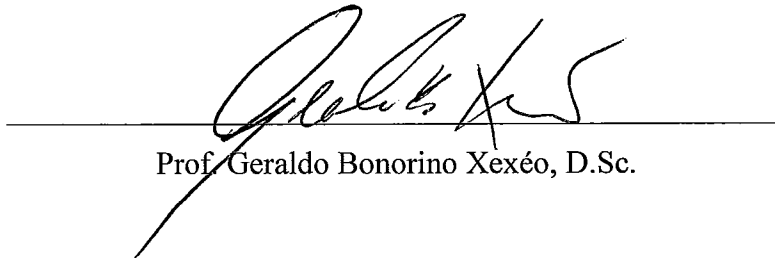
ATRIBUIÇÃO DE GRAU DE SIGILO: UMA ABORDAGEM DE CATEGORIZAÇÃO DE DOCUMENTOS

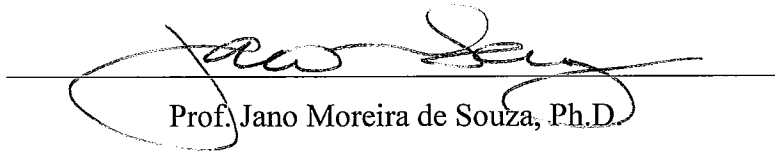
Carla Patricia Mello Lage

DISSERTAÇÃO 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 MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:



Prof. Geraldo Bonorino Xexéo, D.Sc.



Prof. Jano Moreira de Souza, Ph.D.

Prof[a]. Vera Maria Benjamim Werneck, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

SETEMBRO DE 2008

LAGE, CARLA PATRICIA MELLO

Atribuição de Grau de Sigilo: Uma Abordagem de categorização de documentos [Rio de Janeiro] 2008

XII, 95p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 2008)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Categorização Automática de Documentos
2. Classificação de Textos
3. Atribuição de Sigilo

I. COPPE/UFRJ II. Título (série)

Agradecimentos

Dizem que a vida é um navio. Nós paramos no porto, desembarcam alguns, embarcam outros e o nosso navio continua a seguir viagem. E na grande travessia que foi o meu mestrado, rumo aos resultados de pesquisa, muito tenho a agradecer e a muitas pessoas.

À minha mãe, que já terminou sua viagem, mas ainda é meu farol maior em direção ao profissionalismo, à honestidade, ao carinho, ao amor e à maternidade.

Ao meu pai, pela amizade, pelo carinho incondicional e por sempre acreditar em minha capacidade em superar obstáculos.

Ao meu melhor projeto realizado: meu filho Enzo, e ao co-responsável, Mário, meu ex-marido, por ter uma atitude colaborativa para que esta dissertação pudesse ser concluída.

À minha irmã Paula, que teve um papel fundamental nesta dissertação apoiando-me em momentos difíceis, fossem eles de ordem emocional, financeira ou até mesmo logística.

À minha irmã Ohana, que mesmo com o estresse do dia-a-dia incentivou-me constantemente a cuidar de mim mesma.

À Bartira pela amizade de todas as horas e ao Silvio pelo exemplo.

Aos meus chefes e ex-chefes na Marinha, pois sem seu apoio e incentivo não seria possível estar aqui. Ao Comandante (IM) George Hamilton, norte magnético em tecnologia da informação na Marinha do Brasil. Mestre com louvor em banco de dados, foi meu primeiro e grande incentivador.

Ao Aurélio, ao Roberto, ao Eduardo, à Carla, ao João, à Angélica, à Virgínia e todos os outros amigos da Diretoria de Finanças da Marinha e da Pagadoria de Pessoal da Marinha, que tiraram serviço no meu lugar para que eu pudesse assistir às aulas, ou que fizeram o possível e o impossível para não me ligarem durante os ensinamentos. Ao Comando da Força de Superfície que dispôs uma base para experimentos e à minha equipe naquela OM, Roberto, Roberto Oliveira, Canuto, Billy, Gedalias e Dyana pelo apoio. Aos colegas do Comando de Operações Navais e do Centro de Análises de Sistemas Navais por toda a compreensão.

À Rita Vidal, minha irmã de estudos, pelos ouvidos sempre disponíveis e pelo bom coração. Com ela descobri o que é doar atenção e horas de trabalho sem esperar

nada em troca. Aprendi que agindo assim, ganhamos muito mais do que poderíamos esperar.

À professora Myriam do Núcleo de Computação de Alto Desempenho (NACAD), que cedeu espaço em seu laboratório, uma mesa, um computador e acima de tudo, cedeu sua amizade. O convívio com a professora foi marcado por constantes e memoráveis demonstrações de seu excelente desempenho técnico e de sua dedicação à arte de ensinar. Essa profissional extremamente competente ostenta uma humildade no trato com seus pares e alunos, que só "os gigantes" em alguma área estão aptos a ostentar.

Ao amigo Serpa, companheiro de Marinha e de Java, ao Leonardo, pelo matemático apoio, ao Ângelo pela *lógica* atenção, ao Rodrigo, à Valeriana, à Paula, à Júnia e ao Oumar. Todos companheiros de risos e lágrimas, desabafos e abraços, cafés e brindes com Prossecco. Poder aportar neste *cluster* de amizade fez com que eu pudesse administrar a *grid* da minha vida, e abastecer-me de bondade para continuar a jornada.

Ao Yuri, do Grupo de Automação da Engenharia Elétrica (GTA), por alavancar meus conhecimentos e ouvir alguns lamentos. Ao Glaydston de Otimização e Combinatória, pela amizade no início desta jornada.

Ao professor Paulo Roberto de Medeiros, pelo apoio e por perguntar se eu sabia no que estava me metendo quando me inscrevi no mestrado. Respondi que sim. Não sabia.

Aos colegas Rodrigues Neto (Zé) e Ricardo Barros pela fundamental ajuda com o texto e pelo carinho.

À Jonice, pela disponibilidade, por ser um exemplo acadêmico a ser seguido, pelo carinho e principalmente por seus encorajadores sorrisos. À Adriana Vivácqua pela acessibilidade. À Patrícia Fiúza pela pronta atenção. À Carla Góes por me relembrar o quanto eu fiz esse curso por prazer. Ao Rafael Leonardo, meu sobrinho emprestado, aos Rodrigos, aos Vinícius, Melissa, Leandro, Bruno, Diogo, Stainam, Mutaleci, Cadú, Wallace, Wladimir e a todos os outros alunos da linha de Banco de Dados (BD) pelo coleguismo. Às secretárias de BD Patrícia Leal, Vina Guedes, Ana Paula e Carolina Barreiros por toda a paciência e pelo carinho diário. Às secretárias do programa, Solange, Sônia e Cláudia, pela simpatia e pelo pronto atendimento em problemas administrativos.

À Professora Vera Maria Benjamim Werneck, da UERJ, por aceitar fazer parte da banca examinadora.

Ao Professor Jano Moreira de Souza, que possibilitou que meu mestrado pudesse acontecer. O homem mais importante da navegação por bancos. Bancos de areia, bancos de cooperação, recifes de projetos, corais de conhecimento... Muitas vezes me senti feliz por alguns minutos de sua atenção, que invariavelmente significavam horas de estudo e pesquisa. Com ele aprendi o valor da *cooperação* e da *colaboração*. Muito valeram como aprendizado acadêmico e pessoal. Mais do que Gestão do Conhecimento aprendi *Gestão das Emoções*. Por isso, assim como outros alunos e orientados, sinto pelo Prof. Jano um mar de respeito e ondas de admiração.

Deixei para agradecer por último ao meu orientador: DSc. Geraldo Bonorino Xexéo. Não por ele ter sido tudo o que um orientador deve ser, pois ele sabe que o foi. Nem por sua disponibilidade e bom humor em atender-me, pois isso nunca faltou. Devido à sua habilidade em reconhecer e desenvolver talentos me senti muito honrada em ter sido aceita pelo Prof. Xexéo.

Visionário e capaz de quebrar paradigmas, meu orientador há muito venceu o monstro marinho do preconceito. Soube ver além das aparências, não hesitando um minuto sequer, em orientar uma aluna de tempo parcial, não oriunda da UFRJ, sem discriminações.

O simples fato de sua atenção para comigo incentivava-me a continuar quando a pesquisa entrava em *calmaria*. Sua invejável boa disposição dava-me ânimo quando ventos contrários deixavam-me receosa de não chegar ao meu destino. E de *peer* em *peer*, meu orientador conseguiu navegar com minha mente às vezes *fuzzy*.

Não o deixei por último por sua notória capacidade e proficiência lógica. Nem o deixei por último porque chorei quando recebi por e-mail seu "guia para orientados". Identifiquei-me em muitas das situações descritas no guia sobre o comportamento dos alunos, que muitas vezes atrasam a navegação ou deixam seu navio-estudo à deriva.

Meu orientador-amigo, sempre entendeu minhas *criptografadas* dificuldades, porém com carinho de irmão mais velho, nunca deixou que eu as usasse como justificativa para impedir meu desenvolvimento.

Durante meus estudos, o professor Xexéo foi um verdadeiro *lobo-do-mar* e sem ele certamente eu não atravessaria a tempestade da dissertação. Fornecendo coordenadas seguras e mapas de navegação confiáveis, ele ajudou-me a singrar mares de desconhecimento.

Eu o deixei por último porque não tenho palavras que realmente possam expressar todo meu agradecimento. Mesmo assim, MUITO OBRIGADA!

Tantos professores, amigos e colegas da COPPE e da Marinha fizeram parte dessa jornada. Entre idas-e-vindas, todos que conheci, ainda que não estejam mencionados aqui, sem exceção, deixaram marcas na minha alma e no meu coração.

Vou seguir viagem, desejando navegar por águas calmas, em "mar-de-almirante". Na COPPE, um dos lugares que movem o conhecimento neste país, mais do que o aprendizado, aperfeiçoei a mim mesma. Sei que muito mais está por ser descoberto, desenvolvido, testado e executado. Entretanto, meu maior desejo é que esse trabalho possa auxiliar alunos-navegantes vindouros, encurtando-lhes a travessia, nem que seja só um pouquinho.

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

ATRIBUIÇÃO DE GRAU DE SIGILO: UMA ABORDAGEM DE CATEGORIZAÇÃO DE DOCUMENTOS

Carla Patricia Mello Lage

Setembro/2008

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

A partir da necessidade da redução de atribuições de sigilo excessiva em documentos, este trabalho aborda o reconhecimento de padrões utilizando classificação automática de textos para categorizar documentos em português.

Para isso foi utilizada uma coleção de texto real, da Marinha do Brasil (MB), do ano de 2002, sendo composta de um tipo de documento denominado mensagem, específico da MB.

Os categorizadores utilizados foram o k-vizinhos próximos, o Naive Bayes e a Máquina de Vetor Suporte. Embora apenas a utilização dos categorizadores muitas vezes apresente bons resultados, os mesmos são extremamente dependentes dos documentos da amostra. Os resultados obtidos comprovam que a combinação de medidas na seleção de características apresenta ótimos resultados em quaisquer dos subconjuntos separados para teste.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

ALLOCATION DEGREE OF SECRECY: AN APPROACH FOR CATEGORIZATION OF DOCUMENTS

Carla Patricia Mello Lage

September/2008

Advisor: Geraldo Bonorino Xexéo

Department: Computer and Systems Engineering

From necessity to reduce document allocation high degree of secrecy, this work presents the pattern recognition using automatic text classification to categorize Portuguese texts.

This collection is composed with a specific kind of documents used by the Brazilian Navy. It has been used a real text collection from 2002 year.

The classifiers used were the K-nearest neighbor, the Naive Bayes and the Support Vector Machine. Although only applied a common classifier results in good responses, they are extremely dependent from the document samples. The results prove that the feature selection ensemble offer excellent results in any subsets selection for tests.

Sumário

Índice de Figuras

Índice de Tabelas

Capítulo 1 – Introdução

Nos últimos anos é notório o crescimento contínuo do volume de dados eletrônicos disponíveis. Textos, por exemplo, são produzidos aos milhares, no mundo todo, seja em ambientes empresariais, acadêmicos ou domésticos. Como lidar com essa quantidade de informação nesse formato é um dos desafios dos últimos tempos (FELDMAN e SANGER, 2006a).

Formas de processar e analisar esses dados, também fazem parte da descoberta do conhecimento (FAYYAD ET AL., 1996) em textos e vêm sendo estudadas sob diversos aspectos desde o final dos anos 50 (MARON e KUHNS, 1960; ROCCHIO, 1966). Diversas áreas como Recuperação das Informações, Inteligência Artificial, Mineração de textos vêm se integrando para que seja efetiva essa descoberta do conhecimento em bases textuais.

Nesse contexto, uma das áreas que se pode destacar é a Categorização[1] Automática de Documentos de Texto (CADT), onde são atribuídas uma ou mais classes predefinidas a documentos já existentes (SEBASTIANI, 2002a).

Em grandes volumes de dados a categorização manual de textos mostra-se extremamente trabalhosa. Pode-se dizer que cada vez mais, os processos manuais de avaliação e categorização de documentos tornam-se inviáveis, por demandar tempo e gerar resultados que não raramente terminam por se revelarem limitados (SEBASTIANI, 2002f). Além disso, existe a possibilidade de que a mesma seja executada de forma tendenciosa, onde a atribuição de certa categoria a um documento, pode estar fortemente baseada na visão parcial, de quem executa a categorização, sobre determinado assunto.

1.1 – Motivação

Segundo alguns estudos, militares de diversas nacionalidades costumam atribuir graus de sigilo maior que o necessário às informações. Conforme Almeida

[1] Conforme veremos adiante neste trabalho o termo "categorização de textos" refere-se a toda tarefa de enquadramento desses documentos em categorias predefinidas. Neste trabalho o termo "classificação" refere-se à utilização de documentos sigilosos ou "classificados".

(ALMEIDA, 2005a) esse processo é mais evidente na América Latina, especificamente nos países que foram governados por ditaduras militares por muitos anos. Entretanto, mesmo em países que não adotaram esse sistema de governo, como os Estados Unidos, verifica-se que a atribuição do sigilo excessivo faz parte da cultura militar (LEONARD, 2004). Poder-se-ia conjecturar que isso ocorre com os militares por serem treinados para manter sigilo sobre suas operações. Desta forma, este comportamento pode ser estendido para outros casos, como no trato de puras e simples tarefas administrativas.

A Marinha do Brasil (MB) é um dos órgãos que compõem o Ministério da Defesa (DEFESA, 2005). Sua missão é preparar e empregar o Poder Naval, a fim de contribuir para a defesa da Pátria. Pode-se dizer que na MB existem duas grandes perspectivas que se complementam para o cumprimento de suas tarefas. Uma é a perspectiva operativa, composta pelos meios navais, aeronavais, fuzileiros e seus comandos superiores. Essas Organizações Militares (OM) são, em síntese, responsáveis pela operacionalização da missão da MB. A outra perspectiva é a administrativa, que é inerente às diversas outras organizações, e que trata do apoio às tarefas operativas.

Entre as OM existem diversas formas de comunicação. Dentre elas destacamos a utilização de um documento específico para esse fim, denominado **mensagem**[2]. A mensagem é um documento usado para transmitir informações de modo sucinto e objetivo (DGMM, 2005). Embora todas as organizações militares usem este documento, é no setor operativo que ele encontra sua maior utilização. Assim como acontece em outros segmentos do Governo Federal, esses documentos, não raramente, são classificados como sigilosos e com grau de sigilo maior do que o necessário.

Juntando-se a avalanche de documentos gerados anualmente com a excessiva atribuição de sigilo, passamos a ter um grande volume de documentos sigilosos. Com isso, percebem-se alguns efeitos indesejados no dia-a-dia das organizações, e conseqüentemente, ocorre a banalização do emprego dos graus de sigilo.

Por exemplo, quando "tudo" é confidencial, lidar com esses documentos torna-se corriqueiro e muitas vezes os cuidados na salvaguarda dos mesmos acabam por não serem totalmente observados. Isso pode comprometer a segurança da informação, especialmente quando realmente for necessário um tratamento diferenciado a determinado documento. Então, para assegurar-se que determinado tratamento ocorra

---

[2] **Mensagem** é um documento especial utilizado pela Marinha do Brasil. Não confundir com *e-mail* ou mensagem SMS.

adequadamente, acaba-se por aumentar ainda mais o grau de sigilo atribuído (LEONARD, 2007).

Assim, em um efeito cascata, outro problema observado é o acúmulo de documentos sigilosos, que precisam ser mantidos em locais seguros, muitas vezes, por períodos de 10, 20, 30 até 50 anos, com possibilidade de renovação dos períodos de guarda, indefinidamente, para alguns casos.

Quanto maior o grau de sigilo, mais elaborado é o processo de guarda, reprodução, arquivamento, desclassificação e destruição dos documentos (CASA CIVIL, 2002a; EMA, 2002). Na Marinha, estabelece-se que a atribuição de um sigilo superior ao necessário, também implica na adoção de procedimentos possivelmente exagerados, com emprego de pessoal e gasto de material (DGMM,2005). Documentos que não são sigilosos são mais simples de serem arquivados e posteriormente eliminados. Com a redução de atribuição de sigilo, muitas vezes tratamentos diferenciados poderiam ser dispensados, como a sobrecarga de trabalho referente à reclassificação ou desclassificação para as informações sigilosas. Além disso, quando tudo é sigiloso, perde-se uma ferramenta de desinformação.

Pode-se observar que a aplicação do sigilo excessivo é um problema de cunho cultural. Um dos caminhos para a solução integral deste problema é a mudança na cultura organizacional. Esse tipo de mudança é lenta e gradual, e depende do empreendimento de diversos fatores chave, como por exemplo, seu planejamento a longo prazo e execução em etapas; o apoio da alta gerência; a implementação de um processo contínuo; participação em todos os níveis; relacionamentos da mudança com a estrutura, estratégia, sistemas de recompensa e sistemas de controle (TAVARES, 1991). Entretanto, ferramentas de tecnologia da informação são consideradas fortes aliadas nesses processos de mudança, haja vista a sua utilização na maioria das práticas sociais e de comunicação (FREITAS, QUINTANILLA et al., 2004) e sua contribuição para a competitividade e gestão de estratégias nas organizações (ROSSETI,2007).

1.2 – Objetivos

No contexto descrito na seção anterior, a aplicação de Categorização Automática de Documentos de Texto (CADT) é uma possibilidade de auxílio da resolução do problema existente.

Para que o sigilo possa ser reduzido é necessário primeiro que seja identificado o padrão atual com que os documentos já são identificados. A partir daí, devem ser

investigados e testados métodos de categorização existentes e adaptá-los com o propósito de serem utilizados para facilitar a atribuição de sigilo às mensagens.

O objetivo principal deste trabalho é analisar o uso de técnicas de CADT para determinar o grau de sigilo de uma nova mensagem. Em função de uma base de documentos similares já existentes no setor operativo e da análise realizada é proposto um protótipo de um categorizador para a MB.

Existem diversos métodos de categorização consagrados, como por exemplo, Árvores de Decisão (QUINLAN, 2003), Modelos de Markov Escondidos (Hidden Markov Model - HMM)(RABINER, 1989), K-vizinhos mais próximos (k-Nearest Neighbor - kNN) (MITCHELL, 1997), Naive Bayes (DUDA e HART, 1973), Máquina de Vetores Suporte (Suport Vector Machines - SVM)(JOACHIMS, 1998). Dentre esses, foi escolhido para estudo, inicialmente, o k-NN, por ser um dos mais simples algoritmos de categorização. O mesmo é baseado em similaridade, onde a mesma é verificada entre um novo documento e todos os documentos da base (AHA, KIBLER et al., 1991a; MITCHELL, 1997; YANG e LIU, 1999b). O segundo método escolhido foi o Naive Bayes que baseia-se na suposição de independência condicional entre os atributos dada uma classe (DUDA e HART, 1973; FRIEDMAN, GEIGER et al., 1997; LEWIS, 1998). Por último escolheu-se o SVM, que mapeia no espaço n-dimensional os pontos mais próximos à superfície de separação dos conjuntos que representam cada classe (JOACHIMS, 1998b; VAPNIK, 1999; YANG e LIU, 1999c).

Através dos testes dos métodos supracitados, pode-se aplicar o de melhor resultado no desenvolvimento de um mecanismo automático para facilitar a classificação das mensagens. O mecanismo deve ser capaz de determinar, com maior precisão possível, qual o grau de sigilo do documento, e desta forma reduzir os erros cometidos na determinação do sigilo desse tipo de documento.

1.3 – Organização da dissertação

A dissertação encontra-se organizada em cinco capítulos. Neste capítulo é feita uma breve introdução sobre categorização de documentos e o problema da atribuição de sigilo excessivo.

No capítulo 2 é realizada a descrição do problema e a importância da atribuição do sigilo. A coleção utilizada é descrita com detalhamento. Descreve-se também o histórico da base real utilizada, bem como a importância da aplicação deste trabalho para a MB.

Já no capítulo 3 são descritas as áreas que trabalham com descoberta de conhecimento em textos, técnicas aplicadas, algoritmos utilizados de uma forma geral, que consistem da revisão de literatura. São descritos os métodos de aprendizagem e com enfoque na categorização de documentos, incluindo as suas etapas e as medidas de desempenho a serem aplicadas quando de suas utilizações.

No capítulo 4 descreve-se a solução proposta. Nele são expostas as vantagens e desvantagens da utilização de outras possíveis abordagens. Além disso, é nesse capítulo que são enunciados os trabalhos realizados até o momento e as lacunas existentes, algumas parcialmente cobertas por este trabalho. Descrevem-se as etapas necessárias a solução do problema, os experimentos realizados, seus resultados e sua análise.

Por fim, no capítulo 5, é descrito o protótipo de um categorizador para a MB, são feitas as conclusões e proposições para trabalhos futuros.

Capítulo 2 – Sigilo de documentos

Segundo o Dicionário Aurélio a palavra sigilo é oriunda do latim sigillu, 'selo' e significa segredo. Relacionada a sigilo encontra-se a palavra sigiloso [De sigilo + -oso.], que contém ou envolve sigilo; secreto, sigilado (HOLANDA, 1999). De acordo com Teff (TEFFT, 1980) sigilo é "uma obrigatória ou voluntária, porém calculada, ocultação de informação, atividade ou relacionamento". Já Bok (BOK, 1984) resume sigilo, ou segredo, em uma "ocultação intencional".

De acordo com segurança das informações do âmbito da administração pública federal, sigilo além de também significar segredo, remete a palavra ao conceito de "conhecimento restrito a pessoas credenciadas; proteção contra revelação não autorizada" (CASA CIVIL, 2002f).

2.1 – Considerações sobre o sigilo

Segundo SIMMEL (SIMMEL, 1906a) o uso do sigilo é uma forma comum de interagir em sociedade. Todas as pessoas estão envolvidas em alguma forma de sigilo ou controle de informação. A separação psicológica entre pessoas de um grupo, organização ou sociedade e aquelas que não pertencem ao mesmo, também é inerente ao sigilo (BOK, 1984). Famílias não querem ver suas disputas, intimidades, interações privadas ou financeiras serem discutidas fora dos seus lares (DU BOYLAY, 1976). De forma análoga, agem grupos, empresas e governos.

O uso do sigilo permite que seja mantido um monopólio sobre algum tipo de conhecimento (SIMMEL, 1906b). Considera-se ainda que seu uso possibilite certo tipo de associações para evitar repercussão ou destruição política (ELLINGTON, 2004). Quando usado como estratégia, o sigilo pode ter uma conotação agressiva contra rivais, ou defensiva, contra ataques (TEFFT, 1980). Em níveis institucionais públicos ou privados, normalmente os assuntos são conduzidos sob sigilo quando se relacionam com um diferencial competitivo, com interesses de ganhos financeiros ou com a segurança nacional.

Ellington (ELLINGTON, 2004) descreve três modos de proteção da informação independentemente de sua natureza, individual ou institucional: i) A informação é tão bem velada que ninguém pergunta algo sobre ela; ii) A informação é protegida e todas

as perguntas sobre a mesma são recusadas; e por último iii) A informação é protegida e todas as perguntas respondidas com uma mentira. Podemos citar como exemplo a estratégia da "negação plausível" ou "negação capciosa" (*plausible deniability*) (WALTON, 1996) que é muito usada na política, na guerra e nas ações de espionagem, para dar cobertura a atos ilegais ou impopulares. Como exemplo do modo "i" pode-se citar essa mesma doutrina voltada para área de inteligência, onde o objetivo desse tipo de doutrina é negar o acesso à informação sem que o requisitante saiba que o acesso foi negado. Como exemplo do modo "ii" pode-se citar a política da imigração americana que usa essa doutrina, na concessão de vistos. Esse exemplo passa a representar o modo "iii" quando os escalões mais elevados atribuem a responsabilidade aos escalões mais baixos de comando. Quando a União Soviética abateu o avião espião U-2, o presidente Eisenhower tentou usar o método "iii" e descobriu que os Soviéticos já sabiam que sua negação não era plausível. Entretanto, no escândalo contra o Irã e o caso da Guatemala demonstram que alguns insucessos dessa doutrina não significaram que a mesma foi abandonada. Na prática, significa que se os métodos "i" ou "ii" falharem, os oficiais têm o método "iii" como plano backup para assegurar sua posição.

Existem também outros conceitos adotados; referentes à segurança das informações, dos quais podem ser citados, o de "menor conhecimento", onde informações sigilosas devem ser de conhecimento do menor número possível de pessoas e o da "responsabilidade do conhecimento", onde toda pessoa que tomar conhecimento de informação sigilosa torna-se automaticamente responsável pela preservação do seu sigilo (EMA, 2005).

Em Bok (BOK, 1984) são estabelecidas diferentes perspectivas de sigilo pessoal, onde existe a necessidade da manutenção de sigilo sobre si mesmo nos relacionamentos interpessoais e o sigilo do Estado, onde o governo possui um ocultamento intencional de informações sob a responsabilidade de seus órgãos. Neste trabalho foi enfocado o sigilo sob a égide governamental.

Dessa forma, há que se fazer a distinção entre dois requisitos básicos que são necessários para que se tenha acesso às informações. Primeiramente a "credencial de segurança", que é um certificado concedido por uma autoridade competente e que atribui ao portador, uma credencial para determinado grau de sigilo. É importante notar que apenas essa credencial não qualifica um indivíduo para ter acesso a uma informação (EMA, 2005). É imprescindível que o conhecimento de um assunto seja necessário para a execução de suas tarefas. Assim surge a "necessidade de conhecer" o outro requisito

básico para acesso às informações, definida como a "condição pessoal, inerente ao efetivo exercício de cargo, função, emprego ou atividade, indispensável para que uma pessoa possuidora de credencial de segurança, tenha acesso a dados ou informações sigilosos" (CASA CIVIL, 2002e).

2.2 – Sigilo Excessivo

De acordo com Leonard (LEONARD,2007), a superclassificação da informação é considerada em desafio. Embora, a decisão de atribuir um grau de sigilo seja de competência da autoridade que origina o documento, muitas vezes, o excesso pode, paradoxalmente, vir a trazer problemas para o próprio governo, impedindo de serem compartilhadas informações entre órgãos, setores ou o público que teria genuinamente a necessidade de conhecer.

Apesar das normas, publicações e doutrinas na MB (DGMM, 2005; EMA, 2002; EMA, 2005) e no Governo Federal (CASA CIVIL,2002) alertarem para o fato da importância da atribuição correta dos graus de sigilo, evitando a classificação excessiva, os problemas referentes à atribuição excessiva do sigilo continuam ocorrendo.

Junto com a cultura do sigilo navega o conceito de auto-preservação, onde não se deseja a responsabilidade por uma exposição indevida de informações que foi facilitada por um erro na atribuição do sigilo. Assim, o fato que está arraigado na cultura militar é que "na dúvida, peca-se por excesso".

Outro problema do sigilo excessivo é a quantidade de lixo sigiloso, digital ou mesmo em papel. Só para exemplificar, a MB gerou no ano de 2005 mais de 4.000.000 de documentos (CENADEM, 2005). Boa parte desses documentos é sigilosa e recebe um tratamento diferenciado no seu manuseio, por anos a fio. O prazo para a salvaguarda pode chegar até 50 anos ou ser renovado indefinidamente (CASA CIVIL, 2002d), configurando-se um acúmulo desnecessário de documentos classificados.

2.2.1 – Sigilo Excessivo e Transparência Governamental

Outra conseqüência da atribuição de sigilo excessiva refere-se à transparência governamental. Trata-se de um aspecto muito importante atualmente. Dentre as preocupações dos atuais governos democráticos encontra-se a transparência governamental, que é considerada um dos fatores basilares da democracia (ALASDAIR, 2006; DOOREY, 2007b). Embora o dilema sigilo das ações governamentais versus

transparência seja antigo (DOOREY, 2007a), ainda hoje procura-se encontrar uma equação de equilíbrio entre os mesmos.

Segundo a Red de Seguridad y Defensa de América Latina – RESDAL - (RESDAL, 2008), no setor defesa, são identificadas diversas possíveis causas para o déficit de transparência. Dentre elas, cita-se a falta de aderência entre os objetivos de defesa e nacionais, a proteção de interesses corporativos e particulares (ALMEIDA, 2005b) e destaca-se a cultura do sigilo e seus excessos.

2.3 – Sigilo como Requisito Não Funcional

Tradicionalmente em engenharia de software, existem os chamados requisitos funcionais, que definem uma condição ou uma capacidade que um sistema deve estar de acordo, especificando ações que o mesmo deve ser capaz de executar, sem levar em consideração restrições físicas. Assim, os requisitos funcionais modelam o comportamento de entrada e saída de um sistema (PRESSMAN, 2001).

Diferentemente dos supracitados, existem os chamados requisitos não funcionais, que estão relacionados com a qualidade do software desenvolvido (CHUNG, NIXON et al., 1999). Normalmente esses requisitos abrangem o sistema como um todo, interagem entre si, e muitas vezes podem ser subjetivos. Dentre esses podemos citar os que dizem respeito à evolução do sistema, como a manutenibilidade e a escalabilidade, ou à execução do sistema, como o desempenho, a usabilidade e a confiança.

Segundo Somerville (SOMMERVILLE, 2006) a confiança em um sistema depende da disponibilidade em fornecer serviços quando solicitado, da confiabilidade em fornecer serviços conforme esperado, da segurança em operar sem falhas catastróficas e da proteção contra intrusões acidentais ou intencionais. A proteção contra a exposição de informações sigilosas relaciona-se com controle de acesso, com redução de vulnerabilidade e de outras ameaças, constando de um "pacote" de proteção. Assim a manutenção apropriada do sigilo é um requisito capaz de influenciar na qualidade do software.

