RECONHECIMENTO DE PADRÕES TEXTUAIS PARA CATEGORIZAÇÃO AUTOMÁTICA DE DOCUMENTOS

Laila Beatriz Soares Melo

DISSERTAÇÂO 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 ELÉTRICA.

Aprovada por:

_____
Prof. Jorge Lopes de Souza Leão, Dr.Ing

_____
Prof. Antonio Carneiro de Mesquita Filho, Dr.d'État.

_____
Prof.Geraldo Bonorino Xexéo, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

DEZEMBRO DE 2007

MELO, LAILA BEATRIZ SOARES

Reconhecimento de Padrões Textuais para Categorização Automática de Documentos [Rio de Janeiro] 2007

IX, 74p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia Elétrica, 2007)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Categorização de Textos

I. COPPE/UFRJ II. Título (série)

À Bruna, Claudia, Lourdes e Michel

Agradecimentos

- Ao Professor Jorge Lopes de Souza Leão pelo incentivo e orientação ao longo do trabalho;
- Aos amigos Fabiana, Henrique, Ítalo, Marcel, Milton, Newton, Rubens, Sergio, Yuri, Zé pelo apoio e companheirismo;
- Ao professor Antonio Carneiro de Mesquita Filho pelas contribuições com seus conhecimentos ao longo do trabalho;
- Ao professor Geraldo Bonorino Xexéo pelas contribuições com seus conhecimentos;
- Às crianças pela alegria e atenção que me dedicam.

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

RECONHECIMENTO DE PADRÕES TEXTUAIS PARA CATEGORIZAÇÃO AUTOMÁTICA DE DOCUMENTOS

Laila Beatriz Soares Melo

Dezembro/2007

Orientador: Jorge Lopes de Souza Leão

Programa: Engenharia Elétrica

Este trabalho propõe uma abordagem do problema de reconhecimento de padrões textuais aplicada ao processo de classificação automática de documentos. Foram utilizados dois conjuntos de textos: um voltado para a linguagem mais próxima do cotidiano e o outro voltado para a linguagem técnica-científica para permitir a avaliação da abordagem a diferentes tipos de textos. Foram usados dois tipos de classificadores, o Naïves Bayes e as Redes Neurais Artificiais, como métodos comparativos dos resultados obtidos.

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

TEXTUAL PATTERN RECOGNITION FOR AUTOMATIC DOCUMENTS CATEGORIZATION

Laila Beatriz Soares Melo

December/2007

Advisor: Jorge Lopes de Souza Leão

Department: Electrical Engineering

This work proposes an aproach to textual pattern recognition with application to the process of automatic document categorization. Two corpora of texts have been used, one of them next to the daily use and the other a more technical-cientific language, in orter to evaluate the efficiency of the approach to different types of texts. Two types of classifiers were used, the Naïve Bayes and the Artificial Neural Networks and their influence on the results was evaluated.

Índice:

Índice de Figuras:

1. Introdução

Diante da grande quantidade de informação textual existente atualmente em formato eletrônico, tanto na Internet (informações em geral, mensagens de correio, etc) como nas empresas (relatórios, documentação, etc) e diante de seu crescimento diário cada vez maior, a pesquisa em documentos relevantes tem-se tornado uma tarefa difícil, consumidora de tempo e muitas vezes improdutiva diante do que se deseja obter, rapidez e objetividade na obtenção de determinada informação.

O reconhecimento de padrões trata da classificação e da descrição dos objetos. O ser humano constantemente faz uso do reconhecimento de padrões ao reconhecer imagens, sons, etc. A leitura de um texto também é feita através de reconhecimento de padrões, pois através das palavras, frases, o texto é interpretado. O processamento da linguagem natural tem se mostrado uma tarefa difícil, pois ao contrário do ser humano que ao fazer uma leitura dispõe de outras informações (padrões) para compreender e classificar, a máquina tem acesso somente ao padrão textual obtido através de um conjunto de termos selecionados para representar um conceito desejado.

A área da Recuperação da Informação (RI) (BAEZA & YATES,1999) tem sido foco de diversas pesquisas, buscando técnicas automatizadas capazes de organizar e pesquisar documentos textos em linguagem natural (não estruturados), de maneira a obter a partir da consulta de um usuário a informação desejada para fins diferenciados tais como bibliotecas, armazenamento e disponibilidade de informações para setores específicos, disseminação de informações, seleção de documentos, etc.

Com a evolução da área da RI surgiu a Mineração de Textos (MT) ou *Text Mining* (TAN, 1999). Segundo Tan (1999), a área de MT trata da extração de padrões ou conhecimentos interessantes a partir de um conjunto de documentos textuais. Na área de MT, existem diversas abordagens técnicas para organização e extração da

informação, dentre elas, a Categorização Automática de Documentos de Textos (CADT).

A CADT é uma técnica de Processamento da Linguagem Natural (Russel & Norvig, 1995), a qual através de diferentes técnicas de Aprendizado de Máquina aplicadas a distintas coleções de documentos, procura extrair padrões úteis para organizar e recuperar informação dos textos. O processo de categorização de textos ou classificação automática de documentos, foi desenvolvido com o intuito de suprir as necessidades de separar a informação em categorias de conhecimentos, de maneira a permitir a manipulação e a recuperação destes. Quanto mais complexo for o processo de categorização, mais difícil e demorado será o tratamento da informação, obrigando a combinação de técnicas de análise de linguagem natural, recuperação de informação e métodos de análise de dados qualitativos. Atualmente a categorização automática de textos é de grande importância em áreas como, o tratamento e organização da informação de grandes organizações, a triagem e classificação de correio eletrônico, categorização de páginas da Web, etc.

Através da técnica CADT, é possível reduzir o foco da pesquisa de interesse dentro do grande volume de informações disponíveis, já que selecionar documentos dentro de uma base já pré-estabelecida é muito menos dispendioso (YANG & PEDERSON, 1997).

Em linhas gerais, a CADT é uma técnica utilizada para classificar um conjunto de documentos em uma ou mais categorias pré-definidas.(figura 1-1)

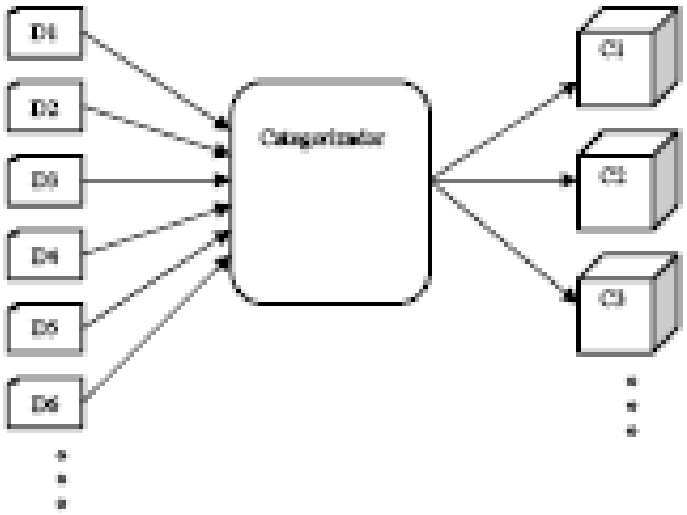


Figura 1-1 – Categorização Automática de Documentos de Textos (CADT)

A tarefa de categorização de textos pode ser dividida em cinco etapas principais (Figura 1- 2):

- Definição da Coleção de Documentos a serem classificados;
- Pré-processamento: consiste de um conjunto de ações para transformar a informação em linguagem não-estruturada (textos), em um conjunto de informações passíveis de serem entendidas para a extração do conhecimento.
- Seleção de características: consiste na seleção das palavras (termos) que melhor representem cada documento, fazendo uso de cálculos, técnicas ou métodos que melhor se apliquem à extração da informação;
- Classificação: consiste em determinar a que classe pertence cada documento, aplicando-se diferentes técnicas de aprendizado de máquina que realizem o reconhecimento de padrões;
- Interpretação dos Resultados.

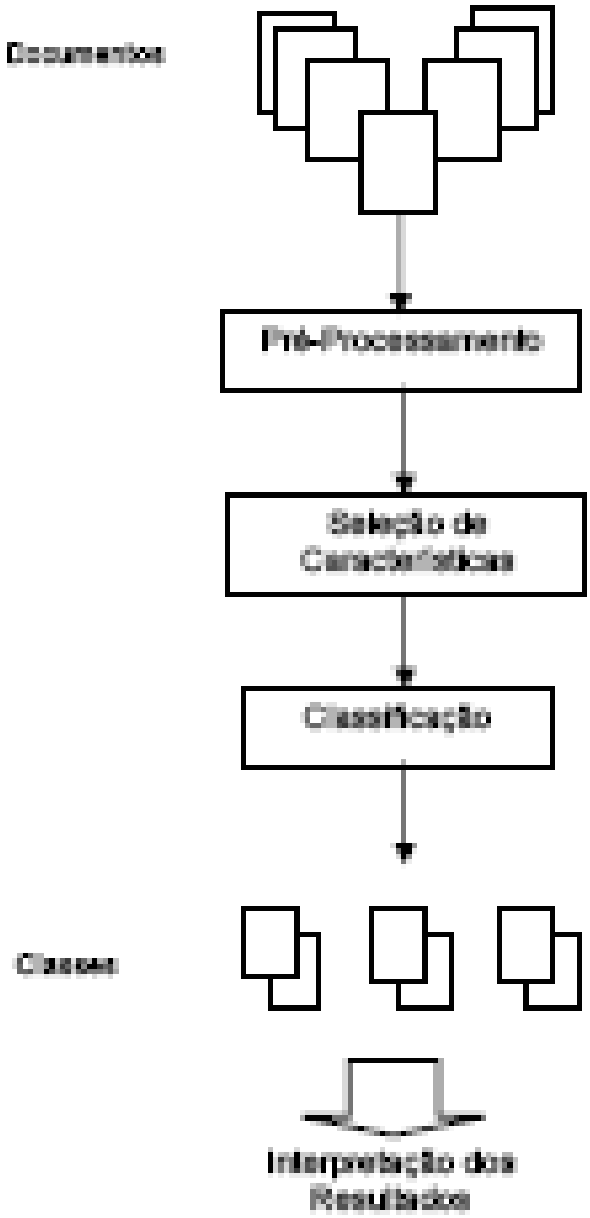


Figura 1- 2 – Etapas da Categorização Automática de Documentos de Textos

Este trabalho propõe uma abordagem do problema de reconhecimento de padrões textuais aplicada ao processo de categorização automática de documentos. Pretende-se avaliar a classificação automática de textos em português fazendo uso de informações lingüísticas para extração de termos, utilizando dois diferentes conjuntos

de documentos (corpora), dos quais um utiliza uma linguagem próxima do cotidiano (textos de jornal) e outro utiliza uma linguagem técnica-científica (textos de dissertações de mestrado e teses de doutorado).

Através de diferentes técnicas aplicadas ao cálculo de relevância dos termos, como escore de relevância e coeficiente de correlação, e diferentes métodos utilizados para a categorização, como o classificador Naïve Bayes e as Redes Neurais Artificiais do tipo Multi-layer Perceptron (MLP), será feita uma comparação dos resultados e avaliação da abordagem proposta.

Corpora utilizados: um é formado por textos do jornal Folha de São Paulo elaborados pelo NILC (Núcleo Interinstitucional de Lingüística Computacional) contendo 855 documentos de textos correspondentes às categorias *esportes*, *imóveis*, *informática*, *política* e *turismo*; o outro conjunto é formado por textos compostos por títulos e resumos das dissertações de mestrado e teses de doutorado da Engenharia Elétrica - COPPE/UFRJ contendo 475 textos correspondentes às categorias *controle*, *microeletrônica*, *processamento de sinais*, *redes de computadores* e *sistemas de potência*.

A Figura 1-3 mostra uma representação da proposta de trabalho com suas diferentes abordagens.

