecido para classificação na respectiva categoria; caso contrário, o documento não é classificado na categoria.

O escore da categoria c_j com relação ao documento de teste $\mathbf{x}_z$ pode ser escrito como:

$$y(\mathbf{x}_z, c_j) = \sum_{\mathbf{x}_i \in k\_NN} sim(\mathbf{x}_z, \mathbf{x}_i)\, \mathrm{I}(\mathbf{x}_i, c_j)\,.$$

onde:

- $y(\mathbf{x}_z, c_j)$ é o escore que representa a verossimilhança da categoria temática c_j para o documento $\mathbf{x}_z$;
- k_NN é o conjunto dos k documentos de treinamento mais similares ao documento de teste $\mathbf{x}_z$;
- $\mathbf{x}_i$ é um dos k documentos de treinamento pertencentes ao conjunto k_NN
- $\mathrm{I}(\mathbf{x}_i, c_j) = \begin{cases} 1, & \text{se o documento } \mathbf{x}_i \text{ pertence à categoria temática } c_j \\ 0, & \text{caso contrário} \end{cases}$
- $sim(\mathbf{x}_z, \mathbf{x}_i)$ corresponde à similaridade entre o documento de teste $\mathbf{x}_z$ e o documento de treinamento $\mathbf{x}_i$. Considerando, por exemplo, a medida de similaridade mais usada, que é a do coseno do ângulo entre dois vetores com o mesmo número de elementos, os cálculos são mostrados a seguir. Sejam:

$$\mathbf{x}_z^t = \left[x_{z1}, \quad \cdots, \quad x_{zp}\right]$$

e

$$\mathbf{x}_i^t = \left[x_{i1}, \quad \cdots, \quad x_{ip}\right],$$

os vetores correspondentes aos documentos de teste e de treinamento, respectivamente. Então,

$$sim(\mathbf{x}_z, \mathbf{x}_i) = \sum_{r=1}^{p} x_{zr} x_{ir} ,$$

que corresponde ao produto escalar entre os vetores $\mathbf{x}_z$ e $\mathbf{x}_i$, considerando que os vetores estão normalizados para comprimento um; caso os vetores não estejam normalizados, a similaridade entre os documentos $\mathbf{x}_z$ e $\mathbf{x}_i$ deve ser calculada como:

$$sim(\mathbf{x}_z, \mathbf{x}_i) = \frac{\sum_{r=1}^{p} x_{zr} x_{ir}}{\sqrt{\sum_{r=1}^{p} x_{zr}^2} \sqrt{\sum_{r=1}^{p} x_{ir}^2}} .$$

2.5.2.1Comentários

- Qualquer função para medir o grau de coincidência entre os valores das coordenadas de dois vetores em um sistema de recuperação da informação por rankeamento pode ser usada como medida de similaridade (SEBASTIANI, 2002).
- Para a utilização do algoritmo *k*-NN, o valor de *k* (o número de vizinhos mais próximos), assim como os pontos de corte para classificação em cada uma das categorias temáticas devem ser determinados em um conjunto de validação ou por meio de validação cruzada. Um exemplo de como isso pode ser feito é encontrado em Lewis et al. (2004, pp. 383).

- Como o método *k*-NN é mais sensível à presença de termos não relevantes do que o algoritmo SVMs, a seleção de termos também pode ser determinada, caso necessário, em um conjunto de validação ou por meio de validação cruzada. Também pode ser encontrado um exemplo para isso em Lewis et al. (2004, pp. 383).
- Debole & Sebastiani (2005) comentam que o cálculo e rankeamento das similaridades é um procedimento extremamente caro do ponto de vista computacional. Além disso, caso se decida trabalhar usando a representação dos documentos por dicionário local, esse procedimento necessita ser realizado *m* vezes - número de categorias do problema -, posto que o mesmo documento terá representações diferentes em cada uma das *m* categorias representadas pelo dicionário local.
- Lan et al. (2006) relatam que apesar do algoritmo ser simples e efetivo, sua desvantagem é a ineficiência computacional em casos de alta dimensionalidade e corpus com grande volume de documentos. Em estudo para comparar métodos de atribuição de pesos aos termos no corpus Reuters-21578, "ModApte", eles comentam que enquanto o algoritmo SVMs foi executado até a dimensão completa – 15.937 termos -, o algoritmo *k*-NN só foi executado com até 1.000 termos, devido à sua ineficiência computacional.

2.6 Avaliação dos Classificadores

Após a construção do classificador, ele é submetido à avaliação em um conjunto de documentos de teste. Em categorização de textos, as medidas mais utilizadas para avaliar o desempenho dos algoritmos de classificação referem-se à

efetividade na classificação, isto é, à habilidade do classificador tomar a decisão correta (classificar corretamente os documentos) ao realizar uma classificação. Para definir essas medidas, considere a matriz de confusão dada na Tabela 2.8, que mostra a distribuição da classificação real (feita pelo especialista, manualmente) versus a predita (feita pelo classificador, automaticamente) para a categoria c_k, considerando os n_{te} documentos de teste:

Tabela 2.8 - Classificação na categoria c_k : classificador x especialista.

(documentos de teste)

Classificação pelo especialista / Classificação pelo classificador	c_k (classe positiva)	$\bar{c}_k$ (classe negativa)
c_k (classe positiva)	VP_k (decisão correta)	FP_k (decisão incorreta)
$\bar{c}_k$ (classe negativa)	FN_k (decisão incorreta)	VN_k (decisão correta)

onde:

VP_k : verdadeiros positivos com relação à c_k : número de documentos da categoria c_k corretamente preditos como categoria c_k pelo classificador;

FP_k : falsos positivos com relação à c_k : número de documentos que não são da categoria c_k, mas foram incorretamente preditos como categoria c_k pelo classificador;

FN_k : falsos negativos com relação à c_k : número de documentos da categoria c_k que foram incorretamente preditos como não sendo da categoria c_k pelo classificador;

VN_k : verdadeiros negativos com relação à classe c_k : número de documentos que

não são da categoria c_k e foram corretamente preditos pelo classificador como não sendo da categoria c_k.

Como comentado na seção 1.3, dependendo do número de categorias que estiverem sendo representadas em $\bar{c}_k$, a quantidade de documentos a ela pertencentes pode ultrapassar em muito a quantidade de documentos em c_k, gerando classes bastante desbalanceadas. Nessas situações, medidas como acurácia ($(VP_k + VN_k)/n_{te}$) ou taxa de erro (1-acurácia) podem ser indicadores fracos de desempenho, a menos que todos os tipos de erros cometidos assim como todos os benefícios dos acertos obtidos tiverem a mesma importância no problema de classificação considerado (vide Tabela 2.8).

Como exemplo de situação onde a acurácia não fornece uma boa indicação para o desempenho do classificador, considere um problema de classificação de e-mail em uma empresa com 50 mensagens do tipo "lixo eletrônico" (c_k - classe positiva) contra 1.950 mensagens do tipo "importante" ($\bar{c}_k$ - classe negativa). Um problema com assimetria de 1:39 (1 documento de c_k para 39 de $\bar{c}_k$).

O classificador trivial majoritário, o que simplesmente classifica todas as mensagens como "importante", apresenta uma acurácia alta (1.950/2.000=0,975), classificando 97,5% das mensagens corretamente. Porém, não consegue identificar nenhum "lixo eletrônico", que é o objetivo do sistema de filtragem.

Em geral, a efetividade na classificação é medida em termos das noções clássicas de precisão (*precision*), revocação (*recall*) e F_1, utilizadas na área de recuperação da informação, e adaptadas para o caso de categorização automática de textos. As medidas são definidas a seguir:

REVOCAÇÃO (*recall*) com relação à categoria c_k: proporção de documentos da categoria c_k corretamente classificados pelo classificador na categoria c_k, ou seja, é a estimativa da probabilidade de um documento pertencente à categoria c_k ser classificado nessa categoria. A revocação é estimada como:

$$R_\kappa = \frac{VP_k}{VP_k + FN_k}$$

PRECISÃO (*precision*) com relação à categoria c_k**:** proporção corretamente classificada na categoria c_k entre os documentos classificados pelo classificador na categoria c_k**,** ou seja, é a estimativa da probabilidade de um documento predito como sendo da classe c_k de fato pertencer a essa categoria. A precisão é estimada como:

$$P_k = \frac{VP_k}{VP_k + FP_k}$$

Analisar o desempenho de um classificador por apenas uma dessas duas medidas, no entanto, pode levar a conclusões enganosas. No problema de classificação de e-mail exposto anteriormente, por exemplo, o classificador trivial que classifica todas as mensagens como “lixo eletrônico” terá índice de revocação perfeito igual a um (50/50), pois todos os “lixos eletrônicos” serão corretamente identificados. Porém, terá índice de precisão baixo (50/2.000=0,025), posto que todas as mensagens do tipo “importante” serão incorretamente classificadas como “lixo eletrônico”. Por outro lado, um outro classificador que classifica cinco “lixos eletrônicos” corretamente e todas as mensagens do tipo “importante” corretamente terá revocação igual a 0,1 (5/50) e precisão perfeita igual a um (5/5). Esse último classificador não identifica nenhuma

mensagem "importante" como "lixo eletrônico", mas deixa de detectar 90% das mensagens do tipo "lixo eletrônico".

Normalmente um classificador exibe uma permuta entre revocação e precisão. É bem conhecido na prática de recuperação da informação que níveis mais altos de precisão podem ser obtidos ao preço de revocação mais baixa e vice-versa. Assim, nem precisão nem revocação fazem sentido isoladamente. Um classificador deve então ser avaliado por uma medida que combina precisão e revocação. A medida mais utilizada para tal fim é denominada medida F_1 e é definida a seguir.

F_1 ou MEDIDA F (*F-measure*) com relação à categoria c_k: corresponde à média harmônica entre revocação (R_k) e precisão (P_k). A medida F_1 dá a mesma importância para revocação e precisão e é definida como:

$$F_{1_k} = \frac{2P_kR_k}{P_k + R_k} = \frac{2VP_k}{2VP_k + FP_k + FN_k}.$$

Observações:

- em sua forma geral, a medida *F*, para β real não-negativo, é definida como:

$$F_{\beta_k} = \frac{(\beta^2 + 1)P_kR_k}{\beta^2 P_k + R_k},$$

onde β pode ser visto como um grau relativo de importância atribuído à precisão ou à revocação. Quando $\beta = 0$, F_{0_k} coincide com P_k; quando $\beta = 0.5$, $F_{0.5k}$ atribui duas vezes mais importância à P_k do que à R_k; quando $\beta = 2$, F_{2k} atribui duas vezes mais importância à R_k do que à P_k; quando $\beta = +\infty$, $F_{+\infty k}$ coincide com R_k;

- o valor onde a precisão é igual à revocação é denominado BEP (*break-even point*). Esse valor corresponde a um valor específico de F_1. Isso é, quando $R_k = P_k$,

$$F1_k = \frac{2P_k{}^2}{2P_k} = P_k = R_k .$$

As medidas de efetividade apresentadas calculam o desempenho do classificador em uma categoria específica. De forma a obter uma medida de efetividade global, que considere o desempenho do classificador no conjunto completo de categorias, é necessário combinar de alguma forma as medidas individuais obtidas.

Na área de categorização de textos, duas medidas de efetividade global são utilizadas: média micro (*micro average*) e média macro (*macro average*). Para definir essas medidas, considere a Tabela 2.9 a seguir:

Tabela 2.9 - Classificação global considerando todas as classes:
Classificador x Especialista (documentos de teste).

Classificação pelo especialista / Classificação pelo classificador	classe positiva	classe negativa
classe positiva	$VP = \sum_{k=1}^{m} VP_k$	$FP = \sum_{k=1}^{m} FP_k$
classe negativa	$FN = \sum_{k=1}^{m} FN_k$	$VN = \sum_{k=1}^{m} VN_k$

As medidas globais utilizadas são definidas conforme segue.

Média Micro (*Microaverage*): as medidas globais para revocação, precisão e F_1 são obtidas somando as decisões individuais em cada categoria:

revocação micro: $$R_{mi} = \frac{VP}{VP + FN}$$

precisão micro: $$P_{mi} = \frac{VP}{VP + FP}$$

F_1 **micro**: $$F_{1_{mi}} = \frac{2VP}{2VP + FP + FN}$$

Média Macro (*Macroaverage*)[20]: as medidas globais para revocação, precisão e F_1 correspondem à média dos resultados individuais obtidos nas diversas categorias

revocação macro: $$R_{ma} = \frac{\sum_{k=1}^{m} R_k}{m}$$

precisão macro: $$P_{ma} = \frac{\sum_{k=1}^{m} R_k}{m}$$

F_1 **macro**: $$F_{1ma} = \frac{\sum_{k=1}^{m} F_{1k}}{m}$$

As medidas calculadas por micro atribuem o mesmo peso para cada classificação de documento, enquanto as calculadas por macro atribuem o mesmo peso para cada um dos *m* problemas de categorização.

Esses dois métodos podem fornecer resultados bem diferentes, especialmente se o número de documentos por categoria for muito desigual, posto que bons desempenhos de classificação em categorias com pequeno número de documentos são enfatizados nas medidas calculadas por macro, porém são diluídos nas medidas calculadas por micro.

[20] Para evitar problemas numéricos no cálculo das medidas por macro, conforme prática comum, R_k (P_k) é assumido igual a 1, quando $VP_k + FN_k$ ($VP_k + FP_k$) for igual a zero (DEBOLE & SEBASTIANI, 2003).

3 Estudo de caso

3.1 Descrição do problema

Neste capítulo é apresentado o estudo sobre a modelagem, por meio da técnica de categorização automática de textos, da indicação de distribuição dos projetos de lei para as comissões permanentes da Câmara Legislativa do Distrito Federal.

No caso sob consideração, os documentos a serem classificados são os projetos de lei (PLs), e as categorias temáticas atribuídas aos PLs são as comissões permanentes que devem apreciá-los, sendo que atualmente são nove essas comissões:

1. Comissão de Constituição e Justiça (CCJ);
2. Comissão de Economia, Orçamento e Finanças (CEOF);
3. Comissão de Assuntos Sociais (CAS);
4. Comissão de Defesa do Consumidor (CDC);
5. Comissão de Defesa dos Direitos Humanos, Cidadania, Ética e Decoro Parlamentar (CDDHCEDP);
6. Comissão de Assuntos Fundiários (CAF);
7. Comissão de Educação e Saúde (CES);
8. Comissão de Segurança (CSEG);
9. Comissão de Desenvolvimento Econômico Sustentável, Ciência, Tecnologia, Meio Ambiente e Turismo (CDESCTMAT).

Dado que todos os projetos de lei passam pela CCJ, para análise da admissibilidade quanto à constitucionalidade, juridicidade, legalidade, regimentalidade, técnica legislativa e redação, ela não será considerada no estudo.

Para entender os dados a serem analisados, é mostrada na Figura 3.1 um exemplo de projeto de lei (documento a ser classificado), que é composto, essencialmente, por um resumo do projeto (ementa), pela matéria a ser disciplinada pela lei (corpo ou texto da lei), e por uma série de argumentos onde o autor procura demonstrar a necessidade ou oportunidade do projeto (justificação).

PROJETO DE LEI N° 5, DE 2003
(Do Deputado ODILON AIRES)

Assegura a expedição de Alvará de Funcionamento para estabelecimentos instalados com os benefícios do Programa de Promoção do Desenvolvimento Econômico Integrado e Sustentável do Distrito Federal PRÓ-DF e dá outras providências.

A CÂMARA LEGISLATIVA DO DISTRITO FEDERAL decreta:

Art. 1º Incluam-se os § 8º e 9º no art. 1º da Lei nº 1.171, de 24 de julho de 1996, com a seguinte redação:

"Art.1º.....

§ 8° - Fica permitida a expedição de Alvará de Funcionamento para mais de uma atividade econômica para os estabelecimentos instalados em áreas destinadas ao Programa de Promoção do Desenvolvimento Econômico Integrado e Sustentável do Distrito Federal PRÓ-DF, desde que o beneficiário, cumulativamente, comprove:

ter implantado o empreendimento dentro do prazo estabelecido no plano de viabilidade técnica, econômica e financeira; e

a efetiva geração do quantitativo de postos de trabalho conforme constante do plano de viabilidade técnica, econômica e financeira".

Art. 2º- Esta lei entra em vigor na data de sua publicação.

Art. 3º- Revogam-se as disposições em contrário.

JUSTIFICAÇÃO

A presente proposição visa assegurar maior dinamização da economia e otimizar a utilização do uso das áreas destinadas aos empreendimentos econômicos. Busca-se também dar nova conformação aos setores do PRÓ-DF, vez que nos tempos atuais a grande tendência é a agregação e uma gama de atividades econômicas complementares como forma de viabilizar investimentos, minimizar custos e ofertar maior com odidade e diversidade de opções para o público e estimular a

concorrência.

Contamos, pois, com o aval dos nobres Deputados para a aprovação do presente Projeto de Lei.

Sala das sessões, em 05.02.2003

Deputado *ODILON AIRES*
PMDB/DF

Figura 3.1 – Exemplo de Projeto de Lei.

O encaminhamento desses PLs às comissões permanentes deve ser feito de acordo com as competências estabelecidas para essas nos artigos 63 a 69-B do Regimento Interno da CLDF, e é realizado a partir da indicação de distribuição efetuada pela Assessoria de Plenário e Distribuição. A Figura 3.2 mostra um exemplo de descrição de uma dessas comissões (categorias temáticas).

Art. 69-B. Compete à Comissão de Desenvolvimento Econômico Sustentável, Ciência, Tecnologia, Meio Ambiente e Turismo analisar e, quando necessário, emitir parecer sobre o mérito das seguintes matérias: *(Artigo acrescido pela Resolução nº 181, de 11/03/2002, e alterado pela Resolução nº 200, de 08/12/2003.)*

a) política industrial;
b) política de incentivo à agropecuária e às microempresas;
c) política de interação com a Região Integrada do Desenvolvimento Econômico do Entorno;
d) política econômica, planos e programas regionais e setoriais de desenvolvimento integrado do Distrito Federal;
e) planos e programas de natureza econômica;
f) estudos, pesquisas e programas de desenvolvimento da ciência e tecnologia;
g) produção, consumo e comércio, inclusive o ambulante;
h) turismo, desporto e lazer;
i) energia, telecomunicações e informática;
j) cerrado, caça, pesca, fauna, conservação da natureza, defesa do solo e dos recursos naturais, proteção do meio ambiente e controle da poluição;
k) desenvolvimento econômico sustentável.

Figura 3.2 – Descrição da categoria temática CDESCTMAT.

Fonte: Regimento Interno da Câmara Legislativa do Distrito Federal.

3.2 Propostas de análises

Dependendo da matéria tratada, um projeto de lei pode ser apreciado por mais de uma comissão permanente, referindo-se o problema de classificação sob estudo a uma classificação *multilabel*, isto é, mais de um rótulo (comissão) pode ser atribuído ao mesmo PL.

Neste trabalho, a classificação *multilabel* dos PLs nas oito comissões permanentes será abordada como oito problemas independentes de classificação binária em cada uma das comissões. Desse modo, um classificador por comissão será construído. Cada classificador responde à seguinte pergunta: - o PL deve ser apreciado por esta comissão? A resposta é sim ou não.

Por meio dessa abordagem, na construção do classificador para uma determinada comissão, os PLs que foram apreciados por essa comissão – independentemente de terem sido apreciados por outras comissões também - vão formar os exemplos positivos da comissão, e os demais PLs, os exemplos negativos.

Para obter classificadores que possam fornecer a maior efetividade possível em termos de classificação, vários modelos serão pesquisados. A escolha do algoritmo de classificação não será objeto de estudo, uma vez que o algoritmo *Support Vector Machines* (SVMs) tem apresentado os melhores resultados em termos de efetividade na classificação de textos, e será o algoritmo utilizado neste trabalho. Também o algoritmo SVMs será utilizado apenas com *kernel* linear, por estudos anteriores não indicarem melhoras significativas em efetividade na classificação usando *kernels* com funções não-lineares (seção 2.5.1.1, comentários sobre o algoritmo SVMs).

A busca pelos melhores modelos será realizada comparando representações vetoriais distintas para os PLs com várias formas de atribuição de pesos aos termos que compõem essas representações.

As representações vetoriais para os PLs serão pesquisadas usando dicionário local e global. Na adoção do dicionário local, uma representação vetorial distinta será construída para cada comissão permanente, usando apenas os termos presentes nos PLs de treinamento que passaram pela respectiva comissão. Com dicionário global, uma única representação vetorial será construída para todas as comissões, usando os termos presentes em pelo menos um dos PLs de treinamento.

Os pesos para os termos (seção 2.3) serão calculados segundo duas abordagens:

I. conforme a proposta usual TF_IDF (fórmula (2.2), seção 2.3):

$$TF\_IDF(t_j, \mathbf{d}_i) = (1 + \log f_{ij}) \log \frac{n}{n_j} \quad ,$$

onde:

- t_j, refere-se ao *j*-ésimo termo do dicionário de termos adotado;
- f_{ij}, é a freqüência do termo t_j no PL $\mathbf{d}_i$;
- n, é o número de PLs do conjunto de treinamento;
- n_j, é o número de PLs do conjunto de treinamento em que o termo t_j está presente;
- $i = 1, \cdots, n$, $j = 1, \cdots, p$.

II. usando atribuição supervisionada de termos (STW). Nessa abordagem, quando o dicionário adotado for o local, o termo $IDF(t_j)$, na fórmula $TF\_IDF(t_j, \mathbf{d}_i)$, será substituído por $f(t_j, c_k)$, onde $f(t_j, c_k)$ é calculada usando uma das medidas de

seleção de termos - qui-quadrado ($\chi^2(t_j, c_k)$), ganho de informação ($IG(t_j, c_k)$), razão de ganho ($GR(t_j, c_k)$), razão de chances ($OR(t_j, c_k)$), abs-logito ($ABSL(t_j, c_k)$) e bi-normal separation ($BNS(t_j, c_k)$) - apresentadas na seção 2.4.1. Isto é:

- $TF\_QUI(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij})\chi^2(t_j, c_k)$
- $TF\_IG(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij})IG(t_j, c_k)$
- $TF\_GR(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij})GR(t_j, c_k)$
- $TF\_OR(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij})OR(t_j, c_k)$
- $TF\_ABSL(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij})ABSL(t_j, c_k)$
- $TF\_BNS(t_j, \mathbf{d}_i, c_k) = (1 + \log f_{ij})BNS(t_j, c_k)$,

onde:

- t_j, refere-se ao *j*-ésimo termo do dicionário de termos adotado;
- f_{ij}, é a freqüência do termo t_j no PL $\mathbf{d}_i$;
- c_k, corresponde a cada uma das comissões permanentes – CEOF, CAS, CDC, CDDHCEDP, CAF, CES, CSEG, CDESCTMAT;
- $k = 1, \cdots, 8$; $j = 1, \cdots, p$; $i = 1, \cdots, n$.

Na adoção do dicionário global, como medida que resume os escores obtidos para o termo t_j nas diversas comissões permanentes c_k, $k = 1, \cdots, 8$, será utilizada a função máximo

$$f_{\max}(t_j) = \max_{k=1,\cdots,8} f(t_j, c_k),$$

pois em trabalhos anteriores, essa foi a função que forneceu os melhores resultados em termos de efetividade na classificação (seção 2.3).

Também será estudado o efeito na efetividade da categorização, do aumento de peso para os termos presentes nas ementas e para os termos relacionados às matérias de competência das comissões permanentes.

Além disso, ainda serão investigadas as representações vetoriais usando o conjunto completo de termos do dicionário adotado e usando seleção de termos, caso onde a mesma função será empregada tanto para selecionar os termos, como para calcular seus pesos pelo método de atribuição supervisionada de termos. Apenas com o método de redução da dimensionalidade pela freqüência de documentos, os termos serão selecionados de forma independente da representação vetorial utilizada.

Conforme prática usual, para dar igual importância aos PLs longos e curtos, os vetores correspondentes aos PLs serão normalizados para terem cumprimento unitário (seção 2.3).

Para viabilizar a realização das análises de interesse, os seguintes dados foram produzidos:

- as palavras foram extraídas dos PLs considerando como documento o projeto de lei como um todo, conforme apresentado na Figura 3.1;
- as palavras foram extraídas dos PLs considerando como documento apenas as ementas;
- palavras que descrevem as matérias de competência de cada uma das comissões permanentes foram extraídas dos PLs usando como base as palavras constantes nos artigos 63 a 69-B do Regimento Interno da CLDF e as relacionadas aos assuntos listados nesses artigos.