2.4 – Documentos na MB

Considera-se documento na MB toda informação registrada em suporte material, suscetível de consulta, estudo, prova e pesquisa. Existem basicamente dois tipos de documentos: os documentos físicos, ligados inseparavelmente ao meio físico em que são registrados; e os documentos eletrônicos, que são as seqüências de bits que podem ser traduzidas por meio de programas de computadores (SGM, 2006).

Os documentos existentes na MB são divididos em administrativos, operativos, publicações e os especiais, como as mensagens, por exemplo (SGM, 2006). Esses documentos são transmitidos pelo Sistema de Comunicações da Marinha (SISCOM), que é o sistema responsável pelas comunicações navais.

As comunicações supracitadas podem ser realizadas em quatro modalidades (voz, texto, dados e imagem). No âmbito das Forças Armadas, essa transmissão ocorre através dos meios de comunicações, que são padronizados em: Ótico, Acústico, Elétrico e Postal. Esses meios de comunicações podem ser subdivididos em canais de comunicações: i) Canais do meio ótico, como por exemplo, bandeiras, holofote, infravermelho, artefatos pirotécnicos, fumígenos etc; ii) Canais do meio acústico, como apito, buzina, megafone, canhão, sinais submarinos; iii) Canais do meio elétrico, como telégrafo, telefone, televisão, fax, dados, radiodados; e iv) Canais do meio postal, como mensageiro e correio.

2.5 – A Mensagem

A mensagem é um tipo de documento especial utilizado na MB que, apesar do nome, não tem relação com os e-mails atuais, nem com os sistemas de troca de mensagens na Internet ou com as mensagens instantâneas utilizadas em celulares. É uma forma de transmitir informações de modo mais sucinto possível e por isso quando assumem a forma de documentos, são pequenos, com formato específico.

2.5.1 – Breve histórico

Desde a utilização de pombos-correio, passando pela utilização de navios para transporte de documentos, até a utilização dos documentos eletrônicos, diversas transformações vieram ocorrendo na utilização e no trâmite dos documentos navais. A partir de 1830, as correspondências, de um modo geral, passaram a ser transportadas em navios específicos para esta tarefa, os chamados transportes de paquetes (LACERDA,

1928). Nessa época as mensagens já existiam mas não especificamente com o formato de documento utilizado hoje em dia.

As mesmas começaram a tomar forma a partir de 1906, com a participação da MB no International Radio Consultative Committee – CCIR (SMITH, 1974). Segundo registros do Serviço de Documentação da Marinha (SDM) estima-se que apenas a partir da década de 20, com o uso da telegrafia (LACERDA,1928), a utilização das mensagens tenha se intensificado, passando a ser registradas em papel e tomando o formato mais similar ao que conhecemos hoje. Assim permaneceram até a implantação do Sistema de Gerência de Documentos Eletrônicos da Marinha (SiGDEM) em 1997. Esse sistema é utilizado até os dias de hoje, e vários documentos, dentre eles as mensagens, são tramitados e armazenados digitalmente dentro das OM.

Conforme já citado, as informações contidas nas mensagens eram e ainda são, transmitidas através de diversos tipos de meios e canais.

2.5.2 – Descrição

Uma peculiaridade das mensagens é que, embora sem assinatura, esses documentos têm a conotação de serem expedidos pelo comandante titular da Organização Militar (OM). Apenas esta característica já seria suficiente para inferirmos a necessidade da correta atribuição de classificação, além é claro, do criterioso trato com as informações nele contidas.

Os documentos do tipo mensagem têm validade de um ano, e não podem servir como referência após esse período. Entretanto, na prática, um assunto perdura por um ou mais anos e elas, não raro, servem como uma "memória" do andamento do assunto, de como ele foi tratado e/ou resolvido. Por isso, obviamente as bases digitais dos anos anteriores são mantidas. Os documentos do tipo mensagem, como têm validade pequena, não são desclassificados. Apesar do texto pequeno, existe um grande volume deste tipo de documento. Seria muito trabalhoso analisar manualmente e desclassificar esse "lixo" digital.

No caso das mensagens, assim como outros documentos do governo federal e da MB, vem escrito o grau de sigilo com a qual foi classificada. Entretanto, diferentemente de outros documentos, caso não seja sigiloso a palavra "ostensivo" não aparece escrita.

Para a tramitação, as mensagens sigilosas devem ser criptografadas utilizando-se os recursos criptológicos em vigor. Devem ainda ser tomados todos os cuidados quanto à segurança da informação digital e à salvaguarda de assuntos sigilosos em vigor.

Para identificarem-se as mensagens é feito uso das informações da data e da hora em que a mesma é expedida. Dá-se a esse tipo de identificação o nome de grupo data-hora. Poderão ser também, identificadas pelo acréscimo do indicativo da OM remetente. Em todos os casos, a informação usada para identificar uma mensagem é tão curta quanto possível. A mensagem é composta por cabeçalho, onde são passadas informações sobre tipos de transmissão, texto e fechamento. Cada parte da mensagem pode possuir componentes, os quais se desdobram em Elementos e Conteúdos que não são obrigatórios, conforme exemplificado na Figura 1, que é um exemplo de uma mensagem ostensiva. Na figura 2 podemos observar o mesmo exemplo com a atribuição de sigilo confidencial.

```
PREFERENCIAL
P-151202Z/FEV/02
DE        DITELM
PARA  ERM
RIO
SVC
GR06
BT
1102Z, "é favorável sugestão dessa ER"
BT
```

Figura 1 Exemplo de documento mensagem ostensiva.

```
PREFERENCIAL
P-151202Z/FEV/02
DE        DITELM
PARA  ERM
RIO
SVC
GR06
BT
CONFIDENCIAL
1102Z, "é favorável sugestão dessa ER"
BT
```

Figura 2 Exemplo de documento mensagem sigilosa.

2.6 – Grau de Sigilo

Segundo o mesmo decreto que regulamenta a salvaguarda de assuntos sigilosos (CASA CIVIL, 2002c; EMA, 2005) os documentos dividem-se em ostensivos e sigilosos. Os ostensivos não possuem classificação e podem ser franqueados ao público. Já os sigilosos são os documentos cujo conhecimento irrestrito ou divulgado pode acarretar qualquer risco à segurança da sociedade, do Estado, da MB ou das demais Forças Armadas, bem como aqueles necessários ao resguardo da inviolabilidade da intimidade da vida privada, da honra ou da imagem das pessoas. A MB, assim como outros órgãos da administração pública federal, também adota as denominações do referido decreto.

O sigilo deve ser atribuído por ocasião da confecção do mesmo, considerando o assunto, a finalidade, os aspectos essenciais, etc. Esta atribuição determina, por conseguinte, os recursos pessoais e materiais que deverão ser utilizados. O grau de sigilo deve ser determinado pela autoridade competente, e que tenha credencial de segurança para tal atribuição. Neste trabalho, iremos utilizar as definições para documentos sigilosos e não sigilosos (EMA, 2002; EMA, 2005), onde na MB são feitas da seguintes forma:

" - Ostensiva - é aquela que contém assuntos que são ou possam ser do conhecimento do público em geral;

- Reservada - é aquela que contém matérias que não devam, imediatamente, ser do conhecimento do público em geral;

- Confidencial - é aquela que contém matérias cujo conhecimento e divulgação possam ser prejudiciais aos interesses da MB;

- Secreta - é aquela que contém matérias que requeiram rigorosas medidas de segurança e cujo teor ou características possam ser do conhecimento de agentes públicos que, embora sem ligação íntima com seu estudo ou manuseio, sejam autorizados a dela tomarem conhecimento em razão de sua responsabilidade funcional; e;

- Ultra-Secreta - é aquela que contém matérias que requeiram excepcionais medidas de segurança e cujo teor só deva ser do conhecimento de agentes públicos ligados ao seu estudo e manuseio."

Após expirado o prazo preconizado pela classificação, os documentos são desclassificados automaticamente, tornando-se ostensivos, a menos que sejam reclassificados. Para isso, normalmente realiza-se uma comissão que periodicamente e manualmente analisa os documentos sigilosos arquivados. Os graus de sigilo podem

também ser revistos a qualquer momento pela autoridade que classificou o documento (EMA, 2005).

2.7 – Atribuição Automática de Sigilo

Existe certa dificuldade em categorizar automaticamente um documento quanto ao sigilo. Diversos assuntos podem ser considerados sigilosos, dependendo muito do contexto em que está inserido e até da temporalidade. Em alguns momentos algo pode ser considerado como sigiloso, como a existência de uma operação militar. Após a ocorrência da mesma, os documentos envolvidos podem não possuir a mesma importância, podendo portanto ser reclassificados ou até desclassificados.

Reduzir a atribuição de sigilo dos documentos ao realmente necessário não é uma tarefa trivial. Uma das formas de se fazer isso seria, inicialmente identificando o padrão de atribuição de sigilo atribuído aos documentos. A partir de uma base existente, este pode ser identificado e utilizado para atribuir sigilo de forma automática. Para reconhecimento e validação desse padrão, pode-se fazer uso da Categorização Automática de Documentos de Texto (CADT).

Capítulo 3 – Categorização Automática de Documentos de Texto (CADT)

Estima-se que, em torno de 80% das informações das companhias estejam armazenadas sob a forma textual (KARANIKAS e THEODOULIDIS, 2002; SPINAKIS, 2005; TAN, 1999). Como textos são naturalmente não estruturados, a aquisição de informações advindas deste tipo de formato é um pouco mais trabalhosa (HEARST, 1999).

A gerência de textos através do esforço humano mostra-se cara e ineficiente, fatos estes que motivaram o desenvolvimento de métodos automáticos, algoritmos e ferramentas para lidar com grande volume de textos (LAGUS, 2000). Neste capítulo serão revisadas algumas das principais formas desenvolvidas, motivadas pelo supracitado problema.

3.1 – A Descoberta de Conhecimento em Textos

Por descoberta de conhecimento entende-se a extração tanto de informação explicitamente declarada, através da estrutura ou regras inerentes ao modelo de dados, quanto de informação implícita existente em fontes de dados(FELDMAN e DAGAN, 1995b). Esse processo de Descoberta de Conhecimentos em Textos (Knowledge Discovery in Textual Databases - KDT) (FELDMAN e DAGAN, 1995a; FELDMAN e SANGER, 2006b) é também chamado - Mineração de dados em textos (Text Data Mining) ou simplesmente Mineração de Textos (Text Mining) (SEBASTIANI, 2002f).

Segundo Hotho et al. (HOTHO, NURNBERGER et al., 2005a), a descoberta do conhecimento é inspirada pela Mineração de Dados (Data Mining), que procura descobrir padrões e pretende extrair conhecimentos úteis de dados não estruturados ou semi-estruturados.

O principal objetivo das técnicas de mineração de textos é a manipulação de documentos em formato texto que se encontram de forma não-estruturada. Transformar textos (formato não estruturado) em tabelas (formato estruturado) tem sido uma abordagem amplamente utilizada para possibilitar o uso da maioria dos algoritmos que utilizam dados estruturados (DIAO, LU et al., 2000).

Como muitas informações estão armazenadas em formato texto (KARANIKAS e THEODOULIDIS, 2002), acredita-se que as técnicas de mineração de textos possuam um grande valor comercial (ARANHA e PASSOS, 2006). De fato, as aplicações deste gênero fornecem uma nova dimensão das informações e que, se bem exploradas, podem se tornar um diferencial no domínio do negócio onde forem aplicadas (FURTADO, 2004). Este processo de Descoberta de Conhecimentos em Textos é especificado em algumas etapas de processamento que devem ser aplicadas ao conjunto de dados a ser utilizado pelo usuário, de maneira que se possa obter algum conhecimento ainda não observado.

Furtado (FURTADO, 2004; KROEZE, MATCHDEL et al., 2003) conceitua que o KDT é igual ao KDD (Knowledge Discovery in Databases), que foi proposto em 1989 por Fayyad (FAYYAD ET AL., 1996). Este refere-se às etapas que produzem conhecimento a partir de dados e, principalmente, à de mineração dos dados, que é a transformação dos dados em informação. Assim, segue-se os principais passos do processo de extração de conhecimento em bases de dados e transforma-se dados de baixo nível em conhecimento de alto nível.

Em diversos trabalhos a mineração de textos é conhecida como uma extensão da mineração de dados ou da descoberta de conhecimento em bases de dados estruturados Dorre et al. (DORRE, GERSTL et al., 1999). Já Hearst (HEARST, 1999) contesta a idéia de que a mineração de textos seja uma evolução da mineração de dados, e sim apenas uma parte da mesma, onde a fonte de dados é não estruturada.

Aranha (ARANHA e PASSOS, 2006) considera que a descoberta de conhecimento de bases de dados textuais é um campo novo e multidisciplinar que inclui conhecimentos de áreas como aprendizado de máquina, inteligência artificial, estatística, visualização dos dados, Lingüística e Ciência Cognitiva.

Outro autores consideram ainda a Mineração de Textos (Text Mining) como uma nova técnica derivada da Recuperação da Informação (Information Retrieval - IR) (HOTHO, NURNBERGER et al., 2005b; KROEZE, MATCHDEL et al., 2003) e do Processamento de Linguagem Natural (Natural Language Processing - NLP) (KAO e POTEET, 2007).

Kroeze (KROEZE, MATCHDEL et al., 2003) e Hearst (HEARST, 1999) definem ainda a *intelligent text mining* como um processo de criação de conhecimento automático e conceituam o processo de investigação definindo-o como um novo tipo de mineração inteligente de texto.

Para Sebastiani (SEBASTIANI, 2006), a tarefa de Classificação de Textos, pode ser subdividida em Categorização de Textos e Clusterização de Textos. Na clusterização são identificadas as categorias existentes e na categorização, de posse de categorias predefinidas, documentos podem ser "enquadrados" em uma ou mais categorias. Entretanto, quando se tratando de categorização, os trabalhos daquele autor também passam a usar os termos "classificação" e "categorização" como sinônimos (SEBASTIANI, 2002d; SEBASTIANI, 2002f; SEBASTIANI, 2006). Em vários trabalhos (LIAO e VEMURI, 2002a; RÁEZ, 2007) os termos "classificação" e "categorização" são usados indistintamente.

Neste trabalho o termo "classificação" refere-se aos graus de sigilo, e o termo "categorização de textos" refere-se a toda tarefa de enquadramento desses documentos em categorias predefinidas. A utilização desses termos é baseada na declaração de Zeredo (ZEREDO, 2002) onde "comumente, documentos de inteligência são "classificados" em função de grau de sigilo, e "categorizados" em função dos conceitos que compõem o conteúdo, ou idéias expressas nestes documentos."

3.2 – Áreas que lidam com informações textuais

Para que seja efetivo o descobrimento de informações textuais, muitas áreas surgem em um âmbito interdisciplinar. Podemos citar a recuperação da informação, a mineração de dados, o processamento de linguagem natural, as abordagens estatísticas, que invariavelmente utilizam-se das mesmas técnicas, algoritmos, e métodos (MOHAMMAD, 2007). Entretanto, o enfoque quanto aos resultados são um tanto quanto distintos, conforme a descrição:

i. Recuperação da Informação – Procura textos que podem conter a informação solicitada. O objetivo principal é o de selecionar palavras adequadas para compor um índice que facilite a localização dos documentos (BAEZA-YATES, 2004). Nesse caso, palavras pouco freqüentes são muito discriminantes, mas ocupam espaço no índice desnecessariamente, pois não recuperam muitos documentos. Por outro lado, se não utilizadas, os documentos não podem ser recuperados através delas. Área de pesquisa em Recuperação de Informação (RI) surgiu na década de 1950, com os sistemas para localização de livros em bibliotecas convencionais. As pesquisas em RI visam desenvolver metodologias que permitam um melhor desempenho dos Sistemas que

usam RI, em identificar a necessidade de informação do usuário e recuperar itens relevantes a esta necessidade (YANG, 1997).

Como subárea de RI podemos citar a Filtragem da Informação ou Filtragem de Texto, que por definição é uma técnica capaz de procurar textos que contêm uma ou mais palavras específicas (SHETH, 1994). Está bastante relacionada com a categorização e a clusterização de documentos e é muito usada na confecção de agentes para recuperação de notícias.

ii. Aprendizado de Máquina (AM) - é uma área da inteligência artificial concebida para desenvolver técnicas que tornem o computador capaz de "aprender" a partir de um conjunto de dados (MITCHELL, 1997).

iii. Processamento de Linguagem Natural (PLN) – Estuda a utilização da linguagem de forma mais natural possível para existir a comunicação com computadores. A tarefa de processar uma linguagem natural permite que os seres humanos comuniquem-se com os computadores, utilizando a linguagem com a qual mais estão habituados (BAEZA-YATES, 2004). Assim, podem se eliminadas a necessidade de adaptação a formas menos amigáveis de interação, como linguagens de máquina.

3.3 – Categorização Automática de Documentos

A CADT é definida como a construção de um modelo de categorização que é capaz de atribuir a um novo documento a uma ou mais classes, definidas anteriormente, a qual ele pertence (ANTONIE e ZAÏANE, 2002c; LIAO e VEMURI, 2002b).

A categorização automática de textos é também definida como a construção de um modelo de categorização que é capaz de atribuir a um novo documento a uma ou mais classes, definidas anteriormente, a qual ele pertence (ANTONIE e ZAÏANE, 2002b; LIAO e VEMURI, 2002c).

Utiliza-se para isso a categorização de documentos que conforme Camargo (CAMARGO, 2007) "A compartimentação propiciada pela classificação permite a organização automática de conteúdos necessária ao desenvolvimento de processos nos quais os executores das etapas não devem ter conhecimento das atividades desenvolvidas por todos, como nas atividades de projetos militares ou de segurança nacional, por exemplo."

Muitas aplicações valem-se das técnicas de aprendizado supervisionado como por exemplo, a indexação automática de documentos, a filtragem de documentos (PAYNE, 1997;LANG, 1995), a generalização automática de meta-dados (PRUDÊNCIO e LUDERMIR, 2007), o uso de catálogos de recursos na web e quaisquer outras aplicações que requeiram organização de documentos (SEBASTIANI, 2002).

Com o aumento constante do número de textos em formato digital (CASTRO, 2000d), a categorização de textos vem assumindo destaque nos cenários acadêmico e empresarial. Considerada uma técnica de aprendizado supervisionado (YANG e LIU, 1999d), o mesmo se dá através dos exemplos que já foram manualmente categorizados. Assim, categorizadores típicos "aprendem" a partir dessa classificação prévia, ou seja, isto é feito a partir de categorias predefinidas (SEBASTIANI, 2002c). Na Figura 2 podemos contextualizar a Categorização de Textos.

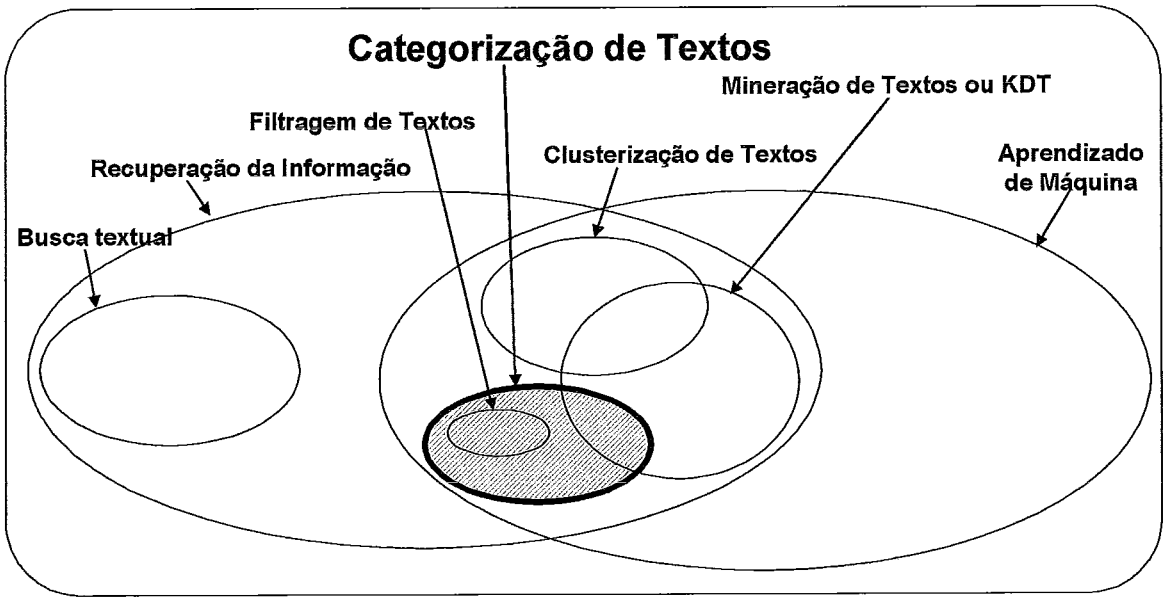


Figura 3 Adaptado de Sebastiani (SEBASTIANI,2002).

A garantia da precisão do método de categorização, em qualquer que seja o contexto de categorização, e a redução do esforço humano para gerar bases pré-classificadas para treinar os categorizadores, têm sido alvo de muitas pesquisas (SEBASTIANI, 2006) e ainda são problemas em aberto.

3.4 – Etapas da Categorização de Documentos

Normalmente os procedimentos para descobertas de conhecimento em textos e também de categorização de documentos, são divididos em etapas ou fases. Alguns autores dividem o trabalho em apenas 3 etapas, considerando o pré-processamento, a representação dos documentos ou a seleção dos dados, e a categorização propriamente dita (CAMARGO, 2007d; DEBOLE e SEBASTIANI, 2002). Outros autores dividem em 4 etapas: o pré-processamento, a representação de documentos, a indução (a categorização) e uma etapa de avaliação ou pós-processamento (CASTRO, 2000c; MONTEJO-RÁEZ, 2005a). Outros ainda incluem uma etapa anterior ao pré-processamento, que seria a coleta dos documentos relevantes (WEISS, INDURKHYA et al., 2004a).

3.4.1 – Coleta de documentos relevantes

A primeira refere-se à coleta de documentos relevantes ao domínio da aplicação do conhecimento a ser adquirido. O sucesso de um sistema de categorização de documentos depende, além dos algoritmos de aprendizado, de uma observação apropriada do problema (WEISS, INDURKHYA et al., 2004b). Normalmente isso acontece através da escolha de uma coleção de documentos que realmente o represente (MONTEJO-RÁEZ, 2005b).

3.4.2 – Pré-Processamento

No pré-processamento os documentos são preparados para serem representados em um formato adequado e assim serem submetidos aos algoritmos de processamento estatístico (SILVA e VIEIRA, 2007c). As técnicas de preparação do texto para a categorização visam retirar tudo o que não é significativo, tornando o texto mais enxuto e a lista de termos das categorias mais sucinta. A vantagem em usar tais técnicas está no fato de que elas reduzem expressivamente o trabalho a ser feito nas fases de seleção de características e de categorização, diminuindo significativamente, assim, o tempo de processamento. Apesar das vantagens, deve-se tomar cuidado com a sua utilização para não comprometer o resultado do processo eliminando termos relevantes à representação do texto (LENNON, PIERCE et al., 1981). As técnicas de identificação de termos, remoção de caracteres inválidos, e remoção de *stopwords* são utilizadas para preparação do texto.

3.4.2.1 – Retirada de stopwords

As palavras que não são passíveis de serem representantes de alguma categoria são conhecidas como *stopwords* ou palavras negativas e podem ser representadas por artigos, pronomes, preposições, advérbios e outras palavras que se apresentem com elevada ou baixa freqüência nos textos.

Para realizar esta tarefa, deve-se elaborar uma lista com todas as *stopwords* referentes ao domínio que o sistema deverá tratar. Essa lista é chamada de *stoplist* (FOX, 1989) e pode ser elaborada manualmente, definindo-se as palavras que não devem aparecer no índice. E também pode ser elaborada de forma automática a partir das palavras com maior e menor freqüência no texto, bem como das que aparecem em maior quantidade de documentos. O sistema de CADT compara cada palavra presente no texto com as da *stoplist*. Caso o termo esteja presente na lista, ele é excluído do texto.

A maioria das palavras mais freqüentemente usadas em línguas, são palavras sem valor de índice, como por exemplo, conjunções, preposições e verbos de ligação. O uso de *stoplist* apropriadas costuma apresentar redução de tempo de processamento e conseqüente melhoria na recuperação da informação (BERG, 1997).

3.4.2.2 – Stemming

Stemming é o processo de combinar as formas diferentes de uma palavra em uma representação comum(HULL, 1998; ORENGO e HUYCK, 2001). Tem como objetivo remover prefixos e sufixos, permitindo a recuperação de variações sintáticas das palavras. Um stem é a parte que resta de um termo quando são retirados seus afixos.

Os algoritmos de *stemming* mais conhecidos são os algoritmos de Porter (PORTER, 1980) e Lovins (LOVINS, 1968) que removem sufixos da língua inglesa. Em português existe uma implementação bastante utilizada; é a do algoritmo *Portuguese Stemmer*, proposto por Orengo e Huyck (ORENGO e HUYCK, 2001).

Apesar de estes algoritmos serem utilizados com o objetivo de aperfeiçoar o desempenho dos processos de recuperação da informação (RI) e mineração de textos, alguns estudos demonstram que a redução de afixos em uma coleção de documentos não apresenta uma melhora significativa no desempenho da tarefa de mineração aplicada (HARMAN, 1991; LENNON, PIERCE et al., 1981). Apresentam entretanto a vantagem da redução no espaço de representação dos dados. Apesar de contribuir de forma relevante para a redução do tamanho do dicionário, seu uso deve ser

criteriosamente avaliado, uma vez que acarreta um alto custo computacional. Portanto, o uso de stems é indicado somente se os benefícios obtidos forem realmente relevantes. Assim, não se utilizou os algoritmos de *stemming* nesse trabalho.

3.4.2.3 – N-grams

Os *n-grams* podem ser definidos como o conjunto de n palavras que ocorrem em seqüência um número significativo de vezes, e são utilizadas como um termo único (BEKKERMAN e ALLAN, 2003a; CAVNAR, 1994).

Funcionam no sentido inverso do *stemming* pois visam agregar significados . Assim, essa seqüência de palavras passa a ser identificada como um termo único que, por conseguinte, deve conter um alto valor preditivo. Exemplificando, a utilização do termo 2-gram, "*text mining*" tem maior valor preditivo do que as palavras isoladas "*text*" e "*mining*"; assim como o 3-gram "*world wide web*" tem maior valor preditivo do que as palavras isoladas "*world*","*wide*" e "*web*" (CASTRO, 2000a).

3.4.3 – Representação de Documentos ou Seleção dos Dados

Nessa etapa são identificados e selecionados os termos mais relevantes nos dados pré-processados. Um dos principais problemas na categorização de documentos é selecionar as palavras ou termos, que sejam realmente representativas da classe (BLUM e LANGLEY, 1997; FAGNI e SEBASTIANI, 2007). Assim, o objetivo dessa etapa é identificar os termos que melhor descrevem o conteúdo de um documento.

3.4.3.1 – Termos

Para a Teoria do Conceito, "termo" é a menor unidade de representação de um conceito (DAHLBERG, 1992), e é constituído por uma palavra ou por um grupo de palavras, não podendo ser divisível na indexação (CABRÉ, BAGOT et al., 2001). Neste trabalho, definimos "termo" como sendo toda seqüência de caracteres alfabéticos entre dois espaços em branco ou sinais de pontuação. Assim, a identificação de termos consiste na separação de todos os termos do texto, individualmente.

3.4.3.2 – Modelos de Representação de Documentos

Existem diversos modelos que podem ser utilizados para representar documentos. A qualidade do texto afeta a escolha do modelo de representação de documentos. Essa escolha, por sua vez, costuma apresentar um grande impacto nos resultados obtidos (KEIKHA.M., RAZAVIAN et al., 2008). As abordagens mais

comuns são o uso de termos e o uso de *n-grams*, descritos em 3.4.2 – A principal desvantagem desses dois tipos de representação é a perda da semântica do documento. Esse problema é abordado em pesquisas de técnicas como a Indexação Semântica Latente - *Latent Semantic Indexing* (LSI) (BERRY, 2003; BERRY et al., 1995) e a Análise Semântica Latente - *Latent Semantic Analysis* (LSA) (DEERWESTER, DUMAIS et al., 1990). Outras duas representações conhecidas são as baseadas em frases e em representação lógica. Existem ainda outros modelos como o Difuso, o Probabilístico (ROBERTSON e JONES, 1976) etc.

Entretanto, a utilização da lista de termos, devido à sua simplicidade é uma consagrada representação e foi utilizada neste trabalho. Dentre suas variantes podemos citar:

i) Modelo Booleano – O modelo booleano (SALTON e BUCKLEY, 1988) percebe a presença ou ausência do termo de indexação no documento, atribuindo de forma binária ("0 ou 1"), caracterizando a presença ou ausência do termo no documento, não possuindo pesos associados aos mesmos. O modelo booleano avalia a presença ou ausência do termo de indexação no documento; portanto, os pesos atribuídos a esses termos são binários (TOMAULT, 2006). Sendo cada categoria representada por um ou mais vetores binários, correspondentes à disjunção de conjunções de termos encontrados nos documentos nela contidos, não há um casamento parcial entre o documento e a categoria, mas apenas uma decisão binária se o documento pertence ou não a esta. A decisão de quantos vetores booleanos serão utilizados para representar uma categoria depende da capacidade em se especificar o conteúdo dos documentos a ela pertencentes apenas por uma conjunção de termos ou por uma disjunção de conjunções de termos.

Os documentos são representados como vetores binários de tamanho n. Cada posição corresponde a um termo usado na indexação dos documentos sendo considerados para a consulta termos conectados por AND, OR e NOT.

Esse modelo é simples de implementar e usar, exigindo menos espaço para armazenamento. Porém é considerado de baixo desempenho por não oferecer uma ordem de relevância dos documentos dentro de uma categoria.

