UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

OESLEI TABORDA RIBAS

CLASSIFICAÇÃO DE SITES A PARTIR DAS ANÁLISES ESTRUTURAL E TEXTUAL

DISSERTAÇÃO

CURITIBA

2013

OESLEI TABORDA RIBAS

CLASSIFICAÇÃO DE SITES A PARTIR DAS ANÁLISES ESTRUTURAL E TEXTUAL

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do grau de "Mestre em Computação" – Área de Concentração: Sistemas de Informação.

Orientador: Celso Antônio Alves Kaestner

CURITIBA

2013

Dados Internacionais de Catalogação na Publicação

R482 Ribas, Oeslei Taborda
Classificação de sites a partir das análises estrutural e textual / Oeslei Taborda Ribas. — 2013.
125 f. : il. ; 30 cm

Orientador: Celso Antônio Alves Kaestner.
Dissertação (Mestrado) – Universidade Tecnológica Federal do Paraná. Programa de Pós-graduação em Computação Aplicada, Curitiba, 2013.
Bibliografia: f. 97-103.

1. Sites da web – Avaliação e classificação. 2. Processamento de textos (Computação). 3. Aprendizado do computador. 4. Redes neurais (Computação). 5. HTML (Linguagem de marcação de documento). 6. Métodos de simulação. 7. Computação – Dissertações. I. Kaestner, Celso Antônio Alves, orient. II. Universidade Tecnológica Federal do Paraná. Programa de Pós-graduação em Computação Aplicada. III. Título.

CDD (22. ed.) 004

Biblioteca Central da UTFPR, Campus Curitiba

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Curitiba
Diretoria de Graduação e Educação Profissional
Departamento Acadêmico de Informática – DAINF
Programa de Pós-Graduação em Computação Aplicada - PPGCA
Mestrado Profissional

Título da Dissertação

"CLASSIFICAÇÃO DE SITES A PARTIR DAS ANÁLISES ESTRUTURAL E TEXTUAL".

por

Oeslei Taborda Ribas

Esta dissertação foi apresentada como requisito parcial à obtenção do grau de MESTRE EM COMPUTAÇÃO APLICADA - Área de Concentração: Ciência da Computação pelo PPGCA - Programa de Pós-Graduação em Computação Aplicada - Mestrado Profissional – da Universidade Tecnológica Federal do Paraná - UTFPR - Câmpus Curitiba, às 14:00 horas do dia 28 de agosto de 2013. O trabalho foi aprovado pela Banca Examinadora, composta pelos professores:

Prof. **Celso Antônio Alves Kaestner, Dr.**
presidente - (UTFPR - CT)

Prof. **Julio Cesar Nievola, Dr.**
(PUC-PR)

Prof. **Gerson Linck Bichinho, Dr,**
(PUC-PR)

Prof. **Robinson Vida Noronha, Dr.**
(UTFPR - CT)

DAINF
Departamento Acadêmico de Informática - DAINF
Programa de Pós-Graduação em Computação Aplicada - PPGCA
Câmpus Curitiba
Bloco B - ramal: 3310-**4644** • Fax: 3310-**4646**

Dedico este trabalho à minha família.
Sem ela nada disto seria possível.

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus por ter permitido que eu chegasse até aqui. Por toda a sabedoria, determinação e força na escolha da direção correta a tomar e por dar-me a oportunidade de conhecer tantas pessoas boas ao longo da minha vida. Agradeço a Ele todas as vitórias e conquistas alcançadas.

Agradeço à minha família, por toda a felicidade, carinho, compreensão, incentivo e apoio incondicional. Agradeço em especial a minha querida mãe Sirlei Aparecida de Paula, minha irmã Priscila Taborda Ribas, minha namorada Jaqueline de Fatima Telma e minha amada filha Laura Bellincanta Taborda Ribas pela compreensão do tempo que tive que ausentar-me para realizar este trabalho.

Ao professor Celso Antônio Alves Kaestner pela excelente orientação. Foi um privilégio tê-lo como orientador. Sou grato por todo o incentivo e pelos ensinamentos que são os alicerces deste trabalho. Sou grato aos professores Gustavo A. G. Lugo, João A. Fabro e Roberto C. Betini pelas sugestões e correções efetuadas nos seminários de acompanhamento do mestrado. Agradeço também ao professor Júlio C. Nievola da PUC-PR e a todos os professores do PPGCA por todo o conhecimento que adquiri em suas disciplinas.

Agradeço ao grande amigo e professor Hermano Pereira por ter me apresentado à pesquisa acadêmica. Ao José Roberto Andrade Júnior pelo apoio e ajuda com as ilustrações inseridas neste trabalho. Sou grato também ao Jean Lima Pierobom por toda a boa vontade e ajuda ao longo desta caminhada. Agradeço também a todos os colegas de trabalho que me incentivaram durante este período: Osvaldo M. Cavalieri, Murilo A. Tosatti, Marco A. Bonato, José A. Salazar, Marcelo F. Guimarães, Luana Boganika Acosta, Fellipe M. Veiga, Airton Kuada, Andre L. de S. Paula e Anderson Augustinho.

À CELEPAR por todo o apoio na realização deste mestrado e principalmente pela contribuição exercida na formação de meu conhecimento técnico e profissional.

RESUMO

RIBAS, Oeslei T. CLASSIFICAÇÃO DE SITES A PARTIR DAS ANÁLISES ESTRUTURAL E TEXTUAL. 126 f. Dissertação – Programa de Pós-Graduação em Computação Aplicada, Universidade Tecnológica Federal do Paraná. Curitiba, 2013.

Com a ampla utilização da *web* nos dias atuais e também com o seu crescimento constante, a tarefa de classificação automática de sítios *web* têm adquirido importância crescente, pois em diversas ocasiões é necessário bloquear o acesso a sítios específicos, como por exemplo no caso do acesso a sítios de conteúdo adulto em escolas elementares e secundárias. Na literatura diferentes trabalhos têm surgido propondo novos métodos de classificação de sítios, com o objetivo de aumentar o índice de páginas corretamente categorizadas. Este trabalho tem por objetivo contribuir com os métodos atuais de classificação através de comparações de quatro aspectos envolvidos no processo de classificação: algoritmos de classificação, dimensionalidade (número de atributos considerados), métricas de avaliação de atributos e seleção de atributos textuais e estruturais presentes nas páginas *web*. Utiliza-se o modelo vetorial para o tratamento de textos e uma abordagem de aprendizagem de máquina clássica considerando a tarefa de classificação. Diversas métricas são utilizadas para fazer a seleção dos termos mais relevantes, e algoritmos de classificação de diferentes paradigmas são comparados: probabilista (Naïve Bayes), árvores de decisão (C4.5), aprendizado baseado em instâncias (KNN - K vizinhos mais próximos) e Máquinas de Vetores de Suporte (SVM). Os experimentos foram realizados em um conjunto de dados contendo sítios de dois idiomas, Português e Inglês. Os resultados demonstram que é possível obter um classificador com bons índices de acerto utilizando apenas as informações do texto âncora dos *hyperlinks*. Nos experimentos o classificador baseado nessas informações atingiu uma Medida-F de 99.59%.

Palavras-chave: classificação de textos, classificação de sítios *web*, aprendizagem de máquina.

ABSTRACT

RIBAS, Oeslei T. WEBSITE CLASSIFICATION USING STRUCTURAL AND TEXTUAL ANALYSIS. 126 f. Dissertação – Programa de Pós-Graduação em Computação Aplicada, Universidade Tecnológica Federal do Paraná. Curitiba, 2013.

With the wide use of the web nowadays, also with its constant growth, task of automatic classification of websites has gained increasing importance. In many occasions it is necessary to block access to specific sites, such as in the case of access to adult content sites in elementary and secondary schools. In the literature different studies has appeared proposing new methods for classification of sites, with the goal of increasing the rate of pages correctly categorized. This work aims to contribute to the current methods of classification by comparing four aspects involved in the classification process: classification algorithms, dimensionality (amount of selected attributes), attributes evaluation metrics and selection of textual and structural attributes present in webpages. We use the vector model to treat text and an machine learning classical approach according to the classification task. Several metrics are used to make the selection of the most relevant terms, and classification algorithms from different paradigms are compared: probabilistic (Naïve Bayes), decision tree (C4.5), instance-based learning (KNN - K-Nearest Neighbor) and support vector machine (SVM). The experiments were performed on a dataset containing two languages, English and Portuguese. The results show that it is possible to obtain a classifier with good success indexes using only the information from the anchor text in hyperlinks, in the experiments the classifier based on this information achieved 99.59% F-measure

Keywords: text classification, website classification, machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ASCII	*American Standard Code for Information Interchange*
AD	Árvore de Decisão
AM	Aprendizado de Máquina
API	*Application Programming Interface*
ARFF	*Attribute-Relation File Format*
ART	*Adaptive Resonance Theory*
CNN	*Cellular neural networks*
FN	Falso Negativo
FP	Falso Positivo
HTML	*HyperText Markup Language*
HTTP	*Hypertext Transfer Protocol*
ICA	*Independent Component Analysis*
IDF	*Inverse Document Frequency*
IETF	*Internet Engineering Task Force*
IP	*Internet Protocol*
KNN	*K-Nearest Neighbor*
MLP	*Multilayer Perceptron*
OSH	*Optimal Separating Hyperplane*
PCA	*Principal Component Analysis*
RFC	*Request for Comments*
RI	Recuperação de Informação
RNA	Rede Neural Artificial
SMO	*Sequential Minimal Optimization*
SOM	*Self-Organizing Map*
SVM	*Support Vector Machine*
TCP	*Transmission Control Protocol*
TF	*Term Frequency*
TFIDF	*Term Frequency Inverse Document Frequency*
TTL	*Time to Live*
URL	*Uniform Resource Locator*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
VSM	*Vector Space Model*

SUMÁRIO

1 INTRODUÇÃO

O número de páginas disponíveis na internet aumenta a cada dia. Um estudo realizado por Gulli e Signorini (2005) informa que existem mais de 10^{10} páginas indexadas pelos mecanismos de buscas. Uma estimativa, com metodologia diferente, é realizada diariamente pelo projeto WorldWideWebSize.com, que informa que a quantidade de páginas da *web* indexável já ultrapassou o valor de 8×10^9 (WORLDWIDEWEBSIZE, 2012). Já segundo a Google (2008) o número de páginas únicas indexadas ultrapassou a marca de 10^{12}, e diariamente surgem mais de 10^9 páginas novas. O crescimento da internet implica no aumento no número de páginas com conteúdo inadequado para determinados públicos, como as de conteúdo adulto para crianças e adolescentes.

Devido a esses fatores surge a necessidade de soluções automáticas que possibilitem controlar ou até mesmo negar o acesso a certos gêneros de sites. Porém, restringir o acesso a esses sites sem interferir no acesso aos demais sites não é uma tarefa simples de ser realizada. Para tentar resolver este problema diversas soluções têm sido propostas. Essas soluções são os filtros de conteúdo que, com base em algum mecanismo, definem se um acesso pode ou não ser realizado.

Um dos mecanismos de filtragem mais simples é a chamada "lista branca". Esse mecanismo consiste em criar manualmente uma lista de sites autorizados. Somente é permitido o acesso a sites que estejam nessa lista, e o acesso aos endereços que não estão na lista é bloqueado. O problema desta solução é a existência de falsos positivos, que ocorrem quando o acesso a um site de conteúdo inofensivo é negado devido a não constar na relação de endereços permitidos (LEE et al., 2005).

Um mecanismo de funcionamento oposto a "lista branca" é a chamada "lista negra". Consiste em criar uma lista de sites não autorizados, aqueles que possuem conteúdo que deve ser bloqueado. Todo o acesso é liberado, sendo bloqueado apenas o acesso aos sites que compõem a lista. A limitação dessa solução é a ocorrência de falso negativo, que acontecem quando é permitido o acesso a um endereço de conteúdo inadequado devido a ele não constar na relação

de bloqueio (HAMMAMI et al., 2006).

Ambas as soluções “lista branca” e “lista negra” sofrem do mesmo problema, que é manter a lista atualizada. A manutenção dessas listas através de um trabalho manual é impraticável e ineficiente, devido ao tamanho da internet e também pelo fato de milhões de páginas surgirem e desaparecerem todo dia.

Um terceiro mecanismo também utilizado é o bloqueio por palavras. Consiste em criar um lista de palavras que são comumente encontradas nos sites que se deseja bloquear. Ao receber uma requisição de acesso se faz uma verificação de todo o conteúdo da página solicitada, procurando por palavras que estão na lista: caso seja encontrada alguma palavra então o acesso é negado; caso contrário o acesso é liberado. A limitação desta solução é que não se leva em consideração o contexto da palavra. Deste modo, sites autorizados que contenham uma única palavra da lista tem o seu acesso negado, enquanto sites que deveriam ser bloqueados mas não possuem nenhuma palavra na lista tem o seu acesso autorizado. Assim, o bloqueio por palavras sofre do problema de ocorrências de falsos positivos e falsos negativos (CAULKINS et al., 2006).

Uma solução de bloqueio de conteúdo inadequado mais eficiente que as apresentadas anteriormente é a classificação automática de sites indesejados, que pode ser vista como uma instância do problema de categorização de sites. A categorização de site consiste em realizar uma análise do conteúdo das páginas e com base nesta análise atribuir um rótulo pré-definido ao site (QI; DAVISON, 2009).

Na literatura vários trabalhos têm sido propostos para melhorar o índice de acerto dos atuais métodos de classificação automática de sites. Esses trabalhos, em geral, utilizam as informações do conteúdo textual da página usando técnicas de Processamento de Linguagem Natural. Entende-se por conteúdo textual aquele que é visualmente disponibilizado para o usuário. Os autores, em sua maioria, propõem novos métodos de classificação que consistem na combinação de um mais algoritmos de aprendizado de máquina, ou então pequenas alterações nesses algoritmos para obter uma performance melhor nesse cenário de uso. Entretanto poucos trabalhos se preocupam com os atributos utilizados para a classificação.

As páginas de internet, diferentemente dos textos comuns, possuem várias informações que podem ser utilizadas na classificação, como os marcadores da linguagem HTML (*HyperText Markup Language*) e os *Hyperlinks*. Essas informações, que representam o conteúdo estrutural da página, se adequadamente selecionadas podem ajudar a aumentar o índice de sítios corretamente classificadas pelo métodos atuais, elevando os índices de acertos para valores aceitáveis para o problema proposto.

Embora existam várias informações contidas nos sítios da internet, nem todas contribuem positivamente para a classificação, algumas podem ter um efeito contrário prejudicando o classificador. Outro aspecto que pode prejudicar o desempenho é a dimensionalidade, que é o número total de atributos utilizados pelo classificador. Deste modo, se faz necessário um estudo que demonstre quais são essas informações - atributos e dimensionalidade - que podem auxiliar o classificador, bem como quais classificadores podem melhor aproveitá-las.

Neste trabalho é realizado um estudo empírico envolvendo classificadores, dimensionalidade, métricas para avaliação de atributos, e seleção de atributos textuais e estruturais. O objetivo deste estudo é encontrar uma combinação desses quatro elementos que permita atingir índices maiores de instâncias corretamente classificadas, sem que seja necessário alterações nos atuais algoritmos de aprendizado de máquina. Este trabalho pretende contribuir com os estudos na área de classificação automática de conteúdo *web*.

1.1 MOTIVAÇÃO

Com a popularização da internet, e o livre acesso a informação que ela proporciona, tornou-se um dos problemas atuais de nossa sociedade a questão da visualização de páginas de conteúdo impróprio por crianças e adolescentes. O fato de conteúdo adulto poder ser acessado facilmente pelas crianças é preocupante para muitos pais. Todavia, não apenas os pais se preocupam com este fato, empregadores também não gostam de pensar que este conteúdo inadequado pode ser acessado por seus funcionários durante o horário de trabalho.

Além disso, é preocupante o fato de sites com conteúdo adulto serem utilizados como vetor para propagação de programas maliciosos, como vírus e programas espiões. Segundo dados do relatório produzido pela Symantec (2010), fabricante de soluções de anti-vírus, os sites com conteúdo pornográfico são responsáveis por 49% das infecções que ocorrem ao navegar na internet. Segundo a empresa, os atacantes estariam utilizando esses sites devido ao grande volume de acesso que eles possuem.

Todos esses fatores evidenciam a necessidade de soluções que possibilitem identificar e coibir o acesso a sites de conteúdo pornográficos. Na literatura inúmeros trabalhos de classificação de conteúdo *web* têm sido propostos para atingir esse objetivo. Entretanto, esses trabalhos não apresentam estudos envolvendo os diferentes classificadores, atributos, dimensionalidade e métrica de avaliação de atributos. O trabalho segue o método empírico, ou seja, os diversos elementos de um sistema de classificação - algoritmo, atributos, dimensões e métrica de avaliação - são testados sobre uma base de sites.

O procedimento utilizado para classificação de conteúdo atualmente encontra-se em fase de implantação, e deverá ser empregado para a filtragem de sítios acessados a partir das escolas públicas do Estado do Paraná, cujo acesso é responsabilidade da Companhia de Tecnologia da Informação e Comunicação do Paraná (CELEPAR). Esta rede é composta por laboratórios de informática em mais de 2.000 escolas públicas de ensino primário e secundário. Embora esteja sendo aplicado para esta situação, o método é genérico o suficiente para ser utilizado em outros ambientes ou para a classificação de páginas de modo geral.

1.2 OBJETIVO GERAL

Este trabalho tem por objetivo contribuir para a solução do problema da classificação automática de sites - que pode ser entendido como um problema de classificação de textos - com base em experimentos empíricos envolvendo classificadores, métricas de seleção de atributos, dimensionalidade e seleção de atributos textuais e estruturais.

1.3 OBJETIVOS ESPECÍFICOS

Os objetivos específicos deste trabalho são:

- Avaliar o desempenho dos classificadores usando atributos selecionados por diferentes métricas e também em dimensões distintas.
- Comparar os métodos de classificação em dois idiomas distintos, Português e Inglês.
- Comparar no problema proposto o índice de classificação de diversos algoritmos de classificação que seguem diferentes paradigmas: probabilista (Naïve-Bayes), árvores de decisão (C4.5), K vizinhos mais próximos (KNN), redes neurais artificiais (MLP) e Máquinas de Vetores de Suporte (SVM).
- Avaliar o desempenho dos classificadores usando diferentes atributos estruturais do HTML (*HyperText Markup Language*).
- Encontrar a dimensionalidade (número de atributos a considerar) ideal a ser utilizada neste gênero de classificação.

1.4 CONTRIBUIÇÕES DO TRABALHO

O presente trabalho apresenta contribuições relacionadas aos seguintes aspectos:

- Realização de experimentos relacionados a classificação de conteúdo *web* segundo a variação de quatro dimensões: algoritmos de classificação, atributos, dimensionalidade (total de atributos a considerar) e métrica de avaliação de qualidade dos atributos. Os resultados mostram que com a escolha adequada desses quatro elementos é possível se obter um classificador com um índice alto de acertos para o problema em questão.

- Construção de uma base de dados de páginas da internet que pode ser utilizado em pesquisas futuras sobre classificação de sites. Uma das dificuldades enfrentadas pelos pesquisadores é a falta de conjunto de dados públicos e amplamente utilizado pela comunidade o que acaba desestimulando a realização de pesquisas na área e o avanço das técnicas atuais. Com a disponibilização dessas informações pretende-se estimular o desenvolvimento de novas pesquisas.

- Desenvolvimento de um classificador confiável que será utilizado em um mecanismo de filtragem de acesso a sítios pelas escolas públicas do Estado do Paraná. O classificador desenvolvido é genérico o suficiente para ser implantado em outros ambientes para outras finalidades.

Os resultados obtidos na primeira parte dos experimentos foram publicados na forma de um artigo científico no IV International Workshop on Web and Text Intelligence (WTI) (RIBAS; KAESTNER, 2012), o qual mostra um comparativo do desempenho dos classificadores, dimensionalidade e métrica de avaliação de atributos utilizando apenas as informações textuais das páginas, sem utilizar as informações estruturais. Tais comparativos também são incluídos nesta dissertação.

1.5 ESTRUTURA DO TRABALHO

O Capítulo 2 mostra alguns conceitos relacionados à classificação automática de textos, explicando assuntos como pré-processamento e seleção de atributos textuais e estruturais. Em seguida, são apresentados os algoritmos de classificação utilizados neste trabalho e as métricas usadas para avaliação desses algoritmos. Por fim, é abordado especificamente a questão de classificadores aplicados a categorização de conteúdo *web*.

O Capítulo 3 aborda a metodologia adotada, e as etapas para o desenvolvimento desta pesquisa. Neste Capítulo também é especificado o ambiente computacional utilizado para execução dos experimentos.

O conjunto de dados construído para execução dos experimentos é abordado no Capítulo

4. São detalhados o processo de construção desse conjunto e também estatísticas sobre os dados nele presentes. Além disso são apresentadas informações sobre o conjunto de dados utilizados por outros pesquisadores.

Os experimentos realizados neste trabalho e a análise dessas informações são detalhadas no Capítulo 5. O primeiro grupo de experimentos utiliza-se apenas das informações textuais presentes na página. O segundo grupo de experimentos analisa o desempenho dos classificados utilizando as informações textuais e estruturais da página. O resultado de cada um dos experimentos é analisado individualmente.

Por fim, o Capítulo 6 apresenta a conclusão sobre o trabalho realizado e as perspectivas de trabalhos futuros.

2 CLASSIFICAÇÃO AUTOMÁTICA DE SÍTIOS WEB

Neste Capítulo serão apresentados os tópicos relacionados à classificação de textos. Primeiramente é apresentado o conceito de classificação automática de textos, depois disso é abordada a classificação automática de conteúdos *web*. Na sessão seguinte são apresentados os atributos estruturais e metadados e em seguida é explicado como é realizado o pré-processamento de texto e a seleção dos atributos textuais. Um exemplo de classificação usando os atributos estruturais e textuais é apresentado. Em seguida, são abordados os algoritmos de classificação utilizados para a tarefa de categorização de conteúdo, e apresentadas as métricas para avaliação desses algoritmos. Por fim, é apresentado o estado da arte no que diz respeito a categorização de sites e por último as considerações finais.

2.1 CLASSIFICAÇÃO AUTOMÁTICA DE TEXTOS

Classificação de textos, também chamada de categorização de textos, é a tarefa de organizar automaticamente um grupo de documentos dentro de categorias previamente definidas. Essa tarefa envolve tópicos relacionados às áreas de RI (Recuperação de Informação) e AM (Aprendizado de Máquina) (YANG, 1999).

Na maioria dos casos a tarefa de aferir rótulos a um conjunto de documentos é uma tarefa simples para um humano, porém para um computador é uma tarefa complexa. No caso do uso de procedimentos automáticos de classificação, a tarefa é complexa devido ao fato de estar relacionada a análise do conteúdo dos documentos (BORKO; BERNICK, 1963).

Segundo Sebastiani (2002) a categorização de texto pode ser formalizada como a tarefa de aproximar uma função alvo desconhecida $F : D \times C \rightarrow \{V,F\}$ (V = verdadeiro e F = falso) que descreve como os documentos devem ser classificados, de acordo com o conhecimento de um especialista. Essa função é chamada de classificador, na qual $C = \{c_1,...,c_{|c|}\}$ é um conjunto finito de categorias e D é um conjunto finito de documentos. Caso $F(d_j,c_i) = V$, então d_j é chamado de exemplo positivo de c_i, enquanto $F(d_j,c_i) = F$ é chamado de exemplo negativo de c_i.

Sebastiani (2002) também afirma que dependendo da aplicação, a classificação de textos pode ser tanto uma tarefa de atribuir um único rótulo ao documento ou de atribuir múltiplos rótulos. No caso de atribuir um único rótulo, apenas um $c_i \in C$ deve ser designado para cada $d_j \in D$. No caso de atribuição de múltiplos rótulos, qualquer número n em que $0 \leq n \leq |C|$ de categorias pode ser designado para um documento $d_j \in D$. O problema de classificação com mais de um rótulo geralmente é tratado como um problema de classificação com $|C|$ classificadores binários independentes. Neste caso, o classificador consiste da composição de $|C|$ classificadores binários.

Um classificador c_i é gerado automaticamente por um processo indutivo, no qual apenas observando as características de um conjunto de documentos previamente classificados como pertencentes a c_i é capaz de descobrir a classe de um documento não conhecido anteriormente. Para realizar a construção do classificador é necessário um conjunto de documentos S em que o valor de cada $F(d_j, c_i)$ seja conhecido para cada $(d_j, c_i) \in S \times C$, ou seja, uma base rotulada de treinamento. Geralmente o conjunto de documentos S é dividido em três subconjuntos: *Tr* (conjunto de treinamento), *Va* (conjunto de validação) e *Te* (conjunto de teste). O *Tr* é o conjunto no qual o classificador irá realizar as observações necessárias para o seu aprendizado. O *Va* é utilizado pelo engenheiro para testar os ajustes do classificador. Por fim, a performance do classificador é finalmente avaliada no *Te* (SEBASTIANI, 2002).

Alguns algoritmos de classificação podem levar horas ou dias para serem treinados. Após treinados o tempo de classificação de um conjunto de novas instâncias é rápido comparado com o tempo de treinamento. Em algumas situações o tempo de *download* de uma página *web*, ou seja, o tempo que o servidor *web* demora para enviar o conteúdo da página ao cliente (*browser*), pode ser superior ao tempo de sua classificação propriamente dito.

2.2 CLASSIFICAÇÃO AUTOMÁTICA DE PÁGINAS WEB

A classificação automática de páginas da internet, também conhecida por categorização de páginas *web*, pode ser definida como o processo de designar a uma página *web* um ou mais rótulos previamente definidos (MITCHELL, 1997).

A classificação de páginas *web* pode ser feita de acordo diversos objetivos: classificação por assunto, classificação funcional, classificação por sentimento entre outros gêneros existentes de classificação. A classificação por assunto tem por objetivo encontrar qual é o assunto ou tema principal abordado pela página, por exemplo: julgar se uma página é sobre “Artes”, “esportes” ou “economia”. A classificação funcional está preocupada em localizar qual é a

função da página, por exemplo: “página pessoal” ou “página de curso”. Já a classificação por sentimento está focada na opinião do autor sobre o assunto tratado na página (QI; DAVISON, 2009).

Este trabalho está focado na classificação por assunto e os demais gêneros de classificação não serão abordados ao longo desta dissertação. Será pesquisada a classificação por assunto por esta permitir a categorização de páginas de acordo com o seu tema, desta forma, possibilitando o desenvolvimento de um classificador capaz de atuar em um filtro de conteúdo *web*.

Comparada com a classificação normal de texto a classificação de páginas *web* é diferente em vários aspectos. Primeiro, a classificação tradicional de texto tipicamente é executada em uma estrutura textual bem definida pelo estilo do autor, enquanto uma coleção de páginas *web* não possui essa característica. Segundo, páginas *web* são documento semi-estruturados em HTML (*HyperText Markup Language*), elas necessitam serem renderizadas para poderem ser visualizadas pelos usuários. Finalmente, em documentos *web* existem *hyperlinks* que o conectam para outros documentos; embora não seja uma característica exclusiva da *web* ela é a principal diferença para um documento de texto comum. Devido a esses aspectos podemos considerar que a classificação de documentos *web* é um problema diferente em relação a classificação tradicional de textos (QI; DAVISON, 2009).

As informações existentes nas páginas *web* como o conteúdo estrutural e *hyperlinks* podem ser utilizadas pelo classificador, embora o seu uso não seja obrigatório em todas as situações. Essas informações em alguns casos podem contribuir para uma melhora na eficiência do algoritmo de classificação.

A sessão 2.3 apresenta quais são os atributos estruturais que podem ser utilizados para realizar a classificação de um documento *web*, enquanto a sessão 2.4 apresenta os atributos textuais que são utilizados para a classificação tradicional de textos e que também podem ser utilizados para a classificação de páginas da *web*.

2.3 ATRIBUTOS ESTRUTURAIS E METADADOS

Um documento *web* conforme mostrado na sessão 2.2 apresenta algumas características que o diferenciam de um documento de texto comum. Entre essas características estão os marcadores HTML (*HyperText Markup Language*). Um marcador HTML é um elemento textual não visual definido entre parênteses angulares (“<” e “>”). Esses marcadores realizam a função de formatação da linguagem.

A Tabela 1 apresenta alguns marcadores que são comuns em um documento *web*. Al-

guns marcadores possuem atributos que podem ser utilizados para representar informações adicionais sobre eles (W3C, 1999).

Tabela 1: Elementos HTML

Elemento	Atributo	Descrição do Elemento
<title>		Título da página
<meta>		Metadados da página
<meta>	description	Descrição do conteúdo da página
<meta>	keywords	Palavras chaves sobre o conteúdo
<h1>, <h2>,... <h6>		Título de sessão em vários tamanhos
<b>		Texto em negrito
<i>		Texto em itálico
<u>		Texto sublinhado
<s>		Texto tachado
		Texto realçado
		Texto enfatizado
<img>		Inclui uma imagem na página
<img>	alt, title	Descrição da imagem
<a>		Link para outro local
<a>	alt, title	Descrição do link

Com base nas informações do HTML e nos textos das páginas é possível extrair outras informações que também podem ser utilizadas pelo classificador, como os metadados e dados de formato de apresentação da página. Abaixo consta a relação de algumas das informações que podem ser obtidas:

- Tamanho do texto: Total de palavras encontradas na página.
- Palavras em destaques: Total de palavras que estão entre marcadores de título ou em negrito.
- Link para páginas: Total de *links* para outras páginas ou outros sites
- Link para imagens ou vídeos: Total de *links* para arquivos de imagens ou de vídeos
- Total de imagens: Total de imagens presentes na página

Outra forma de extração de características são as métricas obtidas a partir de listas de palavras relacionadas ao tópico desejado. Nesse modelo é criado uma lista de palavras usuais sobre a classe que será classificada, por exemplo, caso desejássemos classificar páginas como pertencendo a duas classes: “Esporte” e “Não Esporte”, seria possível criar uma lista de palavras comuns encontradas em sites de esportes, tais como: “futebol”, “automobilismo”, “tênis”, “basquete”, “campeonato”. A partir da lista de palavras é possível obter os seguintes atributos:

- *Description*: Total ou percentual de palavras da lista que estão presentes no atributo *description* do HTML
- *Keywords*: Total ou percentual de palavras da lista que estão presentes no atributo *Keywords* do HTML
- *Title*: Total ou percentual de palavras da lista que estão presentes no elemento *title* do HTML
- *Links*: Total ou percentual de palavras da lista que estão presentes em textos dos *links*
- Imagens: Total ou percentual de palavras da lista que estão presentes em textos descritivos das imagens
- Total palavras: Total ou percentual de palavras da lista que estão presentes nos textos da página, excluindo os textos dos elementos HTML.

No caso excepcional da página não possuir determinado marcador HTML, por exemplo *Keywords*, os atributos númericos relacionados a esse marcador recebem o valor zero.

Além das informações estruturais e textuais que podem ser obtidas da página *web* existem as informações que podem ser obtidas do protocolo HTTP (*Hypertext Transfer Protocol*). O protocolo HTTP é o protocolo utilizado na comunicação entre os navegadores e os servidores *web*. Entre outras definições do protocolo está a utilização da URL (*Uniform Resource Locator*), onde podem constar informações úteis ao classificador (FIELDING et al., 1999).

Uma URL é formada pela concatenação de várias informações, algumas das informações são obrigatórias e outras são opcionais. Uma URL deve possuir o seguinte formato: URL = “protocolo:” “//” *host* [“:” porta] [caminho [“?” *query*]] , no qual:

- protocolo : Define o protocolo a ser utilizado na comunicação, por exemplo HTTP.

- *host*: Domínio ou IP (*Internet Protocol*) do site
- porta: Porta TCP (*Transmission Control Protocol*) utilizada para comunicação, quando não informado é utilizada a porta padrão, porta 80.
- caminho: Localização no *host* do recurso solicitado.
- *query*: Dados passados para o recurso solicitado, geralmente no formato chave/valor.

Os métodos de classificação que utilizam exclusivamente informações do protocolo HTTP possuem algumas vantagens em relação aos demais. Como a quantidade de caracteres em uma URL não é grande, o tempo de extração das informações é pequeno, o que faz com que esses métodos sejam rápidos (KAN; THI, 2005). Além disso, como não utilizam informações do protocolo HTML não é necessário que se tenha acesso ao código fonte da página para classificá-la: é possível categorizá-la mesmo sem ter acesso ao seu conteúdo (KAN, 2004).

2.4 PRÉ-PROCESSAMENTO DE TEXTO E SELEÇÃO DE ATRIBUTOS TEXTUAIS

Em tarefas de processamento de texto normalmente se emprega um modelo padrão para a representação dos documentos. Esse modelo de representação denominado de *Vector Space Model*, também chamado de modelo "*bag-of-words*", foi proposto por Salton et al. (1975). De acordo com esse modelo cada documento $d_i \in D$ é composto por um conjunto de termos indexados. Em toda a coleção de documento D, o conjunto de termos é representado como $T = T_1, ..., T_N$, no qual, N é o total de diferentes termos da coleção. Assim, cada documento d_i corresponde a um vetor N-dimensional, ou $d_i = [w_{i1}, w_{i2}, ..., w_{iN}]$, aonde w_{ij} é o peso do termo t_j no documento d_i.

Existem várias formas para definir o peso w_{ij}, ou seja, o peso do termo t_j no documento d_i. A forma mais simples é a atribuição de pesos booleanos, em que $w_{ij} = 1$ se o termo t_j aparece no documento d_i, caso contrário recebe valor zero (BAEZA-YATES; RIBEIRO-NETO, 1999).

Antes de realizar o treinamento do classificador propriamente dito, deve-se executar alguns processos que visam transformar o documento de forma que o mesmo possa ser tratado pelo algoritmo de classificação. Esses processos são o pré-processamento do texto e a seleção de atributos.

O pré-processamento de texto pode ser definido como uma sequência de passos que objetivam deixar o conteúdo textual na forma de uma sequência de termos. Esses passos são (SOARES et al., 2009):

- *Tokenization*: neste passo é realizada a quebra do fluxo de caracteres em palavras, também chamadas de *tokens*. Nesse momento é realizado a remoção de alguns caracteres, tais como caracteres especiais, números, sinais de pontuação e separação silábica. Também é realizado a conversão de todos os caracteres para minúsculo (ou maiúsculo), assim é possível agrupar palavras idênticas. No caso de páginas *web* nessa fase também ocorre a remoção dos marcadores HTML (BAEZA-YATES; RIBEIRO-NETO, 1999).

- Remoção de *Stopwords*: palavras que são muito comuns em determinado idioma são chamadas de *stopwords*. Elas não ajudam na classificação de textos devido a estarem presentes em vários documentos e carregarem pouco conteúdo semântico, e por isso devem ser removidas. Essas palavras incluem pronomes, artigos, advérbios, preposições, conjunções e qualquer palavra que possua até dois caracteres (BAEZA-YATES; RIBEIRO-NETO, 1999).

- *Stemming*: é o processo que realiza a transformação de cada termo para o radical correspondente, por meio da remoção de prefixos e sufixos dos termos. Algoritmos de *stemming* têm desempenhos diferentes de acordo com o idioma do documento; assim algoritmos escritos para o idioma Inglês dificilmente terão bons resultados para o Português. Esse passo é opcional podendo não ser aplicado em determinadas situações (BAEZA-YATES; RIBEIRO-NETO, 1999).

- *N-Grams*: nesta etapa todo o texto é analisado e agrupado em sequências contínuas de N caracteres. Um *N-gram* com N igual a 1 é chamado de unigrama; com N igual a 2 é denominado bigrama; com N igual a 3 é chamado de trigrama; com N maior que 3 é chamado por 4-*gram*, 5-*gram* e assim por diante (CAVNAR; TRENKLE, 1994). A conversão em *N-grams* também é uma etapa opcional, e seu uso depende da aplicação.

Ao final do pré-processamento os documentos passam a serem representados por uma sequência de termos ou palavras, chamado de vetor de termos. O passo seguinte é a construção de um dicionário, no qual cada termo é representado por um valor numérico. O tamanho do dicionário é igual à quantidade de termos únicos encontrados nos documentos. O próximo passo é a obtenção de um documento codificado onde cada termo é substituído pelo seu correspondente valor no dicionário. A Figura 1 mostra um exemplo de pré-processamento de texto.

Após a conversão dos termos por valores numéricos, os documentos estão prontos para que seja realizada a seleção dos atributos que serão utilizados pelo algoritmo de classificação. Inicialmente todos os termos poderiam ser considerados na representação de um documento, cada um deles correspondendo a um atributo a ser empregado pelo classificador. A seleção

consiste em escolher determinados termos para que eles passem a representar todo o conjunto de documentos. Este é um processo de redução de dimensionalidade, visando diminuir o número de atributos utilizados pelo classificador (PORTER, 1980), (SOARES et al., 2009), (FORMAN, 2003).

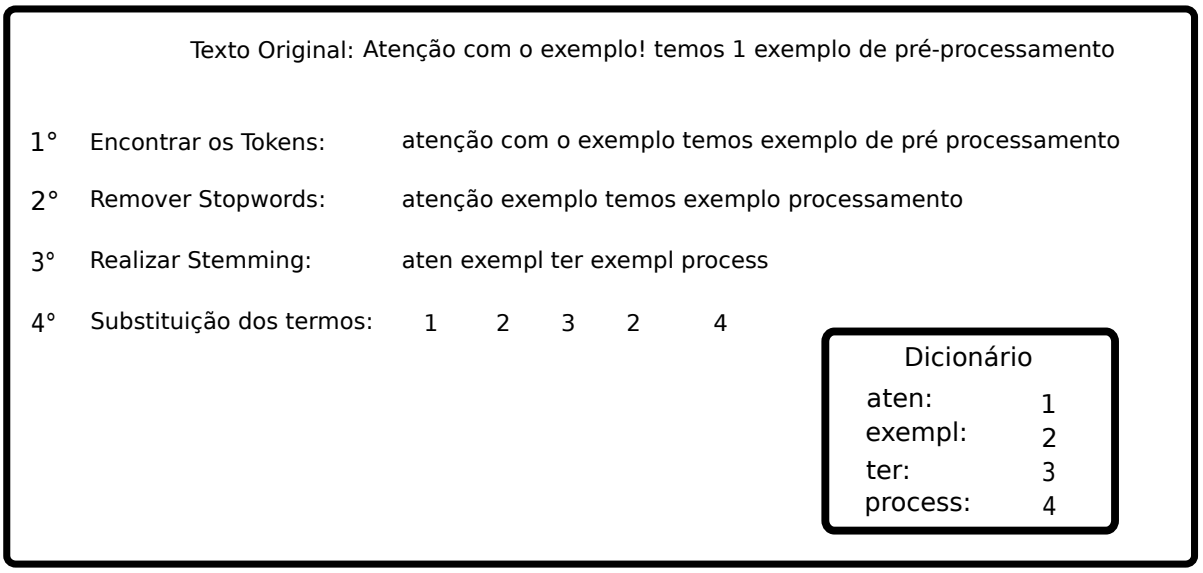

Figura 1: Exemplo de Pré-Processamento de Texto

A seleção de atributos possui vários benefícios potenciais, entre os quais: facilitar a visualização e entendimento dos dados, reduzir os custos com armazenamento, diminuir o tempo de treinamento e aumentar a taxa de acerto do classificador (SEBASTIANI, 2002), (GUYON; ELISSEEFF, 2003).

A forma mais simples de seleção de atributos consiste na abordagem filtro, onde se utiliza uma métrica para estimar a qualidade do atributo. Uma métrica de avaliação é aplicada individualmente a cada atributo, gerando um índice de qualidade para cada um dos atributos em seguida é criado uma lista L de atributos ordenados pelo índice de qualidade. São selecionados os N primeiros atributos da lista L formando o conjunto de atributos que será empregado pelo classificador (FORMAN, 2003). As métricas comumente utilizadas para estimativa de qualidade de atributos são: TF (*Term Frequency*), TFIDF (*Term Frequency Inverse Document Frequency*) e Ganho de Informação.

O TF mede a quantidade de vezes que um termo ocorre em determinado documento ou $tf(t,d) = freq(t,d)$ para um termo t e um documento d. Assim nesta métrica a qualidade do atributo é proporcional a sua frequência nos documentos, quanto maior a frequência maior a qualidade. Para calcular a qualidade do atributo utiliza-se a fórmula 1, na qual: Q_i representa a qualidade do atributo t_i e N é o total de documentos da coleção (SALTON; BUCKLEY, 1988),

(YANG; PEDERSEN, 1997).

$$Q_i = \sum_{j=0}^{N} tf(t_i, d_j) \tag{1}$$

A métrica TFIDF também mede a frequência de termos nos documentos, porém nesta medida os termos que aparecem na maioria dos documentos possuem um peso menor. Nesta métrica a qualidade do atributo é maior quando ele possuir alta frequência em um documento porém não for comum no conjunto de documentos. As equações 3 e 2 demonstram como é realizado o cálculo, conforme pode-se ver o IDF (*Inverse Document Frequency*) é inversamente proporcional ao logaritmo do total de documentos em que o termo aparece, sendo N o total de documentos da coleção e $d(t)$ o total de documentos em que o termo t está presente (SALTON; BUCKLEY, 1988), (YANG; PEDERSEN, 1997).

$$idf(t) = \log(\frac{N}{d(t)}) \tag{2}$$

$$Q_i = \sum_{j=0}^{N} tf(t_i, d_j) \times idf(t_i, N) \tag{3}$$

O Ganho de informação, mede a quantidade de partições de informação obtidas para a predição da categoria através da presença ou ausência de um termo no documento. A qualidade atribuída por esta métrica será maior quanto maior for a capacidade do atributo conseguir separar um conjunto de exemplos em categorias. Esta métrica é utilizada no campo de aprendizagem de máquina, como por exemplo na construção de árvores e regras de decisão (QUINLAN, 1986). As equações 4, 5 e 6 apresentam como é realizado o cálculo.

O ganho mede o quanto um atributo é capaz de separar um conjunto de exemplos em categorias.

$$GanhoInfo(F) = A + B \tag{4}$$

$$A = P(W) \times P(W) \sum_i P(C_i|W) \times \log \frac{P(C_i|W)}{P(C_i)} \tag{5}$$

$$B = P(\bar{W}) \times P(W) \sum_i P(C_i|\bar{W}) \times \log \frac{P(C_i|\bar{W})}{P(C_i)} \tag{6}$$

Nestas fórmulas: $GanhoInfo(F)$ é o ganho de informação da característica F; F é a característica que representa a palavra W; $P(W)$ é a probabilidade de ocorrer a palavra W; $P(\bar{W})$ é a probabilidade de não ocorrer a palavra W; $P(C_i)$ é a probabilidade de ocorrência da i-ésima classe; $P(C_i|W)$ é a probabilidade condicional de ocorrer a i-ésima classe dado a palavra W; $P(C_i|\bar{W})$ é a probabilidade condicional de ocorrer a i-ésima classe dado a não ocorrência da palavra W (YANG; PEDERSEN, 1997). Nas aplicações as probabilidades acima indicadas são estimadas de acordo com as frequências correspondentes.

Por fim, após selecionados os atributos e definido os pesos, a coleção de documentos passa a ser representada por uma matriz de dimensão $M \times N$, na qual M é o total de documentos e N o total de atributos (termos). A matriz serve como informação de entrada dos algoritmos de aprendizado de máquina.

2.5 EXEMPLO DE CLASSIFICAÇÃO DE UMA PÁGINA WEB

A sessão 2.3 mostrou os atributos estruturais e metadados que podem ser utilizados na classificação de sites, enquanto a sessão 2.4 abordou o pré-processamento de texto e a seleção de atributos textuais. Nesta sessão será apresentado um exemplo de obtenção destes atributos de uma página *web*.

Em um exemplo hipotético considere um classificador que tenha como objetivo encontrar sites na internet que falem sobre assuntos relacionados a Copa do Mundo de 2014. Para o classificador em questão foi definido a utilização de lista de palavras chaves, após uma pesquisa em sites relacionados ao tema. Foram escolhidas as seguintes palavras para compor a lista: “copa”, “mundo”, “2014” ,“brasil”,“futebol”.

A página hipotética: www.portaldacopadomundo.gov.br possui o código HTML que consta na Figura 2. Após renderizada pelo navegador esta página produz uma imagem similar a da Figura 3.

Com base no código HTML da página (Figura 2) podem-se obter diversos atributos estruturais e metadados que poderão ser utilizados pelo classificador. A Tabela 2 apresenta atributos que foram extraídos da página de exemplo.

```
<html>

<head>
 <meta name="description" content="Portal da Copa do mundo de 2014. Site com
  informações
  sobre estádios e hotéis das cidades-sede">
 <meta name="keywords" content="portal, copa, mundo, 2014, brasil, sedes, estádios,
  hotéis">
 <title>Portal da Copa do Mundo</title>
 <style type="text/css">
    p {color: green; }
    a {text-decoration: none; color: #01F; font-weight: bold; }
  </style>
<head>

<body>

<center>
 <img src="copa.jpg" title="Emblema da Copa do Mundo de 2014">
</center>

<p> A Copa do Mundo da <a href="http://pt.fifa.com/" title="Site da FIFA" > FIFA </a> é um
dos maiores eventos esportivos do planeta. Em 2014, o Brasil será novamente sede do
torneio. A vigésima Copa do Mundo da FIFA ocorrerá 64 anos depois da edição em que a <a
href="http://www.cbf.com.br/" title="Site da CBF"> seleção </a> nacional se sagrou vice-
campeã mundial em pleno Maracanã. </p>

<p> Neste site você encontra todas as informações sobre o maior evento esportivo do
futebol mundial. Conheça as cidades-sede, os estádios e acompanhe as últimas notícias
sobre a copa. </p>

<body>
<html>
```

Figura 2: Código HTML da Página de Exemplo

Tabela 2: Atributos Extraídos do Código HTML da Página de Exemplo

Atributo	**Valor**
Total de palavras do texto	76
Total de *links* para páginas	2
Total de *links* para imagens	0
Total de imagens na página	1
Total de termos da lista presentes no atributo *description*	3
Total de termos da lista presentes no atributo *keywords*	4
Total de termos da lista presentes no atributo *title*	2
Total de termos da lista presentes em textos dos *links*	0
Total de termos da lista presentes em textos descritivos da imagens	3
Total de termos da lista presentes no texto da página	8
Total de termos da lista presentes na URL	2

Além dos atributos extraídos do código HTML é possível extrair atributos textuais que também podem ser utilizados pelo classificador. A Figura 4 apresenta o processo de extração

Figura 3: Renderização pelo Navegador da Página de Exemplo

dos atributos textuais da página de exemplo (Figura 3). Nesse exemplo foi aplicado o algoritmo de *stemming* proposto por Soares et al. (2009). No dicionário ao lado do termo é apresentando o valor do TF (*Term Frequency*) que representa a frequência com que ocorre o termo no texto.