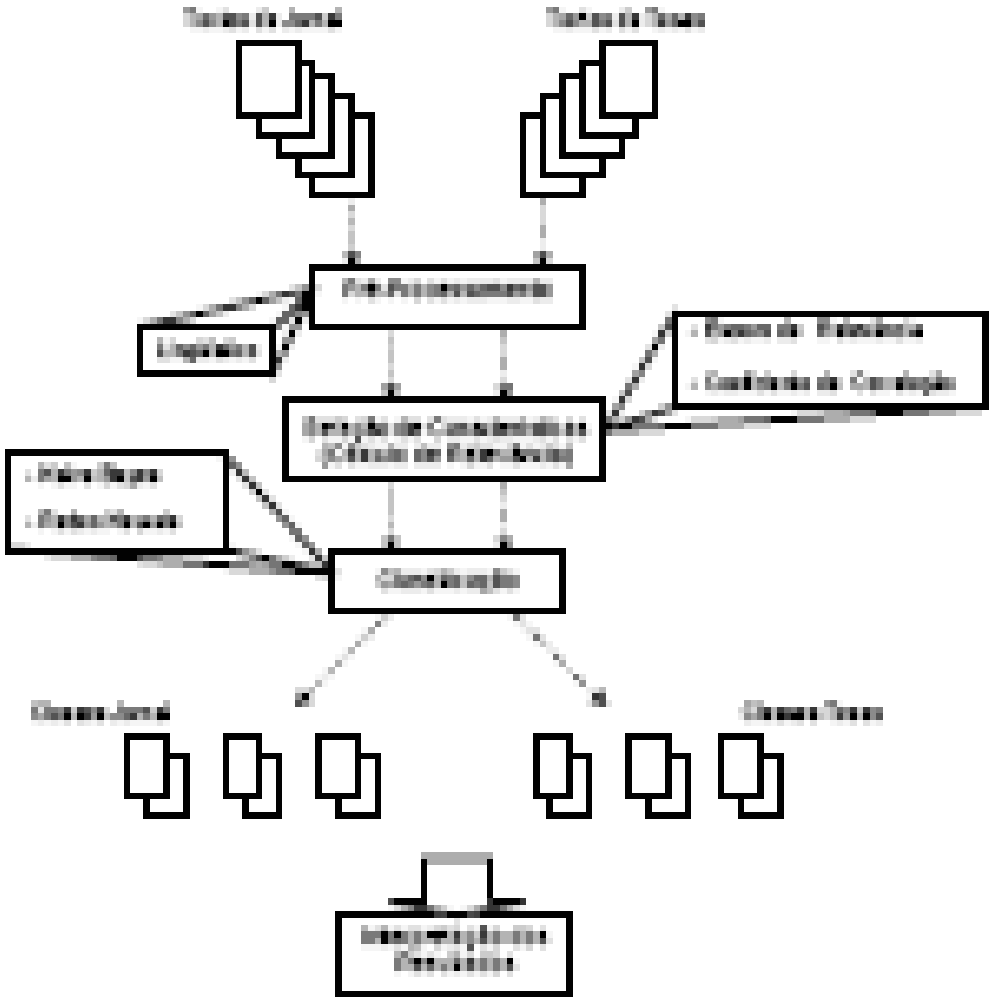


Figura 1- 3 – Proposta de Trabalho e Abordagens

Este trabalho está organizado em sete capítulos cujo conteúdo é relacionado a seguir.

O capítulo um apresenta uma breve introdução definindo o objetivo e a estrutura do trabalho em si.

O capítulo dois fornece uma introdução ao classificador Naïve Bayes.

O capítulo três apresenta uma introdução às redes neurais artificiais com ênfase nas Redes Perceptron de múltiplas camadas (Multi-Layer Perceptron ou MLP).

O capítulo quatro descreve o processo de seleção de características, como o pré-processamento dos textos para extração das características, o modelo de

representação dos documentos, os cálculos de relevância utilizados e a seleção dos termos mais relevantes.

No capítulo cinco são descritos os corpora, as ferramentas utilizadas para a representação dos textos e para a classificação, a implementação computacional destes processos e os experimentos realizados.

O capítulo seis apresenta os resultados obtidos.

O capítulo sete apresenta as conclusões e sugestões de trabalhos futuros.

2. Classificador Naïve Bayes

2.1. Introdução

O classificador Naïve Bayes é baseado no teorema de Bayes e é um dos classificadores mais usados em categorização de textos (McCALLUM, NIGAM, 1998). É um algoritmo para o aprendizado indutivo com abordagem probabilística. É simples, rápido e de fácil implementação. Baseado na probabilidade condicional de determinadas palavras aparecerem em um documento o qual pertence a uma determinada categoria, esta técnica permite calcular as probabilidades de um novo documento pertencer a cada uma das categorias e atribuir a este as categorias de maior probabilidade (LEWIS, RINGUETTE, 1994).

2.2. Conceitos Básicos

O classificador Bayesiano é uma simplificação funcional do classificador ideal Bayesiano. Chamado Naïve por assumir que os atributos são condicionalmente independentes, este classificador assume que existe independência entre as palavras de um texto, ou seja, o método classifica palavras assumindo que a probabilidade de sua ocorrência independe da posição no texto. Apesar desta consideração ser vista como não representativa da realidade, segundo Domingos & Pazzani (1997) a suposição de independência de palavras na maioria dos casos não prejudica a eficiência do classificador.

Cada um dos documentos do conjunto de treinamento é descrito por atributos que indicam a presença ou ausência dos termos $<a_1, a_2,...,a_n>$ e o classificador deverá atribuir a cada um dos documentos a categoria mais provável, por meio de uma função *f* que devolve valores (categorias) pertencentes a um conjunto finito *V*.

O classificador Bayesiano se baseia na suposição simplificada de que vários atributos dos documentos de entrada são condicionalmente independentes, dado o

valor final da função *f* de saída. Isto é, este classificador considera que a probabilidade de ocorrência de uma conjunção de atributos em um dado exemplo é igual ao produtório das probabilidades de ocorrência de cada atributo isoladamente:

$$V_{MAP} = \underset{vj \in V}{\arg\max} P(a_1, a_2, ..., a_n \mid v_j).P(v_j) \quad (2.2\text{-}1)$$

Como os atributos são condicionalmente independentes, dado v, ou seja, as (a_1, a_2, ..., na) são independentes.

$$P(a_1, a_2, ..., a_n \mid v_j) = \prod_i P(a_i \mid v_j) \quad (2.2\text{-}2)$$

Onde:

$P(a_1, a_2, ..., a_n \mid v_j)$ é a probabilidade de ocorrência do conjunto de evidências dada a ocorrência da hipótese (categoria)

Assim o classificador Bayesiano Ingênuo (Naïve Bayes - NB):

$$V_{NB} = \underset{vj \in V}{\arg\max} P(v_j) \cdot \prod_i P(a_i \mid v_j) \quad (2.2\text{-}3)$$

Onde:

V_{NB} é a categoria atribuída ao documento;

v_j é cada um dos possíveis valores (categorias) pertencentes a *V*;

$P(v_j)$ é a probabilidade inicial da ocorrência de cada hipótese;

$P(a_i \mid v_j)$ é a probabilidade de ocorrência de cada evidência dada à ocorrência de uma hipótese (categoria).

Sendo assim, considerando que um documento *D* seja formado por um conjunto de termos t_1 à t_n , a equação 2.2-4 fornece $P(D \mid C_i)$

$$P(D \mid C_i) = \prod_{j=1}^{n} P(t_j \mid C_i) \qquad (2.2\text{-}4)$$

Onde:

$P(t_j \mid C_i)$ é a probabilidade do termo t_j dada uma categoria C_i .

2.3. Modelos

Dentre os modelos estatísticos existentes para os classificadores Naïve Bayes, tem-se o modelo binário que representa um documento como um vetor binário de palavras considerando apenas a ocorrência das palavras no texto e o modelo multinomial, utilizado nessa dissertação, que representa um documento como um vetor de freqüências das palavras no texto. McCallum & Nigam (1998) realizaram experimentos comparando o modelo binário com o modelo multinomial e verificaram que o modelo multinomial apresenta melhores resultados.

3. Redes Neurais

3.1. Introdução

O sistema nervoso é formado por um conjunto extremamente complexo de células, os neurônios. Eles têm um papel essencial na determinação do funcionamento e comportamento do corpo humano e do raciocínio. Os neurônios são formados pelos dendritos, que são um conjunto de ramificações de entrada, pelo corpo central, e pelos axônios que são as ramificações de saída. O potencial do axônio de um neurônio pode se propagar para outro através da existência de um ponto de contacto do seu axônio com um dendrito deste segundo neurônio. Este ponto de contacto é denominado conexão sináptica, ou simplesmente sinapse. As sinapses são unidades estruturais e funcionais elementares que permitem as interações entre neurônios (HAYKIN, 2001).

As redes neurais artificiais são projetadas para simular a estrutura e funcionamento do cérebro humano. São sistemas de processamento de informação intrinsecamente paralelos e distribuídos, constituídos de unidades elementares denominadas neurônios, que têm a capacidade para armazenar conhecimentos experimentais e disponibilizá-los para uso. Possuem habilidade de aprender e generalizar (HAYKIN, 2001).

3.2. Conceitos Básicos

Em uma rede neural, o processamento é feito através da interação de neurônios, também chamados de unidades de processamento ou simplesmente unidades (RUMELHART et al, 1986), que são em geral baseadas no modelo proposto por McCulloch & Pitts para o neurônio humano. De maneira geral, um conjunto de entradas são aplicadas ao neurônio, que responde com uma saída. Cada entrada tem uma influência própria na saída, ou seja, cada entrada tem seu próprio peso na saída.

A conexão de diversos neurônios, organizados em uma ou mais camadas, constitui uma rede neural artificial.

A rede neural artificial, também chamada apenas de rede neural, se assemelha ao cérebro humano em dois aspectos: (HAYKIN, 2001)

- o conhecimento é adquirido pela rede através de um processo de aprendizagem;

- as conexões entre os neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

O neurônio é a unidade de processamento da rede. Cada neurônio gera uma saída a partir da combinação de sinais de entrada recebidos de outros neurônios com os quais está conectado ou a partir de sinais externos. Cada conexão entre dois neurônios possui um peso e estes pesos guardam o conhecimento da rede. A saída de um neurônio é na maior parte dos modelos, o resultado de uma função de ativação aplicada à soma ponderada de suas entradas. Ajustando-se os pesos a rede neural assimila padrões e é capaz de fazer generalizações, produzindo saídas consistentes para entradas não apresentadas anteriormente a rede.(CORREA, 2002)

Neurônios que desempenham função semelhante fazem parte de uma mesma camada. As camadas de uma rede neural podem ser classificadas como entrada, saída ou intermediária. A camada de entrada recebe informações do meio ambiente (documentos). Esta camada apenas propaga esta entrada para a camada seguinte sem nenhuma transformação. A camada de saída transmite a saída para o mundo externo, ou seja, a resposta da rede neural desejada (categoria do documento processado). As camadas intermediárias são as camadas que interligam outras camadas da rede neural, recebendo como entrada as saídas de outra camada e gerando saídas como entradas para outras camadas. Algumas redes não possuem camadas intermediárias e por isso são chamadas de redes de camadas simples.

3.3. Processo de Aprendizagem

Um dos fatores preponderantes para se fazer uso de uma rede neural, é devido a sua capacidade de aprender com seu ambiente e com isso obter melhor performance. Isso é feito com o treinamento caracterizado por um processo iterativo de ajustes aplicados a seus pesos. Essa aprendizagem é executada a partir de um conjunto de regras definidas para a solução de um problema de aprendizado, chamado de algoritmo de aprendizado ou de treinamento.

Existem diferentes tipos de algoritmos de aprendizado específicos para determinados modelos de redes neurais, os quais diferem entre si principalmente pela maneira como os pesos são modificados. (BRAGA, CARVALHO, LUDERMIR, 2000)

O processo de aprendizagem de uma rede neural é caracterizado pela atualização dos valores dos pesos sinápticos de uma rede neural de maneira a obter da rede um padrão de processamento desejado. Nesse contexto existem os seguintes processos de aprendizagem:

- Aprendizado Supervisionado: o processamento desejado para a rede é especificado através de um conjunto de pares ordenados formado por algumas entradas para a rede e as respectivas saídas desejadas. Durante o processo de aprendizagem, é feita uma comparação entre o valor desejado e o valor de saída da rede gerando um erro, o qual é utilizado para ajustar os pesos da rede. Quando este erro é reduzido a valores considerados aceitáveis para o padrão de processamento desejado, é dito que a aprendizagem foi conseguida.