Existe ainda uma variante desse modelo, que é o Booleano Estendido, que visa melhorar o problema das decisões binárias e assemelha-se ao Modelo Espaço Vetorial.

ii) Modelo Espaço Vetorial - O Vector Space Model (VSM) O VSM é uma abordagem típica para a representação de documentos (SALTON, WONG et al., 1975b). Esse modelo advém das técnicas de RI (SEBASTIANI, 2002f) sendo considerado o mais utilizado para a representação de documentos (KELLER, 2006; LIAO e VEMURI, 2002d). Sua popularidade deve-se à sua simplicidade e ao tratamento da proximidade semântica como proximidade espacial. No VSM cada lista de termos é considerada como um vetor de espaço n-dimensional, onde n é o número de distintos termos (RUSSEL; NORVIG, 1995). O conjunto de vetores forma a matriz termo–documento, armazenada, por exemplo, como uma estrutura de índice invertido.

No VSM, os documentos são modelados como Sacos-de-Palavras ou *Bag-Of-Words* (BOW) (BLOEHDORN e HOTHO, 2004). Esse modelo é baseado na freqüência das palavras acrescida de várias normalizações (CANCEDDA, GAUSSIER et al., 2003; LI e JAIN, 1998). Uma variação do mesmo seria a abordagem "*Bag of Bigrams*" (BOB), que considera pares de palavras (LEWIS, 1992).

No VSM, cada texto é representado por um vetor de *BOW*, sendo que cada palavra ou termo está associado a um valor numérico. O valor numérico é fornecido por uma medida geralmente associada à freqüência de ocorrência da palavra ou termos nos textos. Essas palavras do texto são mantidas sem qualquer ordem pré-estabelecida. O tamanho do vetor é o número de termos que ocorre pelo menos uma vez no conjunto de treinamento. Os pesos podem ser binários, onde representam a existência do termo no documento ou não, ou podem ser não binários, onde eles representam o quanto eles contribuem para a semântica do documento (SEBASTIANI, 2002e).

A similaridade entre esses vetores pode ser computada usando a distância euclidiana, a medida do cosseno, o coeficiente de Jaccard ou variantes.

Uma das desvantagens apontadas em sua utilização é o grande espaço gasto para representar um documento e a não relevância dada à relação semântica entre os termos existentes (YUA, XUB et al., 2008).

3.4.4 – Seleção de características

Segundo Mahoney (MAHONEY,2007) a seleção de características é a tarefa de selecionar um subconjunto de características capaz de descrever as classes antes que seja aplicado algum algoritmo de aprendizado. Para esta etapa, é possível adotar uma abordagem semântica e/ou sintática, conhecida por abordagem lingüística, que visa representar o significado do conteúdo do documento através do Processamento de

Linguagem Natural. Pode-se usar também uma abordagem estatística através de cálculos matemáticos aplicados aos termos do documento ou da coleção.

Depois de concluído o processo de preparação do texto, a tarefa é definir o conjunto de termos que melhor representem o assunto a ser categorizado, ocorrendo assim a redução de dimensionalidade. Esta etapa é chamada de seleção de características ou indexação automática (SCHUTZE; HULL; PEDERSEN, 1995). A redução pode ser realizada através da seleção daquelas que melhor representam um documento, de acordo com algum critério. Esta é uma tarefa de certa complexidade, pois se deve selecionar um subconjunto de termos que possa fazer uma discriminação adequada entre as várias categorias, e ao mesmo tempo, ser pequeno o suficiente para que possa ser utilizado pelo classificador. Assim, os métodos automáticos de seleção de características incluem a remoção de termos não informativos e a construção de um conjunto de elementos, termos ou radicais, que representam os textos e facilitam a identificação da categoria a que pertence. É interessante que estes elementos combinem alta representatividade em menor número possível de características.

Outro cuidado que deve ser observado é que dicionários muito grandes tendem a gerar *overfitting*, apresentando baixo desempenho para classificar textos desconhecidos. Dicionários menores, além de evitarem o *overfitting*, reduzem o tempo de processamento dos algoritmos, fato que pode ser observado nos experimentos realizados. *Overfitting* é o fenômeno onde o categorizador gerado é especializado nos dados de treinamento (AHA, KIBLER et al., 1991b), o que pode causar baixo desempenho durante a classificação de novos casos.

Existem diversas medidas de seleção de características que têm sido largamente utilizadas em problemas de categorização de textos (FORMAN, 2002). Dentre elas podemos citar: freqüência de palavras (NG et al., 1997) ou freqüência do termo, o ganho de informação (medida de entropia) (YANG e PEDERSEN, 1997), o coeficiente de correlação (CC) (NG et al., 1997), a técnica do qui-quadrado (NG et al., 1997) e o método de Escore de Relevância (ER) (WIENER et al.,1995). Temos ainda a freqüência absoluta, a freqüência do documento (FD), freqüência inversa de documentos, a freqüência relativa que é a freqüência do termo dividida pelo número de palavras do documento, a separação Bi-Norm e Razão Ímpar (MLADENIC, 1998). Estas são consideradas algumas das medidas mais efetivas. A TFIDF que é a razão ente freqüência do termo e freqüência do documento (SALTON e BUCKLEY, 1988), a CC e a ER serão detalhadas nas subseções seguintes.

3.4.4.1 – TF-IDF

Com base nos cálculos de freqüência dos termos e freqüência dos documentos é possível obter a freqüência do termo - freqüência inversa de documentos, mais conhecida como tf-idf. Através desta medida pode-se aumentar a importância de termos que aparecem em poucos documentos e diminuir a importância de termos que aparecem em muitos documentos. Segundo SALTON (SALTON e BUCKLEY, 1988)os termos de baixa freqüência de documentos, são, em geral, mais discriminantes.

TF-IDF é uma das técnicas mais bem testadas da área de recuperação de informação. Um documento é representado como um vetor de termos e pesos. A computação destes pesos reflete observações empíricas efetuadas sobre os textos. Termos que aparecem frequentemente em um determinado documento (TF = "term frequency" = freqüência do termo), mas raramente nos outros documentos da coleção (IDF = "inverse document frequency" = inverso da freqüência nos documentos), se mostram mais relevantes para a categoria deste documento. O peso TF-IDF de um termo em um documento é o produto de sua freqüência no documento (TF) e o inverso de sua freqüência nos documentos (IDF). Em adição, para prevenir que documentos longos tenham melhor chance de serem recuperados, os vetores de termos são normalizados para alguma unidade padrão, por exemplo, pela divisão da freqüência do termo pelo número de termos do documento. Para a tarefa de recuperação de informação, documentos e as consultas feitas pelo usuário são representados como vetores e submetidos a algum tipo de modelo para a recuperação da informação desejada.

Uma das maneiras mais utilizadas para se identificar o peso do termo é a aplicação da fórmula preconizada dada por:

$$TFIDF(\omega) = TF(\omega) \cdot \log\left(\frac{|D|}{DF(\omega)}\right)$$

Onde : $TF(\omega)$: freqüência da palavra ω no documento

$DF(\omega)$: freqüência de ω em D

D = total de documentos

3.4.4.2 – Coeficiente de Correlação

O coeficiente de correlação foi desenvolvido por Hwee Ng et al. (NG, GOH et al., 1997) para indicar o grau de correlação entre uma palavra e um documento. Para tanto, esse coeficiente leva em conta a quantidade total de documentos de uma coleção,

a quantidade de documentos em que a palavra aparece e a quantidade de documentos em que ela não aparece.

Segundo Cutting (CUTTING et al., 1992), o coeficiente de correlação é derivado do coeficiente $\chi 2$ de Schutze (SCHUTZE, HULL et al., 1995) e corresponde à raiz quadrada do valor obtido por essa métrica. O Coeficiente de Correlação é maior para as palavras que indicam a pertinência de um documento à categoria Cj, enquanto que o $\chi 2$ gera valores maiores não só para este conjunto de palavras mas também para aquelas palavras que indicam a não pertinência à Cj.

A quantidade de documentos em que a palavra aparece é restringida a documentos que pertençam a determinada categoria ou conglomerado, identificando, assim, as palavras mais exclusivas e representativas para eles.

É definido que Coeficiente de Correlação entre o termo t e a classe c é dado por:

$$C = \frac{(N_{r+} \cdot N_{n-} - N_{r-} \cdot N_{n+}) \cdot \sqrt{N}}{\sqrt{(N_{r+} + N_{r-}) \cdot (N_{n+} + N_{n-}) \cdot (N_{r+} + N_{n+}) \cdot (N_{r-} + N_{n-})}}$$

Onde:

N_{r+} = documentos relevantes para Cj que contêm o termo t

N_{r-} = documentos relevantes para Cj que não contêm t

N_{n+} = documentos não relevantes para Cj que contêm t

N_{n-} = documentos não relevantes para Cj que não contêm t

Para representação dos conjuntos de documentos de cada classe são elaborados vetores locais compostos dos n termos mais relevantes de acordo com o cálculo de relevância. Assim, os vetores globais que representarão os documentos serão a junção dos vetores locais de cada classe e servirão de índices para os vetores de cada exemplo e as posições correspondentes representarão a importância do termo no documento.

3.4.4.3 – Escore de Relevância

O Escore de Relevância foi criado por Wiener et Al. (WIENER, PEDERSEN et al., 1995) com base no peso de relevância (relevancy weight) de Salton e Buckley (apud WIENER; PEDERSEN; WEIGEND, 1995). Seu objetivo é o de construir protótipos capazes de representar categorias de documentos em sistemas de categorização de textos.

Pode-se dizer que sua técnica iniciou-se com o trabalho de SALTON (SALTON, WONG et al., 1975a) que propôs a indexação de textos feita a partir de termos com pesos associados. Naquele trabalho, a indexação alcançou melhores resultados por estar definindo, a partir do peso, o grau de importância que o termo tem dentro do texto. A idéia inicial daquele estudo foi calcular a freqüência com que cada termo aparece no documento. Depois, foi determinada a freqüência do termo dentro do texto e dentro da coleção, dando origem à técnica de freqüência inversa de documentos, salientando a idéia de que termos com grande capacidade de representação de conteúdo devem possuir alta freqüência no documento e baixa freqüência na coleção. E com isso, ele definiu a técnica do cálculo do peso de relevância do termo. A partir desses resultados, foi proposto o ER (WIENER, PEDERSEN et al., 1995) para a categorização de textos. A idéia desta técnica está baseada na freqüência de um termo em uma categoria e na sua freqüência nas demais categorias.

A fórmula para o cálculo do escore de relevância é dada por:

$$r_k = \log \frac{w_{tk} / d_t + 1/6}{w_{\bar{t}k} / d_{\bar{t}} + 1/6}$$

Onde:

r_t = escore de relevância do termo k

w_{tk} = o número de documentos pertencentes a uma dada categoria t que contém o termo k em processo de análise;

$w_{\bar{t}k}$ = o número de documentos de outras categorias que contém o termo k em análise

d_t = o número total de documentos de outras categorias

$d_{\bar{t}}$ = o número total de documentos da categoria t

1/6 = uma constante para evitar divisão por zero

Com o ER, as palavras exclusivas dos documentos pertencentes a uma categoria recebem pontos positivos e as palavras de documentos de outras categorias recebem pontos negativos. O processo é repetido para todas as categorias existentes. Ao final do processo, palavras que aparecem em muitas categorias acabam tendo valores muito baixos, por serem pouco discriminantes. Já as palavras que aparecem em poucas

categorias ficam com valores mais altos, indicando que são as mais adequadas para representar essas categorias.

3.4.5 – Métodos de Categorização

Nesta fase, são utilizados métodos que identificam os conceitos no texto e fazem efetivamente a categorização. Esses métodos podem categorizar os documentos em nenhuma, uma ou mais categorias existentes. Quando um método efetua a categorização de textos em apenas uma categoria, diz-se que este método é de classificação binária (single-label)(SEBASTIANI, 2002f). Quando os textos podem ser classificados em mais de uma categoria, diz-se que foi aplicado um método de categorização múltipla (multi-label), podendo-se inclusive definir o grau de pertinência do documento a cada uma das categorias para as quais ele foi classificado (LEWIS, 1998).

A categorização multivalorada, onde um mesmo documento pode ser classificado em mais de uma categoria, não faz parte do nosso escopo por motivos óbvios. O mesmo documento não pode receber simultaneamente duas atribuições de sigilo diferentes ao mesmo tempo.

Existem vários métodos de categorização, mas basicamente o processo é comparar a lista de termos do texto e da categoria, definir se eles são semelhantes, e a partir deste resultado decidir se o texto pertence à categoria.

Em CADT a escolha do algoritmo apropriado por si só já costuma oferecer bons resultados (TEICHERT e MITTERMAYER, 2002). Existem diversos métodos de categorização, e para cada um deles diversos algoritmos que o implementam. Um categorizador pode ter um ou mais algoritmos utilizados em sua implementação. As técnicas de aprendizado freqüentemente utilizadas em categorização automática de textos somam um legado de técnicas (CAMARGO, 2007b; SEBASTIANI, 2002b; SEBASTIANI, 2006)

Serão descritos a seguir os algoritmos mais conhecidos e com desempenho representativo nas publicações. É dada ênfase aos algoritmos utilizados nos experimentos deste trabalho, quais sejam, o K vizinhos próximos (kNN), o Naive Bayes e Máquinas de Suporte a Vetor (SVM).

3.4.6 – Categorizadores baseados em árvores de decisão

As árvores de decisão consistem de nodos que representam os atributos, de arcos provenientes destes nodos e que recebem os valores possíveis para estes atributos, e de nodos folha, que representam as diferentes classes de um conjunto de treinamento(QUINLAN, 2003). São muito utilizadas na construção de categorizadores por serem consideradas formas simples de representar o conhecimento. Quando utilizadas por categorizadores, as classes são identificadas baseadas nos valores de atributos de um conjunto de dados. Alguns dos algoritmos mais utilizados são o ID3 e o C.4.5.

3.4.7 – Categorizadores baseados em Inteligência Artificial

São categorizadores considerados de aprendizado supervisionado e que utilizam técnicas consagradas em Inteligência Artificial como Redes Neurais e Algoritmos Genéticos, sozinhos ou combinados a fim de obter-se os resultados. Na utilização de algoritmos de Redes Neurais, são caracterizadas por se apresentarem como uma forma de computação que são semelhantes, em algum nível, à estrutura do cérebro humano (MITCHELL, 1997; ZHANG, BERG et al., 2006). Já na utilização de algoritmos genéticos (GOLDBERG, 1989) como utilizado em (LOUREIRO, MARGOTO et al., 2005) são automatizadas as buscas por valores de parâmetros que maximizam o desempenho do categorizador gerado. Podemos citar o Clonalg (DE CASTRO e VON ZUBEN, 2002) e o Partical Swarm Optimization (PSO) que é baseado no comportamento social do cardume de peixes ou de aves, como um dos muitos que vêm despontando na utilização em categorização de textos (YU, WANG et al., 2007).

3.4.8 – Categorizadores baseados em modelos probabilísticos

Os categorizadores que utilizam modelos probabilísticos baseiam-se no princípio do ordenamento de probabilidade (ROBERTSON, 1997). Estes métodos seguem a abordagem de ordenar os documentos baseados na probabilidade da relevância com relação à consulta, baseados na distribuição estatística dos termos nos textos. Dentre alguns categorizadores probabilísticos existe o Modelos de Markov Escondidos – Hidden Markov Models (HMM) (RABINER, 1989) que utiliza o processo de Markov com parâmetros desconhecidos. Entretanto, um dos categorizadores mais utilizados é o Naive-Bayes descrito a seguir.

3.4.8.1 – Naive-Bayes

Alguns autores consideram este modelo uma simplificação dinâmica de uma rede Bayseana (YANG e WEBB, 2002). Pode ser usado para determinar a probabilidade de um documento, utilizando valores independentes de entrada e saída (DUDA, HART et al., 2002; DUDA e HART, 1973). Esse categorizador combina a junção das probabilidades de palavras e categorias para estimar a probabilidade de um novo documento pertencer a uma categoria. Desta maneira, o algoritmo calcula a probabilidade de um documento pertencer a classes diferentes e o atribui à classe cuja probabilidade de pertencer é mais alta.

A parte ingênua (naïve) do algoritmo NB é a suposição de independência das características da palavra, ou seja, é assumido que o efeito das características de uma palavra cuja probabilidade condicional está associada a uma categoria é independente das características de outras palavras daquela categoria. Essa abordagem é efetuada a partir do teorema bayes, onde é calculada a probabilidade de diferentes hipóteses à medida que novas evidências são observadas.

$$P(h/D) = \frac{P(D/h)P(h)}{P(D)}$$

Onde:

P(h): probabilidade a priori de h

P(D): probabilidade a priori de D

P(D/h): probabilidade de observar D dado que h aconteceu

O Naive Bayes baseia-se na suposição de independência condicional entre os atributos dada a classe (DUDA, HART et al., 2002) cuja fórmula é dada por:

$$P'(a_i / v_j) \leftarrow \frac{n_c + m \times p}{n + m}$$

Onde:

n : é o número de exemplos para os quais v = vj

n_c: número de exemplos para os quais v = vj e a = ai

p: é a estimativa a priori para P'(ai/vj)

m: é o peso dado a priori (número de exemplos "virtuais")

Entretanto, sabe-se que no mundo real a indepedência condicional quase não ocorre. Assim, embora o algoritmo Naive Bayes tenha se mostrado competitivo quando comparado a outros algoritmos mais complexos para problema de categorização, existem pesquisas que tentam melhorar seus resultados através do relaxamento dessa independência (FRIEDMAN, GEIGER et al., 1997).

A principal vantagem desse classificador é a simplicidade. Entretanto, um dos problemas do Naive Bayes é a assunção de que os atributos são independentes dada a classe, hipótese que nem sempre ocorre no mundo real. Embora apresente um bom desempenho quando o número de exemplos é pequeno, não costuma apresentar melhoras significativas quando o número de exemplos de treino aumenta. O desafio do Naive Bayes é determinar estes parâmetros escondidos a partir de parâmetros conhecidos.

3.4.9 – Categorizadores baseados em instâncias

Categorizadores baseados em instâncias ou ingênuos, recebem esse nome, porque elaboram hipóteses diretamente a partir das próprias instâncias de treinamento. Assim, diferentemente de outros categorizadores, estes não criam um modelo de aprendizagem (ANTONIE e ZAÏANE, 2002a; VAN RIJSBERGEN, 1979). Os dados de treinamento são sempre utilizados para verificar quais são os documentos mais similares ao novo documento (ZHANG, BERG et al., 2006). Ou seja, dado um novo documento, os dados de treinamento são verificados e o mesmo é categorizado por analogia. Uma qualidade de tais categorizadores é que suportam aprendizado incremental, ou seja o número de parâmetros cresce com o conjunto de treinamento.

Exemplos deste tipo de categorizador são o Case-based Reasoning - Raciocínio Baseado em Casos (CBR) (CUNNINGHAM, DOYLE et al., 2003; ZHANG, BERG et al., 2006) e o k-vizinhos próximos (kNN), que é descrito na próxima subseção.

3.4.9.1 – Algoritmo kNN

O algoritmo do k-vizinhos próximos ou kNN, proposto em 1967, é um algoritmo de classificação baseada em similaridade (COVER e HART, 1967; COVER e HART.P.E., 2008). É considerado um dos algoritmos mais populares na categorização de textos (MANNING e SCHTZE, 1999) e continua sendo muito utilizado (BAOLI, SHIWEN et al., 2003; LIU e LIANG, 2008).

Diferentemente de algoritmos como o SVM e Nayve Bayes, o kNN não possui fase de aprendizado (MITCHELL, 1997; ZHANG, BERG et al., 2006). Assim, este

algoritmo simplesmente usa a indexação, construindo um índice invertido para executar a categorização (YANG, ZHANG et al., 2003).

O kNN tem sido largamente usado para resolver problemas de reconhecimento de padrão e classificação de documentos. Assim como o conhecido algoritmo de recuperação Rocchio, também é baseado em similaridade e utiliza o método chamado de Realimentação de Relevantes (ROCCHIO, 1966).

O kNN é um classificador que possui apenas um parâmetro livre (o número de K-vizinhos) que é controlado pelo usuário com o objetivo de obter uma melhor classificação. Para cada documento desconhecido, é verificada a similaridade entre ele e todos os documentos da base, através de uma medida de distância. Para determinar a classe de um elemento que não pertença ao conjunto de treinamento, o classificador kNN procura K elementos (chamados de k-vizinhos próximos) do conjunto de treinamento que estejam mais próximos deste elemento desconhecido, ou seja, que tenham a menor distância.

Verifica-se quais são as classes desses K vizinhos e a classe mais freqüente será atribuída à classe do elemento desconhecido. A classificação é efetuada através da atribuição ao documento desconhecido à classe que for predominante, e é verificado a quais classes pertencem os k documentos mais próximos. Onde k é igual ao número de vizinhos. O kNN também utiliza o modelo convencional do espaço vetorial para representar cada documento como um vetor de termos. Uma das medidas de distância utilizada é a Euclidiana, mas podem ser usadas outras como a Manhatan e etc. (WILSON e MARTINEZ, 1997).

No cálculo da distância é medida a disparidade entre dois elementos de forma a poder identificar quais são os kNN. Existem duas regras de classificação básicas: maioria na votação e peso pela distância. Na primeira, cada elemento tem uma influência igual; a classe escolhida é aquela que possuir mais representantes entre o kNN. Já no peso pela distância, cada k-vizinho tem um peso inversamente proporcional à sua distância.

Além de categorização de textos o kNN é usado também em diversas outras aplicações como, por exemplo, para a detecção de intrusão (LIAO e VEMURI, 2002e).

3.4.10 – Máquina de Suporte a Vetor (SVM)

A máquina de Suporte a Vetor (SVM) foi introduzida na Categorização de Textos por Joachims (JOACHIMS, 1998b) e atualmente é uma das técnicas mais

populares (NÆSS, 2007). A principal característica das SVM's é a determinação automática dos dados de treinamento mais relevantes para o problema abordado, chamados vetores de suporte (SILVA, MONTILHA et al., 2007a).

O SVM aplica uma técnica chamada de Maximização Estrutural de Risco, que visa minimizar a generalização do erro ao invés de utilizar erros empíricos no treinamento dos dados sozinhos (VAPNIK, 1999).

Um algorimo básico SVM deve encontrar uma superfície de decisão que "melhor" separe os dados em duas classes, buscando a maximização da margem existente entre as fronteiras das nuvens de pontos formados por estas classes. No caso da categorização de textos, cada documento é representado por um vetor de termos, em que cada termo corresponde a uma característica (RUSSELL e NORVIG, 2003).

Entretanto, existem basicamente dois tipos de implementação, as lineares, descritas um pouco mais à frente, e as não- lineares ou baseadas em Funções do Kernel, que inserem novas dimensões (MULLER, MIKA et al., 2001). O tipo de implementação linear do SVM é mais básico. Os classificadores lineares separam os dados em duas classes através de um hiperplano de separação.

Segundo HAYKIN (HAYKIN, 2000), uma máquina de vetor suporte depende de duas operações matemáticas:

- O mapeamento não-linear de um vetor de entrada para um espaço de características de alta dimensionalidade, que é oculto da entrada e da saída;
- A construção de um hiperplano ótimo para separar as características descobertas no passo anterior. O hiperplano ótimo de separação das classes é determinado em função das amostras que se encontram entre as margens ($\alpha_i = C$) e nas margens ($0 < \alpha_i < C$). Amostras externas às margens ($\alpha_i = 0$) são ignoradas.

Para o caso de dados linearmente separáveis, o algoritmo simplesmente procura pelo hiperplano de separação com a maior margem. A margem é a separação entre os exemplos positivos e negativos. Ainda nesse caso, pode-se construir um hiperplano que separa os exemplos positivos dos negativos. A função de decisão usada nesses casos é $f(x) = sign((w, x) + b)$ na qual w representa o vetor de pesos da rede, x um vetor de características e b, o bias (VALENTINI e DIETTERICH, 2002).

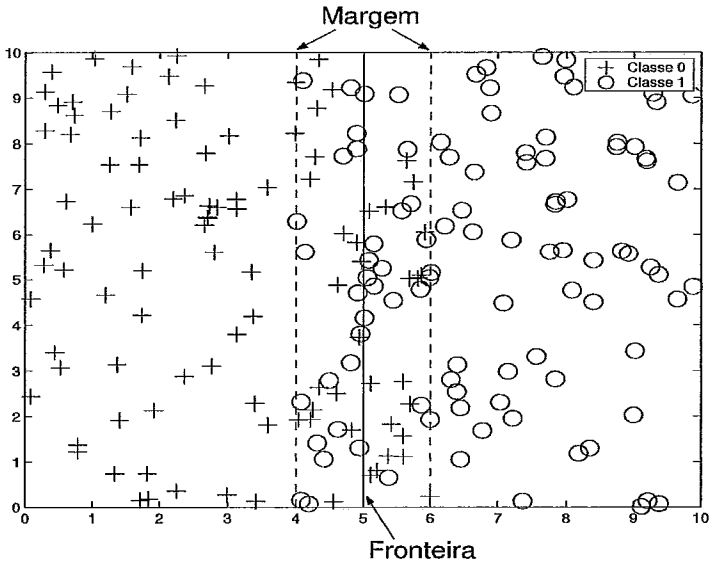


Figura 4 hiperplano não linearmente separável

Os pontos x que pertencem ao hiperplano satisfazem a equação wx+b=0, onde w é normal ao hiperplano e |b|/||w|| é a distância perpendicular do hiperplano até a origem e ||w|| é a norma euclidiana de w. Quando um determinado dado viola a condição yi(xi.w + b) −1 ³ 0 , diz-se que a margem de separação entre as classes é suave. Nesse caso, o dado pode estar dentro da região de separação, no lado correto ou incorreto da superfície de decisão.

Dentre diversas implementações do SVM, podemos citar o SVM-light (JOACHIMS, 1998a), o Least Square-SVM (YANG, ZHANG et al., 2003) e o LIBSVM (CHANG e LIN, 2000). Uma outra implementação muito utilizada é a SMO (Sequential Minimal Optimization). Com a utilização desse algoritmo, a utilização de memória é linear para realizar os treinamentos. Com isso, o SMO permite lidar com grande quantidade de arquivos para treinamento (PLATT, 1999).

3.4.11 – Teorema Não Existe Almoço Grátis (No Free Lunch Theorem)

A idéia do teorema No Free Lunch (NFL) é a de que não existe o melhor algoritmo, e sim o algoritmo que é mais eficiente para um determinado tipo de problema (WOLPERT e MACREADY, 1997). Em outras palavras, segundo essa teoria, não existe algoritmo para a resolução de todos os problemas, seja de otimização, seja de negócios (REINHARDT, 1999) ou até em categorização de textos.

Segundo a teoria do NFL, a afirmação de que um categorizador é inferior ou superior a algum outro é passível de equívoco. O que pode ser afirmado é que algum algoritmo comporta-se melhor que outros com respeito à resolução de uma classe específica de problemas, e como conseqüência comportam-se inadequadamente para outras classes de problemas.

3.5 – Pós-Processamento - Avaliação e interpretação dos resultados

Nesta etapa os resultados são avaliados em função das informações relevantes ou irrelevantes obtidas e interpretados e é verificado se o objetivo proposto foi alcançado. Os dados obtidos são analisados, filtrados e selecionados para que o utilizador os possa manipular. São apresentados os dados obtidos pelos algoritmos, utilizando-se métricas de avaliação originárias da área de RI e baseadas na idéia de relevância, ou seja, se um documento atender a necessidade de informação do usuário, ele é considerado relevante à solicitação do mesmo.

Embora não seja possível evitar o retorno de dados indesejados, podemos avaliar se o retorno da busca foi satisfatório através de diversas medidas de desempenhe, através de métricas como "precisão", "cobertura" e "medida-F" podem ser obtidas através das seguintes relações, discutidas em (VAN RIJSBERGEN, 1979) e melhor explicadas a seguir.

3.5.1 – Medidas de Desempenho

A matriz de confusão (KOHAVI e PROVOST, 1998) é uma das formas de exibição das medições de um categorizador, contendo informações sobre a categorização real e a prevista pelo mesmo. Conforme exemplificado na Figura 5, as células da matriz de confusão contabilizam verdadeiros positivos (VP), falso positivo (FP), verdadeiro negativo (VF) e falso negativo (FN) (CASTRO, 2000b). Assim, o número de ocorrências N = VP + FP + VF + FN.

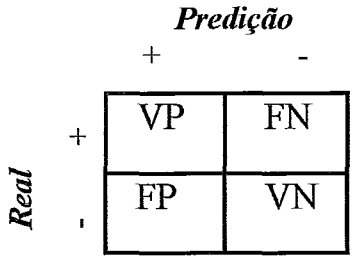


Figura 5 Matriz de Confusão

Utilizando essas células podem-se definir as métricas para decisões binárias como:

i) Abrangência ou Recall: VP / VP + FP. Consiste na avaliação da habilidade do sistema em recuperar os documentos mais relevantes do usuário, medindo a quantidade de itens recuperados, dentre os relevantes na base de dados. A abrangência é dada por:

ii) Precisão ou Cobertura Precision: avalia a habilidade do sistema manter os documentos irrelevantes fora do resultado de uma consulta. A precisão é definida pela razão entre a quantidade de itens recuperados dentre os relevantes e o número total de itens recuperados, isto é:

iii) % de erro = (FP + FN)/N

iii) medida-F ou F-Measure (2*Abrangência*Precisão/(Abrangência+Precisão): combina os valores de precisão e abrangência, de maneira a se obter o desempenho geral do sistema, permitindo um balanceamento entre os dois valores. Essa medida vem sendo cada vez mais utilizada pois a análise separadamente das medidas de precisão e abrangência pode levar a uma avaliação equivocada do sistema, visto que, geralmente, quando se aumenta a precisão, a abrangência do sistema é diminuída. Assim mostra-se um desempenho geral do sistema, permitindo um balanceamento entre os dois valores. A medida-F mede a capacidade de generalização do modelo gerado, permitindo verificar se durante o treinamento este assimilou características significativas do conjunto de treinamento que permitem um bom desempenho em outros conjuntos de dados ou se ocorreu *overfitting*, que é um fenômeno no qual o modelo especializa-se nos dados de treinamento. O valor da medida-f é maximizado quando a precisão ou abrangência são iguais ou muito próximas.