3.3 Preparação dos Dados

3.3.1 Obtenção, preparação, análise e correção do corpus

Em geral, esta é a etapa mais demorada e trabalhosa do processo, pois dificilmente os dados estarão disponíveis e armazenados da forma como necessitamos, exigindo adaptações e correções. No caso sob estudo, as seguintes situações foram encontradas:

- **impossibilidade de obter todos os PLs com o formato de redação inicial - forma como o projeto é apresentado na CLDF -, onde o texto é composto pela ementa, corpo da lei e justificação**: alguns textos utilizados encontram-se na forma de redação final, que é basicamente o formato da lei propriamente dita, onde o PL é composto pela ementa e corpo da lei, sem a justificação. Apesar do corpo da lei e a ementa conterem a essência da matéria disciplinada pela lei, a importância do texto na forma de redação inicial é que a justificação, muitas vezes, fornece fortes indicativos sobre as comissões que devem apreciar o projeto;
- **falta de padronização nas distribuições dos PLs às Comissões**: nos PLs considerados para estudo foram observados projetos referentes a assuntos semelhantes distribuídos de forma diferente às comissões, ora incluindo uma ou outra comissão, ora ignorando. Esta fase exigiu a leitura, análise e correção das classificações nas comissões de todos os projetos de lei do corpus;
- **dificuldade de obtenção de mais dados para complementar a análise**: no início desta pesquisa em 2004, um esforço estava sendo realizado na CLDF para organizar uma base de dados com os textos no formato digitado, sendo que boa

parte dos projetos de lei apresentados em 2003 e 2004 estavam disponíveis nesse formato. Todavia, no final de 2004, considerando os textos dos PLs como dados de acesso privado e com importância restrita à consulta, a instituição passou a disponibilizá-los apenas em formato digitalizado de uma fotocópia dos mesmos. A única forma de obter os textos originais dos PLs seria dentro dos gabinetes dos deputados. Nessa tentativa, além de se deparar com obstáculos administrativos, verificou-se que muitos textos de PLs encontravam-se espalhados pelos computadores particulares de assessores dos deputados e alguns outros haviam sido perdidos. Considerando o tempo extra a ser despendido para a inclusão desses dados - obtenção, consolidação, leitura, análise e correção das classificações dos PLs nas diversas comissões - optou-se por utilizar neste trabalho apenas os textos eletrônicos disponíveis dos PLs, que são os referentes aos anos de 2003 e 2004.

Dado o exposto, o corpus deste estudo é composto por 1.014 textos eletrônicos de projetos de lei - em formatos txt, word e html - apreciados pelas comissões permanentes da Câmara Legislativa do Distrito Federal, nos anos de 2003 e 2004.

Após pesquisar no sistema Legis (sistema de informações legislativas da CLDF) as comissões por onde passou cada PL (as categorias temáticas atribuídas a cada PL) e corrigir as distribuições incorretas observadas, usando como base os artigos 63 a 69-B do Regimento Interno da CLDF, a seguinte distribuição dos 1.014 PLs para as comissões foi observada:

Tabela 3.1 - Distribuição dos 1.014 PLs para as comissões permanentes, em 2003-2004.

Comissão	Número de PLs apreciados
CEOF	590
CAS	454
CDC	102
CDDHCEDP	56
CAF	113
CES	205
CSEG	148
CDESCTMAT	189

Como mostrado na Tabela 3.2 e Figura 3.3 a seguir, desses 1.014 PLs, 355 foram apreciados por apenas uma comissão permanente, 486 por duas comissões, 162 por três comissões, e 11 por quatro comissões. O número médio de comissões que apreciaram cada PL é de 1,83, com desvio padrão de 0,72.

Tabela 3.2 - Freqüências e porcentagens dos 1.014 PLs, segundo o número de comissões que os apreciaram, em 2003-2004.

Número de comissões que apreciaram o PL	Freqüência de PLs	Porcentagem de PLs
1	355	35,01
2	486	47,93
3	162	15,98
4	11	1,08
Total	1.014	100,00

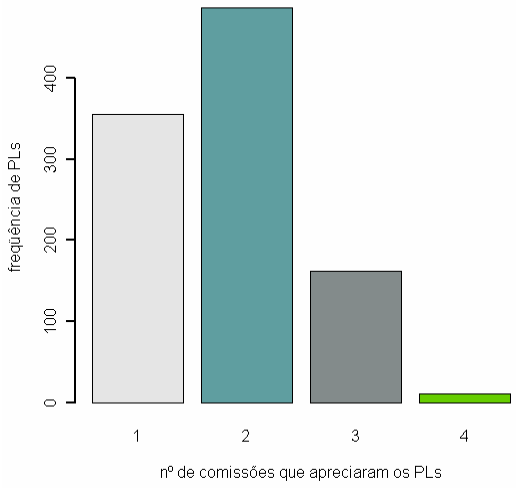


Figura 3.3 - Freqüências dos 1.014 PLs, segundo o número de comissões que os apreciaram, em 2003-2004.

3.3.2 Preparação dos arquivos necessários à análise

A preparação dos dados para estimação dos modelos de classificação foi realizada conforme as seguintes etapas:

- **Extração dos termos presentes nos PLs com suas respectivas freqüências de ocorrência:**

Nesta etapa foi utilizada a ferramenta IDE XML Client, da Temis Text Mining Solutions, versão 2.0, disponível no Núcleo de Transferência de Tecnologia (NTT), do Departamento de Engenharia Civil, da COPPE/UFRJ. O método utilizado pelo Temis para lidar com as variações morfológicas das palavras nos textos é o *lemmatization*, descrito na seção 2.2. Das duas opções disponíveis no programa, foi escolhida a opção "MetaTagging_NAV", que extrai os substantivos (N - Noun), adjetivos (A - Adjective), e verbos (V - Verb) dos textos, com suas respectivas freqüências de ocorrência. A outra opção - "MetaTagging_NV" - extrai apenas os substantivos (N) e verbos (V) dos textos, com suas respectivas freqüências de ocorrência.

A partir da especificação de um diretório contendo os arquivos com os textos de interesse para estudo, o IdeXMLClient gera os termos extraídos desses textos em um arquivo XML. No caso sob estudo, os textos dos 1.014 PLs foram separados em quatro diretórios com aproximadamente 250 PLs cada, pois se todos os PLs fossem colocados em um único diretório apenas um arquivo XML seria gerado, o que dificultaria a manipulação dos resultados obtidos. Os arquivos contendo apenas os textos das ementas dos PLs, por serem pequenos, foram colocados em apenas um diretório.

A Figura 3.4 a seguir mostra um trecho de um dos arquivos XML gerados pelo IdeXMLClient, onde é possível verificar que o documento refere-se ao "PL-2003-00005-RDF" (Figura 3.1) com as palavras, respectivas classes gramaticais e freqüências de ocorrência extraídas do seu texto :

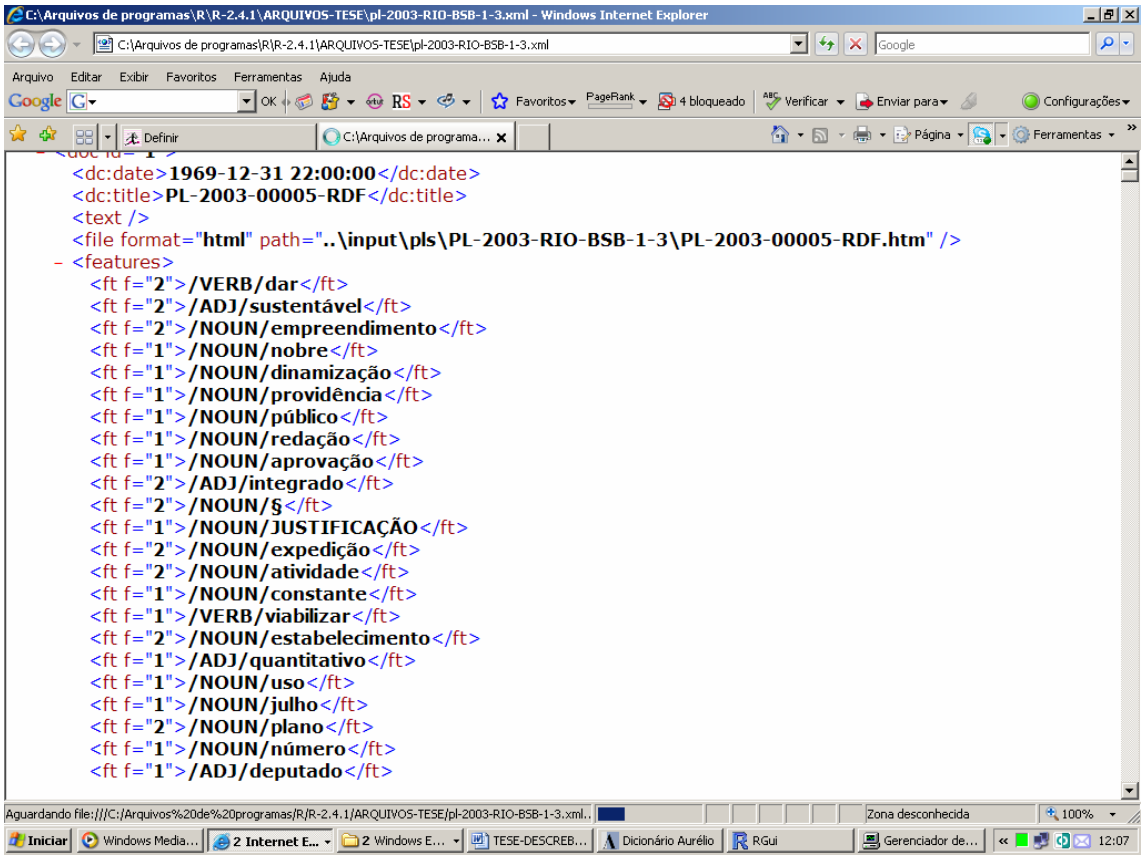


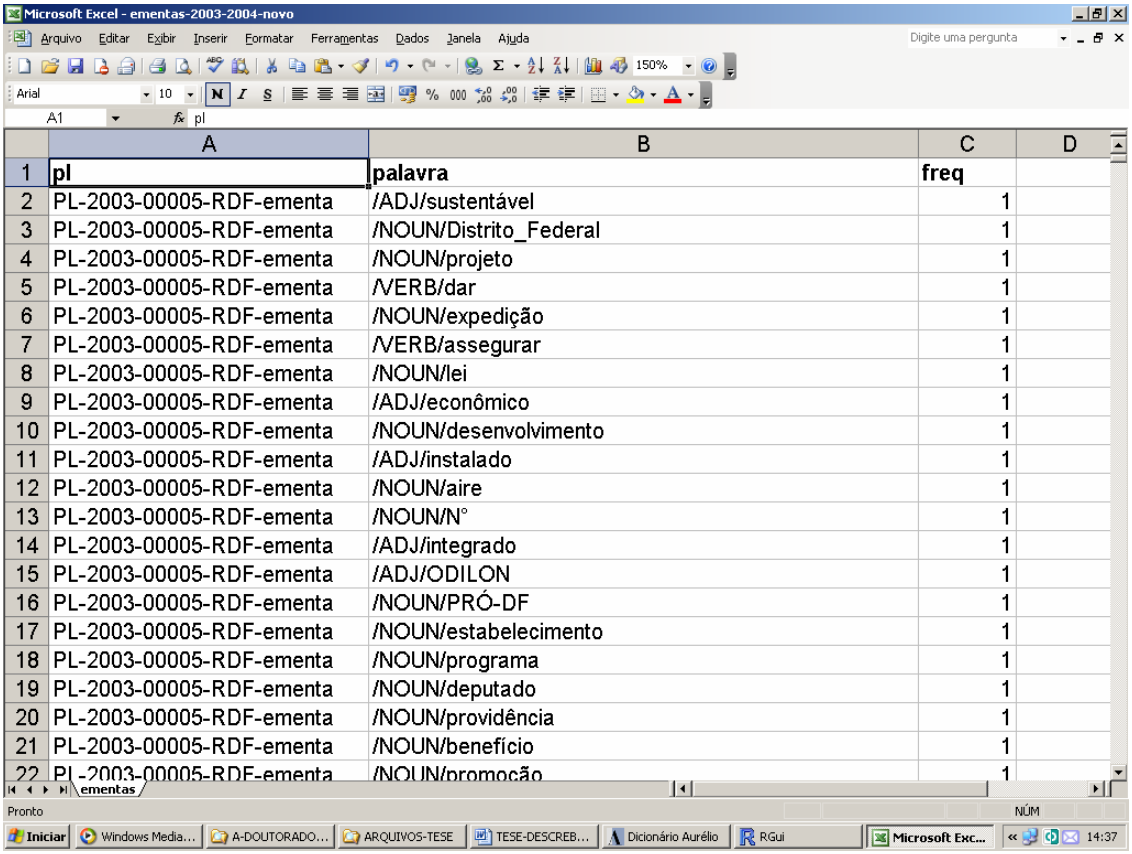
Figura 3.4 – Trecho de um arquivo XML gerado pelo IdeXMLClient.

➢ **Criação dos arquivos necessários à realização das análises de interesse:**

Os arquivos XML gerados pelo IdeXMLClient foram abertos como uma lista XML no programa Microsoft Excel 2003, produzindo um arquivo Excel com uma planilha, no caso dos termos extraídos unicamente das ementas dos PLs, e com quatro

planilhas, quando os termos foram extraídos considerando como documento o PL como um todo.

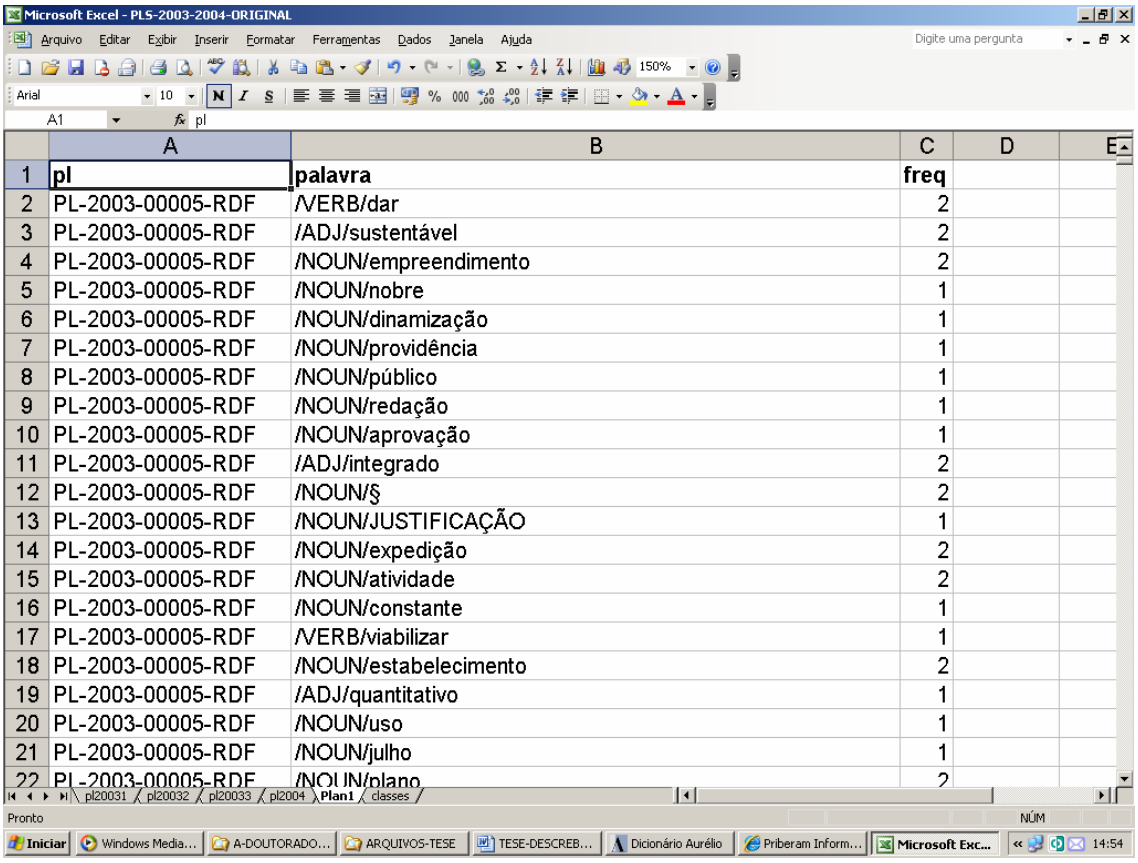
Após a exclusão das colunas desnecessárias e re-nomeação das colunas remanescentes, as planilhas ficaram com as colunas “pl”, “palavra” e “freq”, conforme trechos das planilhas Excel mostradas na Figura 3.5 (dados das ementas) e Figura 3.6 (dados dos PLs) a seguir.



Microsoft Excel - ementas-2003-2004-novo

	A	B	C	D
1	**pl**	**palavra**	**freq**	
2	PL-2003-00005-RDF-ementa	/ADJ/sustentável	1	
3	PL-2003-00005-RDF-ementa	/NOUN/Distrito_Federal	1	
4	PL-2003-00005-RDF-ementa	/NOUN/projeto	1	
5	PL-2003-00005-RDF-ementa	/VERB/dar	1	
6	PL-2003-00005-RDF-ementa	/NOUN/expedição	1	
7	PL-2003-00005-RDF-ementa	/VERB/assegurar	1	
8	PL-2003-00005-RDF-ementa	/NOUN/lei	1	
9	PL-2003-00005-RDF-ementa	/ADJ/econômico	1	
10	PL-2003-00005-RDF-ementa	/NOUN/desenvolvimento	1	
11	PL-2003-00005-RDF-ementa	/ADJ/instalado	1	
12	PL-2003-00005-RDF-ementa	/NOUN/aire	1	
13	PL-2003-00005-RDF-ementa	/NOUN/N°	1	
14	PL-2003-00005-RDF-ementa	/ADJ/integrado	1	
15	PL-2003-00005-RDF-ementa	/ADJ/ODILON	1	
16	PL-2003-00005-RDF-ementa	/NOUN/PRÓ-DF	1	
17	PL-2003-00005-RDF-ementa	/NOUN/estabelecimento	1	
18	PL-2003-00005-RDF-ementa	/NOUN/programa	1	
19	PL-2003-00005-RDF-ementa	/NOUN/deputado	1	
20	PL-2003-00005-RDF-ementa	/NOUN/providência	1	
21	PL-2003-00005-RDF-ementa	/NOUN/benefício	1	
22	PL-2003-00005-RDF-ementa	/NOUN/promoção	1	

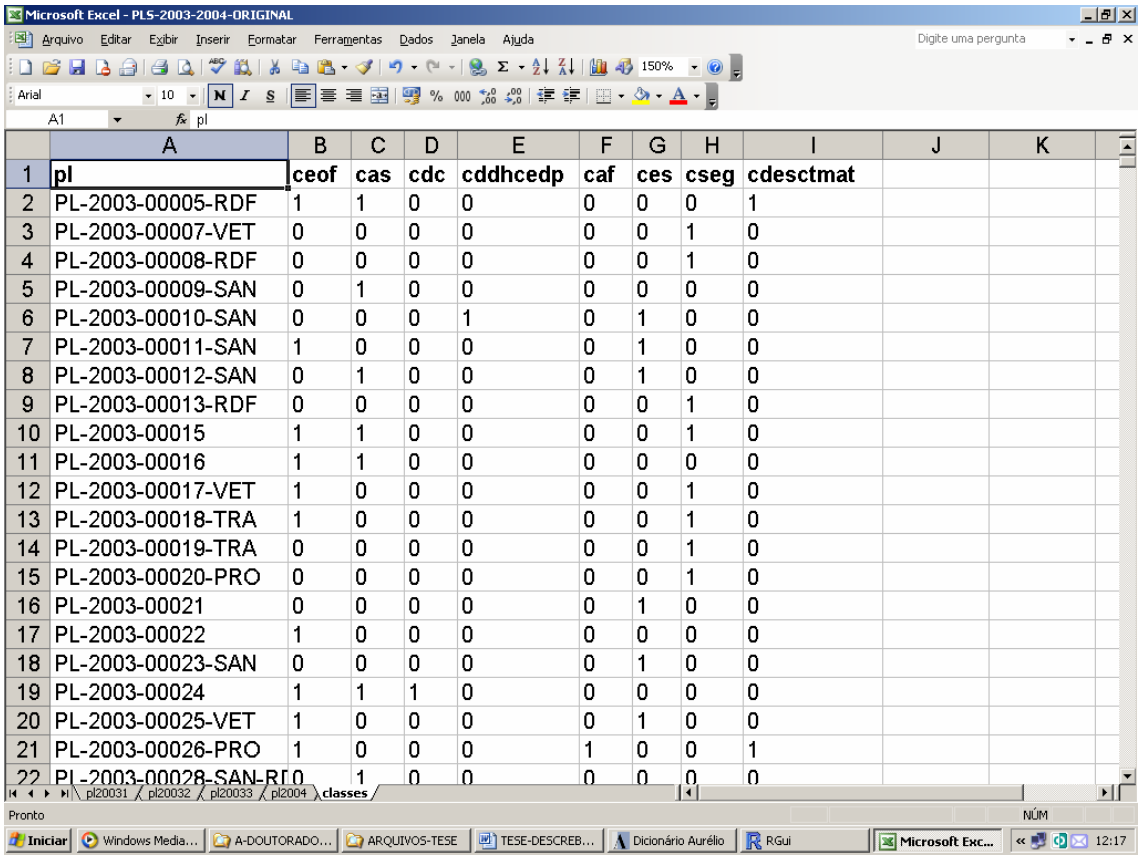
Figura 3.5 – Trecho de uma planilha Excel com dados das ementas dos PLs.



	A	B	C	D	E
1	**pl**	**palavra**	**freq**		
2	PL-2003-00005-RDF	/VERB/dar	2		
3	PL-2003-00005-RDF	/ADJ/sustentável	2		
4	PL-2003-00005-RDF	/NOUN/empreendimento	2		
5	PL-2003-00005-RDF	/NOUN/nobre	1		
6	PL-2003-00005-RDF	/NOUN/dinamização	1		
7	PL-2003-00005-RDF	/NOUN/providência	1		
8	PL-2003-00005-RDF	/NOUN/público	1		
9	PL-2003-00005-RDF	/NOUN/redação	1		
10	PL-2003-00005-RDF	/NOUN/aprovação	1		
11	PL-2003-00005-RDF	/ADJ/integrado	2		
12	PL-2003-00005-RDF	/NOUN/§	2		
13	PL-2003-00005-RDF	/NOUN/JUSTIFICAÇÃO	1		
14	PL-2003-00005-RDF	/NOUN/expedição	2		
15	PL-2003-00005-RDF	/NOUN/atividade	2		
16	PL-2003-00005-RDF	/NOUN/constante	1		
17	PL-2003-00005-RDF	/VERB/viabilizar	1		
18	PL-2003-00005-RDF	/NOUN/estabelecimento	2		
19	PL-2003-00005-RDF	/ADJ/quantitativo	1		
20	PL-2003-00005-RDF	/NOUN/uso	1		
21	PL-2003-00005-RDF	/NOUN/julho	1		
22	PL-2003-00005-RDF	/NOUN/plano	2		

Figura 3.6 – Trecho de uma planilha Excel com dados dos PLs.

Os dados corrigidos sobre as comissões por onde passou cada PL foram colocados em uma planilha Excel no mesmo arquivo contendo as quatro planilhas com as informações referentes aos termos extraídos dos PLs. A Figura 3.7 a seguir mostra um trecho com essas categorizações manualmente efetuadas (categorização feita pelo especialista no assunto), onde o valor "1" significa que o projeto passou pela comissão, e o valor "0" significa que o projeto não passou pela comissão. Os valores "0" e "1" foram empregados para facilitar cálculos a serem realizados com esses dados, e serão substituídos por "-1" e "1", respectivamente, antes de rodar o *software* SVM^{light} , que será utilizado neste trabalho e exige esses valores.



	A	B	C	D	E	F	G	H	I
1	pl	ceof	cas	cdc	cddhcedp	caf	ces	cseg	cdesctmat
2	PL-2003-00005-RDF	1	1	0	0	0	0	0	1
3	PL-2003-00007-VET	0	0	0	0	0	0	1	0
4	PL-2003-00008-RDF	0	0	0	0	0	0	1	0
5	PL-2003-00009-SAN	0	1	0	0	0	0	0	0
6	PL-2003-00010-SAN	0	0	0	1	0	1	0	0
7	PL-2003-00011-SAN	1	0	0	0	0	1	0	0
8	PL-2003-00012-SAN	0	1	0	0	0	1	0	0
9	PL-2003-00013-RDF	0	0	0	0	0	0	1	0
10	PL-2003-00015	1	1	0	0	0	0	1	0
11	PL-2003-00016	1	1	0	0	0	0	0	0
12	PL-2003-00017-VET	1	0	0	0	0	0	1	0
13	PL-2003-00018-TRA	1	0	0	0	0	0	1	0
14	PL-2003-00019-TRA	0	0	0	0	0	0	1	0
15	PL-2003-00020-PRO	0	0	0	0	0	0	1	0
16	PL-2003-00021	0	0	0	0	0	1	0	0
17	PL-2003-00022	1	0	0	0	0	0	0	0
18	PL-2003-00023-SAN	0	0	0	0	0	1	0	0
19	PL-2003-00024	1	1	1	0	0	0	0	0
20	PL-2003-00025-VET	1	0	0	0	0	1	0	0
21	PL-2003-00026-PRO	1	0	0	0	1	0	0	1
22	PL-2003-00028-SAN-RI	0		1	0	0	0	0	0

Figura 3.7 – Trecho da planilha Excel com dados sobre as comissões por onde passou cada PL.

Por fim, foi criado um arquivo Excel contendo oito planilhas - uma por comissão permanente. Cada planilha contém uma lista de termos extraídos dos 1.014 PLs estudados, que inclui termos presentes nos artigos que descrevem as matérias de competência das comissões permanentes e também termos relacionados a essas matérias. A Figura 3.8 a seguir mostra um trecho da planilha contendo as palavras relacionadas às competências da Comissão de Economia, Orçamento e Finanças (CEOF).