- Aprendizado Não Supervisionado: modelos de redes neurais que possuem capacidade de auto-organização e conseguem produzir saídas satisfatórias a partir dos dados de entrada somente, sem que sejam fornecidas as saídas para estes dados de entrada. A aprendizagem é feita pela descoberta de similaridades nos dados de entrada.

3.4. Redes Multilayer Perceptron (MLP)

As redes MLP, também chamadas de *perceptrons* de múltiplas camadas são do tipo *feedforward* e é muito comum serem utilizadas em problemas de classificação. Nestas redes, o sinal de saída de cada neurônio é o resultado da aplicação da função de ativação sobre a soma ponderada dos sinais de entrada. O modelo de um neurônio artificial é apresentado na figura 3.4-1.

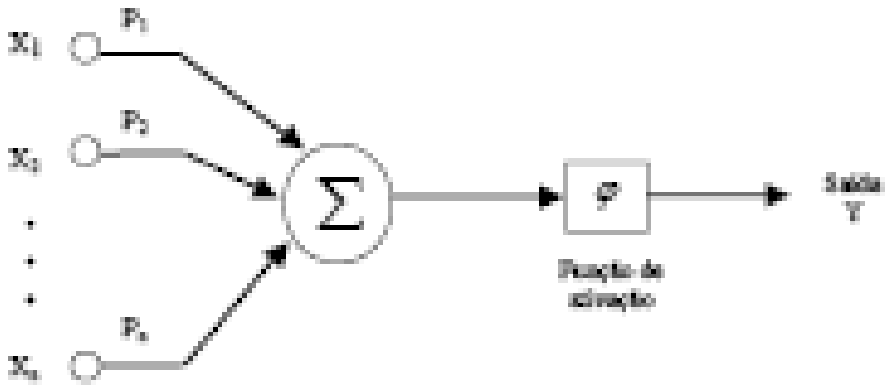


Figura 3.4 -1 – Modelo de um Neurônio Artificial

As redes de uma só camada (*perceptron* simples) são utilizadas para classificação de padrões que sejam linearmente separáveis. Para classificação de padrões não linearmente separáveis são utilizadas redes MLP. São do tipo *feedforward*, o fluxo de dados ocorre da camada de entrada para a camada de saída. Possuem uma ou mais camada ocultas, onde as entradas das unidades das camadas mais elevadas conectam-se com as saídas das unidades da camada imediatamente inferior. Uma rede MLP típica, com uma camada intermediária (oculta) pode ser vista na figura 3.4 - 2.

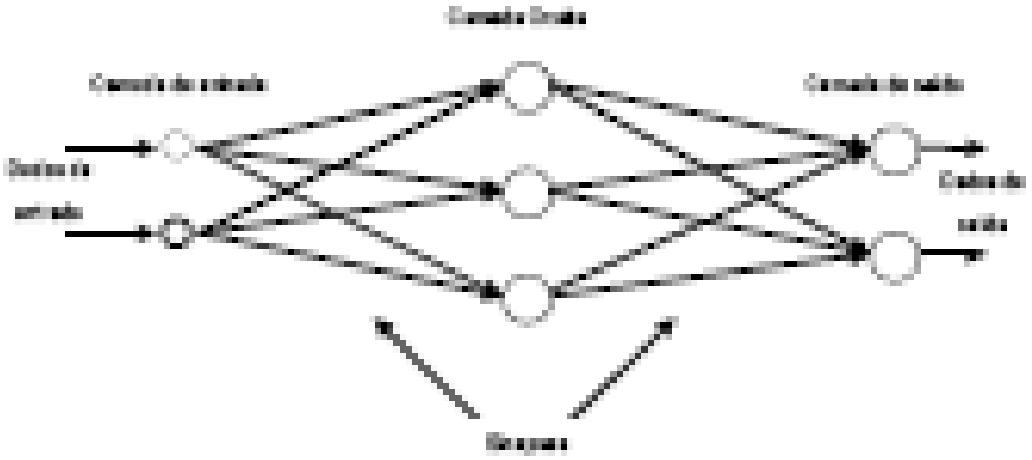


Figura 3.4 -2 – Rede MLP do tipo feedforward com uma camada oculta

Para calcular os pesos adequados a rede a partir de ocorrências do tipo entrada - saída desejada, é necessária uma regra de treinamento.

O algoritmo de retropropagação de erro (também chamado de *algoritmo de backpropagation*) é o mais comumente utilizado dentro dos algoritmos de aprendizagem supervisionada, para treinamento destas redes, pois tem obtido bons resultados quando aplicadas na solução de diversos problemas difíceis (MITCHELL, 1997).

A aprendizagem por retropropagação de erro consiste de dois passos através das diferentes camadas de rede: (HAYKIN, 2001)

- passo para frente (propagação) – onde um padrão de atividade (vetor de entrada) é aplicado aos neurônios da rede e seu efeito se propaga através da rede, camada por camada. Um conjunto de saídas é produzido como resposta da rede. Durante o passo de propagação, os pesos sinápticos da rede são todos fixos.

- passo para trás (retropropagação) - durante o passo para trás, os pesos sinápticos são todos ajustados de acordo com uma regra de correção de erro, ou seja, a resposta real da rede é subtraída de uma resposta desejada para produzir um sinal

de erro. Este sinal de erro é então propagado para trás através da rede, contra a direção das conexões sinápticas. Os pesos sinápticos são ajustados para fazer com que a resposta real da rede se mova para mais perto da resposta desejada, em um sentido estatístico.

Este algoritmo se baseia na regra de aprendizagem por correção de erro e procura minimizar o erro obtido pela saída da rede através do ajuste de pesos.

As redes MLP possuem boa capacidade de generalização, classificando corretamente padrões não utilizados no treinamento ou com ruído. (BRAGA, CARVALHO, LUDERMIR, 2000).

4. Seleção de Características

Segundo Jackson e Moulinier (2002), considerando a problemática geral do processamento informacional da linguagem natural (PLN), em meados da década de 1990 começaram a serem introduzidas sofisticadas abordagens estatísticas no processamento da linguagem natural, trabalhando-se com enormes quantidades de dados lingüísticos, oriundos, por exemplo, de acervos de agências de notícias e páginas da Web. Diante disso, observam uma tendência para o desenvolvimento de programas capazes de executar automaticamente tarefas múltiplas, como por exemplo: selecionar documentos de uma base com enfoque no seu conteúdo, agrupá-los em categorias ou classes e deles extrair determinados conjuntos de informações.

Segundo Moens (2000): “o homem executa a categorização de texto lendo o texto e deduzindo as classes de expressões específicas e seus padrões de contexto. A CADT simula este processo e reconhece os padrões de classificação como uma combinação de características de texto. Estes padrões devem ser gerais o bastante para ter grande aplicabilidade, mas específicos o suficiente para serem seguros quanto à categorização de uma grande quantidade de textos”.

Um conjunto de documentos de textos pode ser representado por um grande número de atributos ou características. Considerando um grande conjunto de atributos, a criação do modelo classificador pode ser prejudicada pelo fato de existirem atributos irrelevantes ou redundantes a uma determinada classificação, conseqüentemente a classificação de novas instâncias será prejudicada.

A seleção de características é o processo de identificação do subconjunto mais representativo, relevante e efetivo dos atributos disponíveis para descrever cada padrão. Basicamente o que a seleção de características faz é reduzir o conjunto de palavras utilizado para representar um documento no processo de classificação.

Diferentes etapas são utilizadas para selecionar as palavras que representarão os documentos a serem classificados.

Como por exemplo:

- remoção de palavras que não teriam significado para o processo de classificação, como as stopwords (preposições, artigos, etc), pois são palavras que se repetem em quase todos os documentos e portanto não são significativas para distinguir cada documento em um processo de classificação;

- formatação dos textos de maneira a obter uma representação estruturada;

- cálculo de relevância para identificar os termos mais significativos;

- redução da dimensionalidade através da seleção dos termos mais relevantes, etc.

Neste trabalho as etapas para tratamento e preparação dos textos podem ser vistas a seguir.

4.1 Pré-processamento

Arquivos textos possuem algumas características difíceis de serem trabalhadas, pois apresentam pouca ou nenhuma estruturação, dificultando o uso de técnicas já conhecidas, e muitas vezes o tamanho do documento compreende milhares de palavras ou termos, tornando o trabalho exaustivo e lento.

O pré-processamento é uma etapa de grande importância na CADT e trabalhosa, pois compreende diversas etapas para transformar o conjunto de documentos em linguagem natural em uma lista de termos úteis e em um formato compatível para a extração do conhecimento. (SILVA, 2004)

A maneira como os documentos são representados e identificados envolve a verificação dos conteúdos, que podem ser analisados automaticamente através de frases e/ou termos que o documento contém. (CORREA, 2002)

Esta análise pode ser realizada através de duas abordagens (RIZZI, 2000):

- abordagem estatística – aplicando-se métodos que incluem seleção e contagem dos termos nos documentos. Neste caso, o termo é o meio de acesso ao documento, e a maneira pela qual eles são identificados e diferenciados.

- abordagem lingüística – aplicando-se métodos semânticos e sintáticos nos textos dos documentos;

O pré-processamento mais comumente utilizado trata da análise léxica (eliminação de dígitos, sinais de pontuação, transformação de maiúscula em minúscula, isolamento dos termos), remoção de termos irrelevantes ou *stopwords* (preposição, artigos, conjunções, etc) (KORFHAGE, 1997; KOWALSKI, 1997; SALTON, 1983), *stemming* ou normalização morfológica dos termos (remoção de afixos e sufixos reduzindo a palavra a seu radical) e seleção dos termos.

Neste trabalho, ao invés das etapas apresentadas acima, pretende-se realizar o pré-processamento fazendo uso de informações lingüísticas para extração das informações pertinentes, ou seja, os termos são extraídos baseados em suas categorias sintáticas e posteriormente são feitas combinações gramaticais que serão utilizadas no processo de classificação como pode ser visto a seguir.

4.1.1. Pré-processamento Lingüístico

Para a extração do conhecimento lingüístico, é necessário fazer a análise sintática dos textos e extrair as informações a partir dela.

O pré-processamento lingüístico é composto dos seguintes passos:

- Análise sintática;

- Extração das categorias gramaticais.

4.1.1.1. Análise Sintática

A sintaxe estuda as regras que governam a formação das frases de uma determinada língua. Estas regras podem ser usadas para a determinação da estrutura sintática das frases geradas.

Uma frase é formada por constituintes (e.g., Sintagma Nominal, Sintagma Verbal, etc), que, por sua vez, são compostos por constituintes de ordem inferior (e.g., Sintagma Preposicional, Sintagma Adverbial, etc), até se chegar às categorias básicas (e.g., substantivo, verbo, etc).

Regras sintáticas determinam a ordem linear dos constituintes na frase, com base na sua categoria sintática. Nas estruturas frasais observa-se uma hierarquia: Sentença (frase), constituintes (componentes sintáticos ou sintagmas), itens lexicais (palavras).

O sintagma é a unidade da análise sintática composta de um núcleo (e.g., um verbo, um nome, um adjetivo, etc) e de outros termos que a ele se unem, formando uma locução que entrará na formação. O nome do sintagma depende da classe da palavra que forma seu núcleo, havendo assim sintagma nominal (núcleo substantivo), sintagma verbal (núcleo verbo), sintagma adjetival (núcleo adjetivo), sintagma adverbial (núcleo advérbio), sintagma preposicional (núcleo preposição) (SILVA, 2004).

A determinação da estrutura sintática das frases é vista como uma etapa central na interpretação de linguagem natural, a partir da qual a frase de entrada pode ser formalmente analisada.

Através da técnica de *parsing* é possível determinar a estrutura sintática da frase sob análise. Um *parser* é um algoritmo que mapeia uma frase na sua estrutura sintática.

Neste trabalho, foi utilizado o *parser* PALAVRAS. Ele trabalha no nível do sintagma, tentando validar o agrupamento de termos que compõe as frases.

O analisador sintático PALAVRAS, desenvolvido para o português por Bick (BICK, 2000), faz parte de um grupo de *parsers*, do projeto Visual Interactive Syntax Learning do Institute of Language and Communication da University of Southern Denmark. Este analisador gera anotação lingüística para textos em língua portuguesa. Através dele, é possível fazer a análise sintática de uma sentença ou de um conjunto de sentenças (textos em linguagem natural). Ele recebe como entrada uma sentença e gera a análise sintática da mesma como pode ser visto a seguir na figura 4.1.1-1.