Embora freqüentemente exista um desbalanceamento entre o número de exemplos positivos e o número de exemplos negativos, existe uma tendência aos números verdadeiros negativos dominarem os acertos e erros de um sistema, podendo levar a uma interpretação errônea dos resultados. Por exemplo, quando os exemplos negativos de uma categoria constituem-se de apenas 1% do conjunto de teste, um

classificador trivial que faz predições negativas para todos os documentos, possui um percentual de acerto de 99%, ou de 1% de erro. Entretanto, de modo sistêmico isto não têm grande valia. Assim, abrangência, precisão e medida-F têm sido mais usadas para a avaliação de categorizadores de texto.

Maiores detalhes sobre medidas de desempenho podem ser vistos em (SEBASTIANI, 2002f; YANG e LIU, 1999a).

Existem ainda abordagens diferentes na avaliação do desempenho dos categorizadores. Há quem, para testar o desempenho dos algoritmos, não utilize a clássica tokenização (*stopwords/stemming*), nem seleção de características ou valoração de peso dos termos, pois são considera comuns a todos (YANG, ZHANG et al., 2003).

3.5.2 – Validação Cruzada

A validação cruzada é uma ferramenta padrão para análise e é um importante recurso para ajudar a desenvolver, ajustar e assegurar a validade dos categorizadores. (KOHAVI, 1995). Dentre os métodos de validação cruzada, como a *holdout* e a *Repeated random sub-sampling,* destacamos o k-*fold* que consiste em dividir o conjunto de dados em k conjuntos mutuamente exclusivos e de tamanhos iguais. Cada subconjunto é fornecido para testes e k-1 subconjuntos para treinamento, repetindo esse "pacote" por k vezes. Por exemplo, em 10-*fold cross-validation,* exige a criação de 10 pares de conjuntos de treinamento e teste.

3.6 – Conclusão

Conforme já citado neste capítulo, existem diversas abordagens para solução de problemas em CADT. Normalmente são consideradas três abordagens para melhoria em resultados com texto. A primeira constitui-se no foco em melhorias da fase de pré-processamento. Na segunda opção, pode-se tentar criar um algoritmo ou aperfeiçoar algum dos diversos existentes.

Na Tabela 1 a seguir apresentamos um resumo das características dos categorizadores utilizados neste trabalho.

Tabela 1 Resumo dos categorizadores utilizados

Categorizador	Principal Característica	Implementação	Desempenho
kNN	Usa uma função de distância, para encontrar o exemplo mais próximo ao do caso a categorizar.	Simples	Bom, porém de execução lenta para conjuntos grandes
Naive Bayes	Assume independência entre os atributos. Usa o teorema de Bayes.	Complexa	Bom para conjunto de amostras pequeno. Não melhora com o aumento das amostras. Treinamento rápido.
SVM	Indica vetores suporte para se detectar uma superfície de separação entre classes	Complexa	Bom desempenho, mas treinamento muito lento.

Mesmo algoritmos já consagrados como kNN e SVM continuam a ser pesquisados em busca de melhorias conforme visto em (TOMAULT, 2006; WANG e CHIANG, 2007a; WANG e CHIANG, 2007b) e no contínuo aprimoramento de implementações como a do LibSVM (CHANG e LIN, 2000). Como terceira e última opção pode-se "atacar" a fase de pós-processamento, melhorando medidas do grau de compreensibilidade e da surpresa do conhecimento descoberto, por exemplo.

O modelo que é gerado é utilizado para identificar a categoria de novos objetos. Apesar de diversos trabalhos na área, devidos aos problemas clássicos de categorização de documentos considera-se que ainda há um grande número de pesquisa a ser desenvolvida na área treinamento.

Capítulo 4 – Proposta de CADT na Atribuição de Sigilo

A atribuição excessiva de sigilo é um problema encontrado em diversas organizações do setor governo de diversos países, que incorrem em acúmulo de documentos sigilosos, os quais recebem tratamento diferenciado e dispendioso. Categorizações manuais são lentas e estão sujeitas a falhas. Assim, identificamos a utilização da Categorização Automática de Documentos como forma de reduzir os erros cometidos na determinação do grau de sigilo.

As aplicações para categorização de documentos envolvem algumas questões já abordadas nos capítulos anteriores. A quantidade de textos que devem ser usados para o treinamento, quais características são mais apropriadas para os diversos tipos de textos, qual categorizador é o mais apropriado e quando os categorizadores devem ser re-treinados, são questões que devem ser respondidas durante o processo de tratamento dos documentos. A quantidade de variações que podem ser propostas é gigantesca. Desse modo, não existem regras fixas que definam os melhores parâmetros para a obtenção de bons resultados.

Como ocorre na maioria dos trabalhos em categorização, a solução proposta deve ser capaz de identificar a classe de textos desconhecidos, a partir do conhecimento obtido de textos previamente classificados. A idéia é evitar que sejam efetuadas atribuições de sigilo mais elevadas do que o necessário. Porém, essas atribuições não poderão ser subestimadas. Devemos proporcionar a um novo documento a atribuição mais adequada possível.

O objetivo principal é investigar e testar métodos de categorização existentes e adaptá-los com o propósito de serem utilizados para facilitar a atribuição de sigilo às mensagens. De posse de um bom método, podemos desenvolver um protótipo de sugestão de atribuição de sigilo.

Desta forma, neste trabalho, a categorização do texto é efetuada automaticamente quanto ao sigilo, em função de uma base de documentos similares já existentes no setor operativo da MB. Como dá-se o nome de corpus ao conjunto de dados reais criteriosamente coletados (SARDINHA, 2004), passaremos a chamar essa coleção de Corpus Mensagens.

4.1 – Sobre o Corpus Mensagens

O Corpus Mensagens utilizado neste trabalho é uma base de dados real, do ano de 2002, formada por um tipo de documento chamado mensagem, que é um documento específico utilizado pela Marinha do Brasil. Os referidos documentos costumam ser usados para transmitir ordens, consultas, informações e respostas, de modo bem sucinto. Entretanto, não chegam a ser de modo telegráfico. O histórico de sua utilização e da formação do documento como utilizado atualmente foi descrito na seção 2.5.2 – desta dissertação.

As mensagens são documentos curtos, normalmente com tamanho entre 1 e 6 Kbyte, do tipo texto, sendo que 93,6% dos documentos tem até 4kbytes. Os textos são pequenos, entretanto, não chegam a ser de modo telegráfico. Também são utilizadas algumas abreviaturas, que têm caráter técnico específico, além de um conjunto padrão de trigramas, que não devem ser confundidos com os *n-grams* descritos na seção 3.4.2.3 –. Exemplificando, os trigramas, que no contexto dos documentos mensagens são abreviações de palavras, temos: INF (informar), CNS (consulto), SOL (solicito), etc.

No Corpus Mensagens, existem documentos sigilosos de graus reservado e confidencial, e textos não sigilosos que são chamados ostensivos. Para evitar comprometimento dos dados as mensagens foram separadas em classes, correspondentes aos graus de sigilo. Os textos foram dispostos em três classes, que passaremos a chamar de Classe 1, Classe 2 e Classe 3. A Classe 1 é composta de 10490, a Classe 2 de 270 documentos e a Classe 3 por 4992, em um total de 15752 documentos em formato de texto puro (.txt). A maior parte dos documentos pertence à classe 1, seguida pela classe 3. Já a classe 2 é bem reduzida e difícil de contextualizar.

4.2 – Soluções possíveis

O objetivo desta seção é mostrar a linha de raciocínio que culminou na solução escolhida. Essa descrição visa a facilitar a continuidade deste trabalho, podendo vir também a facilitar futuras implementações. Assim, nas novas pesquisas, alguns caminhos já percorridos poderão ser descartados e outros aprofundados.

Inicialmente é feita uma breve descrição das possíveis Linhas de Ação (LA) que poderiam ser adotadas na resolução do problema e uma breve análise de cada uma. Posteriormente a solução escolhida é detalhada, apresentando suas vantagens e desvantagens e abordando o porquê de sua escolha.

Possíveis Linhas de Ação (LA) para resolver o problema:

i) LA-01 – Adaptação do algoritmo kNN

Uma hipótese a ser utilizada no tratamento desse tipo de problema seria a de otimizar um algoritmo já consagrado, como o kNN por exemplo, assim como em alguns outros trabalhos (POLAT e GÜNES, 2006a; TOMAULT, 2006). Para obter-se algum tipo de melhoria no kNN normalmente usa-se a regra de categorização, que cuida da relevância de cada um dos k elementos ou a função que calcula a distância entre duas instâncias. Essa última, entretanto, é a mais usada pois, segundo a literatura, é a que obtém melhores resultados (WANG, 2006; YAMADA, YAMASHITA et al., 2006a). Como já existem diversos trabalhos na área, essa linha de ação não foi investigada.

ii) LA-02 – Abordagem Estatística x Análise Sintática

Um outro modo de abordar uma possível solução para o problema seria rotular as sentenças gramaticalmente, com a utilização de verbos, adjetivos, substantivos, isolados ou combinados, conforme abordado em trabalhos correlatos (CAMARGO, 2007a; MELO, 2007b). Nesses trabalhos, os melhores resultados foram obtidos com substantivos. Em Silva (SILVA e VIEIRA, 2007b) a seleção de características baseada em informações lingüísticas também apresentou melhores resultados com substantivos acrescidos de nomes próprios. Já Aizawa (AIZAWA, 2001) declara obter bons resultados com a utilização de um método para incorporar a técnica de PLN no processo de categorização de textos, utilizando-se do modelo probabilístico como seleção de termos e uso dos termos compostos. Na abordagem sintática, o SVM costuma apresentar um maior desempenho quando o número de termos for elevado, enquanto que as AD são melhores no aprendizado com um número reduzido de termos (GABRILOVICH e MARKOVITCH, 2007) apresentam uma discussão detalhada sobre essas diferenças.

Ainda dentro da abordagem de Análise Sintática foi avaliada a possibilidade de utilização da ferramenta *Natural language Toolkit* (NLTK) da Universidade de Princeton (BIRD e LOPER, 2004). No entanto, em português, apresenta fraco desempenho (BICK, 2000; BICK, 2003). Outra ferramenta avaliada para uso foi do analisador sintático PALAVRAS, desenvolvido por Eckhard Bick (BICK, 2000; BICK, 2003) para a língua portuguesa. Ele realiza tarefas como

tokenização, processamento léxico-morfológico, análise sintática e faz parte de um grupo de analisadores sintáticos do *Institute of language and Communication da University of Southern Denmark.* No sítio dessa universidade existe o projeto VISL (Visual Interactive Syntax Learning)(VISL, 2008) onde o envio de documentos pode ser feito via *upload* de arquivo, posteriormente devolvidos com os devidos rótulos gramaticais aplicados pelo PALAVRAS. Entretanto, uma cópia do documento é retida na universidade que provavelmente a utiliza para melhorar a atribuição de rótulos. Devido à restrição em se expor os documentos militares para a universidade, essa opção foi descartada. Como o desenvolvimento de um analisador sintático estava fora do escopo desse trabalho essa LA não foi investigada.

iii) LA-03 – Trabalhar na fase de pré-processamento, com parâmetros e combinação de categorizadores.

Essa LA considera que a abordagem da seleção de características, por exemplo, possui vantagens significativas, tais como sua simplicidade computacional e a interpretabilidade direta do conjunto de características resultante (DASGUPTA, DRINEAS et al., 2007). Entretanto, algumas de suas desvantagens, como não reduzir informação redundante (termos correlacionados) e a exclusão de termos individuais não significativos que poderiam ter grande poder discriminativo em combinação com outros, abrem espaço para pesquisas com técnicas de extração de características. Para tal, foram escolhidas algumas técnicas para a seleção de características. As técnicas foram a Freqüência do Documento (FD), Escore de Relevância (ER) e Coeficiente de Correlação (CC). Com isso, procurou-se abordar uma medida mais simples, a FD, passando por duas consagradas por seus bons desempenhos (ER e CC). Essa escolha foi feita de forma análoga à escolha dos algoritmos para o categorizador, já que foi escolhido o kNN por ser um algoritmo simples, e o Naive Bayes e o SVM por apresentarem ótimos resultados na literatura conforme citado no capítulo 3.

iv) LA-04 – Utilização do Teorema "No Free Lunch"

As possibilidades finalizam-se com esta LA que considera a utilização do teorema "Não Existe Almoço Grátis" ou "No Free Lunch" (NFL) em alguma etapa da categorização de documentos. Segundo essa teoria, não existe um

algoritmo que seja o melhor para todos os problemas, pois é difícil para um categorizador trabalhar com todas as categorias de uma vez (WITTEN e FRANK, 2005a; WOLPERT e MACREADY, 1997). Normalmente esse teorema é aplicável a categorizadores como o kNN e o Naive Bayes que são considerados instáveis em relação ao conjunto de atributos. Ainda, segundo Wolpert (WOLPERT e MACREADY, 1997), para cada ganho em performance numa subclasse de problemas existe outra sub-classe onde a variação é igual, mas de sinal contrário. Assim, utilizamos o teorema para o categorizador reconhecer apenas uma classe de cada vez. Em teoria, isso traria um bom resultado, independentemente do conjunto escolhido.

4.3 – Algumas considerações

Foram analisadas e utilizadas diversas e consagradas técnicas de mineração de dados, recuperação da informação, reconhecimento de padrões e análise estatística. Não foram feitas alterações nos algoritmos dessas técnicas e sim uma concentração de esforços na parte de pré-processamento, que geraram um conjunto de regras para aplicação no Corpus Mensagens. A partir da escolha da LA a ser adotada, qual seja, a LA-03, que é a de melhorarmos a classificação, utilizando-se especificamente da melhoria da seleção de características, foram desenvolvidos testes para achar uma solução que resolvesse o problema. Os procedimentos utilizados e os resultados alcançados são tratados mais à frente.

A categorização de textos, neste trabalho, utiliza a forma mais simples, consistindo na classificação binária, onde é possível classificar os documentos em apenas uma de duas categorias. A mesma mensagem não pode receber simultaneamente duas atribuições de sigilo diferentes.

Os textos usados para validação e testes já estavam classificados. Essa classificação prévia é importante no processo de validação, pois é possível verificar se a classe indicada pelo categorizador é a classe real do texto. Essa informação, ou seja, o grau de sigilo atribuído, obviamente não foi considerada no processo de categorização. Assim, foram retiradas através do processo de remoção de *stopword*, informações atinentes ao grau de sigilo, nos testes desenvolvidos com o algorimo kNN, com o banco de dados MySQL, e nos testes executados na ferramenta Weka para os categorizadores Naive Bayes, SVM e kNN.

Avaliou-se o efeito desse tratamento específico da fase de pré-processamento, mostrando o modelo de tratamento estatístico simples, sem combinações e os resultados após a aplicação do modelo proposto. Sobre os melhores resultados da LA-03 foi aplicada a LA-04, que é o uso do teorema NFL na busca de melhores resultados.

Durante a solução do problema definimos uma nova LA, a LA-05 baseando-se na utilização do kNN sobre os melhores testes executados, buscando-se melhores resultados sobre os já conseguidos com Naive Bayes, SVM e NFL.

Nas subseções seguintes são apresentados os resultados dos testes realizados. para a avaliação da seleção de características com o kNN, Nayve Bayes e SVM. São apresentados gráficos que ilustram os resultados dos modelos, com suas respectivas medidas de precisão, abrangência e medida-F.

4.3.1 – Ferramentas

Como ferramentas de trabalho inicial, foram utilizadas o banco MYSQL versão 5.0, PHP versão 5.2 e Apache versão 2.2.3.

Para processamento dos textos utilizou-se da ferramenta WVTOOL – Word Vector Tool,versão 1.1, envolvida por Michael Wurst (WURST, 2007) na linguagem Java, que realiza a criação de listas de palavras e, através dessas listas, a criação dos vetores numéricos com base nas palavras determinadas. Essa ferramenta foi desenvolvida em módulos independentes (ou classes), possibilitando a inclusão de qualquer outro programa em Java e inclui classes para realização de *stemming* em diversos idiomas, entre eles o português, e classes para tratamento dos arquivos de entrada.

Neste trabalho utilizou-se também a ferramenta Weka, da University of Waikato da Nova Zelândia (WITTEN e FRANK, 2005b), na versão 3.5.6. Essa ferramenta é constituída por uma coleção de algoritmos de AM. Para a executar o SVM utilizamos o Sequential Minimal Optimization (SMO). Já o Nayve Bayes foi executado a partir do categorizador multinomial e o kNN foi usado com o lBK (KIBRIYA, FRANK et al., 2004). O principal formato utilizado para a entrada de dados no software Weka é um arquivo com extensão .arff (Attribute-Relation File Format), gerados pela ferramenta WVTOOL.

4.4 – Procedimentos da Solução – Fase 1

Consideramos como fase 1 o início do processo de conhecimento da base e conseqüente limitação do escopo a ser trabalhado. Iniciamos o processo pelo algoritmo mais simples, o kNN para analisar seus resultados. A partir desses resultados primários, outras técnicas e combinações puderam ser aplicadas e serão descritas nas Fases 2 e 3 das soluções apresentadas.

4.4.1 – kNN

Esse algoritmo é considerado um dos mais simples e foi utilizado para verificar o comportamento do Corpus Mensagens frente a aplicação de algoritmos de categorização de documentos. Como o kNN não utiliza treinamento, fizemos alguns testes e utilizamos toda a base com esse algoritmo.

4.4.2 – kNN- Preparando os dados

Na fase de preparação dos dados foram escolhidos textos aleatoriamente. Foi executada a aplicação de *stoplist* diferentes para a retirada das *stopwords*. A seleção de características nesse primeiro momento não foi executada. A intenção era verificar o comportamento da base ainda que sofrêssemos de *overfitting*.

Foi considerado o uso de 4 situações na etapa de pré-processamento: em um primeiro momento, foram executados os testes para cada k, sem a retirada de *stopwords*. Em seguida, foi utilizada uma *stoplist* composta apenas de conjunções, preposições e verbos de ligação. Na terceira, foram incluídas todos os trigramas navais e os nomes das Organização Militares, neste trabalho designados nomes navais. Por último, acrescentamos à *stoplist* os nomes e sobrenomes de militares da ativa e da reserva.

4.4.3 – kNN – Executando os testes

Na execução:

1. Define-se o número da k vizinhos mais próximos a serem utilizados.

2. Para cada valor da amostragem, seleciona-se somente os vizinhos que possuem maior similaridade que um limiar definido previamente.

3. Calcula-se o número de verdadeiros e falsos positivos, verdadeiros e falsos negativos de toda amostragem.

Nessa amostra do Corpus Mensagem, utilizamos um total de 1570 documentos, distribuídos com 913 para a classe 1, 27 para a classe 2 e 556 para a classe 3.

A implementação do kNN foi utilizada considerando k vizinhos, com os valores de k dentre 1,2,4,8,16,32,64. Foram utilizadas 2 tabelas com a mesma estrutura. Na primeira tabela foi inserido todo o corpus mensagem, com inicialmente 1570 documentos. Na segunda tabela, foram selecionados aleatoriamente, através da função randômica do banco, equivalente a 10% do banco.

Tabela 2 Resultados de % de Erro para o kNN

k	% de Erro
1	39,97
2	39,91
4	**38,5**
8	39,3
16	39,17
32	38,84
64	39,37

Conforme disposto na Tabela 2 os resultados iniciais de percentuais de erro, alcançados com o kNN, não eram promissores. O melhor resultado foi do k=4 com 38,5. Segundo a literatura (BEKKERMAN e ALLAN, 2003b; SILVA, 2007b; YAMADA, YAMASHITA et al., 2006d) bons categorizadores atingem em média, percentuais de 20% de erro, sem muito esforço. Entretanto, ainda era necessário analisar em cada classe quais as medidas que seriam alcançadas para saber e o comportamento do kNN em todas as classes.

Tabela 3 Resultados do categorizador kNN para k=1

1NN		Abrang	Precisão	Medida-F
Sem SW	Classe 1	59,58	53,13	28,08
	Classe 2	0,00	0,00	0,00
	Classe 3	16,73	21,23	9,36
SW (prep + conj)	Classe 1	60,68	53,89	57,08
	Classe 2	0,00	0,00	0,00
	Classe 3	17,45	22,25	19,56
SW (prep+conj + nomes navais)	Classe 1	**89,70**	**63,74**	**74,52**
	Classe 2	0,00	0,00	0,00
	Classe 3	18,71	**52,26**	**27,55**
SW (prep + conj + nomes navais + nomes próprios)	Classe 1	86,64	62,93	72,90
	Classe 2	0,00	0,00	0,00
	Classe 3	**19,24**	45,92	27,12

Conforme a Tabela 3, por ser a classe dominante na base, a classe 1 tem mais possibilidade de ser encontrada. Isso se reflete nos resultados onde as medidas de abrangência, precisão e medida-F que são maiores nessa mesma classe. Embora muitas mensagens façam referências a nomes próprios, a retirada dos mesmos não foi relevante nesse primeiro momento, além de aumentar o tempo de processamento e a utilização de memória. Apenas na classe 3 a abrangência apresentou melhor resultado com a *stoplist* com nomes próprios. Assim, o melhor resultado global apareceu no conjunto que utiliza a retirada de preposições, conjunções e os nomes navais.

Tabela 4 Resultados do categorizador kNN para k=2.

2NN		Abrang	Precisão	Medida-F
Sem SW	Classe 1	58,93	53,11	55,87
	Classe 2	0,00	0,00	0,00
	Classe 3	17,63	21,73	19,46
SW (prep + conj)	Classe 1	60,57	53,74	56,95
	Classe 2	0,00	0,00	0,00
	Classe 3	17,09	21,94	19,21
SW (prep+conj + nomes navais)	Classe 1	**89,70**	**63,83**	**74,59**
	Classe 2	0,00	0,00	0,00
	Classe 3	19,78	54,46	29,02
SW (prep + conj + nomes navais + nomes próprios)	Classe 1	86,20	63,06	72,84
	Classe 2	0,00	0,00	0,00
	Classe 3	**20,14**	46,28	28,07

Para k=2, conforme a Tabela 4, os resultados mantêm-se melhores com a retirada de preposições, conjunções e os nomes navais, o que precisa ser corroborado com outros valores de k. Assim como no resultado anterior, com k=1, na classe 3 a abrangência também apresentou-se melhor com a *stoplist* completa.

Tabela 5 Resultados do categorizador kNN para k=4.

4NN		Abrang	Precisão	Medida-F
Sem SW	Classe 1	57,06	50,93	53,82
	Classe 2	3,70	0,00	0,00
	Classe 3	13,31	16,30	14,65
SW (prep + conj)	Classe 1	55,86	50,65	53,13
	Classe 2	0,00	0,00	0,00
	Classe 3	8,76	17,17	11,60
SW (prep+conj + nomes navais)	Classe 1	**90,47**	62,72	74,08
	Classe 2	0,00	0,00	0,00
	Classe 3	15,65	49,43	23,77
SW (prep + conj + nomes navais + nomes próprios)	Classe 1	90,25	**63,09**	**74,27**
	Classe 2	0,00	0,00	0,00
	Classe 3	**17,27**	**50,79**	**25,77**

Já na Tabela 5, para k=4 os resultados inverteram-se. Apenas na classe 1 a abrangência com a *stoplist* composta de preposições, conjunções e os trigramas e nomes navais, apresentou bom resultado. A stoplist completa, apresentou melhores resultados em todas as medidas da classe 3 e nas medidas de precisão e medida-F da classe 1.

Tabela 6 Resultados do categorizador kNN para k=8.

8NN		Abrang	Precisão	Medida-F
Sem SW	Classe 1	54,98	48,79	51,70
	Classe 2	0,00	0,00	0,00
	Classe 3	9,17	11,09	10,04
SW (prep + conj)	Classe 1	54,22	49,16	51,56
	Classe 2	0,00	0,00	0,00
	Classe 3	6,68	12,90	8,80
SW (prep+conj + nomes navais)	Classe 1	**92,55**	61,95	**74,22**
	Classe 2	0,00	0,00	0,00
	Classe 3	11,33	**48,09**	18,34
SW (prep + conj + nomes navais + nomes próprios)	Classe 1	91,02	**62,39**	74,03
	Classe 2	0,00	0,00	0,00
	Classe 3	**13,85**	46,95	**21,39**

Para k=8, conforme a Tabela 6, na classe 1 a abrangência e a precisão continuaram a apresentar melhores resultados com a retirada de preposições, conjunções e os nomes navais. Mas nessa mesma classe a precisão mostrou-se melhor retirando o conjunto maior de *stoplist* que contém nomes próprios.

Na classe 3, a abrangência e a medida-F apresentaram-se com melhores resultados com a retirada de preposições, conjunções e os nomes navais e nomes próprios, acontecendo também de ter o valor de precisão com melhor resultado usando outra *stoplist* , qual seja, a que não contém os nomes próprios.

Tabela 7 Resultados do categorizador kNN para k=16.

16NN		Abrang	Precisão	Medida-F
Sem SW	Classe 1	54,98	48,04	51,28
	Classe 2	0,00	0,00	0,00
	Classe 3	6,47	8,05	7,18
SW (prep + conj)	Classe 1	53,34	47,42	50,21
	Classe 2	0,00	0,00	0,00
	Classe 3	4,49	8,78	5,94
SW (prep+conj + nomes navais)	Classe 1	**94,41**	**62,10**	**74,92**
	Classe 2	0,00	0,00	0,00
	Classe 3	9,71	**50,47**	16,29
SW (prep + conj + nomes navais + nomes próprios)	Classe 1	93,43	61,99	74,53
	Classe 2	0,00	0,00	0,00
	Classe 3	**10,25**	47,50	**16,86**

Tabela 7, a classe 1 voltou a apresentar melhores resultados em todas a medidas com a retirada de preposições, conjunções e os nomes navais. Já para a classe 3, assim como na tabela anterior, para k=8, a abrangência e a medida-F apresentaram melhores resultados com a retirada de preposições, conjunções e os nomes navais e nomes próprios. Nessa mesma classe, acontece também o valor de precisão com melhor resultado usando outra *stoplist*, qual seja, a que não contém os nomes próprios.

Tabela 8 Resultados do categorizador kNN para k=32.

32NN		Abrang	Precisão	Medida-F
Sem SW	Classe 1	55,42	47,87	51,37
	Classe 2	0,00	0,00	0,00
	Classe 3	5,22	6,62	5,84
SW (prep + conj)	Classe 1	54,87	47,85	51,12
	Classe 2	0,00	0,00	0,00
	Classe 3	3,94	8,02	5,29
SW (prep+conj + nomes navais)	Classe 1	94,74	61,00	74,22
	Classe 2	0,00	0,00	0,00
	Classe 3	5,04	36,36	8,85
SW (prep + conj + nomes navais + nomes próprios)	Classe 1	**95,73**	**61,77**	**75,09**
	Classe 2	0,00	0,00	0,00
	Classe 3	**7,37**	**50,62**	**12,87**

Podemos notar que para k=32 e k=64, conforme as Tabela 8 e 9 , tanto a classe 1 quanto a classe 3 apresentaram melhores resultados com a retirada de preposições, conjunções e os nomes navais e nomes próprios.

Tabela 9 Resultados do categorizador kNN para k=64

64NN		Abrang	Precisão	Medida-F
Sem SW	Classe 1	59,26	48,78	53,51
	Classe 2	0,00	0,00	0,00
	Classe 3	2,52	3,62	2,97
SW (prep + conj)	Classe 1	59,91	49,41	54,16
	Classe 2	0,00	0,00	0,00
	Classe 3	2,41	5,66	3,38
SW (prep+conj + nomes navais)	Classe 1	96,06	61,03	74,64
	Classe 2	0,00	0,00	0,00
	Classe 3	3,78	35,59	6,83
SW (prep + conj + nomes navais + nomes próprios)	Classe 1	**96,60**	**61,38**	**75,06**
	Classe 2	0,00	0,00	0,00
	Classe 3	**4,50**	**42,37**	**8,13**

Para fins de comparação e análise, apresentamos a seguir os gráficos comparativos de cada métrica utilizada, para as classes 1 e 3. Não foram gerados os gráficos da classe 2 dois pois conforme os resultados apresentados não apresentaram valores significativos.

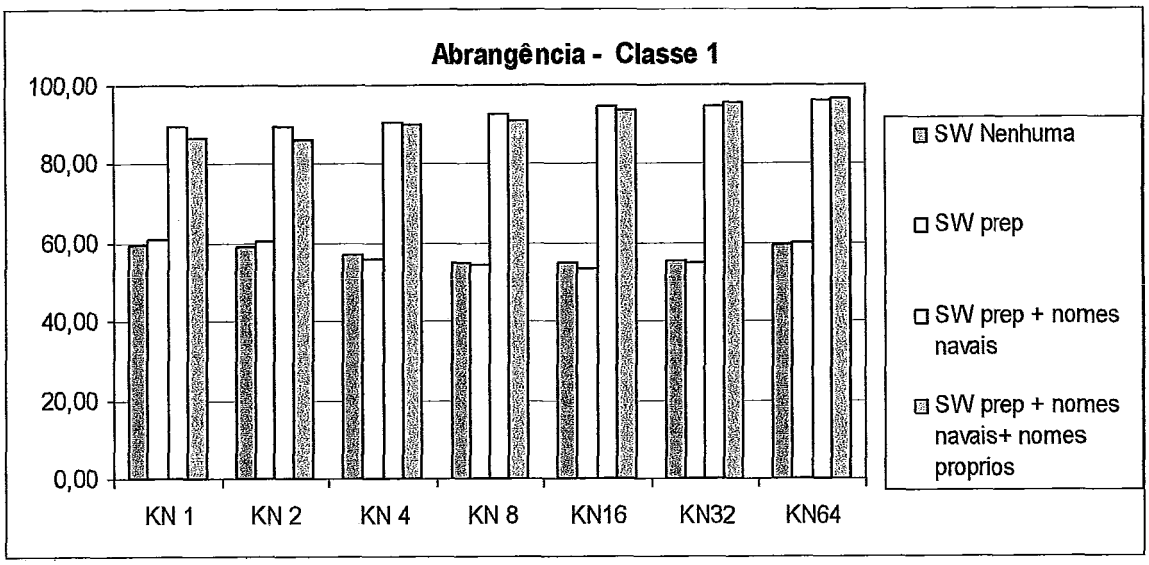


Figura 6 Abrangência da Classe 1

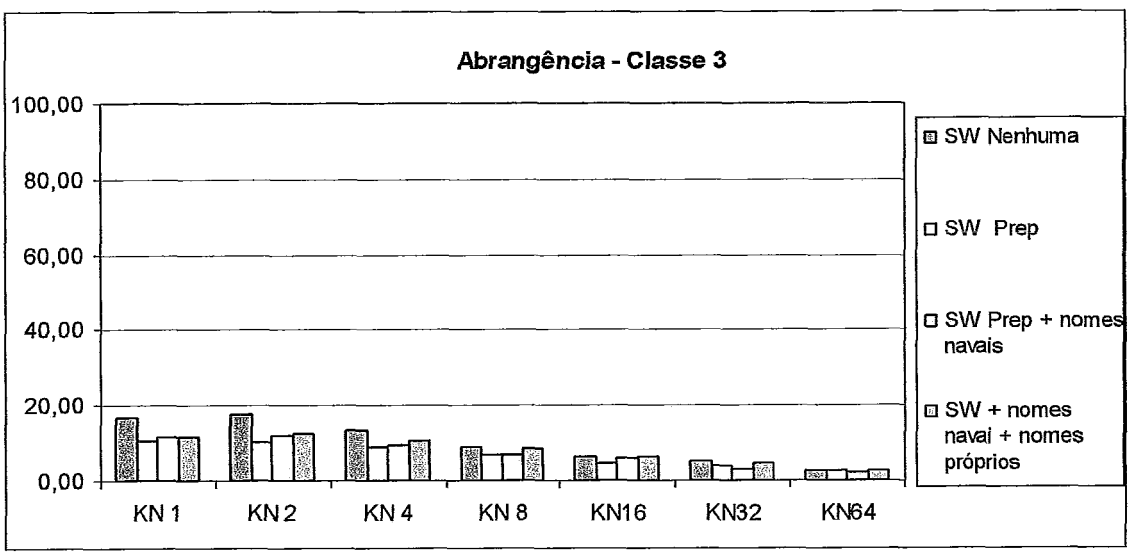


Figura 7 Abrangência da Classe 3.