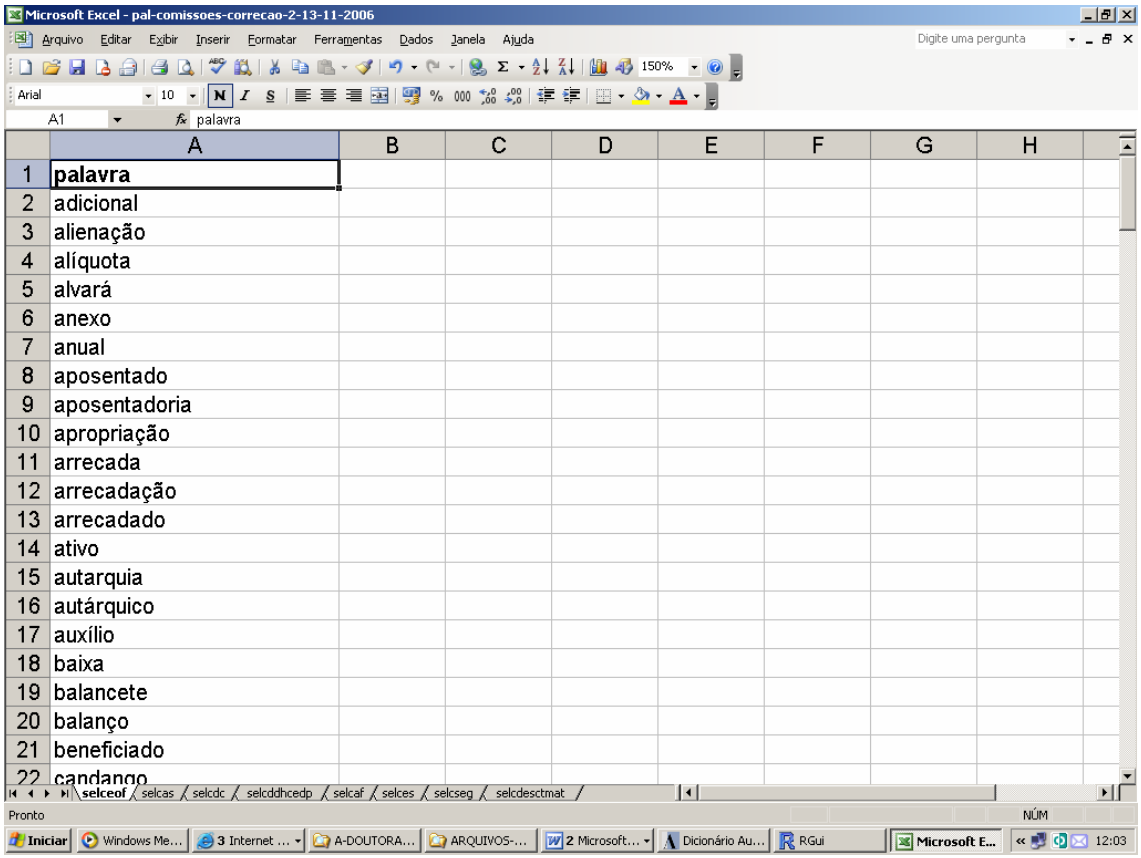


Figura 3.8 – Trecho da planilha Excel com dados relacionados às competências da comissão CEOF.

Desse modo, os arquivos usados para gerar todas as análises desta tese foram: 1) um arquivo Excel com uma planilha contendo os dados extraídos das ementas (Figura 3.5); 2) um arquivo Excel com cinco planilhas, as quatro primeiras contendo os dados extraídos dos PLs (Figura 3.6) e a quinta contendo os dados sobre as classificações dos PLs nas comissões (Figura 3.7); 3) um arquivo Excel com oito planilhas, cada uma contendo palavras relacionadas às competências da respectiva comissão permanente (Figura 3.8).

3.4 Análise dos dados

3.4.1 *Softwares* utilizados e programas desenvolvidos

Para esta fase foram utilizados dois softwares: o *software* livre R - *The* R *Foundation for Statistical Computing*, Versões 2.3.1/2.4.1[21] e o *software* SVM^{light} [22], uma implementação do algoritmo *Support Vector Machines*, cuja utilização é gratuita para propósitos científicos.

O R é um sistema para análises estatísticas e gráficas com facilidades para manipulação de dados multivariados, a partir de cálculos usando operações com matrizes, sendo amplamente utilizado pela comunidade estatística. Ele é um *software* e também uma linguagem de programação considerada como um dialeto da linguagem S, criada pela AT&T Bell Laboratories, e disponibilizada como o *software* S-PLUS comercializado pela Insightful.

O R funciona de forma parecida ao Matlab, contendo várias ferramentas para a área de aprendizado de máquina, como árvores de decisão, redes neurais, regressão logística, análise discriminante, *k*-NN, *k*-means, SOM, etc.

Praticamente no final desta pesquisa[23], verificou-se a inclusão no R do pacote kernlab. Esse pacote é composto por métodos de aprendizado de máquina baseados em kernel, para classificação, regressão, agrupamento, detecção de novidade e redução de dimensionalidade, e inclui uma implementação do algoritmo *Support Vector*

[21] Disponível em: www.r-project.org. Acesso em: 05 fev. 2007.
[22] Disponível em: http://svmlight.joachims.org/. Acesso em: 05 set. 2006.
[23] Ao substituir a versão 2.3.1 (versão sendo usada) pela 2.4.1 (última versão disponível no site, em 05 fev. 2007), devido à necessidade de troca do HD do computador usado para realizar as análises.

Machines. Como essa descoberta ocorreu tardiamente, a implementação do algoritmo SVMs utilizada neste trabalho foi a do SVM^{light}.

A partir de um conjunto de programas especificamente desenvolvidos, o R foi utilizado para: 1) limpeza dos termos sem conteúdo semântico extraídos pela ferramenta IdeXMLClient; 2) seleção dos subconjuntos para validação cruzada; 3) cálculo das métricas de seleção de termos descritas na subseção 2.4.1; 4) cálculo dos pesos para os termos, segundo as propostas listadas na seção 3.2; 5) montagem das matrizes de treinamento e teste, de acordo com as representações vetoriais de interesse para estudo (dicionário global/local, seleção de termos e atribuição de pesos); 6) geração dos arquivos de entrada para o SVM^{light}; 7) cálculo das medidas de performance para o classificador SVMs; e 8) montagem das tabelas e análises gráficas da tese.

As medidas de desempenho para o classificador SVMs foram calculadas no R, uma vez que o SVM^{light} fornece apenas os valores para acurácia, precisão e revocação por categoria (comissão, no caso sob estudo), não calculando a medida F_1 e também não calculando as medidas globais considerando o desempenho do classificador no conjunto completo de categorias (médias micro e macro).

Como o SVM^{light} gera um arquivo contendo os valores da função de decisão para os documentos do conjunto de teste, esses valores foram lidos no R para o cálculo, por comissão, dos valores de acurácia, precisão, revocação, F_1, e dados da matriz de confusão – Tabela 2.8, seção 2.6 (número de PLs da comissão classificados correta e incorretamente pelo classificador, e número de PLs que não são da comissão classificados correta e incorretamente pelo classificador). Essas medidas também foram calculadas para o conjunto completo de categorias pelas médias micro e macro.

Para a classificação de um PL na comissão de interesse, foi considerado o sinal do valor da função de decisão para o respectivo documento, conforme descrito nas especificações do SVM^{light} : valor positivo, classe positiva; valor negativo, classe negativa.

As análises realizadas nesta pesquisa foram rodadas em um computador AMD Sempron TM 2300+ 1.58 GHz, 2,00 GB de RAM.

3.4.2 Validação Cruzada

Com a finalidade de avaliar como deverá ser o desempenho dos classificadores construídos em PLs futuros, ainda não vistos, foi utilizada a validação cruzada estratificada com divisão do corpus em cinco subconjuntos.

Esses subconjuntos foram selecionados seqüencialmente, repetindo cada tentativa de seleção antes de passar para a próxima, até obter em cada subconjunto proporção de PLs por comissão o mais próximo possível à proporção observada no corpus.

Sempre procurando manter essa proporção, os subconjuntos foram aleatoriamente selecionados entre os casos do corpus ainda não selecionados nos subconjuntos anteriores. O número de PLs em cada um dos cinco conjuntos de treinamento e teste formados com os subconjuntos selecionados é mostrado na Tabela 3.3 a seguir.

Tabela 3.3 - Número de PLs por comissão, para os conjuntos de treinamento (TR) e teste (TE) em cada repetição da validação cruzada.

CV	CONJUNTO	COMISSÃO							
		CEOF	CAS	CDC	CDDHCEDP	CAF	CES	CSEG	CDESCTMAT
CV1	TR	472	369	80	46	94	164	119	151
	TE	118	85	22	10	19	41	29	38
CV2	TR	471	365	84	44	89	168	121	152
	TE	119	89	18	12	24	37	27	37
CV3	TR	465	351	83	45	89	160	118	153
	TE	125	103	19	11	24	45	30	36
CV4	TR	482	372	77	45	90	163	116	146
	TE	108	82	25	11	23	42	32	43
CV5	TR	470	359	84	44	90	165	118	154
	TE	120	95	18	12	23	40	30	35

Todos os cálculos referentes às medidas de desempenho dos classificadores foram realizados usando a validação cruzada com os dados conforme as divisões descritas na Tabela 3.3.

Deve-se ressaltar que não foram utilizados procedimentos mais sofisticados devido à dimensão do espaço de atributos que torna lentas as execuções dos programas, e também devido à grande quantidade de arquivos a serem gerados para a realização das análises de interesse.

Para cada representação vetorial estudada, 1.120 arquivos são gerados. Desses, 560 (8*7*5*2) são criados, pois os cálculos são realizados para oito comissões permanentes, sete formas distintas de atribuição de pesos para os termos, cinco repetições da validação cruzada, e dois arquivos - de treinamento e de teste. Os outros 560 arquivos são gerados a partir desses 560, com a execução do programa SVM^{light} - 280 com os modelos de classificação gerados e 280 com os valores da função de decisão para os PLs dos conjuntos de teste.

3.4.3 Análise dos termos extraídos pela ferramenta IdeXmlClient

Devido à limitação no número de linhas das planilhas Excel, para uma melhor visualização dos termos extraídos pela ferramenta IdeXMLClient, o primeiro passo foi juntar no R as quatro planilhas Excel com os dados das palavras contidas nos 1.014 PLs e suas respectivas freqüências de ocorrência, o que gerou uma matriz com 206.003 linhas (registros) e 3 colunas (variáveis: "pl", "palavra" e "freq").

Nesses 1.014 PLs foram identificados 17.211 termos distintos, a partir dos quais as seguintes constatações foram realizadas:

- números escritos por extenso e números arábicos e romanos quando aparecem isolados, como em "oito", "09", "I", "IV" não são identificados como termos. Porém, esses são reconhecidos como termos, quando aparecem junto com algum outro símbolo, como em "1°", "5°-", "XXIII.", "II.", "V-", "1,5%", "22ª"; "00.099.754/0003-02", "10-A";
- são reconhecidos como termos, números referentes a anos, como "2003" e "95";
- são reconhecidos como termos, símbolos como: aspas duplas (" "), aspas simples (' '), "§", "-", "nº", "n.º", "Nº", "&", "§6º", "*";
- as palavras "Art." e "art." são substituídas pela palavra "artículo". Porém, quando a palavra "artigo" é encontrada, ela é reconhecida como "artigo" e não é substituída por "artículo". Se a abreviação "Art." aparece junto com o número do artigo, sem espaço em branco, por exemplo, "Art.5º", a ferramenta reconhece o termo da forma como aparece, ou seja, "Art.5º";
- as vírgulas não são descartadas quando aparecem incorretamente - sem espaço em branco - junto à palavra ou número à sua direita, como em ",TLP", ",14";

- em “Taxa de Limpeza Pública-TLP”, como não existe espaço em branco entre a palavra “Pública” e o “-”, a ferramenta identifica “Pública-TLP” como uma única palavra;
- as aspas são reconhecidas junto às palavras (ex: Constituição”, consumidores.”, “cidade, “Condomínios”, ciência;”, ‘lambe-lambe’). Nessas situações, as palavras são reconhecidas como estão escritas: não há conversão para minúscula, e as palavras também não são reduzidas para o seu *lemma* (ex.: escolares”, se não tivesse as aspas, seria reduzida para escolar);
- palavras que aparecem no texto separadas por /, como em custo/benefício ou artistas/bandas são identificadas como uma única palavra;
- algumas substituições de palavras não acrescentam conteúdo semântico importante para o corpus sob estudo, mas também não causam nenhum prejuízo. Em nomes de cidades e países, como Belo Horizonte e Estados Unidos da América, as palavras Belo e Horizonte são substituídas por Belo_Horizonte, e as palavras Estados, Unidos, da, América por Estados_Unidos;
- as siglas reconhecidas pelo IdeXmlClient são substituídas pelo termo que a ferramenta acredita que ela representa. Porém, o significado da sigla quando escrito por extenso é identificado palavra por palavra. Por exemplo, quando a ferramenta encontra a sigla STF, substitui por Supremo_Tribunal_Federal. Porém, Supremo Tribunal Federal, por extenso, é identificado como três palavras distintas: Supremo, Tribunal e Federal;
- algumas siglas são incorretamente substituídas. Por exemplo, a sigla IPMF é substituída por imposto_provisório_sobre_movimentação_financeira, mas no texto considerado, IPMF significa Igreja Pentecostal Missão da Fé; a sigla TLP é

identificada como Telefones_de_Lisboa_e_Porto, mas, nos textos em que aparece significa Taxa de Limpeza Pública. Essas substituições, mesmo incorretamente realizadas, não causam prejuízo para as análises a serem realizadas, pois a mesma sigla sempre é identificada da mesma forma, não afetando assim o cálculo do número de vezes em que ela aparece nos textos nem a quantidade de textos em que ela aparece;

- "m^2" é identificado como m^2, "m2" é substituído por metro_quadrado, e "metro quadrado" é substituído por "metro" e "quadrar". Essas identificações distintas para o mesmo termo serão unificadas posteriormente;
- a palavra "zinco" é substituída pela palavra zincar. Em "terça parte", a palavra "terça" é substituída pela palavra "terçar". Essas substituições, apesar de incorretas, não causam prejuízo para as análises de interesse.

Com base na análise dos 17.211 termos distintos identificados pela ferramenta IdeXMLClient, as seguintes correções foram realizadas: os símbolos indesejados foram eliminados e as palavras identificadas junto com sinais de pontuação foram corrigidas de forma semi-automática, a partir de um programa desenvolvido no R. As substituições das siglas por seus significados e de uma palavra por outra, mesmo quando incorretamente realizadas são inócuas para a análise de interesse e não foram corrigidas.

Para identificar as correções necessárias, foi gerado pelo R um arquivo txt com os termos distintos extraídos pelo IdeXMLClient. Identificados os símbolos indesejados, esses foram removidos usando um conjunto de comandos do R, e um novo arquivo txt com as correções foi gerado. Nesse novo arquivo gerado, os termos

indesejados remanescentes foram manualmente removidos, gerando o arquivo de termos desejado.

Com a limpeza dos termos gerados pela ferramenta IdeXmlClient, o número de termos distintos no corpus caiu de 17.211 para 14.795, ou seja, 14% dos termos identificados pelo IdeXmlClient representavam símbolos indesejados ou termos que poderiam ter sido agrupados com outros se corretamente identificados.

3.4.4 Escolha de uma representação vetorial base para os PLs

Com o objetivo de definir uma representação vetorial base para os PLs, a partir da qual todos os estudos serão desenvolvidos, a primeira análise consistiu em examinar o desempenho do algoritmo SVMs, considerando os termos extraídos dos PLs de acordo com as classes gramaticais adjetivo, substantivo e verbo.

As seguintes representações vetoriais foram pesquisadas:

1. **verb**: utiliza apenas os verbos extraídos dos PLs. Termos da forma "/VERB/.... ". Exemplo: /VERB/dar;
2. **adj**: utiliza apenas os adjetivos extraídos dos PLs. Termos da forma "/ADJ/...". Exemplo: /ADJ/sustentável.;
3. **noun**: utiliza apenas os substantivos (noun) extraídos dos PLs. Termos da forma "/NOUN/...". Exemplo: /NOUN/empreendimento;
4. **na**: utiliza os substantivos (n) e adjetivos (a) extraídos dos PLs, com as identificações "/NOUN/..." e "/ADJ/...". Exemplos: /ADJ/sustentável; /NOUN/empreendimento;

5. **nav**: utiliza os substantivos (n), adjetivos (a) e verbos (v) extraídos dos PLs, com as identificações "/NOUN/...", "/ADJ/..." e "/VERB/..." (v). Exemplos: /NOUN/empreendimento; /ADJ/sustentável; /VERB/dar.

Para efetuar as comparações de interesse, os pesos para os termos foram calculados segundo as formas de atribuição de pesos descritas na seção 3.2, e a medida de desempenho adotada foi a F_1 macro, pois o objetivo neste trabalho é dar igual importância a todos os problemas de classificação nas comissões, não privilegiando as comissões que, por terem mais exemplos positivos, dominam as medidas computadas por micro.

Nesta fase de definição da representação vetorial base para os PLs, as análises foram realizadas usando apenas as representações vetoriais com dicionário global. Os resultados obtidos são mostrados na Tabela 3.4 e Figura 3.9 a seguir.

Tabela 3.4 – F_1 Macro segundo os pesos e vetores de termos descritos.

Vetor de termos	Nº termos[24]	Peso						
		TF_ABSL	TF_BNS	TF_GR	TF_IDF	TF_IG	TF_OR	TF_QUI
verb	2.263	0,36	0,32	0,238	0,358	0,25	0,371	0,264
adj	3.698	0,573	0,541	0,541	0,551	0,499	0,543	0,54
noun	8.764	0,661	0,637	0,581	0,629	0,59	0,633	0,582
na	12.532	0,688	0,658	0,634	0,64	0,624	0,652	0,634
nav	14.795	0,671	0,641	0,631	0,62	0,625	0,653	0,633

[24]Calculado sobre o corpus pela facilidade de obtenção, e apenas para se ter idéia do número de termos envolvidos em cada representação vetorial estudada. Em todas as tabelas que apresentam esse número, ele é calculado sobre o corpus.

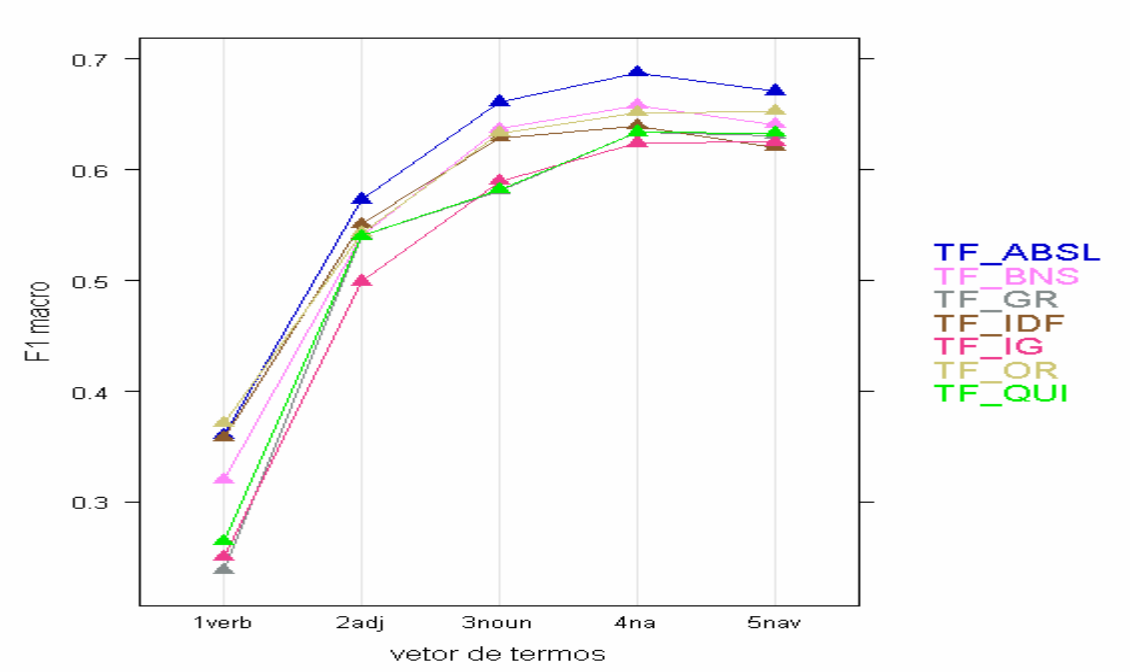


Figura 3.9 – F_1 macro para os vetores de termos descritos no gráfico, com pesos calculados conforme mostrados na legenda.

A partir da Tabela 3.4 e Figura 3.9, pode-se verificar que, conforme esperado, entre as classes gramaticais adjetivo, substantivo e verbo, a dos verbos (vetor de termos "verb") é a menos importante, e a dos substantivos (vetor de termos "noun") a mais importante para representar os PLs, exibindo esta última, performance não muito inferior à representação considerando adjetivos e substantivos (vetor de termos "na"), e à considerando adjetivos, substantivos e verbos (vetor de termos "nav").

Também pode ser constatado, dos dados apresentados, que os maiores valores para a medida F_1 macro são, em geral, atingidos com a representação dos PLs pelos adjetivos e substantivos extraídos de seus textos (vetor de termos "na"). Quando os verbos são incluídos a essa representação (vetor de termos "nav"), o desempenho das SVMs apresenta uma queda ou um resultado praticamente inalterado, segundo todas as formas de atribuição de pesos consideradas.

Nas representações vetoriais estudadas, no entanto, os termos utilizados são compostos pelas palavras juntamente com a identificação de suas respectivas classes gramaticais, pois assim foram extraídos dos PLs (vide Figuras 3.4 a 3.6). Isso representa um inconveniente para análises posteriores a serem realizadas.

Visando, então, investigar a possibilidade de se trabalhar com uma representação vetorial mais simples, eliminando a identificação das classes gramaticais dos termos, a análise realizada foi complementada com a comparação de algumas outras representações vetoriais. Também foi incluída nessa segunda análise, para efeitos de avaliação de desempenho, a representação vetorial usando os termos da forma como foram identificados pela ferramenta IdeXmlClient, sem limpeza. As representações comparadas são descritas a seguir:

1. **orig**[25]: utiliza os substantivos, adjetivos e verbos extraídos dos PLs, com as identificações "/NOUN/...", "/ADJ/..." e "/VERB/...", sem limpeza dos termos extraídos pela ferramenta IdeXmlClient. Exemplos: /NOUN/N°, /ADJ/defensiva", /VERB/dar;

2. **nav**: utiliza os substantivos (n), adjetivos (a) e verbos (v) extraídos dos PLs, com as identificações "/NOUN/...", "/ADJ/..." e "/VERB/...", após limpeza dos termos extraídos pela ferramenta IdeXmlClient. Exemplos: /NOUN/empreendimento; /ADJ/sustentável; /VERB/dar;

3. **snav**: utiliza os substantivos (n), adjetivos (a) e verbos (v) extraídos dos PLs, sem (s) as identificações "/NOUN/...", "/ADJ/...", e "/VERB/...". Exemplos: empreendimento; sustentável; dar;

[25] A única representação vetorial que utiliza os termos sem limpeza é a "orig" .

4. **na**: utiliza os substantivos (n) e adjetivos (a) extraídos dos PLs, com as identificações "/NOUN/..." e "/ADJ/...". Exemplos: /ADJ/sustentável; /NOUN/empreendimento;

5. **sna**: utiliza os substantivos (n) e adjetivos (a) extraídos dos PLs, sem (s) as identificações "/NOUN/..." e "/ADJ/...". Exemplos: sustentável; empreendimento;

6. **nastp**: utiliza os substantivos (n) e adjetivos (a) extraídos dos PLs, com as identificações "/NOUN/..." e "/ADJ/..."; elimina os termos que ocorreram em apenas um PL de treinamento; elimina os termos que não acrescentam conteúdo semântico importante ao domínio considerado (*stop words* (stp)); e padroniza a escrita de alguns termos para os quais são utilizadas mais de uma forma de escrita, como detran/detran-df e art/arts/artículo/artigo;

7. **snastp**: utiliza os substantivos (n) e adjetivos (a) extraídos dos PLs, sem (s) as identificações "/NOUN/..." e "/ADJ/..."; elimina os termos que ocorreram em apenas um PL de treinamento; elimina os termos que não acrescentam conteúdo semântico importante ao domínio considerado (*stop words* (stp)); e padroniza a escrita de alguns termos para os quais são utilizadas mais de uma forma de escrita, como detran/detran-df e art/arts/artículo/artigo.

A diferença entre as representações "nav" (item 2.) e "snav" (item 3.) é que eliminando a identificação das classes gramaticais, a mesma palavra, no mesmo PL, ora identificada como adjetivo, ora como substantivo, pode ser agrupada como uma única palavra, com suas freqüências de ocorrência somadas. Observação semelhante vale para as representações "na" (item 4.) e "sna" (item 5.) .

Na representação "nastp" (item 6.), o agrupamento de palavras no mesmo PL também é feito, porém é devido à padronização da escrita de algumas palavras. Na representação "snastp" (item 7.), o agrupamento de palavras é realizado em razão da

eliminação da identificação da classe gramatical adjetivo/substantivo, assim como da padronização da escrita de algumas palavras.

Os resultados obtidos para as representações vetoriais descritas nos itens 1. a 7. são mostrados na Tabela 3.5 e Figura 3.10 a seguir.