A figura 4.1.1-1 mostra a marcação sintática do *parser* para a sentença “Crianças correm em verdes campos.” submetida ao PALAVRAS.

Figura 4.1.1 -1 – Marcação do analisador sintático PALAVRAS

No exemplo acima verifica-se as seguintes etiquetas morfossintáticas: SUBJ = sujeito; P = predicado; pp = sintagma preposicional; np = sintagma nominal; v*fin = verbo flexionado; n = substantivo; prp = preposição; adj = adjetivo; entre parênteses tem-se a forma canônica da palavra e fora dos parênteses a palavra como aparece no documento.

Com base nas marcações do analisador sintático, um conjunto de programas denominado XTRACTOR foi desenvolvido em cooperação com a Universidade de

Évora. A ferramenta XTRACTOR (GASPERIN, 2003) engloba a análise do corpus por meio do PALAVRAS, e converte a saída do analisador sintático em três arquivos XML(eXtensible Markup Language). XML é uma linguagem de marcação que tem sido utilizada em diversas áreas e aplicações. Em processamento de linguagem natural, essa linguagem é utilizada para agregar aos textos anotações com informações lingüísticas de uma maneira organizada e padronizada (BUITELAAR, 2003) (VILELA, 2005).

Dos arquivos gerados em XML, um contém a lista de palavras do texto e seus identificadores (arquivo WORDS – figura 4.1.1-2), outro contém a estrutura das sentenças (arquivo CHUNKS – figura 4.1.1-3) e o outro contém as informações morfossintáticas do texto (arquivo POS – *part of speech* – figura 4.1.1-4) de interesse para este trabalho.

A figura 4.1.1-2 mostra o arquivo *WORDS* com a lista de palavras e seus identificadores (único para cada termo do texto), gerado a partir da sentença “Crianças correm em verdes campos“.

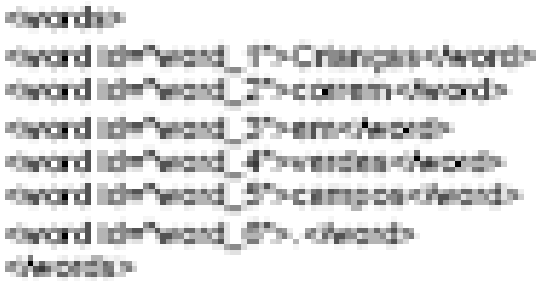


Figura 4.1.1-2– Arquivo *WORDS*

A figura 4.1.1-3 mostra o arquivo *CHUNKS*, que consiste das estruturas e subestruturas sintáticas das sentenças. Um *chunk* representa a estrutura interna da sentença.

Figura 4.1.1-3– Arquivo *CHUNKS*

A figura 4.1.1-4 mostra o arquivo POS (*Part of Speech)* que apresenta as informações morfossintáticas das palavras do texto, onde: o elemento n = substantivo; v = verbo, prp = preposição, adj = adjetivo. A palavra do texto é apresentada em sua forma canônica classificada em gênero e número.

Figura 4.1.1-4– Arquivo POS

Através do arquivo POS gerado em XML, é possível extrair as categorias gramaticais dos textos analisados.

4.1.1.2. Extração das Categorias Gramaticais

A extração das categorias gramaticais é feita através de folhas de estilo XSL (e*Xtensible Stylesheet Language*) aplicadas ao arquivo POS gerado em XML.

XSL é um conjunto de instruções destinadas à visualização de documentos XML, sendo possível transformar um documento XML em diversos formatos como por exemplo HTML, RTF e TXT. A linguagem XSL auxilia a identificação dos elementos de um documento XML, permitindo a simplificação do processamento de transformação desses elementos em outros formatos de apresentação. Uma folha de estilos é composta por um conjunto de regras (chamado *templates*) ativado no processamento de um documento XML.

Utilizando-se folhas de estilo, as categorias gramaticais são extraídas e portanto os termos correspondentes ao documento.

Neste trabalho foram implementadas folhas de estilo para extração das categorias gramaticais: substantivo; substantivo e adjetivo; substantivo e verbo; substantivo e nome próprio; substantivo, nome próprio e adjetivo; substantivo, verbo e adjetivo.

Extraídos os termos dos documentos, faz-se necessário representar a coleção de documentos em um formato estruturado e compacto de maneira a atender as necessidades de processamento, utilizando os chamados termos de indexação.

4.1.2. Representação dos Documentos

O objetivo principal de um modelo de representação de documentos, é a obtenção de uma descrição adequada da semântica do texto, de uma forma que permita a execução correta da tarefa alvo, de acordo com as necessidades do usuário (GEAN e KAESTNER, 2004).

Diversos modelos foram desenvolvidos para a representação de grandes coleções de textos que identificam documentos sobre temas específicos. Um dos modelos utilizados, devido a sua simplicidade e a rapidez com que as operações com vetores são realizadas, é o modelo de espaço vetorial.

Segundo von Wangenheim (2006), “a capacidade de um sistema de realizar o reconhecimento de padrões de forma flexível e adaptável está intimamente associada à idéia de que um sistema de reconhecimento de padrões deve ser capaz de aprender as características e a distribuição dos padrões no espaço vetorial definido por um determinado domínio de aplicação”. Isso implica que o sistema seja conseqüentemente capaz de “aprender como associar um determinado padrão à classe à qual pertence” (WANGENHEIM, 2006).

De acordo com o modelo vetorial de Salton (1975), cada documento é representado por um vetor no espaço T-dimensional, onde T é o número de diferentes termos presentes na coleção. Os valores das coordenadas do vetor que representa o documento estão associados aos termos, e usualmente são obtidos a partir de uma função relacionada à freqüência dos termos no documento e na coleção.

A figura 4.1.2-1 apresenta a relação termo-documento com categorias predefinidas, utilizando este modelo de representação.

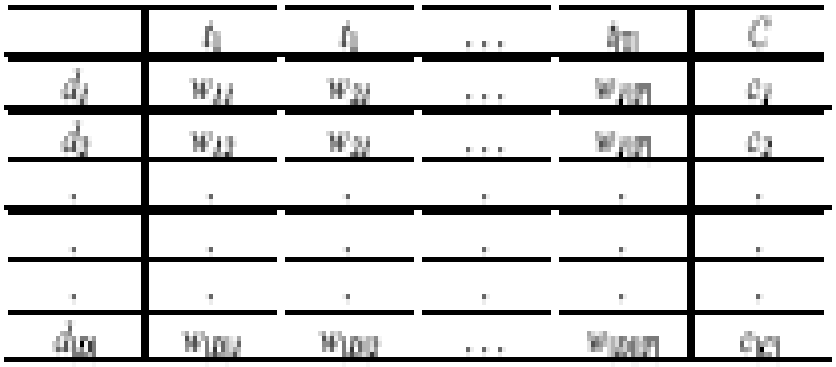


Figura 4.1.2 -1– Relação termo-documento com categorias predefinidas

Onde:

- d_1 a $d_{|D|}$ são documentos da coleção;

- t_1 a $t_{|T|}$ são os termos;

- c_1 a $c_{|C|}$ são as categorias predefinidas;

- w_{11} a $w_{|D||T|}$ são os pesos dos termos.

Segundo este modelo, os termos se tornam dimensões e os valores informam a relevância (peso) dos termos. Assim, neste modelo os documentos são representados por vetores e cada índice corresponde a uma palavra (SALTON, 1988). Um peso é associado a cada palavra para descrever sua relevância no documento. Estas associações de pesos aos termos de indexação funcionam como um grau de similaridade entre os vetores documentos, e entre os vetores documentos e os vetores

representativos das categorias. A similaridade entre dois vetores é obtida aplicando-se o produto interno dividido pelo produto das normas entre esses dois vetores. Cada categoria pode ser representada como um conjunto de vetores resultantes do somatório dos documentos pertencentes aos respectivos subconjuntos delas. Não havendo variância muito grande entre os documentos de uma categoria, apenas um vetor resultante do somatório de todos os documentos a ela pertencentes pode ser utilizado para representá-la.

Em um conjunto de textos, se uma palavra aparece em grande parte deles, não é um bom termo de indexação, porém se esta aparece em alguns poucos, já é significativa para a representação da coleção. Portanto, existem palavras que são mais significativas do que outras, daí a necessidade de se atribuir pesos aos termos (palavras) como medida de relevância dos mesmos para o documento (CORREA,2002).

4.2 Cálculo de Relevância

Concluído o pré-processamento, estabelecidos os termos da coleção e o modelo de representação dos documentos, é necessário definir o conjunto de termos que melhor representem o assunto a ser categorizado através de uma indexação automática.

Este conjunto deve ser estabelecido através de um cálculo de representatividade dos termos, ou seja, estes termos devem ter associados a eles, valores que quantificam sua representatividade na coleção de documentos através de um cálculo de relevância.

Dentre as técnicas existentes para execução desta tarefa, a medida mais comumente usada é o tf-idf que é a freqüência do termo no documento multiplicada pelo inverso da freqüência deste termo na coleção. Neste trabalho foram utilizados o Escore de Relevância (ER) e o Coeficiente de Correlação (CC).

4.2.1. Escore de Relevância

O escore de relevância foi proposto e aplicado inicialmente no estudo de Wiener, Pederson e Weigend (1995) com base no peso de relevância de Salton e Buckley (1983).

Nos estudos de Salton e Buckley (1983), ele calculou a freqüência de cada termo no documento. Posteriormente, calculou a freqüência do termo dentro do documento e da coleção, chegando ao cálculo da freqüência inversa de documentos. Ao verificar que termos com alto grau de representatividade de conteúdo possuem alta freqüência no documento e baixa freqüência na coleção, ele definiu a técnica do cálculo do peso de relevância do termo. Segundo Salton e Buckley (1983), a indexação de textos feita a partir de termos com pesos associados, alcança melhores resultados pois o peso determina o grau de importância do termo dentro do documento.

Baseado neste estudo Wiener propôs o escore de relevância. O escore de relevância se baseia na freqüência dos termos em uma dada categoria e também nas outras categorias da coleção. A partir destes dados é calculada a relevância do termo para uma dada categoria. Termos que aparecem em muitas categorias obtêm valores baixos, por serem pouco discriminantes, enquanto que termos que aparecem em poucas categorias ficam com valores muito altos, podendo então representar a categoria.

O Escore de Relevância do termo t é definido por:

$$r_t = \log \frac{\frac{w_{ct}}{d_c} + \frac{1}{6}}{\frac{w_{\bar{ct}}}{d_{\bar{c}}} + \frac{1}{6}} \qquad (4.2.1\text{-}1)$$

Onde:

- w_{ct} é o número de documentos pertencentes a uma dada categoria (c) que contém o termo t ;

- d_c é o número total de documentos da categoria considerada (c);

- $w_{\bar{ct}}$ é o número de documentos de outras categorias que contém o termo t;

- $d_{\bar{c}}$ é o número total de documentos de outras categorias.

A constante 1/6 aparece na fórmula para eliminar o problema da divisão por zero (caso em que o termo só apareça na categoria considerada e não apareça nas outras categorias)

4.2.2. Coeficiente de Correlação

O coeficiente de correlação foi desenvolvido por Ng et al. (1997) para indicar o grau de correlação entre uma palavra e um documento. Ele 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.

O Coeficiente de Correlação entre o termo *t* e a classe *c* é definido por:

$$C(t,c) = \frac{\left(N_{r+} x N_{n-} - N_{r-} x N_{n+}\right) x \sqrt{N}}{\sqrt{\left(N_{r+} + N_{r-}\right) x \left(N_{n+} + N_{n-}\right) x \left(N_{r+} + N_{n+}\right) x \left(N_{r-} + N_{n-}\right)}} \qquad (4.2.2\text{-}1)$$

Onde:

- Nr+ é o número de documentos relevantes para C_j que contém o termo *t*;

- Nr⁻ é o número de documentos relevantes para C_j que não contém o termo *t;*

- Nn+ é o número de documentos não relevantes para C_j que contém o termo *t*;

- Nn⁻ é o número de documentos não relevantes para C_j que não contém o termo *t*.

Esta medida corresponde à raiz quadrada do valor obtido pela métrica do Qui-quadrado (mede estatisticamente o grau de independência entre o termo e a categoria). O coeficiente de correlação é maior para as palavras que indicam a pertinência de um documento à categoria C_j enquanto a métrica do Qui-quadrado gera valores maiores não só para este conjunto de palavras mas também para aquelas que indicam a não pertinência à C_j.

4.3 Seleção de Atributos

Estabelecidos os termos dos documentos e seus respectivos valores estabelecidos através do cálculo de relevância, faz-se necessário eliminar os termos não representativos e reduzir a dimensionalidade, já que, utilizar todos os termos, além de poder ser inviável computacionalmente, pode também ser um fator de comprometimento da classificação. Cabe salientar também, que o tempo de processamento é proporcional à quantidade de termos utilizados.

Diferentes métodos são utilizados com o intuito de reduzir a quantidade de termos representativos e não perder a qualidade de representatividade, como seleção por peso do termo, seleção por linguagem natural, etc.

Neste trabalho foi utilizada a técnica *seleção por peso do termo*, também chamada de *truncagem*, que tem por objetivo selecionar os termos mais relevantes para representar um documento e eliminar o restante. Através da *truncagem* é possível reduzir a dimensionalidade de maneira a otimizar a performance do classificador e obter um desempenho satisfatório. Esta é uma técnica bastante utilizada por obter bons resultados.

Esta técnica consiste em ordenar os termos por um grau de relevância e os de maior grau são selecionados para a classificação.

5. A Classificação

Através do reconhecimento de padrões é possível classificar. Segundo Wangenheim (2006), o aprendizado de máquina em reconhecimento de padrões é "um método que permite organizar uma seqüência de padrões P_1, P_2, . . . , P_n em vários conjuntos de padrões CP_1, CP_2, . . ., CP_k denominados classes, de tal forma que os padrões organizados em cada conjunto são similares entre si e dissimilares dos padrões armazenados nos outros conjuntos".

Neste capítulo serão descritas as coleções (corpora) utilizadas, as metodologias, técnicas e ferramentas utilizadas para o preparo dos textos de forma a apresentá-los a classificação, assim como a metodologia utilizada na concepção, treinamento e avaliação dos classificadores utilizados.

5.1. Base de Dados Textuais

Dois corpora denominados c*orpus Jornal* e c*orpus Teses*, foram utilizados para a classificação:

- Corpus *Jornal* é um conjunto de textos elaborado pelo Núcleo Interinstitucional de Lingüística Computacional (NILC), composto por 855 textos de artigos jornalísticos do ano de 1994 do jornal Folha de São Paulo divididos em cinco categorias, cada uma com 171 textos. As categorias são: esportes, imóveis, informática, política e turismo.

- Corpus *Teses* é um conjunto de textos compostos por título e resumo, das dissertações de mestrado e teses de doutorado da Engenharia Elétrica da COPPE/UFRJ, composto por 475 textos divididos em cinco categorias, cada uma com 95 textos. As categorias são: controle, microeletrônica, processamento de sinais, redes e sistemas de potência.

5.2. Preparação dos Documentos

Os corpora foram submetidos ao analisador sintático PALAVRAS para obtenção da análise sintática das sentenças. Em seguida, as marcações obtidas pelo *parser* são submetidas à ferramenta XTRACTOR para geração dos arquivos em XML. Aplicando-se folhas de estilo XSL aos arquivos XML gerados, foram obtidas as categorias gramaticais e foram feitas as seguintes combinações de categorias gramaticais: substantivo, substantivo+nome próprio, substantivo+adjetivo, substantivo+verbo, substantivo+nome próprio+adjetivo, substantivo+verbo+adjetivo de cada texto. Este processo pode ser visto na figura 5.2-1.

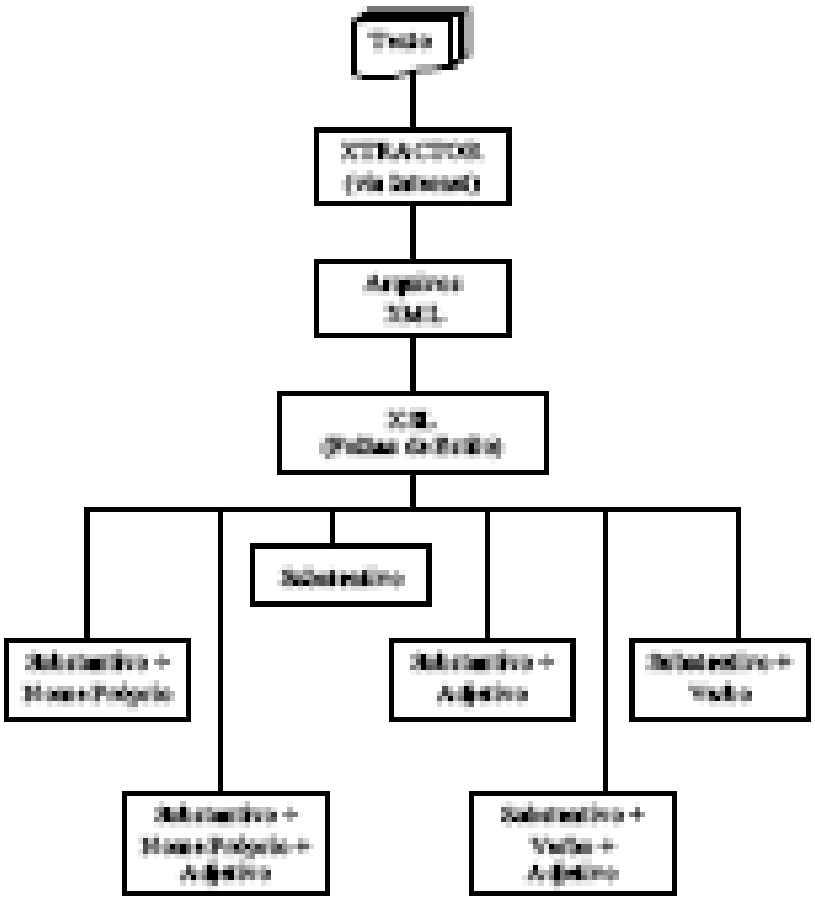


Figura 5.2-1– Processo de extração das categorias gramaticais

As tabelas 5.2-1 e 5.2-2 mostram o número de termos dos corpora Jornal e Teses, após a extração das categorias gramaticais.

A tabela 5.2-1 mostra o número de termos da coleção referente ao corpus Jornal.

Tabela 5.2-1– Corpus Jornal e o número de termos da coleção

[illegible]	
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

A Tabela 5.2-2 mostra o número de termos da coleção referente ao corpus Teses.

Tabela 5.2-2– Corpus Teses e o número de termos da coleção

[illegible]	
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

Para a classificação, é necessário estabelecer um conjunto para treino e um conjunto para teste. Na tentativa de averiguar a robustez dos resultados e obter uma representação mais realista dos conjuntos de treino e teste, evitando resultados específicos a um determinado conjunto escolhido aleatoriamente, foi utilizado o 3-*fold cross validation*. Para cada categoria dos dois corpora, o conjunto de textos foi dividido em treino e teste, sendo 2/3 para treino e 1/3 para teste.

Com relação ao corpus Jornal, os 855 documentos que fazem parte da coleção, foram divididos em cinco classes, sendo 171 documentos por classe. Os documentos de cada classe foram então, divididos em aproximadamente 2/3 para o conjunto de treino e 1/3 para o conjunto de testes.

Com relação ao corpus Teses, os 475 documentos que fazem parte da coleção, foram divididos em cinco classes, sendo 95 documentos por classe. Os documentos de cada classes foram então, divididos em aproximadamente 2/3 para o conjunto de treino e 1/3 para o conjunto de testes.

5.3. Ferramentas Utilizadas

Para o processamento dos textos foi utilizada a ferramenta WVTool, desenvolvida em Java por Michael Wurst e adaptada para o uso neste trabalho. Através desta ferramenta, é feita uma correspondência entre a lista de palavras (termos) e seus valores numéricos obtidos através dos cálculos de relevância (escore de relevância e coeficiente de correlação) aplicados aos documentos.

A partir dos n termos mais relevantes, selecionados através do método de truncagem, são construídos os vetores locais de cada categoria. Unindo-se os vetores locais de cada categoria, é formado o vetor global. Este vetor global gerado vai servir de índice para os vetores de cada exemplo e as posições correspondentes representam a importância da mesma dentro do documento.

Para a construção dos vetores locais foram selecionados os 6, 12, 18, 24, 30 termos mais relevantes de cada categoria, e através da junção dos vetores locais, são gerados respectivamente os vetores globais com 30, 60, 90, 120, e 150 termos. Estas posições já definem a entrada na ferramenta para classificação.

Para a classificação foi utilizada a ferramenta WEKA (Waikato Environment for Knowledge Analysis) (WITTEN and FRANK, 2000). Esta ferramenta possui uma coleção de algoritmos de aprendizado de máquina para resolução de problemas de Data Mining, é implementada em Java (*open source*), suporta métodos de aprendizagem supervisionada e não supervisionada tais como, Árvores de decisão, Redes Neurais Artificiais, Naïve Bayes, Support Vector Machine, K-means, etc.

A ferramenta lê os dados no formato ARFF (formato padrão de arquivo utilizado pela ferramenta). O ARFF consiste basicamente de duas partes:

- Primeira parte: consiste de uma lista de todos os atributos definidos pelos tipos ou valores que ele pode representar;

- Segunda parte: contém uma lista de todas as instâncias, onde os valores dos atributos são separados por vírgula.

5.4. Experimentos

Neste trabalho utilizou-se para a classificação os classificadores Naïve Bayes e as Redes Neurais.

Para o algoritmo Naïve Bayes, foi utilizado o modelo multinomial por apresentar melhores resultados na categorização de textos segundo McCallum & Nigam (1998) e Yang & Liu (1999).

Para as Redes Neurais Artificiais foi utilizada a Rede MLP do tipo feedforward com algoritmo de *backpropagation*. Os parâmetros estabelecidos foram, 0.9 para o valor de momentum, 0.1 para taxa de aprendizado, a condição de parada foi 3000 épocas, o número de neurônios na camada intermediária foi variado entre 2, 4, 8 e 16 neurônios e o número de neurônios na camada de saída corresponde às classes referentes a cada corpus. Estes valores foram baseados nos estudos de SILVA (2004).

Para avaliação foi levado em conta os valores referentes à média do percentual de erro obtido no resultado da classificação dos três conjuntos de teste.

Após realizados os experimentos com os dois classificadores citados acima e os resultados analisados, foi feita uma pesquisa com o intuito de melhorar os resultados obtidos no processo de classificação. Para esta pesquisa foram utilizadas somente as combinações de categorias gramaticais que obtiveram os melhores resultados em cada corpus. Essa pesquisa foi elaborada utilizando-se um ensemble de classificadores. Para cada classificador convencional (Naïve Bayes e Redes Neurais) foi composto um conjunto de cinco classificadores denominados, neste trabalho, Naïve Bayes Especialistas e Redes Neurais Especialistas. As saídas de cada especialista são enviadas para um combinador que produzirá o resultado final, ou seja, a classe vencedora. A suposição é de que os erros sejam minimizados através do uso de múltiplos classificadores ao invés de um único classificador.

A figura 5.4-1 mostra o esquema utilizado para o ensemble de classificadores especialistas que têm a mesma entrada e as saídas individuais são combinadas para produzir uma saída.

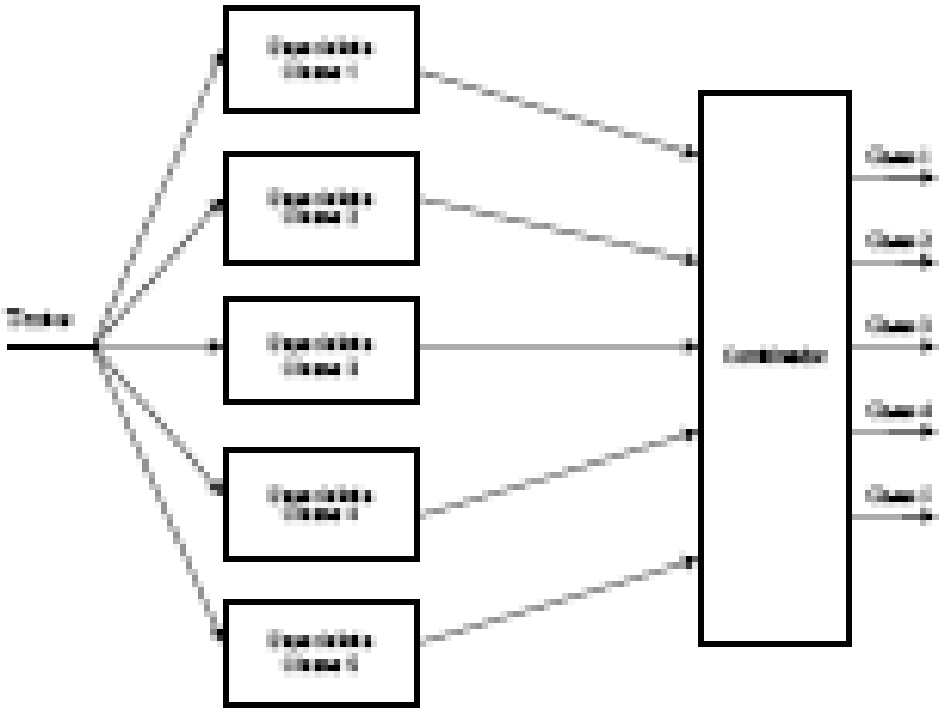


Figura 5.4-1 – Ensemble de Classificadores Especialistas

Existem diferentes maneiras de se combinar as saídas dos especialistas. Neste trabalho foi adotada uma votação majoritária com atribuição de votante por maior grau de certeza. Entre os critérios de votação, a votação majoritária está entre os mais simples e de aplicação mais geral e tem apresentado ótimos resultados se comparado a outros critérios, muitas vezes custosos devido ao processamento e que quando apresentam melhores resultados, os mesmos são pouco significativos.

Inicialmente foi eleito um especialista para cada classe. Cada especialista

identifica se cada padrão de entrada pertence ou não a sua classe e com que grau de certeza. Cada padrão apresentado então, será atribuído à classe cujo especialista apresenta maior grau de certeza.

5.5. Considerações

Na literatura são encontradas diferentes propostas para classificação automática de documentos de textos, envolvendo o desenvolvimento de diferentes técnicas e muitas vezes é feita a comparação com outras, permitindo a avaliação entre diferentes abordagens.

Em SILVA (2004) foi feita uma comparação entre o pré-processamento usual e o pré-processamento baseado em informações lingüísticas.

Em ARANHA (2007) é proposto um modelo automático de enriquecimento dos dados na fase de pré-processamento, transformando o modelo baseado em palavras em um modelo baseado em lexemas, utilizando conhecimentos da área de processamento da linguagem natural e lingüística computacional.

Em LAN (2007) são examinados os caminhos para representação de textos no modelo de espaço vetorial, com o intuito de verificar a performance da categorização de textos.

Em ARCOVERDE (2007) são investigadas evidências que fundamentam a hipótese de que aplicações de métodos que utilizam conhecimento lingüístico é viável. É proposto um modelo de representação de texto fundamentado em sintagmas nominais, cuja representatividade de seus descritores é calculada utilizando-se o conceito de “evidência”, apoiado em modelos estatísticos.

Em CAMARGO (2007) foram combinados os classificadores Naïve Bayes e Máquina de Vetor Suporte para classificação automática de textos.

6. Resultados Obtidos

Neste capítulo serão apresentados os resultados obtidos com os classificadores utilizados.

As tabelas estão organizadas de acordo com a categoria ou combinação de categorias gramaticais. Elas mostram os valores referentes às médias dos percentuais de erro dos três conjuntos de teste obtidos pelos classificadores. Os menores valores estão em negrito para melhor visualização.

As tabelas estão organizadas também, de maneira a se observar o cálculo de relevância e o número de termos utilizados. Para as Redes Neurais, são apresentados também os números de neurônios da camada intermediária utilizados.

6.1 Classificador Naïve Bayes

6.1.1. Corpus Jornal

As Tabelas 6.1.1-1 a 6.1.1-6 mostram os resultados obtidos no processo de classificação. Pode-se verificar que, em todas as combinações de categorias gramaticais, os melhores resultados foram obtidos utilizando-se 150 termos.

Tabela 6.1.1-1 – Corpus Jornal – Substantivo

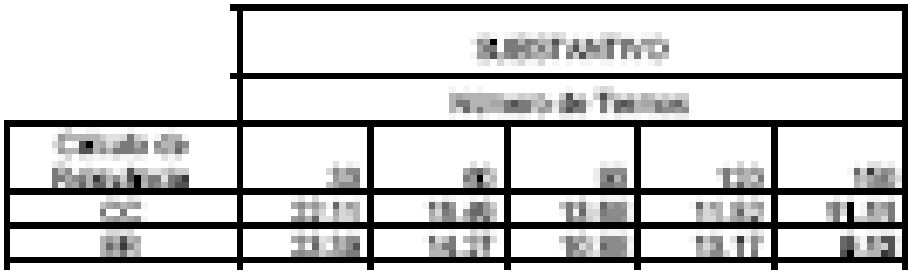


Tabela 6.1.1-2– Corpus Jornal - Substantivo + Adjetivo

	[illegible]				
	[illegible]				
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.1.1-3 – Corpus Jornal - Substantivo + Nome Próprio

	[illegible]				
	[illegible]				
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.1.1-4 – Corpus Jornal - Substantivo + Verbo

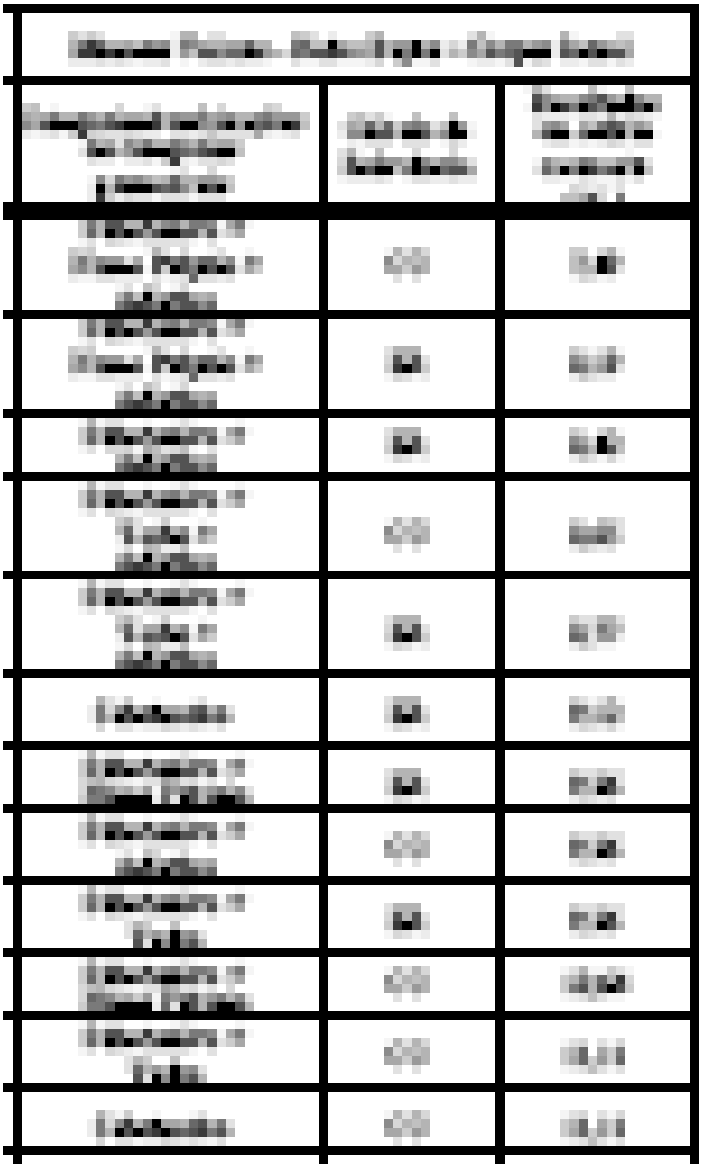
	[illegible]				
	[illegible]				
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.1.1-5 – Corpus Jornal - Substantivo + Verbo + Adjetivo

Tabela 6.1.1-6 – Corpus Jornal - Substantivo + Nome Próprio + Adjetivo

A tabela 6.1.1-7 mostra ordenadamente, os melhores resultados obtidos com o classificador Naïve Bayes para o corpus Jornal. Observa-se que a combinação substantivo + nome próprio + adjetivo apresentou melhor resultado tanto com o coeficiente de correlação quanto com o escore de relevância. É importante observar também que todas as combinações obtiveram bons resultados, sendo a maioria abaixo de 10%.

Tabela 6.1.1-7 – Menores valores - Naïve Bayes – Corpus Jornal



6.1.2. Corpus Teses

As tabelas 6.1.2-1 a 6.1.2-6 mostram os resultados obtidos no processo de classificação do corpus Teses. Os melhores resultados foram obtidos utilizando-se 150 termos assim como no corpus Jornal, porém em algumas combinações de categorias gramaticais, a utilização de 90 termos mais relevantes apresentou melhores resultados.

Tabela 6.1.2 - 1 – Corpus Teses – Substantivo

Tabela 6.1.2 - 2 – Corpus Teses - Substantivo + Adjetivo

Tabela 6.1.2 - 3 – Corpus Teses - Substantivo + Nome Próprio

Tabela 6.1.2 - 4 – Corpus Teses - Substantivo + Verbo

Tabela 6.1.2 -5 – Corpus Teses - Substantivo + Verbo + Adjetivo

Tabela 6.1.2 -6 – Corpus Teses - Substantivo + Nome Próprio + Adjetivo

A tabela 6.1.2-7 mostra ordenadamente, os melhores resultados obtidos com o classificador Naïve Bayes para o corpus Teses. Observa-se que a combinação substantivo + verbo + adjetivo com o escore de relevância, apresentou melhor resultado, diferentemente do corpus Jornal que apresentou melhor resultado com a combinação substantivo + nome próprio + adjetivo e com o coeficiente de correlação.

Tabela 6.1.2-7 – Menores valores - Naïve Bayes – Corpus Teses

[illegible]		
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

6.2 Redes Neurais

6.2.1. Corpus Jornal

As tabelas 6.2.1-1 a 6.2.1--12 mostram os resultados obtidos no processo de classificação utilizando-se as Redes Neurais. O melhor resultado (13.96) foi obtido com a combinação substantivo + nome próprio + adjetivo, com 60 termos e 16 neurônios na camada intermediária. Pode-se observar que os melhores resultados para todas as combinações foram obtidos utilizando-se 60 ou 90 termos, com 8 ou 16 neurônios na camada intermediária.

Tabela 6.2.1-1 – Corpus Jornal - Substantivo – CC

Tabela 6.2.1-2– Corpus Jornal - Substantivo – ER

Tabela 6.2.1-3 – Corpus Jornal - Substantivo + Adjetivo – CC

		SUBSTANTIVO + ADJETIVO				
		Número de Termos				
[illegible]	[illegible]	30	60	90	120	150
CC	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.1-4 – Corpus Jornal - Substantivo + Adjetivo – ER

		SUBSTANTIVO + ADJETIVO				
		Número de Termos				
[illegible]	[illegible]	30	60	90	120	150
ER	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.1-5 – Corpus Jornal - Substantivo + Nome Próprio – CC

		SUBSTANTIVO + NOME PRÓPRIO				
		Número de Termos				
[illegible]	[illegible]	30	60	90	120	150
CC	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.1-6 – Corpus Jornal - Substantivo + Nome Próprio – ER

		[illegible]				
		[illegible]				
[illegible]	[illegible]	30	60	90	120	180
ER	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	**[illegible]**	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.1-7 – Corpus Jornal - Substantivo + Verbo – CC

		[illegible]				
		[illegible]				
[illegible]	[illegible]	30	60	90	120	180
CC	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	**[illegible]**	[illegible]	[illegible]

Tabela 6.2.1-8 – Corpus Jornal - Substantivo + Verbo – ER

		[illegible]				
		[illegible]				
[illegible]	[illegible]	30	60	90	120	180
ER	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	**[illegible]**	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.1-9 – Corpus Jornal - Substantivo + Verbo + Adjetivo – CC

		[illegible]				
		[illegible]				
[illegible]	[illegible]	30	60	90	120	180
CC	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.1-10 – Corpus Jornal - Substantivo + Verbo + Adjetivo – ER

		[illegible]				
		[illegible]				
[illegible]	[illegible]	30	60	90	120	180
ER	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.1-11 – Corpus Jornal - Substantivo + Nome Próprio + Adjetivo – CC

		[illegible]				
		[illegible]				
[illegible]	[illegible]	30	60	90	120	180
CC	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.1-12 – Corpus Jornal - Substantivo + Nome Próprio + Adjetivo – ER

		[illegible]				
		[illegible]				
[illegible]	[illegible]	30	60	90	120	150
[illegible]	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

A tabela 6.2.1-13 mostra ordenadamente, os melhores resultados obtidos no processo de classificação com as Redes Neurais para o corpus Jornal. O melhor resultado foi obtido com a mesma combinação gramatical e o mesmo cálculo de relevância que obteve melhor resultado para o classificador Naïve Bayes.

Tabela 6.2.1-13 – Menores valores – Redes Neurais – Corpus Jornal

[illegible]		
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

6.2.2. Corpus Teses

As tabelas 6.2.2-1 a 6.2.2--12 mostram os resultados obtidos no processo de classificação utilizando-se as Redes Neurais. O melhor resultado foi obtido com a combinação substantivo + verbo + adjetivo, com 60 termos e 16 neurônios na camada intermediária.

Tabela 6.2.2-1– Corpus Teses - Substantivo - CC

Tabela 6.2.2-2– Corpus Teses - Substantivo – ER

Tabela 6.2.2-3 – Corpus Teses - Substantivo + Adjetivo – CC

Tabela 6.2.2-4 – Corpus Teses - Substantivo + Adjetivo – ER

Tabela 6.2.2-5 – Corpus Teses - Substantivo + Nome Próprio – CC

Tabela 6.2.2-6 – Corpus Teses - Substantivo + Nome Próprio – ER

		[illegible]				
		[illegible]				
[illegible]	[illegible]	10	60	80	120	150
ER	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.2-7 – Corpus Teses - Substantivo + Verbo – CC

		[illegible]				
		[illegible]				
[illegible]	[illegible]	10	60	80	120	150
CC	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.2-8 – Corpus Teses - Substantivo + Verbo – ER

		[illegible]				
		[illegible]				
[illegible]	[illegible]	10	60	80	120	150
ER	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.2-9 – Corpus Teses - Substantivo + Verbo + Adjetivo – CC

		[illegible]				
		[illegible]				
[illegible]	[illegible]	30	60	90	120	150
CC	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.2-10 – Corpus Teses - Substantivo + Verbo + Adjetivo – ER

		[illegible]				
		[illegible]				
[illegible]	[illegible]	30	60	90	120	150
ER	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.2-11 – Corpus Teses - Substantivo + Nome Próprio + Adjetivo – CC

		[illegible]				
		[illegible]				
[illegible]	[illegible]	30	60	90	120	150
CC	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.2-12 – Corpus Teses - Substantivo + Nome Próprio + Adjetivo – ER

		[illegible]				
		[illegible]				
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Tabela 6.2.2-13 – Menores valores – Redes Neurais – Corpus Teses

[illegible]		
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

A tabela 6.2.2-13 mostra ordenadamente, os melhores valores obtidos no processo de classificação com as Redes Neurais para o corpus Teses.

O melhor resultado foi obtido com a mesma combinação gramatical que apresentou melhor resultado utilizando-se o classificador Naïve Bayes para o corpus Teses.

6.3 Classificador Naïve Bayes Especialista

6.3.1. Corpus Jornal

A tabela 6.3.1-1 mostra os resultados obtidos no processo de classificação utilizando-se o ensemble de classificadores especialistas com a combinação gramatical que obteve melhor resultado nos classificadores convencionais para o corpus Jornal.

Tabela 6.3.1-1 – Corpus Jornal - Substantivo + Nome Próprio + Adjetivo

	[illegible]				
	[illegible]				
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

A tabela 6.3.1-2 mostra os menores valores obtidos com o classificador Naïve Bayes Especialista para o corpus Jornal:

Tabela 6.3.1 -2 – Menores valores – Naïve Bayes especialista – Corpus Jornal

6.3.2. Corpus Teses

A tabela 6.3.2-1 mostra os valores referentes aos resultados obtidos no processo de classificação utilizando-se a combinação gramatical que obteve melhor resultado nos classificadores convencionais para o corpus Teses:

Tabela 6.3.2-1 – Corpus Teses - Substantivo + Verbo + Adjetivo

A tabela 6.3.2-2 mostra os menores resultados obtidos com o classificador Naïve Bayes Especialista para o corpus Teses:

Tabela 6.3.2-2 – Menores valores – Naïve Bayes especialista – Corpus Teses

6.4 Redes Neurais Especialistas

6.4.1. Corpus Jornal

As tabelas 6.4.1-1 a 6.4.1-3 mostram os valores referentes aos resultados obtidos no processo de classificação utilizando-se a combinação gramatical que obteve melhor resultado nos classificadores convencionais, os números de termos relevantes e números de neurônios na camada intermediária que apresentaram melhores resultados para o corpus Jornal.

O melhor resultado foi obtido com 60 termos e o coeficiente de correlação como nas redes neurais convencionais, porém com 8 neurônios na camada intermediária e não com 16 neurônios.

Tabela 6.4.1-1 – Corpus Jornal - Substantivo + Nome Próprio + Adjetivo – CC

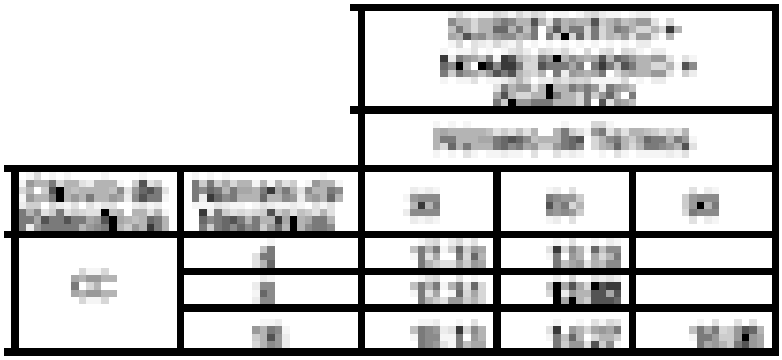
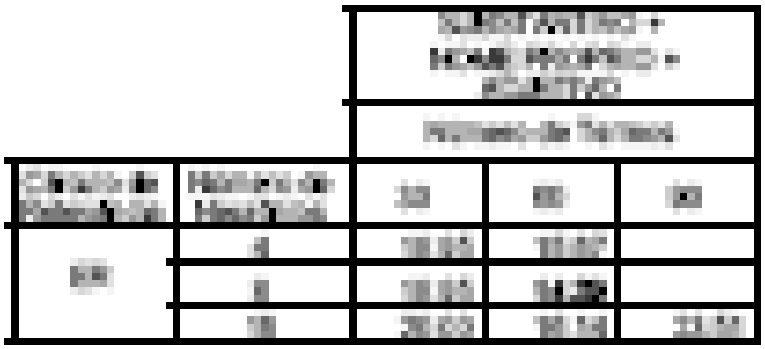
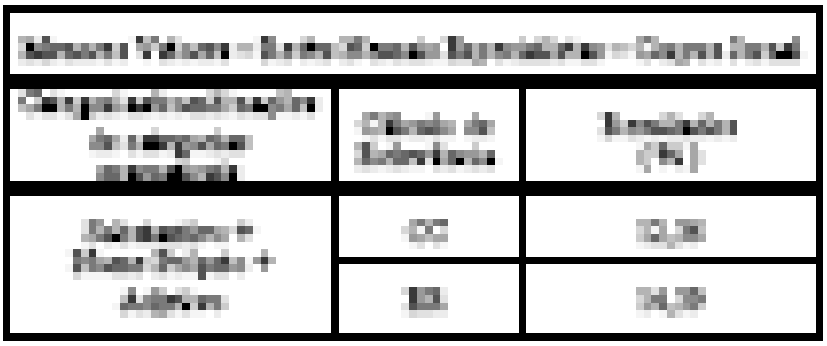


Tabela 6.4.1-2 – Corpus Jornal - Substantivo + Nome Próprio + Adjetivo – ER



A tabela 6.4.1-3 mostra os menores valores obtidos com as redes neurais especialistas para o corpus Jornal:

Tabela 6.4.1-3 – Menores valores – Redes Neurais especialistas – Corpus Jornal



6.4.2. Corpus Teses

As tabelas 6.4.2-1 a 6.4.2-3 mostram os valores referentes aos resultados obtidos no processo de classificação utilizando-se a combinação gramatical que obteve melhor resultado nos classificadores convencionais, os números de termos relevantes e números de neurônios na camada intermediária que apresentaram melhores resultados para o corpus Teses. O melhor resultado foi obtido com 60 termos e o coeficiente de

correlação como nas redes neurais convencionais, porém com 8 neurônios na camada intermediária.

Tabela 6.4.2-1 – Corpus Teses - Substantivo + Verbo + Adjetivo – CC

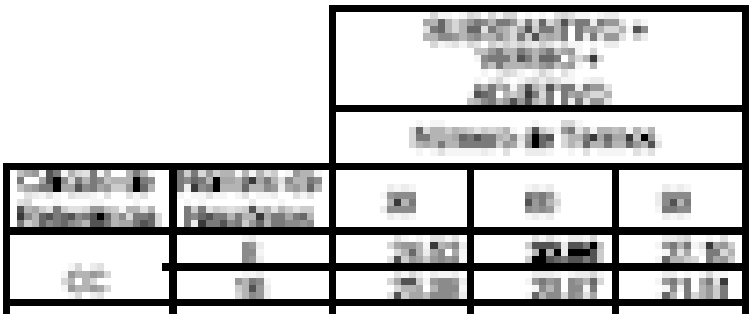
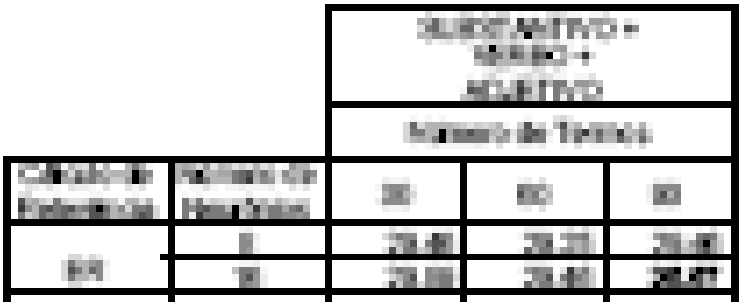


Tabela 6.4.2-2 – Corpus Teses - Substantivo + Verbo + Adjetivo – ER



A tabela 6.4.2-3 mostra os menores valores referentes aos resultados obtidos com as redes neurais especialistas para o corpus Teses:

Tabela 6.4.1-3 – Menores valores – Redes Neurais especialistas – Corpus Teses

[illegible]		
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

6.5 Resultados Comparativos

Nesta seção, serão apresentados os melhores resultados obtidos com os classificadores convencionais e com os classificadores especialistas para efeito de comparação e constatação da melhora obtida.

6.5.1. Naïve Bayes

6.5.1.1 Corpus Jornal

A tabela 6.5.1.1-1 mostra a tabela comparativa dos melhores valores obtidos pelos classificadores Naïve Bayes e Naïve Bayes especialista para o corpus Jornal.

Pela tabela observa-se que houve melhoria para os dois cálculos de relevância utilizados, porém obteve-se uma melhora mais significativa com o coeficiente de correlação.

Tabela 6.5.1.1-1 – Tabela comparativa - Naïve Bayes – Corpus Jornal

[illegible]		
[illegible]		
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

6.5.1.2 Corpus Teses

A tabela 6.5.1.2-1 mostra a tabela comparativa dos melhores valores obtidos pelos classificadores Naïve Bayes e Naïve Bayes especialista para o corpus Teses.

Como no caso do corpus Jornal, verifica-se a melhoria em ambos os cálculos de relevância. Pode-se observar no entanto, que o classificador NaÏve Bayes convencional obteve melhor resultado com o escore de relevância, e o Naïve Bayes especialista obteve melhor resultado com o coeficiente de correlação, passando a apresentar o melhor resultado.

Tabela 6.5.1.2-1 – Tabela comparativa - Naïve Bayes – Corpus Teses

[illegible]		
[illegible]		
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

6.5.2. Redes Neurais

6.5.2.1 Corpus Jornal

A tabela 6.5.2.1-1 mostra a tabela comparativa dos menores valores obtidos pelas redes neurais e redes neurais especialistas para o corpus Jornal.

Tabela 6.5.2.1-1 – Tabela comparativa – Redes Neurais – Corpus Jornal

[illegible]		
[illegible]		
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

As melhoras obtidas para ambos os cálculos de relevância foram equivalentes e o melhor resultado permaneceu para o coeficiente de correlação. Os melhores resultados foram obtidos com 60 termos, porém para as redes neurais convencionais foram necessários 16 neurônios na camada intermediária, enquanto que para as redes neurais especialistas foi obtido melhor resultado com apenas 8 neurônios na camada intermediária.

6.5.2.2 Corpus Teses

A tabela 6.5.2.2-1 mostra a tabela comparativa dos menores valores obtidos pelas redes neurais e redes neurais especialistas para o corpus Teses.

Assim como no corpus Jornal, houve melhora para ambos os cálculos de relevância e a redução de 16 para 8 neurônios na camada intermediária, permanecendo 60, o número de termos para o melhor resultado.

Tabela 6.5.2.2-1 – Tabela comparativa – Redes Neurais – Corpus Teses

[illegible]		
[illegible]		
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

7. Conclusões e Trabalhos Futuros

7.1 Sobre o Trabalho Realizado

O trabalho realizado permitiu verificar que as diferentes propostas para a classificação automática de documentos se mostraram bastante eficientes e contribuem com a literatura para classificação automática de textos em português.

Ambos os classificadores obtiveram muito bons resultados, com destaque para o classificador Naïve Bayes que, neste trabalho, obteve melhor desempenho do que as Redes Neurais Artificiais para ambos os corpora utilizados, evidenciado tanto pelos resultados obtidos , quanto pela rapidez do mesmo no processo de categorização.

Com relação à utilização de categorias gramaticais como meio de extração dos termos dos documentos, verificou-se que gera excelentes resultados. Cabendo ressaltar que as categorias gramaticais substantivo e adjetivo são importantes na extração dos termos e quando adicionada mais uma categoria apresenta ótimos resultados. No caso do corpus Teses, foi verificado que esta categoria a mais foi o *verbo* e para o corpus Jornal foi o *nome próprio*. Acredita-se que o verbo na linguagem do cotidiano tem menos relevância do que na linguagem técnica-científica, pois em geral são verbos de ligação e fazem parte de todos os textos, não sendo, portanto, considerados como discriminantes. Já nos textos do corpus Teses foi possível observar verbos que se repetem de acordo com a área.

Com relação à obtenção de melhores resultados para o corpus Jornal em relação ao corpus Teses, acredita-se que a linguagem de jornal em geral procura seguir um padrão estabelecido pelos editores, é simples e clara de maneira a atingir grande parte da população. Já a linguagem utilizada nos resumos das dissertações de mestrado e teses de doutorado, são linguagens particulares, ou seja, cada aluno gera o seu resumo fazendo uso de um linguajar próprio e portanto, diferenciado, apresentando maior dificuldade para se estabelecer um padrão.

Como proposta de busca do resultado ótimo, ou seja 100% de acerto na classificação, os classificadores especialistas se mostraram como um caminho interessante, já que após um excelente resultado obtido nos classificadores convencionais, conseguiram uma melhora significativa. Em particular para o classificador Naïve Bayes que chegou ao resultado próximo de 95% de acerto. Para as redes neurais que tem um processo mais lento de classificação, o tempo de processamento pôde ser reduzido, pois o melhor resultado obtido foi conseguido com um menor número de neurônios na camada intermediária.

7.2 Sobre o Trabalho que Pode ser Realizado

Um trabalho que pode ser realizado é inserir novos textos ao corpus Teses, submeter o novo corpus ao processo de classificação para verificar se ao aumentar a coleção, consegue-se melhorar as características e portanto estabelecer melhores padrões, obtendo conseqüentemente melhores resultados na classificação, já que os mesmos se mostraram inferiores ao corpus Jornal.

Outro trabalho que pode ser feito é manter o corpus Teses, porém aumentar o número de termos nos vetores locais com o intuito de melhorar as características de cada categoria. Depois submeter o corpus ao processo de categorização para verificar se os resultados da classificação dos corpora Jornal e Teses se aproximam mais, evidenciando a necessidade de um tratamento diferenciado devido aos tipos de textos.

Acredita-se que o uso de um classificador fuzzy na implementação do ensemble de classificadores pode se mostrar uma opção interessante para avaliação de possível melhoria dos resultados.

Outra proposta de trabalho seria utilizar um classificador fuzzy para os mesmos corpora, utilizando as mesmas metodologias, visando comparar os resultados entre

diferentes classificadores e permitindo avaliar o quanto cada classificador se adequa melhor a tarefa de categorização.

Para trabalhos futuros seria interessante levar em conta na classificação a evolução temporal dos textos.

Referências Bibliográficas:

ARANHA, C. N. , 2007, *Uma Abordagem de Pré-processamento para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional*. Tese de Doutorado, Pontifícia Universidade Católica, Rio de Janeiro, RJ, Brasil.

ARCOVERDE, J. M. A. , 2007, *Indução em Filtros Linguisticamente Motivados em Recuperação da Informação*. Dissertação de mestrado, Universidade de São Paulo, SP, Brasil.

BAEZA -YATES, R., RIBEIRO-NETO, B., 1999, *Modern Information Retrieval*. 1 ed., United Kingdom, Addison-Wesley.

BICK, E. , 2000 "The *parsinsg System PALAVRAS: Automatic Gramatical Analysis of Portuguese in a Constraint Grammar Framework*", Århus University. Århus: Århus University Press.

BRAGA, A. P.; CARVALHO, A. P. L. F.; LUDERMIR, T.B. , 2000 *Redes Neurais Artificiais: Teoria e Aplicações*. Livro Técnico Científico. Rio de Janeiro – Brasil

BUITELAAR, P.; DECLERCK, T. , 2003 *Linguistic Annotation for the Semantic Web* In: Siegfried Handschuh and Steefen Staab (eds.) Annotation for The Semantic Web, Frontiers in Artificial Intelligence and Applications Series, Vol 96, IOS Press.

CAMARGO, Y. B. , 2007, *Abordagem Lingüística em Classificação Automática de Textos em Português*. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, RJ, Brasil.

CORREA, R. F. , 2002 *Categorização de Textos Utilizando Redes Neurais – Análise Comparativa com Técnicas Não-conexionistas*. Dissertação de mestrado, Universidade Federal de Pernambuco, PE, Brasil.

DOMINGOS, P.; PAZZANI, M., 1997 *On The Optimality of the Simple Bayesian Classifier Under Zero-one Loss*. Machine Learning, 29 (2/3), 103, 1997.

GASPERIN, C.; VIEIRA, R.; GOULART, R. and QUARESMA, P. , 2003 *Extracting XML Syntactic Chunks from Portuguese Corpora*, Proceedings of the TALN 2003 Workshop Natural Language Processing of Minority Languages and Small Languages – Batz—sur—Mer France June 11 - 14.

GEAN, C. C.; KAESTNER, C. A. A. , 2004 *Classificação Automática de Documentos usando Subespaços Aleatórios e Conjuntos de Classificadores*. In: TIL 2004 - 2º WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, Salvador, Brasil. Anais do SBC 2004 v.1, p.1-8.

HAYKIN, S. , 2001 *Redes Neurais: Princípios e prática* Editora: Bookman. 2ª Edição, Brasil.

JACKSON, P.; MOULINIER, I*. Natural Language Processing for On Line Applications; Text Retrieval; Extraction and Categorization*. Amsterdam / Philadelphia: John Benjamins Publ., 2002.

KORFHAGE, R. R., 1997 *Information Retrieval and Storage*. New York: John Wiley & Sons, 1997. 349p.

KOWALSKI, G. *Information Retrieval Systems – Theory and Implementation*, Kluwer Academic Publishers, 1997.

LAN, M., TAN, C. L., SU, J. LOW, H. B. , 2007 *Text Representation for Text Categorization: A Case of Study.* National University of Singapure.

LEWIS, D. D., RINGUETTE, M. , 1994 *A Comparison of Two Learning Algorithms for Text Categorization* In: Symposium on Document Analysis and IR, ISRI, Las Vegas.

MITCHELL, T.M. *Machine Learning*. WCB/McGraw-Hill, 1997.

McCALLUM, A. K.; NIGAM, K. *A Comparison of event models for naive Bayes text classification*. In: Proceedings of the 1st AAAI Workshop on Learning for Text Categorization, pages 41-48, Madison, USA, 1998.

MOENS, M. F., 2000 *Automatic Indexing and Abstract of Documents Texts*. Masaachusetts: Kluwer Academic Publishers.

NG, H. T. , GOH, W. B., LOW, K. L., 1997 *Feature Selection, Perceptron Learning and a Usability Case Study for Text Categorization*. Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval, 1997, Philadelphia. Philadelphia, PA, USA.

RIZZI, C. B., WIVES, L. K., OLIVEIRA, J. P. M., ENGEL, P. M., (2000) *Fazendo uso da Categorização de Textos em Atividades Empresariais.* In: International Symposium on

Knowledge Management/Document Manegement – ISKDM/ DM 2000, pp. 251 – 268, Curitiba. Brasil.

RUMELHART, D. E., McCLELLAND, J. L. and The PDP Research Group, *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*, vol. 1: Foundations, MIT Press, Cambridge, Masschusetts, USA, 1986

RUSSEL, S. J., NORVIG, P., 1995 Artificial Intelligence A Modern Approach Rio de Janeiro : Prentice-Hall do Brasil.

SALTON, G., MacGILL, M., 1983, Introduction to Modern Information Retrieval. New York, Ed. McGraw-Hill.

SALTON, G., BUCKLEY, C. *Improving Retrieval Performance by Relevance Feedback.* Ithaca, New York. 1987. (Technical Report).

SALTON, G. *Automatic Text Processing*, Addison – Wesley, 1988.

SALTON, G., Wong, A., e Yang, C. S. ,1975 *A Vector Space Model for Automatic Indexing*. In Readings in Information Retrieval, K.Sparck Jones and P.Willet, eds.,Morgan Kaufmann Publishers, Inc., San Francisco

SILVA, C. F., 2004 *Uso de Informações Lingüísticas na etapa de pré-processamento em Mineração de Textos*. Dissertação de M.Sc. Universidade do Vale do Rio dos Sinos. São Leopoldo – Rio Grande do Sul.

TAN, A. *Text mining: The State of the Art and the Challenges*. In: Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases - PAKDD'99, pages 65-70, Beijing, April 1999.

VILELA, R., SIMÕES, A., BICK, E., ALMEIDA, J. J., 2005 Representação em XML da Floresta Sintática. XATA 2005 – III Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas, Braga, Portugal.

WANGENHEIM, A. V., 2006 *Reconhecimento de Padrões (Apostila)*. Santa Catarina, Brasil.

WIENER, E., PEDERSEN, L.O., WEIGEND, A.S. , 1995 *A Neural Network Approach to Topic Spotting*. In: Proceedings of the Symposium on Document Analysis and Information Retrieval, pp.317-332, Las Vegas, US.

WITTEN, I. H.; FRANK, E. *Data mining: Practical Machine Learning tools and tecniches with Java implementations*. Academic Press, 2000.

YANG, Y.; LIU, X. 1999 *A.Re-examination of Text Categorization Methods* In: Proceedings of SIGIR-99 22nd ACM International Conference on Research and Development in Information Retrieval, Berkley, US, 1999.

YANG, Y; PEDERSON,J, 1997 : *A.Comparative Study on Feature Selection in Text Categorization. In Proceedings of 14th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, US(1997), 412-420*