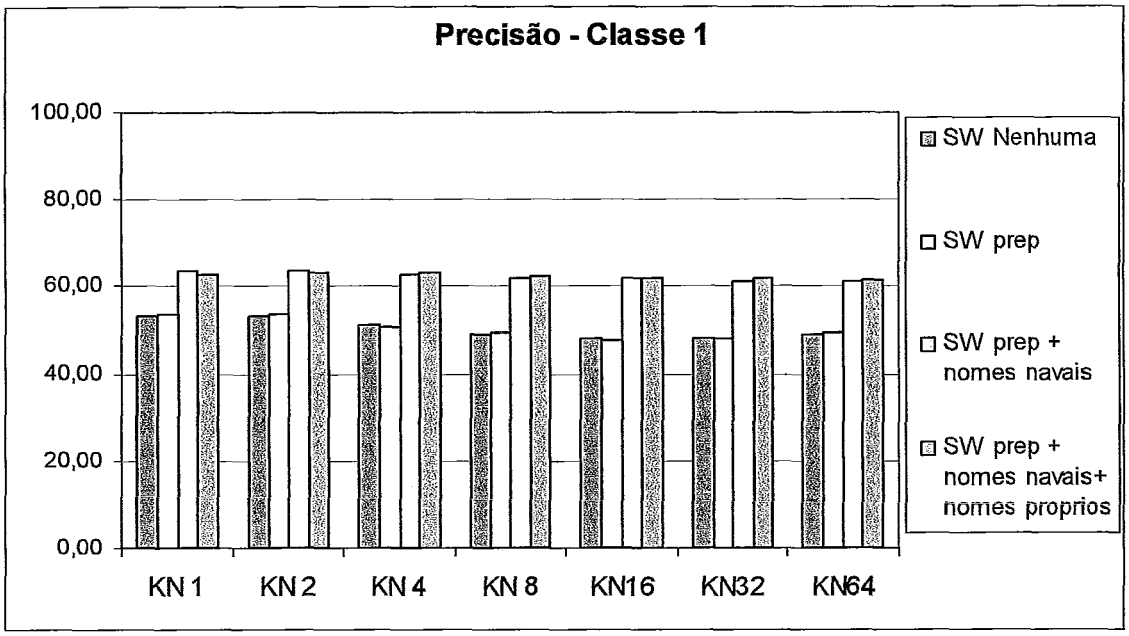


Figura 8 Gráfico de Precisão da Classe 1.

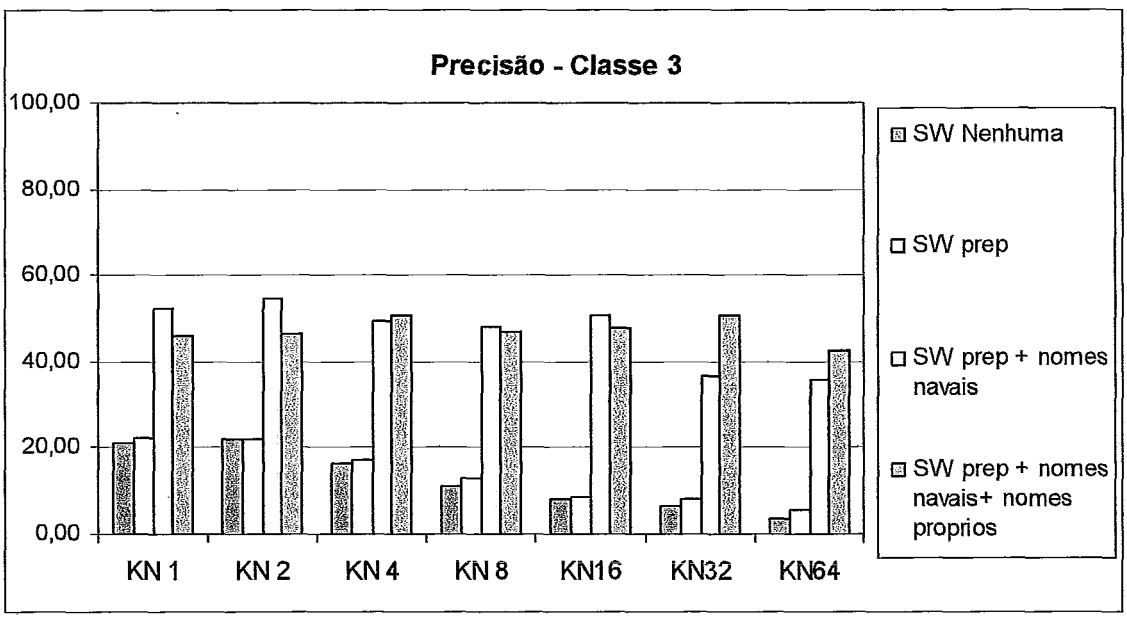


Figura 9 Gráfico de Precisão da Classe 3.

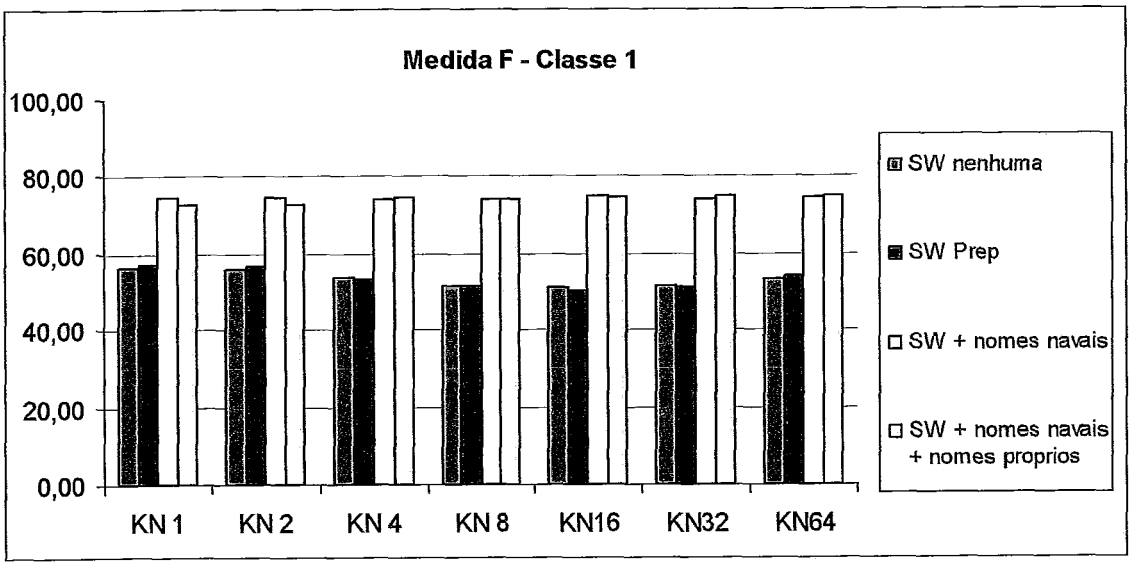


Figura 10 Medida-F da Classe 1.

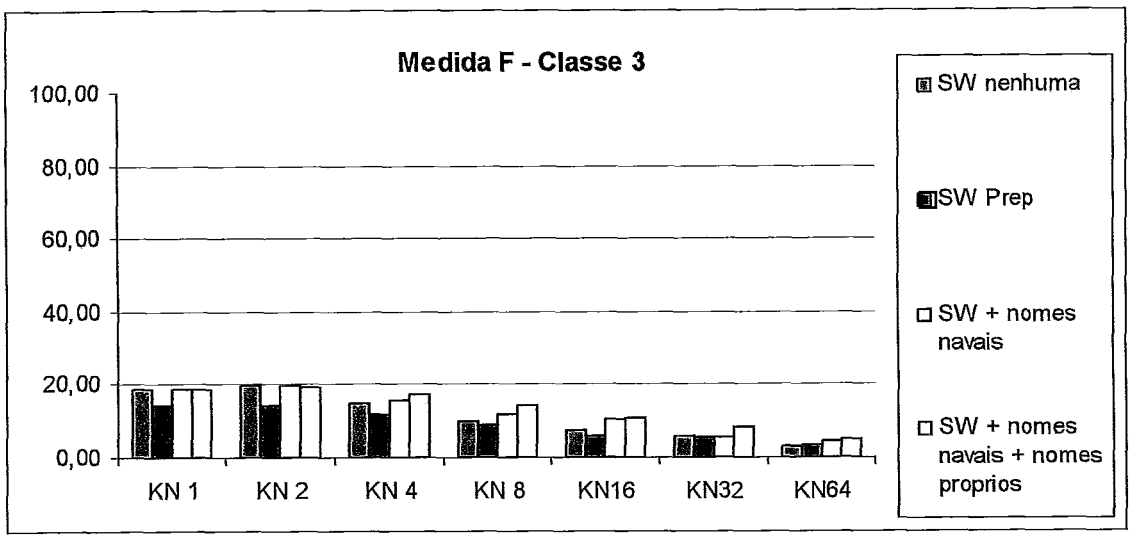


Figura 11 Medida-F da Classe 3.

4.4.4 – kNN – Análise dos Resultados

Diante dos resultados obtidos observa-se que a alternativa com o uso do kNN apresentou bons resultados para a classe 1, porém muito ruins para a classe 3 e ineficazes para a classe 2. Assim, observamos que os resultados com o kNN não pareciam ser tão promissores. Assim, foram iniciados os testes com os outros algoritmos.

Entretanto, a informação sobre a retirada de *stopwords*, conseguida com os testes iniciais, foi bastante relevante. O arquivo completo de *stoplist*, utilizado na última

configuração testada, possui 28.500 palavras. É importante ressaltar que os experimentos com o kNN já demonstram que não haveria ganhos muito superiores na utilização da *stoplist* com os nomes próprios. Para a continuidade dos testes, foi importante decidir sobre a melhor formação da *stoplist*, pois o conjunto de nomes próprios soma 25 mil palavras e aumenta consideravelmente o tempo de processamento. Além disso, uma *stoplist* muito grande acarretava consumo excessivo de memória durante o processamento. Assim, optou-se por utilizar a *stoplist* com preposições, conjunções e nomes navais de 3.500 palavras.

4.5 – Procedimentos da Solução – Fase 2

A partir dos resultados preliminares obtidos com o kNN, iniciamos os testes com o Naive Bayes e o SVM. É importante ressaltar que esse processo difere do que usa o kNN na Fase 1, que foi baseado nos testes com *stoplists*. Nesta segunda fase, cada texto é associado a um vetor de características derivado do próprio texto.

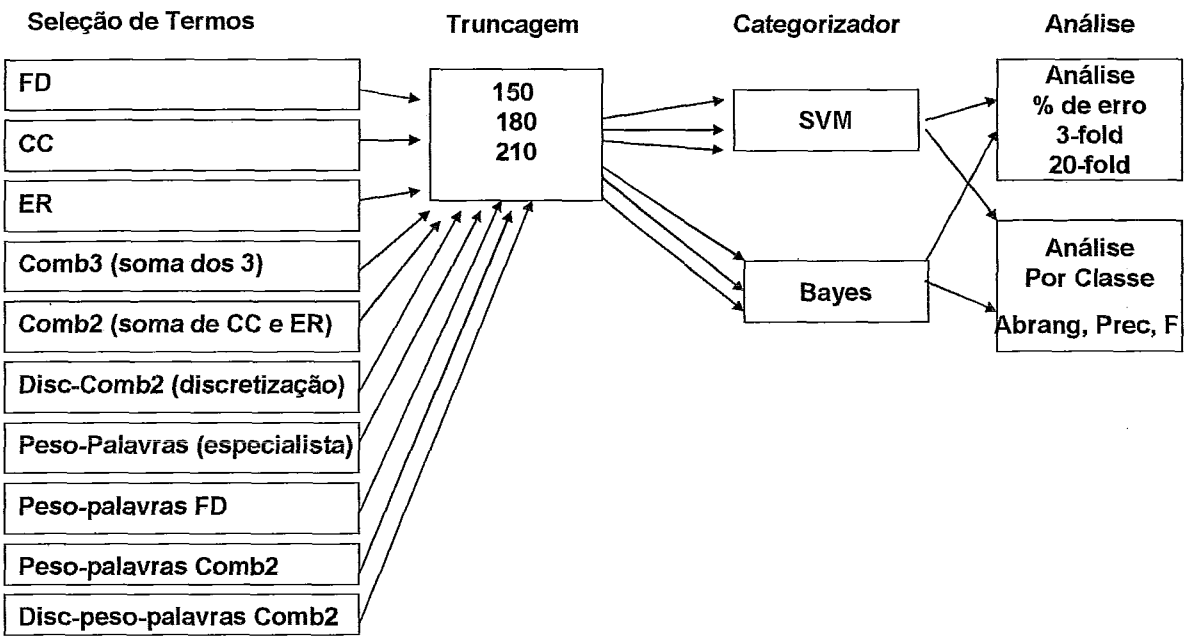


Figura 12 Processo de trabalho na fase 2

Na Figura 12 pode-se observar o processo utilizado nessa fase. Em resumo, para cada seleção de termos, foram selecionados 3 vetores, por exemplo, FD com 150 termos, FD com 180 termos e FD com 210 termos. Cada vetor desses foi submetido a cada um dos 2 categorizadores (bayes e SVM), e em cada categorizador foi realizado teste com validação cruzada com 3-*fold*, para uma análise global. Para uma análise em

cada classe utilizamos as medidas de abrangência, precisão e medida-F. Utilizamos também a validação com o parâmetro 20-*fold*.

As características são palavras que ocorrem no texto, onde cada uma tem um valor numérico baseado na sua freqüência de ocorrência, ou um valor booleano indicando a presença ou ausência da palavra. Geralmente, textos podem ser classificados a partir apenas de um pequeno conjunto dessas características.

Foram aplicadas as técnicas de redução estatística de FD, ER e CC. A partir desses primeiros resultados com FD, CC e ER, vislumbrou-se a possibilidade da utilização da combinação entre as medidas e da aplicação de outras técnicas visando aperfeiçoar os resultados descritos subseqüentemente.

Assim, em Comb3 temos a combinação das três medidas, somadas e submetidas a cada categorizador. Em Comb2 foi utilizada combinação das 2 melhores medidas CC e ER.

Já em Disc-Comb2 aplicou-se a discretização dos valores de cada termo, para valores já normalizados com a fórmula :

$$x_{\substack{norm \\ (0,1)}} = \frac{x - x_{min}}{x_{máx} - x_{min}}$$

Após a normalização, utilizou-se os valores 0, 0.5 e 1, para distanciar mais um termo do outro, dentro de classe.

Em Peso-palavras, foram identificadas palavras reconhecidamente significativas para cada classe com a ajuda de especialistas de cada setor. A partir do conjunto original, optou-se por multiplicar por três cada palavra do conjunto, que aparecesse naquela classe, visando aumentar o peso dessas palavras que semanticamente, representavam melhor cada classe.

Como o resultado de Peso-palavras não foi significativo, em Peso-palavras-FD foi aplicado o peso das palavras na medida FD, para verificarmos se teríamos maiores ganhos. Em Peso-palavras-Comb2 aplicou-se o peso das palavras a combinação de CC e ER. Finalizamos essa etapa de validação cruzada com 3-fold, utilizando a combinação Disc-peso-palavras-Comb2. Nessa última, foi aplicada a discretização sobre o peso das palavras aplicado a melhor combinação das 2 características, quais sejam CC e ER (Comb2).

Executamos a truncagem escolhendo os 50, 60 e 70 termos mais representativos de cada classe. Assim, os vetores globais continham 150, 180 e 210 termos.

O vetor correspondente ao texto é submetido a um categorizador que determina se o texto pertence ou não a determinada categoria que esteja sendo procurada. Nessa fase foram utilizados o SVM e o Bayes.

A validação cruzada foi feita utilizando-se parâmetros de 3-*fold* e 20-*fold*, para verificar se haveria alguma diferença. Assim como em mineração de dados e de textos, a classificação prévia é utilizada inicialmente para a geração do modelo e depois para a verificação de acerto dos textos categorizados.

4.5.1 – SVM e Naive Bayes - Preparando os dados

Para evitar o problema do desbalanceamento de classes, as mesmas foram particionadas em iguais proporções em relação às quantidades de arquivos. Esse procedimento visa a não super-especializar uma classe em detrimento de outra, evitando assim que o categorizador tenha a tendência de classificar novos exemplos como pertencentes à classe majoritária. Como a classe 2 era composta de apenas 270 documentos, este foi o valor inicial escolhido para a execução dos testes.

4.5.2 – SVM e Naive Bayes - Executando os testes

Utilizando a ferramenta Weka, descrita na Seção 4.3.1 – os testes foram executados com a melhor solução para *stoplist*, advinda dos testes com o kNN, composta de preposições, conjunções e nomes navais.

Segundo a literatura, o referido percentual, fica normalmente disposto em torno dos 20% (BEKKERMAN e ALLAN, 2003c; SILVA, 2007a; YAMADA, YAMASHITA et al., 2006c). Entretanto, reduzi-lo a menos de 10% ainda constitui um desafio. Embora já se encontre alguns trabalhos que conseguiram fazê-lo (POLAT e GÜNES, 2006b; YAMADA, YAMASHITA et al., 2006b), eles normalmente acontecem em condições muito específicas e normalmente com maior número de validações cruzadas.

Os resultados alcançados primeiramente com 3-*fold* estão dispostos na Tabela 10. Optamos aqui por enfatizar o percentual de erro, destacando em fonte maior, os menores percentuais de erro conseguidos em cada seleção testada.

Tabela 10 Resultados de % de erro com 3-*fold* para SVM e Naive Bayes.

Classificador - 3-fold	SVM			Bayes		
Seleção//N.termos	150	180	210	150	180	210
CC	17,04	17,04	**16,54**	18,15	18,64	18,15
ER	16,79	16,67	17,04	19,51	18,77	18,02
FD	20,74	20,99	20,49	24,57	25,43	24,69
Comb3	18,64	**16,3**	**16,3**	20,74	19,01	18,77
Comb2	18,02	17,28	16,79	16,42	**16,17**	**16,17**
Disc-Comb2	24,07	23,46	23,33	26,79	26,67	26,17
Peso-palavras	23,7	21,85	22,22	28,64	26,17	25,93
Peso-palavras-FD	21,36	21,11	20,62	24,94	25,68	25,06
Peso-Palavras-Comb2	17,53	16,67	**16,54**	**16,3**	**16,3**	17,04
Disc-peso-palavras-Comb2	17,78	17,04	16,79	16,67	**16,05**	**16,17**

Conforme observado, as combinações dos valores dos termos Comb2 e Comb3 e algumas exceções ocorridas com o SVM, conforme Figura 13, apresentaram bons resultados. Com exceção de CC com 210 termos que apresentou o mesmo resultado que Peso-Palavras-Comb2 para SVM.

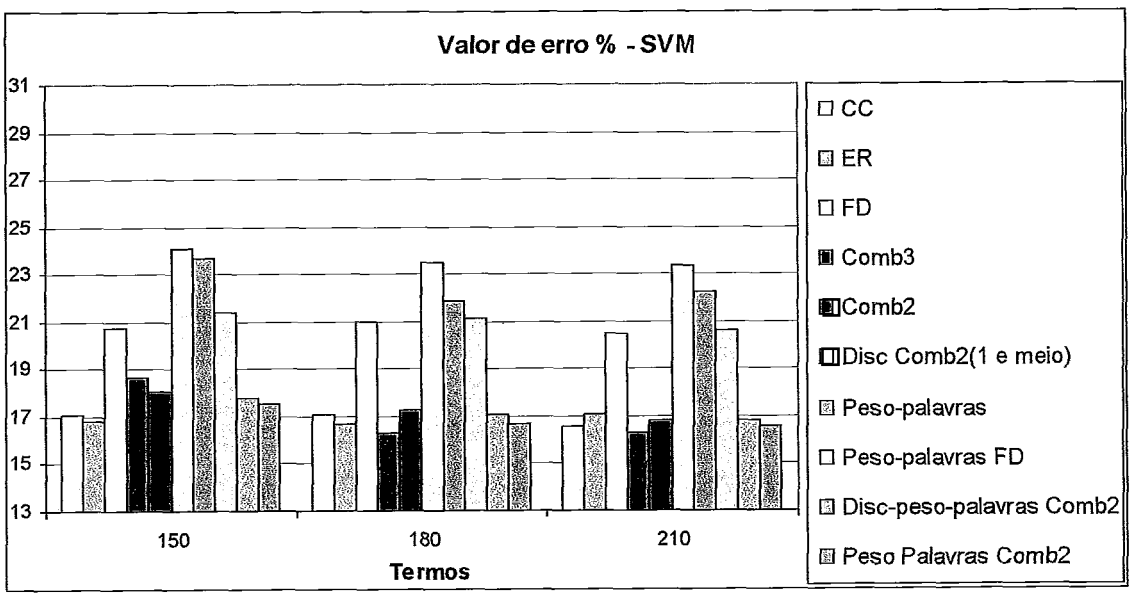


Figura 13 Resultados de % de erros SVM

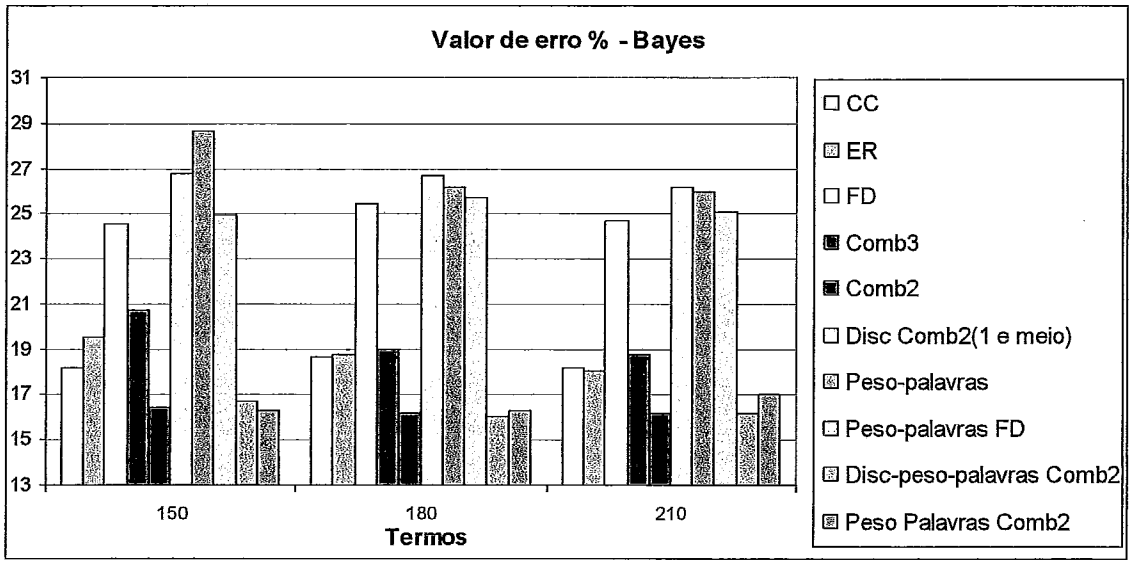


Figura 14 Resultados de % de erro Naive Bayes

Os vetores globais maiores tenderam também a oferecer melhores resultados. O melhor resultado, entretanto, apareceu no Categorizador Naive Bayes, com 16,05% de erro e usando a combinação Disc-peso-palavras-Comb2 conforme Figura 14.

Para verificar a possibilidade da redução do percentual de erros, aos melhores resultados obtidos em cada seleção, foram aplicados a validação com 20-*fold* para cada categorizador.

Tabela 11 Resultados % de erro com 20-*fold* para SVM e Naive Bayes

Categorizador - 20-fold	SVM	Bayes
CC	15.43	17.40
Comb3	16.17	19.01
Comb2	15,3	16.91
Peso-palavras-Comb2	**15,18**	16.66
Disc-peso-palavras-Comb2	15.30	16.79

Era de esperar-se um resultado melhor como aconteceu com o Peso-palavras-Comb2, devido a um maior treinamento conforme demonstrado na Tabela 11.

Os resultados com % de erro ofereceram uma visão geral do desempenho do categorizador com os vetores globais compostos por 150, 180 e 210 termos respectivamente. Para observarmos o desempenho de cada classe foram utilizadas, a exemplo do knn, as medidas de precisão, abrangência e medida-F, em cada classe e em cada categorizador.

Os números com fonte maior na cor preta mostram dentro de cada linha, ou seja, para cada seleção testada, qual o melhor resultado. Já os valores destacados a carmim, são os valores máximos de precisão, abrangência e medida-F, respectivamente, dentro de toda a tabela.

Tabela 12 Resultados *3-fold* para Classe 1 com SVM

Classe 1	SVM								
Termos	150			180			210		
Seleção	Prec	Abrang	F	Prec	Abrang	F	Prec	Abrang	F
CC	**0,98**	0,837	**0,9**	0,982	0,819	0,893	0,954	**0,852**	0,9
ER	**0,95**	0,844	0,896	**0,954**	0,841	0,894	0,947	**0,859**	**0,901**
FD	0,941	0,826	0,88	0,94	0,815	0,873	**0,95**	**0,826**	**0,881**
Comb3	0,966	0,83	0,892	**0,97**	0,833	**0,9**	0,946	**0,844**	0,892
Comb2	**0,97**	0,819	0,889	0,966	0,83	0,892	0,966	**0,841**	**0,899**
Disc Comb2	**0,96**	0,781	0,859	0,951	0,785	0,86	0,935	**0,8**	**0,862**
Peso-palavras	0,963	0,778	0,861	**0,964**	0,789	**0,87**	0,935	**0,8**	0,862
Peso-palavras FD	**0,94**	**0,83**	**0,88**	0,928	0,815	0,868	**0,94**	0,822	0,876
Peso-palavras Comb2	**0,97**	0,83	0,896	0,966	0,837	0,897	0,97	**0,848**	**0,905**
Disc-peso-palavras Comb2	**0,97**	0,815	0,887	0,966	0,83	0,892	0,97	**0,837**	**0,899**

Para o SVM, na Classe 1, conforme tabela Tabela 12, pode-se observar que embora a execução dos testes com o vetor de 210 termos apresente melhores resultados para medida-F. A precisão dessa classe teve seu melhor desempenho atingido no vetor de 150 termos. Pode-se observar que nessa classe, a medida CC com 150 termos obteve um bom resultado geral, estando muito próximo, com 0,90 do melhor valor de medida-f, de 0.905 obtido com o uso de 210 termos em Peso-palavras Comb2.

Tabela 13 Resultados *3-fold* para Classe 2 com SVM

Classe 2	SVM								
Termos	150			180			210		
Seleção	Prec	Abrang	F	Prec	Abrang	F	Prec	Abrang	F
CC	0,745	0,867	0,801	0,749	0,885	0,812	**0,76**	**0,893**	**0,823**
ER	0,773	**0,87**	0,819	0,777	0,867	**0,82**	0,773	0,859	0,814
FD	0,735	0,793	0,763	0,735	0,793	0,763	**0,74**	**0,807**	**0,77**
Comb3	0,741	0,859	0,796	0,773	0,885	0,826	**0,78**	**0,889**	**0,833**
Comb2	0,739	0,881	0,804	0,752	0,878	0,81	**0,77**	**0,881**	**0,819**
Disc-Comb2	0,652	**0,867**	0,744	0,66	0,856	0,745	**0,66**	**0,867**	**0,75**
Peso-palavras	0,648	**0,893**	0,751	0,673	0,878	0,762	**0,68**	0,885	**0,766**
Peso-palavras FD	0,732	0,778	0,754	0,738	0,793	0,764	**0,74**	**0,796**	**0,768**
Peso-palavras Comb2	0,747	0,885	0,81	0,755	**0,89**	0,818	**0,77**	0,878	**0,821**
Disc-peso-palavras Comb2	0,745	0,885	0,809	0,752	**0,89**	0,815	**0,77**	0,881	**0,821**

Na Tabela 13 explicitamente os resultados são melhores com 210 termos em quase todas as combinações executadas. Embora a abrangência seja ligeiramente melhor em CC, é na combinação dos melhores resultados das 3 medidas (CC, ER e FD), descritos em Comb 3 que a classe apresentou melhores resultados.