Tabela 3.5 – F_1 Macro segundo os pesos e vetores de termos descritos na tabela.

Vetor de termos	Nº de termos	Peso						
		TF_ABSL	TF_BNS	TF_GR	TF_IDF	TF_IG	TF_OR	TF_QUI
Orig	17.211	0,671	0,638	0,63	0,614	0,625	0,642	0,633
nav	14.795	0,671	0,641	0,631	0,62	0,625	0,653	0,633
snav	13.776	0,675	0,646	0,63	0,625	0,619	0,65	0,63
na	12.532	0,688	0,658	0,634	0,64	0,624	0,652	0,634
sna	11.590	0,677	0,652	0,637	0,644	0,617	0,646	0,629
nastp	6.477	0,702	0,679	0,637	0,648	0,629	0,668	0,63
snastp	5.723	0,698	0,671	0,633	0,642	0,624	0,661	0,628

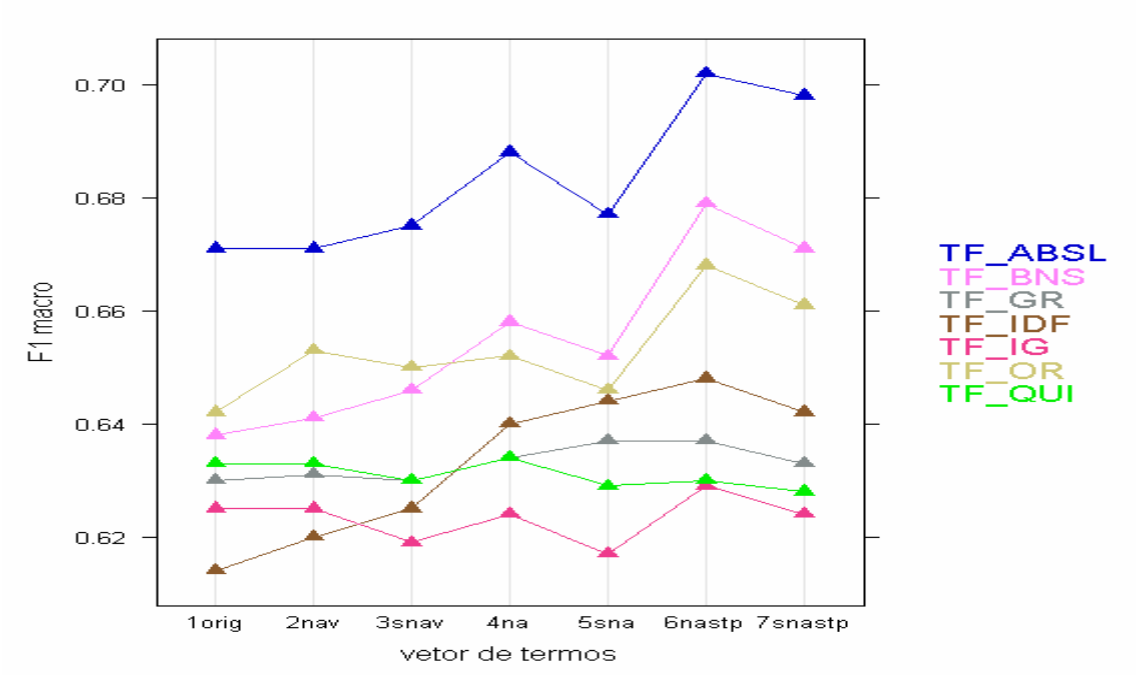


Figura 3.10 – F_1 macro para os vetores de termos descritos no gráfico, com pesos calculados conforme mostrados na legenda.

Um fato surpreendente constatado a partir da Tabela 3.5 e Figura 3.10 apresentados, é que a representação dos PLs com os termos da forma como extraídos pela ferramenta IdeXmlClient (vetor de termos "orig") - incluindo 14% de símbolos indesejados ou termos que poderiam ter sido agrupados com outros se corretamente identificados - apresenta performance próxima a do vetor de termos limpo, sem esses termos (vetor de termos "nav"). Uma possível explicação para tal fato pode ser a limpeza de incorreções semelhantes em praticamente todos os PLs.

Outras duas constatações são evidenciadas e confirmadas nesta segunda análise: 1) o método de atribuição de pesos TF_ABSL, proposto neste trabalho, é o que apresenta o melhor resultado para F_1 macro entre os métodos considerados, e os pesos calculados por TF_BNS, igualmente não utilizados em trabalhos anteriores, também estão entre os melhores resultados exibidos; 2) o desempenho do algoritmo SVMs depende do método de atribuição de pesos utilizado, e esses não apresentam a mesma tendência ao longo dos vetores de termos estudados.

Essas duas últimas constatações citadas guiaram a escolha da representação vetorial base buscada. Assim, comparando inicialmente as cinco primeiras representações vetoriais da Tabela 3.5 e Figura 3.10, verifica-se que o vetor de termos "na" - que considera os adjetivos e substantivos com a identificação das classes gramaticais junto às palavras - continua sendo o melhor candidato para obtenção da representação vetorial base dos PLs, uma vez que apresenta o maior valor para F_1 macro, que é atingido com os pesos calculados por TF_ABSL.

No entanto, na presente pesquisa, além de não importar se a palavra trata-se de um adjetivo ou um substantivo, essa forma de representação dos termos é de difícil manuseio. Um dos inconvenientes seria para a produção da lista de termos que não acrescentam conteúdo semântico importante ao domínio estudado. Outro inconveniente

estaria na elaboração das listas de palavras relacionadas às matérias de competência das comissões permanentes a serem utilizadas no estudo sobre a importância dessas palavras na efetividade da categorização.

A próxima opção, depois do vetor de termos "na", seria utilizar o vetor "sna" (item 5.) para obter o vetor de termos base. Esse vetor apresenta o segundo maior valor para F_1 macro - também atingido com os pesos calculados por TF_ABSL - tendo a vantagem de não utilizar a identificação das classes gramaticais adjetivo/substantivo junto às palavras.

Para melhor avaliar as implicações decorrentes de se utilizar essa segunda opção ao invés da primeira, dois vetores foram criados: "nastp" (item 6.) e "snastp" (item 7.). Um desses dois vetores seria criado em decorrência de se excluir da representação vetorial escolhida – "na" ou "sna", respectivamente -, principalmente, os termos que não acrescentam conteúdo semântico ao domínio estudado. O vetor criado corresponderia à representação vetorial base buscada.

O vetor de termos "snastp" (item 7.) foi construído a partir do "sna" (item 5.), que possuía 11.564 termos. Primeiramente, foram eliminadas as palavras presentes em apenas um dos 1.014 PLs do corpus, que, como pode ser observado na Tabela 3.6 adiante, correspondiam a 47,85% das palavras distintas identificadas. Essas palavras foram eliminadas, pois dificilmente estariam presentes em PLs futuros, não sendo, portanto, consideradas úteis para a construção do classificador.

Após essa eliminação, as 6.031 palavras remanescentes foram analisadas uma a uma para verificar as que não acrescentavam conteúdo semântico ao domínio considerado, e padronizar a escrita das palavras mais relevantes ao problema para as quais foram utilizadas mais de uma forma de redação. Terminado esse processo, o vetor de termos "snastp" ficou com 5.723 palavras.

Tabela 3.6 – Número de PLs por palavra no vetor de termos “sna”: Freqüências e Porcentagens.

Nº de PLs em que a palavra está presente	Freqüência de palavras	Porcentagem	Porcentagem Acumulada
1	5533	47,85	47,85
2	1512	13,08	60,92
3	784	6,78	67,70
4	481	4,16	71,86
5	302	2,61	74,47
6	275	2,38	76,85
7	219	1,89	78,74
8	192	1,66	80,40
9	144	1,25	81,65
10	113	0,98	82,63
11	112	0,97	83,60
12	86	0,74	84,34
13	92	0,80	85,13
14	76	0,66	85,79
15	61	0,53	86,32
16	50	0,43	86,75
17	61	0,53	87,28
18	55	0,48	87,76
19	58	0,50	88,26
20	46	0,40	88,65
21	47	0,41	89,06
22	46	0,40	89,46
23	41	0,35	89,81
24	39	0,34	90,15
≥ 25	1139	9,85	100,00
Total	11564	100,00	-

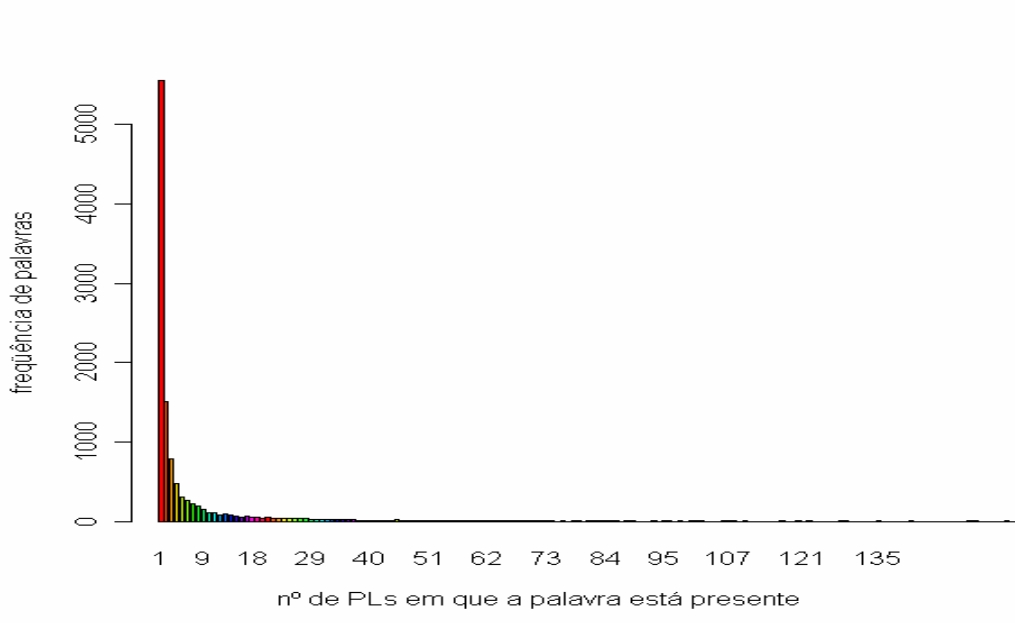


Figura 3.11 – Nº de PLs em que a palavra está presente, considerando os 1.014. PLs do corpus e o vetor de termos “sna” (dados da Tabela 3.6).

A partir do vetor de termos “snastp” (item 7.), que não utiliza a identificação das classes gramaticais adjetivo/substantivo, sendo portanto mais fácil de trabalhar, foi montado o vetor de termos “nastp” (item 6.), que utiliza a identificação das classes gramaticais.

Ao fazer as adaptações necessárias no vetor de termos “snastp” para gerar o “nastp”, mais um inconveniente de se trabalhar com a representação vetorial “na” foi verificado: algumas classes gramaticais para as palavras foram incorretamente identificadas pela ferramenta IdeXmlClient. As palavras “Detran” e “PDOT” (Plano Diretor de Ordenamento Territorial do Distrito Federal), por exemplo, que são substantivos, ora foram identificadas como adjetivo, ora como substantivo. Essa constatação reforçou ainda mais a decisão em favor de se adotar o vetor de termos “snastp” como representação vetorial base para os PLs.

Em comparação ao vetor de termos “nastp”, o vetor “snastp” apresenta valor para F_1 macro 0,57% inferior, quando os pesos são calculados por TF_ABSL, que exibem o melhor desempenho para todos os vetores considerados. Quando os pesos são calculados por TF_BNS, que apresentam o segundo melhor resultado para os vetores comparados, essa queda em F_1 macro é de 1,18%. Esses valores, no entanto, podem ser ainda menores, posto que a classe gramatical para algumas palavras foi identificada incorretamente pela ferramenta IdeXmlClient, sendo que apenas algumas palavras foram analisadas quanto a tal fato.

Dadas as considerações expostas, o vetor escolhido como representação vetorial base para os PLs foi o “snastp”. Esse vetor possui as seguintes características: 1) não apresenta a identificação da classe gramatical junto às palavras e é formado pelos adjetivos e substantivos extraídos dos PLs, padronizando a escrita de algumas palavras redigidas de mais de uma forma e excluindo as palavras que não acrescentam conteúdo

semântico ao domínio considerado; 2) o número de PLs em que cada palavra está presente varia de 2 a 417, com média de 18,85 e desvio padrão de 38,45; 3) o número de palavras por PL varia de 4 a 1.162, com média de 106,4 e desvio padrão de 94,97.

Nas subseções seguintes, usando como base o vetor "snastp", serão apresentados estudos sobre redução de dimensionalidade e sobre aumento de pesos para as palavras presentes nas ementas e para as relacionadas às matérias de competência das comissões permanentes. Primeiro, os estudos serão realizados separadamente. Para a análise final, os melhores resultados obtidos em cada estudo serão considerados em conjunto.

3.4.5 Estudos sobre redução de dimensionalidade

Considerando dicionário global e local, os estudos nesta subseção consistiram em avaliar o efeito da redução de dimensionalidade no desempenho do algoritmo SVMs para os métodos de atribuição de pesos sob avaliação.

No primeiro estudo, foi analisada a redução da dimensionalidade pela freqüência de documentos (número de PLs em que a palavra está presente). Nesse caso, independentemente do método de atribuição de pesos utilizado, os mesmos termos são selecionados para compor o vetor de termos.

No segundo estudo, a mesma métrica usada para compor o método de atribuição supervisionada de termos foi empregada para selecionar os melhores termos. Assim, foi utilizada, por exemplo, a métrica razão de ganho para selecionar os melhores termos para o método de atribuição de pesos TF_GR.

3.4.5.1 Redução da dimensionalidade pela freqüência de documentos

Considerando dicionário global, serão apresentados na Tabela 3.7 e Figura 3.12, os valores de F_1 macro para as sete formas de atribuição de pesos estudadas, reduzindo a dimensionalidade do espaço de termos segundo o número de PLs em que o termo ocorre. A Tabela 3.8 e Figura 3.13 apresentam as mesmas quantidades para dicionário local.

Nos vetores de termos constantes da Tabela 3.7 e Figura 3.12 (Tabela 3.8 e Figura 3.13), o número seguinte às letras "df" ("dl") significa que o vetor é composto pelas palavras que ocorreram em pelo menos essa quantidade de PLs de treinamento. Por exemplo: "df02" ("dl02") - vetor de termos base "snastp" usando dicionário global (dicionário local) - é o vetor composto pelas palavras que ocorrem em dois ou mais PLs.

No caso do dicionário global, independentemente da comissão considerada, o número de termos envolvidos no vetor estudado é o mesmo, pois uma única representação vetorial é obtida para todas as comissões permanentes. Assim, essas quantidades serão apresentadas na própria Tabela 3.7. Todavia, com dicionário local, em cada vetor de termos considerado, uma representação vetorial distinta é obtida para cada comissão permanente. Nesse caso, portanto, o número de termos varia não só com o vetor de termos, como também com a comissão permanente. Dessa forma, esses valores são apresentados separadamente na Tabela 3.9.

Tabela 3.7 - F_1 macro segundo os pesos e vetores de termos descritos, reduzindo a dimensionalidade pela freqüência de documentos, considerando dicionário global.

Vetor de termos	Nº termos	Peso						
		TF_ABSL	TF_BNS	TF_GR	TF_IDF	TF_IG	TF_OR	TF_QUI
df02 (snastp global)	5.723	0,698	0,671	0,633	0,642	0,624	0,661	0,628
df03	4.315	0,698	0,676	0,639	0,661	0,625	0,676	0,632
df04	3.553	0,696	0,672	0,640	0,663	0,625	0,684	0,632
df05	3.089	0,698	0,673	0,636	0,659	0,624	0,675	0,632
df06	2.800	0,695	0,677	0,636	0,668	0,624	0,676	0,632
df07	2.534	0,694	0,675	0,636	0,666	0,628	0,673	0,632
df08	2.323	0,705	0,676	0,629	0,673	0,626	0,672	0,627
df09	2.137	0,71	0,683	0,629	0,677	0,626	0,675	0,627
df10	1.997	0,694	0,677	0,626	0,673	0,625	0,661	0,626
df12	1.778	0,688	0,665	0,621	0,666	0,621	0,648	0,624
df15	1.529	0,669	0,660	0,605	0,663	0,609	0,624	0,608
df20	1.250	0,672	0,660	0,602	0,664	0,610	0,631	0,603
df25	1.039	0,667	0,651	0,597	0,657	0,611	0,621	0,602

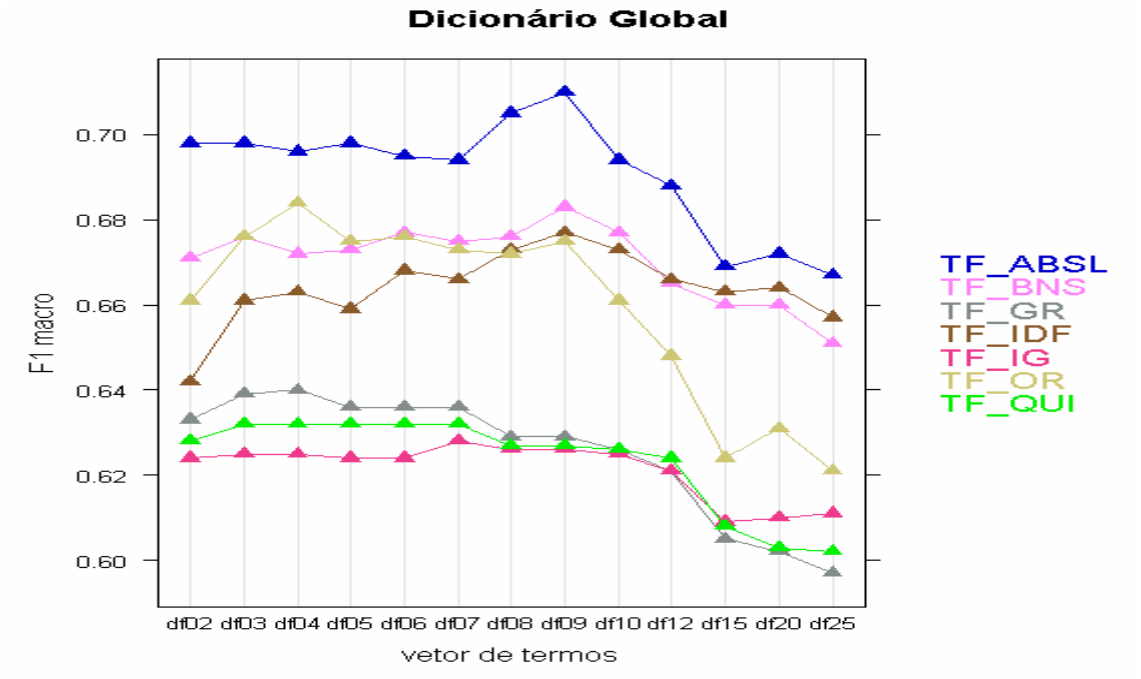


Figura 3.12 - F_1 macro segundo os pesos e vetores de termos descritos, considerando dicionário global.

Tabela 3.8 - F_1 macro segundo os pesos e vetores de termos descritos, reduzindo a dimensionalidade pela freqüência de documentos, considerando dicionário local[26].

Vetor de termos	Peso						
	TF_ABSL	TF_BNS	TF_GR	TF_IDF	TF_IG	TF_OR	TF_QUI
dl02 (snastp local)	0,705	0,702	0,643	0,671	0,643	0,66	0,628
dl03	0,71	0,705	0,643	0,672	0,643	0,653	0,628
dl04	0,717	0,711	0,644	0,675	0,644	0,66	0,624
dl05	0,712	0,71	0,643	0,666	0,643	0,663	0,623
dl06	0,72	0,713	0,647	0,665	0,647	0,67	0,629
dl07	0,724	0,712	0,65	0,672	0,65	0,673	0,63
dl08	0,725	0,715	0,648	0,666	0,648	0,678	0,631
dl09	0,725	0,716	0,65	0,676	0,65	0,683	0,633
dl10	0,704	0,701	0,634	0,672	0,634	0,662	0,614
dl12	0,694	0,694	0,627	0,666	0,627	0,648	0,606
dl15	0,672	0,678	0,591	0,664	0,591	0,61	0,572
dl20	0,666	0,677	0,583	0,663	0,583	0,601	0,568
dl25	0,66	0,666	0,564	0,652	0,564	0,589	0,544

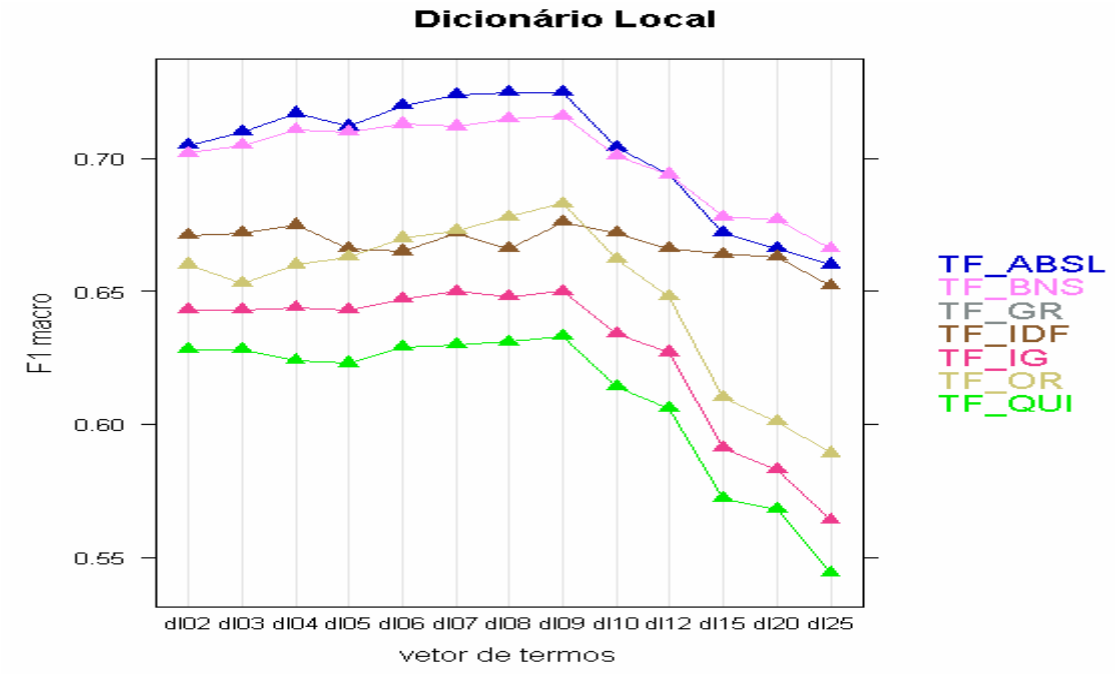


Figura 3.13 - F_1 macro segundo os pesos e vetores de termos descritos, considerando dicionário local

[26] Com dicionário local, os valores para TF_IG e TF_GR são idênticos, uma vez que a diferença no cálculo de IG e GR está na entropia H da categoria, usada como fator de normalização em GR. Localmente a uma determinada categoria, essa diferença é um fator de multiplicação constante. Logo, com a normalização dos pesos, TF_IG e TF_GR tornam-se idênticos.

Tabela 3.9 - Número de termos envolvidos em cada vetor de termos e comissão permanente, considerando dicionário global

Vetor de termos	Nº de termos por comissão							
	CEOF	CAS	CDC	CDDHECDP	CAF	CES	CSEG	CDESCTMAT
dl02 (snastp local)	4.952	4.647	2.620	1.982	3.118	3.822	2.956	4.094
dl03	4.035	3.865	2.383	1.831	2.786	3.262	2.700	3.484
dl04	3.416	3.332	2.183	1.722	2.531	2.898	2.493	3.044
dl05	3.008	2.956	2.062	1.642	2.355	2.632	2.332	2.739
dl06	2.745	2.712	1.957	1.581	2.214	2.444	2.213	2.561
dl07	2.503	2.491	1.836	1.513	2.090	2.268	2.077	2.364
dl08	2.306	2.297	1.732	1.448	1.975	2.116	1.973	2.203
dl09	2.125	2.119	1.655	1.395	1.860	1.978	1.860	2.058
dl10	1.990	1.984	1.592	1.350	1.771	1.870	1.780	1.937
dl12	1.776	1.773	1.465	1.274	1.609	1.690	1.635	1.743
dl15	1.529	1.527	1.315	1.164	1.418	1.483	1.439	1.509
dl20	1.250	1.250	1.145	1.021	1.195	1.229	1.206	1.244
dl25	1.039	1.039	979	898	1.010	1.029	1.021	1.036

Com dicionário global, como pode ser verificado a partir da Tabela 3.7 e Figura 3.12, o maior valor para F_1 macro não é atingido com a mesma redução de dimensionalidade para todas as formas de atribuição de pesos estudadas. Considerando, então, os pesos calculados por TF_ABSL, que apresentam o melhor desempenho para os vetores estudados, o maior valor para F_1 macro é atingido com o vetor de termos utilizando apenas as palavras que ocorrem em nove ou mais PLs de treinamento ("df09"). Essa redução de dimensionalidade também produz o maior valor para F_1 macro com os pesos calculados por TF_BNS e TF_IDF, que estão entre os melhores resultados apresentados para os vetores estudados. Após este pico, o aumento na redução da dimensionalidade começa a provocar uma queda no desempenho da medida F_1 macro, mesmo para as formas de atribuição de pesos que não atingem o pico para F_1 macro com o vetor "df09".