2.6 ALGORITMOS DE CLASSIFICAÇÃO

Nesta sessão será apresentada uma breve descrição dos algoritmos de classificação utilizados neste trabalho. Conforme explicado na sessão 2.1 um classificador é uma função matemática que mapeia dados de entrada em um conjunto finito de categorias. Primeiramente será abordado o classificador probabilístico Naïve Bayes. As redes neurais artificias do tipo MLP (*multilayer perceptron*) são abordadas na sequência. Em seguida, explica-se o funcionamento dos algoritmos de árvore de decisão ID3 e C4.5, e em seguida do classificador KNN (K vizinhos mais próximos) que é um classificador fundamentado no método de aprendizado baseado

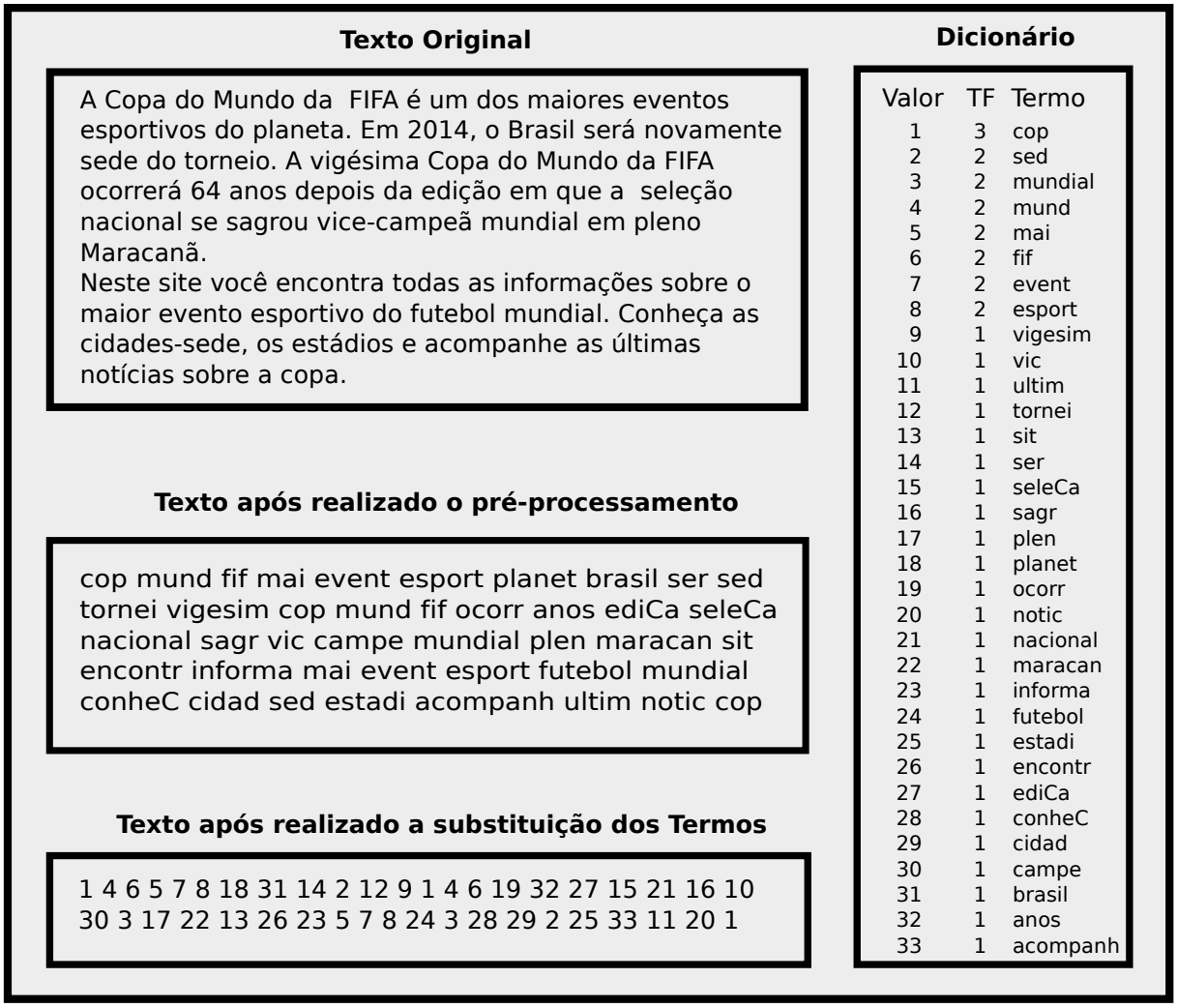

Figura 4: Representação do Processo de Extração de Atributos Textuais da Página de Exemplo

em instâncias. Por fim, aborda-se o SVM (Máquinas de Vetores de Suporte), constituído por um classificador obtido a partir de um processo de otimização que visa maximizar as margens existentes entre os componentes das classes mais próximos da superfície de separação.

2.6.1 NAÏVE BAYES

O classificador Naïve Bayes é um classificador probabilístico baseado na aplicação do teorema de Bayes que relaciona a probabilidade de uma hipótese dada a observação de uma evidência e a probabilidade da evidência dada pela hipótese. Uma característica importante desse classificador é que ele parte do princípio que todos os atributos são independentes. Em outras palavras, o classificador Naïve Bayes acredita que à presença ou a ausência de uma determinada característica da classe não está relacionada a presença ou ausência de nenhuma outra característica desta classe; daí vem o nome Naïve (ingênuo) do algoritmo (MITCHELL, 1997).

O cálculo de probabilidade utilizada por esse classificador é fundamentado pelo teorema de Bayes, representado pela equação 7 (DUDA; HART, 1973).

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{7}$$

Na qual:

- $P(h)$ A probabilidade a priori da hipótese h antes da observação do exemplo de treinamento D que reflete o conhecimento de fundo sobre h. Caso não se disponha deste conhecimento deve-se considerar que todas as hipóteses são equiprováveis.
- $P(D)$ A probabilidade inicial ou probabilidade a priori do exemplo de treinamento D, ou seja, a probabilidade do exemplo D ser observado sem que o conhecimento sobre qual hipótese h é verdadeira para D.
- $P(D|h)$ A probabilidade de D condicional a h, ou seja, a probabilidade de D ocorrer sabendo-se que h é verdadeira.
- $P(h|D)$ A probabilidade a posteriori de h dado D, ou seja, a probabilidade da hipótese h ser verdadeira face a ocorrência do exemplo de treinamento D.

A abordagem utilizada pelo classificador Bayesiano para categorizar uma nova instância consiste em classificá-la com o valor mais provável, Y_s, dado os valores de atributos de entrada $< x_1, x_2, ..., x_n >$ que o descrevem e um conjunto de prováveis valores Y. A fórmula para o cálculo de Y_s é mostrada na Equação 8 (DUDA; HART, 1973).

$$Y_s = \underset{y_j \in Y}{\operatorname{argmax}} P(y_j) \prod_{i=1}^{n} P(x_i|y_j) \tag{8}$$

O Naïve Bayes conhecido por sua simplicidade e eficiência, possui uma estrutura fixa e parâmetros ajustáveis. Embora o princípio da independência de variáveis seja uma simplificação, em muitos casos os resultados obtidos são satisfatórios (MCCALLUM; NIGAM, 1998).

2.6.2 PERCEPTRONS MULTICAMADAS

A *Multilayer Perceptron* (MLP), ou perceptron multicamadas, é uma classe de redes neurais artificiais (RNA) cuja propagação dos cálculos ocorre num único sentido, ou “para a

frente" (*feedforward*) sem a ocorrência de laços. Esse modelo de rede é composto: por uma camada de entrada; por uma ou mais camadas intermediárias, também chamadas de camadas ocultas e por uma camada de saída, responsável diretamente pelos valores de saída da MLP. Os nós dessa rede são constituídos por neurônios simples ou perceptrons, que são conectados da seguinte forma: as saídas dos neurônios de uma camada formam a entrada dos neurônios da camada seguinte (MITCHELL, 1997).

A MLP pode ser vista como uma modificação no algoritmo *perceptron* possibilitando que seja realizado a classificação de dados que não são linearmente separáveis, os quais não poderiam ser corretamente categorizados pelo algoritmo original (MINSKY; PAPERT, 1969).

Neste tipo de rede o sinal de entrada propaga-se para a frente, camada por camada, desde a camada de entrada até a camada de saída. As MLPs têm sido aplicadas com sucesso em diversos problemas de reconhecimento de padrões. Um dos motivos do sucesso da MLP é o seu algoritmo de aprendizado, fundamentado na retropropagação do erro durante a fase de treinamento da rede (*backpropagation*) (HECHT-NIELSEN, 1989).

A aprendizagem por retro propagação de erro é realizada em duas etapas. Na primeira etapa o vetor de entrada é aplicado aos neurônios da primeira camada, propagando o seu efeito para a frente, através da rede camada a camada, usando pesos sinápticos fixos inicialmente, um elemento de compensação ou *bias* e uma função de ativação para os neurônios. Após propagar-se pela camada de saída tem-se a resposta da rede ao vetor apresentado na camada de entrada. Na segunda etapa é aplicado uma regra para cálculo de erro. Levando-se em consideração a diferença entre a resposta produzida pela rede e a desejada, reajustam-se os pesos sinápticos e repete-se o processo de propagação, porém, dessa vez para trás. Este processo é executado iterativamente até que a resposta produzida pela rede apresente erro próximo a limiares aceitáveis para a tarefa (HORNIK et al., 1989).

A Figura 5 apresenta um grafo estrutural de um perceptron de múltiplas camadas com duas camadas ocultas (PAL; MITRA, 1992). Nessa Figura, temos: o vetor de entrada que possui tamanho m, sendo representado pelas variáveis x_i; a indicação das camadas ocultas representadas pela letra k; a camada de saída representada por y. Nota-se que a quantidade de neurônios na camada de saída não precisa necessariamente coincidir com a quantidade de neurônios nas camadas ocultas.

Considerando um problema de classificação com apenas duas classes, podemos tratá-lo utilizando uma rede MLP com uma camada oculta, um dado vetor de entrada $x = (x_1, x_2, ..., x_d)^T$ e um neurônio na camada de saída. Esse problema hipotético pode ser tratado por uma rede MLP que utiliza as equações 9 e 10 (MITCHELL, 1997).

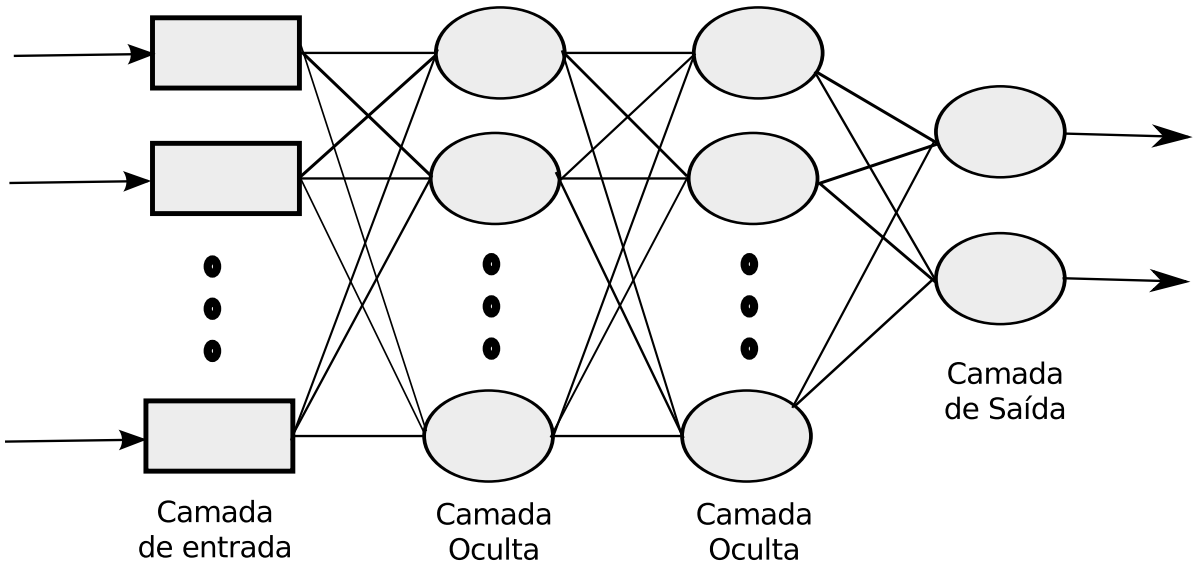

Figura 5: Grafo Arquitetural de um Perceptron Multicamadas com Duas Camadas Ocultas

$$z_j = g_t(\sum_{d=1}^{D} w_{jd}x_d + w_{j0}) \tag{9}$$

$$f = h_t(\sum_{j=1}^{J} v_j z_j + v_0) \tag{10}$$

Nas quais:

- z_j, j=1, 2, ..., J representam as ativações de nós computacionais na camada oculta,
- w_{jd} são os pesos entre a camada de entrada e a camada oculta,
- v_j são os pesos entre a camada oculta e a camada de saída f,
- w_{j0}, v_0 são os *biases* para as camadas oculta e de saída, respectivamente.
- h_t e g_t são funções de ativação de cada neurônio da rede. Em geral são utilizadas funções na forma $tanh(t)$ ou $1/(1+e^{-t})$.

A MLP é uma rede neural que utiliza aprendizado supervisionado. Existem outras redes neurais que utilizam técnicas de aprendizado não supervisionado, por exemplo: (a) *Self-Organizing Map* (SOM) proposta por Kohonen (1982) que após treinada produz uma representação dos dados em uma dimensão menor, sendo úteis para visualização de dados que estão em alta dimensão (FLEXER, 2001); (b) *Fuzzy* ART que implementa lógica *fuzzy* para reconhecimento

de padrões usando a teoria ART (*Adaptive Resonance Theory*. A teoria ART foi introduzida por Grossberg (1976) para explicar como ocorre o processamento de informações pelo cérebro humano. O classificador *Fuzzy* usando esse conceito foi proposto por Carpenter et al. (1991), possuindo como uma de suas principais característica a velocidade de treinamento.

2.6.3 ÁRVORES DE DECISÃO

Árvores de decisão (AD) são métodos para aproximação de funções-alvo baseadas em valores. A representação desta função ocorre através de uma árvore de regras se-então, na qual cada nó representa um teste de atributo e as folhas representam o conceito-alvo (MITCHELL, 1997).

A Figura 6 apresenta um exemplo de uma AD que utiliza informações climáticas para a inferência se será possível jogar tênis durante um determinado dia (MITCHELL, 1997). Na figura cada nó - representado por um retângulo - é um teste sobre um atributo e os ramos que o seguem imediatamente são os possíveis valores que este atributo pode assumir. Os atributos que constam na árvore são: Previsão, Umidade e Vento. O processo de classificação tem início com a apresentação de uma instância de valores de atributos para a árvore, o primeiro teste do nó raiz é então efetuado (Previsão). Após isso, o ramo correspondente ao valor do atributo na instância conduzirá o fluxo de avaliação para o nó seguinte, e assim sucessivamente até que um valor (sim ou não) seja obtido.

Retornando ao exemplo da Figura 6, uma instância com os seguintes valores de atributos: Previsão=Ensolarado, Umidade=Alta e Vento=Forte produzirá como classificação JogarTênis=Não.

Árvores de decisão representam conceitos na forma de disjunções (cláusulas OU $\cup$) e conjunções (cláusulas E $\cap$) de restrições sobre os valores dos atributos. Por exemplo, se o tempo estiver ensolarado e o índice de umidade estiver normal, as condições estarão propícias para se jogar tênis. Usando conjunções e disjunções pode-se representar a regra pela equação 11.

$$Tempo = Ensolarado \cap Umidade = normal \Rightarrow JogarTenis = Sim \tag{11}$$

Os algoritmos que geram árvores de decisão, como o ID3 desenvolvido por Quinlan (1986) e o C4.5 (QUINLAN, 1993), promovem uma busca pelo espaço de hipóteses pela melhor hipótese (árvore) para a representação do conceito-alvo. No C4.5, por exemplo, a busca pela hipótese inicia-se pela escolha do primeiro atributo a ser testado, ou seja, o atributo raiz da

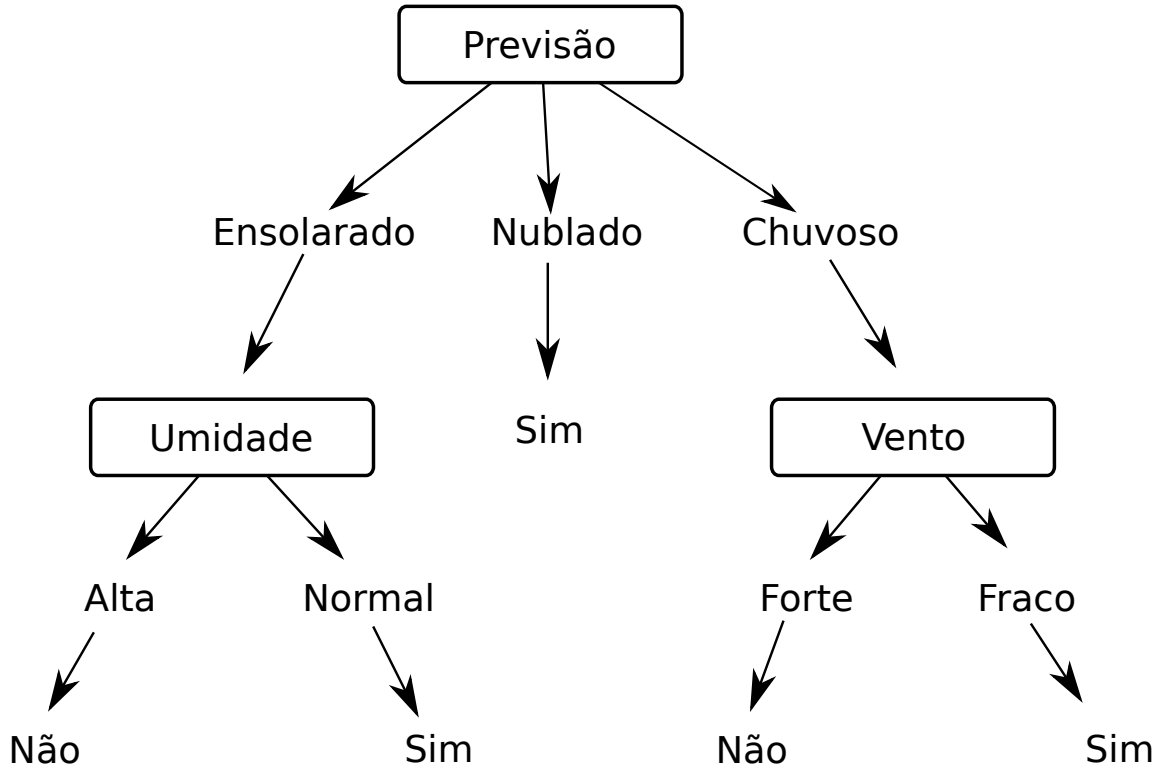

Figura 6: Exemplo de Árvore de Decisão

árvore. Uma avaliação sobre o conjunto de treinamento busca identificar o atributo que isoladamente possui o maior poder de segmentação entre as classes. A avaliação é conduzida sobre uma propriedade estatística denominada ganho de informação (*information gain*).

Para definir o ganho de informação primeiramente é necessário entender o conceito de Entropia. Entropia caracteriza a impureza dos dados de uma coleção e pode ser calculada usando a Equação 12.

$$Entropia(S) = -P\oplus \log P\oplus - P\ominus \log P\ominus \tag{12}$$

No qual, $P\oplus$ é a proporção de exemplos positivos em S e $P\ominus$ a de exemplos negativos. A Entropia será mínima, igual a 0, quando todos os exemplos do conjunto S pertencem à mesma classe. Por outro lado, a Entropia será máxima, igual a 1, quando o conjunto de dados for heterogêneo, ou seja, quando a quantidade de exemplos positivos em S for igual a de negativos. No caso do número de exemplo positivos e negativos serem desiguais a entropia será um valor entre 0 e 1.

Quando o conjunto de dados S possui c classes distintas a Entropia dele pode ser calculada pela Equação 13, na qual: p_i é a proporção de dados em S que pertencem à classe i;

c é o total de classes.

$$Entropia(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{13}$$

Uma vez entendido o conceito de Entropia pode-se explicar a sua relação com o ganho de informação. O atributo de maior ganho é definido a partir de entropia como simplesmente a redução esperada na entropia causada pelo particionamento dos exemplos por este atributo. A Equação 14 explica como realizar o cálculo do ganho de informação.

$$Ganho(S,A) = Entropia(S) - \sum_{x \in P(A)}^{n} \frac{S_x}{S} Entropia(S_x) \tag{14}$$

Na qual:

- $P(A)$: um conjunto dos valores que A pode assumir.
- n: o total de elementos do conjunto $P(A)$.
- x: um elemento do conjunto $P(A)$.
- S_x: o subconjunto de S formado pelos dados em que $A = x$.

Um problema enfrentado pelas técnicas de AD é o superajustamento da hipótese aos dados, o que faz com que o algoritmo tenha uma queda significativa no seu desempenho preditivo. Para evitar que isso ocorra, alguns mecanismos podem ser utilizados como a pré-poda e a pós-poda. A pré-poda tem por objetivo controlar o superajustamento durante a fase de treinamento, o que pode ser feito, por exemplo, por meio do descarte do alguns exemplos. A pós-poda consiste em tratar do superajustamento após a indução do modelo de classificação, isso pode ser feito por meio de corte de alguns dos ramos da AD (MITCHELL, 1997).

2.6.4 K VIZINHOS MAIS PRÓXIMOS

O KNN (do inglês *K-Nearest Neighbor*), ou K vizinhos mais próximos, é um método de classificação que utiliza o paradigma do aprendizado baseado em instâncias. O paradigma tem como pressuposto que se duas instâncias são similares, então elas pertencem à mesma classe. Deste modo, quando uma nova instância é similar a uma instância conhecida, a classe desta é atribuída a nova instância (MITCHELL, 1997).

O KNN não gera um modelo explícito a partir do conjunto de dados de treinamento, em contraposição com métodos de classificação de outros paradigmas, que durante a fase de treinamento geram um modelo explícito de classificação e depois disso podem descartar os dados de treinamento. Devido a isso é denominado de *lazy* (preguiçoso).

O algoritmo KNN relaciona cada uma das instâncias a um ponto em uma espaço *m*-dimensional, sendo que *m* é o número de atributos de entrada que descrevem o conjunto de dados. Assim, quando um novo exemplo necessita ser classificado, a similaridade com os exemplos já conhecidos é calculada por meio da distância de tais instâncias em relação a nova instância.

Diversas medidas de distâncias podem ser utilizadas. Para que a medida de distância seja considerada uma métrica válida ela deve satisfazer as seguintes condições:

- $d(x,y) \geq 0$
- $d(x,y) = 0$, se e somente se $x = y$
- $d(x,y) = d(y,x)$
- $d(x,z) \leq d(x,y) + d(y,z)$

Onde x, y, z são instâncias e $d(x,y)$ a distância entre as instâncias x e y.

Uma das distância que pode ser utilizada no cálculo de similaridade é a distância euclidiana. Para o cálculo desta distância é considerado os valores dos atributos de cada instância. Considerando $x =< x_1, x_2, x_3, ..., x_m >$ e $y =< y_1, y_2, y_3, ..., y_m >$ sendo as instância com seus respectivos vetores de atributos, a distância entre os dois exemplos x e y pode ser cálcula pela Equação 15.

$$d(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{15}$$

Além da distância euclidiana outras métricas podem ser utilizadas, por exemplo: a distância de *Manhattan* representada pela Equação 16; a distância máxima expressada pela Equação 17; a distância de *Minkowski* calculada pela fórmula 18. Usando a distância de *Minkowski* é possível obter a distância euclidiana, para isso basta definir $q = 2$ e $w_i = 1$. A distância de *Manhattan* também pode ser obtida definido o $q = 1$ e $w_i = 1$ na fórmula de *Minkowski* (JAIN; DUBES, 1988).

$$d(x,y) = \sum_{i=1}^{m} |x_i - y_i| \tag{16}$$

$$d(x,y) = \max_{i=1}^{m} |x_i - y_i| \tag{17}$$

$$d(x,y) = \sqrt[q]{\sum_{i=1}^{m} w_i |x_i - y_i|^q} \tag{18}$$

Após realizado o cálculo da distância deve-se definir o valor de k. Quando k é igual a 1, a classe do novo exemplo será igual à classe do exemplo com a menor distância em relação a ele. No caso de k ser maior que 1, é formada uma lista contendo os k vizinhos mais próximos ao novo exemplo. Após formada a lista é utilizado a Equação 19 para determinar a classe a qual o novo exemplo pertence.

$$Y = \operatorname*{argmax}_{v} \sum_{(T_i,Y_i) \in D_z} f(v = Y_i) \tag{19}$$

Na Equação 19, v representa um rótulo, (T_i, Y_i) são exemplos pertencentes à lista dos k vizinhos mais próximos (D_z); $f()$ é uma função que retorna o valor 1 se $v = Y_i$ e 0 caso contrário. De modo geral, a Equação 19 encontra o rótulo que estiver mais presente na lista dos vizinhos mais próximos (COVER; HART, 1967).

A Figura 7 apresenta duas classes dispostas em um espaço bidimensional, a classe das estrelas e a classe dos retângulos, um novo elemento (o circulo) deve ser classificado como pertencente a uma das duas classes. Usando o classificador KNN, com a distância euclidiana como métrica de similaridade, podemos classificar o novo elemento de acordo com os elementos mais próximos. Quando o k é igual a 1, o elemento mais próximo pertence a classe dos retângulos e esta classe será atribuída ao circulo. Usando um k igual a 3 serão utilizados os três elementos mais próximos, todos estão dentro do circulo tracejado, nesta situação um elemento pertecence a classe dos triângulos e dois elementos pertencem a classe dos retângulos sendo então esta classe designada ao novo elemento (COVER; HART, 1967).

Ainda sobre o classificador KNN, um problema que afeta o seu funcionando, não apenas o seu funcionamento mas também os dos demais classificadores, é o aumento da dimensionalidade. A dimensionalidade aumenta na mesma proporção do aumento do total de atributos utilizados. Em espaços com muitas dimensões as amostram se tornam esparsas e poucos similares. Assim tem-se um número maior de objetos distantes um dos outros e também a

ocorrência de objetos equidistantes entre si. Nesse cenário a capacidade de predição do classificador é diminuída, tendo como consequência uma diminuição do índice de acerto do algoritmo de aprendizagem.

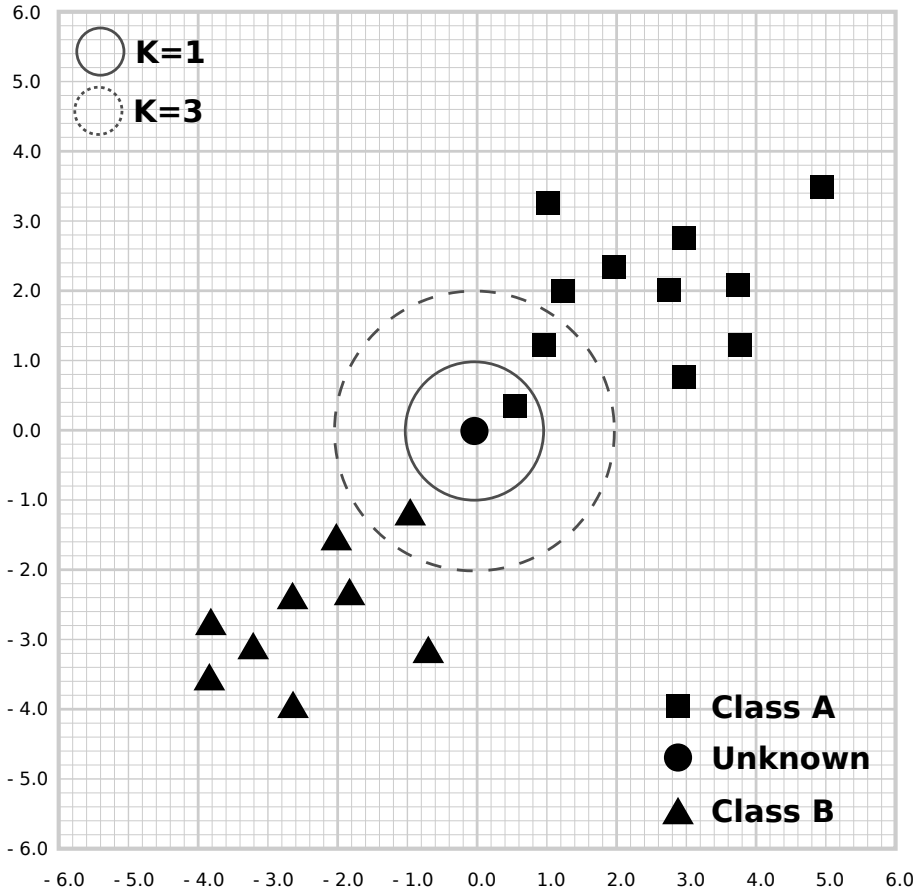

Figura 7: Exemplo de Classificação Utilizando o Método KNN

2.6.5 MÁQUINAS DE VETORES DE SUPORTE

As Máquinas de Vetores de Suporte, ou *Support Vector Machine* (SVM) são um conjunto de algoritmos supervisionados que normalmente são utilizadas para a obtenção de classificadores lineares e binários, ou seja, classificadores que separam exemplos em duas classes, geralmente denominadas de classe positiva e classe negativa (CORTES; VAPNIK, 1995). O uso de SVM para classificação de textos foi inicialmente proposto por Joachims (1998), que realizou um estudo empírico mostrando as vantagens do uso desse classificador em relação aos demais.

O algoritmo de aprendizado baseado em SVM tem por objetivo a separação ótima de classes. Para tanto utiliza-se um processo de otimização. O objetivo é maximizar a "margem", ou distância entre um hiperplano de separação e os pontos (vetores de suporte) mais próximos

à região de separação entre as classes (CRISTIANINI; SHAWE-TAYLOR, 2000).

A Figura 8 mostra dois possíveis separadores lineares de classes. Diante de diversas possibilidades para obtenção de uma função separadora de classes, deve-se avaliar qual é a função que faz a melhor distinção entre duas categorias diferentes de dados. A função de decisão que faz a melhor separação é aquela que apresenta a maior margem entre as duas classes analisadas. A margem pode ser definida como a soma das distâncias entre os pontos de ambas as classes que são mais próximos a função de separação. Para facilitar o entendimento do conceito de margem, introduz-se o conceito de *Support Vectors*, que dão nome ao algoritmo (CRISTIANINI; SHAWE-TAYLOR, 2000).

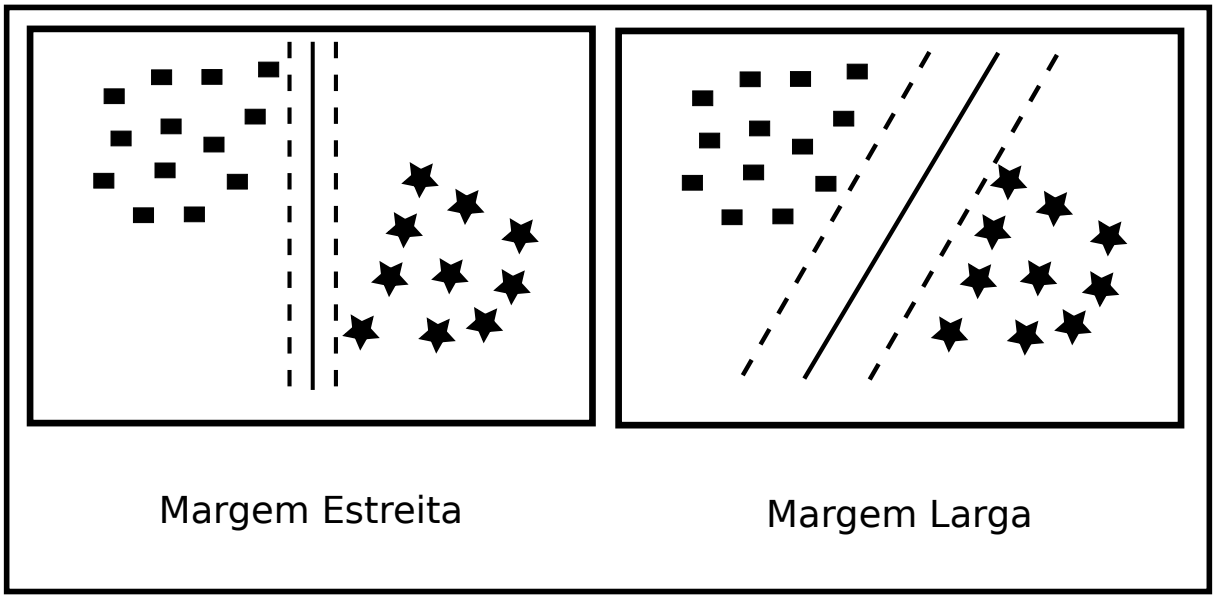

Figura 8: Exemplo de Dois Separadores de Classes e suas Margens

Vetores de Suporte (ou *Support Vectors*) são os pontos de ambas as classes que estão mais próximos do separador de classes. A determinação destes Vetores de Suporte é fundamental para o estabelecimento da função separadora das classes, pois o algoritmo faz uso destes dados para gerar a classificação. Na Figura 8 verifica-se, portanto, que o classificador da direita é o mais adequado, pois o separador ótimo de classes apresenta a maior margem entre a função de separação e os Vetores de Suporte.

A função de separação das classes também é chamada de hiperplano de separação. No caso dos atributos do espaço amostral estarem distribuídos em um espaço bidimensional o hiperplano de separação será uma reta, de dimensão única. O hiperplano de separação ótima é conhecido como *Optimal Separating Hyperplane* (OSH) (BURGES, 1998).

O treinamento de uma SVM pode ser visto como a resolução de um problema de otimização. Considere um conjunto de dados de treinamento linearmente separável com N

instâncias rotuladas $(x_1, y_1), ..., (x_n, y_n)$. Cada elemento do conjunto tem o rótulo da classe igual a $+1$ para classe positiva e a -1 para classe negativa. A função de classificação é um hiperplano, $f(x) = wx + b$, capaz de separar linearmente as classes, na qual os parâmetros w e b podem ser otimizados durante o treinamento SVM de modo a maximizar a separação entre as classes. Assim as equações 20 e 21 resolvem o problema de otimização:

$$Minimizar : ww \tag{20}$$

$$Sujeito\ a : y_i(wx_i + b) \geq 1, i = 1, ..., N \tag{21}$$

Esse problema de otimização é denominado forma *primal*. Geralmente a forma *primal* é transformada em um outro problema, chamado de *dual*, que é expressada pelas equações 22 e 23.

$$Maximizar\ Q(a) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y_i y_j (x_i \times x_j) \tag{22}$$

$$Sujeito\ a\ : \sum_{i=1}^{N} a_i y_i = 0; a_i \geq 0; i = 1, ..., N \tag{23}$$

Uma modificação chamada de *soft margin* SVM possibilita encontrar um hiperplano de separação para conjuntos de dados não linearmente separáveis por conta de algum erro de rotulação, de ruídos nos dados ou mesmo pela natureza intrínseca dos dados. O *trade-off* entre maximizar a margem e permitir erros no conjunto de treinamentos é expresso por uma constante $C > 0$. A função a ser otimizada permanece a mesma, porém é adicionada ao problema *dual* a restrição $a_i \leq C$ (BOSER et al., 1992).

Para resolver o problema da não linearidade do espaço amostral, foi proposta uma projeção dos dados amostrais em um espaço dimensional maior através de funções Kernel, sendo que a partir deste novo espaço amostral o algoritmo de classificação SVM é utilizado. Na Figura 9 é mostrado os dados em um espaço bidimensional sendo projetados em um espaço tridimensional através de uma função Kernel (CRISTIANINI; SHAWE-TAYLOR, 2000).

Um Kernel corresponde a um produto escalar que normalmente encontra-se em espaço dimensional superior a qual os atributos foram inferidos. Neste novo espaço dimensional espera-se que os pontos mapeados passem a ser linearmente separáveis. A aplicação de funções Kernel otimiza a atuação da função de separação de classes.

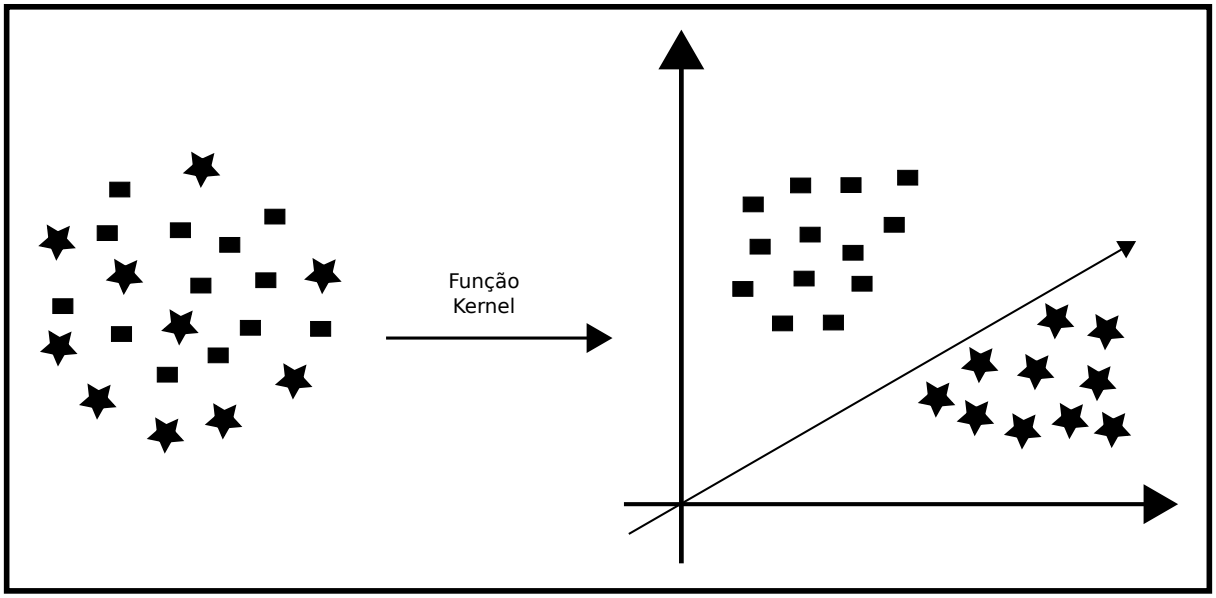

Figura 9: Exemplo de Dados de Entrada Coletados em um Espaço 2D sendo Redistribuídos em um Espaço 3D por Aplicação de uma Função Kernel

Existem vários algoritmos para resolver o problema de otimização quadrático que surge das SVMs, a maioria deles utilizando a heurística para quebrar o problema em pedaços menores e mais gerenciáveis. Um método comum para resolver este problema é o *Sequential Minimal Optimization* (SMO) (PLATT, 1999). O SMO é um algoritmo interativo para resolver o problema de otimização que surge durante o treinamento das SVMs. O SMO é utilizado por ferramentas populares de aprendizado de máquinas, como a LIBSVM (CHANG; LIN, 2011) e WEKA (HALL et al., 2009).

Na literatura, diversas aplicações foram feitas utilizando SVM. No reconhecimento de padrões, por exemplo, aplicações deste algoritmo no reconhecimento de letras escritas à mão, reconhecimento de objetos e categorização de documentos (JOACHIMS, 1998). Os resultados da aplicação desta técnica são comparáveis aos obtidos por outros algoritmos de aprendizado, como as RNA (Redes Neurais Artificiais) e, em inúmeras tarefas, têm se mostrado superiores.

2.7 MÉTRICAS PARA AVALIAÇÃO DE CLASSIFICADORES

Segundo Sebastiani (2002) um classificador pode ser avaliado por três métricas: (a) eficiência no treinamento, baseada no tempo levado para realizar o treinamento; (b) eficiência da classificação, trata-se do tempo requerido para classificar um elemento e (c) efetividade, métrica que mede a taxa de acerto do classificador. Para a tarefa de classificação de textos a efetividade costuma ser o critério mais utilizado pois é o critério utilizado para calcular o índice de instâncias corretamente classificadas. Nesta sessão se apresenta como ela pode ser calculada.

Considere o problema de classificação que envolva apenas duas classes, chamadas P e N, Positiva e Negativa respectivamente. Pode-se obter quatro situações possíveis para um classificador, são elas:

- (a) VP (Verdadeiro Positivo): Ocorre quando a instância é Positiva sendo classificada como tal.
- (b) FN (Falso Negativo): Ocorre quando a instância é Positiva porém é classificada como Negativa.
- (c) VN(Verdadeiro Negativo): Ocorre quando é Negativa sendo classificada como tal.
- (d) FP (Falso Positivo) : Ocorre quando a instância é Negativa porém é classificada como Positiva.

Desta forma, com um classificador binário e um conjunto de instâncias pode-se construir uma matriz de confusão de dimensão dois. A Figura 10 mostra uma matriz de confusão gerada por duas classes. Com base nessas informações são obtidas as métricas para avaliação de classificadores, descritas abaixo (FAWCETT, 2006), (BAEZA-YATES; RIBEIRO-NETO, 1999):

- *Cobertura* também conhecido como *Recall, TPrate*: é calculado pela equação 24, na qual VP é o total de Verdadeiro Positivo e FN é o total de Falso Negativo.

$$Cobertura = \frac{VP}{(VP+FN)} \tag{24}$$

- Taxa de falsos positivos ou *FPrate*: obtida pela equação 25, na qual FP é o total de Falso Positivo e VN o total de Verdadeiro Negativo.

$$FPrate = \frac{FP}{(FP+VN)} \tag{25}$$

- *Acurácia* expressada pela equação 26. Sendo: VP o total de Verdadeiro Positivo; VN o total de Verdadeiro Negativo; P o total de instâncias Positivas; N o total de instâncias Negativas.

$$Acuracia = \frac{(VP+VN)}{(P+N)} \tag{26}$$

- *Precisão* calculado pela equação 27. Sendo VP o total de Verdadeiro Positivo e FP o total de Falso Positivo.

$$Precisão = \frac{VP}{(VP+FP)} \tag{27}$$

- *Medida-F* conhecida também como *F-measure, F1-score* ou *F-score*: calculada através da média harmônica entre Precisão e Cobertura, conforme equação 28. Sendo a Precisão definida pela equação 27 e Cobertura definida pela equação 24.

$$Medida\text{-}F = 2\frac{(Precisão \times Cobertura)}{(Precisão + Cobertura)} \tag{28}$$

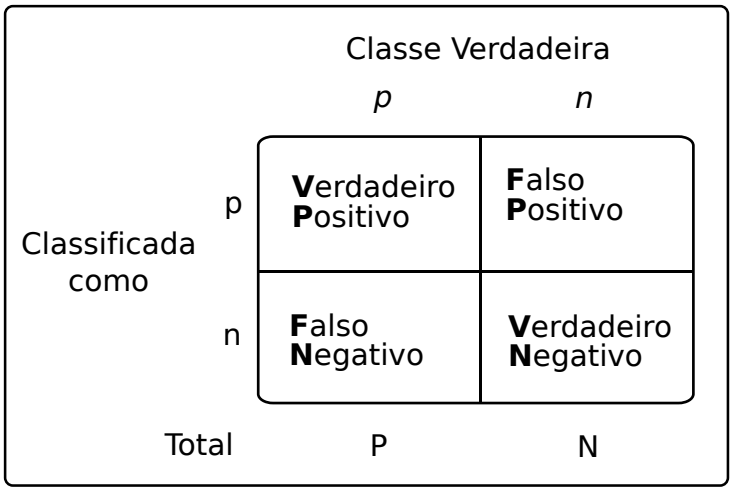

Figura 10: Matriz de Confusão

Embora alguns trabalhos utilizem a acurácia como principal métrica, ela não é a mais adequada. Caso a distribuição de classes esteja desbalanceada o resultado da acurácia é tendencioso. Cobertura e Precisão são boas métricas, porém os classificadores podem ser arbitrariamente ajustados para enfatizar o Cobertura ao custo de diminuir a Precisão, ou o contrário. Deste modo, apenas uma combinação entre Precisão e Cobertura pode apresentar resultados relevantes. Essa combinação pode ser obtida pelo uso da métrica Medida-F, que combina em um único valor Precisão e Cobertura (SEBASTIANI, 2002).

2.8 ABORDAGENS UTILIZADAS PARA A CLASSIFICAÇÃO DE SITES

Nesta sessão são apresentadas diferentes abordagens para resolução do problema de categorização automática de sites, e em particular aquelas que empregam aprendizagem de máquina.

Kim e Nam (2006) propuseram um sistema de classificação hierárquico. Ao todo existem quatro níveis, sendo o nível zero atribuído as páginas que não possuem conteúdo adulto, os níveis de 1 ao 3 são designados as páginas que apresentam material pornográfico. Quanto menor o nível menor é a quantidade de pornografia presente na página. O processo de categorização é realizado em dois passos. No primeiro passo são removidos todos os marcadores HTML permanecendo apenas o conteúdo textual. O texto presente na página é então comparado com um dicionário de palavras indesejadas, caso não seja encontrada nenhuma palavra do dicionário a página é classificada como nível zero, caso contrário é executado o segundo passo. No segundo passo é realizada a categorização como nível 1, 2 ou 3. Isto é realizado pelo algoritmo de aprendizado de máquina SVM (*support vector machine*) (CORTES; VAPNIK, 1995). O modelo é gerado utilizando-se de atributos textuais extraídos, dos documentos, pelo método TF/IDF (*term frequency inverse document frequency*). Os autores utilizaram em seu trabalho uma base contendo 20.000 páginas coletadas da internet, com conteúdos em Inglês e Coreano. Aplicando o modelo proposto, na base de teste, foi atingido a acurácia de 87,07%.

O trabalho de Gao et al. (2008) propõe a utilização de KNN (*k-nearest neighbor*) com SVM (CORTES; VAPNIK, 1995) para a categorização dos sites. No trabalho propõe-se a eliminação das etiquetas do HTML durante a fase de pré-processamento, o texto resultante é colocado em um vetor de palavras, sem a realização de *stemming*. Em seguida, utiliza-se o algoritmo KNN para classificação dos exemplos de treinamento. Durante a fase de treinamento os exemplos que são incorretamente classificados pelo KNN são eliminados, os exemplos remanescentes são utilizados para a construção do classificador SVM. Os experimentos demonstraram que a execução em conjunto dos dois classificadores resulta em uma taxa de acerto maior do que a execução isolada de um único classificador. Nos experimentos foi utilizada uma base de dados contendo 1400 páginas em Inglês e 1400 em Chinês. A acurácia, nos experimentos, foi de 96,5% para páginas em Inglês e de 91,0% para as páginas em Chinês.

O trabalho de Polpinij et al. (2006) apresenta um estudo comparativo dos algoritmos de aprendizado de máquina Naïve Bayes e SVM (CORTES; VAPNIK, 1995) para o problema de categorização de páginas *web*. O pré-processamento realizado nos textos segue o mesmo método utilizado por Gao et al. (2008), remove-se os marcadores do HTML e as palavras resultantes são colocadas em um vetor, técnica conhecida como *bag-of-words*. Com o vetor de palavras o autor realizou dois experimentos. No primeiro experimento foi utilizado mais de 6.000 elementos do vetor. No segundo experimento foi aplicado a técnica de *N-Gram* para formar um vetor de palavras com bigramas depois disso, foram selecionados pouco menos de 200 elementos do vetor. Nos dois experimentos a performance do SVM foi superior ao Naïve Bayes. O SVM apresentou a mesma performance, independente do tamanho de vetor utilizado,

o mesmo não aconteceu com Naïve Bayes que teve performance pior quando o número de elementos do vetor era menor. A base utilizada para os testes era constituída de 1400 páginas em Inglês e 1400 páginas em tailandês. A acurácia do SVM atingiu índices de 100,00% para conteúdo em Inglês e 97,5% para tailandês, enquanto o Naïve Bayes atingiu 100,00% e 89,67%, respectivamente para documentos em Inglês e Tailandês.

Lee e Luh (2008) desenvolveu um método de classificação de conteúdo *web*, usando o método estatístico do qui-quadrado invertido, para geração de lista de sites de conteúdo adulto. O trabalho também propôs um mecanismo de atualização automática da lista de sites. Para que ocorra a classificação é criada uma lista de palavras, na qual palavras que constem frequentemente em sites adultos recebem peso positivo, enquanto as demais palavras recebem pesos negativos. Todo o conteúdo HTML é removido, após isso, os termos dos textos são comparados com a lista. Depois disso, um método de cálculo recebe a informação dos termos dos textos que estão presentes na lista e de seus respectivos pesos. O resultado do cálculo classifica a página em três possíveis classes: "Pornográfica", "Não Pornográfica" e "Incerta". As páginas classificadas como "Pornográfica" são colocadas na lista de sites adultos, periodicamente os sites da lista são visitados procurando *links* para sites que ainda não foram classificados. O modelo proposto foi aplicado a um conjunto de 5.000 páginas, nos idiomas Inglês e Chinês. A acurácia da classificação foi de 96,88% para conteúdos em Chinês e de 96,58% para o Inglês.

A classificação baseada em atributos do HTML é proposta no trabalho de Lee et al. (2005). No método proposto primeiro é criado uma lista de palavras comuns, encontradas nas classes que deseja-se classificar. Após isso, são extraídas quatro informações das páginas: conteúdo textual, título da página, conteúdo descritivo das imagens e cabeçalho da página, composto pelas informações *description* e *keywords*. Para extrair essas informações são utilizados os marcadores HTML: *title*, *description*, *keywords* e *img*. Em seguida, são calculadas métricas baseadas na presença ou ausência dos termos da lista nas informações extraídas. Por fim, as métricas geradas são repassadas para duas redes neurais que vão gerar o modelo de classificação. Uma das redes utilizadas foi do tipo *Fuzzy* ART (*Adaptive Resonance Theory*) e a outra foi do tipo SOM (*Self-Organizing Map*). Para execução dos experimentos foi construída uma base de dados contendo 4.786 páginas em Inglês e 1.464 em Chinês. O melhor resultado obtido foi com a rede neural do tipo SOM que obteve uma acurácia, na base de teste, de 95,00%.

Wu et al. (2009) utilizaram uma abordagem de classificação fundamentada em pesos para alguns marcadores do HTML. Os marcadores utilizados no trabalho foram *title*, H1, H2, H3, H4, H5, H6, B, U e I. Na fase de pré-processamento é realizada a extração desses marcadores, os demais são descartados e os termos restantes são considerados como texto da página.

A seleção dos atributos textuais é realizada com base na teoria *Rough Set* desenvolvida por Pawlak (1982) e Pawlak (2002). Após selecionados os atributos é executado um algoritmo Naïve Bayes de classificação. Esse algoritmo é adaptado para atribuir pesos diferentes de acordo com a etiqueta HTML em que o atributo está presente. O resultado dos experimentos demonstra que o uso da teoria *Rough Set* proporciona um índice de classificação melhor do que quando é utilizado os métodos TF, IDF ou TFIDF. Nos experimentos realizados foi feito uso de uma base de dados própria, em chinês, contendo 600 documentos. O classificador com uso de pesos obteve uma acurácia, nos experimentos, de 84,4%.

Caulkins et al. (2006) desenvolveram um método estatístico para a categorização de sites na internet. No método criado o conteúdo da página é dividido em duas partes: cabeçalho e corpo. O cabeçalho consiste de todo o texto presente dentro do elemento <head> do HTML, enquanto o corpo consiste do texto presente dentro do elemento <body>. Após realizada a divisão ocorre a deleção de todas as *tags* do HTML, permanecendo apenas o conteúdo textual. Com as informações textuais é criada uma lista de palavras para cada classe, essa lista é formada pelos termos mais comuns encontrados nas páginas que pertencem a classe em questão. Por fim, métricas estatísticas são geradas usando as informações das listas e do local em que o termo ocorre, corpo ou cabeçalho. As métricas são utilizadas como entrada de uma função que determina a qual classe pertence a página. A categorização da página é realizada pelo método estatístico desenvolvido pelo próprios pesquisadores, não é utilizado algoritmo de aprendizado de máquina. O trabalho usou para testes uma base contendo 930 documentos em Inglês, nesta base obteve-se uma Cobertura de 90,00%.

No trabalho de Hu et al. (2007) é proposto a análise das imagens para identificar sites pornográficos. O algoritmo de árvore de decisão C4.5 (QUINLAN, 1993) é utilizado para dividir as páginas, de acordo com o seu conteúdo, em três grupos: páginas com textos contínuos, páginas com textos não contínuos e páginas com imagens. Os textos contínuos são aquelas em que existe uma relação semântica e lógica entre as palavras, já textos não contínuos não apresentam essas propriedades. Cada um dos grupos é então analisado por um classificador específico. Os textos contínuos são analisados por uma *Cellular neural networks* (CNN) proposta por Chua e Yang (1988b) e Chua e Yang (1988a). A aplicação de CNN tem por objetivo encontrar a relação semântica entre as palavras. Os textos não contínuos são processados por um classificador Naïve Bayes, que calcula a probabilidade do texto ser pornográfico. As páginas com imagens são classificadas por um algoritmo que analisa o contorno dos objetos. Por último, a teoria do Bayes é utilizada para combinar o resultado dos classificadores de texto e imagem. Os classificadores não utilizam as informações estruturais das páginas, apenas as informações textuais. Os autores realizaram os testes em uma base de dados contendo 1500 páginas em

Inglês. No experimento mais completo foi atingido a acurácia de 93,50%.

Hammami et al. (2006) propuseram um classificador com base em análise de imagem, análise textual e informações estruturais. O classificador proposto utiliza duas listas, uma "lista negra" de sites e outra "lista negra" de palavras. O conteúdo textual é comparado com o conteúdo da lista de palavras, dessa comparação é extraída a informação de quantas palavras da lista constam no texto na página. Das informações estruturais é comparado o texto presente nos marcadores HTML <a> <img> *keywords* e URL (*Uniform Resource Locator*) com o conteúdo da lista de palavras, além dessas informações, também é verificado quantos *links* direcionam para a "lista negra" de sites. A análise de imagens procura identificar quantas imagens do site são pornográficas. As métricas geradas por todas essas análises são então processadas por um conjunto de algoritmos de árvore de decisão, incluindo o ID3 (QUINLAN, 1986) e o C4.5 (QUINLAN, 1993). Por fim, uma equação pondera o peso da cada algoritmo e fornece a classificação final para o site. Nos experimentos foram utilizadas duas bases de dados, uma delas com 400 sites e outra 12.311 sites pornográficos. Os sites escolhidos possuem conteúdo nos idiomas Inglês, Francês, Alemão, Espanhol e Italiano. A acurácia do sistema de classificação foi de 97,40% na base menor e de 95,62% na base maior.

Um classificador especializado em detectar páginas de conteúdo adulto foi proposto por Ahmadi et al. (2011). Os autores propõem a classificação das páginas em três categorias: permitida, imoral e pornográfica. As páginas categorizadas como imoral são aquelas que apresentam conteúdo ofensivo, como palavras de baixo calão. Para cada categoria é criada uma lista de termos comuns, além disso, também é criada uma "lista negra" com URLs de sites pornográficos. A classificação é realizada usando atributos extraídos dos conteúdos textuais, estruturais e visuais. O conteúdo textual é comparado com as listas de palavras, o mesmo procedimento é realizado para o conteúdo estrutural presente nas *tags* HTML <a>, <img>, *description* e *keywords*. Ainda durante a análise do conteúdo estrutural é contabilizado quantos *links* da página direcionam para URLs da "lista negra". As imagens são processadas por uma rede neural MLP (*Multilayer Perceptron*) que tem por objetivo encontrar figuras pornográficas. As métricas extraídas das informações textuais e estruturais são processadas por uma árvore de decisão ID3 (QUINLAN, 1986). O resultado dos algoritmos MLP e ID3 servem como entrada de um algoritmo que determinará a classe da página. O método proposto foi testado em uma base com 5.000 páginas nos idiomas Inglês e Persa, a acurácia atingida foi de 92,00%.

Outros trabalhos relacionados ao tema são: (a) Chen e Wu (2010) que propuseram um modelo de extração de atributos textuais para classificação de sites; (b) Zhong et al. (2010) os quais desenvolveram um método de descoberta automática de novos atributos textuais para

categorização de sites pornográficos; (c) Weitzner (2007) que apresentou um estudo sobre as diferentes propostas existentes na sociedade para identificação e bloqueio de conteúdo impróprio para crianças; (d) Zhou et al. (2005) que criaram uma metodologia, usando conteúdo textual e *Hyperlinks*, para localização de sites de grupos extremistas e grupos de intolerância; (e) Agarwal et al. (2006) propuseram um sistema de bloqueio de conteúdo inadequado baseado em informações da URL e o do código HTML; (f) Sam et al. (2007) realizam um estudo de categorização de sites comparando o desempenho de algoritmos de ICA (*Independent Component Analysis*) (COMON, 1994) e PCA (*Principal Component Analysis*) (PEARSON, 1901); (g) Santos et al. (2013) propuseram um classificador que utiliza algoritmos de compressão para representação dos dados ao invés do modelo VSM (*Vector Space Model*); (h) Ling et al. (2008) desenvolveram um classificador independente da linguagem, possibilitando que o treinamento seja realizado em um idioma e aplicado em outro e (i) Ho e Watters (2005) realizaram um estudo estatístico visando identificar as características das páginas de conteúdo adulto.

A Tabela 3 apresenta um resumo dos resultados obtidos pelos trabalhos apresentados nesta sessão. A acurácia informada na tabela refere-se ao experimento que obteve o melhor resultado. No caso dos autores que realizaram testes em mais de um idioma, foi escolhido apenas o idioma que atingiu o maior índice de instâncias corretamente classificadas.

Tabela 3: Relação de Trabalhos Relacionados a Classificação de Sites

Trabalho	Atributos Textuais	Atributos Estruturais	Imagens	Classificador	Acurácia
Kim e Nam (2006)	Sim	Não	Não	SVM	87,07%
Gao et al. (2008)	Sim	Não	Não	KNN e SVM	96,5%
Polpinij et al. (2006)	Sim	Não	Não	Naïve Bayes e SVM	100,00%
Lee e Luh (2008)	Sim	Não	Não	Método Qui-quadrado	96.88%
Lee et al. (2005)	Sim	Sim	Não	Rede Neural	95,00%
Wu et al. (2009)	Sim	Sim	Não	Bayesiano	84,4%
Caulkins et al. (2006)	Sim	Sim	Não	Método próprio	90,00%
Hu et al. (2007)	Sim	Não	Sim	C4.5, Rede Neural e Naïve Bayes	93,50%
Hammami et al. (2006)	Sim	Sim	Sim	Árvore de Decisão	97,40%
Ahmadi et al. (2011)	Sim	Sim	Sim	Rede Neural e Árvore de Decisão	92,00%

A Tabela 5 contém informações do idioma e tamanho dos conjuntos de dados utilizados pelos autores dos trabalhos relacionados. Todas essas bases são bases próprias construídas pelos próprios autores dificultando assim comparações entre os trabalhos.

Analisando os trabalhos apresentados nesta sessão alguns fatos podem ser observados:

- Conforme os dados da Tabela 3 existem vários trabalhos que utilizaram métodos parecidos, porém obtiveram resultados distintos. Esse fato ocorre devido aos métodos não terem sido aplicados na mesma base de dados. A performance do classificador é sensível aos dados de treinamento e teste. Um classificador terá uma performance melhor quando

Tabela 4: Conjuntos de Dados Utilizados nos Trabalhos Relacionados

Trabalho	Idioma	Quant. de páginas
Kim e Nam (2006)	Inglês e Coreano	20.000
Gao et al. (2008)	Inglês	1.400
Gao et al. (2008)	Chinês	1.400
Polpinij et al. (2006)	Inglês	1.400
Polpinij et al. (2006)	Tailandês	1.400
Lee e Luh (2008)	Inglês e Chinês	5.000
Lee et al. (2005)	Inglês	4.786
Lee et al. (2005)	Chinês	1.464
Wu et al. (2009)	Chinês	600
Caulkins et al. (2006)	Inglês	930
Hu et al. (2007)	Inglês	1.500
Hammami et al. (2006)	Inglês, Francês, Alemão, Espanhol e Italiano	400
Hammami et al. (2006)	Inglês, Francês, Alemão, Espanhol e Italiano	12.311
Ahmadi et al. (2011)	Inglês e Persa	5.000

as classes de testes possuírem atributos semelhantes entre si e diferentes em relação as outras classes.

- Alguns autores propuseram a análise das imagens para detectar sites pornográficos. Em geral, os algoritmos utilizados realizam essa detecção procurando tons de pele nas imagens. Essa abordagem apresenta bons resultados na identificação de conteúdo adulto, porém, dificilmente terá os mesmos resultados caso o objetivo seja localizar sites relacionados a outros assuntos, como por exemplo: economia ou política. Assim essa abordagem não é recomendada para um classificador que tenha como propósito ser genérico.

- As propostas que utilizam informações apenas do conteúdo textual deixam de utilizar informações úteis presentes no código HTML da página. Essas informações quando foram utilizadas nos outros trabalhos apresentaram uma melhora no índice de classificação.

- Os trabalhos utilizam diferentes algoritmos de aprendizado de máquina para realizar a classificação, como SVM, redes neurais, Naïve Bayes, algoritmos de agrupamento e árvores de decisão. Porém nenhum dos trabalhos apresenta experimentos que justifiquem o porque foi escolhido determinado classificador e não outro. Acredita-se que tais experimentos sejam relevantes, pois os classificadores utilizados são bem distintos o que pode interferir nos resultados.

- Os estudos que realizam classificação utilizando algum método de seleção de atributos textuais não apresentam comparações envolvendo as diferentes métricas existentes na literatura para seleção de atributos, além disso não informam de que forma foi deter-

minado o total de atributos a serem selecionados. Esses dois parâmetros: métrica de seleção e dimensionalidade são muito importantes pois podem influenciar nos índices de classificação.

- O resultado do classificador depende do idioma do conteúdo, isso fica evidenciado pelos trabalhos que realizam experimentos em bases com mais de um idioma. Esse fato está relacionado a fatores linguísticos que influenciam no pré-processamento de texto. Não foi encontrado na literatura experimentos específicos para o idioma Português.
- Os trabalhos que fazem uso das informações estruturais não apresentam comparações sobre os atributos selecionados. Desta forma, não é evidenciado se um determinado atributo contribui positivamente para a construção do modelo de classificação.

2.9 CONSIDERAÇÕES FINAIS

Este Capítulo apresentou tópicos relacionados a classificação automática de conteúdo *web*. Embora a categorização de páginas *web* possa ser vista com uma forma de classificação automática de textos ela difere em alguns pontos. Um documento HTML apresenta uma estrutura diferente de um documento comum de texto. Ela é composto por marcadores, hiperligações, conteúdo multimídia, entre outros. Essa estrutura, rica em informações, fornece uma gama enorme de possibilidade de construção de classificadores.

O conteúdo estrutural pode auxiliar o trabalho do classificador, possibilitando que a página seja categorizada com base na análise de parte de seu conteúdo. A diversidade de marcadores e outras informações presentes nos sítios *web* propicia o surgimento de inúmeras propostas para categorização. Assim torna-se relevante a elaboração de estudos que apresentem comparações do desempenho dos classificadores com uso de diferentes marcadores.

Além dos marcadores, um aspecto também importante são as métricas de avaliação de atributos. Essas métricas são as responsáveis por estimar a qualidade que determinado atributo pode possuir no que diz respeito a classificação. Cada uma dessas métricas possui requisitos diferentes de avaliação. Deste modo é importante que sejam realizados estudos que apresentem quais métricas melhor se adaptam a este gênero de classificação.

As métricas de avaliação ponderam a qualidade dos atributos, enquanto a dimensionalidade representa quantos desses atributos serão efetivamente utilizados na classificação. A quantidade de atributos também pode afetar os resultados atingidos pelo classificador. Portanto, também é importante a realização de estudos que demonstrem qual é a dimensionalidade ideal para o problema proposto.

Na sessão anterior foram apresentadas diferentes abordagens para a resolução do problema de categorização de conteúdo *web*. Pode-se observar nesses trabalhos o uso de algoritmos de classificação de paradigmas distintos. Poucos são os trabalhos que apresentam estudos que justificam a escolha de determinado paradigma em detrimento de outrem.

Para aferir a qualidade dos resultados apresentados pelos algoritmos de categorização inúmeras métricas podem ser utilizadas. Segundo Sebastiani (2002) a Medida-F é a métrica ideal para uso em trabalhos relacionados a classificação de textos. Porém, conforme observado nos trabalhos apresentados na sessão 2.8 a Acurácia é a métrica mais utilizada nos trabalhos relacionados.

Devido a todos os pontos acima expostos identificam-se que continuam abertas algumas questões relacionadas a classificação automática de sítios *web*. Este trabalho visa contribuir para a área com a apresentação de estudos empíricos que objetivam respondem algumas dessas questões.

3 METODOLOGIA

Neste Capítulo será apresentada e detalhada a metodologia utilizada para o desenvolvimento deste trabalho, com alguma ênfase aos requisitos tecnológicos necessários à efetiva implantação de proposta em um ambiente real. A sessão 3.1 relata a necessidade de proposição de um ambiente para implantação do classificador proposto. Na sessão 3.2 é abordada a questão do conjunto de dados utilizados nos experimentos. Já as sessões 3.3 e 3.4 abordam respectivamente assuntos relacionados ao pré-processamento de texto e algoritmos de classificação e seleção de atributos utilizados neste trabalho. Por fim, a sessão 3.5 comenta como será feita a análise dos resultados.

3.1 AMBIENTE PARA IMPLANTAÇÃO

O foco deste trabalho está na classificação automática de sites e tem como um dos seus objetivos a aplicação desta classificação no controle de acesso a sites indesejados. Desta forma, é necessário que o método de classificação proposto possa ser implementado em um ambiente de rede de computadores que propicie o bloqueio de sites categorizados como inadequados.

Alguns métodos de classificação, como os que analisam o conteúdo presente em vídeos, podem levar muito tempo para realizar a classificação de um sítio como um todo. Outros métodos que podem não ser eficientes em relação ao tempo de classificação são aqueles que utilizem informações de sites ancestrais e descendentes em relação ao site que esta sendo classificado.

Portanto é necessário que o classificador desejado possa trabalhar adequadamente em ambientes tempo-real. A classificação de um nova instância deve ocorrer em poucos segundos para que não comprometa a navegação do usuário.

Em relação ao ambiente de rede para implantação do classificador é desejável que seja robusto, escalável e ao mesmo tempo de performance adequada para que não cause demasiado atraso na navegação *web*. O sítio acessado pelo usuário pode não ter sido classificado ainda, o

ambiente proposto deve tratar desta situação permitindo que o usuário possa navegar pelo sítio até que seja realizada a classificação ou então esperar até que a classificação ocorra, dependendo da política de acesso adotada pela instituição.

Assim, ambiente de implantação e método de classificação devem ser compatíveis para que a solução de filtro de conteúdo funcione apropriadamente. Devido a isto, uma das primeiras atividades realizadas foi a elaboração de um ambiente para implantação do algoritmo de classificação que está sendo proposto. No Apêndice A é apresentado o ambiente sugerido para uso do método de classificação proposto.

3.2 AQUISIÇÃO DO CONJUNTO DE DADOS

Visto que este trabalho propõe um sistema de categorização baseado em aprendizagem supervisionada é necessário um conjunto de dados rotulados para treinamento do classificador. Os dados devem ser compostos por páginas *web* de duas classes, pornográficas e não pornográficas. Além da questão das classes as páginas devem possuir conteúdo nos idiomas Português e Inglês. Os idiomas distintos são necessários para mensurar o desempenho do método proposto em línguas diferentes, uma vez que o desempenho do classificador pode variar devido a peculiaridades de cada idioma.

Outro requisito importante para o conjunto de dados é que as páginas estejam disponíveis em seu formato HTML (*HyperText Markup Language*). Desta forma, é possível extrair e avaliar o conteúdo individual de cada etiqueta (*tag*) HTML. Com isso pode-se, por exemplo, avaliar o desempenho do classificador usando como dado de entrada a combinação de duas ou mais etiquetas da linguagem HTML.

Para efeitos de comparação com outros trabalhos da literatura o ideal seria utilizar uma mesma base de dados já utilizada por outros pesquisadores. Visando encontrar um base de dados pública para o domínio do problema foi realizado uma busca em diferentes locais. Entretanto não foi encontrada nenhuma base de dados, disponíveis para uso e comparação com outros trabalhos, que atendesse os requisitos necessários.

Conforme já diagnosticado por Qi e Davison (2009) a inexistência de um conjunto de dados padrão é uma significativa desvantagem as pesquisas relacionadas a classificação de conteúdo na *web*. Por este motivo constatou-se a necessidade da construção de um conjunto de dados que possuísse as características para a execução dos experimentos deste trabalho. O Capítulo 4 aborda a construção deste conjunto de dados, explicando a metodologia utilizada e o resultado atingido.

3.3 PRÉ-PROCESSAMENTO

Neste trabalho esta sendo realizado a categorização de páginas *web*. Como os dados a serem pré-processados são páginas de internet o fluxo é um pouco diferente do pré-processamento de um texto comum.

Um dos primeiros problemas enfrentados para o pré-processamento de páginas *web* é a questão da codificação dos caracteres. Apesar da existência de uma RFC (*Request for Comments*), produzida pela IETF (*Internet Engineering Task Force*), recomendando que os protocolos de comunicação da internet suportem a codificação Unicode UTF-8 (YERGEAU, 2003) ainda é comum encontrar servidores *web* que não suportam essa codificação. A falta de uma codificação de caracteres do alfabeto latino adotada como padrão dificulta a análise das páginas.

Tanto o protocolo HTTP (FIELDING et al., 1999) quanto o protocolo HTML (W3C, 1999) especificam maneiras para que seja informado qual é a codificação adotada no documento que esta sendo transmitido. Porém, é comum encontrar servidores em que esses dados não são informados, ou documentos HTML que não apresentam tais informações. Outro problema que ocorre é quando a codificação informada não corresponde com a utilizada, por exemplo, o HTML informa que o conteúdo do documento está em UTF-8 porém o documento foi gravado usando a codificação ISO-8859-1. Caso a codificação não seja corretamente descoberta a análise do documento fica comprometida, principalmente para o caso de palavras acentuadas. Todos esses fatores fazem com que seja necessário realizar testes heurístico para descobrir a codificação correta.

Uma particularidade do pré-processamento de documentos da internet são as etiquetas do HTML. O padrão (W3C, 1999) define os requisitos para que um documento HTML seja considerado válido. Contudo, são poucos sites que seguem as regras definidas pelo padrão. Os navegadores (*browsers*) conseguem interpretar e exibir o conteúdo de um documento HTML mesmo que ele não seja um documento válido, isso faz com que os desenvolvedores não se preocupem em corrigir os documentos e deixá-los aderentes as normas. Devido a este fato, durante o pré-processamento é necessário que o programa utilizado para extrair as informações das páginas consiga obter os dados mesmo em um documento que não siga as especificações do protocolo.

Após superados esses dois problemas, codificação e desrespeito as normas do protocolo HTML, teve início o pré-processamento da base propriamente dito. Foram executadas as seguintes fases:

- *Tokenization*: consiste na quebra do fluxo de caracteres em palavras, também chamadas de *tokens*. Neste trabalho os *tokens* foram separados por caracteres não alfabéticos, tais como: números, símbolos e caracteres especiais.
- Remoção de *Stopwords*: remoção de palavras comuns do idioma. Neste trabalho foi utilizado uma lista de palavras, construída por Soares et al. (2008).
- *Stemming*: processo que transforma cada termo para o radical que o originou.

Para o processo de *stemming* nos textos em Inglês foi utilizado o algoritmo proposto por Porter (1980). Já para o Português foi utilizado o algoritmo proposto por Soares et al. (2009), que é uma adaptação do algoritmo do Porter (1980) para este idioma.

Por fim, os textos analisados estão prontos para a seleção de atributos e posteriormente serem processadas pelos algoritmos de aprendizagem de máquina, essas tarefas são melhores descritas na sessão seguinte 3.4.

3.4 SELEÇÃO DE ATRIBUTOS E CLASSIFICADORES

Uma vez os dados preprocessados, conforme descrito na sessão 3.3, teve início o processo de seleção de atributos textuais. A seleção de atributos é um dos pontos determinantes para o resultado do classificador. Embora algumas vezes não seja realizada corretamente, esse é um ponto fundamental que pode impactar positivamente ou negativamente no resultado da classificação.

A Figura 11 apresenta o processo para que as informações da página HTML possam ser processadas pelos algoritmos de classificação. Na primeira etapa o conjunto de documentos HTML passa por um processo que separa os termos presentes em várias *tags* do HTML, em seguida é realizado o pré-processamento de cada um desses conjuntos de *tags*. Na segunda etapa são aplicadas as métricas de avaliação de atributos a estes conjuntos de termos (separados por *tags*). Na terceira etapa é feita a seleção de *N* atributos. Por último, são utilizados os atributos selecionados para geração dos arquivos a serem processados pelos algoritmos de classificação.

As métricas de avaliação de atributos utilizadas foram: TF (*Term Frequency*), TFIDF (*Term Frequency Inverse Document Frequency*) e Ganho de Informação. A sessão 2.4 detalha seu funcionamento.

Utilizou-se uma abordagem filtro, na qual os atributos são selecionados a priori, de forma independente do algoritmo de classificação. Esta abordagem foi considerada adequada

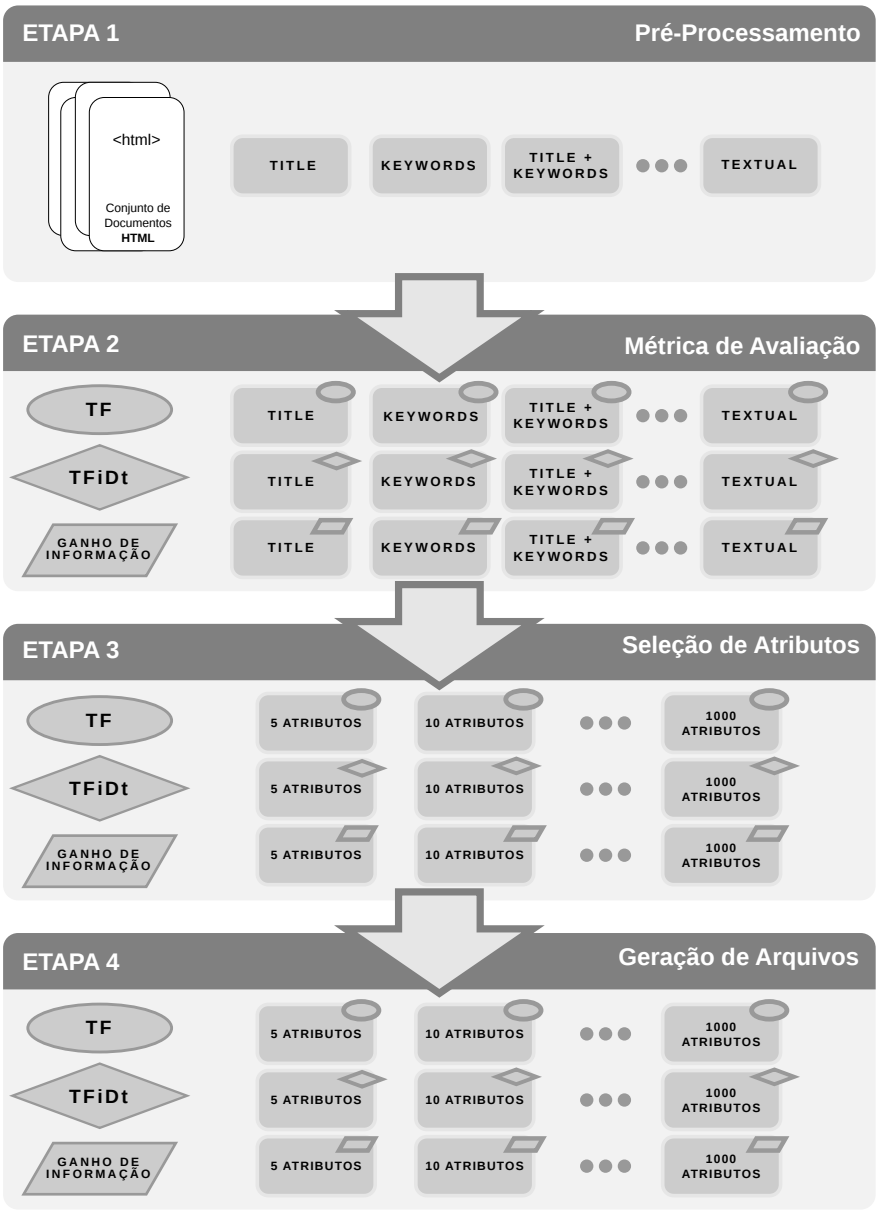

Figura 11: Fluxo para Geração dos Arquivos de Atributos

para este problema devido aos requisitos de performance (tempo real) exigidos. Essa abordagem funciona da seguinte forma: os atributos (termos) avaliados individualmente pela métrica de avaliação são colocados em uma lista *L*. A lista é ordenada de acordo com o índice de qualidade de cada atributo; são selecionados os *N* primeiros atributos da lista *L* formando uma nova lista que será utilizada pelos algoritmos de classificação. Neste trabalho foi utilizado um valor de *N* variando entre 5 e 1000.

Após utilizar a abordagem filtro, foi usado o modelo "*bag-of-words*" para representação das páginas HTML do conjunto de dados *D*. Usando os atributos selecionados pelo filtro, temos $T = T_1, ..., T_N$, no qual, *N* é o total de diferentes atributos (termos) selecionados. As-

sim, cada página HTML d_i, na qual $d_i \in D$, corresponde a um vetor N-dimensional, ou $d_i = [w_{i1}, w_{i2}, ..., w_{iN}]$, aonde w_{ij} é o peso do termo t_j no documento d_i. Atribuiu-se pesos booleanos para cada elemento w_{ij}, em que $w_{ij} = 1$ se o termo t_j aparece no documento d_i ou $w_{ij} = 0$ caso não conste.

Os arquivos gerados para os classificadores, quarta etapa do processo mostrado na Figura 11, estão no formato ARFF (*Attribute-Relation File Format*). O formato ARFF é um arquivo de texto codificados em ASCII (*American Standard Code for Information Interchange*) que descreve uma lista de instâncias que possuem um conjunto de atributos. Foi utilizado este formato devido a sua facilidade de compreensão e compatibilidade com o *software* Weka (HALL et al., 2009).

A escolha dos classificadores utilizados nos experimentos foi embasada no estudo desenvolvido por Wu et al. (2008) e também na pesquisa realizada nos trabalhos relacionados, descrita na sessão 2.8. Os requisitos para escolha foram: ser um algoritmo de aprendizagem supervisionada; ser amplamente utilizado pela comunidade científica; possuir bons resultados na classificação de páginas *web* ou de texto.

Com base nessa pesquisa e nos requisitos definidos foram selecionados os seguintes algoritmos de classificação:

(a) O probabilístico Naïve Bayes.

(b) O algoritmo baseado em instâncias KNN, com K=1.

(c) O classificador baseado em árvores de decisão C4.5.

(d) O classificador SVM, utilizando a implementação SMO (*Sequential Minimal Optimization*) empregando uma função de núcleo (kernel) polinomial.

(e) O algoritmo de redes neurais MLP (*Multilayer Perceptron*).

Na execução dos experimentos foi utilizado o *software* Weka (HALL et al., 2009) devido a ele possuir uma base abrangente de algoritmos de aprendizagem de máquina, além de ser amplamente utilizada no meio acadêmico. Outra vantagem do software é a sua API (*Application Programming Interface*) que possibilita a execução automática de experimentos.

3.5 ANÁLISE DOS RESULTADOS

Após selecionados os algoritmos de aprendizagem de máquina teve início a fase de execução dos experimentos. Para avaliação dos resultados foi utilizada a métrica Médida-F.

Conforme explicado por Sebastiani (2002) a Medida-F é ideal para mensurar a qualidade de um classificador de texto.

Inicialmente foi definido quais classificadores seriam testados, bem como quais métricas de avaliação de atributos seriam utilizadas, faltando apenas a definição do total de atributos a ser utilizado. O total de atributos ou dimensionalidade é um aspecto importante para a classificação, pois, um número muito pequeno ou muito grande pode dificultar o aprendizado do classificador. Para definir qual seria o intervalo abrangido pelos experimentos foram realizados testes empíricos aumentando a dimensionalidade até o momento em que constatou-se uma tendência de estabilização dos resultados.

Os experimentos realizados abrangeram testes envolvendo a seleção de 5 a 1000 atributos. Conforme já esperado, alguns classificadores tiveram uma tendência de diminuição da Medida-F com o aumento da dimensionalidade, enquanto outros apresentaram uma tendência de estabilização.

No Capítulo 5 são apresentados os resultados detalhados dos experimentos. Nesse mesmo Capítulo são apresentadas informações sobre: qual classificador apresentou os melhores resultados; a melhor métrica de seleção de atributos textuais para o problema em questão; a quantidade ideal de atributos textuais; o desempenho do algoritmo de classificação para as páginas em Português em relação a páginas em Inglês.

4 BASE DE DADOS

Este trabalho utiliza algoritmos de aprendizagem supervisionada. Este paradigma de aprendizagem necessita que existam dados previamente rotulados para que se possa aplicar o algoritmo de aprendizado. Esta sessão apresentará informações relacionadas aos dados que foram utilizados para execução dos experimentos. Já a sessão 4.2 aborda os detalhes relacionados a construção da base de dados própria, necessária para execução dos experimentos. Por fim, a sessão 4.3 mostra algumas estatísticas relacionadas a essa base.

4.1 AUSÊNCIA DE BASE DE DADOS PADRÃO

Nos trabalhos relacionados não foi encontrado uma base única utilizada por todos os pesquisadores, conforme pode-se constatar pelas descrições presentes na sessão 2.8. Este problema de ausência de um conjunto de dados padronizados disponível para uso também foi diagnosticado no trabalho de Ahmadi et al. (2011). Segundo o autor, não existe uma base única utilizada pelos pesquisadores desta área, os trabalhos publicados neste campo, em geral, são baseados em conjuntos de dados específicos, os quais geralmente não estão disponíveis ao público.

Segundo Qi e Davison (2009) a ausência de um conjunto de dados padrão, que represente a *web*, é uma significante desvantagem para as pequisas de classificação de conteúdo da internet. Os autores também afirmam que a inexistência de dados apropriados tem atrasado o progresso das pesquisas relacionados a categorização de páginas.

Os trabalhos relacionados descritos na sessão 2.8 não utilizam uma base única, cada pesquisador utilizou um base diferente. Todas as bases utilizadas foram construídas pelos próprios pesquisadores com critérios e metodologias próprias. Em geral, as bases foram construídas com informações extraídas por pesquisas realizadas na internet. A Tabela 5 apresenta informações a respeito desses conjuntos de dados.

Devido a não ter sido encontrado um conjunto de dados que atendesse os requisitos

Tabela 5: Tamanho dos Conjuntos de Dados Usados nos Trabalhos Relacionados

Trabalho	Idioma	Qtde de páginas
(KIM; NAM, 2006)	Inglês e Coreano	20.000
(GAO et al., 2008)	Inglês	1.400
(GAO et al., 2008)	Chinês	1.400
(POLPINIJ et al., 2006)	Inglês	1.400
(POLPINIJ et al., 2006)	Tailandês	1.400
(LEE; LUH, 2008)	Inglês e Chinês	5.000
(LEE et al., 2005)	Inglês	4.786
(LEE et al., 2005)	Chinês	1.464
(WU et al., 2009)	Chinês	600
(CAULKINS et al., 2006)	Inglês	930
(HU et al., 2007)	Inglês	1.500
(HAMMAMI et al., 2006)	Inglês, Francês, Alemão, Espanhol e Italiano	400
(HAMMAMI et al., 2006)	Inglês, Francês, Alemão, Espanhol e Italiano	12.311
(AHMADI et al., 2011)	Inglês e Persa	5.000

julgou-se necessário a construção de uma base de dados. Mesmo os conjunto de dados listados na Tabela 5, caso fossem públicos, não poderiam ser utilizados devido a não possuírem algumas características requiridas.

Conforme pode-se observar pela Tabela 5 a relação de bases utilizadas pelos trabalhos apresentados na sessão 2.8 não contemplam dados no idioma Português. Com isso não seria possível realizar experimentos para esse idioma, comprometendo um dos objetivos deste trabalho que é a execução de experimentos de classificação de conteúdo no idioma Português.

Outro requisito para o conjunto de dados é que as páginas estejam em seu formato HTML (*HyperText Markup Language*) da mesma forma que as páginas originais disponíveis na internet. Isso é importante para que seja possível extrair e analisar o conteúdo de cada marcador HTML individualmente. Deste modo, pode-se avaliar o desempenho de diferentes classificadores usando combinações desses marcadores. Algumas bases utilizadas nos trabalhos relacionados não possuem dados neste formato. Deste modo, mesmo que essas bases fossem públicas não seria possível a sua utilização, devido a não possuírem as informações necessárias para execução dos experimentos.

4.2 CONSTRUÇÃO DA BASE DE DADOS.

Conforme exposto anteriormente na sessão 4.1 foi necessário criar um novo conjunto de dados para a execução dos experimentos. A base construída possui sítios em dois idiomas, Inglês e Português. O idioma Inglês foi escolhido devido a ser o mais utilizado na internet, pos-

suindo o maior número de páginas disponíveis na *web* (GREFENSTETTE; NIOCHE, 2000). Já o idioma Português foi escolhido devido a ser a língua oficial do Brasil e também por não ter sido encontrado na literatura trabalhos neste idioma. A escolha de dois idiomas é importante pois a classificação de textos é sensível às peculiaridades da língua. Desta forma, um classificador treinado usando páginas de determinado idioma dificilmente vai ter a mesma performance quando for aplicado em um segundo idioma.

Em relação as classes consideradas no problema, foi definido que a base seria constituída de páginas de duas classes, pornográficas e não pornográficas. As páginas de conteúdo pornográfico são aquelas que possuem conteúdo sexuais, não recomendado para menores de idade, de acordo com os objetivos da aplicação. Já a classe não pornográficas é composta pelas demais páginas, tais como páginas de notícias, esportes, economia, entre outras. Assim, o classificador deve ser capaz de identificar corretamente as páginas de conteúdo pornográfico, possibilitando o bloqueio a sites com tais conteúdo.

Ao todo foram selecionadas 4.000 páginas. As páginas foram extraídas de duas fontes de informações. A primeira fonte de dados foi um relatório de acesso de um conjunto de *proxys* utilizados por diversas escolas de ensino fundamental e médio, essa lista foi essencial para a descoberta de páginas em Português, e ligado diretamente à aplicação desejada do projeto. A segunda fonte foi a lista dos sites de maior audiência do mundo, produzida pela empresa Alexa que possui uma rede de sensores coletando informações de acesso dos usuários (ALEXA, 2012). A lista produzida pela Alexa (2012) foi essencial para a descoberta de sites no idioma Inglês. Assim, foi utilizado as páginas que possuíam o maior número de acessos no conjunto de *proxys* e também as páginas com a maior posição no *ranking* produzido pela empresa Alexa.

Todas as páginas foram então classificadas manualmente, pelo pesquisador, em uma das duas categorias: pornográfica e não pornográfica. Nesse processo de classificação foi utilizado uma abordagem conservadora, caso a página apresentasse um único elemento pornográfico ela era então categorizada como pornográfica. Além disso não foram utilizadas páginas em que o conteúdo não fosse bem definido, e que pudessem apresentar dúvidas sobre qual deveria ser a sua classe.

As páginas selecionadas pertencem a uma lista de 400 sites, em média foi utilizado 10 páginas de cada site. Para definição do total de páginas selecionadas foi utilizado como referência o tamanho das bases utilizadas nos trabalhos relacionados, conforme dados apresentados na Tabela 5.

A Tabela 6 apresenta informações sumarizadas sobre o conjunto de dados, tais como o total de páginas e também a distribuição entre as classes.

Tabela 6: Conjunto de Dados Construído para os Experimentos

Classe	Páginas em Português	Páginas em Inglês
Pornográfica	1.000	1.000
Não Pornográfica	1.000	1.000

4.3 ESTATÍSTICAS SOBRE A BASE DE DADOS

Esta subsessão apresenta algumas estatísticas da base de sítios utilizada.

A Tabela 7 contêm informações sobre o total de *tokens* do conjunto de dados antes do pré-processamento. A Tabela apresenta os totais agrupados por idioma (Português e Inglês) e por classe (pornográfica, não pornográfica). A coluna *Tokens* mostra o total encontrado em cada um dos grupos. A coluna seguinte mostra o total de *tokens* únicos encontrados nos respectivos conjuntos de dados. A última coluna da Tabela apresenta a relação de *tokens* únicos em relação ao total de *tokens* encontrados.

Tabela 7: Total de *Tokens* do Conjunto de Dados Antes do Pré-Processamento

Idioma	Classe	*Tokens*	*Tokens* Únicos	% de *Tokens* Únicos
Português	Pornográfica	1.165.382	35.756	3,07%
Português	Não Pornográfica	2.374.449	130.982	5,52%
Português	Todas	3.539.831	152.548	4,31%
Inglês	Pornográfica	2.106.265	71.694	3,40%
Inglês	Não Pornográfica	2.048.679	107.097	5,23%
Inglês	Todas	4.154.944	157.106	3,78%

O total de termos após realizado o pré-processamento é mostrado na Tabela 8. As etapas utilizadas no pré-processamento são descritas na sessão 2.4. A estrutura da Tabela 8 e a semântica das colunas é semelhante a utilizada na Tabela 7, na qual: a primeira e a segunda coluna mostram a informação do idioma e classe respectivamente; a terceira coluna mostra o total de termos; a quarta coluna mostra o total de termos únicos e a última coluna da Tabela apresenta a relação de termos únicos em relação ao total de termos encontrados.

Conforme os dados das Tabelas 7 e 8 pode-se observar que a classe pornográfica é a que possui o menor percentual de termos e *tokens* únicos, o que pode ser explicado devido as páginas da outra classe serem de assuntos variados, não concentrados no mesmo tema.

Ainda analisando os dados apresentados por essas Tabelas pode-se observar que ocorreu um decréscimo significativo do percentual de termos únicos após o pré-processamento. Isso ocorreu principalmente devido a aplicação do algoritmo de *stemming* que visa transformar cada *token* para o seu radical. Com os números apresentados é possível avaliar a eficiência do algo-

ritmo de *stemming*.

Tabela 8: Total de Termos do Conjunto de Dados Após o Pré-Processamento

Idioma	Classe	Termos	Termos Únicos	% de Termos Únicos
Português	Pornográfica	772.476	10.359	1,34%
Português	Não Pornográfica	1.492.443	47.696	3,20%
Português	Todas	2.264.919	51.236	2,26%
Inglês	Pornográfica	1438249	27.941	1,94%
Inglês	Não Pornográfica	1353743	44.609	3,30%
Inglês	Todas	2791992	59.676	2,14%

Os percentuais de redução do texto após o pré-processamento são apresentados na Tabela 9. Conforme os dados é possível constatar que a tendência de redução do total de termos é a mesma entre os grupos, variando entre 31,72% e 37,15%. O total de termos únicos também apresenta uma tendência de redução, porém maior, variando de 58,35% à 71,03%.

Tabela 9: Percentual de Redução de Termos Após o Pré-Processamento

Idioma	Classe	% de Redução do Total de Termos	% de Redução do Total de Termos Únicos
Português	Pornográfica	33,71%	71,03%
Português	Não Pornográfica	37,15%	63,59%
Português	Todas	36,02%	66,41%
Inglês	Pornográfica	31,72%	61,03%
Inglês	Não Pornográfica	33,92%	58,35%
Inglês	Todas	32,80%	62,02%

A Tabela 10 mostra os percentuais de presença dos marcadores HTML (*HyperText Markup Language*) nas páginas do conjunto de dados. As colunas da Tabela apresentam as informações de cada idioma e classe, enquanto as linhas representam os marcadores ou conjunto de marcadores. A sessão 2.3 apresenta uma relação dos principais marcadores utilizados pela linguagem e seus respectivos significados.

Abaixo uma breve explicação de cada marcador ou conjunto de marcadores utilizados na Tabela 10:

- Âncora: Contém o texto presente entre os símbolos <a> e do marcador HTML <a>.
- *Description*: Possui o texto do elemento *content* do marcador *description*.
- Destaque: Representa a junção dos textos presentes nos marcadores utilizados para destaque de informações da página *web*, são eles: <b>, <i>, <u>, <s>, e .

- Imagem: Contêm os textos dos atributos *alt* e *title* do marcador <img>.
- *Keyword*: Possui o texto do elemento *content* do marcador *keyword*.
- *Link*: Contêm os textos dos atributos *alt* e *title* do marcador <a>.
- *Title*: Contém o texto presente no marcador <title>, localizado no cabeçalho da página.
- Título: Representa a junção dos textos presentes nos marcadores utilizados para simbolizar títulos de textos, são eles: <h1>, <h2>, <h3>, <h4>, <h5> e <h6>.

Tabela 10: Percentual de Uso dos Marcadores HTML no Conjunto de Dados

Marcador	Português Não Pornográfica	Português Pornográfica	Inglês Não Pornográfica	Inglês Pornográfica
Âncora	99,90%	99,90%	100,00%	99,80%
Description	72,00%	66,80%	75,50%	85,30%
Destaque	86,70%	86,70%	77,50%	81,00%
Imagem	91,90%	96,50%	92,90%	89,90%
Keyword	62,60%	63,40%	57,60%	72,70%
Link	80,90%	88,60%	76,50%	76,50%
Title	100,00%	99,80%	99,40%	99,60%
Título	82,90%	81,10%	93,30%	74,10%

Os dados presentes na Tabela 10 mostram que os elementos HTML mais populares nas páginas são os marcadores *title* e <a>, o último referenciado na Tabela como Âncora. Esses atributos estão presentes em praticamente todas as páginas do conjunto de dados. O alto percentual de uso desses marcadores fazem deles bons candidatos de fonte de informações para os classificadores.

O atributo Imagem também possui alto percentual de presença, variando de 89,90% a 96,50%. O valor elevado deste marcador indica que a maioria das páginas do conjunto de dados possuem figuras e não apenas informações textuais.

A Tabela 10 também mostra que os marcadores de metadados *Description* e *Keyword* não são tão populares quanto os demais, principalmente o marcador *Keyword*. Esses marcadores representam metadados da página HTML, são utilizados para sumarizar as informações ali presentes. Porém, a baixa frequência com que estão presentes prejudica o desempenho de um classificador que vier a utilizá-los como única fonte de dados para categorização.

Estudos desenvolvidos por outros pesquisadores já discutiram a pouca utilidade de determinados marcadores, a citar:

(a) Pierre (2000) descreveu como uma das dificuldades para a categorização de páginas o pouco uso de alguns atributos como *Description* e *Keyword*;

(b) Pierre (2001) realizou um estudo que mostra que os marcadores *Description* e *Keyword* são os mais ausêntes nas páginas da internet;

(c) Alimohammadi (2004) desenvolveu um estudo sobre a *web* iraniana, os resultados da pesquisa evidenciam que os marcadores de metadados são pouco usados nos sítios daquele país;

(d) Craven (2004) pesquisou o uso do marcador *Description* em páginas de diferentes idiomas, os resultados mostram que o percentual de uso varia de acordo com o idioma, porém são sempre percentuais menores comparados com o de outros atributos;

(e) Shen et al. (2003) compararam o desempenho de classificadores com o uso de diferentes marcadores. Os resultados obtidos demonstram que classificadores baseados apenas em informações dos atributos *Description* e *Keyword* possuem um desempenho inferior comparado a classificadores que fazem uso de outros atributos. Um dos motivos para o baixo desempenho é a quantidade de atributos faltantes quando se utiliza informações desses dois marcadores.

As pesquisas citadas acima demonstraram que alguns marcadores são menos populares que outros. As informações presentes na Tabela 10 complementam essas pesquisas mostrando os percentuais de uso de alguns marcadores não abordados nos estudos anteriores, além disso, mostram informações segmentadas por idioma e classe.

Um fator que pode explicar porque os marcadores de metadados *Description* e *Keyword* estão sendo pouco utilizados é que alguns mecanismos de busca não mais utilizam as informações presentes em tais elementos (GOOGLE, 2009). Antigamente as informações presentes nesses marcadores eram utilizadas pelos algoritmos de busca para definir se a página seria exibida ou não quando um usuário procurasse determinado termo, além de definir a posição da página na ordem dos resultados. Alguns sítios estavam manipulando as informações desses metadados com o objetivo de ganhar posições nos resultados das buscas. Devido a isso algumas empresas de buscas passaram a não mais utilizar esses marcadores.

Além das informações apresentadas na Tabela 10 também foi realizado um estudo a respeito dos sites ligados por meio de *Hyperlinks* ao conjunto de dados construído. Esse estudo visava identificar a classe a qual pertenciam os sites pais e filhos das páginas pertencentes ao conjunto de dados.

Site pai pode ser definido como um site que possui uma ou mais hiperligações para a página que está sendo classificada, geralmente expressadas pelo marcador HTML <a>. En-

quanto um site filho pode ser definido como um site referenciado pela página que está sendo classificada. A Figura 12 mostra um exemplo de site pai e site filho, mostrando a sua relação com a página em análise (QI; DAVISON, 2009).

Os sites filhos foram identificados apenas com a análise do conteúdo HTML das páginas da base de dados. Já a descoberta de sites pais foi realizada com uso de mecanismos de busca na internet. As ferramentas de busca procuram apenas as páginas por elas indexadas, deste modo, acredita-se que a quantidade real de sites pais existentes seja superior ao identificado por essas ferramentas.

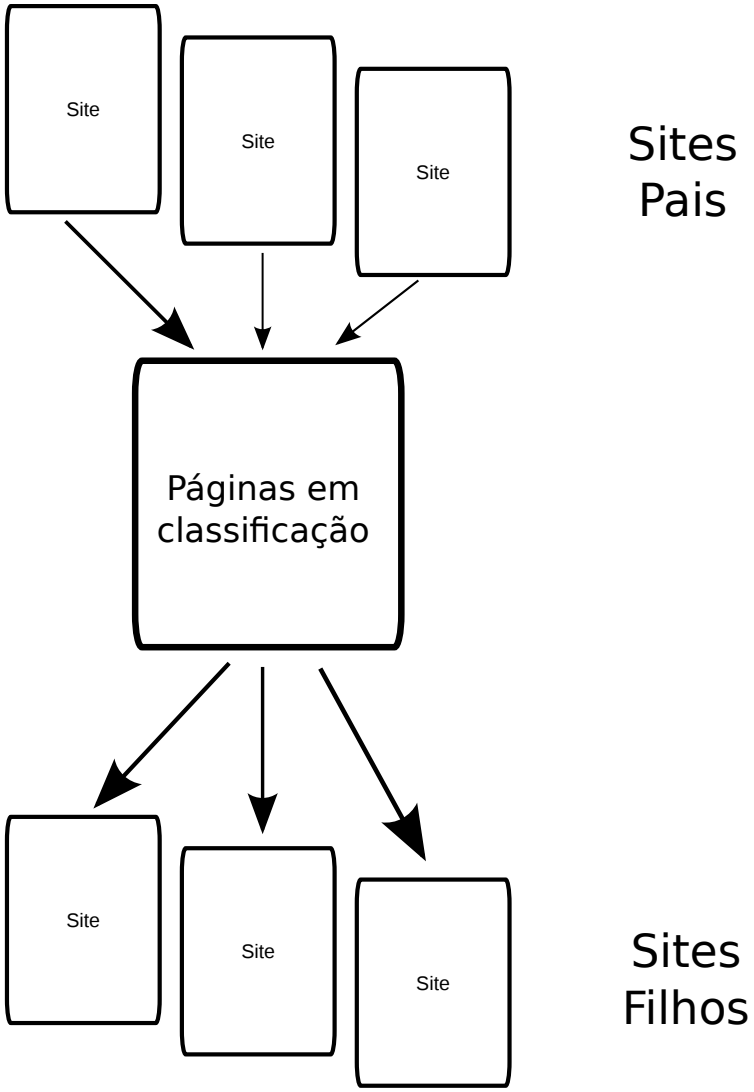

Figura 12: Exemplo de Sites Pais e Filhos

O percentual de páginas que possuem sites pais com conteúdo pornográfico é mostrado na Tabela 11. Os dados estão agrupados por idioma e classe. Conforme pode-se observar as

páginas da classe não pornográficas são as que possuem o maior número de sites pais. Merece destaque as páginas pornográfica em Português que apresentam o menor número de sites pais, apenas 67,00%.

Continuando a análise dos dados da Tabela 11 pode-se constatar uma alta relação da classe dos sites pais com a classe da página em análise, principalmente para a classe pornográfica. No idioma Português 67,00% das páginas pornográficas possuem sites pais, sendo que 60,00% são sites pais com conteúdo pornográfico. A tendência é a mesma para o idioma Inglês, 92,00% das páginas possuem sites pais, sendo que 82,00% possuem pais pornográficos. Esse percentual de páginas que possuem sites pais pornográficos correspondem a um percentual sobre o total de páginas existentes nos referidos idiomas e classes.

Tabela 11: Percentual de Páginas do Conjunto de Dados que Possuem Sites Pais com Conteúdo Pornográfico

Idioma	Classe	% de Páginas que Possuem Sites Pais	% de Páginas que Possuem Sites Pais Pornográficos
Português	Não Pornográfica	98,00%	10,00%
Português	Pornográfica	67,00%	60,00%
Inglês	Não Pornográfica	100,00%	17,00%
Inglês	Pornográfica	92,00%	82,00%

A Tabela 12 apresenta o percentual de páginas que possuem sites filhos. Os dados também estão agrupados por idioma e classe. Chama a atenção o fato de 100% das páginas de todos os grupos possuírem páginas filhas. Esse alto percentual já era esperado, pois conforme os dados da Tabela 10 aproximadamente todas as páginas do conjunto de dados possuem texto âncora.

Nos dados apresentados pela Tabela 12 fica evidente a relação existente entre a classe da página em análise e a classe dos sites filhos. No idioma Inglês a relação ficou próxima de 100,00%, de todos os sites filhos 99,00% eram pornográficos. Igualmente alta foi a relação para as páginas pornográficas em Português, nas quais 97,00% dos sites filhos eram pornográficos.

Alguns trabalhos anteriores já demonstraram a existência de uma relação entre a classe das páginas e suas hiperligações. Entre eles estão:

(a) Shen et al. (2006) propuseram um classificador que usava um método de obtenção de informações implícitas de ligações entre as páginas, os resultados atingidos demonstram que tais informações contribuem para a classificação;

(b) Riboni (2002) apresentou um classificador baseado em informações presentes nas

Tabela 12: Percentual de Páginas do Conjunto de Dados que Possuem Sites Filhos com Conteúdo Pornográfico

Idioma	Classe	% de Páginas que Possuem Sites Filhos	% de Páginas que Possuem Sites Filhos Pornográficos
Português	Não Pornográfica	100,00%	1,00%
Português	Pornográfica	100,00%	97,00%
Inglês	Não Pornográfica	100,00%	2,00%
Inglês	Pornográfica	100,00%	99,00%

hiperligações entre as páginas (marcador âncora);

(c) Chakrabarti et al. (2002) estudaram a estrutura de ligações entre os sítios, os dados apresentados demonstram que as páginas tendem a serem ligadas a outros páginas que possuam o mesmo tema;

(d) Menczer (2005) realizou um estudo que mostra a relação entre as ligações e o conteúdo das páginas ligadas.

Os estudos apresentados acima já demonstravam uma relação entre as páginas e suas ligações, sites pais e sites filhos. Os dados apresentados nas Tabelas 11 e 12 ajudam a comprovar tal suposição, além disso, demonstram que a relação é mais forte para determinadas classes em detrimento a outras.

5 EXPERIMENTOS E ANÁLISE DOS RESULTADOS

Neste Capítulo serão apresentados os resultados dos experimentos realizados e também uma análise desses resultados.

Todos os os experimentos foram executados em um computador com processador Intel Xeon 3.00GHz, 2GB de memória RAM, sistema operacional GNU/Linux Debian kernel 2.6.32 em uma arquitetura de 32 bits. Na execução foi utilizado o conjunto de dados descrito na sessão 4, sendo que 50% dos registros foram reservados para treinamento e os outros 50% para testes.

A sessão 5.1 apresenta os experimentos que comparam as métricas de avaliação de atributos. Na sessão 5.2 são apresentados os experimentos envolvendo testes relacionados ao número total de atributos utilizados na classificação, denominada aqui de dimensionalidade. Em seguida, a sessão 5.3 realiza uma comparação de performance entre os algoritmos de classificação. Na sequência a sessão 5.4 apresenta comparações entre os atributos estruturais presentes nas páginas *web*. Por fim, a sessão 5.5 apresenta as considerações finais a respeito dos resultados dos experimentos.

5.1 MÉTRICAS DE AVALIAÇÃO DE ATRIBUTOS

Nesta sessão serão detalhados os experimentos realizados com as métricas de avaliação de atributos. As métricas de avaliação de atributos utilizadas na seleção foram: TF, TFIDF e Ganho de informação. O funcionamento dessas métricas está detalhado na sessão 2.4.

Os experimentos foram realizados utilizando algoritmos de aprendizado de máquina de diferentes paradigmas; ao todo cinco algoritmos foram testados. A quantidade de atributos textuais selecionados, ou dimensionalidade, variou de 10 atributos até 200 atributos. Na seleção foi utilizado apenas o conteúdo textual das páginas HTML e não foram utilizados os textos presentes dentro dos marcadores da linguagem.

A Tabela 13 contém os resultados dos experimentos no conjunto de dados em Português. Na primeira coluna da tabela consta a informação sobre a dimensionalidade utilizada,

as demais colunas apresentam o resultado da Medida-F para cada uma das métricas de avaliação de atributos usando os diferentes classificadores.

Tabela 13: Medida-F na Base em Português: Experimento com Métricas de Avaliação de Atributos

Qtde	TF					TFIDF					Ganho de Informação				
Atributos	C4.5	KNN	MLP	Naive	SMO	C4.5	KNN	MLP	Naive	SMO	C4.5	KNN	MLP	Naive	SMO
10	0.9439	0.9559	0.9459	0.9189	0.9379	0.93	0.9289	0.9299	0.9179	0.9148	0.9709	0.9829	0.9779	0.9709	0.9739
20	0.9489	0.9799	0.9689	0.9439	0.9559	0.9459	0.9749	0.9709	0.9499	0.9509	**0.9739**	0.9859	0.9829	0.9579	0.9799
30	0.9519	0.9769	0.9699	0.9579	0.974	0.9479	0.9799	0.9759	0.9559	0.9569	0.9659	0.9879	0.9869	0.9639	0.9869
40	0.9549	0.9769	0.9799	0.9649	0.9779	0.9509	0.9789	0.9719	0.9649	0.9719	0.9609	0.9869	0.9849	0.9689	0.9859
50	0.9579	**0.9909**	0.9829	0.9679	0.9799	0.9499	**0.9879**	0.9719	0.9599	0.9759	0.9529	0.9849	0.9849	0.9669	0.9869
60	0.9689	0.9899	0.9819	0.9699	0.9749	**0.9549**	0.9869	0.9809	0.9649	0.9769	0.9689	**0.9909**	**0.9899**	0.9689	**0.9939**
70	0.9499	0.9879	0.9849	0.9749	0.9869	**0.9549**	0.9799	0.9799	0.9659	0.9839	0.9529	**0.9909**	0.9859	0.9679	0.9919
80	0.9499	0.9889	0.9849	0.9749	0.9849	**0.9549**	0.9819	0.9839	0.9719	0.9839	0.9599	0.9899	0.9849	0.9759	0.9869
90	0.9599	0.9869	0.9889	0.9809	0.9909	**0.9549**	0.9759	0.9799	0.9769	0.9839	0.9599	**0.9909**	0.9839	0.9779	0.9879
100	0.9599	0.9819	0.9869	0.9799	0.9889	**0.9549**	0.9749	0.9789	0.9759	0.9779	0.9599	0.9899	0.9839	0.9779	0.9849
110	**0.9699**	0.9869	0.9939	0.9839	0.9959	**0.9549**	0.9789	0.9849	0.9849	0.9789	0.9599	0.9849	0.9859	0.9799	0.9839
120	0.9679	0.9869	0.9939	0.9859	**0.9979**	0.9529	0.9749	0.9839	0.9829	0.9849	0.9599	0.9879	0.9869	0.9789	0.9829
130	0.9679	0.9859	0.9969	0.9829	**0.9979**	0.9479	0.9739	0.9819	0.9849	0.9799	0.9599	0.9859	0.9849	0.9789	0.9879
140	0.9599	0.9859	0.9969	0.9849	0.9969	0.9479	0.9789	0.9839	0.9849	0.9819	0.9609	0.9829	0.9879	0.9829	0.9879
150	0.9669	0.9869	**0.9979**	0.9839	0.996	0.9509	0.9829	0.9859	**0.9869**	0.9849	0.9609	0.9809	0.9859	0.9829	0.9869
160	0.9669	0.9859	0.9959	0.9839	0.996	0.9509	0.9809	0.9859	0.9839	0.9859	0.9609	0.9789	0.9859	0.9849	0.9879
170	0.9669	0.9849	0.9959	0.9829	0.996	0.9509	0.9829	0.9819	0.9829	0.9859	0.9609	0.9809	0.9869	0.9859	0.9879
180	0.9569	0.9879	0.9959	0.9849	0.996	0.9509	0.9819	0.9859	0.9849	**0.9869**	0.9609	0.9839	0.9859	0.9869	0.9879
190	0.9569	0.9869	0.9949	**0.9869**	0.996	0.9509	0.9809	**0.9869**	0.9849	0.9839	0.9619	0.9839	0.9879	**0.9879**	0.9889
200	0.9559	0.9859	0.9939	0.9849	0.9969	0.9509	0.9819	0.9859	**0.9869**	0.9859	0.9619	0.9849	**0.9899**	**0.9879**	0.9899

A Tabela 14 mostra os resultados dos experimentos no conjunto de dados em Inglês. A sequência das informações apresentadas é a mesma da utilizada na Tabela 13.

Tabela 14: Medida-F na Base em Inglês: Experimento com Métricas de Avaliação de Atributos

Qtde	TF					TFIDF					Ganho de Informação				
Atributos	C4.5	KNN	MLP	Naive	SMO	C4.5	KNN	MLP	Naive	SMO	C4.5	KNN	MLP	Naive	SMO
10	0.8764	0.8837	0.8898	0.8371	0.8622	0.8555	0.8576	0.8650	0.8556	0.8660	0.8745	0.8937	0.9044	0.8992	0.8943
20	0.8928	0.8998	0.8939	0.8660	0.8803	0.8800	0.8814	0.9015	0.8809	0.8951	0.8863	0.9185	0.9216	0.8981	0.9228
30	0.9268	0.9329	0.9379	0.8900	0.9197	0.8828	0.8849	0.9089	0.8830	0.9136	0.9004	0.9125	0.9268	0.9042	0.9146
40	0.9279	0.9409	0.9479	0.8962	0.9278	0.8982	0.8949	0.9128	0.8930	0.9257	0.8972	0.9105	0.9218	0.8991	0.9207
50	0.9399	0.9539	0.9479	0.9043	0.9368	0.9168	0.9369	0.9529	0.9001	0.9308	0.9207	0.9439	0.9409	0.8990	0.9449
60	0.9259	0.9669	0.9579	0.9103	0.9359	0.9229	0.9519	0.9589	0.9012	0.9308	0.9359	0.9469	0.9489	0.8969	0.9489
70	0.9429	0.9679	0.9539	0.9083	0.9329	0.9188	0.9389	0.9669	0.9083	0.9338	0.9359	0.9589	0.9489	0.8980	0.9549
80	0.9389	0.9719	0.9589	0.9073	0.9389	0.9259	0.9469	0.9629	0.9042	0.9308	0.9299	0.9579	0.9499	0.8969	0.9479
90	0.9258	0.9749	0.9659	0.9073	0.9509	0.9289	0.9559	0.9609	0.9032	0.9418	0.9319	0.9669	0.9549	0.8980	0.9559
100	0.9389	0.9789	0.9689	0.9093	0.9539	0.9329	0.9449	0.9469	0.9073	0.9469	0.9429	0.9669	0.9649	0.8990	0.9619
110	0.9509	0.9769	0.9689	0.9042	0.9529	0.9319	0.9569	0.9529	0.9113	0.9519	0.9429	0.9549	0.9539	0.9000	0.9579
120	0.9529	0.9789	0.9729	0.9042	0.9649	0.9269	0.9689	0.9669	0.9144	0.9559	0.9429	0.9629	0.9519	0.9000	0.9599
130	0.9519	0.9789	0.9669	0.9093	0.9609	0.9239	0.9709	0.9639	0.9164	0.9589	0.9469	**0.9799**	0.9659	0.9021	0.9689
140	0.9429	0.9789	0.9689	0.9114	0.9599	0.9219	0.9689	0.9609	0.9175	0.9559	**0.9639**	0.9789	0.9739	0.9031	0.9749
150	0.9329	0.9809	0.9689	0.9114	0.9629	0.9249	0.9689	0.9629	0.9175	0.9579	**0.9639**	**0.9799**	0.9769	0.9021	0.9719
160	0.9349	**0.9829**	0.9649	0.9134	0.9659	0.9289	0.9709	0.9609	0.9144	0.9509	0.9559	0.9789	0.9789	0.9041	0.9769
170	0.9349	0.9779	0.9659	0.9144	0.9619	0.9359	0.9719	0.9639	0.9144	0.9489	0.9569	0.9739	0.9779	0.9011	0.9779
180	**0.9549**	0.9779	0.9709	0.9134	0.9629	**0.9469**	0.9749	0.9639	0.9174	0.9619	0.9579	0.9749	0.9809	**0.9052**	0.9809
190	0.9459	0.9789	**0.9809**	**0.9154**	**0.9679**	0.9459	0.9769	0.9699	0.9174	0.9649	0.9579	0.9789	**0.9839**	0.9042	0.9809
200	0.9459	0.9769	**0.9809**	0.9144	0.9639	0.9419	**0.9779**	**0.9729**	**0.9194**	**0.9659**	0.9579	0.9789	0.982	0.9031	**0.9819**

A Figura 13 apresenta um gráfico comparando os resultados das três métricas de avaliação de atributos utilizando o classificador baseado em árvore de decisão C4.5, a versão disponível no Weka para o algoritmo C4.5. Conforme o gráfico é possível constatar que no

geral os melhores resultados foram obtidos pela métrica TF. A métrica Ganho de Informação também apresentou bons resultados, principalmente no idioma Português.

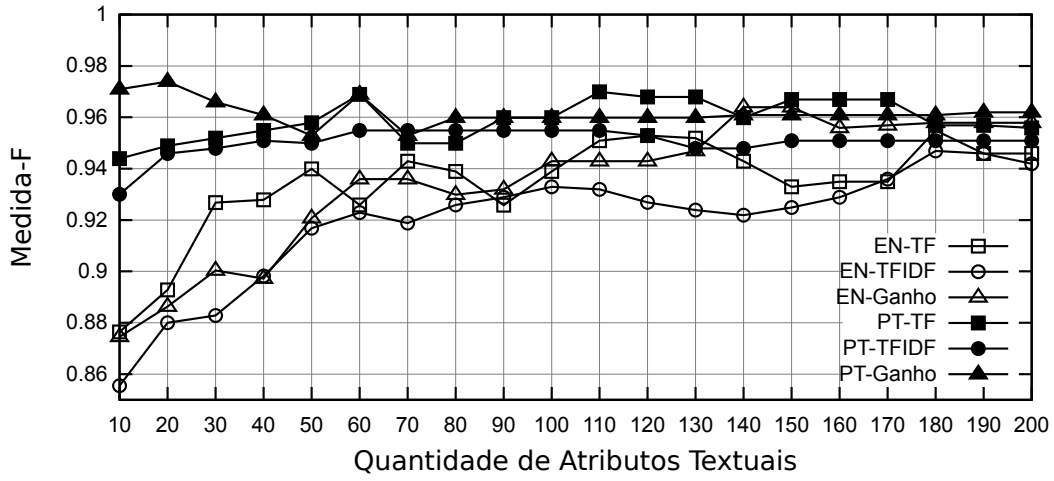

Figura 13: Desempenho do Classificador C4.5 Utilizando Diferentes Métricas na Seleção de até 200 Atributos

Os resultados das métricas de avaliação de atributos usando o classificador KNN é mostrado pelo gráfico presente na Figura 14. O desempenho das três métricas ficou muito próximo quando a quantidade de atributos era superior a 130, principalmente no idioma Português. No geral o desempenho da métrica TF foi superior, sendo mais visível essa diferença no idioma Inglês

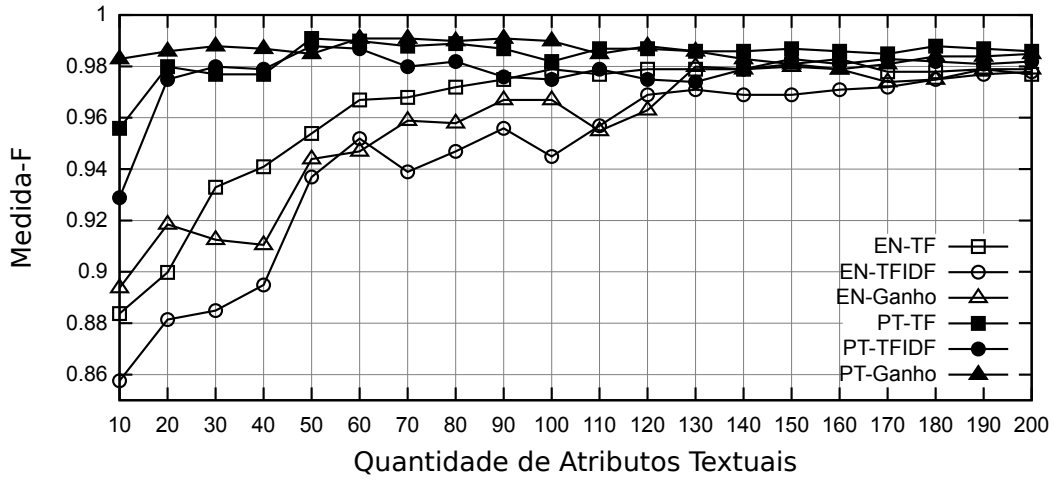

Figura 14: Desempenho do Classificador KNN Utilizando Diferentes Métricas na Seleção de até 200 Atributos

O desempenho das métricas utilizando o classificador MLP é apresentado na Figura 15. Para o idioma Português os melhores resultados foram atingidos usando a métrica TF, o mesmo não se repetiu com o idioma Inglês o qual obteve os melhores resultados com a métrica Ganho de Informação.

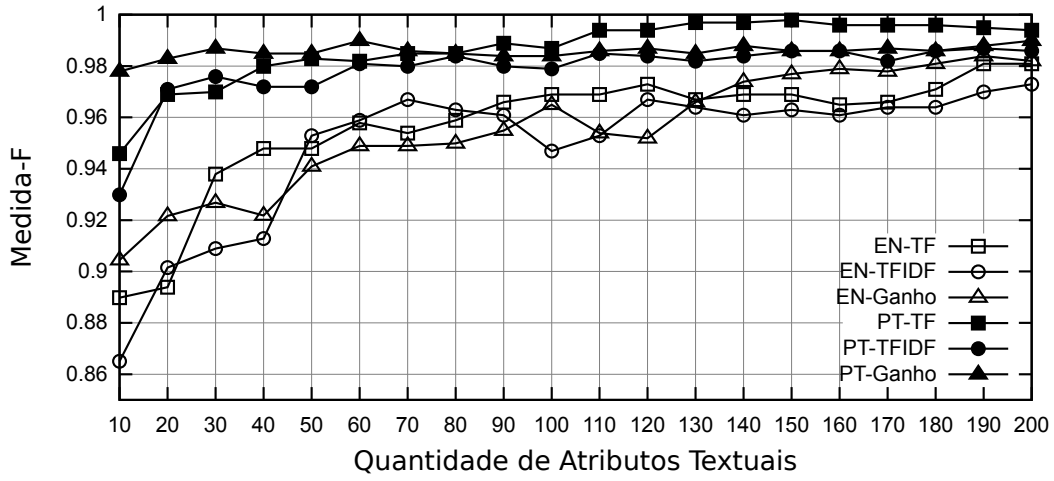

Figura 15: Desempenho do Classificador MLP Utilizando Diferentes Métricas na Seleção de até 200 Atributos

A Figura 16 mostra um gráfico comparando os resultados das métricas utilizando o classificador Naïve Bayes. As diferença dos resultados das três métricas no idioma Português foram pequenas, algumas vezes imperceptíveis visualmente. Já no conjunto de dados em Inglês fica evidente que a TFIDF obteve os melhores resultados, também fica evidente que o Ganho de Informação apresentou os piores resultados.

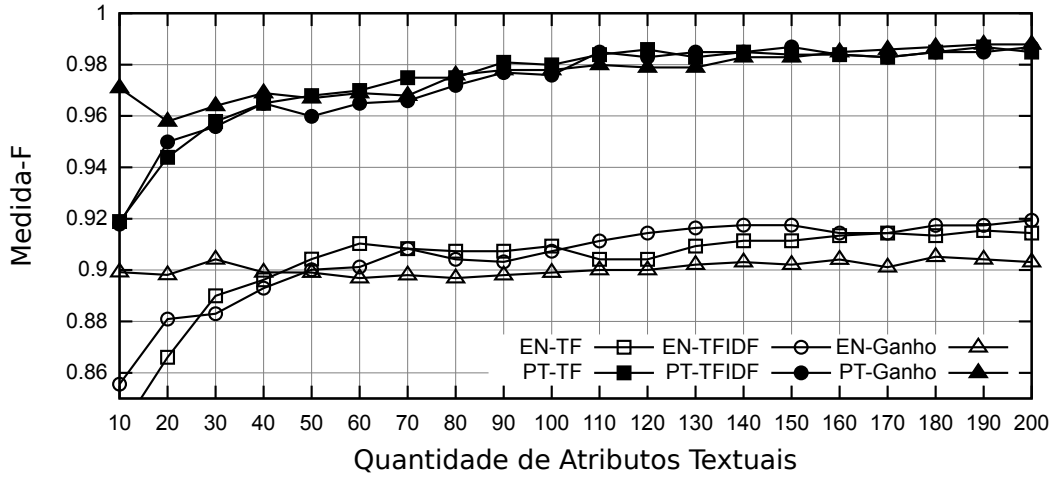

Figura 16: Desempenho do Classificador Naïve Bayes Utilizando Diferentes Métricas na Seleção de até 200 Atributos

Por último, a Figura 17 apresenta um gráfico de desempenho das métricas de avaliação de atributos usando o classificador SMO (implementação disponível no Weka para o SVM). O desempenho das métricas usando o classificador SMO foi bem semelhante ao desempenho utilizando o classificador MLP. Os melhores desempenhos para Português foram obtidos com a métrica TF, enquanto no Inglês os melhores resultados foram obtidos com a métrica Ganho de Informação.

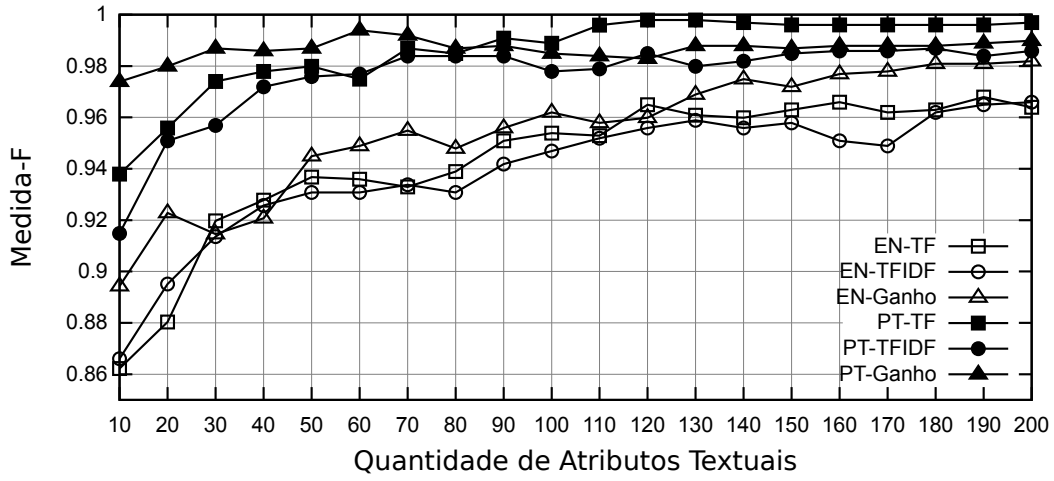

Figura 17: Desempenho do Classificador SMO Utilizando Diferentes Métricas na Seleção de até 200 Atributos

De acordo com os dados apresentados nas Tabelas 13 e 14 alguns pontos podem ser destacados:

(a) A TFIDF apresentou os piores resultados comparado com as demais métricas, sendo melhor apenas quando foi utilizado o algoritmo Naïve Bayes no conjunto de dados em Inglês.

(b) No geral a métrica TF obteve os melhores resultados nos experimentos realizados. O Ganho de Informação também apresentou bons resultados, alguns deles próximos aos atingidos pela TF.

(c) Usando os classificadores MLP e SMO no conjunto de dados em Inglês os melhores resultados foram atingidos pela métrica Ganho de Informação, porém conforme gráficos dos respectivos classificadores é possível verificar uma tendência de aumento da performance da métrica TF conforme a dimensionalidade aumenta.

(d) Em algumas situações os resultados dos classificadores usando diferentes métricas foi imperceptível, por exemplo: no caso do classificador Naïve Bayes no idioma Português.

5.2 ANÁLISE RELACIONADA À DIMENSIONALIDADE

Nesta sessão serão considerados experimentos que levam em conta uma dimensionalidade - número de atributos - maior do que o considerado na sessão anterior. Nos experimentos desta sessão será utilizada uma dimensionalidade variando de 100 atributos até 1000 atributos.

O objetivo desses experimentos é verificar quais serão os resultados dos classificadores

e das métricas de avaliação de atributos em uma dimensionalidade maior, visando identificar qual é a dimensionalidade ideal para cada um dos classificadores nesse gênero de classificação textual.

Nos testes foram analisados três métricas de avaliação de atributos - TF, TFIDF e Ganho de Informação - e cinco algoritmos de classificação. Os atributos textuais selecionados pertenciam apenas ao conteúdo textual das páginas HTML, não foram utilizadas as informações presentes nos marcadores da linguagem.

A Tabela 15 contém os resultados dos experimentos no conjunto de dados em Inglês. Na primeira coluna da tabela consta a informação da dimensionalidade utilizada no experimento, as demais colunas apresentam o resultado da Medida-F para cada uma das métricas de avaliação de atributos utiliza pelos respectivos classificadores.

Tabela 15: Medida-F na Base em Inglês: Experimento Envolvendo Dimensionalidade

Qtde	TF					TFIDF					Ganho de Informação				
Atributos	C4.5	KNN	MLP	Naive	SMO	C4.5	KNN	MLP	Naive	SMO	C4.5	KNN	MLP	Naive	SMO
100	0.9389	0.9789	0.9689	0.9093	0.9539	0.9329	0.9449	0.9469	0.9073	0.9469	0.9429	0.9669	0.9649	0.8990	0.9619
200	0.9459	0.9769	0.9809	0.9144	0.9639	0.9419	0.9779	0.9729	0.9194	0.9659	0.9579	0.9789	0.982	0.9031	0.9819
300	**0.9489**	0.9789	0.9849	0.9245	0.9819	0.9489	0.9749	0.9839	0.9306	0.9809	0.9629	**0.9859**	0.9849	0.9154	**0.9839**
400	**0.9489**	**0.9819**	0.9889	0.9397	0.9879	**0.9569**	0.9739	0.9899	0.9407	0.9839	**0.9639**	0.9849	0.9739	0.9144	0.9769
500	0.9429	0.9799	0.9909	0.9418	0.9859	0.9549	0.9759	0.9909	0.9367	0.9849	0.9629	0.9839	0.9799	0.9174	0.978
600	0.9429	0.9789	0.9879	0.9539	0.9879	0.9509	**0.9809**	**0.9919**	0.9428	0.9869	**0.9639**	0.9769	0.9799	0.9235	0.9789
700	**0.9489**	0.9789	0.992	**0.9549**	0.9849	0.9489	0.9799	0.9869	0.9468	0.9869	**0.9639**	0.9769	0.9839	0.9256	0.9799
800	0.9449	0.9769	0.9899	0.9519	0.9859	0.9389	0.9789	0.9879	0.9488	**0.9879**	**0.9639**	0.9689	**0.9879**	0.9347	0.9809
900	0.9449	0.9769	**0.9949**	0.9539	0.9879	0.9389	0.9759	0.9909	0.9469	0.9869	**0.9639**	0.9679	0.9869	**0.9448**	0.9809
1000	0.9449	0.9749	0.9879	0.9529	**0.9889**	0.9429	0.9749	0.9909	**0.9489**	**0.9879**	0.9629	0.9649	0.9849	0.9438	**0.9839**

A Tabela 16 apresenta os resultados dos experimentos no conjunto de dados em Português. A sequência das informações apresentadas é a mesma da Tabela 15, sendo a primeira coluna contendo a dimensionalidade e as demais contendo os resultados para cada métrica de avaliação de atributos.

Tabela 16: Medida-F na Base em Português: Experimento Envolvendo Dimensionalidade

Qtde	TF					TFIDF					Ganho de Informação				
Atributos	C4.5	KNN	MLP	Naive	SMO	C4.5	KNN	MLP	Naive	SMO	C4.5	KNN	MLP	Naive	SMO
100	**0.9599**	0.9819	0.9869	0.9799	0.9889	0.9549	0.9749	0.9789	0.9759	0.9779	0.9599	**0.9899**	0.9839	0.9779	0.9849
200	0.9559	**0.9859**	0.9939	0.9849	0.9969	0.9509	0.9819	0.9859	0.9869	0.9859	0.9619	0.9849	0.9899	0.9879	0.9899
300	0.9559	**0.9859**	0.9949	0.9889	**0.9989**	0.962	**0.9839**	0.9929	0.9879	**0.9989**	**0.9629**	0.9809	**0.9979**	**0.9919**	**0.9949**
400	0.9559	0.9819	0.998	0.9869	0.9979	**0.9679**	0.9819	0.9959	0.986	0.998	0.9609	0.9779	0.9969	**0.9919**	0.9939
500	0.9559	0.9769	0.9969	0.9889	0.9969	0.9559	0.9799	**0.9969**	0.9879	0.998	0.9579	0.9789	0.9969	0.9899	0.9939
600	0.9539	0.9769	0.9969	**0.9899**	0.998	0.9559	0.9789	**0.9969**	0.9879	0.998	0.9579	0.9799	0.998	0.988	0.9929
700	0.9539	0.9759	**0.9989**	0.9889	0.998	0.9539	0.9769	0.998	**0.9899**	0.9959	0.9579	0.9759	0.9969	0.9869	0.9929
800	0.9539	0.9789	**0.9989**	0.9879	**0.9989**	0.9539	0.9729	0.998	0.9889	0.998	0.9549	0.9739	0.9969	0.9869	0.9919
900	0.9569	0.9719	0.998	0.9879	0.998	0.9539	0.9709	0.998	0.9869	0.9969	0.9549	0.9719	0.998	0.988	0.9939
1000	0.9589	0.9689	0.998	0.9889	0.998	0.9559	0.9679	0.9949	0.9869	0.9959	0.9549	0.9709	0.9969	0.9869	0.9929

A Figura 18 apresenta um gráfico comparando o desempenho do classificador C4.5 em

diferentes dimensionalidades. Conforme o gráfico é possível verificar que no geral ocorre uma tendência de estabilização dos resultados. Tanto para o idioma Português quanto para o Inglês os melhores índices foram obtidos com a métrica Ganho de Informação.

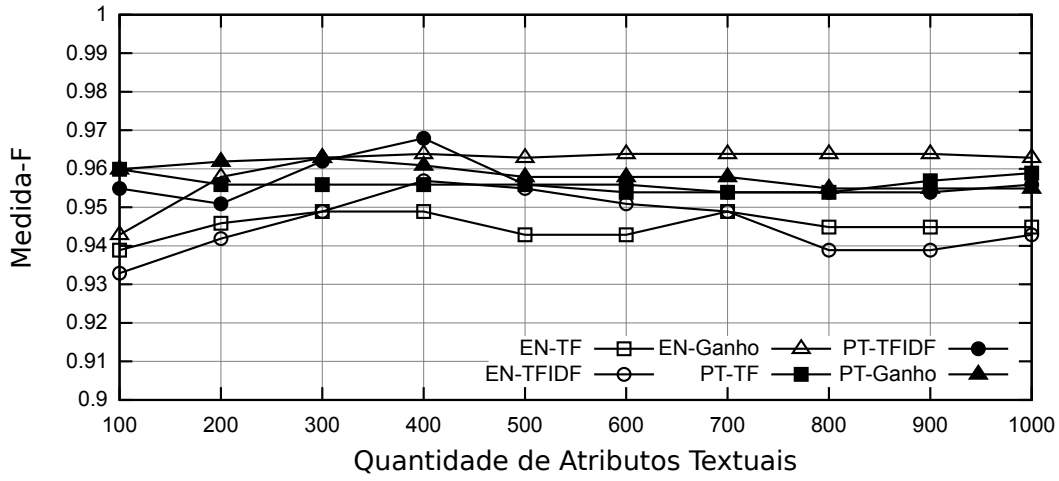

Figura 18: Desempenho do Classificador C4.5 Utilizando Diferentes Métricas na Seleção de até 1000 Atributos

Os resultados com diferentes dimensionalidades do classificador KNN são mostrados pelo gráfico presente na Figura 19. O classificador apresentou uma tendência de queda dos resultados, para os dois idiomas, conforme o número de atributos aumentava. O resultado das métricas foram próximas, em geral bons resultados foram atingidos entre 200 a 300 atributos.

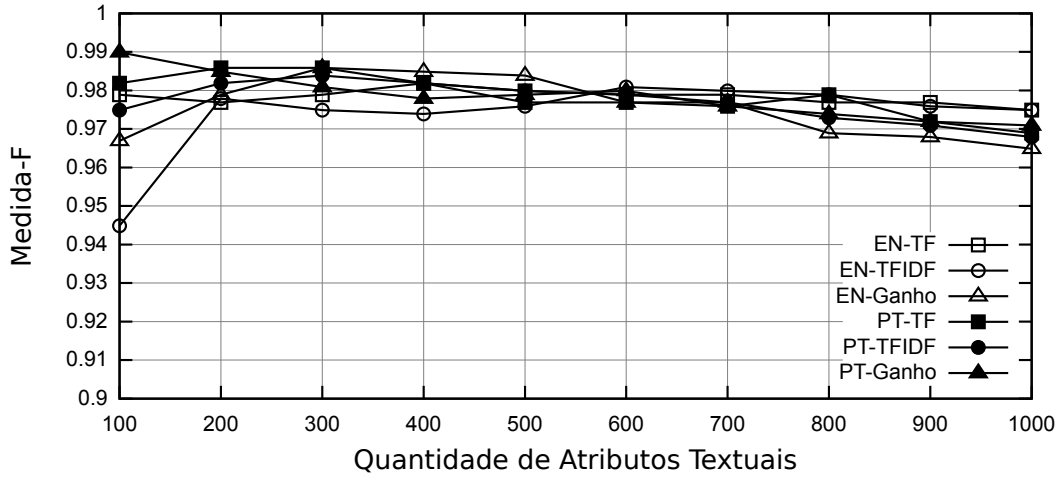

Figura 19: Desempenho do Classificador KNN Utilizando Diferentes Métricas na Seleção de até 1000 Atributos

O desempenho com diferentes quantidades de atributos utilizando o classificador MLP é apresentado na Figura 20. Conforme o gráfico é possível constatar uma tendência de aumento da Medida-F conforme aumentava a quantidade de atributos. Os resultados atingidos pelas

diferentes métricas foram bem próximos, principalmente para o idioma Português. Acima de 400 atributos já foi possível atingir uma Medida-F superior a 0.99 para ambos os idiomas.

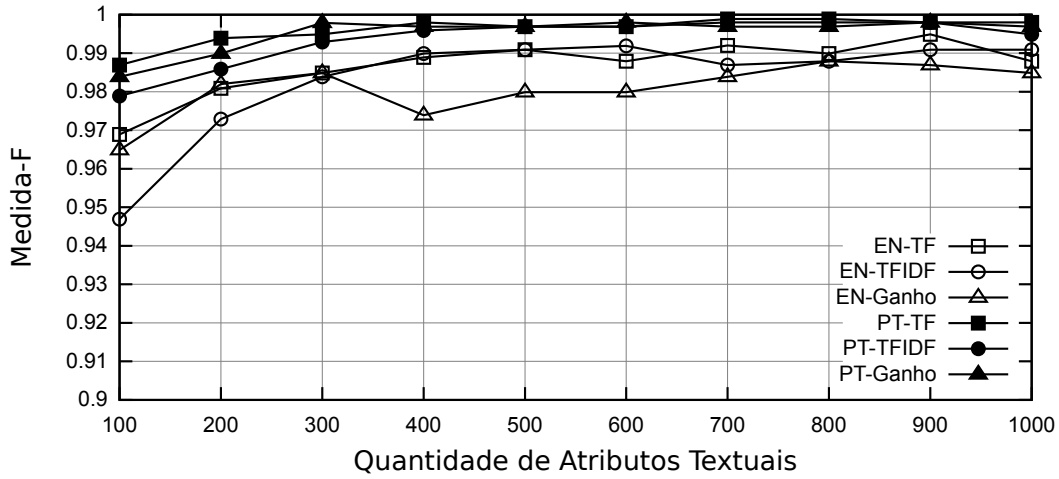

Figura 20: Desempenho do Classificador MLP Utilizando Diferentes Métricas na Seleção de até 1000 Atributos

A Figura 21 mostra um gráfico comparando os resultados de diferentes quantidades de atributos usando o classificador Naïve Bayes. Para o idioma Português os melhores resultados foram atingidos com uma quantidade entre 300 e 400 atributos, após atingido o ápice os resultados ficaram oscilando. Já para o idioma Inglês ocorreu uma tendência inicial de aumento da Medida-F e logo após uma estabilização dos resultados, nesse idioma os melhores resultados foram apresentados com a métrica TF.

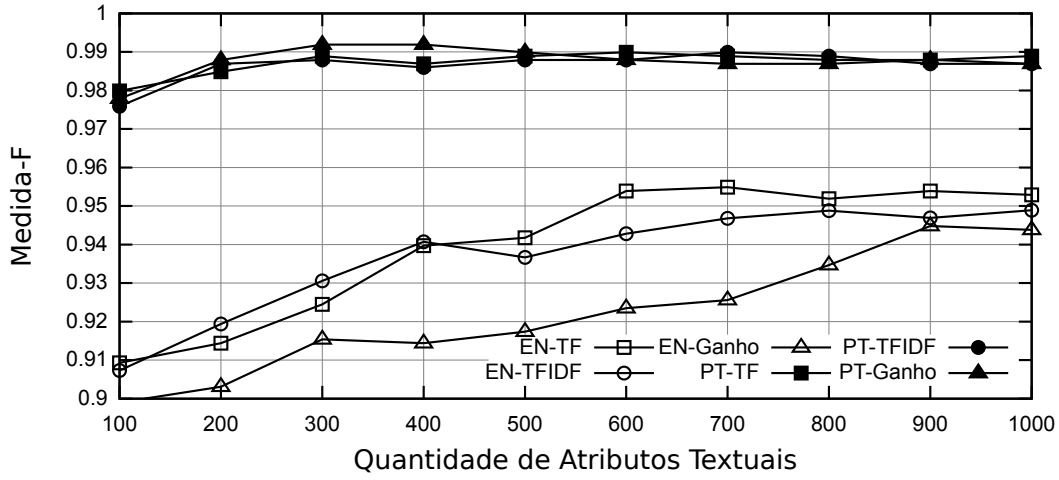

Figura 21: Desempenho do Classificador Naïve Bayes Utilizando Diferentes Métricas na Seleção de até 1000 Atributos

Por último, a Figura 22 apresenta um gráfico de desempenho de diferentes dimensionalidades usando o classificador SMO. Nesse classificador as métricas de avaliação de atributos

apresentaram resultados bem similares, sendo que inicialmente ocorreu uma tendência de aumento da Medida-F e logo após uma estabilização. O melhor resultado no idioma Português foi atingido com 300 atributos, enquanto para o idioma Inglês a Medida-F mais alta foi atingida com 1000 atributos.

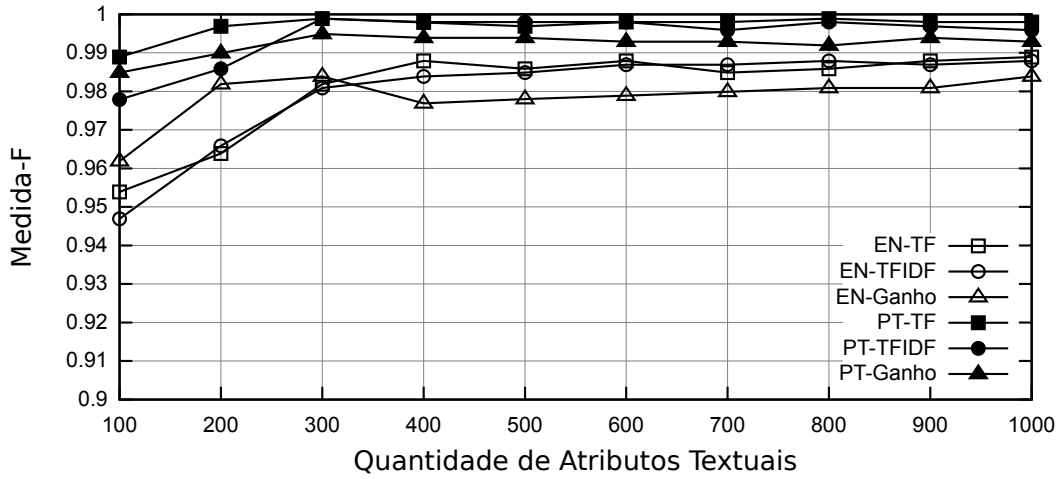

Figura 22: Desempenho do Classificador SMO Utilizando Diferentes Métricas na Seleção de até 1000 Atributos

Com base nos dados apresentados nas Tabelas 15 e 16 pode-se constatar os seguintes fatos:

(a) O classificador KNN apresentou uma tendência de diminuição da Medida-F conforme a dimensionalidade aumentava, o que demonstra que existe uma dimensionalidade ideal para esse classificador, após determinado ponto os resultados tendem a piorar ao invés de melhorar. Esse tipo de comportamento já era esperado para o classificador KNN devido a maneira que ocorre o seu aprendizado.

(b) Os demais classificados apresentaram uma tendência de aumento inicial da Medida-F de acordo com o aumento da quantidade de atributos, em seguida apresentaram uma tendência de estabilização dos resultados. Assim, aumentar a quantidade de atributos depois de certo ponto mostrou-se ineficaz no que diz respeito a melhora nos resultados do classificador.

(c) Com o aumento da dimensionalidade as diferentes métricas de avaliação de atributos apresentaram resultados bem similares, o que demonstra que a métrica de avaliação apresenta uma influência maior nos resultados quando a dimensionalidade é menor. Um dos motivos que pode explicar esse fato é que quando a dimensionalidade é alta existe uma probabilidade maior que mesmo diferentes métricas acabem por escolherem os mesmos atributos.

(d) Os resultados apresentados pelos classificadores SMO e MLP foram bem similares, ambos demonstram a mesma tendência de acordo com a quantidade de atributos.

5.3 DESEMPENHO DOS CLASSIFICADORES

Esta sessão apresentará os resultados dos experimentos que comparam o desempenho dos classificadores. Os classificadores selecionados para os experimentos foram:

(a) O probabilístico Naïve Bayes.

(b) O algoritmo baseado em instâncias KNN, com K=1.

(c) O classificador baseado em árvores de decisão C4.5.

(d) O classificador SVM, utilizando a implementação SMO (*Sequential Minimal Optimization*) empregando uma função de núcleo (kernel) polinomial.

(e) O algoritmo de redes neurais MLP (*Multilayer Perceptron*).

O primeiro experimento realizado visava identificar o tempo de treinamento de cada um dos algoritmos de classificação. Nesse experimento foi utilizado apenas a base em Inglês usando somente a informação contida no marcador <title> da linguagem HTML. A métrica usada no processo de seleção de atributos foi o Ganho de Informação, com dimensionalidade de 10 até 1000 atributos.

A Tabela 17 apresenta os resultados desse primeiro experimento. A primeira coluna mostra o total de atributos utilizados e as demais os tempos de treinamento de cada um dos respectivos classificadores.

A Figura 23 mostra um gráfico com os tempos de treinamento de cada classificador. Conforme os dados desse gráfico e também da Tabela 17 é possível constatar que os algoritmos C4.5, KNN, Naïve Bayes e SMO possuem um tempo de treinamento de poucos segundos enquanto o classificador MLP pode levar algumas horas na fase de treinamento, dependendo da quantidade de atributos. Esses resultados são importantes para mensurar o tempo necessário para a realização de testes futuros com os algoritmos.

O segundo experimento realizado visava comparar a eficiência na classificação de cada um dos algoritmos de aprendizagem de máquina. Para medir essa eficiência da classificação foi utilizada a Medida-F. Nos testes foram utilizados apenas o conteúdo textual das páginas *web*, não foi utilizado o conteúdo presente nos marcadores do HTML.

Os resultados desses experimentos constam nas Tabelas 13 e 14 da sessão 5.1. Os gráficos desses resultados serão apresentados nesta sessão comparando o desempenho individual de cada classificador.

A Figura 24 mostra o desempenho dos classificadores utilizando a métrica Ganho de

Tabela 17: Tempo de Treinamento dos Classificadores

Qtde Atributos	C4.5	KNN	MLP	Naïve Bayes	SMO
10	0h 0m 0s 64ms	0h 0m 0s 548ms	0h 0m 2s 942ms	0h 0m 0s 60ms	0h 0m 0s 113ms
20	0h 0m 0s 67ms	0h 0m 0s 825ms	0h 0m 6s 698ms	0h 0m 0s 56ms	0h 0m 0s 87ms
30	0h 0m 0s 84ms	0h 0m 1s 4ms	0h 0m 12s 6ms	0h 0m 0s 80ms	0h 0m 0s 153ms
40	0h 0m 0s 111ms	0h 0m 1s 214ms	0h 0m 19s 4ms	0h 0m 0s 106ms	0h 0m 0s 258ms
50	0h 0m 0s 138ms	0h 0m 1s 388ms	0h 0m 27s 82ms	0h 0m 0s 133ms	0h 0m 0s 249ms
60	0h 0m 0s 165ms	0h 0m 1s 606ms	0h 0m 36s 851ms	0h 0m 0s 158ms	0h 0m 0s 255ms
70	0h 0m 0s 190ms	0h 0m 1s 767ms	0h 0m 48s 433ms	0h 0m 0s 189ms	0h 0m 0s 275ms
80	0h 0m 0s 193ms	0h 0m 1s 958ms	0h 1m 2s 468ms	0h 0m 0s 212ms	0h 0m 0s 431ms
90	0h 0m 0s 415ms	0h 0m 2s 322ms	0h 1m 17s 512ms	0h 0m 0s 253ms	0h 0m 0s 423ms
100	0h 0m 0s 263ms	0h 0m 2s 368ms	0h 1m 33s 401ms	0h 0m 0s 266ms	0h 0m 0s 468ms
110	0h 0m 0s 287ms	0h 0m 2s 610ms	0h 1m 51s 347ms	0h 0m 0s 291ms	0h 0m 0s 445ms
120	0h 0m 0s 314ms	0h 0m 2s 822ms	0h 2m 11s 743ms	0h 0m 0s 312ms	0h 0m 0s 485ms
130	0h 0m 0s 341ms	0h 0m 2s 995ms	0h 2m 34s 176ms	0h 0m 0s 344ms	0h 0m 0s 495ms
140	0h 0m 0s 371ms	0h 0m 3s 206ms	0h 2m 58s 747ms	0h 0m 0s 369ms	0h 0m 0s 480ms
150	0h 0m 0s 395ms	0h 0m 3s 389ms	0h 3m 26s 306ms	0h 0m 0s 397ms	0h 0m 0s 603ms
160	0h 0m 0s 424ms	0h 0m 3s 574ms	0h 3m 58s 925ms	0h 0m 0s 425ms	0h 0m 0s 561ms
170	0h 0m 0s 449ms	0h 0m 3s 765ms	0h 4m 39s 568ms	0h 0m 0s 448ms	0h 0m 0s 568ms
180	0h 0m 0s 478ms	0h 0m 3s 898ms	0h 5m 29s 491ms	0h 0m 0s 478ms	0h 0m 0s 592ms
190	0h 0m 0s 507ms	0h 0m 4s 160ms	0h 6m 1s 451ms	0h 0m 0s 502ms	0h 0m 0s 754ms
200	0h 0m 0s 522ms	0h 0m 4s 336ms	0h 6m 41s 559ms	0h 0m 0s 585ms	0h 0m 0s 733ms
300	0h 0m 0s 819ms	0h 0m 6s 554ms	0h 18m 15s 241ms	0h 0m 0s 810ms	0h 0m 1s 30ms
400	0h 0m 1s 111ms	0h 0m 8s 359ms	0h 32m 7s 644ms	0h 0m 1s 72ms	0h 0m 1s 358ms
500	0h 0m 1s 411ms	0h 0m 10s 13ms	1h 3m 0s 869ms	0h 0m 1s 335ms	0h 0m 1s 860ms
600	0h 0m 2s 739ms	0h 0m 11s 282ms	1h 13m 8s 542ms	0h 0m 1s 696ms	0h 0m 2s 198ms
700	0h 0m 4s 358ms	0h 0m 11s 898ms	1h 54m 45s 466ms	0h 0m 2s 112ms	0h 0m 2s 482ms
800	0h 0m 3s 63ms	0h 0m 12s 743ms	2h 55m 35s 672ms	0h 0m 2s 319ms	0h 0m 2s 688ms
900	0h 0m 3s 106ms	0h 0m 13s 396ms	2h 43m 42s 228ms	0h 0m 2s 533ms	0h 0m 3s 197ms
1000	0h 0m 4s 18ms	0h 0m 13s 955ms	3h 23m 14s 642ms	0h 0m 3s 132ms	0h 0m 3s 430ms
Tempo total	0h 0m 26s 403ms	0h 2m 17s 955ms	14h 50m 10s 14ms	0h 0m 20s 673ms	0h 0m 26s 671ms

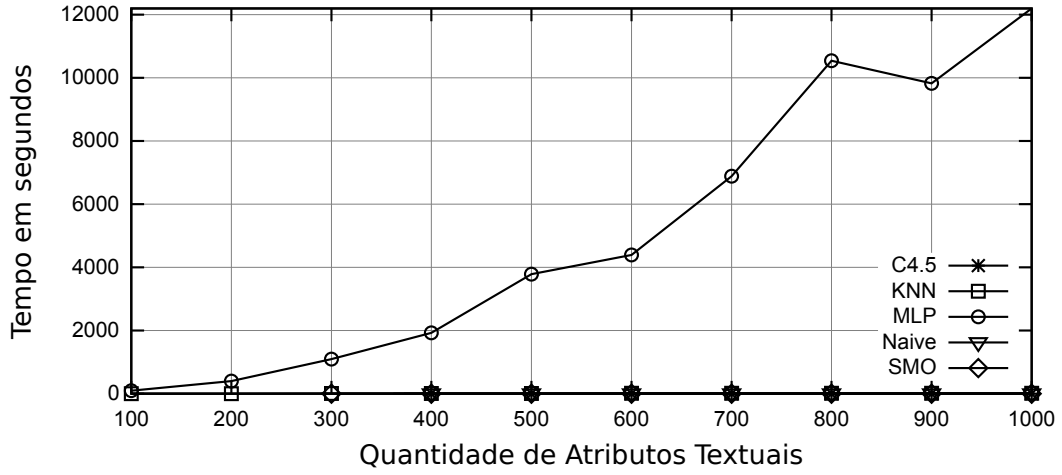

Figura 23: Tempo de Treinamento dos Classificadores

Informação no idioma Inglês. Conforme o gráfico, é possível verificar que os classificadores C4.5 e Naïve Bayes tiveram os piores resultados, principalmente o classificador Naïve Bayes, os demais classificadores apresentaram resultados bem similares.

O desempenho dos classificadores usando a métrica de avaliação de atributos TF no conjunto de dados em Inglês é mostrado na Figura 25. De acordo com o gráfico os piores resultados foram obtidos pelo classificador C4.5 e Naïve Bayes. Os melhores resultados foram

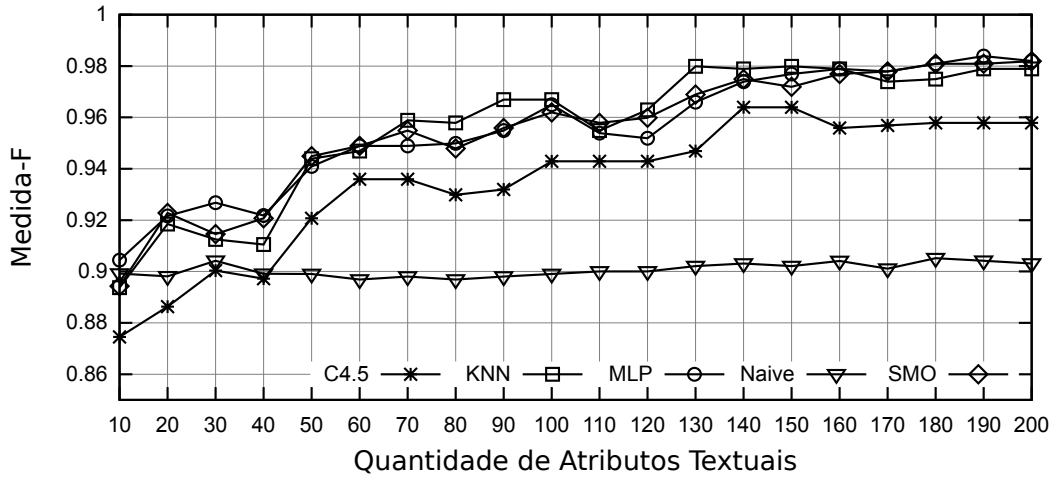

Figura 24: Desempenho dos Classificadores Usando a Métrica Ganho de Informação no Conjunto de Dados em Inglês

conquistados pelo classificador KNN, em seguida MLP e SMO.

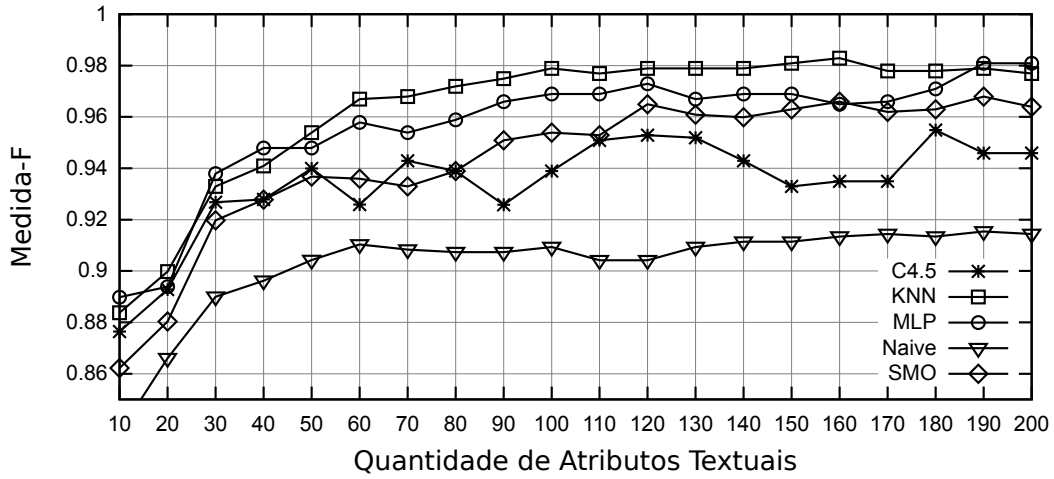

Figura 25: Desempenho dos Classificadores Usando a Métrica TF no Conjunto de Dados em Inglês

A Figura 26 apresenta o resultado dos classificadores utilizando atributos selecionados pela métrica TFIDF nas páginas em Inglês. Nesse experimento o classificador Naïve Bayes obteve o pior resultado, da mesma forma o C4.5 também não obteve bons resultados. Já o KNN atingiu os melhores índices, enquanto MLP e SMO apresentaram resultados similares.

O resultado dos classificadores usando os atributos selecionados pela métrica Ganho de Informação na base em Português é mostrado na Figura 27. Nesse experimento todos os algoritmos de aprendizagem de máquina atingiram resultados similares, com exceção do C4.5 que ficou abaixo dos resultados atingidos pelos demais.

O desempenho dos algoritmos de classificação na base em Português utilizando a métrica de avaliação de atributos TF é apresentado na Figura 28. Conforme o gráfico o classifi-

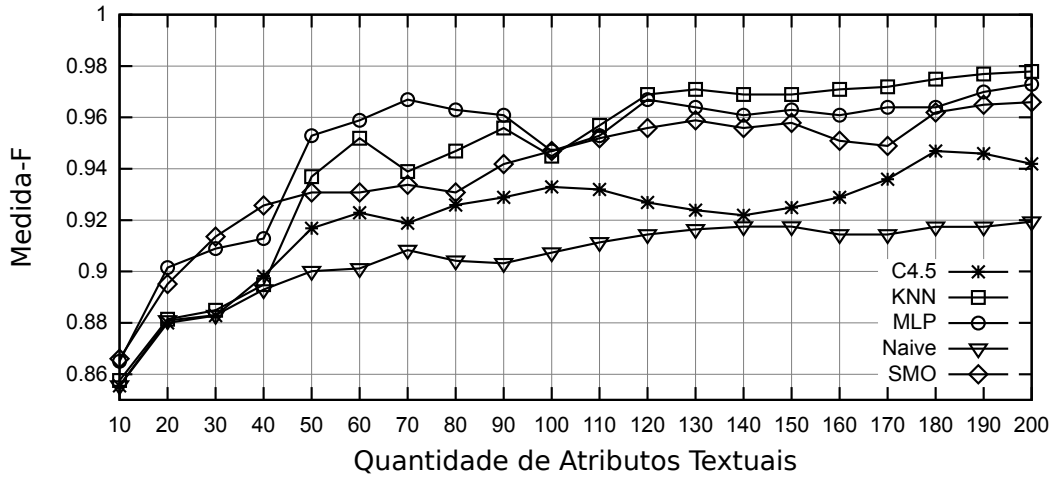

Figura 26: Desempenho dos Classificadores Usando a Métrica TFIDF no Conjunto de Dados em Inglês

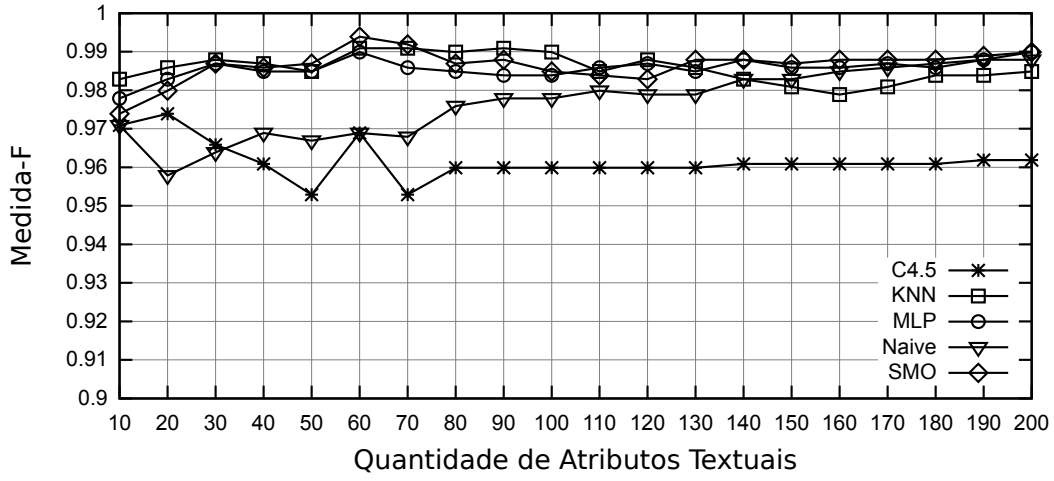

Figura 27: Desempenho dos Classificadores Usando a Métrica Ganho de Informação no Conjunto de Dados em Português

cador C4.5 obteve o pior resultado, já KNN e Naïve Bayes apresentaram resultados similares e intermediários. Os melhores resultados foram obtidos pelos classificadores MLP e SMO.

A Figura 29 mostra o resultado dos classificadores na base em Português usando para seleção de atributos a métrica TFIDF. De acordo com o gráfico o pior resultado foi do classificador C4.5, os demais apresentaram resultados bem próximos quando a dimensionalidade era superior a 100 atributos.

Analisando as informações apresentadas a respeito dos classificadores pode-se notar os seguintes fatos:

(a) Os classificadores MLP e SMO obtiveram os melhores resultados da Medida-F, apresentando bons resultados em todos os cenários de testes. A Medida-F de ambos os classifi-

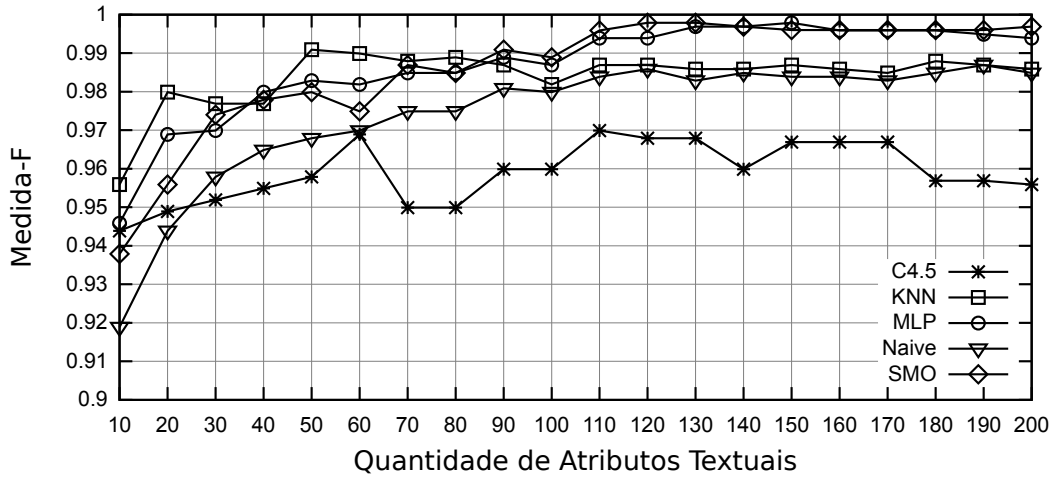

Figura 28: Desempenho dos Classificadores Usando a Métrica TF no Conjunto de Dados em Português

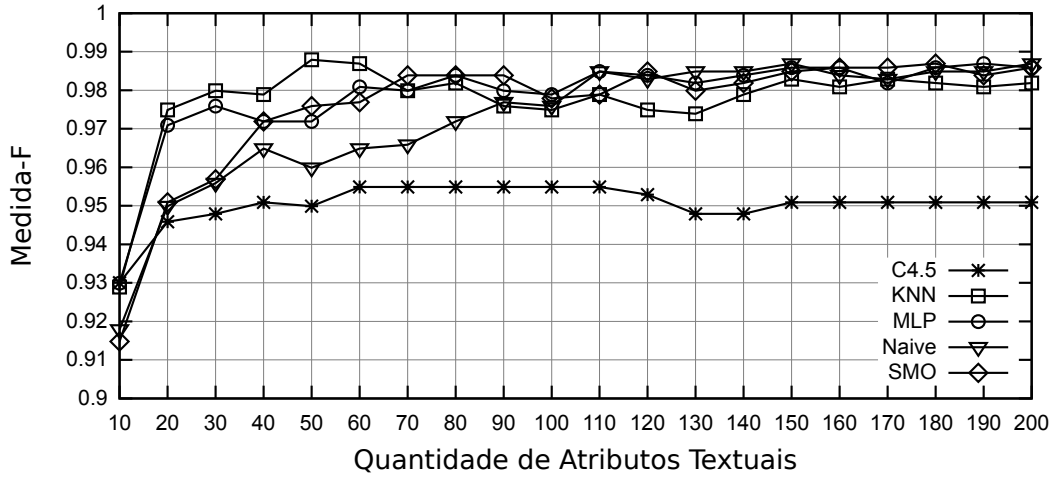

Figura 29: Desempenho dos Classificadores Usando a Métrica TFIDF no Conjunto de Dados em Português

cadores foi similar, principalmente quando a dimensionalidade era superior a 100 atributos.

(b) O algoritmo de aprendizagem de máquina KNN também obteve bons resultados, um pouco inferior, porém ainda próximos aos resultados obtidos pelos algoritmos MLP e SMO.

(c) O classificar Naïve Bayes conseguiu bons resultados apenas no idioma Português, para o idioma Inglês os resultados foram inferiores aos demais classificadores.

(d) O algoritmo C4.5 apresentou resultados ruins em todos os cenários de testes, sempre inferior a média de resultados obtidos pelos demais algoritmos de aprendizagem.

(e) Em geral os classificadores apresentaram piores resultados para o conjunto de dados em Inglês quando comparado com os dados em Português. Esse fato ocorreu independentemente da métrica de avaliação de atributos utilizada. Um motivo que pode explicar esse fato

é que o número de termos únicos do conjunto de dados em Inglês é muito maior do que os existentes para o Português, conforme os dados apresentados na sessão 4.3.

5.4 ATRIBUTOS ESTRUTURAIS

Nas sessões anteriores foram apresentados experimentos com métricas de avaliação de qualidade de atributos, dimensionalidade e também classificadores. Em todos esses experimentos foram utilizados apenas os textos presentes dentro do elemento <body> do HTML, não foi utilizado o texto presente nos demais elementos. O texto localizado no interior do marcador <body> é aquele apresentado ao usuário ao acessar a página *web*.

Nesta sessão serão mostrados experimentos que avaliam o desempenho do classificador utilizando diferentes atributos do HTML. O objetivo desses experimentos é verificar o quanto cada marcador do HTML pode contribuir para a correta categorização da página.

Para avaliar os marcadores do HTML foi utilizado a métrica de avaliação de atributos TF e o classificador SMO. A TF foi escolhida devido a ter apresentado bons resultados, conforme dados da sessão 5.1. Do mesmo modo o classificador SMO foi selecionado devido aos seus resultados nos experimentos anteriores, mostrados na sessão 5.3.

Esta sessão analisará apenas os resultados dos algoritmos de aprendizagem de máquina SMO, entretanto também foram realizados experimentos com os demais algoritmos abordados neste trabalho. Os resultados dos experimentos com os demais algoritmos constam no Apêndice B.

Os textos dos seguintes marcadores foram avaliados:

(1) *title*: Texto do elemento <title>, representa o título da página HTML.

(2) *description*: Texto presente no atributo *description* do elemento <meta>, localizado no cabeçalho da página. Possui a função de descrever o conteúdo da página.

(3) *keyword*: Texto presente no atributo *keyword* do elemento <meta>, localizado no cabeçalho da página. Reservado para inserir palavras chaves relacionadas ao conteúdo da página.

(4) âncora: Conteúdo textual entre os elementos de *Hyperlinks* <a> e . O conteúdo entre esses elementos geralmente é destacado para representar a ligação entre duas páginas.

(5) imagem: Conteúdo dos marcadores *alt* e *title* do elemento <img>. Esses textos

são utilizados para descrever o conteúdo das imagens contidas na página.

(6) *link*: Conteúdo dos marcadores *alt* e *title* do elemento de *Hyperlinks* <a>. Esses textos são utilizados para descrever os *Hyperlinks* presentes na página.

(7) título: Conteúdo textual dos elementos de títulos <h1>, <h2>, <h3>, <h4>, <h5> e <h6>. Esses elementos são utilizados para definir títulos dos textos presentes na página.

A sessão 2.3 apresenta informações a respeito da função estrutural de cada um desses marcadores utilizados na avaliação.

Além do conteúdo individual dos marcadores também foram realizados experimentos com combinações deles. O conteúdo dos marcadores foram combinados entre eles e também com o conteúdo textual da página, que é aquele conteúdo não presente nos marcadores estruturais. Os conjuntos de combinações realizadas foram:

Conjunto A: Combinação dos textos presentes em todos os elementos do cabeçalho da página (*title*, *description* e *keyword*).

Conjunto B: Combinação de todos os elementos localizados no elemento <body> do HTML (âncora, imagem, *link*, título e elementos de destaque de texto).

Conjunto C: Junção do conteúdo de todos os elementos de destaque de texto (<b>, <i>, <u>, <s>, e).

Conjunto D: União do conteúdo do Conjunto A com o conteúdo textual da página.

Conjunto E: Combinação do Conjunto A e Conjunto B com o conteúdo textual da página.

Conjunto F: Combinação do Conjunto B com o conteúdo textual da página.

Conjunto G: Combinação do Conjunto A com o Conjunto B.

A Tabela 18 apresenta os resultados dos experimentos no conjunto de dados em Português. A primeira coluna mostra a quantidade de atributos selecionados para o experimento, a segunda coluna mostra o resultado do experimento utilizando apenas o conteúdo textual da página, a coluna 3 até a coluna 9 apresenta os resultados utilizando conjunto de marcadores, por fim as colunas 10 até 16 contêm os resultados quando é utilizado marcadores individuais.

Na Tabela 19 são apresentados os resultados dos experimentos no conjunto de dados em Inglês. A distribuição das informações nessa Tabela é a mesma da utilizada na Tabela 18, sendo que na primeira coluna é mostrado a quantidade de atributos, na segunda o resultado

Tabela 18: Medida-F na Base em Português: Experimento com Atributos Estruturais

Qtde Atributos	Texto	Conjunto de Marcadores							Marcadores						
		A	B	C	D	E	F	G	1	2	3	4	5	6	7
10	0.9379	0.9438	0.9359	0.7705	0.9529	0.9639	0.9539	0.9599	0.9186	0.7524	0.7386	0.9239	0.8419	0.8608	0.7939
20	0.9559	0.9629	0.948	0.7859	0.9609	0.9659	0.9539	0.9669	0.9408	0.7632	0.7440	0.9499	0.8525	0.8650	0.8171
30	0.974	0.9649	0.9579	0.8022	0.9739	0.9709	0.9669	0.9759	0.9519	0.7734	0.7548	0.9639	0.8546	0.8744	0.8364
40	0.9779	0.9639	0.9639	0.8076	0.9829	0.9799	0.9759	0.9839	0.9559	0.7780	0.7606	0.9689	0.8566	0.9012	0.8454
50	0.9799	0.9649	0.9689	0.8099	0.9849	0.9799	0.9769	0.9869	0.9549	0.7768	0.7690	0.9859	0.8716	0.9062	0.8503
60	0.9749	0.9639	0.9769	0.8132	0.9909	0.9829	0.9769	0.9859	0.9549	0.7730	0.7713	0.996	0.8778	0.9042	0.8513
70	0.9869	0.9659	0.9769	0.8141	0.9949	0.9909	0.9799	0.9919	0.9569	0.7776	0.7636	0.994	0.8761	0.9103	0.8545
80	0.9849	0.9639	0.9789	0.8174	0.9959	0.9909	0.9809	0.9949	0.9539	0.7776	0.7657	0.9919	0.8780	0.9103	0.8535
90	0.9909	0.9649	0.9929	0.8085	0.9949	0.9949	0.9819	0.9949	0.9539	0.7799	0.7747	**0.9959**	0.8770	0.9063	0.8536
100	0.9889	0.9629	0.9939	0.8085	0.9949	**0.9969**	0.9859	0.996	0.9549	0.7831	0.7747	**0.9959**	0.8852	0.9125	0.8556
110	0.9959	0.9629	0.9959	0.8134	**0.9989**	0.998	0.9959	0.996	0.9559	0.7829	**0.7759**	**0.9959**	0.8872	0.9135	0.8525
120	**0.9979**	0.9649	0.9959	0.8182	**0.9989**	0.998	0.996	0.998	0.9559	0.7839	0.7747	0.9939	0.8882	0.9094	0.8556
130	**0.9979**	**0.9669**	0.9949	0.8225	**0.9989**	0.998	**0.9969**	0.998	0.9559	0.7839	0.7724	0.9939	0.8871	0.9104	0.8546
140	0.9969	0.9649	0.9979	0.8239	**0.9989**	0.998	**0.9969**	0.9969	0.9559	0.7861	0.7736	0.9949	0.9095	0.9136	0.8546
150	0.996	0.9639	**0.9989**	0.8209	**0.9989**	0.998	0.9949	0.9969	0.9559	0.7888	0.7736	0.9939	0.9085	0.9145	0.8547
160	0.996	0.9649	0.9969	0.8171	0.9969	0.998	0.9949	0.9969	0.9559	**0.7891**	0.7724	0.9939	0.9206	0.9156	0.8548
170	0.996	0.9639	0.9969	0.8119	0.9969	**0.9969**	0.9959	**0.9989**	0.9559	**0.7891**	0.7726	0.9939	0.9206	**0.9166**	**0.8620**
180	0.996	0.9659	0.9979	0.8223	0.9969	0.9949	0.9959	**0.9989**	0.9559	0.7872	0.7726	0.9939	0.9226	0.9165	0.8559
190	0.996	**0.9669**	0.9979	0.8275	0.9959	0.9959	**0.9969**	**0.9989**	0.9569	0.7853	0.7726	0.9939	**0.9246**	0.9135	0.8560
200	0.9969	0.9659	0.9979	**0.8336**	0.9959	0.9959	**0.9969**	**0.9989**	**0.9579**	0.7853	0.7717	0.9949	**0.9246**	0.9145	0.8590

dos experimentos utilizando o conteúdo textual da página, as colunas 3 até 9 apresentam os resultados quando é utilizado conjunto de marcadores e nas colunas seguintes a Medida-F para quando é feito uso de marcadores individuais.

Tabela 19: Medida-F na Base em Inglês: Experimento com Atributos Estruturais

Qtde Atributos	Texto	Conjunto de Marcadores							Marcadores						
		A	B	C	D	E	F	G	1	2	3	4	5	6	7
10	0.8622	0.8830	0.8785	0.5702	0.8877	0.9008	0.8775	0.9038	0.8071	0.8107	0.7687	0.8445	0.7599	0.7428	0.6673
20	0.8803	0.9155	0.9077	0.5491	0.8996	0.9279	0.9127	0.9299	0.8350	0.8139	0.8088	0.8803	0.7767	0.7590	0.6827
30	0.9197	0.9186	0.9187	0.6328	0.9478	0.9539	0.9419	0.9529	0.8448	0.8461	0.8077	0.9216	0.7855	0.7783	0.7073
40	0.9278	0.9247	0.9318	0.6302	0.9429	0.9629	0.9358	0.9639	0.8514	0.8526	0.8121	0.9236	0.7827	0.7787	0.7202
50	0.9368	0.9307	0.9398	0.6173	0.9559	0.9649	0.9459	0.9619	0.8562	0.8597	0.8188	0.9308	0.7816	0.7839	0.7430
60	0.9359	0.9398	0.9469	0.6438	0.9529	0.9649	0.9529	0.9649	0.8586	0.8629	0.8211	0.9459	0.7830	0.7859	0.7475
70	0.9329	0.9428	0.9529	0.6420	0.9489	0.9619	0.9529	0.9659	0.8654	0.8599	0.8179	0.9489	0.7830	0.7977	0.7441
80	0.9389	0.9418	0.9559	0.6636	0.9629	0.9619	0.9459	0.9659	0.8695	0.8600	0.8211	0.9709	0.7832	0.8015	0.7515
90	0.9509	0.9418	0.9569	0.6666	0.9669	0.9649	0.9539	0.9659	0.8746	0.8663	0.8211	0.9699	0.7837	0.7991	0.7536
100	0.9539	0.9368	0.9629	0.6682	0.9709	0.9639	0.9559	0.9769	0.8736	0.8673	0.8222	0.9629	0.7810	0.8021	0.7911
110	0.9529	**0.9469**	0.9649	0.6703	0.9769	0.9709	0.9519	0.9789	0.8859	0.8703	0.8232	0.9639	0.7770	0.8031	0.7893
120	0.9649	0.9459	0.9699	0.6712	**0.9809**	0.9729	0.9619	0.9829	0.8858	0.8685	0.8232	0.9649	0.7823	0.8040	0.7936
130	0.9609	0.9459	0.9759	0.6823	0.9769	0.9729	0.9699	0.9829	0.8910	0.8671	0.8232	0.9689	0.7812	0.8054	0.7898
140	0.9599	0.9429	0.9749	0.6877	0.9739	0.9759	0.9629	0.9819	0.8972	0.8683	0.8243	0.9749	0.7922	0.7998	0.8004
150	0.9629	0.9429	0.9809	0.6950	0.9779	0.9769	0.9609	0.9859	0.8921	0.8673	0.8243	0.9699	0.7853	0.8069	0.8002
160	0.9659	0.9439	0.9829	0.7049	0.9759	0.9769	0.9619	0.9869	0.8941	0.8673	0.8243	0.9689	0.7899	0.8110	0.8061
170	0.9619	0.9449	0.9739	0.7130	0.9759	0.9749	0.9669	0.9879	0.8962	0.8663	0.8254	0.9679	0.7922	0.8081	0.8186
180	0.9629	0.9429	0.9819	0.7145	0.976	0.974	0.9659	0.9869	0.9003	0.8684	0.8254	0.9719	0.7952	0.8139	0.8220
190	**0.9679**	**0.9469**	0.9829	**0.7198**	0.9769	0.9779	0.9689	0.9889	0.9053	0.8704	0.8244	0.9749	**0.7963**	**0.8160**	0.8204
200	0.9639	0.9449	**0.9839**	0.7191	0.9749	**0.9799**	**0.9719**	**0.9899**	**0.9084**	**0.8714**	**0.8255**	**0.9789**	0.7948	0.8056	**0.8387**

Os resultados das Tabelas 18 e 19 serão comparados entre si, utilizando como base os valores atingidos pela classificação que usou apenas a parte textual, informação que consta na segunda coluna das tabelas. Para facilitar a visualização dos dados foram criados quatro grupos de comparações para cada idioma, em cada grupo os dados são comparados com os valores atingidos pelo classificador base.

A Figura 30 apresenta os resultados do grupo de comparações 1 para o idioma Inglês, já a Figura 31 mostra os resultados do mesmo grupo porém para o idioma Português. Conforme os dados apresentados pelos dois gráficos pode-se constatar que o Conjunto C - composto pelos marcadores de destaque de texto - obteve os piores resultados do grupo. Enquanto o Conjunto A - composto pelos marcadores do cabeçalho da página - atingiu resultados próximos aos atingidos pelo classificador que utilizou apenas as informações textuais da página. O Conjunto B que é formado pelas informações dos marcadores de âncora, imagem, *link*, título e destaque obteve bons resultados, no idioma Português os valores foram similares aos atingidos pelo classificador utilizado como base das comparações e para o Inglês os resultados foram superiores a esse classificador.

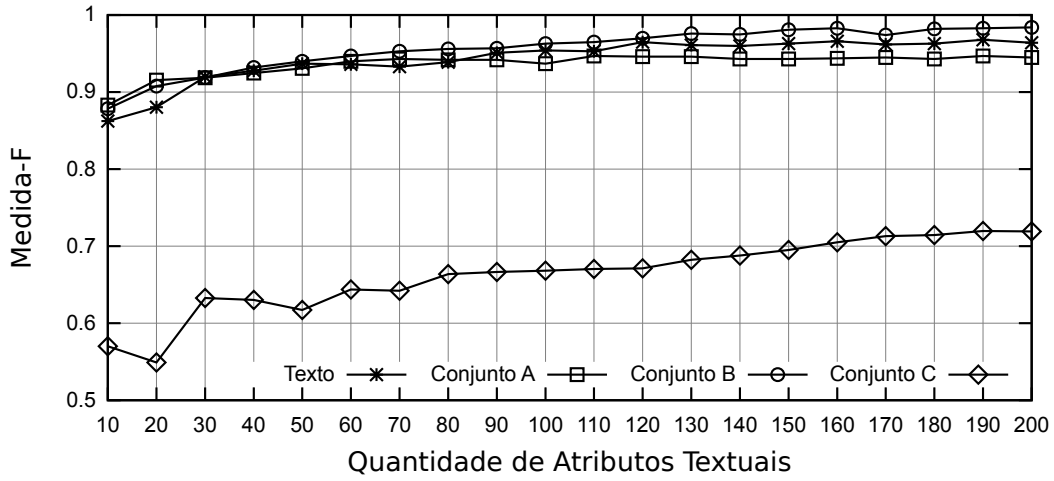

Figura 30: Desempenho do Classificador SMO no Conjunto de Dados em Inglês: Grupo de Comparações 1

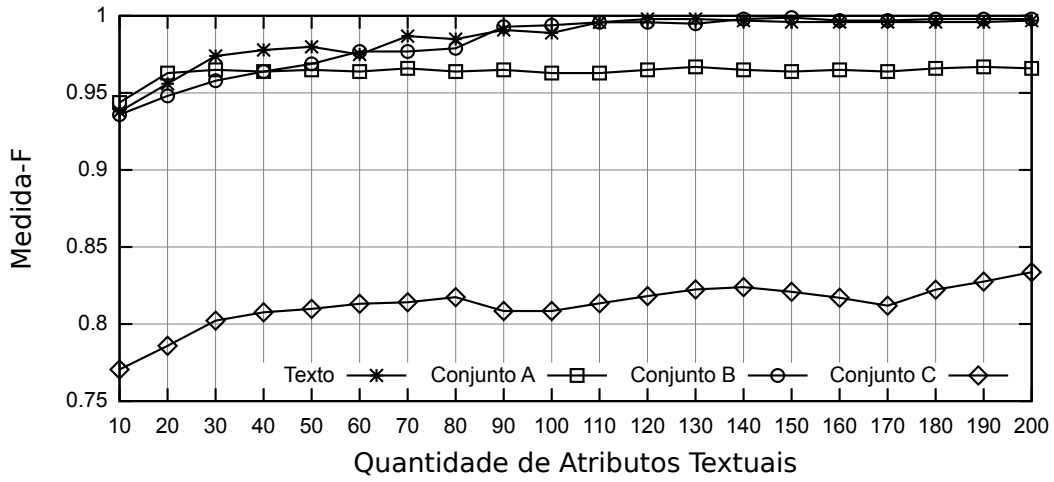

Figura 31: Desempenho do Classificador SMO no Conjunto de Dados em Português: Grupo de Comparações 1

Os resultados da Medida-F para o grupo de comparação 2 são exibidos nas Figuras

32 e 33, respectivamente resultados nos idiomas Inglês e Português. Nesse segundo grupo de comparações os resultados ficaram mais próximos entre os conjuntos avaliados, ficando um pouco mais distante no idioma Inglês e similares no idioma Português. O Conjunto G - composto pelas informações dos marcadores do cabeçalho da página e também dos marcadores âncora, imagem, *link*, título e destaque - obteve os melhores resultados, superiores até mesmo ao classificador utilizado como base das comparações. Os demais conjuntos (D, E, F) atingiram resultados similares, principalmente para o Português.

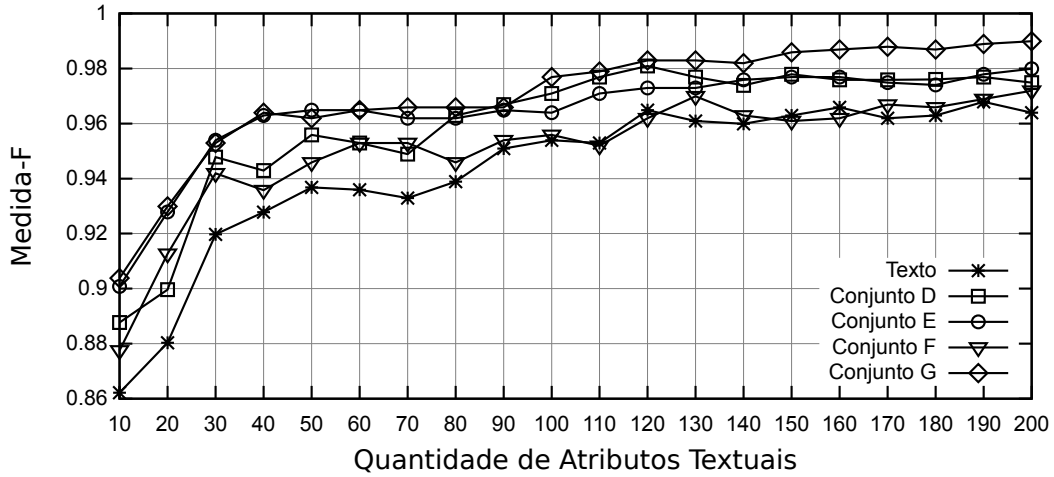

Figura 32: Desempenho do Classificador SMO no Conjunto de Dados em Inglês: Grupo de Comparações 2

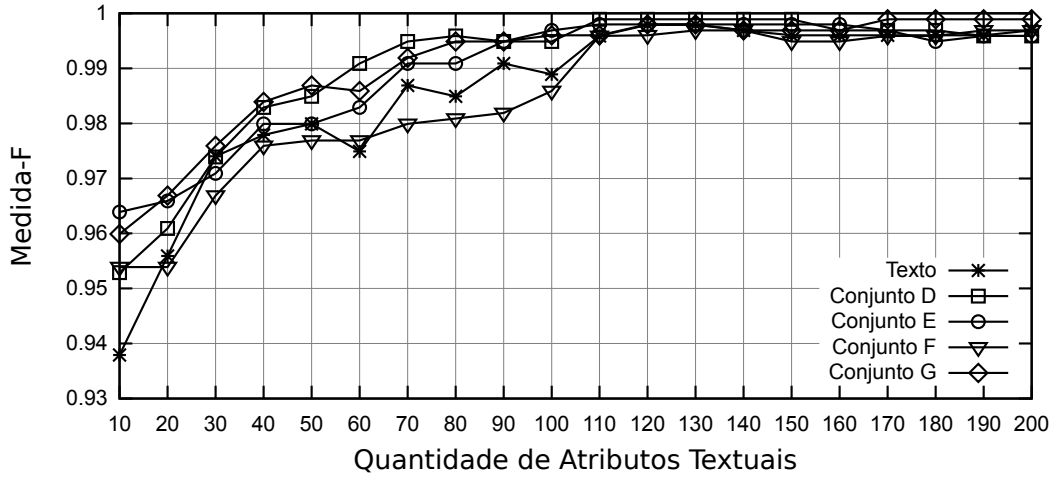

Figura 33: Desempenho do Classificador SMO no Conjunto de Dados em Português: Grupo de Comparações 2

Os resultados para o terceiro grupo de comparações são mostrados nas Figuras 34 e 35, respectivamente usando conjunto de dados nos idiomas Inglês e Português. Nenhum dos atributos avaliados atingiu um resultado superior aos atingidos pelo classificador base que utilizava apenas informações textuais. Os piores resultados foram obtidos pelo classificador que

usou apenas as informações do marcador *keyword*. Os resultados dos classificadores baseados em informações dos marcadores *title* e *description* foram intermediários em relação aos demais marcadores, sendo que o primeiro obteve resultados melhores que o segundo. O baixo desempenho dos classificadores que usaram os marcadores *title* e *description* pode ser explicado pelas estatísticas do conjunto de dados (sessão 4.3), as quais mostram que esses marcadores não são utilizados em todas as páginas.

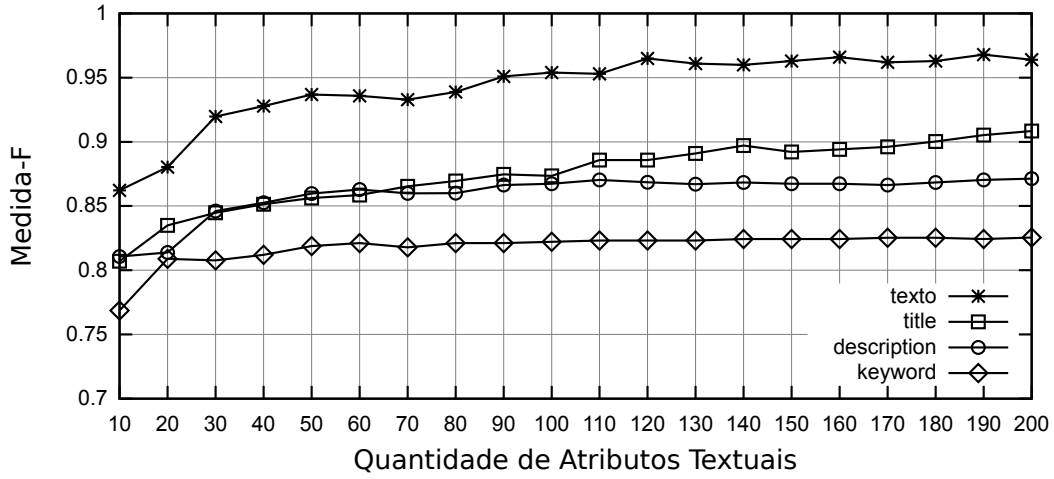

Figura 34: Desempenho do Classificador SMO no Conjunto de Dados em Inglês: Grupo de Comparações 3

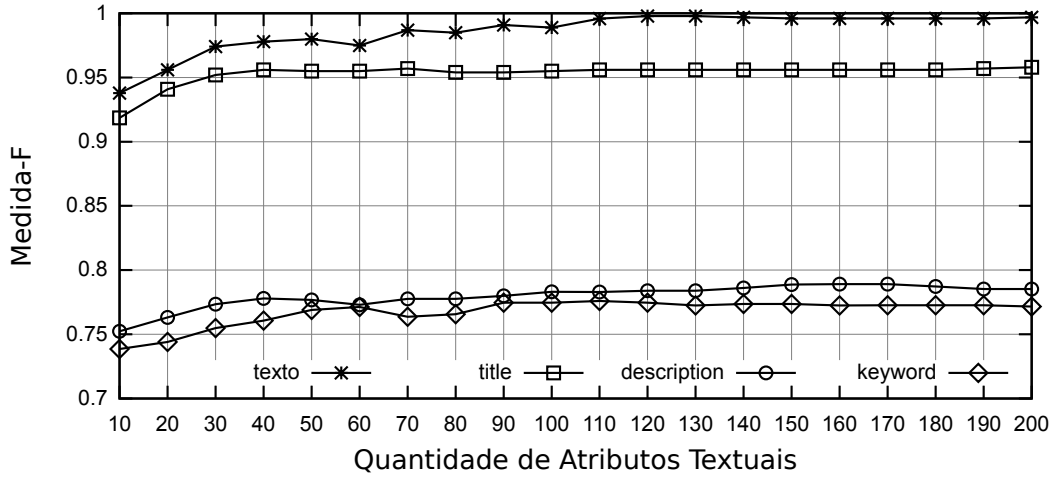

Figura 35: Desempenho do Classificador SMO no Conjunto de Dados em Português: Grupo de Comparações 3

Na Figura 36 é apresentado os resultados do último grupo de comparação, o grupo 4, enquanto a Figura 37 mostra os resultados do mesmo grupo porém para o idioma Português. De acordo com os dados apresentados pelos gráficos é possível observar que os marcadores *link* e imagem obtiveram resultados inferiores ao classificador base. Resultado semelhante foi obtido pelo classificador que utilizava as informações dos marcadores de título. Já o classificador que

usava as informações do marcador âncora atingiu resultados superiores ao classificador base no idioma Inglês e resultados similares no idioma Português.

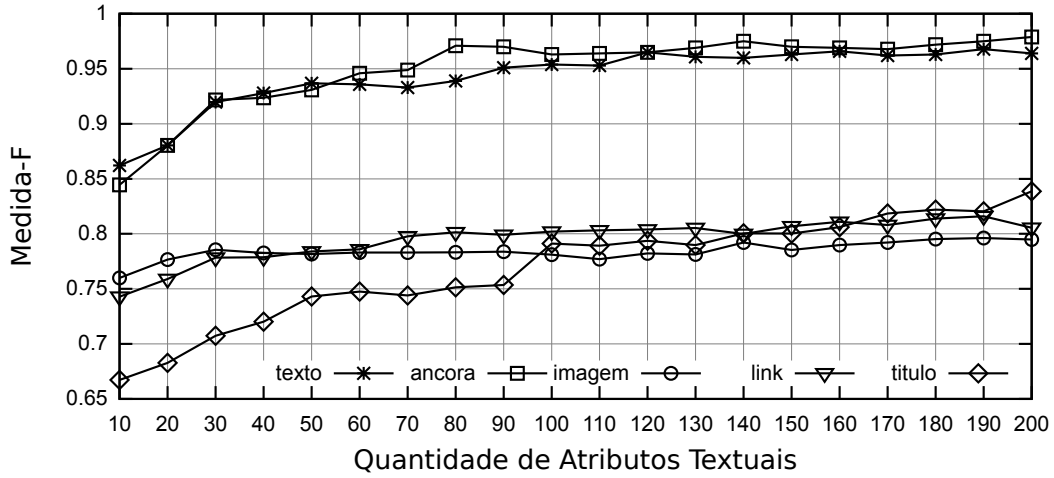

Figura 36: Desempenho do Classificador SMO no Conjunto de Dados em Inglês: Grupo de Comparações 4

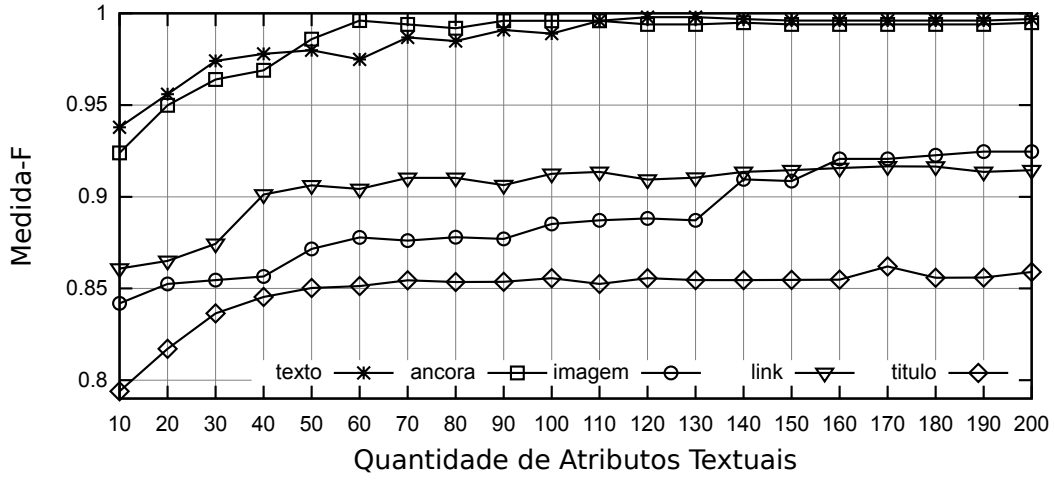

Figura 37: Desempenho do Classificador SMO no Conjunto de Dados em Português: Grupo de Comparações 4

Analisando as informações apresentadas pelas Tabelas 18 e 19 e também pelos gráficos apresentados neste Capítulo alguns fatos podem ser notados:

(a) O classificador que utilizou informações provenientes do marcador âncora obteve os melhores resultados comparado com outros classificadores que utilizaram informações provenientes de um único marcador. Os resultados no idioma Inglês foram superiores até mesmo ao classificador utilizado como base nas comparações, que é o classificador baseado apenas nas informações textuais. No idioma Português os resultados do classificador que utilizou as informações do texto âncora foram similares ao classificador utilizado como base.

(b) Os demais marcadores, a exceção do âncora, não atingiram resultados da Medida-F superiores aos atingidos pelo classificador base. Um dos motivos que explica esse fato é que alguns desses marcadores como *description* e *keyword* não estão presentes em todas as páginas do conjunto de dados. A sessão 4.3 exibe uma Tabela que contém informações sobre o uso de cada um desses marcadores de acordo com o idioma e classe da página.

(c) Em relação aos experimentos com conjuntos de marcadores, observa-se que os Conjuntos B e G conquistaram resultados melhores que o classificador base. Esses dois conjuntos são compostos por informações provenientes do marcador âncora, assim, acredita-se que o bom desempenho ocorreu devido a esse uso.

(d) Os conjuntos D, E e F também apresentaram resultados melhores que o classificador base, sendo mais visível essa melhora no idioma Inglês. Esses conjuntos são compostos por informações do marcador âncora e/ou das informações utilizadas pelo classificador base.

(e) Os únicos conjuntos que tiveram resultados inferiores ao classificador base foram os conjunto A e C. Não por acaso também são os únicos conjuntos que não utilizam informações provenientes do marcador âncora ou informações utilizadas pelo classificador base.

5.5 CONSIDERAÇÕES FINAIS

Este Capítulo mostrou o resultado de experimentos envolvendo classificadores, métricas de avaliação de atributos, dimensionalidade e atributos estruturais do HTML.

Em todos os experimentos os resultados para o conjunto de dados em Português foram melhores que os resultados atingidos no conjunto de dados em Inglês. Acredita-se que esse fato tenha ocorrido devido a base em Português possuir um número menor de termos únicos principalmente para uma de suas classes, conforme dados apresentados na sessão 4.3.

Outro fato observado é que a Medida-F atingida nos experimentos foi superior ao que era esperado para um problema de classificação textual. Um dos motivos que contribui para esse índice é o fato da base conter textos de duas categorias com vocabulários bem distintos. A categoria pornográfica possui termos que, em geral, não são utilizados pela outra categoria presente na base.

Foi constatado que os classificadores que usaram apenas informações textuais atingiram bons índices de instâncias corretamente classificadas. Na base em Português foi atingido uma Medida-F de 99.79% e Acurácia de 99.8%, já na base em Inglês a Medida-F atingida foi de 96.79% e Acurácia de 96.8%. Os valores conquistados na base em Português são melhores

que os resultados de nove dos dez trabalhos apresentados no Capítulo 2, cujo os resultados estão sumarizados na Tabela 3. Utilizando como base os resultados na base em Inglês temos que a Acurácia atingida foi superior a sete dos dez trabalhos referenciados na Tabela 3.

Notou-se ainda que utilizando uma dimensionalidade entre 200 a 300 atributos é possível atingir bons índices de acerto na categorização. Quando utilizado uma dimensionalidade maior não ocorreu um aumento proporcional na Medida-F, por outro lado ocorreu um aumento no tempo para treinamento dos algoritmos de aprendizado de máquina.

Por fim, com relação aos atributos estruturais constatou-se que as informações dos *hyperlinks*, mais especificamente do texto âncora, contribuíram de maneira significativa na construção de um classificador mais eficiente no que diz respeito ao índice de instâncias corretamente classificadas. Conforme os experimentos realizados é possível afirmar que um classificador que utilize apenas as informações do texto âncora atinge uma Medida-F superior a um classificador que utilize qualquer outro marcador HTML. Os resultados de um classificador baseado exclusivamente no texto âncora também são iguais ou superiores a classificadores que utilizem outras informações textuais presentes na página.

6 CONCLUSÕES E PERSPECTIVAS

Neste trabalho foi apresentado um estudo empírico sobre métodos de classificação automática de páginas *web*. O estudo realizado teve como objetivo contribuir com os métodos atuais de classificação através de experimentos com quatro aspectos envolvidos no processo de categorização: algoritmos de aprendizagem de máquina, dimensionalidade (total de atributos utilizados pelo classificador), métricas de avaliação de qualidade de atributos e seleção de atributos textuais e estruturais presentes nas páginas HTML.

Inicialmente identificou-se a necessidade da construção de um conjunto de dados para que os experimentos fossem viáveis. Esse conjunto foi elaborado com sítios de dois idiomas - Português e Inglês - com um total de 4.000 páginas *web*. Após isso foram extraídas algumas informações sobre esses dados, informações essas importantes para a interpretação dos resultados de alguns métodos de classificação. Destacam-se entre essas informações: o percentual que mostra a baixa utilização dos marcadores *description* e *keyword* comparado ao uso dos demais marcadores; a quantidade menor de *tokens* únicos para o idioma Português em relação ao Inglês; a alta relação existente entre a categoria das páginas e de seus sites pais e sites filhos.

Visando possibilitar a execução da classificação em tempo real, impedindo o acesso a sites de categorias não permitidas, foi sugerido no Apêndice A um ambiente de rede de computadores para implementação dos métodos de categorização abordados neste trabalho. Além disso foram realizados experimentos mostrando o tempo necessário para o treinamento de cada um dos classificadores, sabe-se que o tempo para predição de uma instância é significativamente menor que o tempo de treinamento, assim é possível afirmar que todos os classificadores abordados neste trabalho podem ser utilizados no ambiente sugerido.

Sobre os resultados dos experimentos observou-se uma grande variação da Medida-F conforme os métodos utilizados. Em geral, obteve-se bons resultados utilizando: o classificador SMO, a métrica TF, dimensionalidade inferior a 200 atributos e informações textuais presentes na página. Usando esses elementos foi possível atingir uma Medida-F de 99,79% na base de dados em Português e de 96,79% no conjunto de dados em Inglês. Valores iguais e algumas

vezes um pouco superiores a esses também foram atingidos por outras combinações, porém de modo geral essa combinação de elementos foi a que apresentou os melhores índices.

Ainda sobre a variação dos resultados, em uma dimensionalidade de até 200 atributos e utilizando apenas as informações textuais tem-se uma variação de 91,48% a 99,79% para o Português e de 83,71% a 98,39% para o Inglês. Uma oscilação de 8,31% e de 14,68% respectivamente para os idiomas Português e Inglês. Essa variação demostra a importância da escolha correta dos classificadores, métricas de avaliação de atributos e dimensionalidade a ser utilizada na construção do classificador para o problema em questão, bem como da importância da realização de experimentos envolvendo esses elementos.

Nos experimentos envolvendo as métricas de avaliação de qualidade de atributos a TF (*Term Frequency*), no geral, obteve os melhores resultados. A métrica Ganho de Informação conquistou bons resultados, alguns deles próximos ou um pouco superiores a TF. Enquanto a TFIDF (*Term Frequency Inverse Document Frequency*) teve o pior desempenho comparada as demais métricas. Deste modo a TF foi a que proporcionou uma melhor redução de dimensionalidade sem comprometer os índices de acerto dos classificadores.

Com relação ao aumento da dimensionalidade observou-se que os algoritmos de aprendizagem de máquina tiveram um comportamento distinto. O algoritmo KNN apresentou um tendência de queda da Medida-F quando a quantidade de atributos foi superior a determinado patamar, os demais algoritmos apresentaram uma tendência de estabilização. Assim, aumentar a dimensionalidade depois de certo ponto não teve como consequência uma melhora nos resultados do classificador, em alguns casos ocorreu justamente o oposto. Notou-se também que conforme a dimensionalidade aumentava as diferentes métricas de avaliação de atributos passaram a atingir resultados similares, isso demostra que quanto menor a dimensionalidade maior é a influência das métricas nos resultados dos classificadores.

Acerca dos classificadores, MLP e SMO obtiveram os melhores resultados nos conjuntos de testes realizados, apresentando resultados semelhantes entre eles. O classificador KNN também obteve bons resultados, alguns deles próximos aos atingidos pelo MLP e SMO. Já o classificador Naïve Bayes apresentou um bom desempenho apenas nas páginas em Português, enquanto o classificador C4.5 apresentou desempenho inferior aos demais classificadores independente da métrica de avaliação de atributos utilizada ou idioma.

Os experimentos com atributos estruturais mostraram que a exceção do texto âncora nenhum outro marcador individual obteve desempenho superior ao classificador base. O classificador utilizado como base utilizava apenas informações textuais, desconsiderando aquelas presentes nos marcadores estruturais. Marcadores localizados no cabeçalho da página como

keyword e *description* atingiram uma Medida-F inferior a 80.00% na base em Português, índice esse bem abaixo do registrado pelo classificador base que foi de 99.79%. Além desses experimentos com atributos individuais também foram feitas comparações envolvendo o uso combinado de dois ou mais marcadores individuais, nessas comparações os únicos conjuntos que atingiram um índice de classificação melhor que o classificador base foram aqueles que eram compostos por informações do texto âncora.

O texto âncora presente no marcador de *hyperlink* é utilizado para sinalizar a ligação de um documento HTML com outro, esse texto segundo os experimentos realizados demonstrou-se ser de grande valia para a tarefa de categorização de páginas *web*. No conjunto de dados em Português o classificador que utilizou exclusivamente essas informações atingiu uma Medida-F de 99.59% enquanto o classificador base atingiu o índice de 99.79%. Nas páginas em Inglês esse classificador obteve uma Medida-F de 97.89%, índice superior ao atingido pelo classificador base que foi de 96.79%. Os resultados dos experimentos demonstram que é possível construir um classificador com alto índice de acerto utilizando apenas as informações do texto âncora, sem o uso das demais informações da página.

Além dos bons índices de acerto o uso do texto âncora possui outra vantagem que é o tempo de pré-processamento da página. O pré-processamento apenas dessas informações é mais rápido do que ter que processar todas as informações textuais existentes no documento. Para uma aplicação de classificação em tempo real essa características é uma grande vantagem.

Devido aos resultados apresentados ao longo dos experimentos, levando em consideração a questão do tempo de processamento e o percentual de acerto do método utilizado, decidiu-se por construir um classificador com os seguintes elementos: algoritmo de aprendizagem de máquina SMO; métrica de avaliação de atributos TF; dimensionalidade de duzentos atributos e informações textuais apenas do texto âncora. Esse classificador foi implantado em uma rede de computadores utilizada para prover acesso a internet.

O classificador construído foi implantado, com sucesso, utilizando o ambiente de redes sugerido no Apêndice A. As análises sequentes a implantação evidenciaram a viabilidade da solução proposta, bem como a eficiência do método desenvolvido. O classificador manteve um bom índice de acerto, próximo ao atingido na base de testes.

Os resultados apresentados encorajam o desenvolvimento de outros trabalhos que abordem a classificação com classes diferentes das que foram aqui utilizadas, por exemplo, trabalhos que auxiliem na identificação de sítios maliciosos, como os utilizados por criminosos para captura de informações pessoais dos usuários.

Outra possibilidade de estudo é a elaboração de propostas de classificadores para páginas que não possuem textos, apenas conteúdo multimídia. Para esses casos um classificador que utiliza apenas informações textuais é ineficiente. Um dos fatores que pode ser utilizado nessas situações é a informação da classe dos sítios pais e filhos, as estatísticas do conjunto de dados demonstraram que existe uma relação da classe da página com a classe dos sítios que estão conectados a ela. Assim um classificador que utilize essas informações tende a apresentar bons resultados. Neste trabalho não foi utilizado essa informação devido aos resultados atingidos serem satisfatórios, entretanto para outros cenários isso pode vir a ser útil.

Igualmente interessante é o desenvolvimento de estudos que tratem da questão do aprendizado contínuo do classificador, como manter o mesmo índice de acerto da categorização em um ambiente tão dinâmico como a internet. Uma das possibilidades a ser estudada é atualização periódica do conjunto de dados com as informações atuais das páginas *web*, seguida do treinamento do classificador com tais dados.

REFERÊNCIAS

AGARWAL, N.; LIU, H.; ZHANG, J. Blocking objectionable web content by leveraging multiple information sources. **ACM SIGKDD Explorations Newsletter**, ACM, New York, NY, USA, v. 8, n. 1, p. 17–26, jun. 2006. ISSN 1931-0145.

AHMADI, A.; FOTOUHI, M.; KHALEGHI, M. Intelligent classification of web pages using contextual and visual features. **Applied Soft Computing**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 11, n. 2, p. 1638–1647, mar. 2011. ISSN 1568-4946.

ALEXA. **The top 500 sites on the web. (http://www.alexa.com/) - Acesso em 10/05/2012**. 2012.

ALIMOHAMMADI, D. Measurement of the presence of keywords and description meta-tags on a selected number of iranian web sites. **Online Information Review**, v. 28, n. 3, p. 220–223, 2004.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. 1st. ed.Addison Wesley, 1999. Paperback. ISBN 020139829X.

BORKO, H.; BERNICK, M. Automatic document classification. **Journal of the Association for Computing Machinery**, ACM, New York, NY, USA, v. 10, p. 151–162, April 1963. ISSN 0004-5411.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the fifth annual workshop on Computational learning theory**. New York, NY, USA: ACM, 1992. (COLT '92), p. 144–152. ISBN 0-89791-497-X.

BURGES, C. J. C. A tutorial on support vector machines for pattern recognition. **Data Mining and Knowledge Discovery**, Kluwer Academic Publishers, Hingham, MA, USA, v. 2, n. 2, p. 121–167, jun. 1998. ISSN 1384-5810.

CARPENTER, G.; GROSSBERG, S.; ROSEN, D. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. **Neural Networks**, Elsevier Science Ltd., Oxford, UK, UK, v. 4, n. 6, p. 759–771, 1991. ISSN 0893-6080.

CAULKINS, J. P.; DING, W.; DUNCAN, G.; KRISHNAN, R.; NYBERG, E. A method for managing access to web pages: filtering by statistical classification (fsc) applied to text. **Decision Support Systems**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 42, p. 144–161, October 2006. ISSN 0167-9236.

CAVNAR, W. B.; TRENKLE, J. M. N-Gram-Based Text Categorization. In: **In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval**. 1994. p. 161–175.

CHAKRABARTI, S.; JOSHI, M. M.; PUNERA, K.; PENNOCK, D. M. The structure of broad topics on the web. In: **WWW '02: Proceedings of the 11th international conference on World Wide Web**. New York, NY, USA: ACM Press, 2002. p. 251–262. ISBN 1581134495.

CHANG, C.-C.; LIN, C.-J. Libsvm: A library for support vector machines. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM, New York, NY, USA, v. 2, n. 3, p. 27:1–27:27, maio 2011. ISSN 2157-6904.

CHEN, Y.; WU, O. Semi-automated feature selection for web text filtering. In: **Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on**. 2010. v. 6, p. 2513 –2517.

CHUA, L.; YANG, L. Cellular neural networks: applications. **Circuits and Systems, IEEE Transactions on**, v. 35, n. 10, p. 1273–1290, 1988. ISSN 0098-4094.

CHUA, L.; YANG, L. Cellular neural networks: theory. **Circuits and Systems, IEEE Transactions on**, v. 35, n. 10, p. 1257–1272, 1988. ISSN 0098-4094.

COMON, P. Independent component analysis, a new concept? **Signal Processing**, Elsevier North-Holland, Inc., Amsterdam, The Netherlands, The Netherlands, v. 36, n. 3, p. 287–314, abr. 1994. ISSN 0165-1684.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, Springer Netherlands, v. 20, p. 273–297, 1995. ISSN 0885-6125. 10.1007/BF00994018.

COVER, T.; HART, P. Nearest neighbor pattern classification. **Information Theory, IEEE Transactions on**, v. 13, n. 1, p. 21–27, 1967. ISSN 0018-9448.

CRAVEN, T. C. Variations in use of meta tag descriptions by web pages in different languages. **Information Processing and Management**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 40, n. 3, p. 479–493, jan. 2004. ISSN 0306-4573.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An introduction to support Vector Machines: and other kernel-based learning methods**. New York, NY, USA: Cambridge University Press, 2000. ISBN 0-521-78019-5.

DUDA, R. O.; HART, P. E. **Pattern Classification and Scene Analysis**. 1. ed.John Wiley & Sons, 1973. Hardcover. ISBN 0471223611.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, Elsevier Science Inc., New York, NY, USA, v. 27, n. 8, p. 861–874, jun. 2006. ISSN 01678655.

FIELDING, R.; GETTYS, J.; MOGUL, J.; FRYSTYK, H.; MASINTER, L.; LEACH, P.; BERNERS-LEE, T. Rfc 2616: Hypertext transfer protocol - http/1.1. **Internet Engineering Task Force (IETF)**, 1999.

FLEXER, A. On the use of self-organizing maps for clustering and visualization. **Intelligent Data Analysis**, IOS Press, Amsterdam, The Netherlands, The Netherlands, v. 5, n. 5, p. 373–384, out. 2001. ISSN 1088-467X.

FORMAN, G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. **Journal of Machine Learning Research**, v. 3, p. 1289–1305, mar. 2003.

GAO, Z.; LU, G.; DONG, H.; WANG, S.; WANG, H.; WEI, X. Applying a novel combined classifier for hypertext classification in pornographic web filtering. In: **International Conference on Internet Computing in Science and Engineering, 2008. ICICSE '08.** 2008. p. 270 –273.

GOOGLE. **1 trillion web page milestone hit (http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html) - Acesso em 01/11/2011**. 2008.

GOOGLE. **Google does not use the keywords meta tag in web ranking. (http://googlewebmastercentral.blogspot.com.br/2009/09/google-does-not-use-keywords-meta-tag.html) - Acesso em 10/02/2013**. 2009.

GREFENSTETTE, G.; NIOCHE, J. Estimation of english and non-english language use on the www. In: **In Recherche d' Information Assistée par Ordinateur (RIAO**. 2000. p. 237–246.

GROSSBERG, S. Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions. **Biological Cybernetics**, v. 23, n. 4, p. 187–202, ago. 1976. ISSN 0340-1200.

GULLI, A.; SIGNORINI, A. The indexable web is more than 11.5 billion pages. In: **Special interest tracks and posters of the 14th international conference on World Wide Web**. New York, NY, USA: ACM, 2005. (WWW '05), p. 902–903. ISBN 1-59593-051-5.

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

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. The WEKA data mining software: an update. **Special Interest Group on Knowledge Discovery and Data Mining Explorer Newsletter**, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, nov. 2009. ISSN 1931-0145.

HAMMAMI, M.; CHAHIR, Y.; CHEN, L. Webguard: a web filtering engine combining textual, structural, and visual content-based analysis. **IEEE Transactions on Knowledge and Data Engineering**, v. 18, n. 2, p. 272 – 284, feb. 2006. ISSN 1041-4347.

HECHT-NIELSEN, R. Theory of the backpropagation neural network. In: **Neural Networks, 1989. IJCNN., International Joint Conference on**. 1989. p. 593–605 vol.1.

HO, W.; WATTERS, P. Identifying and blocking pornographic content. In: **Data Engineering Workshops, 2005. 21st International Conference on**. 2005. p. 1181–1181.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. **Neural Networks**, v. 2, n. 5, p. 359–366, jan. 1989. ISSN 08936080.

HU, W.; WU, O.; CHEN, Z.; FU, Z.; MAYBANK, S. Recognition of pornographic web pages by classifying texts and images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 29, n. 6, p. 1019 –1034, june 2007. ISSN 0162-8828.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988. ISBN 0-13-022278-X.

JOACHIMS, T. Text categorization with support vector machines: learning with many relevant features. In: NÉDELLEC, C.; ROUVEIROL, C. (Ed.). **Proceedings of ECML-98, 10th European Conference on Machine Learning**. Chemnitz, DE: Springer Verlag, Heidelberg, DE, 1998. p. 137–142.

KAN, M. Y. Web page classification without the web page. In: **Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters**. New York, NY, USA: ACM, 2004. (WWW Alt. '04), p. 262–263. ISBN 1-58113-912-8.

KAN, M. Y.; THI, H. O. N. Fast webpage classification using url features. In: **Proceedings of the 14th ACM international conference on Information and knowledge management**. New York, NY, USA: ACM, 2005. (CIKM '05), p. 325–326. ISBN 1-59593-140-6.

KIM, Y.; NAM, T. An efficient text filter for adult web documents. In: **Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference**. 2006. v. 1, p. 3 pp. –440.

KOHONEN, T. Self-organized formation of topologically correct feature maps. **Biological Cybernetics**, Springer Berlin / Heidelberg, v. 43, n. 1, p. 59–69, jan. 1982. ISSN 0340-1200.

LEE, L.-H.; LUH, C.-J. Generation of pornographic blacklist and its incremental update using an inverse chi-square based method. **Information Processing and Management**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 44, n. 5, p. 1698–1706, set. 2008. ISSN 0306-4573.

LEE, P.; HUI, S.; FONG, A. An intelligent categorization engine for bilingual web content filtering. **IEEE Transactions on Multimedia**, v. 7, n. 6, p. 1183 – 1190, dec. 2005. ISSN 1520-9210.

LING, X.; XUE, G. R.; DAI, W.; JIANG, Y.; YANG, Q.; YU, Y. Can chinese web pages be classified with english data source? In: **WWW '08: Proceeding of the 17th international conference on World Wide Web**. New York, NY, USA: ACM, 2008. p. 969–978. ISBN 978-1-60558-085-2.

MCCALLUM, A.; NIGAM, K. A comparison of event models for Naive Bayes text classification. In: **AAAI-98 Workshop on Learning for Text Categorization**. 1998. p. 41–48.

MENCZER, F. Mapping the semantics of web text and links. **Internet Computing, IEEE**, v. 9, n. 3, p. 27–36, 2005. ISSN 1089-7801.

MINSKY, M. L.; PAPERT, S. A. **Perceptrons**The MIT Press, 1969. Paperback. ISBN 0262631113.

MITCHELL, T. M. **Machine Learning**. 1nd. ed.McGraw Hill, 1997. ISBN 0070428077.

PAL, S.; MITRA, S. Multilayer perceptron, fuzzy sets, and classification. **Neural Networks, IEEE Transactions on**, v. 3, n. 5, p. 683–697, 1992. ISSN 1045-9227.

PAWLAK, Z. Rough sets. **International Journal of Computer and Information Science**, v. 11, p. 341–356, 1982.

PAWLAK, Z. Rough sets and intelligent data analysis. **Information Sciences - Informatics and Computer Science**, Elsevier Science Inc., New York, NY, USA, v. 147, n. 1-4, p. 1–12, out. 2002. ISSN 0020-0255.

PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, v. 2, n. 6, p. 559–572, 1901.

PIERRE, J. Practical issues for automated categorization of web sites. In: **Electronic Proc. ECDL 2000 Workshop on the Semantic Web**. 2000.

PIERRE, J. M. On the automated classification of web sites. **Computing Research Repository (CoRR)**, cs.IR/0102002, 2001.

PLATT, J. C. Using Analytic QP and Sparseness to Speed Training of Support Vector Machines. In: **In Neural Information Processing Systems 11**. 1999. v. 11, p. 557–563.

POLPINIJ, J.; CHOTTHANOM, A.; SIBUNRUANG, C.; CHAMCHONG, R.; PUANGPRONPITAG, S. Content-based text classifiers for pornographic web filtering. In: **IEEE International Conference on Systems, Man and Cybernetics, 2006. SMC '06.** 2006. v. 2, p. 1481 –1485.

PORTER, M. F. An algorithm for suffix stripping. **Program**, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, v. 14, n. 3, p. 130–137, 1980. ISSN 0033-0337.

QI, X.; DAVISON, B. D. Web page classification: Features and algorithms. **ACM Computing Surveys**, ACM, New York, NY, USA, v. 41, p. 12:1–12:31, February 2009. ISSN 0360-0300.

QUINLAN, J. R. Induction of Decision Trees. **Machine Learning**, Springer Netherlands, v. 1, n. 1, p. 81–106, mar. 1986. ISSN 08856125.

QUINLAN, R. J. **C4.5: programs for machine learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0.

RIBAS, O. T.; KAESTNER, C. A. A. Classificao automtica de stios web usando anlise textual. In: **IV International Workshop on Web and Text Intelligence - WTI**. Curitiba, PR: Anais do WTI 2012, 2012.

RIBONI, D. Feature Selection for Web Page Classification. In: **In EURASIA-ICT 2002 Proceedings of the Workshop**. 2002.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing e Management**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 24, n. 5, p. 513–523, ago. 1988. ISSN 0306-4573.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. **Communications of the ACM**, ACM, New York, NY, USA, v. 18, n. 11, p. 613–620, nov. 1975. ISSN 0001-0782.

SAM, L. Z.; MAAROF, M. bin; SELAMAT, A.; SHAMSUDDIN, S. Features extraction for illicit web pages identifications using independent component analysis. In: **Intelligent and Advanced Systems, 2007. ICIAS 2007. International Conference on**. 2007. p. 139–144.

SANTOS, I.; GALáN-GARCíA, P.; SANTAMARíA-IBIRIKA, A.; ALONSO-ISLA, B.; ALABAU-SARASOLA, I.; BRINGAS, P. Adult content filtering through compression-based text classification. In: HERRERO, Á.; SNá?EL, V.; ABRAHAM, A.; ZELINKA, I.; BARUQUE, B.; QUINTIáN, H.; CALVO, J. L.; SEDANO, J.; CORCHADO, E. (Ed.). **International**

Joint Conference CISIS?12-ICEUTE´12-SOCO´12 Special Sessions Springer Berlin Heidelberg, 2013, (Advances in Intelligent Systems and Computing, v. 189). p. 281–288. ISBN 978-3-642-33017-9.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM Computing Surveys**, ACM, New York, NY, USA, v. 34, n. 1, p. 1–47, mar. 2002. ISSN 0360-0300.

SHEN, D.; CONG, Y.; SUN, J.-T.; LU, Y.-C. Studies on chinese web page classification. In: **Machine Learning and Cybernetics, 2003 International Conference on**. 2003. v. 1, p. 23–27 Vol.1.

SHEN, D.; SUN, J.-T.; YANG, Q.; CHEN, Z. A comparison of implicit and explicit links for web page classification. In: **Proceedings of the 15th international conference on World Wide Web**. New York, NY, USA: ACM, 2006. (WWW '06), p. 643–650. ISBN 1-59593-323-9.

SOARES, M.; PRATI, R.; MONARD, M. Pretext ii - descrição da reestruturação da ferramenta de pré-processamento de textos. **ICMC-USP - Technical Report**, v. 7, n. 4, p. 472 –477, aug. 2008. ISSN 1548-0992.

SOARES, M.; PRATI, R.; MONARD, M. Wci 02 improvements on the porter's stemming algorithm for portuguese. **Latin America Transactions, IEEE (Revista IEEE America Latina)**, v. 7, n. 4, p. 472 –477, aug. 2009. ISSN 1548-0992.

SYMANTEC. **Internet Security Threat Report, Volume 16 (http://www.symantec.com/business/threatreport/) - Acesso em 01/11/2011**. 2010.

W3C. **HTML 4.01 Specification - http://www.w3.org/TR/1999/REC-html401-19991224/**. 1999.

WEITZNER, D. Free speech and child protection on the web. **Internet Computing, IEEE**, v. 11, n. 3, p. 86 –89, may-june 2007. ISSN 1089-7801.

WORLDWIDEWEBSIZE. **The size of the World Wide Web (http://www.worldwidewebsize.com/) - Acesso em 08/12/2012**. 2012.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; YU, P. S.; ZHOU, Z.-H.; STEINBACH, M.; HAND, D. J.; STEINBERG, D. Top 10 algorithms in data mining. **Knowledge and Information Systems**, Springer London, New York, NY, USA, v. 14, n. 1, p. 1–37, jan. 2008. ISSN 0219-1377.

WU, Y.; SHE, K.; ZHU, W.; YUE, X.; LUO, H. A web text filter based on rough set weighted bayesian. In: **Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009. DASC '09.** 2009. p. 241 –245.

YANG, Y. An Evaluation of Statistical Approaches to Text Categorization. **Information Retrieval**, Kluwer Academic Publishers, v. 1, n. 1/2, p. 69–90, 1999.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: FISHER, D. H. (Ed.). **Proceedings of ICML-97, 14th International Conference on Machine Learning**. Nashville, US: Morgan Kaufmann Publishers, San Francisco, US, 1997. p. 412–420.

YERGEAU, F. Rfc 3629: Utf-8, a transformation format of iso 10646. **Internet Engineering Task Force (IETF)**, 2003.

ZHONG, P.; CHANG, Y.; XIAO, Q. Research on automatic construction of the dynamic chinese porn feature knowledge system. In: **Computer Design and Applications (ICCDA), 2010 International Conference on**. 2010. v. 4, p. V4–344 –V4–348.

ZHOU, Y.; REID, E.; QIN, J.; CHEN, H.; LAI, G. Us domestic extremist groups on the web: link and content analysis. **Intelligent Systems, IEEE**, v. 20, n. 5, p. 44 – 51, sept.-oct. 2005. ISSN 1541-1672.

APÊNDICE A – AMBIENTE

Conforme diagnosticado na sessão 3.1 é necessário que exista um ambiente de rede de computadores que proporcione a implantação do método de classificação proposto neste trabalho. O ambiente de rede sugerido para implantação do mecanismo de classificação proposto pode ser visualizado na Figura 38. Nesse ambiente existem três elementos que são essenciais, são eles: um servidor de *proxy*, um servidor de avisos e um servidor de filtro de conteúdo.

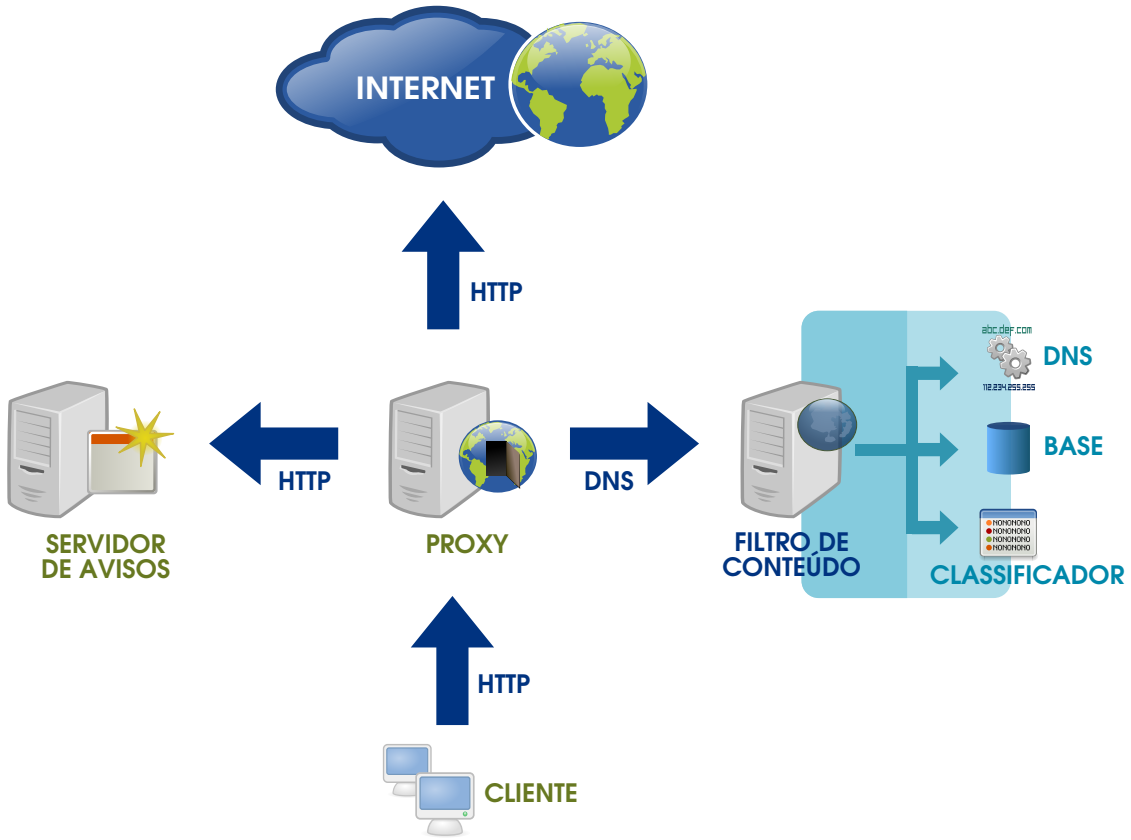

Figura 38: Ambiente para Implantação do Classificador

O servidor *proxy* é o responsável por receber todas as requisições de páginas *web* solicitadas pelos clientes (estações dos usuários), o *proxy* recebe as requisições via protocolo HTTP e encaminha as respostas aos solicitantes. O filtro de conteúdo analisa as requisições,

verificando se o sítio que o cliente deseja acessar pertence a uma categoria de acesso permitido. A comunicação entre o servidor *proxy* e o filtro de conteúdo ocorre via protocolo DNS (*Domain Name System*), o filtro de conteúdo possui uma base de dados de sites já classificados, caso o site solicitado ainda não tenha sido classificado é executado o classificador. Por último tem-se o servidor de avisos que trata-se de um servidor HTTP que tem por objetivo fornecer uma página HTML que informa o usuário que o acesso a determinado sítio não é permitido. O servidor de avisos comunica-se com o *proxy* via protocolo HTTP.

O *proxy* é o elemento central do ambiente, responsável por comunicar-se com os demais elementos e por intermediar o acesso do usuário ao sítio da internet caso o acesso seja autorizado. Já o filtro de conteúdo é responsável pela execução do algoritmo de aprendizagem de máquina (classificador), pela atualização e consulta a base de sites classificados e também por realizar a resolução de nomes utilizando o protocolo DNS.

O filtro de conteúdo ao receber, do *proxy*, uma requisição de resolução de nome (DNS) para um site pertencente a uma categoria permitida fornece ao *proxy* o IP (*Internet Protocol*) verdadeiro do sítio. Caso a solicitação seja para um site de categoria não autorizada o filtro de conteúdo retorna o IP do servidor de avisos, o qual irá envia uma página HTML com uma mensagem advertindo o usuário que não é permitido o acesso ao sítio solicitado.

A Figura 39 apresenta o diagrama de atividades do ambiente de rede, mostrando como ocorre a interação entre os diferentes componentes do ambiente.

Conforme o diagrama no caso de um sítio não ter sido classificado é autorizado o seu acesso até que ocorra a sua classificação, uma vez classificado o acesso somente será autorizado caso a categoria a qual pertence o sítio esteja autorizada. Isso é possível através da manipulação do campo TTL (*Time to Live*) da resposta do DNS. O uso desse mecanismo permite o usuário acessar o sítio até que ocorra a classificação, sem perceber a ação do filtro de conteúdo.

Ainda é possível mudar o comportamento do filtro de conteúdo negando todos os acessos a sítios ainda não categorizados. Enquanto o usuário espera que ocorra a classificação do sítio é possível, por exemplo, enviar uma página HTML informando-o que esta sendo executado a classificação e solicitar que seja aguardado um tempo pré-determinado. A política de acesso a sites ainda não categorizadas pode ser definida pela instituição e implementada no filtro de conteúdo, não sendo necessário mudanças no ambiente de rede.

O ambiente sugerido viabiliza o uso do mecanismos de classificação automática de sites, possibilitando a sua implantação desde de uma rede pequena até uma rede com milhares de usuários simultâneos. A sua implantação ocorre com apenas três componentes. Esses com-

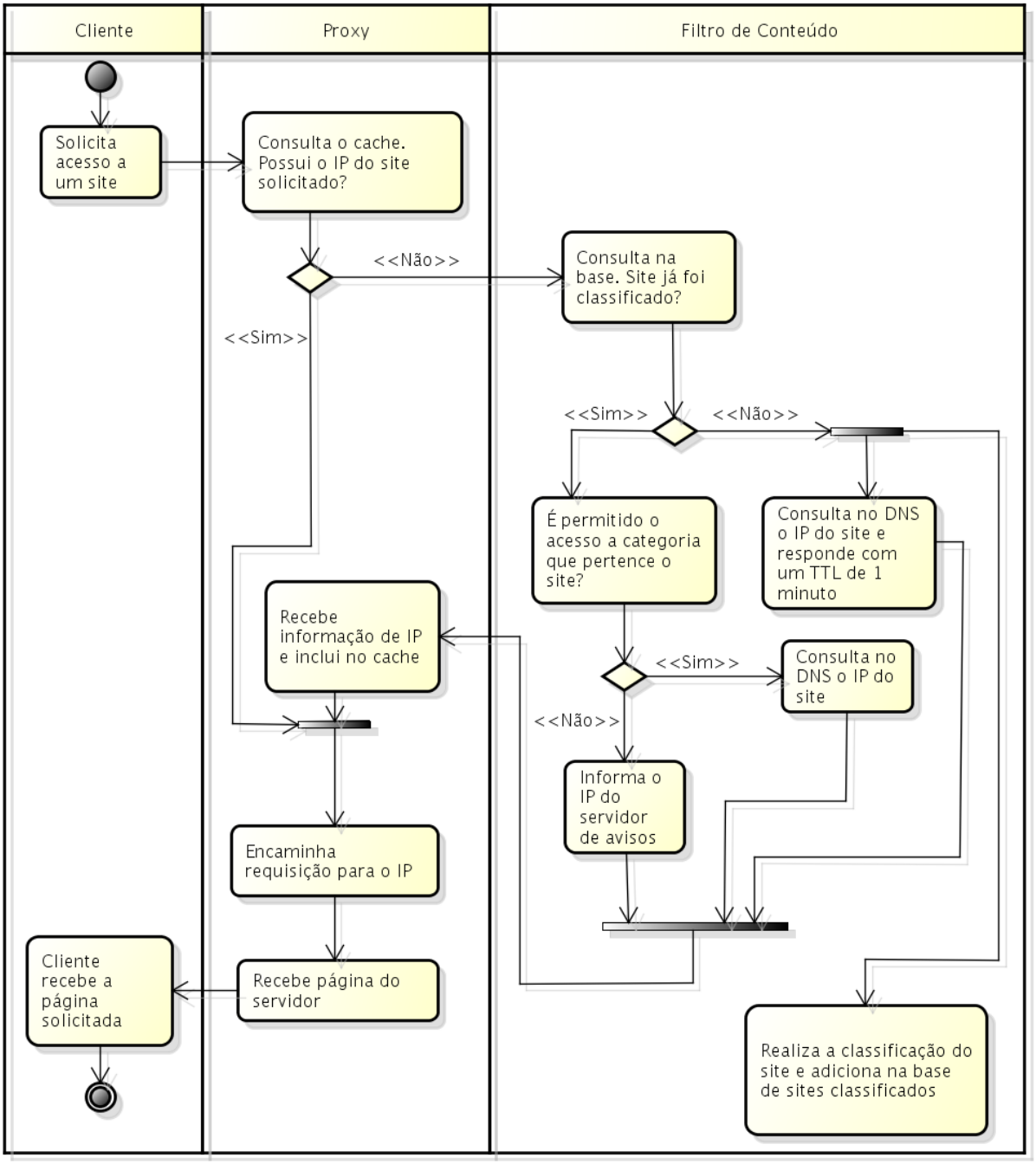

Figura 39: Diagrama de Atividades do Ambiente

ponentes podem estar localizados no mesmo servidor no caso de uma rede pequena ou então em diversas máquinas no caso de uma rede de grande porte. Isso torna a arquitetura de rede escalável pois novos componentes podem ser adicionados de acordo com a demanda. O ambiente pode ser dimensionado de acordo com o volume de acesso dos usuários.

Além disso, o ambiente sugerido também cria um ponto central de controle de acesso. É possível criar políticas que proíbam ou liberem o acesso a uma determinada categoria de sites, por exemplo sites de conteúdo adulto. Essa política pode ser aplicada para toda a rede

sem que seja necessário alterações nas estações dos usuários ou mesmo instalação de softwares adicionais.

A Figura 40 mostra as etapas que envolvem a classificação de uma página. Ao todo são seis etapas e cada uma delas será brevemente descritas abaixo:

- (A) *Download* da página: nessa etapa é realizado o *download* da página a ser classificada, dependendo do *link* de rede essa etapa é a que consome o maior tempo para ser realizada. Nessa etapa também é identificado a codificação de caracteres utilizada pela página.
- (B) Identificar marcadores HTML: nesse momento é identificado o texto pertencente a cada marcador da linguagem HTML utilizada pela página.
- (C) Determinar idioma: a determinação do idioma é feita usando a técnica de contagem de *stopword*. Com base nessa técnica será atribuido ao documento o idiomar que tiver o maior número de *stopword* presentes.
- (D) Pré-processamento: nesse momento é executado todas as etapas do pré-processamento apresentadas na sessão 2.4.
- (E) Extração de atributos: nessa etapa são extraídos os atributos que serão utilizados pelo classificador.
- (F) Classificação: nessa etapa é executado o algoritmo de aprendizagem de máquina usando os atributos extraídos pela etapa anterior. Ao final desta etapa o classificador atribui um rótulo a página.

A Figura 40 apresenta as etapas da classificação de uma página, enquanto a Figura 41 mostra como ocorre a predição de uma página nova, bem como o processo de treinamento do classificador. Conforme pode-se ver na Figura são dois processos distintos: o treinamento do classificador e a predição de uma página de classe desconhecida.

O treinamento é a primeira etapa a ser concluída, anterior a predição, ocorre em um conjunto de dados previamente pré-processados e rotulados. O tempo de treinamento depende de vários fatores como o algoritmo de aprendizagem (classificador), quantidade de atributos e total de instâncias. Como produtos do treinamento temos o modelo de seleção de atributos e o modelo classificador, ambos serão utilizados na etapa de predição.

A predição de uma página é um processo posterior ao treinamento. O tempo gasto com a predição é extremamente baixo se comparado ao tempo necessário para o treinamento. Em

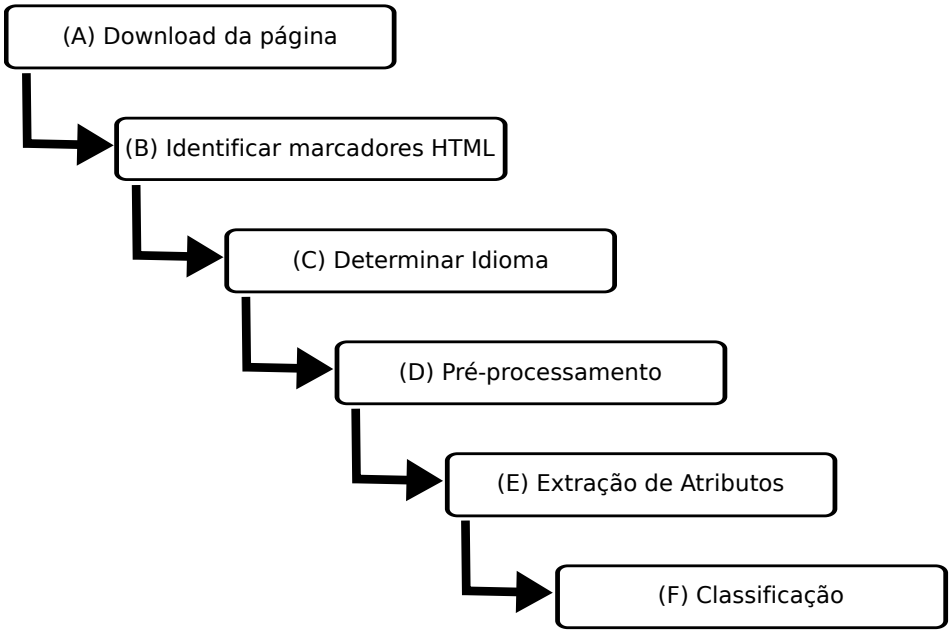

Figura 40: Etapas Envolvendo a Classificação

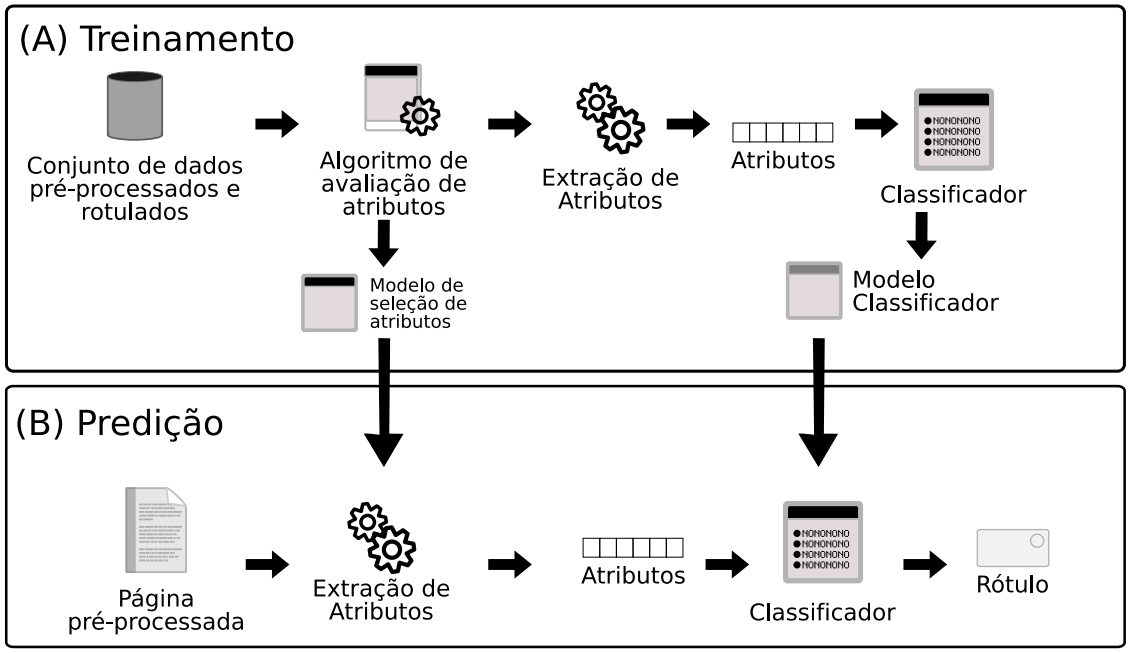

Figura 41: Predição de uma Nova Página

geral o processo de predição ocorre em milissegundos. Nesta etapa são utilizados os modelos produzidos durante o treinamento e no final do processo é designado um rotulo a instância que esta sendo classificada.

APÊNDICE B – DESEMPENHO DOS ALGORITMOS DE CLASSIFICAÇÃO

Neste apêndice estão os resultados dos experimentos dos algoritmos de aprendizagem de máquina com diferentes atributos estruturais. Em todos os experimentos foi utilizado a métrica de avaliação de qualidade de atributos TF (*Term Frequency*), essa métrica foi escolhida devido aos bons resultados apresentados nos experimentos anteriores (sessão 5.1).

Os atributos estruturais avaliados nos experimentos estão melhores descritos na sessão 5.4. A notação utilizada para simbolizar os atributos nas Tabelas deste apêndice também é a mesma da utilizada na sessão 5.4.

As Tabelas deste apêndice utilizam a seguinte ordem das informações: primeira coluna informa o total de atributos selecionados para o experimento; segunda coluna mostra os resultados quando o classificador utilizou apenas informações textuais desconsiderando aquelas informações presentes nos atributos estruturais; coluna 3 até coluna 9 apresenta os resultados quando utilizado conjunto de marcados, composto por informações de vários marcadores; coluna 10 até 16 contêm os resultados quando é utilizado marcadores individuais.

Os resultados apresentados nas Tabelas deste apêndice também estão dispostos em forma de gráficos, apresentados na sequência das Tabelas. Cada um dos Gráficos mostra informações de um experimento realizado para um determinado algoritmo de aprendizagem de máquina, idioma e grupo de quatro ou cinco atributos ou conjunto de atributos estruturais. Os resultados apresentados são sempre comparados com o resultado do classificador base, o qual não utilizou informações dos atributos estruturais.

Tabela 20: Experimento com Atributos Estruturais na Base em Português Utilizando o Classificador Naïve Bayes

Qtde Atributos	Texto	Conjunto de Marcadores							Marcadores						
		A	B	C	D	E	F	G	1	2	3	4	5	6	7
10	0.9189	**0.9298**	0.9299	0.7612	0.9469	0.9559	0.9479	0.9479	0.9028	0.7288	0.7133	0.9179	0.8242	0.8537	0.7785
20	0.9439	0.9115	0.9439	0.7805	0.9579	0.9579	0.9489	0.9519	0.9178	0.7361	0.7389	0.9369	0.8344	0.8576	0.8171
30	0.9579	0.9124	0.9449	0.7889	0.9609	0.9629	0.9539	0.9549	0.9188	0.7500	0.7454	0.9429	0.8401	0.8597	0.8291
40	0.9649	0.9134	0.9499	0.7999	0.9679	0.9629	0.9559	0.9639	0.9178	0.7543	0.7498	0.9408	0.8520	0.8641	0.8337
50	0.9679	0.9216	0.9569	0.7955	0.9679	0.9639	0.9589	0.9679	0.9208	0.7578	0.7528	0.9469	0.8518	0.8681	0.8356
60	0.9699	0.9216	0.9599	0.8003	0.9709	0.9669	0.9669	0.9659	0.9228	0.7691	0.7540	0.9529	0.8548	0.8679	0.8354
70	0.9749	0.9195	0.9609	0.7939	0.9759	0.9649	0.9639	0.9749	0.9238	0.7695	0.7531	0.9559	0.8333	0.8700	0.8386
80	0.9749	0.9195	0.9659	0.7983	0.9749	0.9679	0.9619	0.9759	**0.9268**	0.7712	0.7531	0.9589	0.8436	0.8700	0.8379
90	0.9809	0.9195	0.9689	0.8106	0.9779	0.9689	0.9679	0.9719	**0.9268**	**0.7723**	0.7563	0.9559	0.8358	0.8720	0.8378
100	0.9799	0.9195	0.9689	0.8095	0.9829	0.9719	0.9699	0.9739	**0.9268**	**0.7723**	0.7572	0.9619	0.8339	0.8720	0.8357
110	0.9839	0.9205	0.9709	0.8095	0.9869	0.9759	0.9749	0.9769	**0.9268**	**0.7723**	0.7572	0.9659	0.8361	0.8720	0.8441
120	0.9859	0.9185	0.9739	0.8066	0.9869	0.9739	0.9789	0.9769	**0.9268**	**0.7723**	0.7572	0.9669	0.8382	0.8710	0.8452
130	0.9829	0.9195	0.9729	0.8057	0.9869	0.9789	0.9789	0.9809	**0.9268**	**0.7723**	0.7648	0.9659	0.8380	0.8700	0.8463
140	0.9849	0.9205	0.9729	0.8107	0.9849	0.9799	0.9809	0.9809	**0.9268**	**0.7723**	0.7648	0.9749	0.8390	0.8709	0.8424
150	0.9839	0.9215	0.9739	0.8096	0.9849	0.9809	0.9809	0.9829	**0.9268**	**0.7723**	**0.7659**	0.9749	0.8590	0.8741	0.8414
160	0.9839	0.9215	0.9749	0.8107	0.9869	0.9829	0.9809	0.9809	**0.9268**	**0.7723**	**0.7659**	0.9759	0.8643	**0.8751**	0.8444
170	0.9829	0.9215	**0.9769**	0.8087	0.9859	**0.9869**	0.9809	**0.9839**	**0.9268**	**0.7723**	**0.7659**	0.9759	0.8654	0.8741	0.8444
180	0.9849	0.9215	0.9759	0.8109	0.9859	0.9859	0.9809	0.9809	**0.9268**	**0.7723**	**0.7659**	**0.9769**	**0.8673**	0.8709	0.8444
190	**0.9869**	0.9215	0.9759	0.8109	**0.9879**	0.9859	0.9819	0.9829	**0.9268**	**0.7723**	**0.7659**	**0.9769**	0.8610	0.8719	0.8464
200	0.9849	0.9215	0.9759	**0.8118**	**0.9879**	0.9849	**0.9839**	0.9819	**0.9268**	**0.7723**	**0.7659**	**0.9769**	0.8610	0.8719	**0.8495**

Tabela 21: Experimento com Atributos Estruturais na Base em Inglês Utilizando o Classificador Naïve Bayes

Qtde Atributos	Texto	Conjunto de Marcadores							Marcadores						
		A	B	C	D	E	F	G	1	2	3	4	5	6	7
10	0.8371	0.8821	0.8454	0.5827	0.8636	0.8778	0.8496	0.8777	0.8087	0.8198	0.7649	0.8073	**0.7421**	0.7284	0.6675
20	0.8660	0.8982	0.8636	0.5827	0.9047	0.9066	0.8782	0.8965	0.8122	0.8167	0.7813	0.8452	0.7261	**0.7361**	0.6945
30	0.8900	0.8870	0.8785	0.6165	0.9206	0.9206	0.8932	0.9063	0.8173	0.8295	0.7894	0.8606	0.7260	0.7321	0.7098
40	0.8962	0.9044	0.8713	0.6309	0.9227	0.9185	0.9083	0.9114	0.8193	0.8315	0.7966	0.8701	0.7271	0.7316	0.7216
50	0.9043	0.9053	0.8763	0.6484	0.9286	0.9236	0.9062	0.9103	0.8193	0.8441	0.8070	0.8701	0.7246	0.7221	0.7199
60	0.9103	0.9114	0.8783	0.6482	0.9297	0.9225	0.9052	0.9113	0.8193	0.8461	0.8070	0.8754	0.7204	0.7220	0.7194
70	0.9083	0.9053	0.8814	0.6439	0.9307	0.9225	0.9062	0.9103	0.8193	0.8471	0.8081	0.8690	0.7201	0.7250	0.7208
80	0.9073	0.9114	0.8763	0.6600	0.9337	0.9215	0.9063	0.9093	0.8193	0.8482	**0.8091**	0.8740	0.7053	0.7221	0.7181
90	0.9073	0.9104	0.8773	0.6520	0.9317	0.9215	0.9011	0.9113	0.8193	0.8501	**0.8091**	0.8751	0.7025	0.7234	0.7232
100	0.9093	0.9135	0.8794	0.6655	0.9357	0.9236	0.9031	0.9154	0.8193	0.8511	**0.8091**	0.8782	0.7001	0.7208	0.7238
110	0.9042	0.9125	0.8803	0.6728	0.9357	0.9276	0.9031	0.9154	0.8193	0.8511	**0.8091**	0.8782	0.7025	0.7192	0.7217
120	0.9042	0.9155	0.8793	**0.6956**	0.9367	0.9317	0.9083	0.9144	0.8203	**0.8520**	0.8068	0.8762	0.7025	0.7225	0.7128
130	0.9093	0.9165	0.8814	0.6922	**0.9388**	**0.9327**	0.9093	0.9144	0.8203	0.8510	0.8068	0.8783	0.7040	0.7278	0.7152
140	0.9114	0.9165	0.8814	0.6908	0.9367	**0.9327**	0.9103	0.9164	0.8203	0.8510	0.8079	0.8824	0.7086	0.7244	0.7183
150	0.9114	0.9165	0.8814	0.6936	0.9367	0.9317	**0.9113**	0.9144	0.8203	0.8510	0.8079	0.8824	0.7053	0.7262	0.7165
160	0.9134	0.9175	0.8834	0.6906	0.9337	0.9307	0.9103	0.9123	0.8203	**0.8520**	0.8079	0.8793	0.7037	0.7137	0.7132
170	0.9144	**0.9185**	**0.8844**	0.6933	0.9367	0.9306	0.9092	0.9133	0.8203	**0.8520**	0.8079	0.8814	0.7047	0.7158	0.7241
180	0.9134	**0.9185**	0.8803	0.6910	0.9377	0.9276	0.9092	0.9164	0.8203	**0.8520**	0.8079	0.8845	0.6968	0.7167	0.7241
190	**0.9154**	**0.9185**	0.8814	0.6919	**0.9388**	0.9276	0.9103	0.9184	**0.8213**	**0.8520**	0.8079	**0.8866**	0.6946	0.7140	0.7250
200	0.9144	**0.9185**	0.8824	0.6917	**0.9388**	0.9296	0.9103	**0.9195**	**0.8213**	**0.8520**	0.8079	**0.8866**	0.6940	0.7201	**0.7279**

Tabela 22: Experimento com Atributos Estruturais na Base em Português Utilizando o Classificador KNN

Qtde Atributos	Texto	Conjunto de Marcadores							Marcadores						
		A	B	C	D	E	F	G	1	2	3	4	5	6	7
10	0.9559	0.9418	0.9429	0.7630	0.9669	0.9679	0.9579	0.9619	0.9176	0.7636	0.7434	0.9439	0.8429	0.8644	0.7879
20	0.9799	0.9599	0.9829	0.7832	0.9849	0.9919	0.9859	0.9909	0.9509	0.7590	0.7515	0.9819	0.8544	0.8776	0.8368
30	0.9769	**0.9649**	0.9899	0.8060	0.9879	0.992	0.9859	**0.9969**	0.9529	0.7668	0.7608	0.9889	0.8553	0.8837	0.8425
40	0.9769	0.9639	0.9869	0.8045	0.9929	0.9939	0.9909	**0.9969**	0.9519	0.7677	0.7632	0.9869	0.8584	0.9104	0.8462
50	**0.9909**	0.9559	0.9899	0.8116	0.9859	**0.9969**	0.9919	**0.9969**	0.9549	0.7691	0.7641	0.9899	0.8806	0.9135	0.8543
60	0.9899	0.9559	**0.9939**	0.8109	**0.9939**	**0.9969**	**0.9939**	0.9959	0.9549	0.7647	0.7667	**0.9919**	0.8892	0.9155	0.8597
70	0.9879	0.9619	0.9899	0.8096	0.9929	0.9939	0.9909	**0.9969**	0.9559	0.7628	0.7676	0.9899	0.8901	0.9154	0.8629
80	0.9889	0.9619	0.9889	0.8146	0.9909	0.9929	0.9899	0.9939	0.9559	0.7626	**0.7688**	0.9889	0.8911	0.9155	0.8608
90	0.9869	0.9619	0.9879	0.8336	0.9909	0.9929	0.9899	0.9939	0.9569	0.7649	0.7678	0.9889	0.8961	0.9185	0.8599
100	0.9819	0.9609	0.9859	0.8347	0.9889	0.9939	0.9889	0.9929	0.9569	0.7685	0.7678	0.9879	0.9033	0.9185	0.8582
110	0.9869	0.9619	0.9859	0.8340	0.9899	0.9899	0.9849	0.9899	**0.9579**	0.7683	0.7620	0.9879	0.9023	**0.9205**	0.8603
120	0.9869	0.9639	0.9889	0.8381	0.9919	0.9909	0.9879	0.9899	0.9478	**0.7693**	0.7620	0.9869	0.8974	**0.9205**	0.8582
130	0.9859	0.9639	0.9849	**0.8443**	0.9909	0.9929	0.9899	0.9899	0.9559	0.7298	0.7671	0.9869	0.8975	0.9116	0.8633
140	0.9859	0.9629	0.9879	0.8399	0.9889	0.9929	0.9889	0.9939	0.9499	0.7305	0.7671	0.9879	0.9207	0.9126	0.8652
150	0.9869	0.9619	0.9869	0.8399	0.9879	0.9929	0.9889	0.9919	0.9499	0.7461	0.7671	0.9879	0.9227	0.9126	**0.8704**
160	0.9859	0.9619	0.9869	0.8370	0.9889	0.9929	0.9909	0.9879	0.9509	0.7370	0.7671	0.9889	0.9327	0.9126	0.8693
170	0.9849	0.9589	0.9849	0.8409	0.9889	0.9899	0.9879	0.9879	0.9489	0.7374	0.7671	0.9889	0.9307	0.9155	0.8674
180	0.9879	0.9609	0.9859	0.8380	0.9899	0.9899	0.9879	0.9869	0.9489	0.7385	0.7671	0.9859	0.9307	0.9145	0.8685
190	0.9869	0.9609	0.9849	**0.8443**	0.9909	0.9889	0.9879	0.9899	0.9499	0.7244	0.7671	0.9859	**0.9338**	0.9145	0.8239
200	0.9859	0.9549	0.9849	**0.8443**	0.9899	0.9889	0.9879	0.9899	0.9509	0.7180	0.7671	0.9859	0.9308	0.9135	0.8687

Tabela 23: Experimento com Atributos Estruturais na Base em Inglês Utilizando o Classificador KNN

Qtde Atributos	Texto	Conjunto de Marcadores							Marcadores						
		A	B	C	D	E	F	G	1	2	3	4	5	6	7
10	0.8837	0.8878	0.8847	0.5885	0.9098	0.9189	0.8926	0.9249	0.8054	0.8213	0.7661	0.8547	0.7605	0.7461	0.6714
20	0.8998	0.9165	0.9087	0.6095	0.9349	0.9399	0.9147	0.9529	0.8380	0.8232	0.7977	0.8883	0.7602	0.7719	0.6963
30	0.9329	0.9146	0.9068	0.6327	0.9609	0.9489	0.9189	0.9479	0.8596	0.8422	0.8083	0.9228	0.7742	0.7827	0.7110
40	0.9409	0.9195	0.9308	0.6472	0.9649	0.9639	0.9409	0.9629	0.8588	0.8429	0.8102	0.9309	0.7707	0.7962	0.7086
50	0.9539	0.9196	0.9459	0.6709	0.9589	0.9669	0.9469	0.9759	0.8639	0.8412	0.8147	0.9539	0.7627	0.8030	0.7442
60	0.9669	0.9236	0.9609	0.6687	0.9729	0.9709	0.9559	0.9779	0.8589	0.8491	0.8167	0.9679	0.7664	0.8008	0.7440
70	0.9679	0.9236	0.9709	0.6837	0.9759	0.9799	0.9749	0.9839	0.8578	0.8522	0.8177	0.9709	0.7690	0.8075	0.7530
80	0.9719	0.9277	0.9729	0.6966	0.9789	0.9809	0.9719	0.9809	0.8600	0.8573	0.8177	0.9729	0.7815	0.8050	0.7625
90	0.9749	0.9349	0.9729	0.7019	0.9829	0.9849	0.9799	0.9829	0.8613	**0.8592**	0.8177	0.9719	0.7736	0.8070	0.7685
100	0.9789	0.9309	0.9739	0.7078	0.9839	0.9839	0.9819	0.9819	0.8704	0.8517	0.8199	0.9749	0.7660	0.8028	0.7862
110	0.9769	0.9358	0.9729	0.7115	0.9859	0.9829	0.9789	0.9829	0.8695	0.8518	0.8188	0.9749	0.7648	0.8047	0.7886
120	0.9789	0.9368	0.9739	0.7359	0.9859	0.9839	0.9809	0.9849	0.8770	0.8506	0.8209	0.9749	0.7885	0.8050	0.7951
130	0.9789	0.9328	0.9829	0.7290	0.9829	0.9869	0.9799	0.9859	0.8750	0.8496	0.8209	0.9769	0.7829	0.8047	0.8069
140	0.9789	0.9308	0.9839	0.7365	0.9849	**0.9879**	0.9789	0.9879	0.8802	0.8475	0.8201	0.9739	0.7852	0.8088	0.8111
150	0.9809	0.9328	0.9819	0.7401	**0.9879**	0.9859	0.9819	0.9859	0.8801	0.8517	0.8211	0.9749	0.7910	0.8076	0.8138
160	**0.9829**	0.9328	0.9819	0.7406	0.9859	0.9859	0.9809	0.9889	0.8802	0.8498	0.8221	0.9739	**0.7914**	0.8064	0.8107
170	0.9779	0.9308	0.9829	**0.7429**	0.9829	0.9859	0.9809	**0.9899**	0.8822	0.8498	0.8232	0.9769	0.7855	0.8085	0.8167
180	0.9779	0.9308	**0.9849**	0.7408	0.9829	0.9819	0.9809	0.9889	0.8884	0.8447	0.8232	0.9849	0.7856	0.8143	0.8352
190	0.9789	0.9358	**0.9849**	0.7389	0.9829	0.9839	0.9819	0.9889	0.8843	0.8495	0.8232	0.984	0.7814	0.8133	0.8407
200	0.9769	**0.9378**	0.9799	0.7391	0.9849	0.9849	**0.9849**	0.9889	**0.8894**	0.8485	**0.8233**	**0.9869**	0.7807	**0.8165**	**0.8473**

Tabela 24: Experimento com Atributos Estruturais na Base em Português Utilizando o Classificador C4.5

Qtde Atributos	Texto	Conjunto de Marcadores							Marcadores						
		A	B	C	D	E	F	G	1	2	3	4	5	6	7
10	0.9439	0.9053	0.9369	0.7630	0.9629	0.9589	0.9499	0.9589	**0.8928**	0.7421	0.7173	0.9359	0.8232	0.8477	0.7961
20	0.9489	0.9092	0.9429	0.7635	0.9659	0.9839	0.9559	0.9629	**0.8928**	0.7541	0.7264	0.9449	0.8286	0.8580	0.8073
30	0.9519	0.8996	0.9509	0.7616	0.9679	0.9699	0.9619	0.9759	**0.8928**	0.7447	0.7264	0.9539	0.8286	0.8650	0.8143
40	0.9549	0.9057	0.9559	0.7662	0.9779	0.9879	0.9569	0.9759	**0.8928**	0.7447	0.7264	0.9669	0.8232	0.8941	**0.8299**
50	0.9579	0.9146	0.9509	0.7719	0.9729	0.9819	0.9659	0.9749	**0.8928**	0.7519	0.7273	0.9659	0.8232	0.8872	0.8231
60	0.9689	**0.9186**	0.9649	0.7710	0.9759	0.9699	0.9549	0.9749	**0.8928**	**0.7594**	0.7273	0.9859	0.8297	0.8872	0.8223
70	0.9499	**0.9186**	0.9649	0.7701	0.9829	0.9819	0.9589	0.9779	**0.8928**	**0.7594**	0.7273	0.9859	0.8321	0.8871	0.8241
80	0.9499	**0.9186**	**0.9699**	**0.7745**	0.9839	0.9829	0.9589	**0.9859**	**0.8928**	**0.7594**	0.7273	0.9849	0.8321	0.8841	0.8241
90	0.9599	**0.9186**	0.968	0.7682	0.9789	0.9829	0.9589	0.9749	**0.8928**	**0.7594**	0.7273	0.9809	0.8374	0.8870	0.8241
100	0.9599	**0.9186**	0.9619	0.7682	0.9739	0.9809	0.9589	0.9749	**0.8928**	**0.7594**	0.7273	0.9839	0.8496	**0.8973**	0.8166
110	**0.9699**	**0.9186**	0.9619	0.7682	0.988	0.9809	0.9589	0.9749	**0.8928**	**0.7594**	0.7273	0.9799	0.8496	0.8962	0.8175
120	0.9679	**0.9186**	0.9589	0.7738	**0.9899**	**0.9889**	0.9649	0.9839	**0.8928**	**0.7594**	0.7273	0.9799	0.8496	0.8962	0.8175
130	0.9679	**0.9186**	0.9689	0.7738	**0.9899**	0.992	**0.9699**	0.9829	**0.8928**	**0.7594**	**0.7347**	0.9799	0.8496	0.8962	0.8182
140	0.9599	**0.9186**	0.976	0.7738	0.988	0.992	**0.9699**	0.992	**0.8928**	**0.7594**	**0.7347**	0.9769	**0.8687**	0.8972	0.8182
150	0.9669	**0.9186**	0.976	0.7738	0.988	0.992	**0.9699**	0.992	**0.8928**	**0.7594**	**0.7347**	0.9769	0.8677	0.8972	0.8182
160	0.9669	**0.9186**	0.9679	0.7738	0.9859	0.992	**0.9699**	0.992	**0.8928**	**0.7594**	**0.7347**	0.9779	0.8677	0.8972	0.8182
170	0.9669	**0.9186**	0.9679	0.7738	0.9859	0.9779	0.9649	0.992	**0.8928**	**0.7594**	**0.7347**	0.9779	0.8677	0.8972	0.8182
180	0.9569	**0.9186**	0.968	0.7738	0.9859	0.9779	0.9649	0.992	**0.8928**	**0.7594**	**0.7347**	0.9779	0.8677	0.8972	0.8182
190	0.9569	**0.9186**	0.968	0.7738	0.9859	0.9779	0.9519	0.992	**0.8928**	**0.7594**	**0.7347**	**0.9869**	0.8677	0.8972	0.8182
200	0.9559	**0.9186**	0.968	0.7738	0.9759	0.9779	0.9519	0.992	**0.8928**	**0.7594**	**0.7347**	**0.9869**	0.8677	0.8972	0.8182

Tabela 25: Experimento com Atributos Estruturais na Base em Inglês Utilizando o Classificador C4.5

Qtde Atributos	Texto	Conjunto de Marcadores							Marcadores						
		A	B	C	D	E	F	G	1	2	3	4	5	6	7
10	0.8764	0.8701	0.8912	0.5585	0.9005	0.9096	0.8883	0.9077	0.7896	0.7912	0.7487	0.8531	0.7344	0.7376	0.6492
20	0.8928	0.9002	0.9036	0.5757	0.9198	0.9258	0.9098	0.9268	0.7896	0.7833	0.7717	0.8702	0.7594	0.7580	0.6562
30	0.9268	0.9066	0.9129	0.6151	0.9539	0.9429	0.9149	0.9389	0.7896	0.8029	0.7696	0.9197	**0.7601**	0.7674	0.6963
40	0.9279	0.9076	0.9229	0.616	0.9499	0.9459	0.9219	0.9559	0.7907	**0.8114**	**0.7744**	0.9309	0.7430	0.7723	0.7058
50	0.9399	0.9076	0.9399	0.6339	0.9419	0.9429	0.9258	0.9559	0.7918	0.8064	**0.7744**	0.9499	0.7430	0.7729	0.6934
60	0.9259	0.9086	0.9459	0.6345	0.9469	0.9439	0.9338	0.9529	**0.7941**	0.8103	**0.7744**	0.9639	0.7430	0.7729	0.7091
70	0.9429	0.9086	0.9389	0.6362	0.9489	0.9459	0.9358	0.9569	**0.7941**	0.8103	**0.7744**	**0.9679**	0.7430	0.7808	0.7091
80	0.9389	0.9086	0.9449	0.6454	0.9499	0.9439	0.9328	0.9639	**0.7941**	0.8103	**0.7744**	0.9649	0.7426	**0.7819**	0.7091
90	0.9258	0.9055	0.9489	0.6454	0.9529	0.9419	0.9349	0.9689	**0.7941**	**0.8114**	**0.7744**	0.9549	0.7430	0.7717	0.7046
100	0.9389	0.9055	0.9529	0.6447	0.9559	0.9459	0.9469	0.9719	**0.7941**	**0.8114**	**0.7744**	0.9619	0.7440	0.7717	0.7164
110	0.9509	**0.9117**	0.9429	0.6425	0.9559	0.9559	0.9439	0.9739	**0.7941**	**0.8114**	**0.7744**	0.9619	0.7575	0.7727	0.7175
120	0.9529	**0.9117**	0.9269	0.6460	0.9579	0.9599	0.9549	**0.9759**	**0.7941**	**0.8114**	**0.7744**	0.9619	0.7579	0.7728	0.7175
130	0.9519	**0.9117**	0.9569	0.6460	0.9579	0.9599	0.9549	**0.9759**	0.7884	**0.8114**	**0.7744**	0.9609	0.7579	0.7728	0.7216
140	0.9429	**0.9117**	0.9519	0.6460	0.9569	**0.9629**	0.9489	0.9749	0.7884	**0.8114**	**0.7744**	0.9609	0.7588	0.7728	0.7439
150	0.9329	**0.9117**	0.9519	0.6460	0.9599	**0.9629**	0.9479	0.9749	0.7884	**0.8114**	**0.7744**	0.9609	0.7579	0.7728	0.7454
160	0.9349	**0.9117**	0.9409	**0.6478**	0.9599	**0.9629**	0.9499	0.974	**0.7941**	**0.8114**	**0.7744**	0.9609	0.7579	0.7662	0.7503
170	0.9349	**0.9117**	0.9409	0.6456	0.9599	**0.9629**	0.9499	0.974	0.7884	**0.8114**	**0.7744**	0.9609	0.7587	0.7651	0.7584
180	**0.9549**	**0.9117**	0.9539	0.6456	**0.9619**	0.9619	0.9489	0.9669	0.7884	**0.8114**	**0.7744**	0.9629	0.7575	0.7651	0.7584
190	0.9459	**0.9117**	0.9539	0.6456	0.9569	0.9619	**0.9619**	0.9669	0.7940	**0.8114**	**0.7744**	0.9639	0.7575	0.7690	0.7584
200	0.9459	**0.9117**	**0.9669**	0.6460	0.9569	0.9619	**0.9619**	0.9719	0.7940	**0.8114**	**0.7744**	0.9609	0.7575	0.7700	**0.7619**

Tabela 26: Experimento com Atributos Estruturais na Base em Português Utilizando o Classificador MLP

Qtde Atributos	Texto	Conjunto de Marcadores							Marcadores						
		A	B	C	D	E	F	G	1	2	3	4	5	6	7
10	0.9459	0.9428	0.9449	0.7609	0.9609	0.9629	0.9619	0.9649	0.9186	0.7636	0.7425	0.944	0.8429	0.8634	0.7920
20	0.9689	0.9599	0.9789	0.7917	0.9749	0.9879	0.9719	0.9809	0.9539	0.7617	0.7526	0.9649	0.8554	0.8777	0.8339
30	0.9699	0.9609	0.9809	0.8026	0.9849	0.9869	0.9719	0.9769	0.9549	0.7649	0.7653	0.9749	0.8565	0.8786	0.8477
40	0.9799	0.9629	0.9719	0.8099	0.9789	0.9909	0.9839	0.9849	0.9529	0.7713	0.7699	0.9809	0.8575	0.9134	0.8525
50	0.9829	0.9629	0.9819	0.8147	0.9829	0.9809	0.9799	0.9859	0.9529	0.7784	0.7701	0.9899	0.8764	0.9074	0.8545
60	0.9819	0.9599	0.9909	0.8135	0.9839	0.9789	0.9769	0.99	0.9509	0.7746	0.7722	0.9929	0.8909	0.9134	0.8566
70	0.9849	0.9629	0.9929	0.8100	0.9909	0.9819	0.9809	0.9819	0.9539	0.7748	0.7713	0.9929	0.8920	0.9144	0.8588
80	0.9849	0.9639	0.9929	0.8170	0.9959	0.9899	0.9829	0.9919	0.9559	0.7772	0.7722	0.9929	0.8941	0.9175	0.8557
90	0.9889	0.9639	0.9929	0.8365	0.9929	0.9889	0.9829	0.9929	0.9539	0.7792	0.7701	0.9909	0.8960	0.9186	0.8589
100	0.9869	0.9649	0.9939	0.8333	0.9949	**0.9969**	0.9879	0.9939	0.9539	0.7813	0.7722	**0.9969**	0.9043	0.9185	0.8538
110	0.9939	0.9639	0.9929	0.8345	0.9969	0.9929	0.9929	0.9959	0.9569	0.7839	0.7734	0.9959	0.9063	0.9215	0.8608
120	0.9939	**0.9669**	0.9929	0.8311	0.9979	0.9949	**0.9979**	0.9959	0.9549	**0.7869**	0.7713	0.9959	0.9042	**0.9225**	0.8620
130	0.9969	0.9639	0.9949	0.8356	0.9979	0.996	**0.9979**	0.996	0.9549	0.7426	0.7726	0.9949	0.9033	0.9126	0.8686
140	0.9969	0.9629	0.9959	0.8405	**0.9989**	0.996	0.9969	0.998	0.9559	0.7536	0.7717	0.9949	0.9256	0.9156	0.8592
150	**0.9979**	0.9619	**0.9969**	0.8394	0.9959	0.996	0.9949	0.998	0.9549	0.7626	0.7717	0.9899	0.9277	0.9155	0.8703
160	0.9959	0.9619	**0.9969**	0.8366	0.9969	0.9929	0.9929	**0.9989**	**0.9599**	0.7442	**0.7738**	0.9919	0.9398	0.9195	0.8671
170	0.9959	0.9579	**0.9969**	0.8423	0.9959	0.9939	0.9949	**0.9989**	0.9579	0.7649	0.7726	0.9939	**0.9408**	0.9164	0.8681
180	0.9959	0.9649	0.9959	0.8359	0.9969	0.9939	0.9939	0.998	0.9579	0.7489	0.7728	0.9919	0.9398	0.9206	0.8682
190	0.9949	0.9639	0.9959	0.8421	0.9969	0.9939	0.9929	0.998	0.9569	0.7519	**0.7738**	0.9919	**0.9408**	0.9166	0.8671
200	0.9939	0.9639	0.9959	**0.8430**	0.9969	0.9909	0.9929	**0.9989**	0.9579	0.7503	0.7726	0.9939	0.9327	0.9144	**0.8774**

Tabela 27: Experimento com Atributos Estruturais na Base em Inglês Utilizando o Classificador MLP

Qtde Atributos	Texto	Conjunto de Marcadores							Marcadores						
		A	B	C	D	E	F	G	1	2	3	4	5	6	7
10	0.8898	0.8899	0.8915	0.5832	0.9139	0.9189	0.8813	0.9239	0.8074	0.8153	0.7646	0.8552	0.7613	0.7470	0.6744
20	0.8939	0.9135	0.9166	0.5802	0.9229	0.9399	0.9259	0.9469	0.8328	0.8165	0.7990	0.8985	0.7721	0.7708	0.7010
30	0.9379	0.9216	0.9279	0.6142	0.9649	0.9629	0.9329	0.9579	0.8469	0.8493	0.8135	0.9348	0.7759	0.7864	0.7235
40	0.9479	0.9297	0.9469	0.6174	0.9599	0.9649	0.9469	0.9659	0.8544	0.8513	0.8132	0.9399	0.7832	0.7848	0.7250
50	0.9479	0.9307	0.9399	0.6566	0.9659	0.9709	0.9499	0.9689	0.8618	0.8521	0.8199	0.9549	0.7823	0.8050	0.7462
60	0.9579	0.9337	0.9659	0.6647	0.9689	0.9679	0.9549	0.9769	0.8671	0.8518	0.8210	0.9609	0.7755	0.8034	0.7426
70	0.9539	0.9327	0.9639	0.6705	0.9679	0.9699	0.9519	0.97	0.8682	0.8571	0.8221	0.9649	0.7843	0.8048	0.7494
80	0.9589	0.9388	0.9679	0.6865	0.9729	0.9659	0.9559	0.9769	0.8703	0.8603	0.8231	0.9719	0.7877	0.8097	0.7393
90	0.9659	0.9399	0.9669	0.6856	0.9739	0.9689	0.9599	0.9719	0.8747	0.8632	0.8231	0.9759	0.7711	0.8104	0.7472
100	0.9689	0.9278	0.9669	0.7037	0.9729	0.9729	0.9639	0.9739	0.8839	0.8623	0.8231	0.9779	0.7759	0.8065	0.7619
110	0.9689	0.9368	0.9739	0.7298	0.9769	0.9749	0.9599	0.9779	0.8859	0.8663	0.8240	0.9789	0.7704	0.8076	0.7952
120	0.9729	0.9408	0.9759	0.7434	0.9819	0.9699	0.9649	0.9819	0.8921	0.8652	0.8240	0.9789	0.7730	0.8127	0.7599
130	0.9669	0.9418	**0.9869**	0.7356	0.9759	0.9769	0.9679	0.9809	0.8712	0.8673	0.8231	0.9779	0.7820	0.8059	**0.8032**
140	0.9689	0.9428	0.9829	**0.7487**	0.9749	0.9789	0.9639	0.9849	0.8849	0.8703	0.8240	**0.9819**	0.7389	0.8146	0.7582
150	0.9689	0.9398	0.9839	0.7394	0.9809	0.9779	0.9679	0.9859	0.8982	0.8672	**0.8262**	0.9769	**0.7995**	0.8211	0.7426
160	0.9649	0.9398	0.9849	0.7434	0.9729	0.9809	0.9689	0.9889	0.8942	0.8703	0.8251	0.9779	0.7631	**0.8233**	0.7558
170	0.9659	**0.9429**	0.9819	0.7367	0.9779	0.9799	0.9719	0.9879	0.8828	0.8693	0.8240	0.9759	0.7683	0.8077	0.7477
180	0.9709	**0.9429**	**0.9869**	0.6393	0.9839	0.9819	0.9729	0.9879	0.8913	0.8662	0.8219	0.9799	0.7721	0.8146	0.6496
190	**0.9809**	0.9257	0.9839	0.6934	0.9849	0.9819	0.9769	0.9879	**0.9055**	0.8661	0.8230	0.9789	0.7731	0.8144	0.7472
200	**0.9809**	0.9398	0.9809	0.6036	**0.9859**	**0.9849**	**0.9789**	**0.9899**	0.9032	**0.8723**	0.8219	0.9799	0.7874	0.8194	0.7553

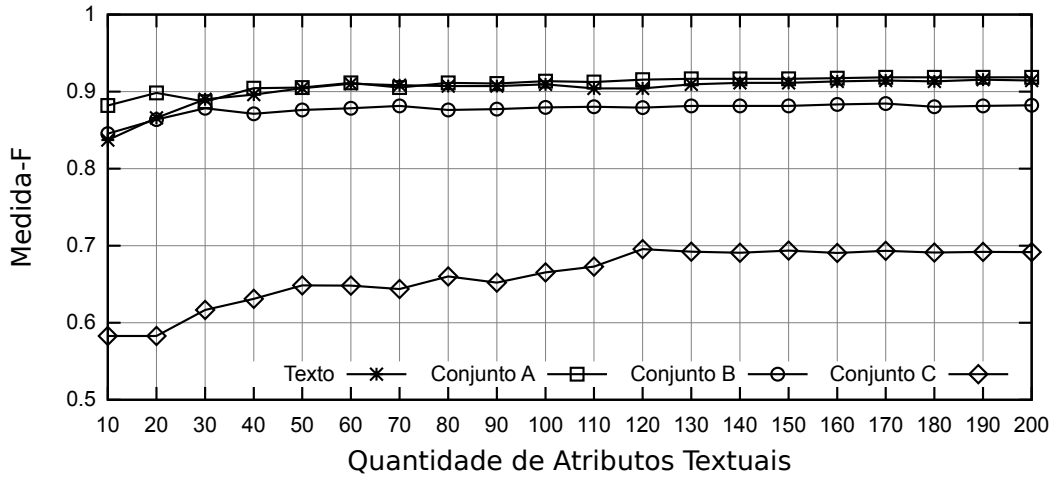

Figura 42: Desempenho do Classificador Naïve Bayes no Conjunto de Dados em Inglês: Grupo de Comparações 1

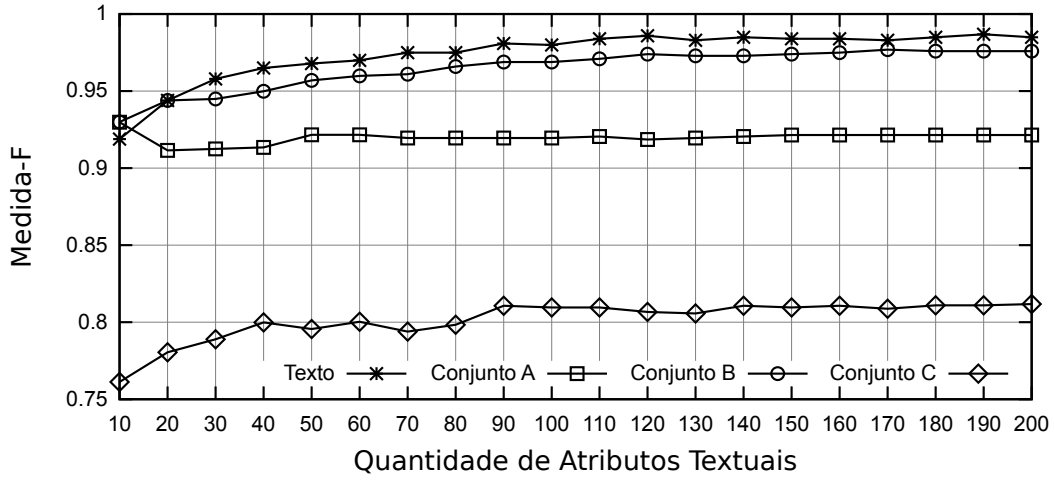

Figura 43: Desempenho do Classificador Naïve Bayes no Conjunto de Dados em Português: Grupo de Comparações 1

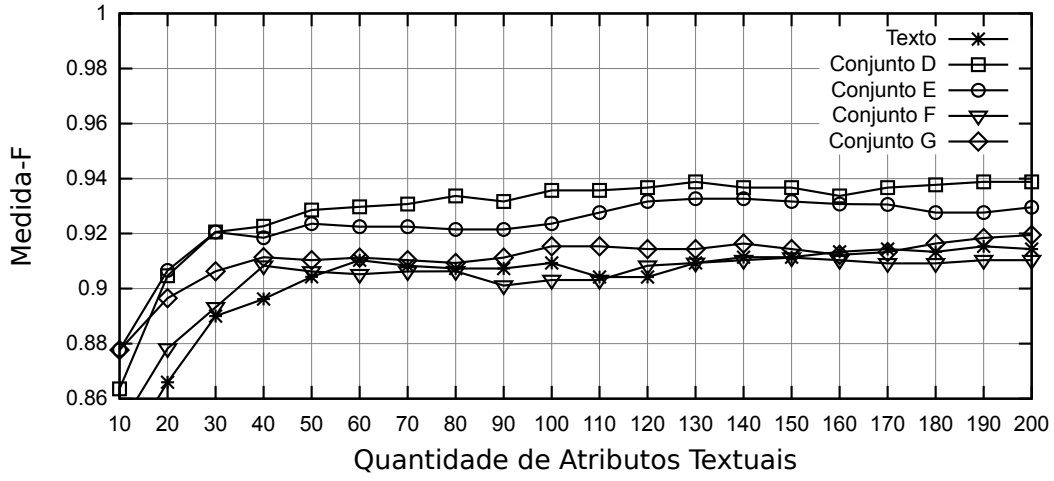

Figura 44: Desempenho do Classificador Naïve Bayes no Conjunto de Dados em Inglês: Grupo de Comparações 2

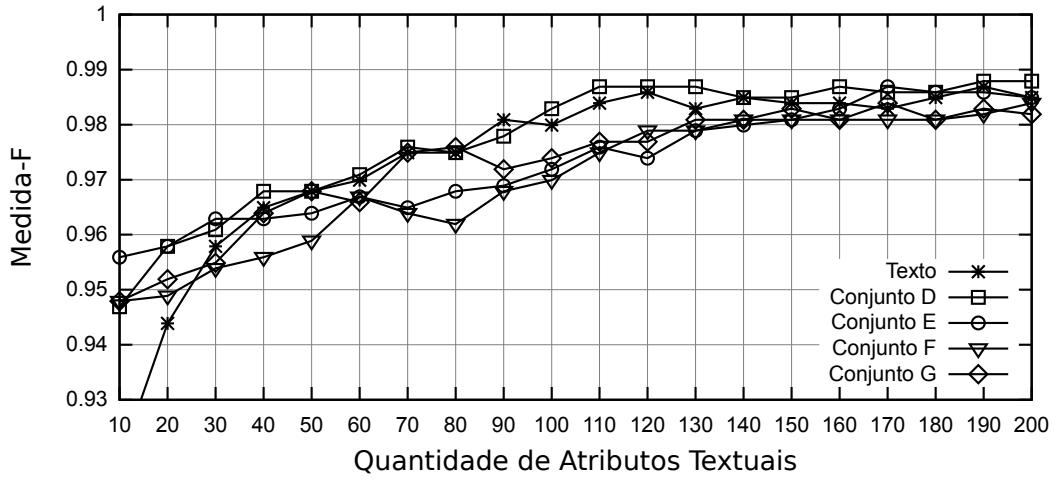

Figura 45: Desempenho do Classificador Naïve Bayes no Conjunto de Dados em Português: Grupo de Comparações 2

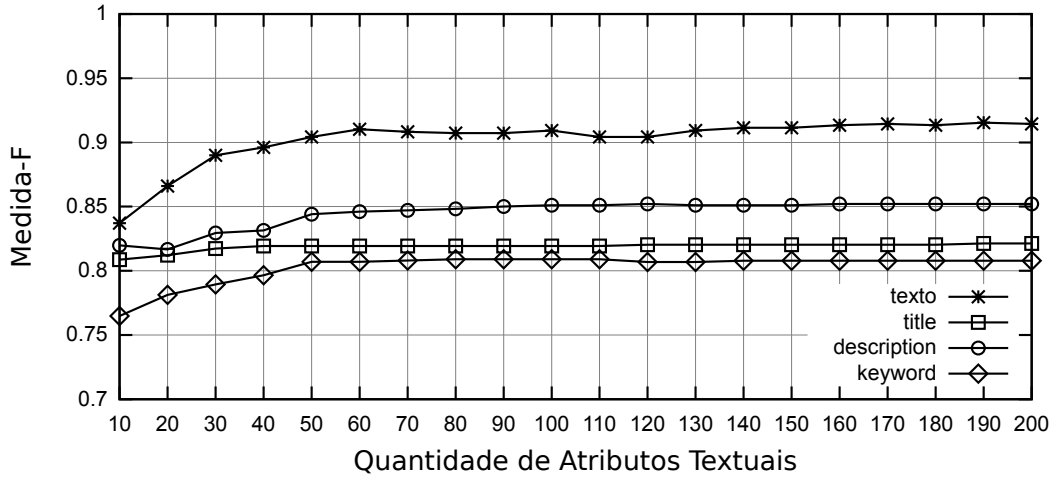

Figura 46: Desempenho do Classificador Naïve Bayes no Conjunto de Dados em Inglês: Grupo de Comparações 3

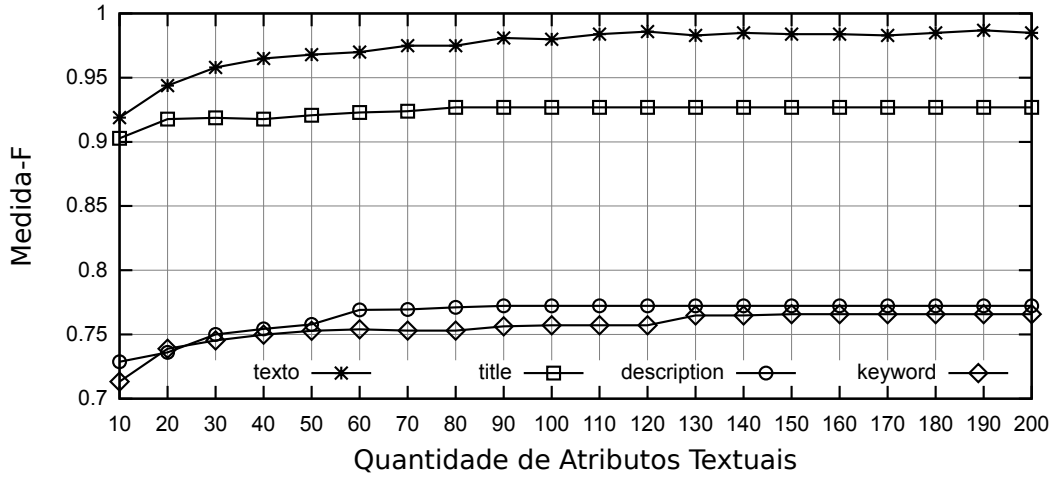

Figura 47: Desempenho do Classificador Naïve Bayes no Conjunto de Dados em Português: Grupo de Comparações 3

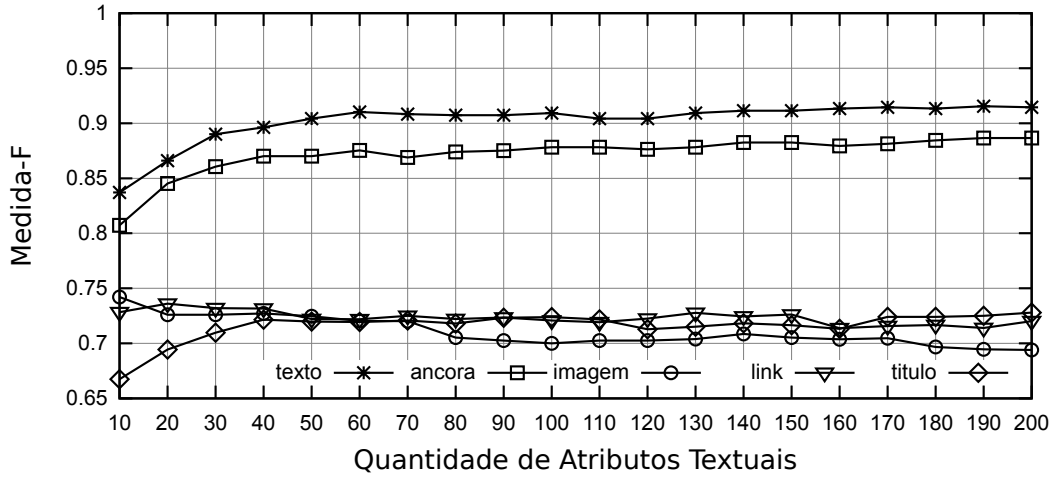

Figura 48: Desempenho do Classificador Naïve Bayes no Conjunto de Dados em Inglês: Grupo de Comparações 4

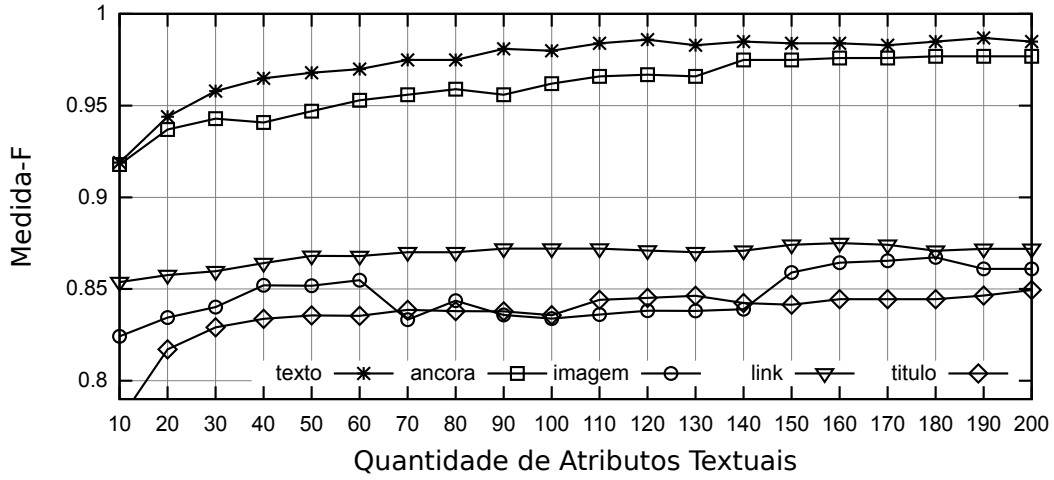

Figura 49: Desempenho do Classificador Naïve Bayes no Conjunto de Dados em Português: Grupo de Comparações 4

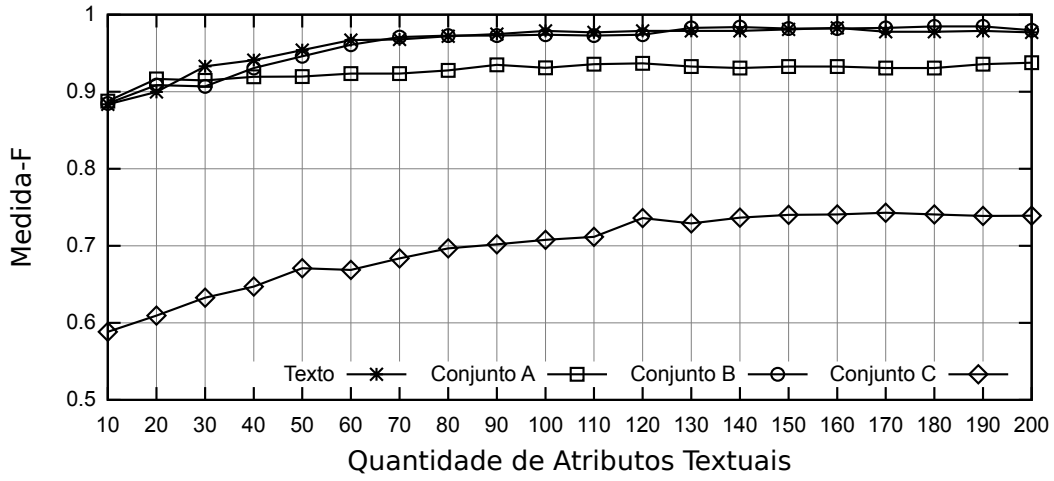

Figura 50: Desempenho do Classificador KNN no Conjunto de Dados em Inglês: Grupo de Comparações 1

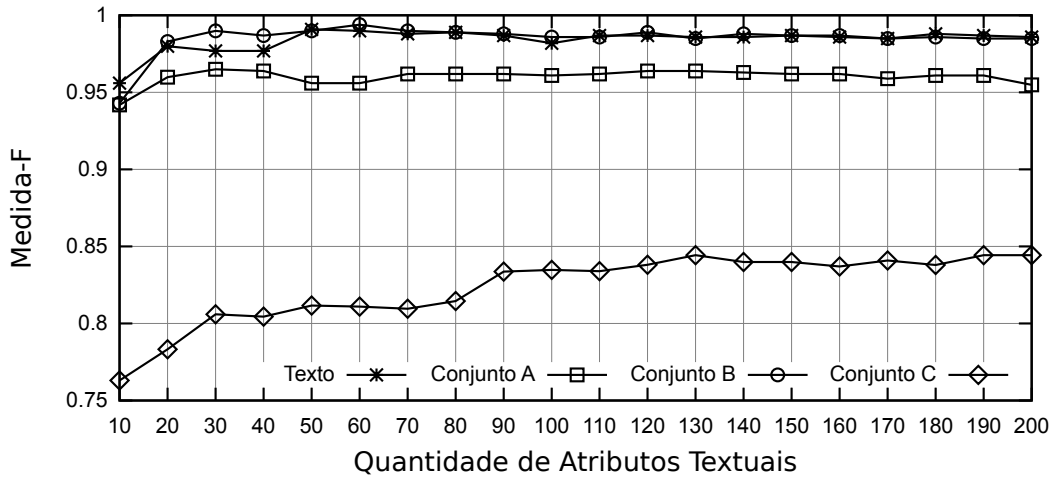

Figura 51: Desempenho do Classificador KNN no Conjunto de Dados em Português: Grupo de Comparações 1

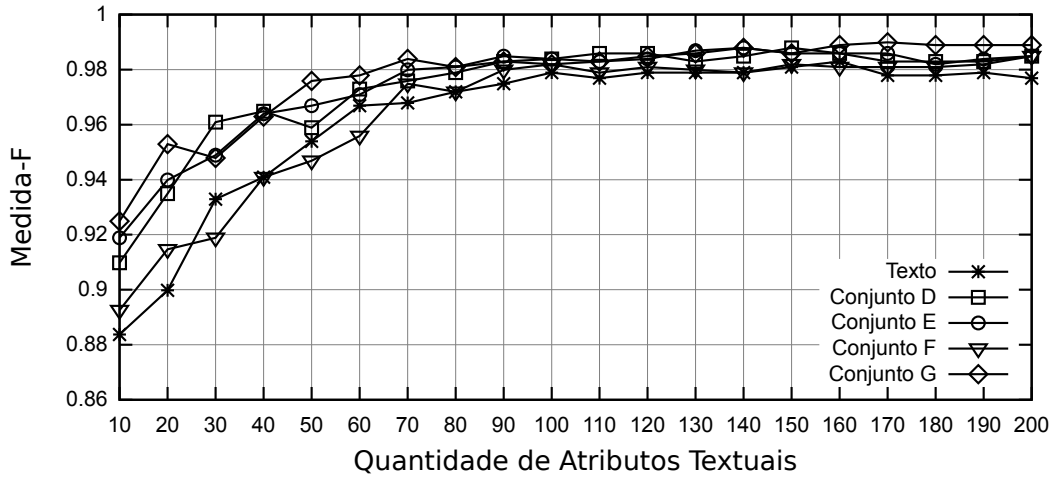

Figura 52: Desempenho do Classificador KNN no Conjunto de Dados em Inglês: Grupo de Comparações 2

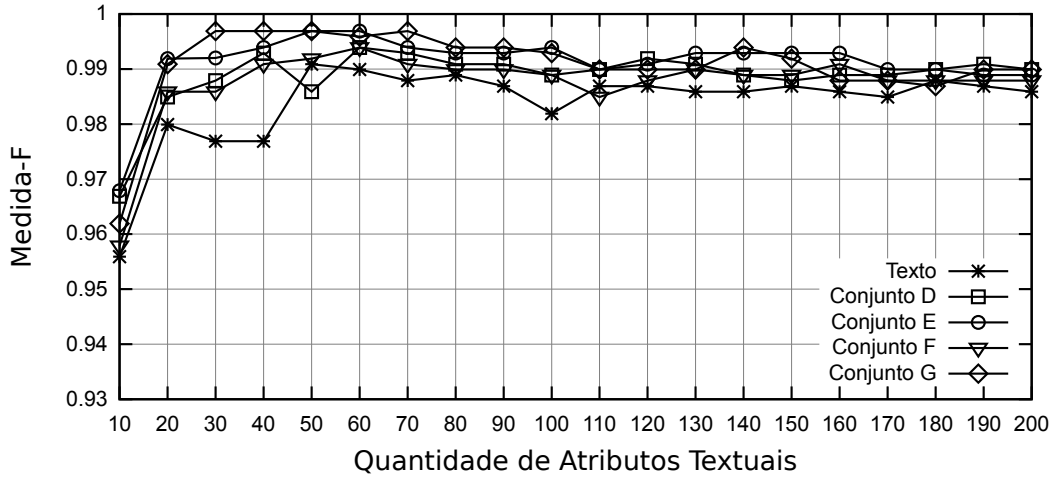

Figura 53: Desempenho do Classificador KNN no Conjunto de Dados em Português: Grupo de Comparações 2

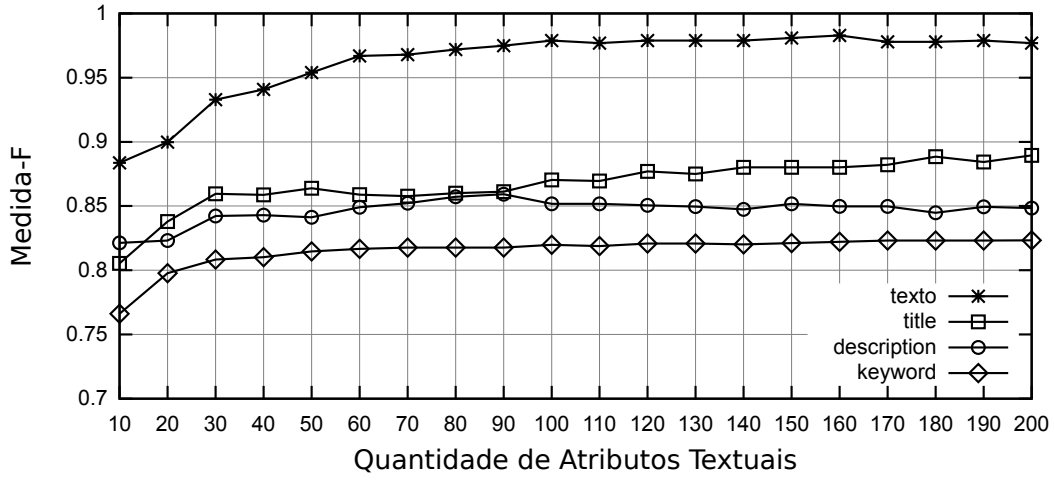

Figura 54: Desempenho do Classificador KNN no Conjunto de Dados em Inglês: Grupo de Comparações 3

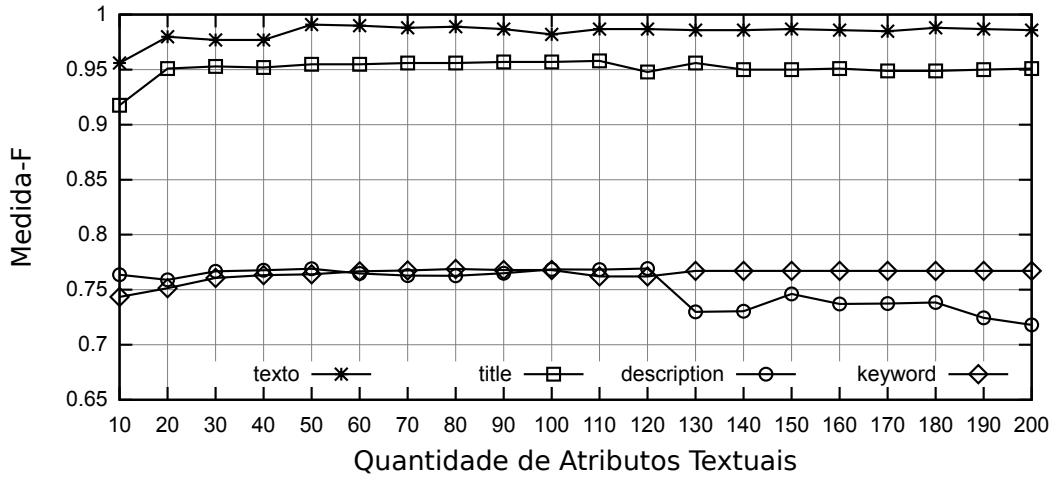

Figura 55: Desempenho do Classificador KNN no Conjunto de Dados em Português: Grupo de Comparações 3

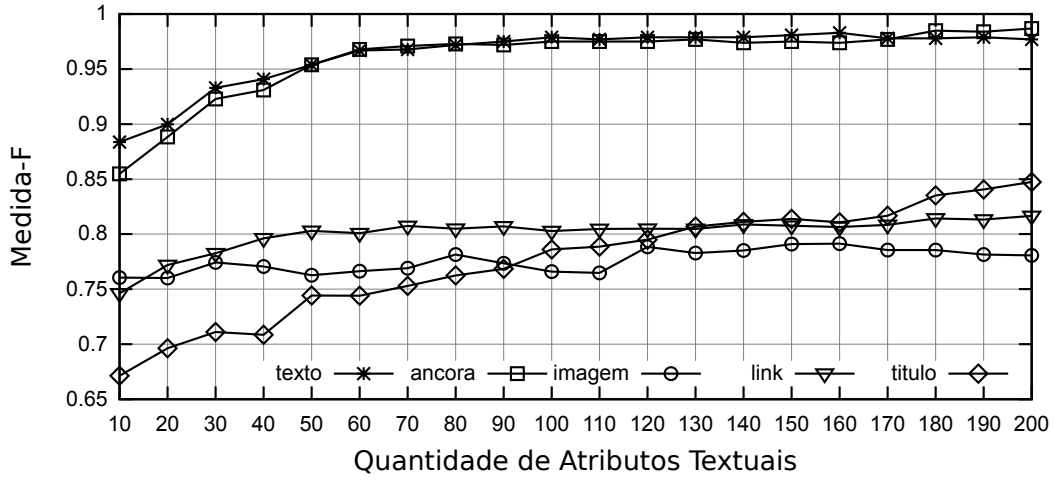

Figura 56: Desempenho do Classificador KNN no Conjunto de Dados em Inglês: Grupo de Comparações 4

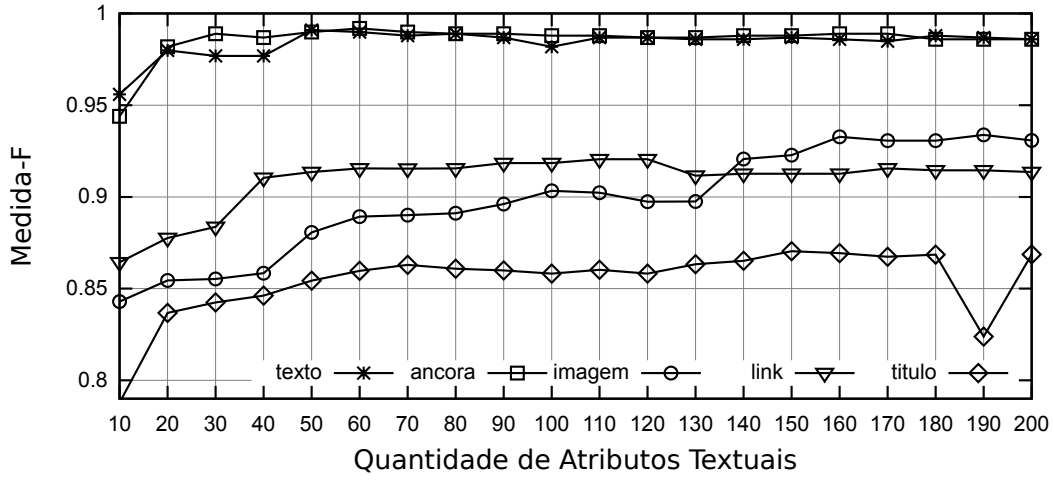

Figura 57: Desempenho do Classificador KNN no Conjunto de Dados em Português: Grupo de Comparações 4

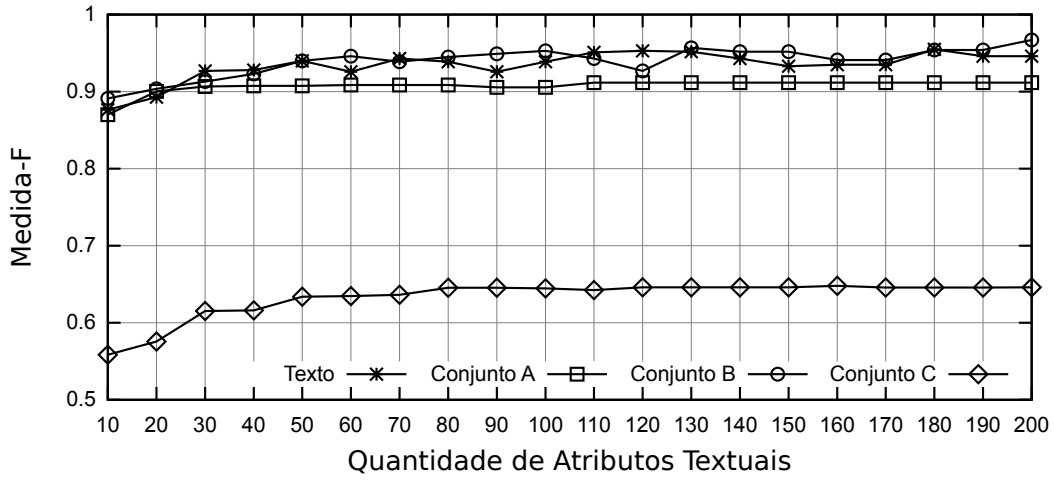

Figura 58: Desempenho do Classificador C4.5 no Conjunto de Dados em Inglês: Grupo de Comparações 1

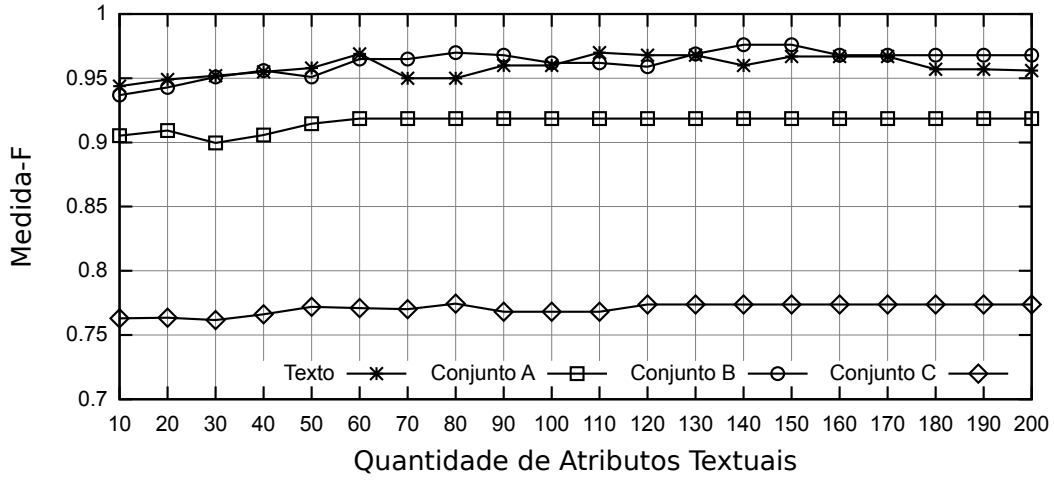

Figura 59: Desempenho do Classificador C4.5 no Conjunto de Dados em Português: Grupo de Comparações 1

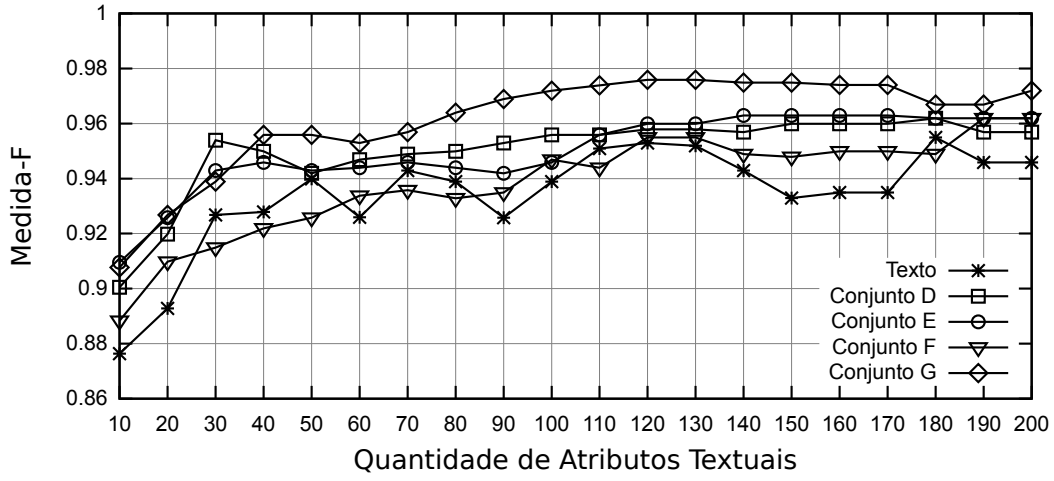

Figura 60: Desempenho do Classificador C4.5 no Conjunto de Dados em Inglês: Grupo de Comparações 2

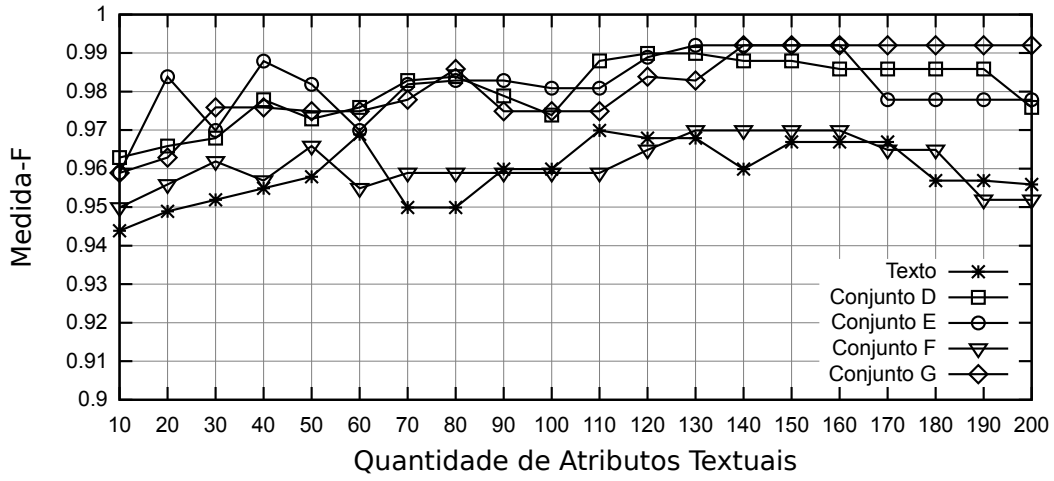

Figura 61: Desempenho do Classificador C4.5 no Conjunto de Dados em Português: Grupo de Comparações 2

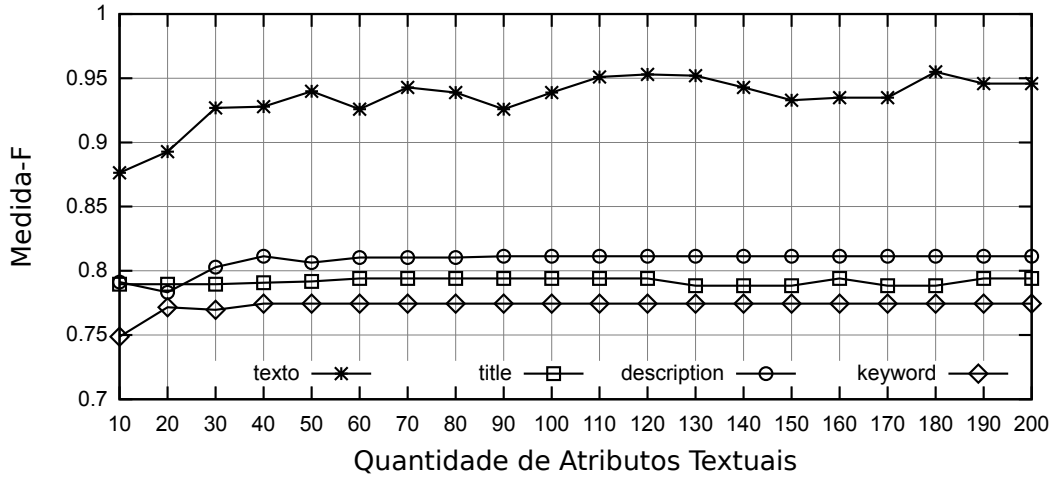

Figura 62: Desempenho do Classificador C4.5 no Conjunto de Dados em Inglês: Grupo de Comparações 3

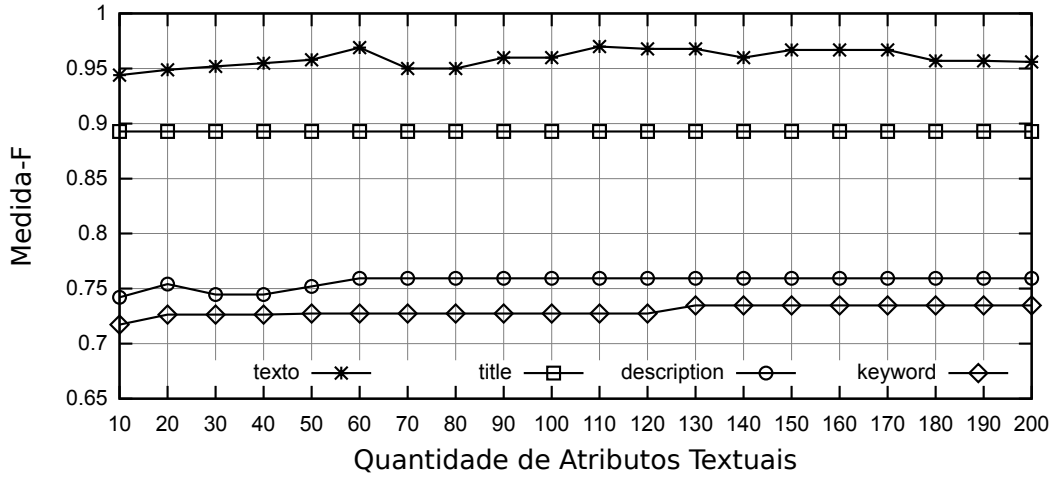

Figura 63: Desempenho do Classificador C4.5 no Conjunto de Dados em Português: Grupo de Comparações 3

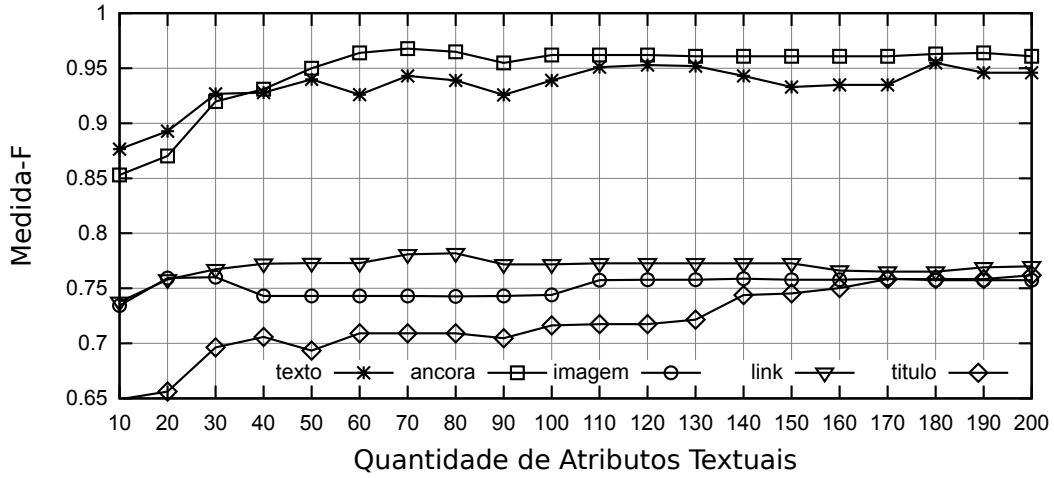

Figura 64: Desempenho do Classificador C4.5 no Conjunto de Dados em Inglês: Grupo de Comparações 4

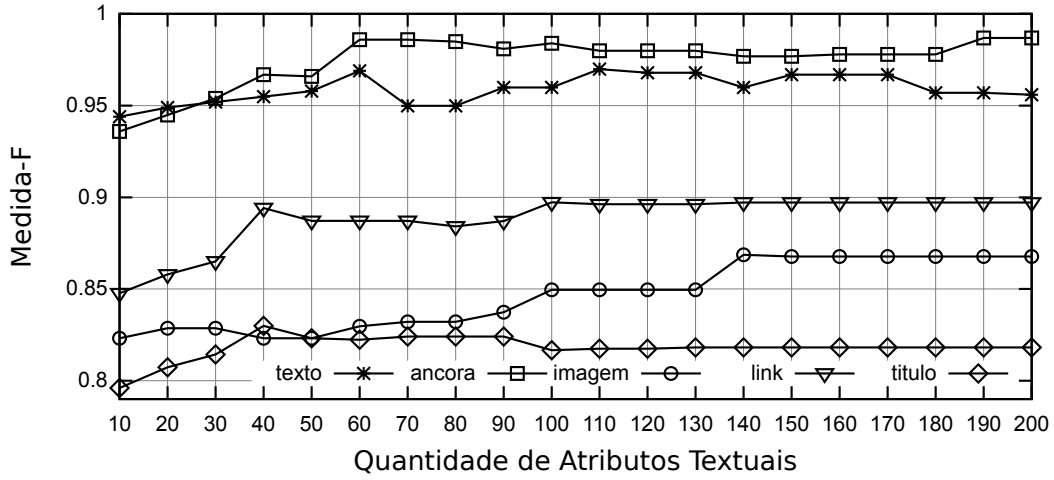

Figura 65: Desempenho do Classificador C4.5 no Conjunto de Dados em Português: Grupo de Comparações 4

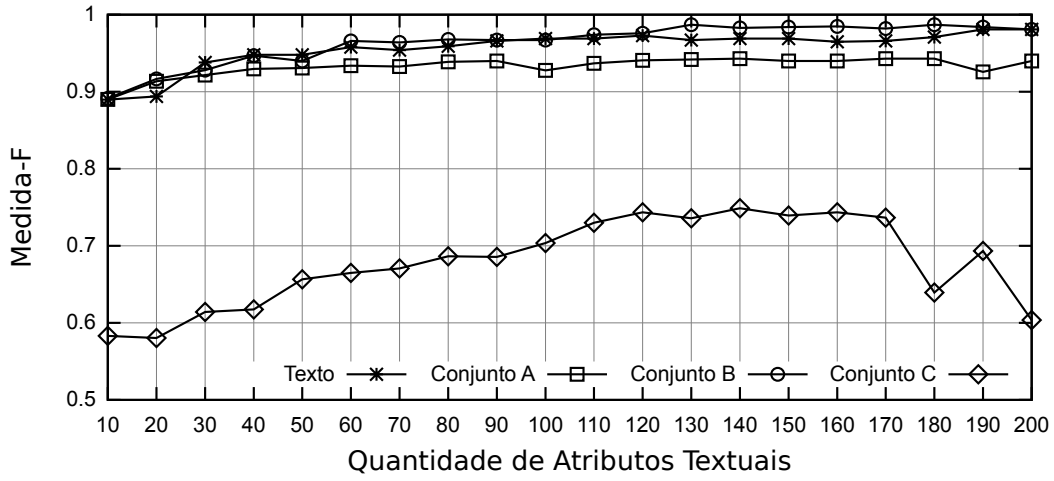

Figura 66: Desempenho do Classificador MLP no Conjunto de Dados em Inglês: Grupo de Comparações 1

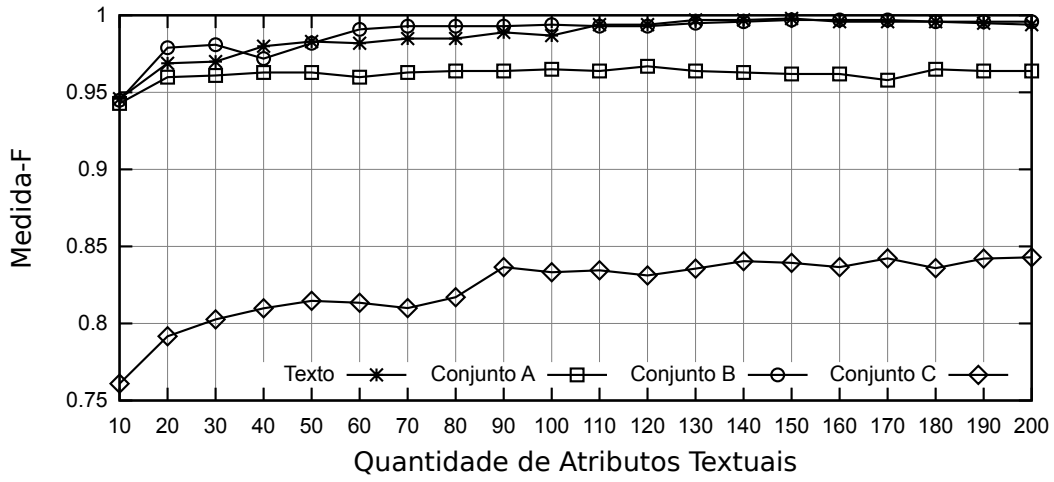

Figura 67: Desempenho do Classificador MLP no Conjunto de Dados em Português: Grupo de Comparações 1

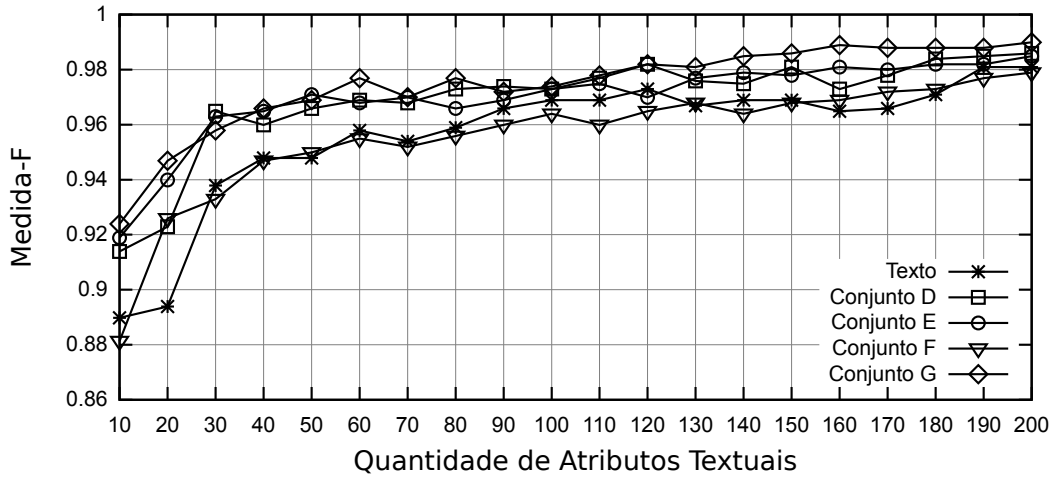

Figura 68: Desempenho do Classificador MLP no Conjunto de Dados em Inglês: Grupo de Comparações 2

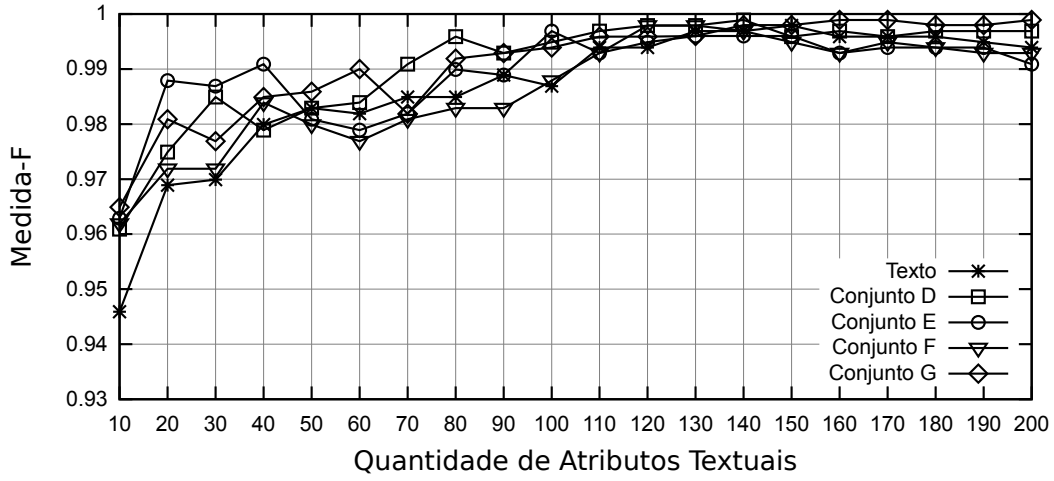

Figura 69: Desempenho do Classificador MLP no Conjunto de Dados em Português: Grupo de Comparações 2

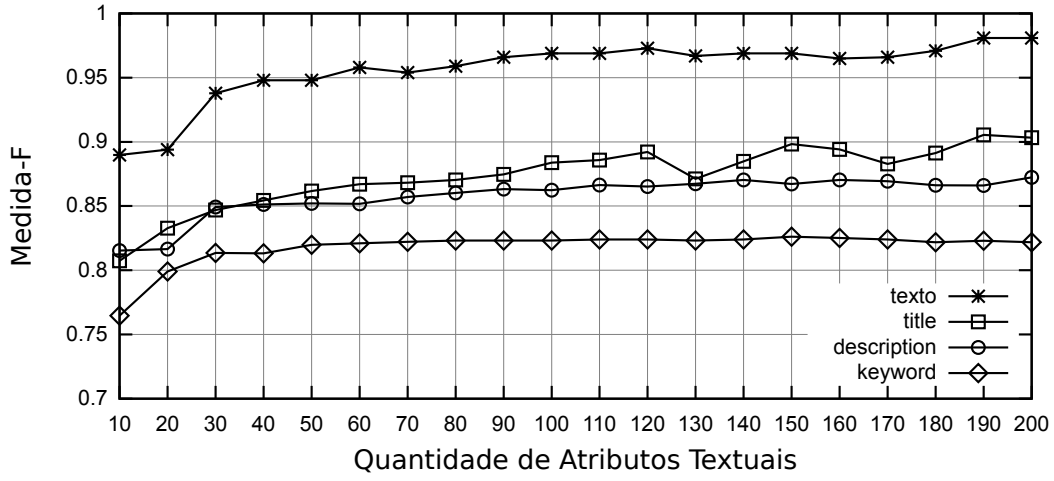

Figura 70: Desempenho do Classificador MLP no Conjunto de Dados em Inglês: Grupo de Comparações 3

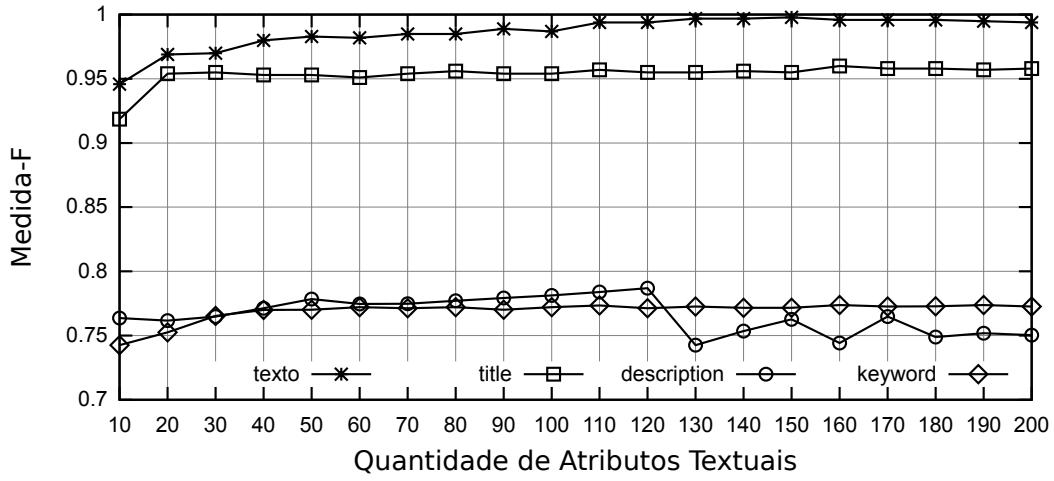

Figura 71: Desempenho do Classificador MLP no Conjunto de Dados em Português: Grupo de Comparações 3

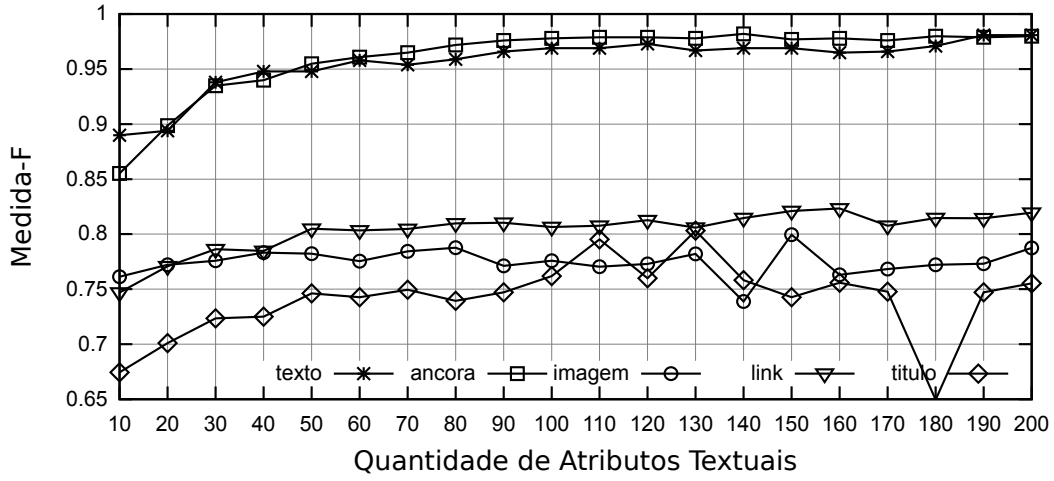

Figura 72: Desempenho do Classificador MLP no Conjunto de Dados em Inglês: Grupo de Comparações 4

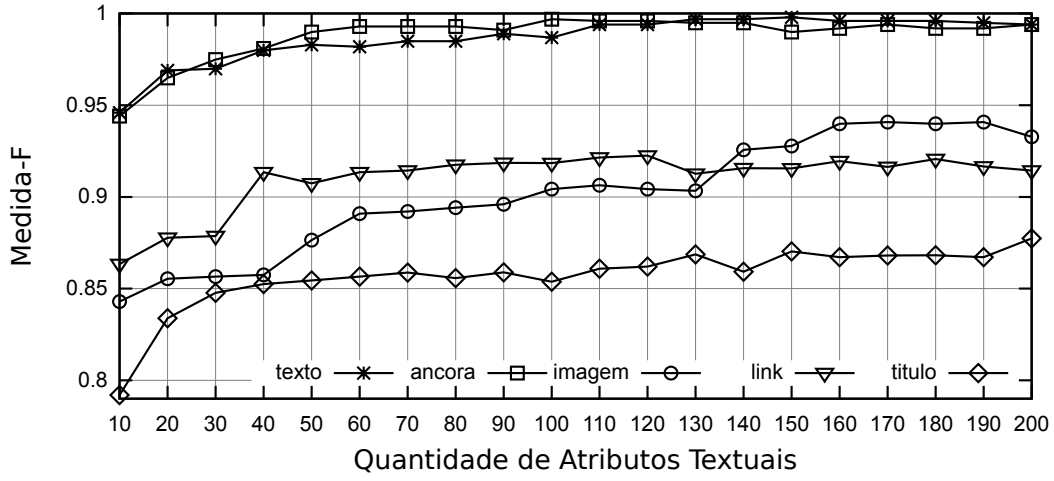

Figura 73: Desempenho do Classificador MLP no Conjunto de Dados em Português: Grupo de Comparações 4