Tabela 14 Resultados *3-fold* para Classe 3 com SVM

Classe 3	SVM								
Termos	150			180			210		
Seleção	Prec	Abrang	F	Prec	Abrang	F	Prec	Abrang	F
CC	0,797	**0,785**	**0,791**	0,797	**0,785**	0,791	**0,81**	0,759	0,784
ER	0,79	0,781	0,786	0,79	**0,793**	**0,791**	0,785	0,77	0,778
FD	0,727	0,759	**0,743**	0,723	**0,763**	0,742	**0,73**	0,752	0,741
Comb3	0,766	0,752	0,759	0,796	**0,793**	**0,794**	**0,798**	0,778	0,788
Comb2	0,785	0,759	0,772	**0,8**	0,774	0,784	0,792	0,774	0,783
Disc-Comb2	0,739	0,63	0,68	0,747	0,656	**0,698**	**0,76**	0,633	0,691
Peso-palavras	0,759	0,619	0,682	**0,77**	0,678	**0,722**	0,769	0,652	0,705
Peso-palavras FD	0,715	0,752	0,733	0,724	0,759	0,741	**0,728**	0,763	**0,745**
Peso-palavras Comb2	0,788	0,759	0,774	**0,81**	0,77	**0,789**	0,787	0,778	0,782
Disc-peso-palavras Comb2	0,787	0,767	0,777	**0,8**	0,77	**0,786**	0.728	0.763	0.745

Na Classe 3, Tabela 14, pode-se observar que diferentemente dos resultados das classes anteriores, o vetor de 180 termos apresentou os melhores resultados de um modo geral. Em Comb 3 também observamos os melhores valores para abrangência e medida-F.

Foram executados os testes para cada classe com Naive Bayes e serão descritos a seguir:

Tabela 15 Resultados *3-fold* para Classe 1 com Naive Bayes

Classe 1	Bayes								
Termos	150			180			210		
Seleção	Prec	Abrang	F	Prec	Abrang	F	Prec	Abrang	F
CC	**0,955**	0,863	**0,907**	0,943	0,852	0,895	0,926	**0,878**	0,901
ER	**0,947**	0,856	**0,899**	0,92	0,852	0,885	0,917	**0,859**	0,887
FD	**0,919**	0,796	0,853	0,919	0,796	0,853	0,912	**0,807**	**0,857**
Comb3	**0,969**	0,804	0,879	0,961	0,822	**0,886**	0,949	**0,826**	0,883
Comb2	**0,959**	0,867	0,911	0,955	0,863	0,907	0,959	**0,87**	**0,913**
Disc-Comb2	**0,794**	0,8	0,797	0,793	0,807	0,8	0,786	**0,819**	**0,802**
Peso-palavras	0,766	0,789	0,777	**0,785**	**0,8**	**0,793**	0,779	0,796	0,788
Peso-palavras FD	**0,919**	0,796	**0,853**	0,918	0,793	0,851	0,9	**0,804**	0,849
Peso-palavras Comb2	**0,959**	**0,87**	**0,913**	0,955	0,867	0,909	0,944	**0,87**	0,906
Disc-peso-palavras Comb2	0,955	0,867	0,909	0,955	0,863	0,907	**0,959**	**0,87**	**0,913**

Na Classe 1, conforme Tabela 15, os melhores resultados de precisão encontram-se no vetor de 150 termos, enquanto que em abrangência podemos observar que o vetor de 210 termos obteve melhor desempenho. A medida-F acabou por distribuir-se entre os dois conjuntos de termos. Aqui não existiu pelos valores obtidos uma seleção que se destaque em todas as medidas.

Já para Classe 2, conforme a Tabela 16, apesar de os melhores resultados de cada seleção concentrarem-se no vetor de 210 termos, é com o vetor de 180 termos que temos o valor máximo atingido na tabela para todas as medidas.

Tabela 16 Resultados *3-fold* para Classe 2 com Naive Bayes

Classe 2	Bayes								
Termos	150			180			210		
Seleção	Prec	Abrang	F	Prec	Abrang	F	Prec	Abrang	F
CC	0,799	**0,752**	**0,775**	0,804	0,744	0,773	**0,812**	0,737	0,773
ER	0,773	0,756	0,764	0,79	0,767	0,778	**0,795**	**0,789**	**0,792**
FD	0,702	**0,707**	**0,705**	0,698	0,685	0,692	**0,724**	0,67	0,696
Comb3	0,797	0,726	0,76	**0,835**	0,73	0,779	0,821	**0,748**	**0,783**
Comb2	0,808	0,793	0,8	0,821	**0,8**	**0,808**	**0,823**	0,774	0,798
Disc-Comb2	0,706	0,711	0,708	0,704	0,715	0,71	**0,709**	**0,741**	**0,725**
Peso-palavras	0,687	0,681	0,684	0,698	0,744	0,72	**0,706**	**0,756**	**0,73**
Peso-palavras FD	0,698	**0,693**	0,695	0,7	0,674	0,687	**0,724**	0,67	**0,696**
Peso-palavras Comb2	0,813	0,789	0,801	0,814	**0,79**	**0,803**	**0,822**	0,752	0,785
Disc-peso-palavras Comb2	0,804	0,789	**0,796**	**0,827**	0,796	**0,811**	0,826	0,774	0,799

Os resultados da Classe 3 estão dispostos na Tabela 17, onde de forma análoga à classe 2 com naive bayes, os resultados por seleção foram melhores com o vetor de 210 termos. Entretanto, nessa classe, os melhores valores de precisão, abrangência e medida-F, encontram-se um em cada vetor, tendendo a seleção Comb 2 a ser melhor por apresentar duas das medidas, precisão e medida-F com os melhores resultados dessa tabela.

Tabela 17 Resultados *3-fold* para Classe 3 com Naive Bayes

Classe 3	Bayes								
Termos	150			180			210		
Seleção	Prec	Abrang	F	Prec	Abrang	F	Prec	Abrang	F
CC	0,728	0,841	0,78	0,722	**0,84**	0,778	**0,735**	0,841	0,784
ER	0,719	0,804	0,759	0,742	**0,82**	0,778	**0,758**	0,811	0,784
FD	**0,674**	0,759	0,714	0,656	0,756	0,702	0,657	**0,781**	0,714
Comb3	0,674	0,848	0,751	0,691	**0,88**	0,773	**0,708**	0,863	0,778
Comb2	**0,802**	0,848	0,761	0,76	0,856	0,805	0,756	**0,87**	**0,809**
Disc-Comb2	0,695	**0,685**	0,69	0,701	0,678	0,689	**0,717**	0,656	0,685
Peso-palavras	0,686	**0,67**	0,678	0,733	**0,67**	0,7	**0,739**	**0,67**	0,703
Peso-palavras FD	**0,669**	0,763	0,713	0,65	0,763	0,702	0,655	**0,774**	0,71
Peso-palavras Comb2	0,759	0,852	0,803	**0,762**	0,852	0,804	0,745	**0,867**	0,801
Disc-peso-palavras Comb2	**0,76**	0,844	0,8	0,758	0,859	0,806	0,753	**0,87**	0,808

Como o valor da medida-F é maximizado quando a precisão ou abrangência são iguais ou muito próximas, para melhor visualização e comparação entre os resultados obtidos nas classes com o SVM e o Naive Bayes, utilizamo-nos novamente desses gráficos.

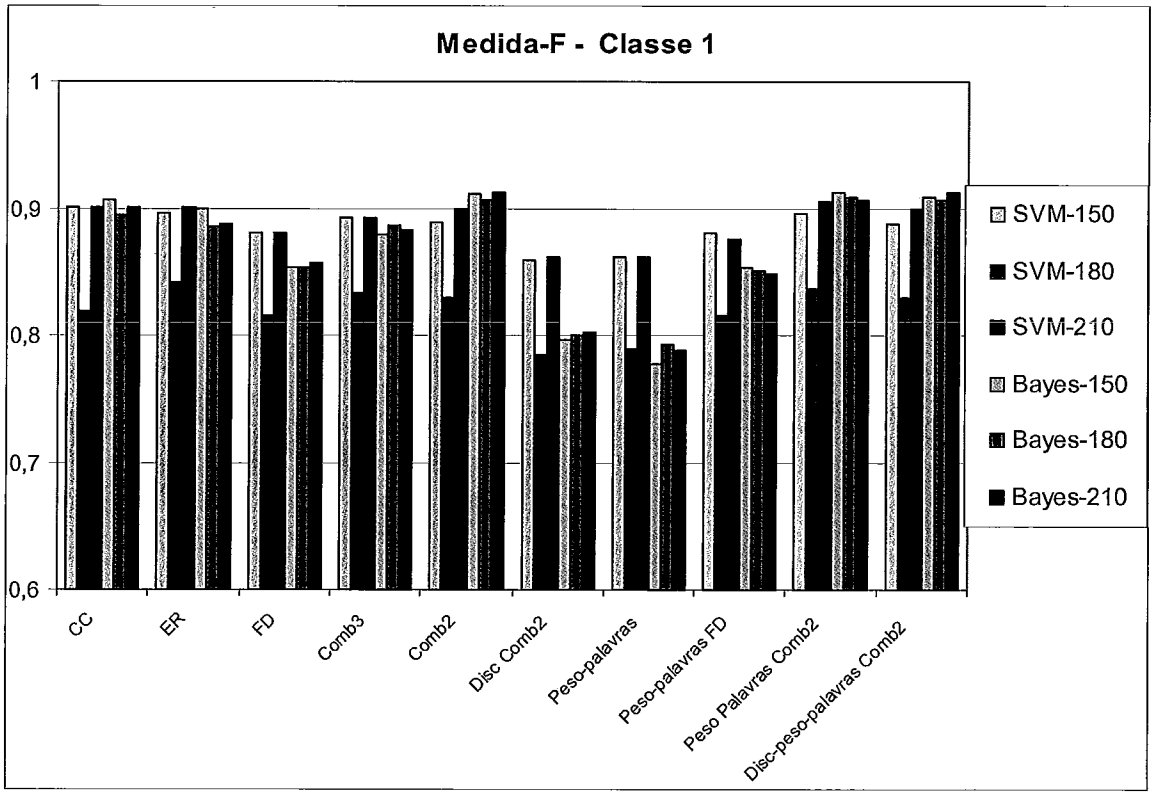


Figura 15 Medida-F da classe 1 de SVM e Bayes

No comparativo apresentado no gráfico da Figura 15 observa-se que na classe 1, embora o SVM com 210 termos apresente bons percentuais em todos os tipos de seleção testadas, a opções Comb2 , Peso-palavras-comb2 e Disc-peso-palavras-comb2 apresentaram melhores resultados com Bayes também com 210 termos.

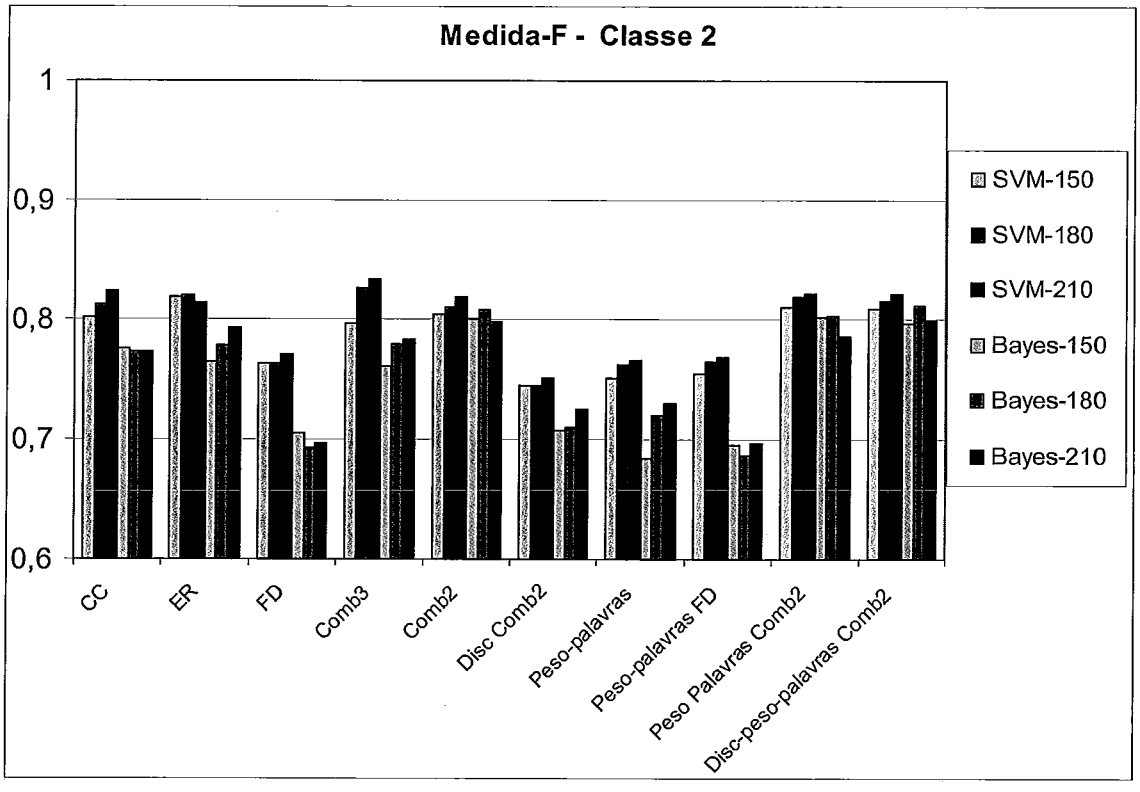


Figura 16 Medida-F da classe 2 de SVM e Bayes

Já para a classe 2, pode-se observar no gráfico da Figura 16, o SVM com 210 termos destaca-se em todos os tipos de seleção testados. Agora o melhor resultado apareceu na seleção Comb3, seguido por Peso-palavras-Comb2 e Disc-peso-palavras Comb2.

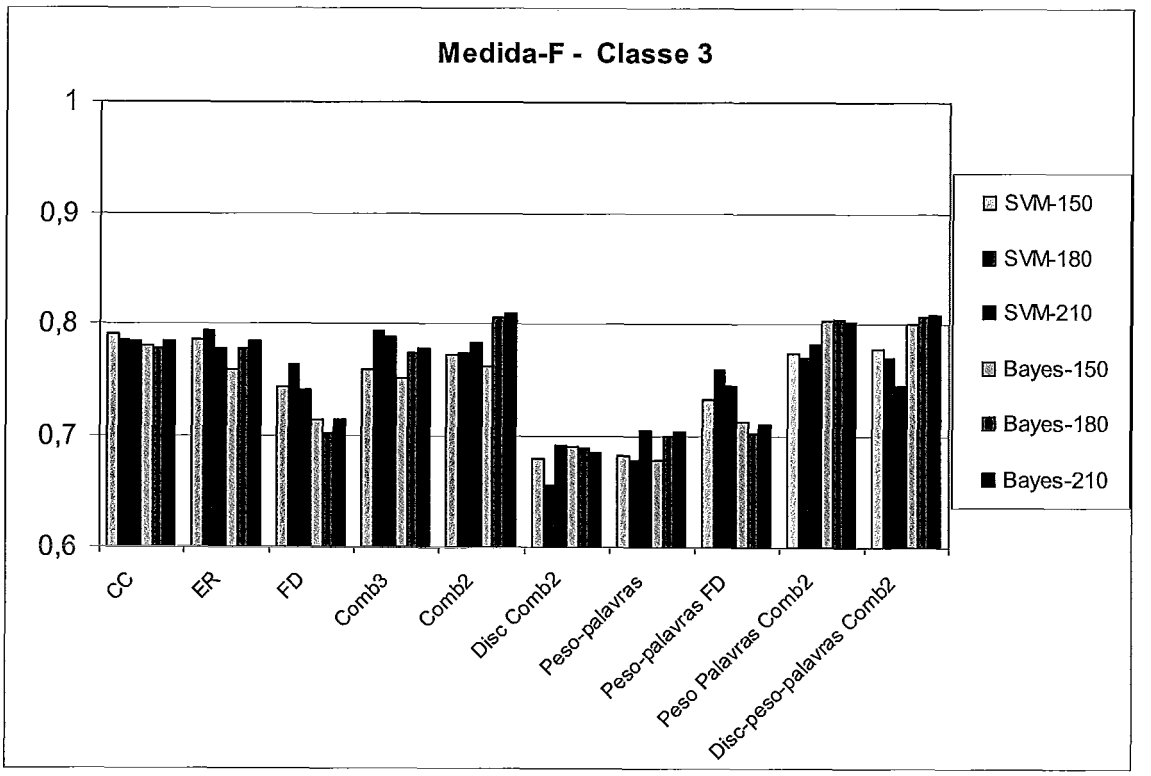


Figura 17 Medida-F da classe 3 de SVM e Bayes

No gráfico da classe 3 podemos observar na Figura 17 que diferentemente das classes 1 e 2 o SVM-210 não sobressaiu em todas as classes. Entretanto Bayes-210 obteve melhores resultado em Comb2, seguido por Disc-peso-palavras-comb2 e Peso-palavras-comb2.

4.5.3 – SVM e Naive Bayes – Análise Inicial dos Resultados

Conforme demonstrado nos resultados anteriores, a taxa de percentual de erro faz uma amostragem global do resultado do SVM e do Naive Bayes. Para verificar o desempenho dentro de cada classe foram utilizadas as medidas de precisão, abrangência e medida-F de forma análoga ao realizado para o kNN.

Os percentuais de medida-F foram maiores na classe 1, seguidas da classe 2 e por último a classe 3. Em uma comparação dos três algoritmos por classe, identificamos que a classe 1 foi melhor com os 3 algoritmos, havendo entretanto uma inversão de resultados na classe 2. Enquanto que com o kNN utilizando-se de todo o conjunto, essa classe mal pode ser localizada,com o SVM e com o Nayve Bayes, a classe 2 foi mais bem categorizada do que a classe 3.

Podemos observar que as combinações nos três gráficos anteriores, que essas combinações Comb2, Disc-peso-palavras-comb2 e Peso-palavras-comb2 mostram-se

vantajosas em todas as classes. Entretanto os percentuais de categorização da classe 3 estão mais baixos que das outras classes e podem ser melhorados.

4.5.4 – SVM e Naive Bayes – Aplicando o Teorema NFL

Aos melhores resultados da Seção anterior foi aplicado o teorema NFL. Conforme LA-04 citado na seção 4.2 – a idéia desse teorema é de que o categorizador resultará em melhores índices caso precise reconhecer apenas uma classe por vez.

Para isso, o teorema foi aplicado em 2 momentos. Primeiramente aplicou-se nas técnicas na fase de pré-processamento. Assim, para cada classe considerou-se uma classe e a soma das outras classes como não-classe. Utilizamos apenas 2 classes de cada vez. Por exemplo, para escolhermos os termos da classe 1, usamos a própria e o conjunto formado pelas não-classes 2 e 3 conforme demonstrado na Figura 18. O mesmo conceito foi aplicado a cada classe. Com isso, os vetores locais e depois o global foram gerados de modo mais especializado, utilizando-se apenas a identificação de cada classe e desprezando-se a não classe conforme demonstrado na Figura 19.

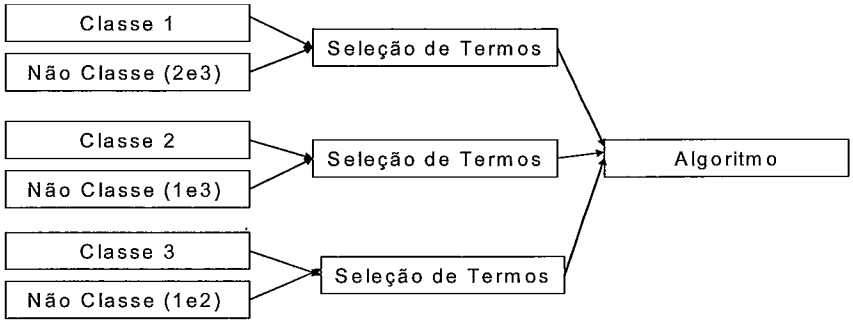


Figura 18 NFL aplicado na Seleção de Características

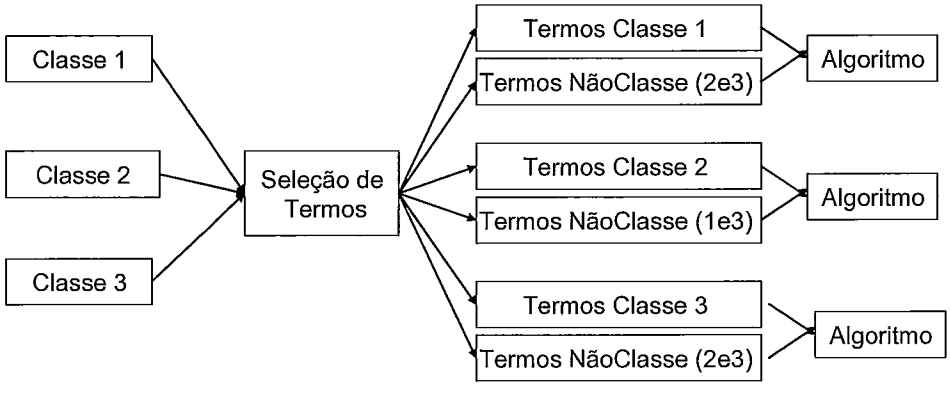


Figura 19 NFL aplicado ao Categorizador

Os resultados dos percentuais de erro das duas arquiteturas estão descritos na Tabela 18, onde foram aplicados para 210 termos e com o melhor percentual para Comb2. Em nenhuma das duas abordagens NFL os resultados foram promissores.

Tabela 18 Resultados de % de erro do NFL

NFL - 20-*fold*	SVM	Bayes
Seleção de Características	27.14	28.29
Categorizador	**23.97**	28.46

4.6 – Procedimentos da Solução – Fase 3

Visando melhorar os percentuais na categorização, resolveu-se aumentar a quantidade de amostras usadas para treino. Os documentos surgiram dos primeiros meses do ano de 2003, uma vez que o ano não deve influenciar no conteúdo dos documentos. Conseguiu-se aumentar cada amostragem de 270 para 490 documentos por classe. Além disso, como a maior parte dos documentos é muito pequena, resolveu-se utilizar os maiores documentos disponíveis em cada classe.

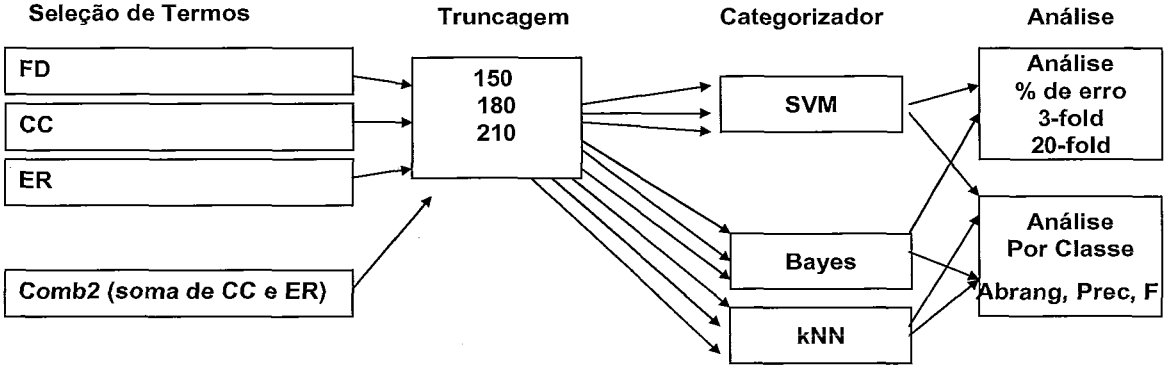


Figura 20 Processo de trabalho na fase 3

Na Figura 20 pode-se observar o processo utilizado na fase 3. O processo é praticamente o mesmo da fase 2, porém desta feita, fizemos primeiro um comparativo entre os resultados do Bayes e do SVM. Posteriormente passamos a comparar o SVM e o kNN.

4.6.1 – SVM e Naive Bayes – Otimizando resultados

Esperava-se com a melhoria nos documentos do Corpus, pudessem ser obtidos melhores resultados, uma vez que o categorizador poder-se-ia dispor a princípio, de maiores e melhores exemplos.

Tabela 19 Resultado 3-*fold* de % de erro para amostragem maior

Termos		SVM			Bayes	
Seleção	150	180	210	150	180	210
CC	19.59	18.84	17.619	28,98	30,27	28,03
ER	17.483	17.89	**15.23**	28,1	29,52	27,76
FD	17.00	18.843	15.64	28,1	28,57	**27,62**
Comb2	19.932	16.93	15.30	28,16	30,27	27,82

Com uma melhor amostragem, os resultados dos percentuais de erros descritos na Tabela 19 apresentaram-se inferiores quando da utilização do algoritmo Naive Bayes. Já o SVM apresentou bons resultados e foi melhor examinado em cada classe visando descobrir se haveria um melhor resultado em cada uma delas em relação à amostragem normal de 270 documentos, utilizada nos testes anteriores.

Tabela 20 Medidas das Classes 1, 2 e 3 para com Corpus de 490 - SVM

	SVM								
Termos	150			180			210		
Seleção	Prec	Abrang	F	Prec	Abrang	F	Prec	Abrang	F
Classe 1									
CC	0,93	0,814	0,868	**0,932**	0,812	0,868	0,931	**0,827**	**0,876**
ER	0,927	0,824	0,873	0,916	0,824	0,868	**0,94**	**0,851**	**0,892**
FD	**0,916**	0,827	0,869	0,888	0,81	0,847	0,915	**0,839**	0,875
Comb2	0,921	0,804	0,858	0,938	0,833	0,882	**0,94**	**0,843**	**0,89**
Classe 2									
CC	0,7	0,871	0,776	0,72	0,861	0,784	**0,73**	**0,878**	**0,795**
ER	0,756	**0,847**	0,799	0,738	**0,847**	0,789	**0,84**	0,831	**0,836**
FD	0,767	0,851	0,807	0,742	0,827	0,782	**0,78**	**0,859**	**0,817**
Comb2	0,699	0,859	0,771	0,747	0,867	0,803	**0,77**	**0,884**	**0,824**
Classe 3									
CC	0,826	0,727	0,773	0,816	0,761	0,788	**0,85**	**0,767**	**0,806**
ER	0,812	0,804	0,808	**0,831**	0,792	0,811	0,78	**0,861**	**0,819**
FD	0,822	0,812	0,817	0,82	0,798	0,809	**0,85**	**0,833**	**0,84**
Comb2	0,823	0,739	0,778	0,833	0,792	0,812	**0,85**	**0,814**	**0,83**

Verificando os resultados de precisão, abrangência e medida-F das classes 1, 2 e 3 para SVM, na Tabela 20 pode-se observar um padrão onde os melhores resultados de todas as classes concentram-se nos com 210 termos, em todas as seleções testadas.

Para efeito de comparação entre os resultados SVM dos conjuntos de 270 e 490 documentos, seus valores foram exibidos nos gráficos a seguir.

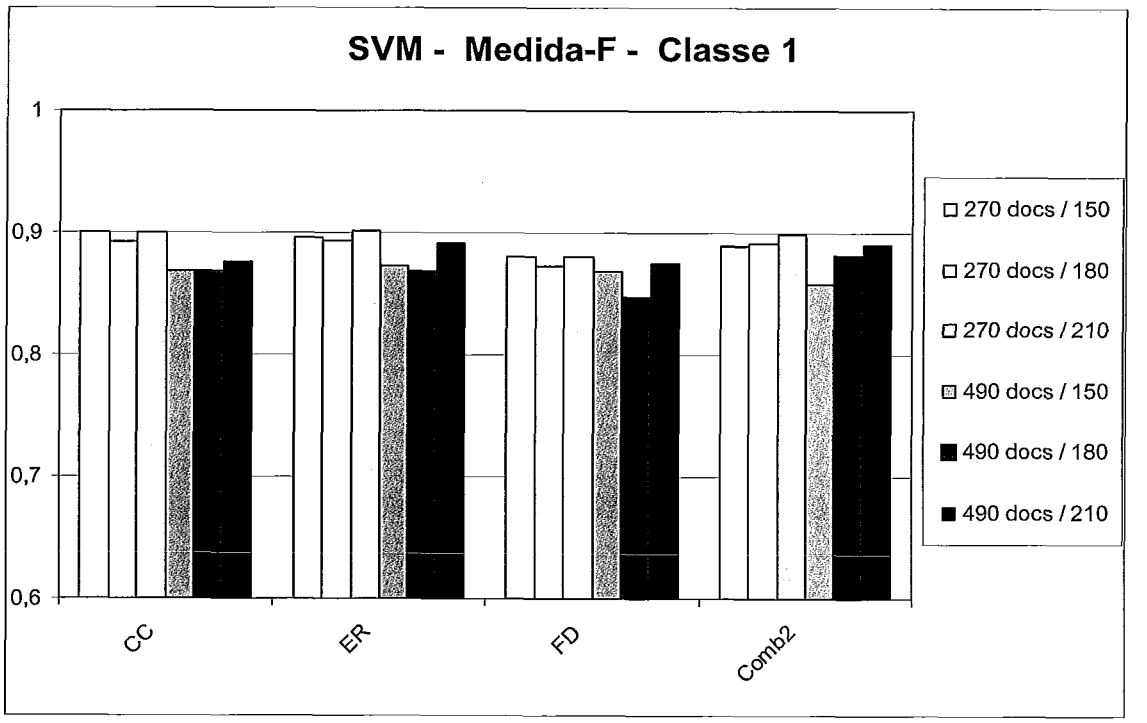


Figura 21 Medida –F comparativa entre conjuntos de documentos da Classe 1.

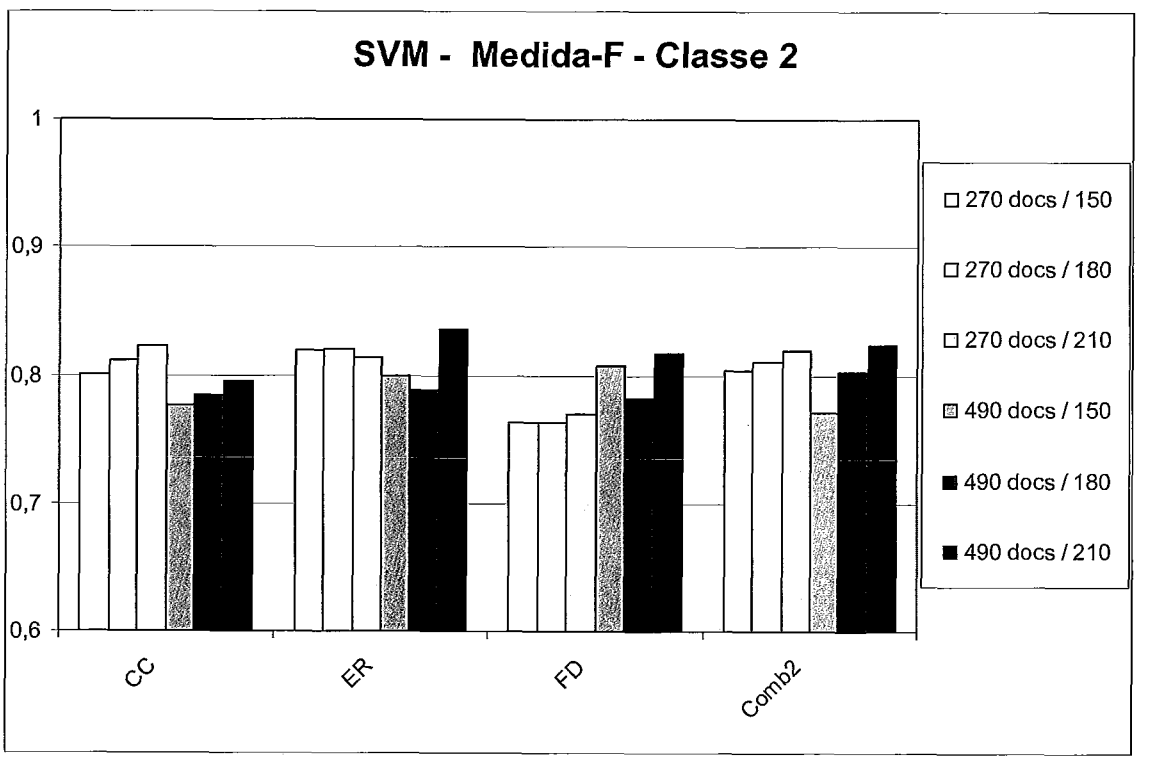


Figura 22 Medida –F comparativa entre conjuntos de documentos da Classe 2.

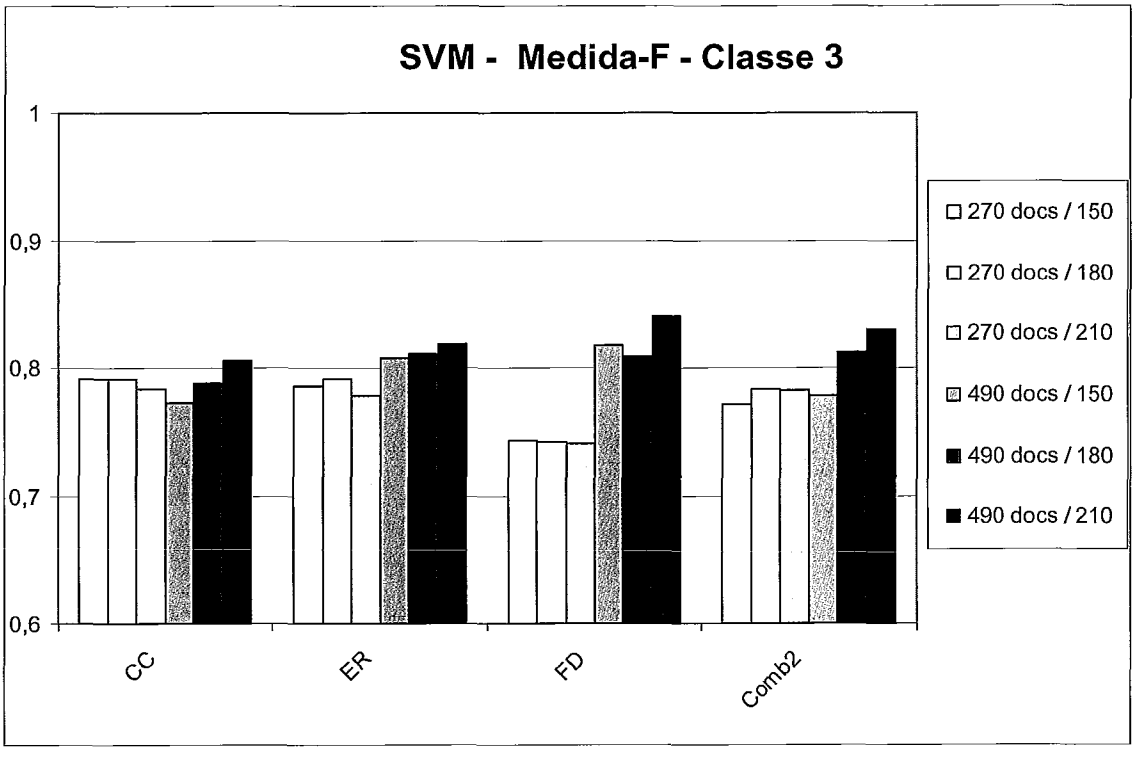


Figura 23 Medida –F comparativa entre conjuntos de documentos da Classe 3.

Pode-se observar através do Gráfico da Figura 21 que na comparação entre os conjuntos de documentos da Classe 1, o conjunto original com 270 documentos apresentou melhores resultados. Os valores de todos os conjuntos continuam maiores para essa classe. Entretanto, para as classes 2 e 3, conforme Figura 22 e Figura 23, os resultados para conjuntos maiores foram melhores com vetores de 210 termos.

Resumindo, na classe 1, a melhoria da amostra não obteve ganhos. Outra observação sobre os testes com o SVM, é que o percentual de erro, em torno de 15% , ainda é alto. Assim, retomamos os testes com o kNN.

4.6.2 – kNN – Testes finais

Sobre os melhores resultados, ou seja, aqueles em que houve um maior tratamento na fase de pré-processamento, resolveu-se voltar a aplicar o kNN. Assim, poder-se-ia verificar, o quanto os resultados anteriores do mesmo categorizador seriam otimizados com o referido tratamento.

Os resultados de percentual de erro para 20-*fold* com k=1 estão descritos na Tabela 21.

Tabela 21 Resultado 20-*fold* de % de erro para kNN.

Termos	kNN		
Seleção	150	180	210
CC	11,29	11,97	11,36
ER	10,81	10,88	**10,13**
FD	11,36	11,08	11,29
Comb2	10,81	10,61	**10,54**

Agora com o kNN na amostragem do Corpus Mensagem de 490 documentos, apresentam-se os resultados de abrangência, precisão e medida-F na Tabela 22. O vetor de 210 termos ainda apresenta melhores resultados de uma forma geral, porém não tão unânime quanto nos resultados com SVM descritos na Tabela 20.

Tabela 22 Medidas das Classes 1, 2 e 3 para Corpus de 490 - kNN

	kNN - 490								
Termos	150			180			210		
Seleção	Prec	Abrang	F	Prec	Abrang	F	Prec	Abrang	F
Classe 1									
CC	0,925	**0,884**	0,904	0,916	0,871	0,893	**0,93**	0,882	**0,905**
ER	0,917	0,882	0,899	0,925	0,876	0,899	**0,93**	**0,886**	**0,905**
FD	0,887	**0,898**	0,892	**0,901**	0,89	**0,9**	0,876	0,896	0,886
Comb2	0,918	**0,894**	0,906	0,927	0,88	0,903	**0,93**	0,884	**0,907**
Classe 2									
CC	**0,852**	0,916	0,883	0,85	0,916	0,882	0,85	**0,922**	**0,885**
ER	**0,862**	0,92	0,89	0,847	0,935	0,888	0,86	**0,939**	**0,898**
FD	**0,901**	0,888	0,894	0,879	**0,916**	**0,9**	0,895	0,892	0,894
Comb2	**0,857**	0,931	0,892	0,851	0,935	0,891	0,853	**0,937**	**0,893**
Classe 3									
CC	**0,888**	**0,861**	**0,875**	0,878	0,853	0,865	0,886	0,855	0,87
ER	0,899	**0,873**	0,886	0,91	0,863	0,886	**0,92**	0,871	**0,893**
FD	0,872	0,873	0,873	0,888	0,861	0,875	**0,89**	**0,873**	**0,882**
Comb2	0,905	0,851	0,877	**0,91**	**0,867**	**0,89**	0,906	0,863	0,884

Os índices de todas as medidas são superiores aos do SVM e são demonstrados com a medida-F nos gráficos a seguir.

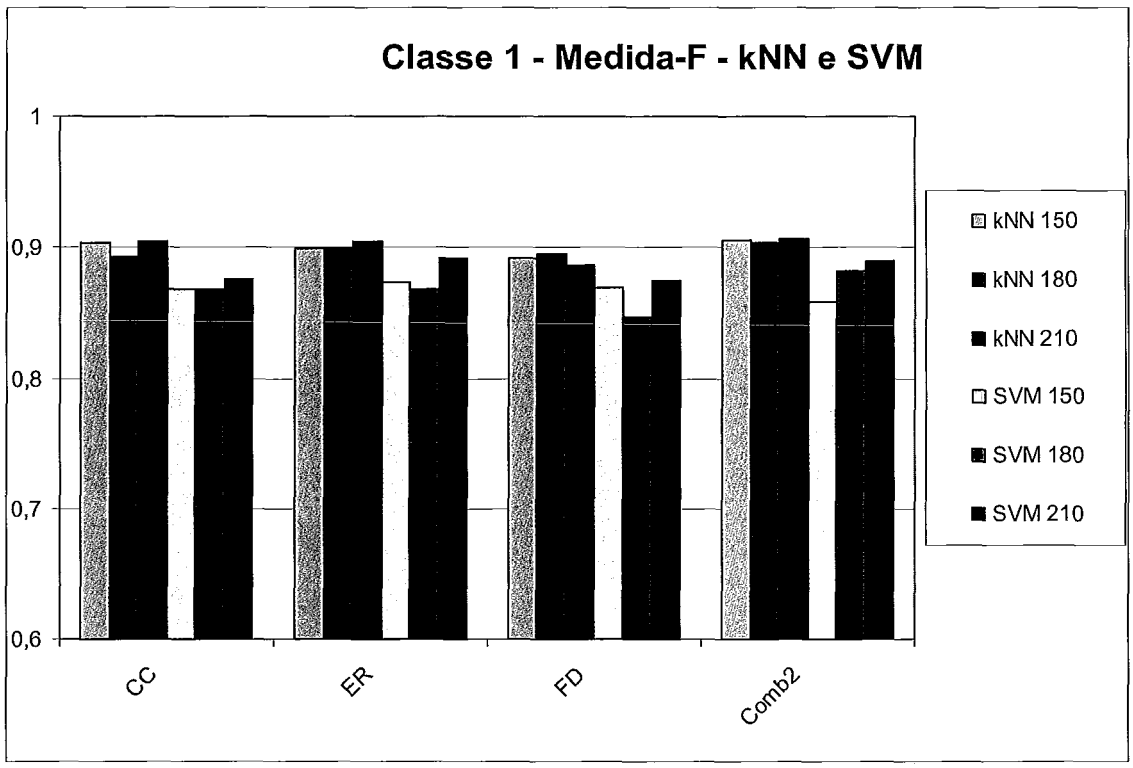


Figura 24 Classe 1 - Medida –F comparativa entre SVM e kNN

No gráfico da Figura 24 pode ser observado que o kNN apresenta na classe 1, melhor desempenho, independentemente do número de termos dos vetores, quais sejam, 150, 180 ou 210. Entretanto o vetor de 210 obteve melhores resultados em todas as seleções, excetuando-se em FD.

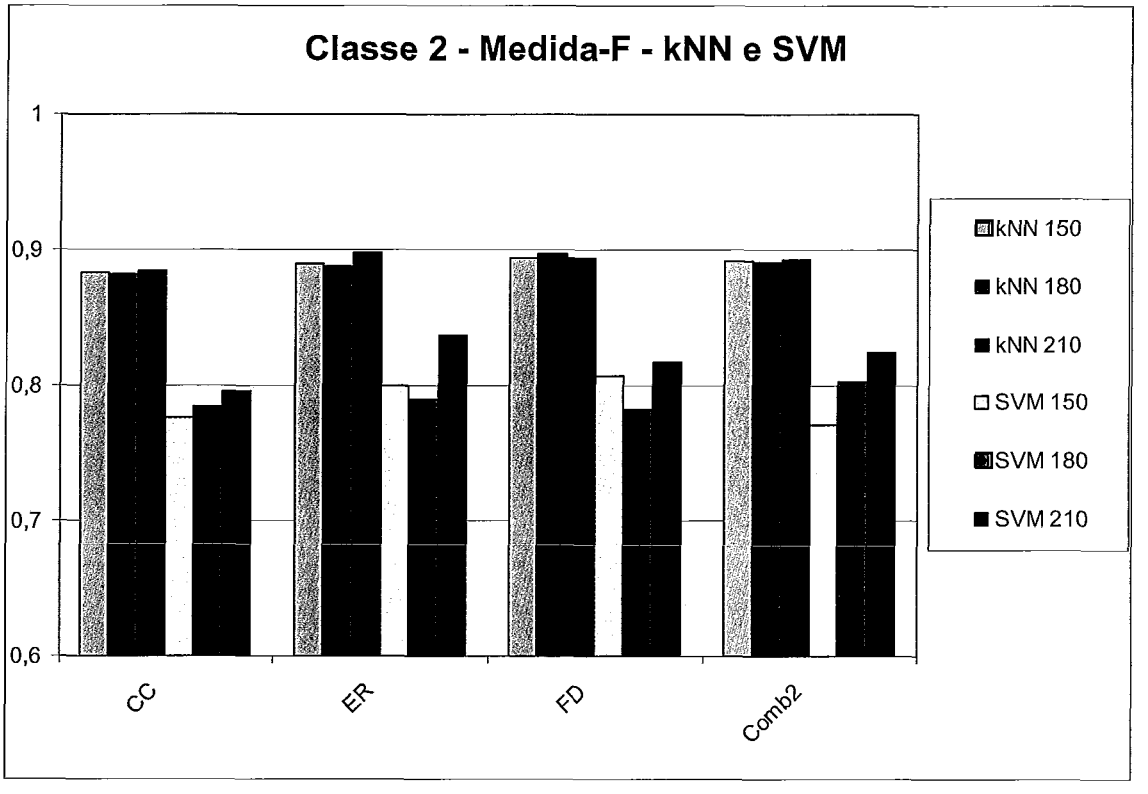


Figura 25 Classe 2 - Medida –F comparativa entre SVM e kNN

De acordo com os gráficos das Figura 24 e 25 pode-se observar que assim como na classe 1 os percentuais do kNN foram maiores, entretanto, nesta classe o diferença entre os dois categorizadores foi bem mais acentuada, mostrando a clara vantagem do kNN nessa classe.

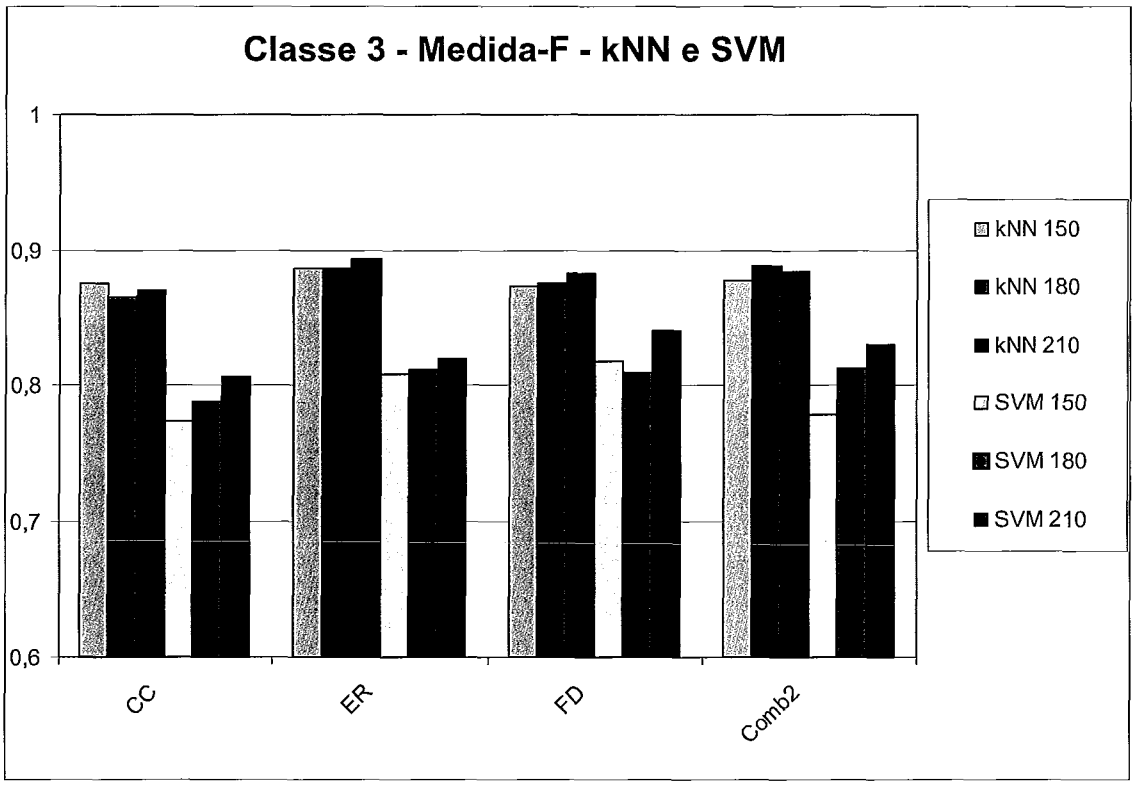


Figura 26 Classe 3 - Medida –F comparativa entre SVM e kNN

Na Figura 26, repete-se na classe 3 o ótimo desempenho do kNN em relação ao SVM, embora nessa classe os percentuais kNN estejam um pouco abaixo dos atingidos nas classes 1 e 2.

Em todas as classes os percentuais de medida-F foram superiores com o kNN dominantemente com o vetor de 210 termos.

4.6.3 – SVM, Naive Bayes e kNN – Análise dos Resultados Finais

De acordo com os resultados da Tabela 18 a teoria NFL descrita na seção 3.4.11 – não apresentou bons resultados. Isto não quer dizer que ela não funcione, e sim que neste corpus e com esses algoritmos, mesmo privilegiando cada classe, em detrimento de outras, o categorizador não obteve bons resultados.

Podemos avaliar empiricamente que para esta coleção o impacto da aplicação de tarefas específicas de pré-processamento antes da execução do algoritmo de classificação foi extremamente positivo. De acordo com os resultados apresentados, os melhores resultados e que pouco variaram independentemente do conjunto aplicado foram os obtidos nas combinações Comb2 e suas varições Peso-palavras Comb2 e Disc-Peso-palavras Comb2. Estas representam respectivamente a combinação de CC e ER, a

aplicação do peso das palavras sobre a combinação a combinação de CC e ER, e a discretização sobre esse peso conforme descrito na Figura 27.

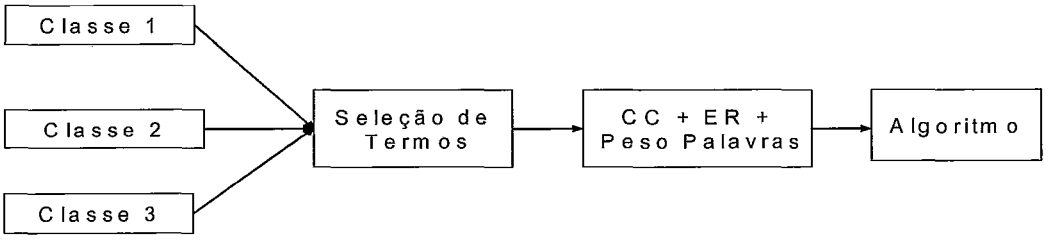


Figura 27 Arquitetura da melhor solução em todas as amostras

Conforme os dados alcançados com a coleção de 490 documentos pode ser observado que os resultados do SVM melhoraram em relação ao corpus inicial de 270 documentos. Esse resultado era esperado também para o Naive Bayes, que diferentemente, obteve resultados piores que os do conjunto original.

O tratamento na fase de pré-processamento ofereceu bons resultados com todos os categorizadores testados. Entretanto, os melhores resultados foram obtidos com o kNN que revelou até esse passo ser o método mais adequado para a solução.

Podemos observar que obtivemos um percentual de erro menor para o uso de 490 documentos e com 210 termos no vetor global, apenas com ER. Entretanto, a medida-F para as classes 1 e 2 foi melhor também utilizando a opção Comb2.

Cabe ressaltar que os resultados obtidos foram baseados em dados reais. Alguns autores sugerem que a utilização de massas de dados preparadas para tratamento de texto tais como as conhecidas coleções internacionais Yahoo, Reuters, Oshumed, TREC, e a nacional CETENF Folha de São Paulo, permite melhores índices de resultados (CAMARGO, 2007; MELO, 2007; SILVA, 2007; SILVA, MONTILHA et al., 2007; SILVA e VIEIRA, 2007).

4.6.4 – Protótipo para Categorização de Sigilo

Com base nos experimentos descritos, o protótipo desenvolvido para a MB apenas sugere a classe à qual o novo documento deve pertencer. A exemplo da ferramenta VISL (BICK, 2000; VISL, 2008), o documento é submetido via *upload.*

Devemos ressaltar que na construção do categorizador, utilizamos as ferramentas da fase 1 descritos na seção 4.4 –. A construção deste protótipo visou apenas exemplificar, para alguns usuários da MB, o uso do CADT na atribuição automática de sigilo.

A recomendação de sigilo é exibida na tela conforme demonstrado na Figura 28.

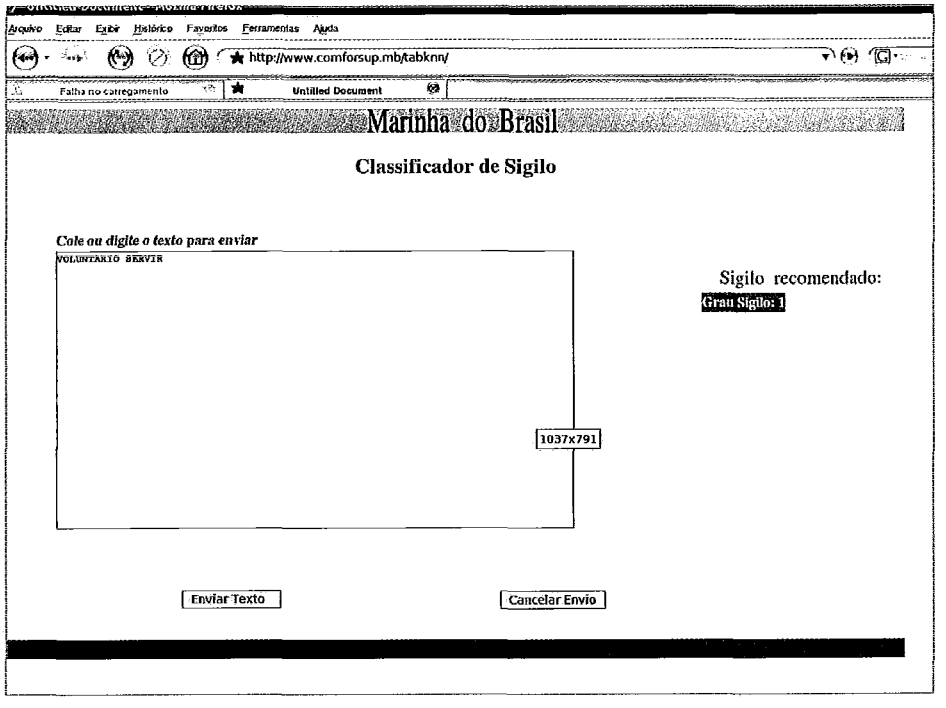


Figura 28 Tela do protótipo desenvolvido para a MB

4.7 – Analisando a Confiabilidade do Categorizador

Na realização de trabalhos com informações sensíveis, é recomendável que tenhamos alguma forma de analisar a vulnerabilidade a que essas informações podem estar expostas. Quanto menor for o grau de vulnerabilidade das informações, maior será a confiabilidade no uso do categorizador.

Em muitos domínios, os erros possuem "custos" diferenciados, como por exemplo, nos domínios de crédito onde o custo de fornecer um crédito incorretamente não é o mesmo da perda de não fornecer o crédito a um bom cliente. Outro exemplo é o domínio de detecção de fraudes, onde o custo de uma investigação inútil não é o mesmo custo de uma perda causada pela falta de investigação de uma fraude real.

No caso deste trabalho, de forma análoga aos exemplos citados acima, o "custo" de superclassificar um documento é diferente do custo de subclassificá-lo. No primeiro caso, da superclassificação, temos o "custo" como o manuseio diferenciado, de restrição de conhecimento e de armazenamento por mais tempo etc., conforme citado na

introdução deste trabalho. De acordo com o problema da excessiva atribuição de sigilo, já sabemos previamente que é mais fácil ocorrer o erro de uma mensagem que deveria ser "reservado" ter uma classificação "confidencial", do que o inverso.

Entretanto, o segundo caso, o de riscos de erro tendendo a uma subclassificação, ocorre o aumento do grau de vulnerabilidade à informação do documento. Assim, nosso "custo" é a diminuição da confiabilidade dos usuários no categorizador, o que poderia vir a desacreditar seu uso.

Tabela 23 Grau de Vulnerabilidade

		Estimado		
		Classe 1	Classe 2	Classe 3
Real	Classe 1	Baixo	Baixo	Baixo
	Classe 2	Médio	Baixo	Baixo
	Classe 3	Alto	Médio	Baixo

Na Tabela 23 apresentamos a relação com os níveis de vulnerabilidade com os resultados apresentados. Podemos observar que quanto mais alta a classe (ou sigilo) do documento, maior vulnerabilidade ela apresenta em relação à classe mais baixa. Assim, aqui o principal problema é o caso da subclassificação, onde quanto maior a distância entre os níveis de confidencialidade, maior será o impacto na vulnerabilidade. Ex: Classificar um documento "confidencial" como "reservado" é bem menos prejudicial que classificá-lo como "ostensivo".

De posse dessas informações executamos uma análise na matriz de confusão dos 2 melhores resultados com o algorimo kNN visando verificar a confiabilidade de acordo com a vulnerabilidade da corpus em questão descrita na Tabela 23.

```
=== Confusion Matrix ===
a  b  c  <-- classified as
433  32  25 |  a = OST
12 459  19 |  b = RES
20  47 423 |  c = CONF
```

Figura 29 Matriz deConfusão de kNN 210 Comb2

De acordo com os resultados apresentados na matriz de confusão exemplificada na Figura 29 podemos observar o comportamento de cada amostra de 490 documentos por classe. Na classe 1, dos 490 documentos foram corretamente categorizados 433, e

incorretamente categorizados 55 (32 + 25 soma dos FN) têm grau de vulnerabilidade baixo. Na classe 2 foram corretamente categorizados 459 documentos e incorretamente 19 com grau de vulnerabilidade baixo e 12 com grau médio. Já para a classe 3, foram corretamente classificadas 423 e incorretamente 59 com grau baixo e 20 com grau alto.

Tabela 24 Análise dos erros e seus graus de vulnerabilidade na matriz de confusão

Vulnerabilidade	Qtde Doctos	Percentual
Grau baixo	76	5,17
Grau Médio	59	4,01
Grau Alto	20	1,36
Total	155	10,54

Para melhor compreensão da matriz de confusão colocamos os percentuais de erro dispostos na Tabela 24. Esse percentuais foram obtidos aplicando-se a tabela de graus de vulnerabilidade sobre a matriz de confusão.

Tabela 25 Graus de Vulnerabilidade dos melhores resultados

	kNN		SVM	
	ER	Comb2	ER	Comb2
% Erro total	10,14	10,54	14,15	14,08
% Grau Baixo	4,90	5,17	7,41	7,82
% Grau Médio	3,81	4,01	6,26	5,71
% Grau Alto	1,43	1,36	0,48	0,54

Do mesmo modo como foi realizada a análise de erros e graus de vulnerabilidade apresentada na Tabela 25, fizemos um comparativo dos 4 melhores resultados obtidos na realização desse trabalho, sendo os 2 com o SVM e 2 com o kNN, na Tabela 25 Graus de Vulnerabilidade dos melhores resultados Esses resultados forma obtidos com 490 documentos por classe, com vetores de 210 termos e 20-fold. Devemos lembrar que o menor percentual de 0,48 representa em termos quantitativos 7 documentos em 490, enquanto 1,43 representam 21 documentos. Já para os documentos com baixo grau de vulnerabilidade saltam de 72 para 115 no pior caso (SVM-Comb2).

Há que ser considerada uma forma de atribuição de pesos ou "custo" para cada grau de vulnerabilidade. Deve-se ainda lembrar que o aumento nos FN e FP incorre em um aumento na taxa de erro do categorizador, o que a priori não é desejável. O estabelecimento de *threshold* por grau de vulnerabilidade em cada classe também pode ser utilizado. Todas essas questões relativas a confidencialidade e a matrizes sensíveis ao custo por si só constituem um campo para estudos e não foram considerados aqui.

Capítulo 5 – Conclusões e Trabalhos Futuros

Em consonância com orientações do governo federal (CASA CIVIL, 2002b), esse trabalho visa evitar que, a princípio, no âmbito das mensagens da MB, novos documentos recebam grau de sigilo maior que o necessário. O trabalho foi realizado com uma base de dados do setor operativo, considerado sensível, do ponto de vista da segurança das informações. Tratamos textos sigilosos e não sigilosos e algumas possibilidades para categorizá-los automaticamente.

5.1 – Conclusões

Durante a realização deste trabalho, foi descrito o problema da atribuição excessiva do sigilo que ocorre no âmbito do governo federal de um modo geral. Para a melhor compreensão da relevância do tema, foi descrita com um maior detalhamento a importância deste trabalho e do Corpus Mensagem, para a MB e que foi considerada uma boa amostragem dos assuntos que tramitam com regularidade. Foram revistos a descoberta de conhecimento em textos, as técnicas aplicadas e os algoritmos utilizados em categorização de documentos de uma forma geral.

Foram também descritas as técnicas de classificação implementadas, detalhando os algoritmos kNN, Naive Bayes e SVM, utilizados no trabalho, como também foi realizado um detalhamento sobre as métricas de avaliação utilizadas nesse tipo de processo.

No capítulo sobre a Proposta de CADT na Atribuição de Sigilo, fez-se uma descrição sobre o caminho percorrido para chegar à idéia a ser trabalhada nessa dissertação. Foram detalhados os programas utilizados para as etapas necessárias à solução do problema. Avaliou-se a eficácia dos algoritmos supracitados na resolução do problema proposto. A realização de experimentos e análise de seus resultados culminou com o desenvolvimento de protótipo de um categorizador para a MB.

Foram identificados textos que melhor representam a base para a realização dos testes com os categorizadores. Baseado em padrões identificados a partir de base já classificada, foi construído um protótipo de um categorizador para a MB que, a partir de um conjunto de regras, sugere o grau de sigilo de um novo documento. Esse protótipo usa o algoritmo kNN. A princípio, a intenção era evoluir o protótipo utilizando o Naive Bayes. Embora o SVM iniciasse apresentando melhores resultados, o Naive Bayes é

considerado mais barato do ponto de vista computacional, e em geral apresenta bom desempenho. Porém, após a realização dos testes finais, verificou-se, que os melhores resultados e que menos variaram dentro das classes, foram os alcançados pelo kNN. É importante ressaltar que o bom desempenho só se deu após tratamento dos textos na fase de pré-processamento.

Com este trabalho pode-se categorizar um novo documento, quanto ao sigilo, com um valor razoável de acerto. Estima-se que quanto mais o categorizador for utilizado, e conseqüentemente treinado, melhores serão os resultados. Entretanto, cabe ressaltar que esta categorização realizada foi feita baseada nos padrões atuais de atribuição de sigilo da MB. Embora isso possa ser feito de forma totalmente automática, na prática, a intervenção humana para a ratificação da atribuição ainda deverá ser utilizada.

Assim, as maiores contribuições deste trabalho são a demonstração da possibilidade de categorizar mensagens automaticamente quanto ao sigilo, e a definição do uso do categorizador kNN, nessa tarefa, como a melhor opção dentre os categorizadores estudados.

5.2 – Trabalhos futuros

Como o reconhecimento de padrões deste tipo de coleção possui uma grande importância estratégica, é possível evoluir o protótipo para um categorizador que poderá integrar o SiGDEM (Sistema de Gerência de Documentos Eletrônicos da Marinha) e ou SGC (Sistema Gerenciador de Comunicações) na tramitação das mensagens. Assim, poder-se-á melhorar a aplicação do grau de sigilo também para outros documentos de secretaria e operativos.

Outro ponto a ser abordado diz respeito ao formato dos documentos tratados. Seria interessante que futuras implementações considerassem outros formatos de documentos a serem trabalhados, como por exemplo, o formato .doc do MS-Word. Muitas das coleções do mundo real e na MB encontram-se nesse modo. O tratamento de textos diagramados é um pouco mais trabalhoso devido aos seus respectivos caracteres de controle, figuras, gráficos, formatação, erros de digitação e grafia. Embora se possa excluir os caracteres de controle incluindo-os na *stoplist*, o número de caracteres imprevisíveis é grande, o que, invariavelmente aumenta a taxa de erro do categorizador. Por outro lado, passar todas as bases das empresas para texto plano, embora possível, não parece ser uma solução adequada. Deve-se levar em consideração que novos

documentos vão sendo acrescentados, e necessitariam também ser submetidos constantemente a conversões.

Cabe ressaltar que a solução aqui encontrada pode ser utilizada em situações similares como para verificar precedência das mensagens. Além da minimização de erros nos novos documentos, em uma etapa futura, poderão ser revistas massas de documentos arquivados, auxiliando as comissões de triagem na desclassificação e eliminação dos documentos. Entretanto, deve-se lembrar que esse "lixo digital" pode servir para outros tipos de extrações de conhecimento. Podemos citar o reconhecimento de padrões referentes às áreas de logística e operações, com a categorização, por exemplo, de textos que dizem respeito à condição de eficiência do navio. Assim, novas funcionalidades podem ser extraídas da mesma aplicação de categorização automática de documentos.

Foram identificadas possibilidades de melhorias na atribuição do grau de sigilo na MB. Dentre elas podemos citar a identificação de quem atribuiu o Grau Sigilo e o setor a que pertence. Outra das possibilidades que podem ser exploradas é a colocação do sigilo em função do tempo. Por exemplo, antes de ocorrer uma operação naval, detalhes do transporte de material, ou de autoridades que participarão de determinado evento podem ser um assunto com certo grau de sigilo. Depois de ocorrido, ou seja, em outra fatia do tempo, a mesma informação, ou no caso, o mesmo documento, já não possui tanta importância. Podem-se citar outras situações como a de o assunto já ter sido divulgado na mídia, já haver ocorrido o evento, entre outras.

Uma boa prática adotada nos EUA é a utilização de um órgão, o Information Secutiry Oversight Office (ISOO, 2007) que cuida e define com um pouco mais de precisão assuntos que podem ser sigilosos ou não. Assim, já existe inclusive uma pré-definição detalhada de níveis de classificação, que auxilia a atribuição do sigilo para assuntos tais como planos militares, informações de governos estrangeiros, fontes de inteligência, métodos criptológicos, projetos nucleares, capacidades e vulnerabilidades etc.

Vários outros trabalhos podem ser desenvolvidos a partir da construção do categorizador para a MB. Com o descobrimento do padrão de atribuição de sigilo, os patamares dessa atribuição poderão ser reduzidos. As verificações de distorções nas atribuições do sigilo adicionalmente podem ser usadas para sugerir procedimentos na atribuição de sigilo para as Organizações Militares.

Com relação à confiabilidade do uso do categorizador, seria de grande utilidade que fossem efetuados estudos sobre as questões que envolvem as matrizes sensíveis ao custo. Assuntos como a atribuição de pesos ou "custo" para cada grau de vulnerabilidade e o estabelecimento de *threshold* para cada grau, deverão ser aprofundados no intuito de aumentar a confiabilidade de sua utilização na instituição.

Conforme citado anteriormente, a área de CADT está em franco desenvolvimento, existindo muitos campos em aberto e possibilidades de auxílio na automatização de processos de documentos textuais. Assim, esse trabalho não pretende ser conclusivo, funcionando, a propósito, como um ponto de partida para a discussão sobre atribuições de sigilo na MB e sobre o tratamento do mesmo através da categorização de documentos.

Referências

AHA, D., KIBLER, D., ALBERT, M., 1991, "Instance-Based Learning Algorithms", *Mach.Learn.*, v. 6, n. 1, pp. 37-66.http://portal.acm.org/citation.cfm?id=104717

AIZAWA, A., 2001, "Linguistic Techniques to Improve the Performance of Automatic Text Categorization", Tokyo, JP

ALASDAIR, R., 2006, "Blacked out: government secrecy in the information age". *Cambridge University Press*, New York

ALMEIDA, C. W. D., 2005, "Transparência do orçamento de defesa - O Caso Brasileiro", Papeles de Investigacion RESDAL, Buenos Aires, Argentina.

ANTONIE, M., ZAÏANE, O. R., 2002, "Text Document Categorization by Term Association", ICDM. IEEE Computer Society, Washington, DC, USA

ARANHA, C., PASSOS, E., 2006, "A Tecnologia de Mineração de Textos", Rio de Janeiro, Brasil

BAEZA-YATES, R., 2004, "Challenges in the Interaction of Information Retrieval and Natural Language Processing".

BAOLI, L., SHIWEN, Y., QIN, L., 2003, "An Improved k-Nearest Neighbor Algorithm for Text Categorization", Shenyang, China

BEKKERMAN, R.,ALLAN, J., 2003, *Using Bigrams in Text Categorization.* CIIR Technical Report IR-408, 2004

BERG, C. N., 1997, "Developing a Corpus Specific Stoplist Using Quantitative Comparison". *Storming Media*, AIR FORCE INST OF TECH WRIGHT-PATTERSONAFB OH SCHOOL OF ENGINEERING,USA

BICK, E., 2000, "*The parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*". *Arhus University Press*, Denmark

BICK, E., 2003, "*A Constraint Grammar Based Question Answering System for Portuguese.* ". *LNAI SpringerVerlag*, Proceedings of 11° Portuguese Conference on Artificial Intelligence

BIRD, S., LOPER, E., 2004, "NLTK: The Natural Language Toolkit". *Association for Computational Linguistics*, Barcelona

BLOEHDORN, S., HOTHO, A., 2004, "Text Classification by Boosting Weak Learners based on Terms and Concepts".http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/proceedings/&toc=comp/proceedings/icdm/2004/2142/00/2142toc.xml&DOI=10.1109/ICDM.2004.10077

BLUM, A. L.,LANGLEY, P., 1997, "Selection of relevant features and examples in machine learning", *Artificial Intelligence*, v. 97, n. 1, pp. 245-271.http://portal.acm.org/citation.cfm?id=270626

BOK, S., 1984, "Secrets: On the Ethics of Concealment and Revelation". *Vintage Books*, New York

CABRÉ, M. T., BAGOT, R. E., PLATRES, J. V., 2001, "Automatic term detection: A review of current systems". *Jonh Benjamins Publishing Company*, France

CAMARGO, Y. B. L., 2007, *ABORDAGEM LINGUÍSTICA NA CLASSIFICAÇÃO AUTOMÁTICA DE TEXTOS EM PORTUGUÊS*, Tese de Dissertação de M.Sc., Programa de Engenharia Elétrica - COPPE - UFRJ, Rio de Janeiro

CANCEDDA, N., GAUSSIER, E., GOUTTE, C., et al., 2003, "Word-Sequence Kernels".http://citeseer.ist.psu.edu/cancedda02wordsequence.html

CASA CIVIL, 2002, "DECRETO N. 4.553 - DE 27 DE DEZEMBRO DE 2002 ". *Presidência da República - Casa Civil*, D.O.U. - Diário Oficial da União; Poder Executivo, de 30 de dezembro de 2002

CASTRO, P. F. D., 2000, "Categorização Automática de Textos", Rio de Janeiro, Brasil.

CAVNAR, W. B., 1994, "Using an N-Gram-Based Document Representation With a Vector Processing Retrieval Model". *TREC*, NIST SP 500-225, p. 269-77

CENADEM, 2005, "Banco de Cases do CENADEM - Área Governo Federal". *acessado em 15/02/2008.*, http://www.cenadem.com.br/bcases_gov_federal.php.

CHANG, C. C., LIN, C. J., 2000, "LIBSVM: a Library for Support Vector Machines (Version 2.31)".http://citeseer.ist.psu.edu/chang01libsvm.html

CHUNG, L., NIXON, B., YU, E., et al, 1999, *Non-Functional Requirements in Software Engineering (THE KLUWER INTERNATIONAL SERIES IN SOFTWARE ENGINEERING Volume 5) (International Series in Software Engineering)*, Springer.http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0792386663

COVER, T.,HART, P., 1967, "Nearest neighbor pattern classification", *Information Theory, IEEE Transactions on*, v. 13, n. 1, pp. 21-27.http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1053964

COVER, T. M., HART.P.E., 2008, "Nearest Neighbor Classifiers"

CUNNINGHAM, P. +., DOYLE, D. +., LOUGHREY, J., 2003, "An Evaluation of the Usefulness of Case-Based Explanation".http://dx.doi.org/10.1007%2F3-540-45006-8_12

DAHLBERG, I., 1992, "Knowledge organization and terminology : philosophical and linguistic bases", Bulgária

DASGUPTA, A., DRINEAS, P., HARB, B., et al, 2007, "Feature selection methods for text classification", pp. 230-239,ACM.http://portal.acm.org/citation.cfm?id=1281192.1281220

DE CASTRO, L. N.,VON ZUBEN, F. J., 2002, "Learning and optimization using the clonal selection principle", *Evolutionary Computation, IEEE Transactions on*, v. 6, n. 3, pp. 239-251.http://dx.doi.org/10.1109/TEVC.2002.1011539

DEBOLE, F., SEBASTIANI, F., 2002, "Supervised Term Weighting for Automated Text Categorization"

DEERWESTER, S., DUMAIS, S., LANDAUER, T., et al, 1990, "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science*, v. 41, n. 6, pp. 391-407.http://citeseer.ist.psu.edu/deerwester90indexing.html

DEFESA, M. D., 2005, "Regimento Interno do Comando da Marinha", Ministério da Defesa, Decreto nº 5.417.

DGMM, 2005, "Procedimentos de Comunicações - DGMM-0550 (revisão-4)", Diretoria Geral de Material da Marinha, Rio de Janeiro, Brasil.

DIAO, Y., LU, H., WU, D., 2000, "A Comparative Study of Classification Based Personal E-mail Filtering". *Lecture Notes In Computer Science. Springer-Verlag*, London

DOOREY, T. J., 2007, "Intelligence Secrecy and Transparency: Finding the Proper Balance from the War of Independence to the War on Terror". *Center for Contemporary Conflict at the Naval Postgraduate School*, Monterey, California

DORRE, J., GERSTL, P., SEIKERT, R., 1999, "Text mining: Finding nuggets in mountains of textual data", San Diego,USA

DU BOYLAY, J., 1976, "Lies, Mockery, and Family Integrity in MEDITERRANEAN FAMILY STRUCTURES". *Cambridge University Press*, Cambridge, UK

DUDA, R., HART, P., STORK, D., 2002, "Pattern Classification". *Willey Interscience*

DUDA, R.,HART, P., 1973, *Pattern Classification and Scene Analysis*, {John Wiley & Sons Inc}.http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0471223611

ELLINGTON, T. C., 2004, "Official Secrecy: Self, State and Society". *PHD Thesi, I University of Maryland, EUA.*

EMA, 2002, "Manual de Publicações da Marinha - EMA-411", Estado Maior da Armada, Brasília, DF, Brasil.

EMA, 2005, "Normas para Salvarguarda de Materiais Controlados, Dados,Informações, Documentos e Materiais Sigilosos na Marinha - EMA-414", Estado Maior da Armada, Brasília, DF, Brasil.

FAGNI, T., SEBASTIANI, F., 2007, "On the Selection of Negative Examples for Hierarchical Text Categorization", Poznan, PL

FAYYAD ET AL., 1996, "Advances in knowledge discovery and data mining", AAAI Press/MIT Press

FELDMAN, R., DAGAN, I., 1995, "Knowledge Discovery in Textual Databases (KDT)". *In First international conference on knowledge discovery (KDD'95)*, Montreal, Canada

FELDMAN, R.,SANGER, J., 2006, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, {Cambridge University Press}.http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0521836573

FORMAN, G., 2002, "Choose Your Words Carefully: An Empirical Study of Feature Selection Metrics for Text Classificatio", Helsinki, Finland.http://citeseer.ist.psu.edu/675643.html

FOX, C., 1989, "A stop list for general text"

FREITAS, R. A., QUINTANILLA, L. W., NOGUEIRA, et al., 2004, "Portais Corporativos: uma ferramenta estratégica para a Gestão do Conhecimento". *Brasport*, Rio de Janeiro

FRIEDMAN, N., GEIGER, D., GOLDSZMIDT, M., 1997, "Bayesian Network Classifiers", *Machine Learning*, v. 29, n. 2-3, pp. 131-163.http://citeseer.ist.psu.edu/friedman97bayesian.html

FURTADO, M. I. V., 2004, "Inteligência competitiva para o ensino superior privado: Uma abordagem através da mineração de textos", Brasil

GABRILOVICH, E.,MARKOVITCH, S., 2007, "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis", *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 6-12

GOLDBERG, D., 1989, *Genetic Algorithms in Search, Optimization, and Machine Learning*,{Addison-Wesley Professional}.http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0201157675

HARMAN, D., 1991, "How effective is suffixing"

HAYKIN, S., 2000, "Redes Neurais: Princípios e Práticas". *Ed. BOOKMAN*

HEARST, M., 1999, "Untangling text data mining".http://citeseer.ist.psu.edu/hearst99untangling.html

HOLANDA, A. B., 1999, "Dicionário Aurélio Eletrônico". *Nova Fronteira e Lexicon Informática*, Rio de Janeiro, RJ. Brasil

HOTHO, A., NURNBERGER, A., PAASS, G., 2005, "A Brief Survey of Text Mining"

HULL, D., 1998, "Stemming algorithms: A case study for detailed evaluation", *Journal of the American Society for Information Science*, v. 47, n. 1, pp. 70-84.http://dx.doi.org/10.1002/(SICI)1097-4571(199601)47:1%3C70::AID-ASI7%3E3.0.CO;2-%23

ISOO, 2007, "Marking Classified National Security Information", *www.archives.gov/isoo/training/**marking**-booklet.pdf*, acessado em 10/03/2008

JOACHIMS, T., 1998a, "Making large-scale support vector machine learning practical". In olkopf, C. B., *Advances in Kernel Methods: Support Vector Machines*, MIT Press, Cambridge, MA.http://citeseer.ist.psu.edu/joachims98making.html

JOACHIMS, T., 1998b, "Text categorization with support vector machines: learning with many relevant features", pp. 137-142,Springer Verlag, Heidelberg, DE.http://citeseer.ist.psu.edu/141654.html

KAO, A., POTEET, S. R., 2007, "Natural Language Processing and Text Mining". *Springer-Verlag*

KARANIKAS, H., THEODOULIDIS, B., 2002, "Knowledge Discovery in Text and Text Mining Software", Manchester, UK

KEIKHA.M., RAZAVIAN, N. S., OROUMCHIAN, F., et al., 2008, "Survey of Text Mining II - Document Representation and Quality of Text: An Analysis". *Springer*, London

KELLER, M., 2006, "Machine Learning Approaches to Text Representation using Unlabeled Data". *Ecole Polytechnique Federale de Lausanne*, Lausanne,FR

KIBRIYA, A., FRANK, E., PFAHRINGER, B., et al., 2004, "Multinomial Naive Bayes for Text Categorization Revisited"

KOHAVI, R., PROVOST, F., 1998, "On Applied Research in Machine Learning". *Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, Boston.

KOHAVI, R., 1995, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", pp. 1137-1145.http://citeseer.ist.psu.edu/kohavi95study.html

KROEZE, J. H., MATCHDEL, M. C., ., et al., 2003, "Differentiating data and text-mining terminology". *South African Institute for Computer Scientists and Information Technologists*, Republic of South Africa

LAGUS, K., 2000, "Text Mining with the WEBSOM". *Department of Computer Science and Engineering*, University of Technology (Espoo, Finland)

LENNON, M., PIERCE, D., TARRY, B., et al., 1981, "An evaluation of some conflation algorithms for information retrieval"

LEONARD, J. W., 2004, "Formal Statement". *Information Security Oversight Office - National Archives and Records Administration,* http://www.archives.gov/isoo/speeches-and-articles/formal-statement-august-2004.html, acessado em 30/03/2008

LEONARD, J. W., 2007, "Formal Statement" *Information Security Oversight Office - National Archives and Records Administration, intelligence.house.gov/Media/PDFS/Leonard071207.pd,* , acessado em 30/03/2008

LEWIS, D., 1992, "An evaluation of phrasal and clustered representations on a text categorization task". *ACM Press*, New York, US.

LEWIS, D., 1998, "Naive (Bayes) at forty: The independence assumption in information retrieval", pp. 4-15,Springer Verlag, Heidelberg, DE.http://citeseer.ist.psu.edu/397172.html

LI, Y. H.,JAIN, A. K., 1998, "Classification of Text Documents", *The Computer Journal*, v. 41, n. 8, pp. 537-546.http://dx.doi.org/10.1093/comjnl/41.8.537

LIAO, Y., VEMURI, V., 2002, "Using Text Categorization Techniques for Intrusion Detection", San Francisco, CA

LIU, J.,LIANG, C., 2008, "Text Categorization of Multilingual Web Pages in Specific Domain".http://dx.doi.org/10.1007/978-3-540-68125-0_96

LOUREIRO, S. M., MARGOTO, L. R., VAREJÃO, F. M., et al., 2005, "Um mecanismo automático para busca de parâmetros de técnicas de classificação utilizando algoritmos genéticos". *XXV Simpósio Brasileiro de Computação - SBC*, V Encontro Nacional de Inteligência Artificial - ENIA

LOVINS, J. B., 1968, "Development of a stemming algorithm"

MANNING, C.,SCHTZE, H., 1999, *Foundations of Statistical Natural Language Processing*,TheMIT Press.http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0262133601

MARON, M. E.,KUHNS, J. L., 1960, "On Relevance, Probabilistic Indexing and Information Retrieval", *J.ACM*, v. 7, n. 3, pp. 216-244.http://portal.acm.org/citation.cfm?id=321035

MELO, L. B. S. D., 2007, "Reconhecimento de Padrões Textuais para Categorização". *COPPE*, Rio de Janeiro-Brasil.

MITCHELL, T. M., 1997, "Machine Learning". *Ed. McGraw-Hill*, USA

MLADENIC, D., 1998, "Machine Learning on Non-homogeneous Distributed Text Data". *DSc Thesis. University of Ljubljana. Faculty of Computer and Information Science.*

MOHAMMAD, M. A., 2007, "Text Mining :A Burgeoning Quality Improvement Tool", The Graduate School of Applied Mathematics of the Middle East Technical University, Ankara, Turkey

MONTEJO-RÁEZ, A., 2005, "Automatic Text Categorization in High Energy Physics Domain", Editorial da Universidade de Granada. Spain.

MULLER, K. R., MIKA, S., RATSCH, G., et al, 2001, "An introduction to kernel-based learning algorithms", *Neural Networks, IEEE Transactions on*, v. 12, n. 2, pp. 181-201.http://dx.doi.org/10.1109/72.914517

NÆSS, A. B., 2007, "Bayesian Text Categorization". *Norwegian University of Science and Technology*, Master Thesis on Science in Physics and Mathematics

NG, H., GOH, W., LOW, K., 1997, "Feature selection, perceptron learning, and a usability case study for text categorization", pp. 67-73,ACM Press, New York, US.http://citeseer.ist.psu.edu/404607.html

ORENGO, V. M., HUYCK, C. R., 2001, "A stemming algorithim for the portuguese language", *Laguna de San Raphael*

PLATT, J., 1999, "Fast training of support vector machines using sequential minimal optimization", pp. 185-208.http://portal.acm.org/citation.cfm?id=299105

POLAT, K., GÜNES, S., 2006, "The effect to diagnostic accuracy of decision tree classifier of fuzzy and k-NN based weighted pre-processing methods to diagnosis of erythemato-squamous diseases"

PORTER, M. F., 1980, "An algorithm for suffix stripping", *Program: electronic library & information systems*, v. 40, n. 3, pp. 211-218.http://dx.doi.org/10.1108/00330330610681286

PRESSMAN, R., 2001, *Software Engineering: A Practitioner's Approach*, {McGraw-Hill.http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-0&path=ASIN/0072496681

PRUDÊNCIO, R., LUDERMIR, T., 2007, "Aprendizagem Ativa para Seleção de Exemplos em Meta-Aprendizado". *Congresso da SBC -- Interação entre as Ciências: Desafio para a Tecnologia da Informação, Rio de Janeiro, Brasil.*

QUINLAN, J. R., 2003, "Induction of Decision Trees", *Mach.Learn.*, v. 1, n. 1, pp. 81-106.http://portal.acm.org/citation.cfm?id=637969

RABINER, L. R., 1989, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, v. 77, n. 2, pp. 257-286.http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=18626

RÁEZ, A., 2007, "Using linguistic information as features for text categorization", Italy

REINHARDT, F., 1999, "Market Failure and the Environmental Policies of Firms: Economic Rationales for "Beyond Compliance" Behavior", Harvard Business School, Boston, MA, USA

RESDAL, 2008, "Red de Seguridad y Defensa de América Latina"

ROBERTSON, S. E., 1997, "The probability ranking principle in IR", pp. 281-286.http://portal.acm.org/citation.cfm?id=275701

ROBERTSON, S. E.,JONES, S., 1976, "Relevance weighting of search terms", *Journal of the American Society for Information Science*, v. 27, n. 3, pp. 129-146.http://dx.doi.org/10.1002/asi.4630270302

ROCCHIO, J. J., 1966, "Document Retrieval Systems-Optimizationand Evaluation. PhD thesis.", Cambridge, MA

RUSSELL, S.,NORVIG, P., 2003, *Artificial Intelligence, 2nd ed*, Prentice Hall

SALTON, G., WONG, A., YANG, C. S., 1975, "A vector space model for automatic indexing", *Commun.ACM*, v. 18, n. 11, pp. 613-620.http://portal.acm.org/citation.cfm?id=361220

SALTON, G.,BUCKLEY, C., 1988, "Term-weighting approaches in automatic text retrieval", *Inf.Process.Manage.*, v. 24, n. 5, pp. 513-523.http://portal.acm.org/citation.cfm?id=54260

SARDINHA, T. B., 2004, "Linguística de Corpus", São Paulo, Brasil.

SCHUTZE, H., HULL, D., PEDERSEN, J., 1995, "A Comparison of Classifiers and Document Representations for the Routing Problem", pp. 229-237.http://citeseer.ist.psu.edu/schutze95comparison.html

SEBASTIANI, F., 2002a, "Automated Text Categorization:Tools, Techniques and Applications". *Centre National de Recherche Technologique,Rennes, France* , Istituto di Elaborazione della Informazione,Consiglio Nazionale delle Ricerche 56124 Pisa, Italy

SEBASTIANI, F., 2002b, "Machine Learning in Automated Text Categorization"

SEBASTIANI, F., 2006, "Classification of Text, Automatic". *Elsevier*, Universitá di Padova,Padova, Italy

SHETH, B., 1994, *A Learning Approach to Personalized Information Filtering*.http://citeseer.ist.psu.edu/112107.html

SILVA, A. A. S., 2007, "AÎURI: UM PORTAL PARA MINERAÇÃO DE TEXTOS INTEGRADO A GRIDS COMPUTACIONAIS", Dissertação de Mestrado, COPPE/UFRJ, RJ. Brasil

SILVA, B. C., MONTILHA, G., MACHADO, L. H. R., et al., 2007, "Introdução ao Processamento das Línguas Naturais e Algumas Aplicações", NILC - ICMC-USP - São Carlos, SP, Brasil

SILVA, C. F., VIEIRA, R., 2007, "Categorização de Textos da Língua Portuguesa com Árvores de Decisão, SVM e Informações Lingüísticas". *XXVII Simpósio da Sociedade Brasileira de Computação*, Rio de Janeiro, Brasil

SIMMEL, G., 1906, "The Sociology of Secrecy and of Secret Societies". *The University of Chicago Press*, USA

SOMMERVILLE, I., 2006, *Software Engineering: (Update) (8th Edition) (International Computer Science Series)*, {Addison Wesley}.http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0321313798

SPINAKIS, A., 2005, "Text Mining: A Powerful Tool for Knowledge Management". *QUANTOS SARL*, Greece

TAN, A. H., 1999, "Text mining: The state of the art and the challenges"

TAVARES, M. D. G. P., 1991, "Cultura organizacional: uma abordagem antropológica da mudança". *Ed. Qualitymark*, São Paulo, Brasil.

TEFFT, S., 1980, "Secrecy: a cross-cultural perspective". *Human Sciences Press*, New York

TEICHERT, T.,MITTERMAYER, M. A., 2002, "Text mining for technology monitoring", v. 2, pp. 596-601

TOMAULT, T., 2006, "kNN, Rocchio and Metrics for Information Filtering at TREC-10".http://citeseer.ist.psu.edu/611385.html

VALENTINI, G., DIETTERICH, T. G., 2002, "Bias-Variance Analysis and Ensembles of SVM". *Springer-Verlag*, London, UK

VAN RIJSBERGEN, C. J., 1979, *Information Retrieval, 2nd edition*, Dept. of Computer Science, University of Glasgow.http://citeseer.ist.psu.edu/vanrijsbergen79information.html

VAPNIK, V. N., 1999, "An overview of statistical learning theory", *Neural Networks, IEEE Transactions on*, v. 10, n. 5, pp. 988-999.http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=788640

VISL, 2008, "Visual Interactive Syntax Learning - Portuguese"

WALTON, D., 1996, "Plausible Deniability and Evasion of Burden of Proof". *Kluwer Academic Publishers. Printed in the Netherlands., Department of Philosophy - University of Winnipeg- CA*

WANG, H., 2006, "Nearest neighbors by neighborhood counting". *IEEE Transactions*

WANG, T., CHIANG, H., 2007, "Fuzzy support vector machine for multi-class text categorization". *Inf. Process. Manage.*

WEISS, S., INDURKHYA, N., ZHANG, T., et al, 2004, *Text Mining: Predictive Methods for Analyzing Unstructured Information*, Springer.http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0387954333

WIENER, E., PEDERSEN, J., WEIGEND, A., 1995, "A neural network approach to topic spotting", pp. 317-332.http://citeseer.ist.psu.edu/wiener95neural.html

WILSON, R.,MARTINEZ, T., 1997, "Improved Heterogeneous Distance Functions", *Journal of Artificial Intelligence Research*, v. 6, pp. 1-34.http://citeseer.ist.psu.edu/wilson97improved.html

WITTEN, I.,FRANK, E., 2005, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*,http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&amp

WOLPERT, D. H.,MACREADY, W. G., 1997, "No free lunch theorems for optimization", *Evolutionary Computation, IEEE Transactions on*, v. 1, n. 1, pp. 67-82.http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=585893

WURST, M., 2007, "The Word & Web Vector Tool - WVTOOL"

YAMADA, T., YAMASHITA, K., ISHII, K., et al., 2006, "Text Classification by Combining Different Distance Functions with Weight"

YANG, Y.,LIU, X., 1999, "A re-examination of text categorization methods", pp. 42-49.http://citeseer.ist.psu.edu/361822.html

YANG, Y., ZHANG, T., KISIE, B., 2003, "A Scalability Analysis of Classifiers in Text Categorization", Toronto,Canadá

YANG, Y.,PEDERSEN, J., 1997, "A comparative study on feature selection in text categorization", pp. 412-420,Morgan Kaufmann Publishers, San Francisco, US.http://citeseer.ist.psu.edu/yang97comparative.html

YANG, Y., WEBB, G., 2002, "A Comparative Study of Discretization Methods for Naive-Bayes Classifiers".http://citeseer.ist.psu.edu/588336.html

YU, J., WANG, S., XI, L., 2007, "Evolving artificial neural networks using an improved PSO and DPSO", *Neurocomputing*, v. 71- In Press, Corrected Proof, pp. 4-6.http://dx.doi.org/10.1016%2Fj.neucom.2007.10.013

YUA, B., XUB, Z., LI, C., 2008, "Latent semantic analysis for text categorization using neural network". *Elsevier*, Xi'an Jiaotong University, Xi'an 710049, China

ZEREDO, L. L., 2002, "TEXT MINING: Encontrando Sentindo na Teia de Informações Não-estruturadas". *SBGC*, São Paulo, SP.

ZHANG, H., BERG, A. C., MAIRE, M., et al, 2006, "SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition", v. 2, pp. 2126-2136.http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1641014