Na adoção do dicionário local, como pode ser visto na Tabela 3.8 e Figura 3.13, independentemente da forma de atribuição de pesos utilizada, o maior valor para F_1 macro é atingido com o vetor de termos "dl09", que utiliza apenas as palavras que ocorreram em nove ou mais PLs de treinamento. Com os pesos calculados por TF_ABSL - que apresentam o maior valor para F_1 macro - o máximo também ocorre com o vetor de termos "dl08". Para os pesos calculados por TF_IG/TF_GR – que estão entre as piores performances observadas - o máximo também ocorre com o vetor de termos "dl07". Após o pico em "dl09", assim como no caso do dicionário global, o aumento na redução da dimensionalidade começa a provocar uma queda no desempenho da medida F_1 macro para todas as formas de atribuição de pesos estudadas.

Das constatações feitas e considerando a escolha de um único valor para a redução da dimensionalidade pela freqüência de documentos, a melhor opção parece ser a representação vetorial que utiliza apenas as palavras que ocorrem em nove ou mais PLs, principalmente, por ser essa a que apresenta o maior valor para F_1 macro usando tanto dicionário global como local.

Comparada ao vetor de termos base "snastp" usando dicionário global ("df02"), a representação "df09" apresenta ganhos de 1,72% com TF_ABSL, 1,79% com TF_BNS e 5,45% com TF_IDF. A representação "dl09", em relação ao vetor de termos base "snastp" usando dicionário local ("dl02"), apresenta ganhos de 2,84% para TF_ABSL, 1,99% para TF_BNS, 1,09% para TF_GR, 0,75% para TF_IDF, 1,09% para TF_IG, 3,48% para TF_OR e 0,80% para TF_QUI.

3.4.5.2 Redução da dimensionalidade pelas demais métricas de seleção de termos

Neste estudo, a mesma métrica utilizada para compor o método de atribuição supervisionada de termos é utilizada para selecionar os melhores termos. Como observado nos estudos de Debole & Sebastiani (2003) e Lan et al. (2006), na redução da dimensionalidade por essa estratégia, o melhor desempenho para o algoritmo SVMs é obtido sem redução de dimensionalidade, o que pode ser constatado na Tabela 3.10 e Figura 3.14.

A Tabela 3.10 e Figura 3.14 mostram o desempenho do algoritmo SVMs medido por F_1 macro, considerando a dimensão completa, e a seleção dos 4.500 (**rd4500**), 3.500 (**rd3500**) e 2.500 (**rd2500**) melhores termos, usando dicionário global. A dimensão completa - 5.723 termos - corresponde ao número de termos do vetor de termos base "snastp", considerando dicionário global.

Tabela 3.10 - F_1 macro segundo os pesos e vetores de termos descritos, considerando dicionário global.

Vetor de Termos	Nº termos	Peso					
		TF_ABSL	TF_BNS	TF_GR	TF_IG	TF_OR	TF_QUI
1snastp	5.723	0,698	0,671	0,633	0,624	0,661	0,628
2rd4500	4.500	0,697	0,637	0,633	0,624	0,655	0,628
3rd3500	3.500	0,697	0,637	0,633	0,624	0,655	0,628
4rd2500	2.500	0,661	0,628	0,633	0,623	0,626	0,628

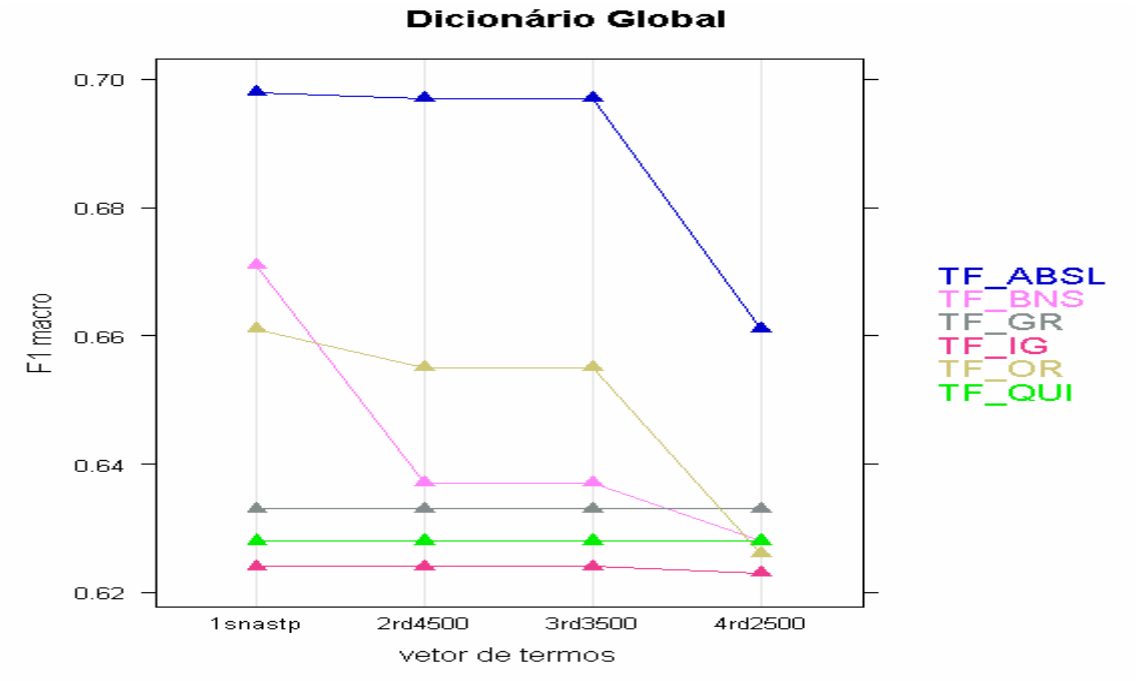


Figura 3.14 - F_1 macro segundo os pesos e vetores de termos descritos, considerando dicionário global.

No estudo conduzido por Yang & Pedersen (1997), eles comentam ter realizado a redução da dimensionalidade apenas até onde nenhum documento fosse eliminado. No caso sob estudo, já na seleção dos 4.500 melhores termos ocorre a eliminação de PLs pela métrica razão de chances (OR).

Apesar desse fato, os resultados da seleção até os 2.500 melhores termos serão apresentados a fim de mostrar o comportamento das métricas de seleção de termos abs_logito e bi-normal separation com as formas de atribuição supervisionada de pesos TF_ABSL e TF_BNS, respectivamente, ainda não utilizadas em trabalhos anteriores.

Embora a seleção de termos tenha sido realizada escolhendo os 4.500, 3.500, 2.500, 1.500, 1.000 e 500 melhores termos, segundo cada uma das métricas estudadas, a medida F_1 macro não foi calculada na seleção dos melhores 1.500 termos

ou menos, devido à grande quantidade de PLs excluídos por não conterem nenhuma das palavras selecionadas.

Durante o processo de redução de dimensionalidade, foi observado o seguinte comportamento para as métricas estudadas: com 4.500 termos, pela métrica razão de chances (OR), em uma das rodadas da validação cruzada, um PL de treinamento foi excluído, e em outra rodada, um PL de teste foi excluído. Com 3.500 termos selecionados o número de PLs excluídos usando essa métrica aumentou, e também a métrica abs_logito (ABSL) começou a excluir PLs. Com 2.500 termos selecionados, aumentou ainda mais o número de PLs excluídos por essas duas métricas. Com 1.500 termos selecionados, a métrica binormal separation (BNS) começou a excluir PLs. Com 1.000 e com 500 termos, todas as métricas excluem PLs. Porém, a quantidade de PLs excluídos pelas métricas razão de chances e abs_logito é muito superior às demais.

No caso de utilização do dicionário local, uma representação vetorial distinta é obtida para cada comissão permanente. Como a representação vetorial com o menor número de termos - 1.982 termos - é a da comissão CDDHCEDP, para selecionar o mesmo número de termos para todas as comissões, foram selecionados inicialmente os 1.000 melhores termos segundo cada métrica. Dado que para essa comissão e com essa redução de dimensionalidade ocorreu a exclusão de PLs segundo todas as métricas estudadas, a medida F_1 macro não foi calculada.

3.4.6 Estudos sobre aumento de peso para algumas palavras

Nesta subseção são apresentados os resultados do estudo sobre aumento de peso para as palavras presentes nas ementas e para as relacionadas às matérias de competência das comissões permanentes. Os estudos foram realizados usando dicionário global e local e se basearam nas idéias apresentadas em Apté et al. (1994) e em Schweighofer et al. (2002).

Apté et al. (1994) estudaram dobrar a freqüência de ocorrência das palavras presentes nos títulos das reportagens do corpus Reuters-22173, visando enfatizá-las, considerando que essas podem fornecer dicas adicionais a respeito das categorias às quais pertencem as reportagens. Usando um sistema baseado em regras de decisão, com dicionário local, a representação dos pesos dos termos pela freqüência, contando em dobro as palavras presentes nos títulos, apresentou uma melhora de 2% em efetividade (*breakeven point*) em relação à representação dos pesos dos termos pela freqüência, sem dar ênfase às palavras presentes nos títulos.

Schweighofer et al. (2002), em aplicação da técnica de análise de agrupamento para organizar documentos relativos à legislação européia, pesquisaram dobrar o valor de TF_IDF para os 204 conceitos mais importantes do domínio estudado, a fim de melhorar a representação vetorial desses documentos. Eles destacaram que esse procedimento gerou grupos de documentos (*clusters*) e rótulos para os grupos (cada grupo é rotulado com as palavras-chave que melhor representam seus documentos) mais focados no pequeno tesauro adotado e menos no texto dos documentos.

No presente trabalho, as seguintes estratégias foram investigadas:

- no estudo sobre aumento de peso para as palavras presentes nas ementas, a freqüência de ocorrência dessas palavras nos respectivos PLs foi multiplicada por dois;
- para as palavras relacionadas às matérias de competência das comissões permanentes, foram estudados aumentos de pesos de 30%, 50%, 70% e 100%. Com dicionário local, foram aumentados os pesos apenas das palavras relacionadas às matérias de competência da comissão representada pelo dicionário; na adoção do dicionário global, como uma única representação vetorial é usada para todas as comissões, o aumento de peso foi efetuado para as palavras relacionadas às matérias de competência de pelo menos uma das comissões permanentes.

Os resultados do estudo considerando dicionário global são apresentados na Tabela 3.11 e Figura 3.15, onde a seguinte nomenclatura é utilizada:

- **snastp**: vetor de termos base;
- **em2**: vetor de termos contando em dobro (2) a freqüência de ocorrência das palavras presentes nas ementas (em);
- **co30**; **co50**; **co70**; e **co100**: vetores de termos aumentando os pesos das palavras relacionadas às matérias de competência das comissões permanentes (co) em 30% (30), 50% (50), 70% (70) e 100% (100), respectivamente.

Tabela 3.11 - F_1 macro segundo os pesos e vetores de termos descritos, considerando dicionário global.

Vetor de	Peso						
termos	TF_ABSL	TF_BNS	TF_GR	TF_IDF	TF_IG	TF_OR	TF_QUI
1snastp	0,698	0,671	0,633	0,642	0,624	0,661	0,628
2em2	0,698	0,683	0,637	0,664	0,628	0,664	0,635
3co30	0,708	0,673	0,631	0,673	0,622	0,666	0,633
4co50	0,706	0,681	0,626	0,678	0,623	0,665	0,635
5co70	0,703	0,682	0,623	0,684	0,623	0,666	0,631
6co100	0,701	0,685	0,621	0,685	0,622	0,665	0,629

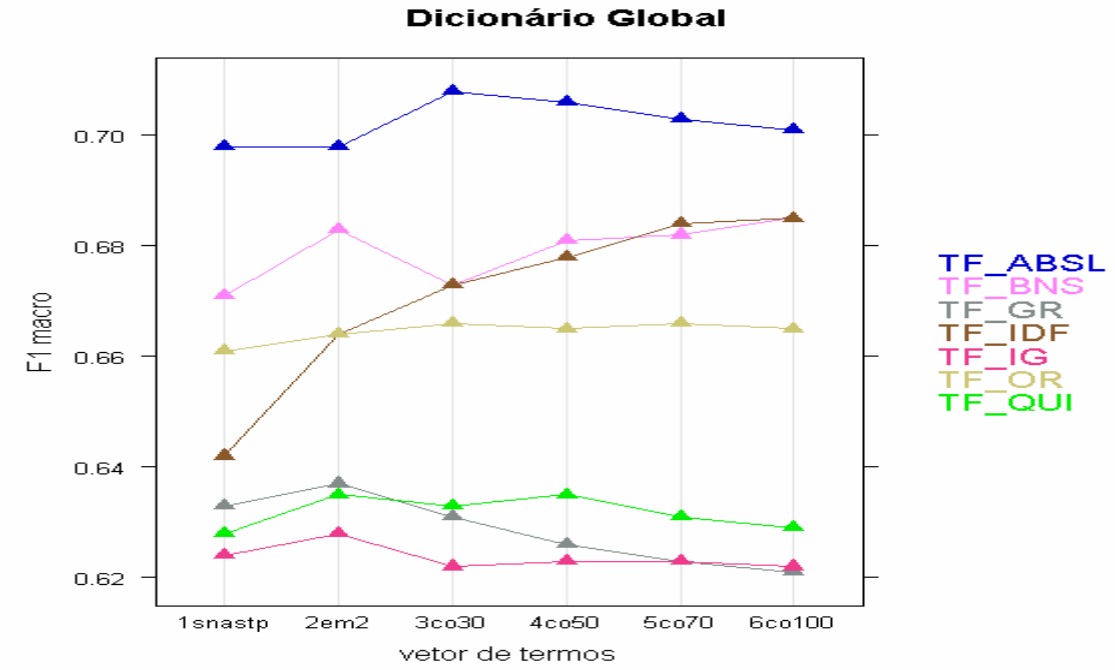


Figura 3.15 - F_1 macro segundo os vetores de termos expostos e pesos calculados de acordo com o descrito na legenda, considerando dicionário global.

Na Tabela 3.12 e Figura 3.16 a seguir, “dl02” é o vetor de termos base “snastp” usando dicionário local. Os demais vetores de termos são os correspondentes aos do dicionário global, só que com dicionário local.

Tabela 3.12 - F_1 macro segundo os pesos e vetores de termos descritos, considerando dicionário local.

Vetor de	Peso						
termos	TF_ABSL	TF_BNS	TF_GR	TF_IDF	TF_IG	TF_OR	TF_QUI
1dl02	0,705	0,702	0,643	0,671	0,643	0,66	0,628
2em2	0,709	0,71	0,646	0,675	0,646	0,662	0,629
4co30	0,712	0,707	0,635	0,686	0,635	0,653	0,613
5co50	0,71	0,704	0,623	0,69	0,623	0,649	0,603
6co70	0,707	0,707	0,616	0,697	0,616	0,646	0,599
7co100	0,697	0,695	0,604	0,702	0,604	0,641	0,593

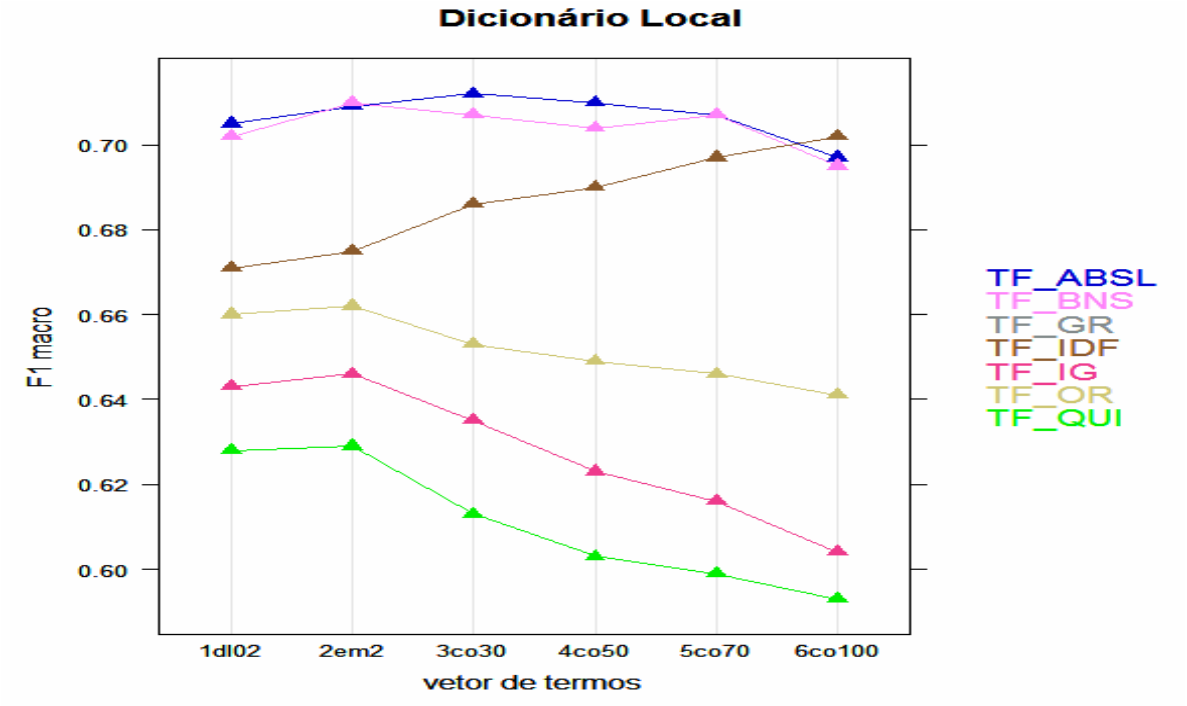


Figura 3.16 - F_1 macro segundo os vetores de termos expostos e pesos calculados de acordo com o descrito na legenda, considerando dicionário local.

A partir dos dados apresentados nas Tabelas 3.11 e 3.12 e Figuras 3.15 e 3.16, obtidos usando dicionário global e local, seguem as constatações:

- **Estudo sobre aumento de peso para as palavras presentes nas ementas**

Comparativamente aos respectivos vetores de termos base "snastp" e "dl02", considerando dicionário global e local, a estratégia de dobrar no PL a freqüência de ocorrência das palavras presentes nas ementas - vetor "em2" - produziu uma melhora no desempenho do algoritmo SVMs, medido por F_1 macro, para todas as formas de atribuição de pesos consideradas, exceto para TF_ABSL, que com dicionário global não apresentou alteração em F_1 macro.

Com essa estratégia, os maiores aumentos em F_1 macro foram atingidos usando dicionário global. Nesse caso, apesar de não ter sido observado ganho para F_1 macro com os pesos calculados por TF_ABSL, que apresentaram o melhor resultado, foram observados ganhos de 1,79% para os pesos calculados por TF_BNS, e de 3,43% para os pesos calculados por TF_IDF, que apresentaram o segundo e terceiro melhores resultados, respectivamente. Com dicionário local, o maior aumento em F_1 macro (1,14%) foi atingido pelos pesos calculados por TF_BNS, que apresentaram o melhor resultado com essa estratégia.

- **Estudo sobre aumento de peso para as palavras relacionadas às matérias de competência das comissões permanentes**

A estratégia de aumentar o peso das palavras relacionadas às matérias de competência das comissões permanentes produziu efeitos distintos para as sete formas de atribuição de pesos estudadas.

Das quatro estratégias de aumento de peso pesquisadas para as palavras relacionadas às matérias de competência das comissões permanentes, a que aumenta em 30% o peso dessas palavras – "co30" - apresentou o maior valor para F_1 macro, tanto com dicionário global como com dicionário local. Em ambos os casos, esse

valor máximo foi atingido pelos pesos calculados por TF_ABSL. Com dicionário global, o ganho foi de 1% em relação ao vetor de termos base “snastp”; com dicionário local, o ganho foi de 1,43% em relação ao vetor de termos base “dl02”.

Para os pesos calculados por TF_BNS, que, com dicionário local, apresentaram valores de F_1 macro bem próximos aos exibidos pelos pesos calculados por TF_ABSL, as melhores performances foram obtidas com aumentos de 30% e 70% para as palavras relacionadas às matérias de competência das comissões permanentes. Com relação ao vetor de termos base “dl02”, esse ganho foi de 0,71%.

Os métodos de atribuição de pesos TF_IG e TF_GR apresentam perdas - com relação aos vetores de temos base “snastp” e “dl02”, usando dicionário global e local, respectivamente – com todas as estratégias de aumento de pesos consideradas para as palavras relacionadas às matérias de competência das comissões permanentes. Com dicionário local, as perdas, em relação ao vetor de termos base “dl02”, também foram observadas para os pesos calculados por TF_OR e TF_QUI.

Dos dados apresentados e considerando a escolha de uma única estratégia de aumento de peso para as palavras relacionadas às matérias de competência das comissões permanentes, a melhor opção parece ser a que aumenta em 30% o peso dessas palavras, por essa ter apresentado o maior valor para F_1 macro – atingido pelos pesos calculados por TF_ABSL - usando tanto dicionário global como local.

3.4.7 Estudo final combinando redução de dimensionalidade com aumento de peso para algumas palavras

Nesta análise final, os melhores resultados obtidos nas análises sobre aumento de pesos para as palavras presentes nas ementas e para as relacionadas às matérias de competência das comissões permanentes serão combinados com o melhor resultado obtido para a redução da dimensionalidade do espaço de termos pela freqüência de documentos. Para efeitos comparativos, também serão apresentadas os dados referentes às representações vetoriais base - "snastp" e "dl02" - usando dicionário global e local, respectivamente.

No caso do dicionário global, ainda serão apresentados os valores de F_1 macro ao representar os PLs apenas pelas ementas. Apesar de alguns PLs terem sido eliminados por não conterem nenhuma das palavras do vetor de termos base "snastp", os valores de F_1 macro foram calculados para mostrar o desempenho de classificação dos PLs nas comissões usando só as ementas. Esse estudo não foi realizado com dicionário local, devido à grande quantidade de PLs excluídos.

Os seguintes vetores de termos serão comparados usando dicionário global:

- **ementas**: utiliza apenas os termos presentes nas ementas;
- **snastp**: vetor de termos base;
- **df09**: utiliza os termos que ocorreram em pelo menos nove PLs de treinamento (df09);
- **em2**: dobra (2) a freqüência de ocorrência das palavras presentes nas ementas (em);

- **em2f9**: utiliza apenas as palavras que ocorreram em nove ou mais PLs de treinamento (f9), e conta em dobro a freqüência de ocorrência das palavras presentes nas ementas (em2);

- **co30**: aumenta em 30% (30) o peso das palavras relacionadas às matérias de competência das comissões permanentes (co);

- **c30f9**: utiliza apenas as palavras que ocorreram em nove ou mais PLs de treinamento (f9), e aumenta os pesos das palavras relacionadas às matérias de competência das comissões permanentes em 30% (c30);

- **c30e2**: dobra a freqüência de ocorrência das palavras presentes nas ementas (e2) e aumenta em 30% os pesos das palavras relacionadas às matérias de competência das comissões permanentes (c30);

- **c30e2f9**: utiliza apenas as palavras que ocorreram em nove ou mais PLs de treinamento (f9), dobra a freqüência de ocorrência das palavras presentes nas ementas (e2), e aumenta em 30% os pesos das palavras relacionadas às matérias de competência das comissões permanentes (c30).

Observação: para melhorar a apresentação gráfica do nome do vetor, nas representações vetoriais "c30f9", "c30e2" e "c30e2f9", as seguintes alterações foram realizadas: a) a nomenclatura utilizada para o aumento de 30% para os pesos das palavras relacionadas às matérias de competência das comissões permanentes foi alterada de "co30" para "c30"; b) a nomenclatura utilizada para a contagem em dobro da freqüência de ocorrência das palavras presentes nas ementas foi alterada de "em2" para "e2"; e c) a nomenclatura utilizada para os termos que ocorreram em pelo menos nove PLs de treinamento foi alterada de "df09/dl09" para "f9".

Os resultados são apresentados na Tabela 3.13 e Figura 3.17.

Tabela 3.13 - F_1 macro segundo os pesos e vetores de termos descritos, considerando dicionário global.

Vetor de termos	Peso						
	TF_ABSL	TF_BNS	TF_GR	TF_IDF	TF_IG	TF_OR	TF_QUI
0ementas[27]	0,59	0,568	0,542	0,568	0,544	0,57	0,538
1snastp	0,698	0,671	0,633	0,642	0,624	0,661	0,628
2df09	0,71	0,683	0,629	0,677	0,626	0,675	0,627
3em2	0,698	0,683	0,637	0,664	0,628	0,664	0,635
4em2f9	0,713	0,681	0,636	0,686	0,63	0,688	0,634
5co30	0,708	0,673	0,631	0,673	0,622	0,666	0,633
6c30f9	0,714	0,685	0,628	0,686	0,625	0,677	0,629
7c30e2	0,714	0,688	0,634	0,68	0,623	0,673	0,639
8c30e2f9	0,719	0,685	0,631	0,698	0,628	0,689	0,637

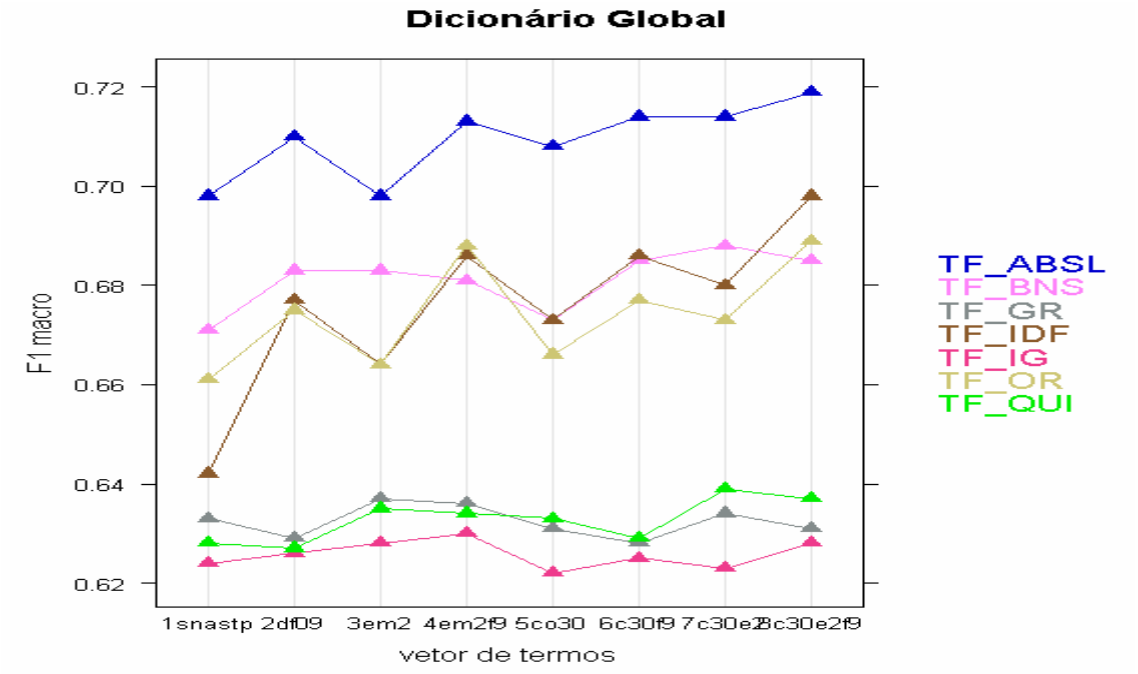


Figura 3.17 - F_1 macro segundo os pesos e vetores de termos descritos considerando dicionário global.

Na Tabela 3.14 e Figura 3.18 a seguir, os vetores de termos apresentados são os correspondentes ao do dicionário global, apresentados na Tabela 3.13 e Figura

[27] Por apresentarem desempenhos bem inferiores às demais representações vetoriais, os dados referentes às ementas não serão apresentados na Figura 3.17, para clareza da apresentação.

3.17, só que usando dicionário local. Com dicionário local, nomenclaturas diferentes são utilizadas para o vetor de termos base “snastp” e para o vetor de termos usando apenas as palavras que ocorreram em nove ou mais PLs de treinamento, que são denominados, respectivamente, de “dl02” e “dl09”.

Tabela 3.14 - F_1 macro segundo os pesos e vetores de termos descritos, considerando dicionário local

vetor de	Peso						
Termos	TF_ABSL	TF_BNS	TF_GR	TF_IDF	TF_IG	TF_OR	TF_QUI
1dl02	0,705	0,702	0,643	0,671	0,643	0,66	0,628
2dl09	0,725	0,716	0,65	0,676	0,65	0,683	0,633
3em2	0,709	0,71	0,646	0,675	0,646	0,662	0,629
4em2f9	0,726	0,724	0,656	0,685	0,656	0,688	0,63
5co30	0,712	0,707	0,635	0,686	0,635	0,653	0,613
6c30f9	0,722	0,721	0,642	0,684	0,642	0,676	0,617
7c30e2	0,714	0,704	0,631	0,704	0,631	0,655	0,611
8c30e2f9	0,722	0,722	0,64	0,699	0,64	0,679	0,616

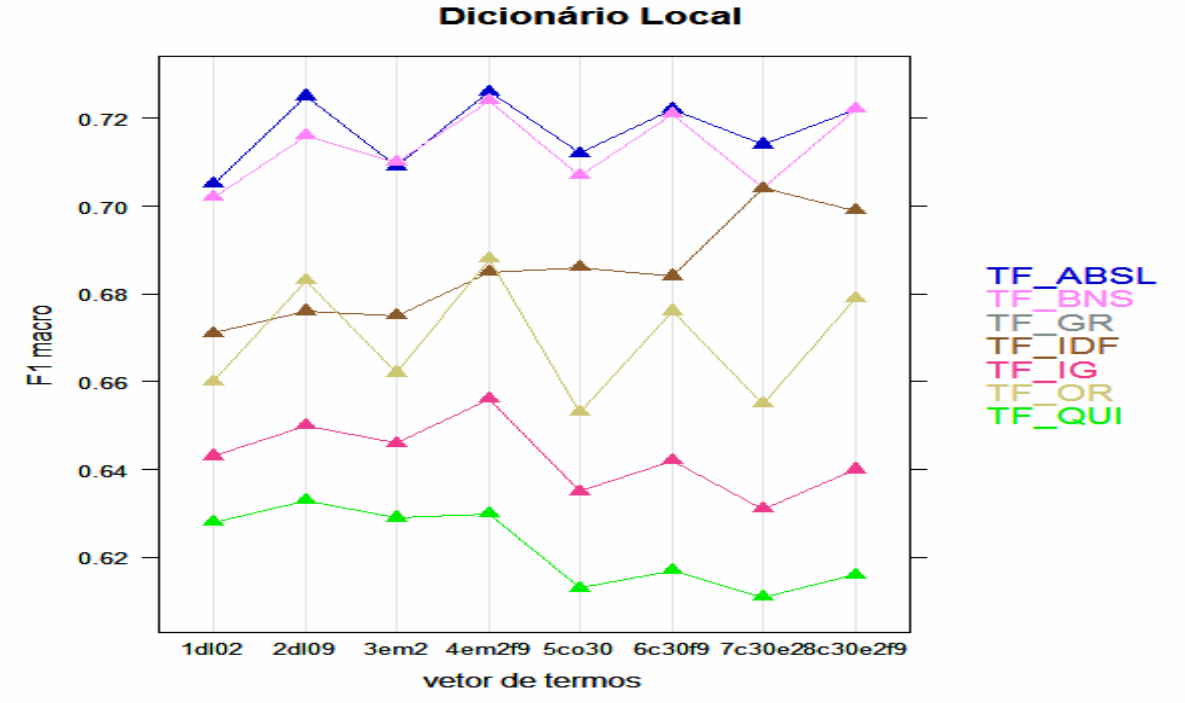


Figura 3.18 - F_1 macro segundo os pesos e vetores de termos descritos, considerando dicionário local.

No caso de utilização do dicionário global, como pode ser verificado na Tabela 3.13 e Figura 3.17, de todas as representações comparadas neste trabalho, o

maior valor para F_1 macro é atingido pelos pesos calculados por TF_ABSL usando a representação "c30e2f9". Essa representação vetorial, a partir da representação "snastp", considera apenas as palavras presentes em nove ou mais PLs, aumenta em 30% o peso das palavras relacionadas às matérias de competência das comissões permanentes, e dobra a freqüência de ocorrências das palavras presentes nas ementas. Em relação à representação vetorial base "snastp", com pesos também calculados por TF_ABSL, a representação vetorial "c30e2f9" exibe desempenho 3% superior em termos de F_1 macro.

Considerando dicionário local, o maior valor para F_1 macro é atingido usando a representação "em2f9" com pesos também calculados por TF_ABSL. A partir da representação vetorial "dl02", a representação "em2f9" considera apenas as palavras presentes em nove ou mais PLs e conta em dobro a freqüência de ocorrência das palavras presentes nas ementas. Essa representação vetorial exibe desempenho em termos de F_1 macro 2,98% superior, comparada à representação vetorial base usando dicionário local ("dl02") com pesos também calculados por TF_ABSL.

De todas as representações vetoriais estudadas, considerando tanto dicionário global como local, a "em2f9" usando dicionário local é a que apresenta o maior valor para F_1 macro. Em relação à representação vetorial base "snastp" usando dicionário global, essa representação vetorial apresenta desempenho 4% superior, em termos de F_1 macro, considerando os pesos calculados por TF_ABSL, que apresentaram o melhor resultado.

Por fim, os resultados obtidos para as representações vetoriais que exibiram as melhores performances em termos de F_1 macro serão apresentados, por comissão permanente, nas Figuras 3.19 a 3.26 a seguir. Para efeitos comparativos, também serão

apresentadas as representações vetoriais base usando dicionário global (“snastp”) e local (“dl02”). Os seguintes vetores de termos serão comparados:

1. **snastp**: vetor de termos base usando dicionário global;
2. **dgc30e2f9**: utiliza dicionário global (dg), apenas palavras que ocorreram em nove ou mais PLs de treinamento (f9), dobra a freqüência de ocorrência das palavras presentes nas ementas (e2), e aumenta em 30% os pesos das palavras relacionadas às matérias de competência das comissões (c30);
3. **dl02**: vetor de termos base “snastp” usando dicionário local;
4. **dlem2f9**: utiliza dicionário local (dl), palavras que ocorreram em nove ou mais PLs de treinamento (f9), e dobra a freqüência de ocorrência das palavras presentes nas ementas (em2).

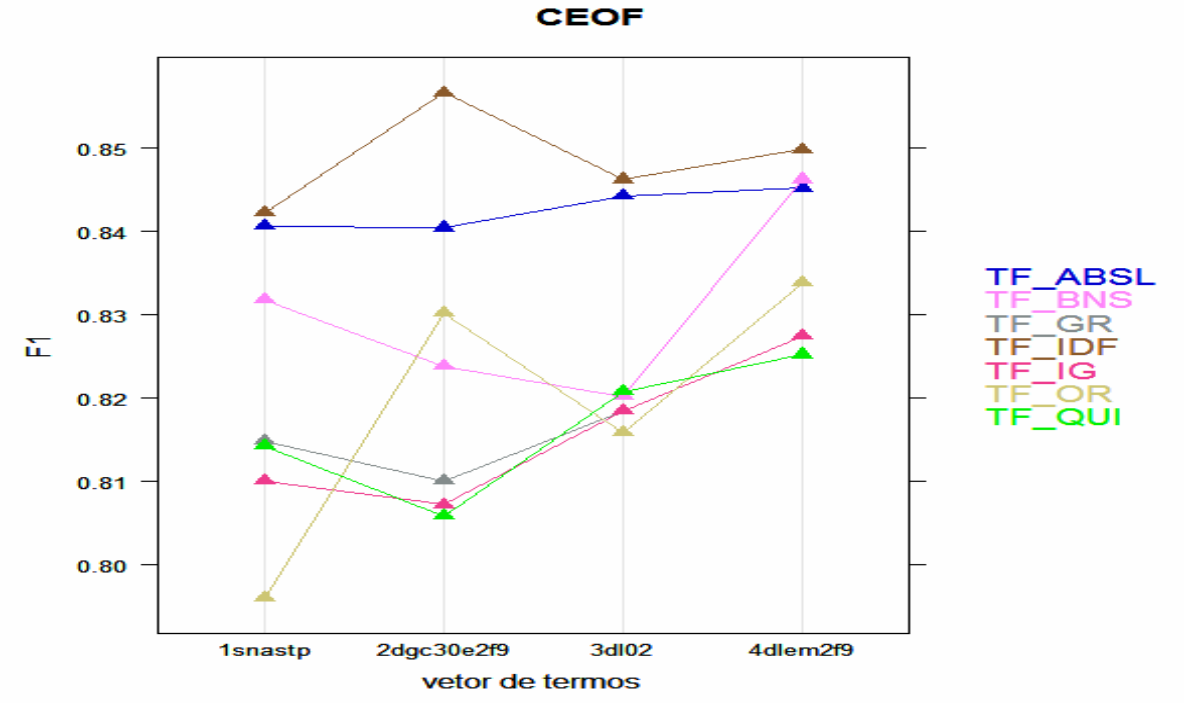


Figura 3.19 - F_1 segundo os vetores de termos expostos, e pesos calculados de acordo com o descrito na legenda – CEOF.

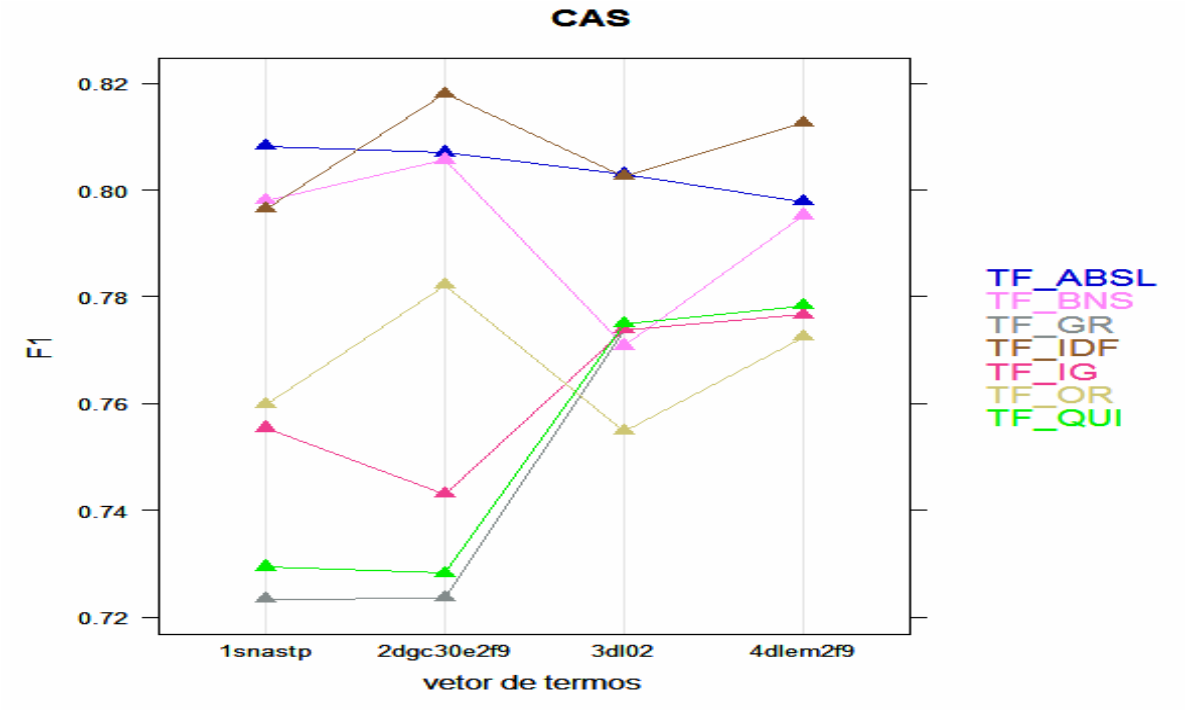


Figura 3.20 - F_1 segundo os vetores de termos expostos, e pesos calculados de acordo com o descrito na legenda – CAS.

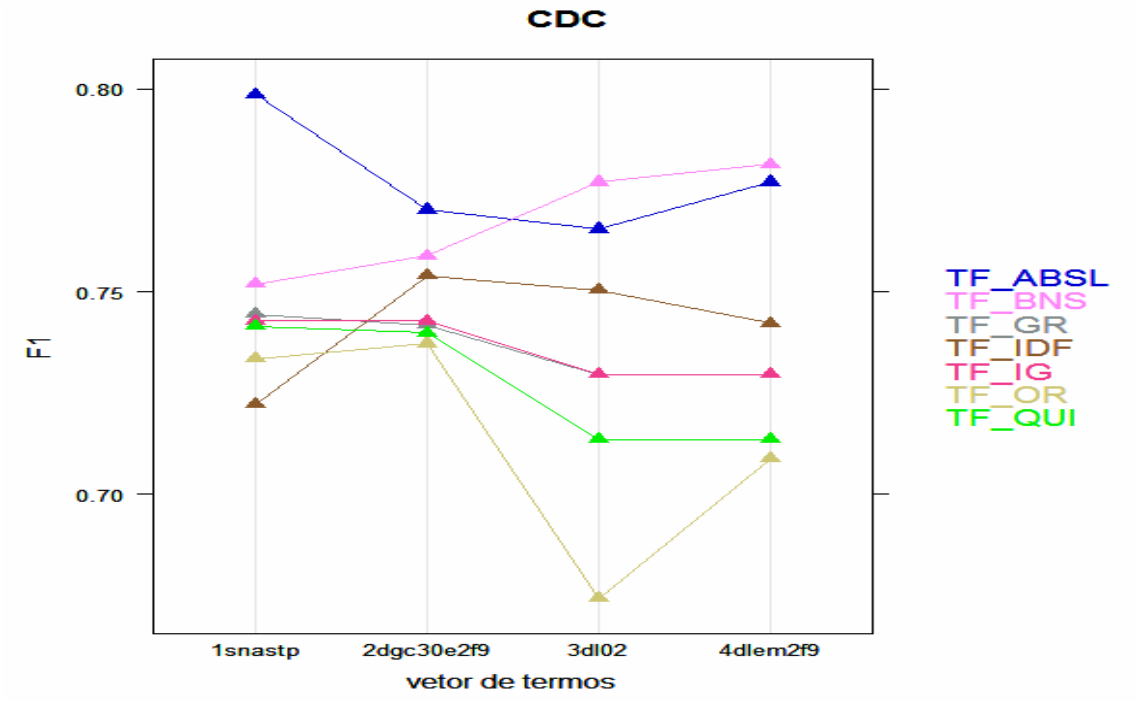


Figura 3.21 - F_1 segundo os vetores de termos expostos, e pesos calculados de acordo com o descrito na legenda – CDC.

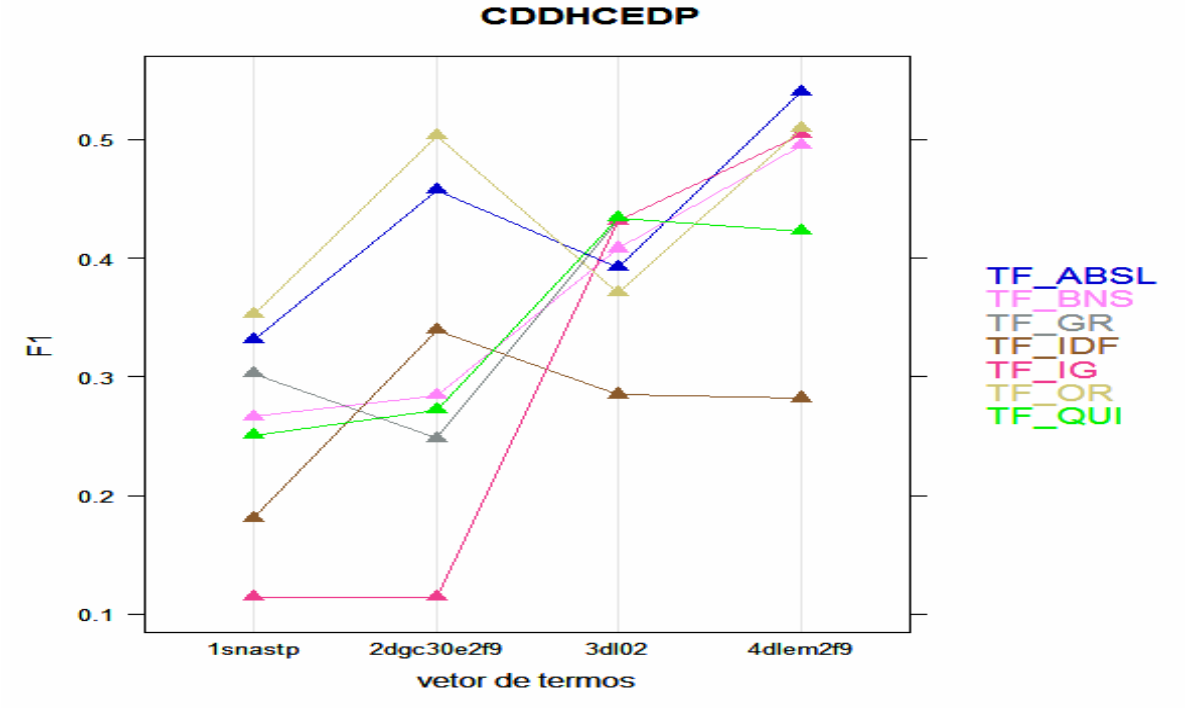


Figura 3.22 - F_1 segundo os vetores de termos expostos e pesos calculados de acordo com o descrito na legenda – CDDHCEDP.

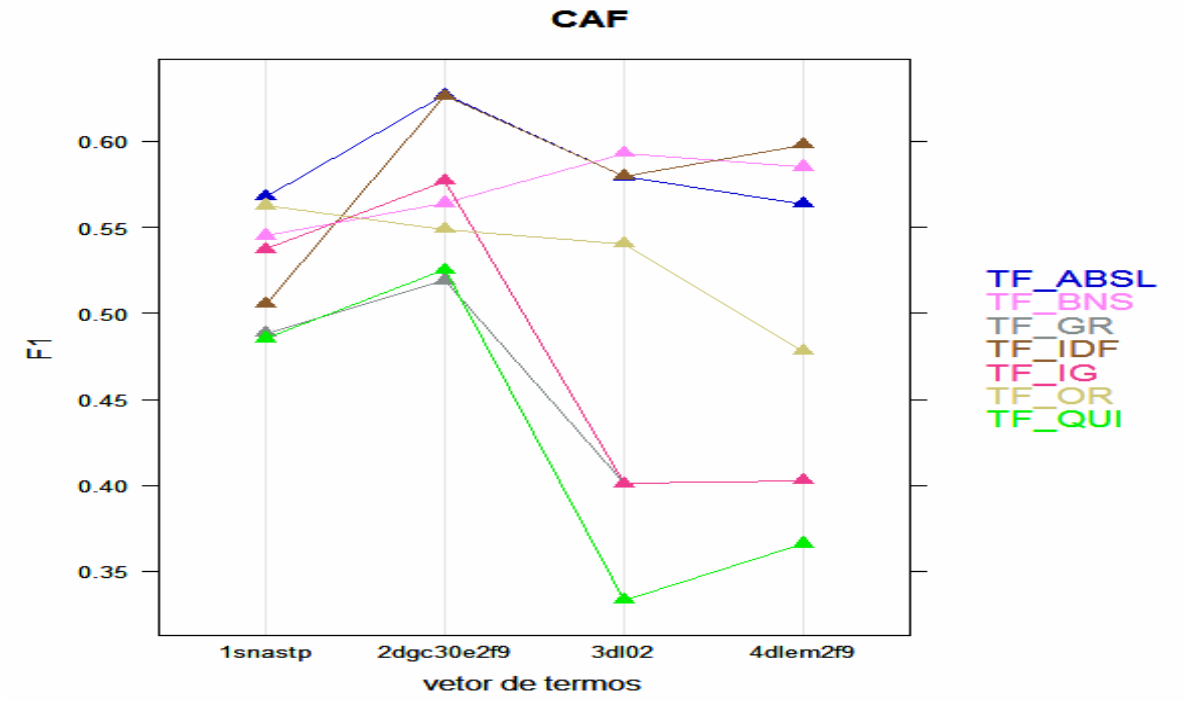


Figura 3.23 - F_1 segundo os vetores de termos expostos e pesos calculados de acordo com o descrito na legenda – CAF.

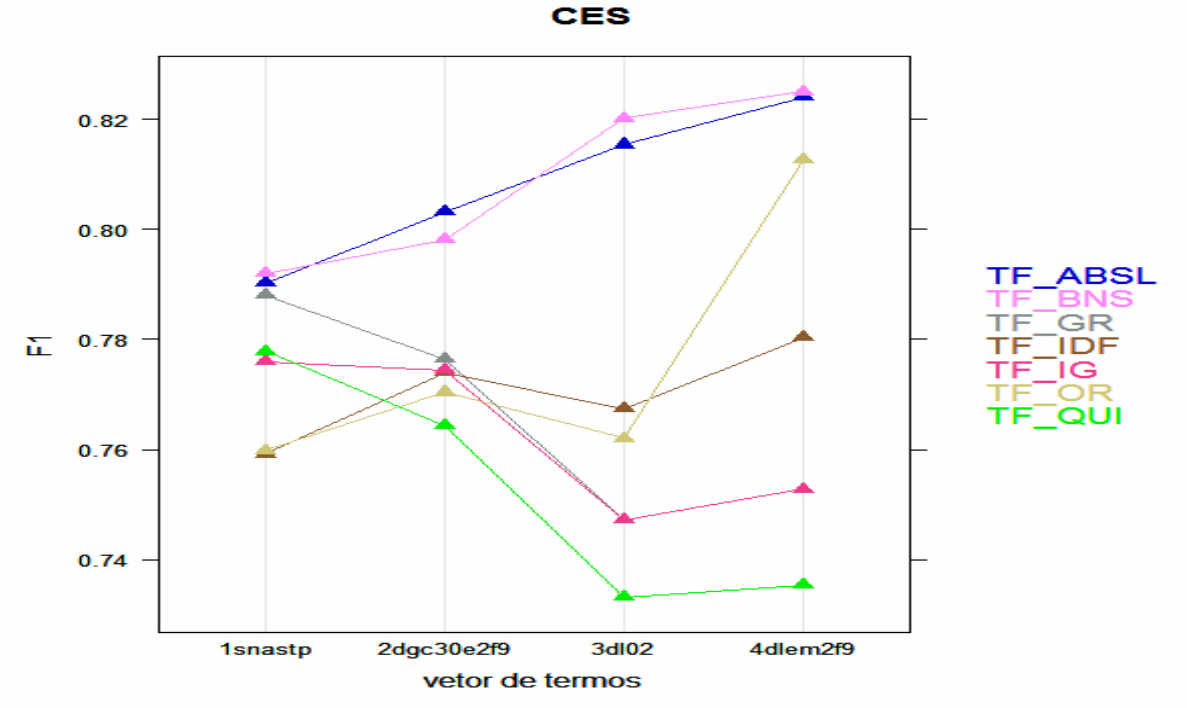


Figura 3.24 - F_1 macro segundo os vetores de termos expostos, e pesos calculados de acordo com o descrito na legenda – CES.

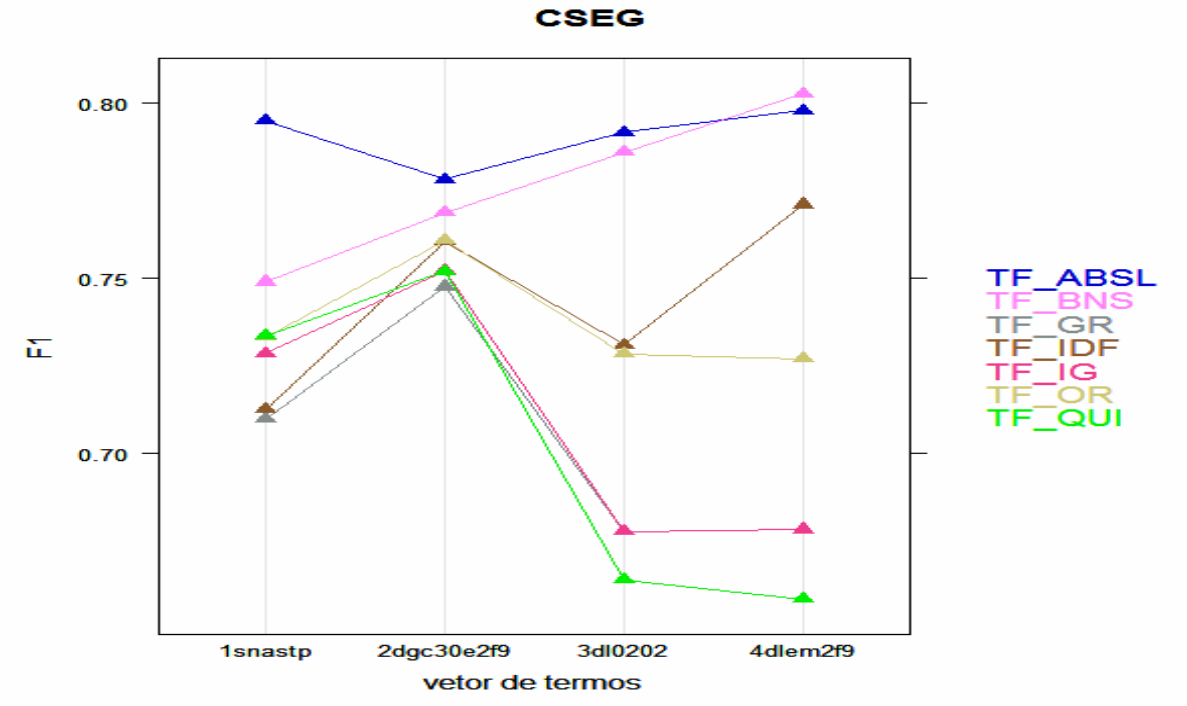


Figura 3.25 - F_1 segundo os vetores de termos expostos, e pesos calculados de acordo com o descrito na legenda – CSEG.

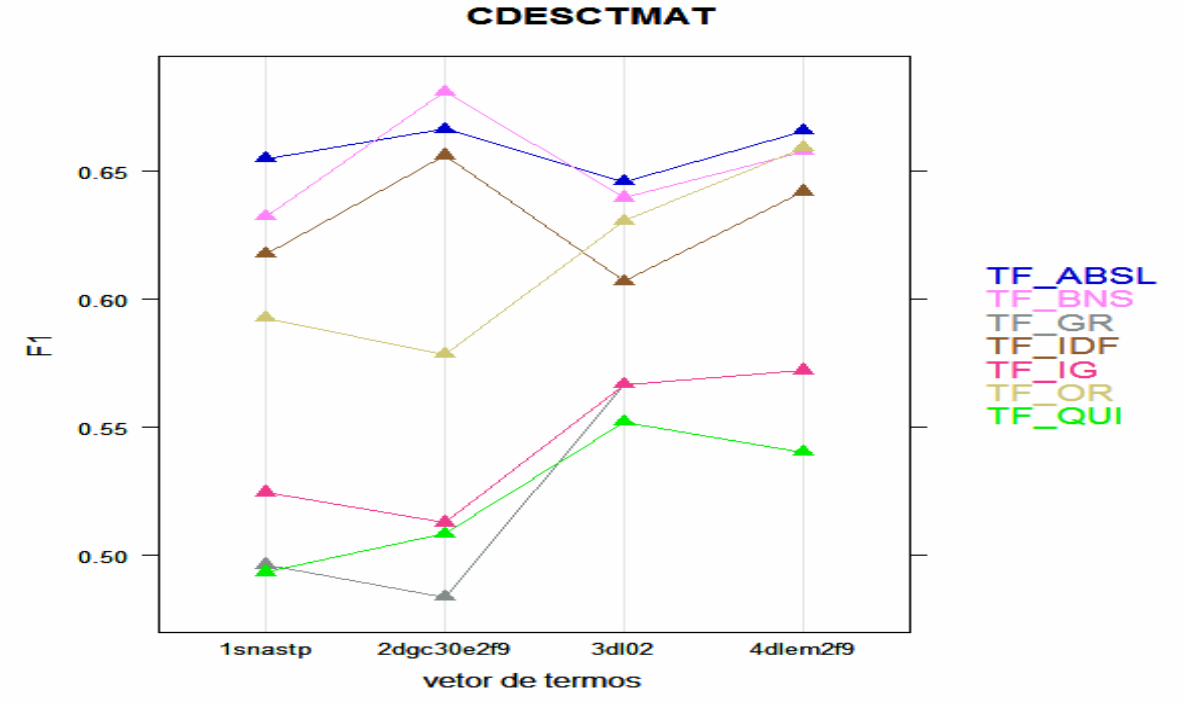


Figura 3.26 - F_1 segundo os vetores de termos expostos e pesos calculados de acordo com o descrito na legenda – CDESCTMAT.

Das Figuras 3.19 a 3.26 apresentadas, pode-se verificar que o maior valor para F_1 não é atingido nas comissões permanentes, nem pela mesma representação vetorial, nem pelo mesmo método de atribuição de pesos. Em cada comissão permanente, o maior valor para F_1 é atingido conforme representação vetorial e pesos listados a seguir:

- CEOF e CAS: representação “dgc30e2f9”, com pesos calculados por TF_IDF;
- CDC: representação “snastp”, com pesos calculados por TF_ABSL;
- CDDHCEDP: representação “dlem2f9”, com pesos calculados por TF_ABSL;
- CAF: representação “dgc30e2f9”, com pesos calculados por TF_ABSL/TF_IDF;
- CES: representação “dlem2f9”, com pesos calculados por TF_BNS;
- CSEG: representação “dlem2f9”, com pesos calculados por TF_BNS;

- CDESCTMAT: representação “c30e2f9”, com pesos calculados por TF_BNS.

Como se pode constatar, a maior parte das comissões permanentes atinge valor máximo para F_1 com as representações vetoriais “dgc30e2f9” e “dlem2f9”. Das análises gráficas, ainda se pode verificar que os pesos calculados por TF_BNS e TF_ABSL estão entre os melhores resultados apresentados para todas as comissões permanentes, ao contrário dos pesos calculados por TF_IDF, que, para algumas comissões, apresenta desempenho bem inferior aos melhores resultados atingidos. Assim, para escolher uma única representação vetorial com a mesma forma de atribuição de pesos para todas as comissões permanentes, serão analisadas essas quatro combinações - “dgc30e2f9” e “dlem2f9” com pesos calculados por TF_BNS e TF_ABSL - juntamente com a combinação que forneceu o melhor resultado para cada comissão.

As Tabelas 3.15 a 3.22 a seguir apresentam, por comissão permanente, representação vetorial e peso atribuído, as seguintes quantidades[28]: 1) VP – verdadeiros positivos - número de PLs da comissão que foram corretamente classificados pelas SVMs na comissão; 2) FN – falsos negativos - número de PLs da comissão que foram incorretamente classificados pelas SVMs como não sendo da comissão; 3) VN – verdadeiros negativos - número de PLs de outras comissões que foram corretamente classificados pelas SVMs como não sendo da comissão; 4) FP – falsos positivos - número de PLs que são de outras comissões, mas que foram incorretamente classificados pelas SVMs como sendo da comissão; 5) acurácia; 6) precisão; 7) revocação e 8) F_1.

[28] Média dos cinco resultados obtidos na validação cruzada, com base em um númeroa médio de 202,8 PLs de teste.

Tabela 3.15 – Medidas de desempenho para CEOF, segundo os pesos e vetores de termos descritos.

CEOF									
peso	vetor de termos	VP	FN	VN	FP	acurácia	precisão	revocação	F_1
TF_IDF	2dgc30e2f9	101,800	16,200	66,800	18,000	0,831	0,851	0,864	0,857
TF_ABSL	2dgc30e2f9	99,600	18,400	65,400	19,400	0,814	0,837	0,845	0,840
TF_BNS	2dgc30e2f9	98,000	20,000	62,800	22,000	0,793	0,818	0,832	0,824
TF_ABSL	5dlem2f9	99,400	18,600	67,000	17,800	0,821	0,849	0,843	0,845
TF_BNS	5dlem2f9	99,400	18,600	67,200	17,600	0,822	0,850	0,843	0,846

Tabela 3.16 – Medidas de desempenho para CAS, segundo os pesos e vetores de termos descritos.

CAS									
peso	vetor de termos	VP	FN	VN	FP	acurácia	precisão	revocação	F_1
TF_IDF	2dgc30e2f9	71,000	19,800	100,400	11,600	0,845	0,859	0,783	0,818
TF_ABSL	2dgc30e2f9	68,400	22,400	101,600	10,400	0,838	0,868	0,756	0,807
TF_BNS	2dgc30e2f9	70,400	20,400	98,400	13,600	0,832	0,838	0,778	0,806
TF_ABSL	5dlem2f9	69,400	21,400	98,200	13,800	0,827	0,834	0,766	0,798
TF_BNS	5dlem2f9	68,600	22,200	98,800	13,200	0,825	0,839	0,757	0,795

Tabela 3.17 – Medidas de desempenho para CDC, segundo os pesos e vetores de termos descritos.

CDC									
peso	vetor de termos	VP	FN	VN	FP	acurácia	precisão	revocação	F_1
TF_ABSL	1snastp	14,400	6,000	181,200	1,200	0,964	0,923	0,704	0,799
TF_ABSL	2dgc30e2f9	14,000	6,400	180,400	2,000	0,958	0,877	0,687	0,770
TF_BNS	2dgc30e2f9	13,400	7,000	180,800	1,600	0,957	0,894	0,664	0,759
TF_ABSL	5dlem2f9	15,200	5,200	179,000	3,400	0,957	0,820	0,743	0,777
TF_BNS	5dlem2f9	15,200	5,200	179,200	3,200	0,958	0,830	0,743	0,781

Tabela 3.18 – Medidas de desempenho para CDDHCEDP, segundo os pesos e vetores de termos descritos.

CDDHCEDP									
peso	Vetor de termos	VP	FN	VN	FP	acurácia	precisão	revocação	F_1
TF_ABSL	5dlem2f9	4,400	6,800	191,000	0,600	0,963	0,893	0,391	0,540
TF_ABSL	2dgc30e2f9	3,800	7,400	191,000	0,600	0,960	0,909	0,334	0,458
TF_BNS	2dgc30e2f9	2,200	9,000	191,200	0,400	0,954	0,860	0,194	0,284
TF_ABSL	5dlem2f9	4,400	6,800	191,000	0,600	0,963	0,893	0,391	0,540
TF_BNS	5dlem2f9	4,000	7,200	191,000	0,600	0,961	0,893	0,354	0,495

Tabela 3.19 – Medidas de desempenho para CAF, segundo os pesos e vetores de termos descritos.

CAF									
peso	vetor de Termos	VP	FN	VN	FP	acurácia	precisão	revocação	F_1
TF_ABSL	2dgc30e2f9	11,800	10,800	177,200	3,000	0,932	0,793	0,529	0,627
TF_ABSL	2dgc30e2f9	11,800	10,800	177,200	3,000	0,932	0,793	0,529	0,627
TF_BNS	2dgc30e2f9	10,000	12,600	177,400	2,800	0,924	0,787	0,447	0,564
TF_ABSL	5dlem2f9	10,800	11,800	175,400	4,800	0,918	0,698	0,481	0,564
TF_BNS	5dlem2f9	11,200	11,400	175,800	4,400	0,922	0,730	0,500	0,585

Tabela 3.20 – Medidas de desempenho para CES, segundo os pesos e vetores de termos descritos.

CES									
peso	vetor de Termos	VP	FN	VN	FP	acurácia	precisão	revocação	F_1
TF_BNS	5dlem2f9	33,600	7,400	155,000	6,800	0,930	0,832	0,820	0,825
TF_ABSL	2dgc30e2f9	30,200	10,800	157,800	4,000	0,927	0,887	0,739	0,803
TF_BNS	2dgc30e2f9	29,400	11,600	158,600	3,200	0,927	0,907	0,720	0,798
TF_ABSL	5dlem2f9	33,400	7,600	155,200	6,600	0,930	0,836	0,816	0,824
TF_BNS	5dlem2f9	33,600	7,400	155,000	6,800	0,930	0,832	0,820	0,825

Tabela 3.21 – Medidas de desempenho para CSEG, segundo os pesos e vetores de termos descritos.

CSEG									
peso	vetor de Termos	VP	FN	VN	FP	acurácia	precisão	revocação	F_1
TF_BNS	5dlem2f9	23,400	6,200	167,800	5,400	0,943	0,815	0,792	0,803
TF_ABSL	2dgc30e2f9	21,000	8,600	169,800	3,400	0,941	0,864	0,710	0,778
TF_BNS	2dgc30e2f9	20,200	9,400	170,400	2,800	0,939	0,881	0,683	0,769
TF_ABSL	5dlem2f9	23,000	6,600	168,000	5,200	0,942	0,819	0,780	0,798
TF_BNS	5dlem2f9	23,400	6,200	167,800	5,400	0,943	0,815	0,792	0,803

Tabela 3.22 – Medidas de desempenho para CDESCTMAT, segundo os pesos e vetores de termos descritos.

CDESCTMAT									
peso	vetor de termos	VP	FN	VN	FP	acurácia	precisão	revocação	F_1
TF_BNS	2dgc30e2f9	20,800	17,000	162,600	2,400	0,904	0,902	0,552	0,681
TF_ABSL	2dgc30e2f9	20,800	17,000	161,000	4,000	0,896	0,840	0,553	0,666
TF_BNS	2dgc30e2f9	20,800	17,000	162,600	2,400	0,904	0,902	0,552	0,681
TF_ABSL	5dlem2f9	21,200	16,600	160,000	5,000	0,893	0,812	0,564	0,666
TF_BNS	5dlem2f9	20,800	17,000	160,200	4,800	0,893	0,816	0,551	0,658

Considerando como satisfatórios neste trabalho valores acima de 0,70[29] para acurácia, precisão, revocação e F_1, as seguintes constatações podem ser realizadas a partir das Tabelas 3.15 a 3.22 apresentadas:

- as comissões CEOF, CAS, CDC, CES e CSEG apresentam bons resultados gerais com relação a todas as medidas de desempenho calculadas;
- os valores mais baixos para F_1 são exibidos para a comissão CDDHCEDP, que também apresenta valores muito baixos para revocação. Considerando a representação vetorial "dlem2f9" em combinação com os pesos calculados por TF_ABSL, que apresenta o maior valor para F_1, a revocação é de apenas 0,391, ou

[29] Neste trabalho foram considerados satisfatórios valores acima de 0,70 para as medidas de desempenho usadas para avaliar a classificação automática, posto que mesmo entre humanos ocorrem divergências. De acordo com Cleverdon (1991), nas decisões de dois especialistas sobre a relevância de um conjunto de documentos em relação a determinado assunto, a coincidência pode ser de apenas 60%.

seja, apenas 39% dos PLs que são da CDDHCEDP são corretamente classificados como sendo dessa comissão. Por outro lado, para essa mesma configuração, praticamente todos os PLs que são de outras comissões são corretamente identificados como não sendo da CDDHCEDP. Além disso, essa comissão também apresenta valores muito bons – na faixa de 0,90 - para acurácia e precisão;

- para a comissão CAF, apesar dos valores para revocação e F_1 serem um pouco melhores que para a comissão CDDHCEDP, eles ainda são baixos. A revocação apresentada - em torno de 0,50 - significa que dos PLs dessa comissão somente 50% são corretamente classificados pelo classificador nessa comissão. Ao contrário, a acurácia apresentada para a comissão é muito boa – na faixa de 0,93 - e também é boa a precisão – próxima de 0,80. Ademais, a taxa de acerto no reconhecimento dos PLs que não são da comissão também é muito boa;
- a comissão CDESCTMAT, apesar de exibir valor para F_1 próximo de 0,70 e valores muito bons para acurácia e precisão - em torno de 0,90 -, apresenta valor baixo – na faixa de 0,50 - para revocação;

Na busca da possível causa para o fraco desempenho exibido para as comissões CDDHCEDP, CAF e CDESCTMAT, principalmente em termos de revocação, o primeiro passo foi analisar a quantidade de exemplos disponíveis para essas comissões no estudo. Dessa análise, a seguinte constatação foi feita: a comissão CDDHCEDP, que apresentou os piores resultados, é a que possui o menor número de PLs no corpus (56¨PLs); a comissão CAF, que apresentou o segundo pior resultado geral, tem 113 PLs no corpus - mais PLs que a comissão CDC, que apresentou resultados gerais bem melhores; a comissão CDESCTMAT, que apresentou o terceiro pior resultado geral, possui 189 PLs no corpus - mais PLs que a comissão CSEG e que a comissão CDC, que apresentaram resultados gerais bem superiores.

Como mostrado no parágrafo anterior, a justificativa para as baixas performances observadas não está apenas na escassez de exemplos disponíveis para a comissão, pois comissões com menor número de exemplos apresentaram performances superiores para as medidas estudadas.

A explicação mais plausível parece ser que as comissões estudadas apresentam graus de dificuldade diferentes para classificação, sendo que a quantidade de PLs disponíveis para estudo nessas três comissões mostrou-se insuficiente para bem caracterizar um PL a elas pertencente. Assim sendo, o mais recomendável é obter mais dados, e reconstruir os classificadores para essas três comissões. Esse processo deve ser repetido até que um resultado considerado satisfatório seja obtido.

Julgando que, em termos práticos, os resultados considerados satisfatórios são os apresentados pelas comissões CEOF, CAS, CDC, CES e CSEG, a escolha de uma única representação vetorial e forma de atribuição de pesos será feita com base nos resultados apresentados para essas comissões.

Analisando, então, os dados mostrados nas Tabelas 3.15, 3.16, 3.17, 3.20 e 3.21, é possível verificar que as melhores combinações "representação vetorial/peso atribuído" são dadas por "dlem2f9"/TF_BNS e "dlem2f9"/TF_ABSL – que exibem resultados praticamente idênticos -, pois essas são as únicas combinações que apresentam valores acima de 0,70 para acurácia, precisão, revocação e F_1. Deve-se observar que essas mesmas formas de atribuição de pesos e representação vetorial devem ser utilizadas ao refazer o treinamento para as comissões "CDDHCEDP", "CAF" e "CDESCTMAT". Como pode ser verificado nas Figuras 3.22, 3.23 e 3.26, as combinações consideradas estão entre os melhores resultados apresentados para essas comissões também.

A seguir é mostrada na Tabela 3.23, para as comissões CEOF, CAS, CDC, CES e CSEG, o valor mínimo, a média, o valor máximo e o desvio padrão obtidos para as medidas acurácia, precisão, revocação e F1 nas cinco repetições da validação cruzada, usando as combinações "dlem2f9"/TF_BNS e "dlem2f9"/TF_ABSL. Nesta tabela é possível verificar que em nenhuma das repetições da validação cruzada e nenhuma das medidas consideradas foi observado valor abaixo de 0,70. Deve-se ressaltar que o mesmo não ocorre com as representações vetoriais base "snastp", com dicionário global, e "dl02", com dicionário local.

Tabela 3.23 – Mínimo, média, máximo e desvio padrão considerando as cinco repetições da validação cruzada, segundo os pesos, medidas de performance e comissões permanentes analisadas (continua).

Peso	Medida de desempenho	Estatística	Comissão				
			CEOF	CAS	CDC	CES	CSEG
TF_BNS	acurácia	min	0,798	0,807	0,946	0,91	0,93
		média	0,822	0,825	0,958	0,93	0,943
		max	0,857	0,847	0,97	0,956	0,956
		dp	0,023	0,017	0,01	0,019	0,012
TF_ABSL	acurácia	min	0,798	0,807	0,946	0,91	0,926
		média	0,821	0,827	0,957	0,93	0,942
		max	0,842	0,847	0,97	0,956	0,96
		dp	0,016	0,015	0,009	0,018	0,0153
TF_BNS	precisão	min	0,836	0,802	0,722	0,762	0,75
		média	0,850	0,839	0,830	0,832	0,815
		max	0,869	0,892	1	0,867	0,885
		dp	0,015	0,034	0,113	0,042	0,05
TF_ABSL	precisão	min	0,828	0,805	0,722	0,762	0,742
		média	0,849	0,834	0,820	0,836	0,819
		max	0,868	0,884	1	0,886	0,885
		dp	0,0173	0,031	0,11	0,049	0,055
TF_BNS	revocação	min	0,8	0,718	0,722	0,756	0,75
		média	0,843	0,757	0,743	0,820	0,792
		max	0,892	0,793	0,773	0,929	0,852
		dp	0,035	0,028	0,022	0,068	0,039

Tabela 3.24 – Mínimo, média, máximo e desvio padrão considerando as cinco repetições da validação cruzada, segundo os pesos, medidas de performance e comissões permanentes analisadas (conclusão).

Peso	Medida de desempenho	Estatística	Comissão				
			CEOF	CAS	CDC	CES	CSEG
TF_ABSL	revocação	min	0,792	0,738	0,722	0,756	0,719
		média	0,843	0,766	0,743	0,816	0,780
		max	0,867	0,805	0,773	0,929	0,852
		dp	0,032	0,026	0,023	0,072	0,049
TF_BNS	F1	min	0,83	0,783	0,722	0,78	0,774
		média	0,846	0,795	0,781	0,825	0,803
		max	0,865	0,807	0,864	0,897	0,836
		dp	0,017	0,011	0,054	0,045	0,03
TF_ABSL	F1	min	0,828	0,785	0,722	0,78	0,754
		média	0,846	0,798	0,777	0,824	0,798
		max	0,856	0,81	0,864	0,897	0,852
		dp	0,011	0,01	0,053	0,044	0,044

Para melhor compreensão dos resultados obtidos, serão mostrados a seguir, no formato de matriz de confusão (Tabela 2.8), os dados comparativos entre a classificação manual (especialista) e a classificação automática (classificador), para as comissões CEOF, CAS, CDC, CES e CSEG. Os dados são referentes à representação vetorial "dlem2f9", com pesos calculados por TF_BNS, exibidos na última linha das Tabelas 3.15, 3.16, 3.17, 3.20 e 3.21 (de forma análoga, pode-se analisar os dados referentes à combinação "dlem2f9/TF_ABSL").

Como descrito na subseção 3.4.2, para avaliar o desempenho do classificador SVMs, foi utilizada a validação cruzada estratificada, com divisão do corpus em cinco subconjuntos. Assim, dos 1.014 PLs do corpus, foram utilizados, em

média, 811,2[30] PLs para construção do modelo de categorização e 202,8[31] PLs para avaliação do desempenho do modelo construído.

Os resultados obtidos nos 202,8 PLs utilizados para testar o desempenho do algoritmo SVMs são apresentados a seguir, lembrando que os resultados apresentados referem-se a médias, pois as estimativas foram obtidas por validação cruzada.

- **Comissão CEOF:**

Classificação pelo especialista / Classificação pelo classificador	*CEOF*	$\overline{CEOF}$	total
CEOF	99,4	17,6	117
$\overline{CEOF}$	18,6	67,2	85,8
total	118	84,8	202,8

- dos 202,8 PLs de teste, 118 foram apreciados pela CEOF;
- dos 118 PLs apreciados pela CEOF, o classificador decidiu corretamente que 99,4 deles eram da CEOF, ou seja, 84,3% (revocação) dos PLs apreciados pela CEOF foram corretamente identificados como CEOF pelo classificador;
- dos 84,8 PLs que não tratavam de matéria de competência da CEOF, o classificador decidiu corretamente que 67,2 deles realmente não eram da CEOF, ou seja, 79,2% dos PLs não apreciados pela CEOF foram corretamente identificados pelo classificador como não sendo da CEOF;
- dos 202,8 PLs de teste, 166,6 PLs (99,4+67,2) foram corretamente classificados pelo classificador como CEOF ou não CEOF ($\overline{CEOF}$), ou seja, 82,2%

[30] em quatro repetições da validação cruzada, o conjunto de treinanento foi formado por 811 PLs, e em uma repetição foi formado por 812 PLs. Portanto, o número médio de exemplos de treinamento usados para construção dos cinco modelos de categorização foi de 811,2.
[31] em quatro repetições da validação cruzada, o conjunto de teste foi formado por 202 PLs e em uma repetição foi formado por 201 PLs. Portanto, o número médio de exemplos de teste usados para avaliar os cinco modelos construídos foi de 202,8.

(acurácia) dos PLs foram classificados corretamente como sendo ou como não sendo da CEOF;

- dos 117 PLs que o classificador identificou como sendo da CEOF, 99,4 deles realmente foram apreciados pela CEOF, ou seja, 85% (precisão) dos PLs indicados pelo classificador para serem apreciados pela CEOF realmente tratavam de máterias de competência da CEOF.

- **Comissão CAS:**

Classificação pelo especialista / Classificação pelo classificador	*CAS*	$\overline{CAS}$	total
CAS	68,6	13,2	81,8
$\overline{CAS}$	22,2	98,8	121
total	90,8	112	202,8

- dos 202,8 PLs de teste, 90,8 foram apreciados pela CAS;
- dos 90,8 PLs apreciados pela CAS, o classificador decidiu corretamente que 68,6 deles eram da CAS, ou seja, 75,7% (revocação) dos PLs apreciados pela CAS foram corretamente identificados como CAS pelo classificador;
- dos 112 PLs que não tratavam de matéria de competência da CAS, o classificador decidiu corretamente que 98,8 deles realmente não eram da CAS, ou seja, 88,2% dos PLs não apreciados pela CAS foram corretamente identificados pelo classificador como não sendo da CAS;
- dos 202,8 PLs de teste, 167,4 PLs (68,6+98,8) foram corretamente classificados pelo classificador como CAS ou não CAS ($\overline{CAS}$), ou seja, 82,5% (acurácia) dos PLs de teste foram classificados corretamente pelo classificador como sendo ou como não sendo da CAS;

- dos 81,8 PLs que o classificador identificou como sendo da CAS, 68,6 deles realmente foram apreciados pela CAS, ou seja, 83,9% (precisão) dos PLs indicados pelo classificador para serem apreciados pela CAS realmente tratavam de máterias de competência da CAS.

- **Comissão CDC:**

Classificação pelo especialista / Classificação pelo classificador	*CDC*	$\overline{CDC}$	total
CDC	15,2	3,2	18,4
$\overline{CDC}$	5,2	179,2	184,4
total	20,4	182,4	202,8

- dos 202,8 PLs de teste, 20,4 foram apreciados pela CDC;
- dos 20,4 PLs apreciados pela CDC, o classificador decidiu corretamente que 15,2 deles eram da CDC, ou seja, 74,3% (revocação) dos PLs apreciados pela CDC foram corretamente identificados como CDC pelo classificador;
- dos 182,4 PLs que não tratavam de matéria de competência da CDC, o classificador decidiu corretamente que 179,2 deles realmente não eram da CDC, ou seja, 98,2% dos PLs não apreciados pela CDC foram corretamente identificados pelo classificador como não sendo da CDC;
- dos 202,8 PLs de teste, 194,4 PLs (15,2+179,2) foram corretamente classificados pelo classificador como CDC ou não CDC ($\overline{CDC}$), ou seja, 95,8% dos PLs foram classificados corretamente como sendo ou como não sendo da CDC;
- dos 18,4 PLs que o classificador identificou como sendo da CDC, 15,2 deles realmente foram apreciados pela CDC, ou seja, 83% (precisão) dos PLs

indicados pelo classificador para serem apreciados pela CDC realmente tratavam de máterias de competência da CDC.

- **Comissão CES:**

Classificação pelo especialista / Classificação pelo classificador	*CES*	$\overline{CES}$	total
CES	33,6	6,8	40,4
$\overline{CES}$	7,4	155	162,4
total	41	161,8	202,8

- dos 202,8 PLs de teste, 41 foram apreciados pela CES;
- dos 41 PLs apreciados pela CES, o classificador decidiu corretamente que 33,6 deles eram da CES, ou seja, 82% (revocação) dos PLs apreciados pela CES foram corretamente identificados como CES pelo classificador;
- dos 161,8 PLs que não tratavam de matéria de competência da CES, o classificador decidiu corretamente que 155 deles realmente não eram da CES, ou seja, 95,8% dos PLs não apreciados pela CES foram corretamente identificados pelo classificador como não sendo da CES;
- dos 202,8 PLs de teste, 188,6 PLs (33,6+155) foram corretamente classificados pelo classificador como CES ou não CES ($\overline{CES}$), ou seja, 93% (acurácia) dos PLs foram classificados corretamente como sendo ou como não sendo da CES;
- dos 40,4 PLs que o classificador identificou como sendo da CES, 33,6 deles realmente foram apreciados pela CES, ou seja, 83,2% (precisão) dos PLs indicados pelo classificador para serem apreciados pela CES realmente tratavam de máterias de competência da CES.

- **Comissão CSEG:**

Classificação pelo especialista / Classificação pelo classificador	*CSEG*	$\overline{CSEG}$	total
CSEG	23,4	5,4	28,8
$\overline{CSEG}$	6,2	167,8	174
total	29,6	173,2	202,8

- dos 202,8 PLs de teste, 29,6 foram apreciados pela CSEG;
- dos 29,6 PLs apreciados pela CSEG, o classificador decidiu corretamente que 23,4 deles eram da CSEG, ou seja, 79,2% (revocação) dos PLs apreciados pela CSEG foram corretamente identificados como CSEG pelo classificador;
- dos 173,2 PLs que não tratavam de matéria de competência da CSEG, o classificador decidiu corretamente que 167,8 deles realmente não eram da CSEG, ou seja, 96,9% dos PLs não apreciados pela CSEG foram corretamente identificados pelo classificador como não sendo da CSEG;
- dos 202,8 PLs de teste, 191,2 PLs (23,4+167,8) foram corretamente classificados pelo classificador como CSEG ou não CSEG ($\overline{CSEG}$), ou seja, 94,3% (acurácia) dos PLs foram classificados corretamente como sendo ou como não sendo da CSEG;
- dos 28,8 PLs que o classificador identificou como sendo da CSEG, 23,4 deles realmente foram apreciados pela CSEG, ou seja, 81,5% (precisão) dos PLs indicados pelo classificador para serem apreciados pela CSEG realmente tratavam de máterias de competência da CSEG.

4 Conclusões e trabalhos futuros

4.1 Conclusões

Nesta tese foi apresentada a metodologia de categorização automática de textos desenvolvida para a indicação automática de distribuição dos projetos de lei para as comissões permanentes da Câmara Legislativa do Distrito Federal.

Também foi proposta uma métrica para seleção de termos denominada de abs-logito, a qual corrige a assimetria apresentada pela métrica razão de chances (*odds ratio*), que privilegia a seleção dos termos mais prevalentes na categoria de interesse.

A métrica proposta - abs-logito - juntamente com a métrica *bi-normal separation* – que apresentou excelentes resultados na pesquisa realizada por Forman (2003) – foram utilizadas para construir as medidas de atribuição de pesos TF_ABSL e TF_BNS, respectivamente, ainda não utilizadas em trabalhos publicados. Essas medidas foram criadas com base na idéia proposta por Debole & Sebastiani (2003), de incluir no cálculo dos pesos para os termos a importância desses para a discriminação das categorias. Segundo essa proposta, o valor IDF - na fórmula TF_IDF - é substituído pelo escore calculado por uma das métricas de seleção de termos. No caso sob estudo, foram usadas as métricas abs-logito e *bi-normal separation*.

As medidas de atribuição de pesos TF_ABSL e TF_BNS foram comparadas com a medida usual de atribuição de pesos TF_IDF, e com as medidas - TF_IG, TF_GR, TF_OR e TF_QUI - utilizadas em trabalhos anteriores -, que substituem o valor de IDF pelo escore calculado de acordo com as métricas de seleção de termos ganho de informação, razão de ganho, razão de chances e qui-quadrado, respectivamente.

Empregando validação cruzada estratificada com cinco subconjuntos, as sete formas de atribuição de pesos citadas no parágrafo anterior foram comparadas usando o algoritmo *support vector machines* com dicionário global e local, redução da dimensionalidade por seleção de termos e aumento de peso para as palavras presentes nas ementas dos PLs, e para as palavras relacionadas às matérias de competência das comissões permanentes.

Igualmente aos resultados obtidos nos trabalhos de Apté et al. (1994) e Ng et al. (1997), e ao contrário dos resultados obtidos por Debole & Sebastiani (2003), neste trabalho os resultados usando dicionário local apresentaram desempenho superior ao usando dicionário global.

Com dicionário global, a forma de atribuição de pesos proposta TF_ABSL apresentou os maiores valores para F_1 macro para todas as representações estudadas, exceto para a representação que considera apenas os verbos - que apresentou o pior desempenho de todas as representações consideradas –, onde TF_OR exibiu desempenho melhor.

Usando dicionário local, os melhores resultados foram exibidos pelos pesos calculados por TF_ABSL e TF_BNS, que apresentaram desempenhos, em termos de F_1 macro, superiores aos apresentados por TF_IDF, em praticamente todas as representações vetoriais estudadas.

As formas de atribuição de pesos TF_QUI e TF_GR, que no trabalho de Debole & Sebastiani (2003) apresentaram os melhores desempenhos em termos de F_1 macro usando o algoritmo SVMs com dicionário global, exibiram resultados sistematicamente inferiores às duas medidas de atribuição de pesos propostas neste trabalho, com dicionário global e local.

A estratégia de contar em dobro a freqüência de ocorrência das palavras presentes nas ementas apresentou ganhos em relação à utilização do vetor de termos base, principalmente quando essa estratégia foi combinada com a redução da dimensionalidade pela freqüência de documentos, considerando apenas as palavras presentes em nove ou mais PLs. Os ganhos foram de 2,15% com dicionário global e de 2,98% com dicionário local, para os pesos calculados por TF_ABSL, que exibiram os maiores valores para F_1 macro.

Também a estratégia de aumentar em 30% o peso das palavras relacionadas às matérias de competência das comissões permanentes apresentou ganhos em relação à utilização do vetor de termos base, principalmente quando essa estratégia foi combinada com a redução da dimensionalidade pela freqüência de documentos, considerando apenas as palavras presentes em nove ou mais PLs. Os ganhos foram de 2,29% usando dicionário global e de 2,41% usando dicionário local, para os pesos calculados por TF_ABSL, que exibiram os maiores valores para F_1 macro.

As estratégias de contar em dobro a freqüência de ocorrência das palavras presentes nas ementas e de aumentar em 30% o peso das palavras relacionadas às matérias de competência das comissões permanentes, combinadas com a redução da dimensionalidade pela freqüência de documentos usando apenas as palavras presentes em nove ou mais PLs apresentou ganhos - em relação à utilização do vetor de termos base - de 3% com dicionário global e de 2,41% com dicionário local, para os pesos calculados por TF_ABSL, que exibiram os maiores valores para F_1 macro.

Dos três estudos sobre aumento de pesos comentados, o que conta em dobro a freqüência de ocorrência das palavras presentes nas ementas, considera apenas as palavras presentes em nove ou mais PLs e usa dicionário local foi o que apresentou o maior valor para F_1 macro. Esse resultado considerado como o melhor analisando o

conjunto completo de categorias também foi confirmado como a melhor estratégia ao analisar categoria por categoria. Para essa estratégia, usando tanto os pesos calculados por TF_ABSL como os calculados por TF_BNS, em nenhuma das cinco repetições da validação cruzada o valor para acurácia, precisão, revocação e F_1 foi inferior a 0,70 - valor a partir do qual o resultado foi considerado satisfatório neste trabalho – para as comissões CEOF, CAS, CDC, CES e CSEG.

Os resultados obtidos confirmaram a viabilidade prática da proposta, com a ressalva de que os classificadores para as comissões CDDHCEDP, CAF e CDESCTMAT, que apresentaram valores abaixo dos considerados satisfatórios neste trabalho, principalmente em termos de revocação, devem ser re-estimados após a obtenção de mais dados, antes de serem colocados em uso na prática.

4.2 Trabalhos futuros

Para confirmar o excelente desempenho apresentado pelos métodos de atribuição de pesos TF_ABSL e TF_BNS - propostos neste trabalho -, uma das direções de pesquisa futura consistiria em aplicar esses métodos a outras coleções de documentos, como a Reuters 21578 - dados utilizados nos estudos sobre atribuição supervisionada de pesos conduzidos por Debole & Sebastiani (2003 e 2005) – ou a Reuters Corpus, Volume 1, por exemplo.

Outra direção de pesquisa seria no sentido de desenvolver toda a metodologia para categorização automática de textos usando o *software* livre R. O presente trabalho foi desenvolvimento utilizando três *softwares*: 1) o *software* comercial IDE XML Client da Temis Text Mining Solutions, versão 2.0, para extração dos termos

dos textos dos PLs; 2) o *software* livre R - *The* R *Foundation for Statistical Computing*, versões 2.3.1/2.4.1, para limpeza dos termos extraídos pelo IdeXMLClient, seleção dos subconjuntos para validação cruzada, seleção de termos segundo as métricas descritas na seção 2.4.1, montagem das matrizes de treinamento e teste de acordo com as representações vetoriais de interesse para estudo, geração dos arquivos de entrada para o SVM^{light}, cálculo das medidas de performance para o classificador SVM e análises gráficas; e 3) o *software* SVM^{light}, uma implementação do algoritmo *Support Vector Machines*, para gerar os valores da função de decisão para os documentos do conjunto de teste.

Como verificado praticamente no final deste trabalho, o *software* R também possui uma implementação do algoritmo SVMs, sendo necessário para a sua utilização, fazer uma adaptação nos programas desenvolvidos para gerar os arquivos de entrada para o SVM^{light}. Além disso, da experiência adquirida com o desenvolvimento dos programas necessários à realização das análises propostas neste trabalho, parece totalmente viável a implementação de um algoritmo de *stemming*, por exemplo, no R. Assim, toda a análise desenvolvida nesta tese utilizando três *softwares* poderá ser realizada utilizando apenas um *software*, que é de distribuição livre.

Adicionalmente às complementações citadas, outra linha de pesquisa relevante consistiria em estudar algoritmos específicos para tarefas de classificação *multilabel*, posto que a abordagem adotada neste trabalho, que considerou o problema *multilabel* de classificação nas oito comissões permanentes como oito problemas independentes de classificação binária, não leva em consideração as correlações existentes entre as diferentes comissões que apreciam cada PL.

Em estudo preliminar realizado sobre o assunto, verificou-se a descrição na literatura de vários métodos especificamente desenvolvidos para solucionar o problema

multilabel. Dessa forma, seria interessante aprofundar a investigação sobre estratégias eficientes e de fácil implementação para executar tal tarefa.

Referências bibliográficas

AGRESTI, A., 1999, "On the logit confidence intervals for the odds ratio with small samples", **Biometrics**, v. 55, issue 2 (Jun.), pp. 597-602.

AHUJA, N., YANG M. H., ROTH D., 2001, **Support Vector Machines for Vision**. Disponível em: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/YANG1/cvonline2/cvonline2.html. Acesso em: 24 jan. 2007.

APTÉ, C., DAMERAU, F., WESS, S. H., 1994, "Automated learning of decision rules for text categorization", **ACM Transactions on Information Systems**, v.12, n. 3 (Jul.), pp. 233-251.

BERRY, M. W., DRMAČ, Z., JESSUP, E. R., 1999, "Matrices, vector spaces, and information retrieval", **SIAM Review**, v. 41, n. 2, pp. 335-362.

BERRY, M. W., DUMAIS, S. T., O'BRIEN, G. W., 1995, "Using linear algebra for intelligent information retrieval", **SIAM Review**, v. 37, n. 4 (Dec.), pp. 573-595.

BURGES, C. J. C., 1998, "A tutorial on support vector machines for pattern recognition". **Data Mining and Knowledge Discovery**, v. 2, pp. 121-167, Boston, Kluwer Academic Publishers.

CLEVERDON, C. W., 1991, "The significance of the Cranfield tests on index languages", In: **Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval**, Chicago, Illinois, USA, pp. 3-12.

DEBOLE, F., SEBASTIANI, F., 2003, "Supervised term weighting for automated text categorization". In: **Proceedings of SAC-03, 18the ACM Symposium on Applied Computing**, Melbourne, USA, pp. 784-788.

DEBOLE, F., SEBASTIANI, F., 2005, "An analysis of the relative hardness of Reuters-21578 Subsets", **Journal of the American Society for Information Science and Technology**, v. 56, n. 6, pp. 584-596.

DEERWESTER, S., DUMAIS, S. T., HARSHMAN, R., 1990, "Indexing by latent semantic analysis", **Journal of the American Society for Information Science**, v. 41, issue 6, pp. 391-407.

DISTRITO FEDERAL (Brasil). Câmara Legislativa. **Regimento Interno da Câmara Legislativa do Distrito Federal** - 6ª ed. consolidada. - Brasília: CLDF, 2005. Disponível em http://www.cl.df.gov.br/portal/legislacao/**regimento-interno**-julho-de-2005.pdf. Acesso em: 4 out., 2006.

EYHERAMENDY, S., MADIGAN, D., 2005, "A novel feature selection score for text categorization", In: **Proceedings of the International Workshop on Feature Selection for Data Mining: Interfacing Machine Learning and Statistics (in conjunction with 2005 SIAM International Conference on Data Mining)**, Newport Beach, CA, USA, April.

FORMAN, G., 2003, "An extensive empirical study of feature selection metrics for text classification", **Journal of Machine Learning Research**, v. 3, pp. 1289-1305.

GIRÁLDEZ, I., PUERTAS, E., MARIA GÓMEZ, J., 2002, "HERMES: Intelligent multilingual news filtering based on language engineering for advanced user profiling". In: **Proceedings of the Multilingual Information Access and Natural Language Processing Workshop**, VIII Iberoamerican Conference on Artificial Intelligence (IBERAMIA), pp. 81-88.

GÓMEZ HIDALGO, J. M., 2002, “Evaluating cost-sensitive unsolicited bulk email categorization”. In: **Proceedings of SAC-02, 17th ACM Symposium on Applied Computing**, pp. 615-620, Madrid, ES.

GÓMEZ HIDALGO, J.M., 2003, “Text Representation for Automatic Text Categorization”, **Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003)**, Madrid, ES, April. Disponível em: http://www.esi.uem.es/~jmgomez/tutorials/eacl03/index.html. Acesso em: 05 dez. 2006.

GÓMEZ HIDALGO, J.M., PUERTAS SANZ, E., BUENAGA RODRÍGUEZ, M. et al., 2002, “Text filtering at POESIA: A new Internet content filtering tool for educational environments”, **Procesamiento del Lenguaje Natural**, n. 29, pp. 291-292.

HAYES, P. J., ANDERSEN, P. M., NIRENBURG, I. B. et al., 1990, “TCS: A shell for content-based text categorization”. In: **Proceedings of the Sixth Conference on Artificial Intelligence Applications,** pp. 320-326, Santa Barbara, California, USA.

JOACHIMS, T., 1998, “Text categorization with support vector machines: learning with many relevant features”. In: **Proceedings of ECML-98, 10th European Conference on Machine Learning**, n. 1398 in Lecture Notes in Computer Science, pp. 137-142.

JOACHIMS, T., 1999, “Making large-Scale SVM learning practical”, **Advances in Kernel Methods: Support Vector Learning**, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press.

JOACHIMS, T., 2005, “A support vector method for multivariate performance measures”. In: **Proceedings of the 22 nd International Conference on Machine Learning (ICML)**, v. 119, pp. 377-384, Bonn, Germany.

KOHAVI, R., 1995, “A study of cross-validation and bootstrap for accuracy estimation and model selection”. In: **Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)**, pp. 1137-1143, Montreal, Canada.

LAN, M., TAN, C. L., LOW, H. B., 2006, “Proposing a New Term Weighting Scheme for Text Categorization”. In: **Proceedings of the 21th National Conference on Artificial Intelligence**, pp. 763-768, Boston, MA, USA, Jul.

LANCASTER, F. W., 1993, **Indexação e resumos: teoria e prática**. Brasília: Briquet de Lemos/Livros.

LEOPOLD, E., KINDERMANN, J., 2002, “Text categorization with support vector machines. How to represent texts in input spaces?”, **Machine Learning**, v. 46, issue 1-3, pp. 423-444.

LEWIS, D. D., **Reuters-21578 text categorization test collection**. Disponível em: http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt. Acesso em: 18 set. 2006.

LEWIS, D. D., **RCV1-v2/LYRL2004: The LYRL2004 Distribution of the RCV1-v2 Text Categorization Test Collection (14-Oct-2005 Version)**. Disponível em: http://www.jmlr.org/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm. Acesso em: 18 set. 2006.

LEWIS, D. D., YANG, Y., ROSE, T. et al., 2004, “RCV1: A new benchmark collection for text categorization research”. **Journal of Machine Learning Research**, v. 5, pp. 361-397.

LOPES, M. C. S., 2004, **Mineração de dados textuais utilizando técnicas de clustering para o idioma português**. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

MLADENIC, D., 1998, "Turning Yahoo! into an automatic Web page classifier". In: **Proceedings of the 13th European Conference on Artificial Intelligence (ECAI-13)**, pp. 473-474, Brighton, UK.

MOSCHITTI, A., BASILI, R., 2004, "Complex linguistic features for text classification: a comprehensive study". In: **Proceedings of ECIR-04, 26th European Conference on Information Retrieval Research**, Sunderland, UK.

NG, H. T., GOH, W. B., LOW, K. L., 1997, "Feature selection, perceptron learning, and a usability case study for text categorization". In: **Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval**, pp. 67-73, Philadelphia, USA.

OSUNA, E. E., FREUND, R., GIROSI, F., 1996, **Support Vector Machines: Training and Applications**, A.I. Memo, 1602, MIT A.I. Lab.

OZGÜR, A.; ÖZGÜR, L.; GÜNGÖR, T., 2005, "Text categorization with class-based and corpus-based keyword selection". In: **Proceedings of ISCIS-05, 20th International Symposium on Computer and Information Sciences**, n. 3733 in Lecture Notes in Computer Science, pp. 607-616.

SALTON, G., BUCKLEY, C., 1988, "Term-weighting approaches in automatic text retrieval", **Information Processing and Management: an International Journal**, v.24, issue 5, pp. 513-523.

SCHAPIRE, R E., SINGER Y., 2000, "BoosTexter: A Boosting-based system for text categorization", **Machine Learning**, v. 39, n. 2/3, pp. 135-168.

SCHWEIGHOFER, E., RAUBER, A., DITTENBACH, M., 2001, “Automatic text representation, classification and labeling in european law”. In: **Proceedings of ICAIL-2001, International Conference on Artificial Intelligence and Law**, pp.78-87, St. Louis, Missouri.

SCHWEIGHOFER, E., HANEDER, G., RAUBER, A. et al., 2002, “Improvement of Vector Representations of Legal documents with Legal Ontologies”. In: **Proceedings of BIS-2002, 5th International Conference on Business Information Systems**, Poznan, Poland, April.

SEBASTIANI, F., 2002, “Machine learning in automated text categorization”, **ACM Computing Surveys**, v. 34, n. 1 (Mar.), pp. 1-47.

SOUCY P., MINEAU G. W., 2005, “Beyond tfidf weighting for text categorization in the vector space model”. In: **Proceedings of IJCAI-05, 9th International Joint Conference on Artificial Intelligence**, pp. 1130-1135, Edinburgh, Scotland, UK.

TANG, L., LIU, H., 2005, “Bias Analysis in Text Classification for Highly Skewed Data”. In: **Proceedings of ICDM-05, 5th IEEE International Conference on Data Mining**, pp. 781-784.

TANTRUM, J., 2003**, Model based and hybrid clustering of large datasets**. Ph.D. dissertation, University of Washington, Washington, USA.

TEAHAN, W. J., 2000, “Text classification and segmentation using minimum cross-entropy”. In: **Proceeding of RIAO-00, 6th International Conference “Recherche d'Information Assistee par Ordinateur"**, Paris, FR.

WELLING, M., **Support vector machines**. Disponível em: http://www.ics.uci.edu/~welling/classnotes/papers_class/SVM.pdf. Acesso em: 21 jan. 2007.

WIENER E., PEDERSEN J. O., WEIGEND A. S., 1995, “A neural network for topic spotting”. In: **Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval**, pp. 317-332, Las Vegas, NV, USA.

YANG, Y., Liu, X., 1999, “A re-examination of text categorization methods”. In: **Proceedings of SIGIR-99, 22th ACM International Conference on Research and Development in Information Retrieval**, pp. 42-49, Berkeley, CA, USA.

YANG, Y., PEDERSEN, J.O., 1997, “A comparative study on feature selection in text categorization”. In: **Proceedings of ICML-97, 14th International Conference on Machine Learning**, pp. 412-420, Nashiville, TN, USA.

YETISGEN-YILDIZ, M.; PRATT, W., 2005, “The Effect of Feature Representation on MEDLINE Document Classification”. In: **Proceedings of the American Medical Informatics Association Fall Symposium (AMIA-05)**, Washington D.C., USA, October.

ZHANG, M., ZHOU, Z., 2006, “Multilabel neural networks with applications to functional genomics and text categorization”, **IEEE Transations on knowledge and data engineering**, v. 18, n. 10, pp. 1338-1351.

ZHANG T., OLES F., 2001, “Text categorization based on regularized linear classification methods”, **Information Retrieval**, v. 4, pp. 5-31.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo