5
Experimentos e Resultados

A idéia por trás dos experimentos realizados neste trabalho não foi a simples construção de classificadores eficazes, mas a exploração de diferentes abordagens ao problema de classificação, com variações nos conjuntos de atributos e o teste de diferentes classificadores. Partindo desta exploração, foi possível uma avaliação concreta de diferentes estratégias de classificação.

Foram utilizados 5 conjuntos distintos de páginas para os experimentos, criados conforme explicado no capítulo 4. Esses conjuntos foram rotulados de "Treino", "Teste1", "Posts+Comments", "Teste2_Gr" e "Teste2_Ing". Todos os conjuntos são disjuntos, e foram utilizados para todos os experimentos. Apesar dos atributos de texto não terem muito significado para o conjunto "Teste2_Gr", que não possui páginas em inglês, executar a classificação desse conjunto com esses atributos fornece um balizador interessante para os resultados obtidos.

Cada experimento representa a utilização de um determinado conjunto de atributos extraídos dos conjuntos de páginas mencionados acima. Para cada um deles, foram construídos 3 classificadores sobre o conjunto "Treino": uma árvore de decisão, uma rede neural e uma SVM. Em todos os experimentos, foi utilizada uma estratégia de validação cruzada com 10 quebras (10-*fold cross-validation*). O método de árvore de decisão adotado foi o J48, capaz de trabalhar com atributos numéricos e realizar a separação não-binária de atributos nos ramos. Todas as redes neurais utilizadas foram *multi-layer perceptrons* com 4 saídas, classificando as entradas em uma de 4 classes. A SVM foi treinada através do método de *Sequential Minimal Optimization*[30], e foram treinadas diversas SVMs para realizar as comparações binárias das classes. Após a sua construção, estes classificadores foram testados sobre os outros conjuntos.

Todos os classificadores foram treinados e submetidos aos conjuntos de testes através da ferramenta WEKA. A utilização de uma ferramenta amplamente adotada e reconhecida evita o desenvolvimento e todas as validações de corretude de implementação que seriam necessárias. Permite também a avaliação da

qualidade de classificadores "*off-the-shelf*", ou seja, que estão disponíveis para uso geral.

Em todos os casos foram utilizados os parâmetros padrão que já vêm pré-configurados no WEKA. Não houve nenhuma alteração nos parâmetros nem testes com outras parametrizações, no intuito de avaliar a qualidade sem a preocupação de chegar a um classificador bastante otimizado. A combinação das SVMs utilizada foi a disponibilizada automaticamente pelo próprio WEKA, sem interferência manual.

Não foram realizados em nenhum momento testes sobre conjuntos especiais, nem a incorporação de conjuntos de teste ao conjunto de treinamento. Em todos os experimentos, apenas o conjunto "Treino" foi utilizado para treinar e validar os classificadores (através da validação cruzada). Todos os outros conjuntos foram utilizados apenas para testes dos classificadores gerados.

Abaixo, estão descritos os diferentes experimentos realizados, com as variações no conjunto de atributos utilizados. São apresentados também os resultados obtidos em cada um deles, e uma breve discussão sobre os mesmos.

5.1. Classificação Funcional com Atributos Estruturais

O objetivo deste experimento é avaliar a capacidade preditiva do conjunto inicial de atributos estruturais selecionados para classificação. Para este experimento, o conjunto de atributos estruturais descrito na seção 4.1 foi extraído dos 5 conjuntos de páginas ("Treino", "Teste1", "Posts+Comments", "Teste2_Gr" e "Teste2_Ing").

A seguir, é apresentada uma tabela com os valores de acurácia dos classificadores para cada um dos conjuntos de páginas. Em seguida, são exibidas as matrizes de confusão para os mesmos.

	Treino	Teste1	Posts+Comments	Teste2_Gr	Teste2_Ing
Árvore de Decisão	**97.86%**	92.43%	68.50%	61.32%	56.37%
Rede Neural	97.56%	**93.51%**	**71.00%**	**69.50%**	**56.76%**
SVM	89.83%	88.69%	67.50%	66.04%	55.21%

Tabela 3 – Acurácia por classificador por conjunto de informação. Em negrito estão sinalizados os melhores resultados para cada conjunto

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	241	0	3	3
Notícia	1	248	1	0
Portal de Notícia	7	4	234	0
Blog Posts	1	0	1	239
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	239	0	7	1
Notícia	0	243	2	5
Portal de Notícia	3	4	236	2
Blog Posts	0	0	0	241
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	217	3	22	5
Notícia	8	238	1	3
Portal de Notícia	22	11	212	0
Blog Posts	11	14	0	213

Tabela 4 – Matriz de Confusão por classificador para o conjunto “Treino”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	239	1	9	4
Notícia	4	242	4	0
Portal de Notícia	34	15	210	1
Blog Posts	4	5	1	249
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	239	0	11	3
Notícia	2	237	3	8
Portal de Notícia	24	9	219	3
Blog Posts	3	0	0	256
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	223	1	20	9
Notícia	10	236	1	3
Portal de Notícia	30	14	210	1
Blog Posts	16	10	0	233

Tabela 5 – Matriz de Confusão por classificador para o conjunto “Teste1”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	84	3	13	0
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	17	25	5	53
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	75	6	17	2
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	22	10	4	64
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	71	4	25	0

Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	22	10	4	64

Tabela 6 – Matriz de Confusão por classificador para o conjunto “Posts + Comments”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	70	2	14	1
Notícia	21	42	25	7
Portal de Notícia	4	4	83	0
Blog Posts	6	29	20	0
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	69	2	8	8
Notícia	11	49	18	17
Portal de Notícia	4	4	83	0
Blog Posts	4	5	12	20
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	69	2	12	4
Notícia	6	59	27	3
Portal de Notícia	3	6	82	0
Blog Posts	0	28	17	0

Tabela 7 – Matriz de Confusão por classificador para o conjunto “Teste2_Gr”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	51	0	5	1
Notícia	2	33	5	2
Portal de Notícia	3	4	43	0
Blog Posts	22	49	20	19
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	45	3	5	4
Notícia	0	33	5	4
Portal de Notícia	2	5	43	0
Blog Posts	4	76	4	26
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	50	1	4	2
Notícia	0	39	3	0
Portal de Notícia	3	6	41	0
Blog Posts	1	91	5	13

Tabela 8 – Matriz de Confusão por classificador para o conjunto “Teste2_Ing”

Sobre o conjunto “Treino”, o classificador que apresentou o melhor resultado foi a Árvore de Decisão, seguida da Rede Neural, e da SVM. Todos os classificadores apresentaram resultados semelhantes em termos de acurácia para esse conjunto.

No conjunto "Teste1", os resultados obtidos no treino se mantiveram consistentes. Ocorreu uma pequena redução nas acurácias da Árvore de Decisão e da Rede Neural, e a manutenção da acurácia da SVM. Através das matrizes de confusão, é possível observar que a maior dificuldade de classificação, ou seja, o ponto onde ocorrem a maioria dos erros, está na distinção entre blogs e portais de notícias. Para esse conjunto de informações, essa foi a classe com maiores taxas de erro em todos os classificadores (17,6% para a árvore de decisão, 14,1% para a rede neural e 17,6% para a SVM).

Para o conjunto "Posts+Comments", é possível observar uma queda significativa no patamar de acurácia, caindo de aproximadamente 90% no primeiro conjunto de testes para aproximadamente 70% neste. A rede neural ainda se apresenta como o classificador com melhores índices de acurácia, se mantendo assim nos outros conjuntos de teste também.

Observando as matrizes de confusão para o conjunto "Teste2_Gr", é possível observar que existe dificuldades de distinção, entre notícias e portais de notícias e blogs, e entre blog posts e notícias. Em nenhum outro conjunto a dificuldade de distinguir notícias de portais de notícias e blogs se repete, o que sugere que as notícias do conjunto "Teste2_Gr" são estruturalmente diferentes das notícias dos outros conjuntos.

No conjunto "Teste2_Ing", a queda de acurácia é ainda mais acentuada, passando para a faixa de 55%. Diferente do ocorrido no conjunto "Teste1", a maior dificuldade de classificação neste caso foi a distinção dos posts das outras classes. Através das matrizes de confusão, é possível observar que, descartando-se a classe de posts, os índices de acurácia seriam muito mais elevados. Para a rede neural, por exemplo, o índice de acurácia descartada a classe de Blog Posts, seria de 73,63% (201 exemplos classificados corretamente sobre 273 exemplos existentes).

A dificuldade de classificação dos posts deve-se ao fato de que nestes conjuntos, conforme mencionado no capítulo 4, os Posts foram selecionados por possuírem um número maior de comentários, o que dificulta sua classificação (conforme discutido na seção 3.2).

Apesar da redução na acurácia, os índices atingidos representam ganhos significativos sobre uma classificação ou palpite aleatório, para o qual a probabilidade de acerto seria de 25% (1 chance em 4).

5.2. Classificação Funcional com Atributos de Texto (80%)

O objetivo deste segundo experimento é avaliar a qualidade dos classificadores construídos utilizando apenas atributos de texto. Aqui, o conjunto de atributos utilizados é o das palavras que aparecem em pelo menos 80% dos documentos de uma classe (conforme mencionado no capítulo 4).

A seguir, é apresentada uma tabela com os valores de acurácia de cada um dos classificadores para os diferentes conjuntos de teste e, em seguida, as matrizes de confusão para cada um deles.

	Treino	Teste1	Posts+Comments	Teste2_Gr	Teste2_Ing
Árvore de Decisão	**96.23%**	**75.42%**	**67.50%**	**34.59%**	47.1%
Rede Neural	73.86%	68.83%	61.50%	21.70%	49.81%
SVM	66.83%	65.68%	56.00%	33.66%	**57.53%**

Tabela 9 – Acurácia por classificador por conjunto de informações. Em negrito estão sinalizados os melhores resultados para cada conjunto

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	241	1	1	4
Notícia	3	236	6	5
Portal de Notícia	3	9	231	2
Blog Posts	2	0	1	238
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	191	20	3	33
Notícia	2	246	2	0
Portal de Notícia	1	158	86	0
Blog Posts	11	25	2	203
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	149	24	15	59
Notícia	6	181	53	10
Portal de Notícia	2	77	159	7
Blog Posts	10	37	26	168

Tabela 10 – Matriz de Confusão por classificador para o conjunto “Treino”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	212	4	9	28
Notícia	12	164	63	11
Portal de Notícia	24	44	171	16
Blog Posts	22	6	11	220
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**

Blog	172	24	6	51
Notícia	2	241	6	1
Portal de Notícia	7	174	72	2
Blog Posts	20	20	4	215
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	133	20	20	80
Notícia	4	185	52	9
Portal de Notícia	4	83	163	5
Blog Posts	13	28	31	187

Tabela 11 – Matriz de Confusão por classificador para o conjunto “Teste1”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	82	4	5	9
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	47	0	0	53
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	73	10	6	11
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	40	6	4	50
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	58	6	9	27
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	33	4	9	54

Tabela 12 – Matriz de Confusão por classificador para o conjunto “Post + Comments”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	64	1	14	8
Notícia	47	0	48	0
Portal de Notícia	44	0	46	1
Blog Posts	37	0	8	0
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	3	1	0	83
Notícia	0	23	0	72
Portal de Notícia	0	22	0	69
Blog Posts	0	2	0	43
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	0	0	21	66
Notícia	0	2	90	3
Portal de Notícia	0	0	80	11
Blog Posts	0	0	19	26

Tabela 13 – Matriz de Confusão por classificador para o conjunto “Teste2_Gr”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**

Blog	43	2	5	7
Notícia	3	20	16	3
Portal de Notícia	2	17	30	1
Blog Posts	50	14	17	19
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	42	7	2	6
Notícia	2	39	0	1
Portal de Notícia	1	32	17	0
Blog Posts	31	36	12	31
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	36	8	4	9
Notícia	1	34	4	3
Portal de Notícia	1	22	25	2
Blog Posts	29	13	14	54

Tabela 14 – Matriz de Confusão por classificador para o conjunto “Teste2_Ing”

Utilizando este conjunto de atributos, o classificador que apresentou os melhores resultados foi a árvore de decisão, obtendo a maior acurácia em todos os conjuntos de páginas menos o último (“Teste2_Ing”). Nesse último conjunto, o classificador com os melhores resultados foi a SVM.

No primeiro conjunto de páginas (“Teste1”), o primeiro ponto de observação é a redução nos níveis de acurácia dos classificadores. Enquanto que para o conjunto de atributos estruturais a acurácia estava no patamar de 90%, aqui ela se encontra na faixa de 65-75%. Essa redução na acurácia ocorre também para todos os outros conjuntos, sugerindo que o poder de classificação dos atributos de texto é menor do que o dos atributos estruturais. Os atributos de texto utilizados estão fortemente ligados ao conjunto de treino, devido ao método de extração adotado, o que prejudica a qualidade da classificação dos outros conjuntos de páginas.

A baixa acurácia no conjunto “Teste2_Gr” já era esperada, uma vez que os termos utilizados como atributos são da língua inglesa, e nenhuma das páginas deste conjunto são desta língua.

Diferente do caso anterior, neste experimento é possível observar, através das matrizes de confusão, dois pontos de dificuldade significativos nas classificações, a confusão de blog posts com blogs e a confusão de portais de notícias com notícias. A explicação por trás dessas dificuldades está no fato de que os posts vêm de blogs, assim como as notícias vêm de portais. Os blogs incorporam, ao longo do seu texto, trechos dos posts que os compõe, e a mesma coisa acontece com os portais de notícias. Assim, os termos permitem a

diferenciação de blogs e portais de notícias, por exemplo, mas não a diferenciação de blogs e blog posts.

É interessante observar que a degradação do índice de acurácia do conjunto "Teste1" para o conjunto "Posts+Comments" e para o conjunto "Teste2_Ing" foi bem menor do que no experimento anterior. Olhando para as matrizes de confusão, é possível observar que a dificuldade de segmentação dos blog posts e blogs, assim como a dificuldade de segmentação de notícias e portais de notícias, está presente em todos os conjuntos de páginas. Diferente do experimento anterior, a dificuldade de classificação de blog posts não ocorre apenas quando os posts têm grandes quantidades de comentários. Ela é constante em todos os corpus utilizados para testes, e até mesmo no corpus de treino.

A baixa acurácia no último conjunto de testes ("Teste2_Ing") é novamente um resultado da distorção nas proporções de páginas neste conjunto. Nele, das 259 páginas, 110 são blog posts, uma classe que os classificadores têm dificuldade de segmentar corretamente. Assim, o percentual de erro nesta classe influencia fortemente a acurácia total do classificador, uma vez que esta classe representa 42,47% do total. Caso as páginas estivessem igualmente distribuídas, os resultados obtidos seriam melhores. Para a SVM, por exemplo, descartando-se a classe de blog posts, a acurácia seria de 63,75% (95 de 149 exemplos classificados corretamente).

5.3. Classificação Funcional com Atributos de Texto (50%)

Neste experimento, o objetivo é avaliar os classificadores construídos em cima de um conjunto de atributos textuais que são as palavras que aparecem em pelo menos 50% das páginas de cada uma das classes. Esse é o segundo experimento com atributos de texto. Além de comparar os resultados com os obtidos com atributos estruturais (na seção 5.1), deseja-se também comparar os diferentes conjuntos de atributos de texto.

Abaixo está exibida a acurácia de cada um dos classificadores para os diferentes conjuntos de informações, e as matrizes de confusão para os mesmos.

	Treino	Teste1	Posts+Comments	Teste2_Gr	Teste2_Ing
Árvore de Decisão	**97.05%**	**82.99%**	60.50%	**42.77%**	48.26%
Rede Neural	26.55%	25.37%	0.50%	29.87%	16.99%
SVM	89.62%	82.60%	**66.00%**	21.7%	**60.23%**

Tabela 15 – Acurácia por classificador por conjunto de informações. Em negrito estão sinalizados os melhores resultados para cada conjunto

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	238	1	2	6
Notícia	1	244	4	1
Portal de Notícia	2	5	237	1
Blog Posts	3	1	2	235
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	9	238	0	0
Notícia	0	250	0	0
Portal de Notícia	0	245	0	0
Blog Posts	1	238	0	2
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	199	3	3	42
Notícia	2	220	28	0
Portal de Notícia	0	18	223	4
Blog Posts	2	0	0	239

Tabela 16 – Matriz de confusão por classificador para o conjunto “Treino”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	197	18	15	23
Notícia	11	216	18	5
Portal de Notícia	17	32	200	6
Blog Posts	21	2	5	231
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	8	243	0	2
Notícia	1	248	0	1
Portal de Notícia	0	255	0	0
Blog Posts	2	255	0	2
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	169	12	7	65
Notícia	3	208	38	1
Portal de Notícia	4	30	212	9
Blog Posts	6	0	2	251

Tabela 17 – Matriz de confusão por classificador para o conjunto “Teste1”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	59	8	12	21
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	36	1	1	62

Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	1	98	6	1
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	6	94	0	0
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	72	6	3	19
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	38	2	0	60

Tabela 18 – Matriz de confusão por classificador para o conjunto “Posts + Comments”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	74	0	8	5
Notícia	39	1	52	3
Portal de Notícia	37	0	51	3
Blog Posts	32	0	3	10
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	0	87	0	0
Notícia	0	95	0	0
Portal de Notícia	0	91	0	0
Blog Posts	0	45	0	0
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	9	0	2	76
Notícia	0	2	10	83
Portal de Notícia	1	0	16	74
Blog Posts	2	0	1	42

Tabela 19 – Matriz de confusão por classificador para o conjunto “Teste2_Gr”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	37	6	8	6
Notícia	2	25	9	6
Portal de Notícia	4	10	31	5
Blog Posts	46	3	29	32
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	1	55	0	1
Notícia	0	42	0	0
Portal de Notícia	0	50	0	0
Blog Posts	4	105	0	1
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	41	7	2	7
Notícia	1	36	3	2
Portal de Notícia	1	10	39	0
Blog Posts	43	11	16	40

Tabela 20 – Matriz de confusão por classificador para o conjunto “Teste2_Ing”

A primeira diferença significativa entre este conjunto de atributos e os apresentados nas seções anteriores (5.1 e 5.2) é a baixa qualidade de resultados apresentados pela rede neural. Enquanto a rede neural se mostrou o melhor classificador de todos no primeiro experimento, e apresentou resultados competitivos com os outros classificadores no segundo, neste experimento sua performance foi muito fraca.

Isso ocorreu devido ao grande número de atributos utilizados neste experimento, em especial com relação aos dois primeiros (307 contra 23 e 78, respectivamente). Com este grande número de atributos, o treino da rede neural não terminou de maneira satisfatória (tendo permanecido em execução por algumas horas). Uma vez não tendo sido treinado corretamente, é lógico concluir que os resultados obtidos não são significativos.

Embora a rede neural tenha sido o classificador que mais sofreu devido ao incremento no número de atributos, tanto a árvore de decisão quanto a SVM também incorreram em muito mais tempo de processamento durante o seu processo de treinamento.

No entanto, os resultados obtidos, para todos os conjuntos de páginas, com estes dois classificadores foram melhores do que os resultados obtidos com o conjunto de atributos de texto 80%, apresentado anteriormente. O processo de extração de atributos de texto depende de um ponto de corte arbitrário. Esse ponto de corte, de 80% no caso anterior, pode ter causado a remoção de termos relevantes para a classificação. A redução deste ponto de corte neste experimento para 50% garantiu a inclusão desses termos, permitindo que o processo de treinamento dos classificadores determinasse sua relevância ou não.

Apesar das melhoras com relação ao conjunto de atributos de texto apresentados anteriormente, os resultados obtidos neste experimento ainda se mostraram inferiores aos obtidos com a utilização de atributos estruturais.

Os atributos estruturais, ao longo destes três primeiros experimentos, se mostraram significativamente melhores do que os atributos de texto. Em primeiro lugar, são totalmente independentes de linguagem, necessitando apenas da interpretação do código HTML de uma página para a sua extração. Em segundo lugar, foram utilizados apenas 23 atributos estruturais, contra 78 atributos de texto no conjunto 80% e 307 atributos no conjunto 50%. O aumento no conjunto de atributos causa um problema de explosão do tempo de treinamento dos

classificadores. Por último, os atributos de texto precisam ser recalculados sempre que o conjunto de treino for alterado, uma vez que eles dependem diretamente deste conjunto. Os atributos estruturais, no entanto, são independentes das páginas.

Outra dificuldade encontrada com os atributos de texto é a sua dependência do idioma. Enquanto os atributos estruturais podem ser extraídos e analisados de qualquer página, os atributos de texto utilizados neste trabalho estão associados ao inglês. Observando a matriz de confusão dos classificadores para o conjunto "Teste2_Gr", nota-se que as classes foram atribuídas de forma incorreta para praticamente todos os casos. A rede neural, por exemplo, classificou todos os exemplos desse conjunto na mesma classe, alguns corretamente e outros incorretamente.

5.4. Classificação Funcional com Seleção de Melhores Atributos

Com os experimentos anteriores, foi possível estabelecer a superioridade dos atributos estruturais sobre os atributos de texto na tarefa de classificação funcional de documentos da Web. O conjunto de atributos utilizado no experimento da seção 5.1 consiste de 23 atributos, selecionados manualmente.

O objetivo deste experimento é verificar os resultados da classificação quando é realizada a seleção dos "melhores" atributos através de métodos automatizados. Existem diversos métodos que atendem a estas características, disponíveis na ferramenta de classificação utilizada. Dentro destes, foram selecionados os métodos de ranqueamento por *Gain Ratio*, ranqueamento por *Information Gain*, *CFS Subset Evaluation* e ranqueamento pela estatística chi-quadrada com relação a classe. A execução destes métodos resultou em 3 ordenações dos atributos (de cada um dos ranqueamentos) e um conjunto de 10 atributos. Foi verificada então a interseção diferentes conjuntos de resultados e a posição dos atributos nos ranques, e foram selecionados os 8 atributos melhor posicionados.

	Melhores Atributos
1	URL Length
2	URL Depth
3	Has ATOM Feed

4	Has RSS Feed
5	Total Script Length
6	Total Script Pieces
7	META Tag Count
8	Average Script Length

Tabela 21 – Melhores atributos, selecionados através de métodos de seleção automatizada

Foi realizada a extração destes atributos dos conjuntos de páginas mencionados anteriormente, chegando aos conjuntos utilizados neste experimento.

A seguir, é exibida uma tabela com a acurácia de cada classificador para cada um dos diferentes conjuntos de páginas. Na seqüência, têm-se as matrizes de confusão para os mesmos.

	Treino	**Teste1**	**Posts+Comments**	**Teste2_Gr**	**Teste2_Ing**
Árvore de Decisão	**95.93%**	**92.04%**	56.00%	**65.72%**	**49.42%**
Rede Neural	90.84%	89.38%	56.50%	53.77%	47.49%
SVM	80.67%	75.91%	**57.00%**	64.78%	46.72%

Tabela 22 – Acurácia por classificador por conjunto de informações. Em negrito estão sinalizados os melhores resultados para cada conjunto

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	231	2	12	3
Notícia	1	248	1	0
Portal de Notícia	13	6	226	0
Blog Posts	2	0	1	238
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	210	6	28	3
Notícia	4	242	0	4
Portal de Notícia	20	10	214	1
Blog Posts	3	11	0	227
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	205	5	36	1
Notícia	18	221	0	11
Portal de Notícia	49	13	183	0
Blog Posts	13	44	0	184

Tabela 23 - Matriz de confusão por classificador para o conjunto “Treino”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	229	1	19	4
Notícia	1	240	8	1
Portal de Notícia	29	10	213	3
Blog Posts	3	1	1	254
Rede Neural				

Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	217	12	21	3
Notícia	2	239	4	5
Portal de Notícia	29	17	206	3
Blog Posts	6	6	0	247
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	208	9	34	2
Notícia	24	206	2	18
Portal de Notícia	70	24	161	0
Blog Posts	16	46	0	197

Tabela 24 - Matriz de confusão por classificador para o conjunto “Teste1”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	69	2	26	3
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	21	32	4	43
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	56	9	31	4
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	20	21	2	57
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	79	1	20	0
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	18	42	5	35

Tabela 25 - Matriz de confusão por classificador para o conjunto “Posts + Comments”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	63	3	19	2
Notícia	5	72	13	5
Portal de Notícia	11	6	74	0
Blog Posts	0	30	15	0
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	55	2	20	10
Notícia	33	51	6	5
Portal de Notícia	16	10	65	0
Blog Posts	14	24	7	0
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	68	3	14	2
Notícia	0	64	28	3
Portal de Notícia	11	6	74	0
Blog Posts	0	29	16	0

Tabela 26 – Matriz de confusão por classificador para o conjunto “Teste2_Gr”

Árvore de Decisão

Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	38	0	15	4
Notícia	0	35	4	3
Portal de Notícia	6	3	41	0
Blog Posts	0	92	4	14
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	36	4	14	3
Notícia	1	37	4	0
Portal de Notícia	5	8	37	0
Blog Posts	6	90	1	13
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	51	1	3	2
Notícia	0	39	3	0
Portal de Notícia	10	9	31	0
Blog Posts	0	105	5	0

Tabela 27 - Matriz de confusão por classificador para o conjunto “Teste2_Ing”

A árvore de decisão se mostrou o melhor classificador para todos os conjuntos de dados, com exceção do conjunto “Posts+Comments”. Neste, o melhor classificador foi a SVM. A acurácia mais alta foi atingida pela árvore de decisão sobre o conjunto “Teste1”, e a mais baixa foi atingida pela SVM sobre o conjunto “Teste2_Ing”.

Para todos os conjuntos de informações, a acurácia dos classificadores se mostrou inferior a obtida quando foram utilizados todos os atributos. Apesar da seleção dos melhores atributos, teoricamente aqueles com maior poder preditivo, a redução do conjunto de atributos causou uma degradação na qualidade dos classificadores.

Um dos principais motivos desta degradação é que a seleção dos melhores atributos é realizada com relação a um conjunto de informações. Neste caso, foi selecionado o conjunto de treino para realizar a filtragem de atributos. Assim, os melhores atributos utilizados são os melhores quando consideramos o conjunto “Treino”, mas não necessariamente o são quando olhamos para os outros conjuntos. A seleção de melhores atributos, portanto, está atrelada ao conjunto de treino utilizado.

Embora tenha sido verificada uma redução nos níveis de acurácia obtidos, esta redução foi mínima, e pouco significativa. As grandes tendências observadas no primeiro experimento, de confusão de portais de notícias com outras classes, por exemplo, se mantiveram. Além disso, nos conjuntos “Teste2_Gr”, “Teste2_Ing” e “Posts+Comments”, a dificuldade de classificação dos posts se

tornou mais pronunciada com a redução do número de atributos. É interessante observar que eles foram mais confundidos com notícias do que com blogs ou portais de notícias.

5.5. Classificação Funcional com Atributos "Não-Tecnológicos"

Diversos dos atributos estruturais selecionados refletem a existência ou não de tecnologias específicas dentro das páginas analisadas, como a presença de links para feeds RSS ou ATOM. Embora estes atributos estejam entre os mais efetivos, segundo o processo de seleção de melhores atributos apresentado na seção 5.4, eles são transitórios, no sentido de que conforme a tecnologia de construção de páginas e distribuição de conteúdo for evoluindo, eles tendem a desaparecer ou perder sua eficácia.

O objetivo deste experimento, portanto, é observar o comportamento e a qualidade dos classificadores quando a classificação é realizada deixando de lado estes atributos relacionados com tecnologia.

A seguir é apresentada uma tabela com esse novo conjunto de atributos estruturais, já desconsiderando os tecnológicos. Ele é um subconjunto do original.

	Atributos Estruturais
1	URL Length
2	URL Depth
3	Tag Count
4	Average Tag Depth
5	Meta Tag Count
6	Anchor Tag Count
7	Link Tag Count
8	Style Tag Count
9	Script Tag Count
10	Image Tag Count
11	Total Text Pieces
12	Total Text Length
13	Average Text Length
14	Total Anchor Text Length
15	Average Anchor Text Length
16	Uses External Components
17	Comment Count
18	Total Script Pieces
19	Total Script Length
20	Average Script Length

Tabela 28 – Atributos estruturais "não-tecnológicos", ou seja, desconsiderados os atributos relacionados com tecnologias específicas.

Do conjunto original, 3 atributos foram removidos: O indicador de se a página possui ou não um link para um *feed* RSS, o indicador de se a página possui ou não um link para um *feed* ATOM e o indicador de se a página utiliza ou não a tecnologia de CSS. Esses 3 atributos possuem o maior relacionamento com tecnologias específicas da Internet, e, portanto, têm a maior chance de se tornarem irrelevantes ao longo do tempo.

Abaixo é apresentada a acurácia de cada classificador para os diferentes conjuntos de páginas, e as matrizes de confusão dos mesmos.

	Treino	Teste1	Posts+Comments	Teste2_Gr	Teste2_Ing
Árvore de Decisão	**97.46%**	87.61%	**79.50%**	63.21%	57.92%
Rede Neural	95.42%	**90.66%**	62.00%	64.78%	**58.30%**
SVM	85.15%	82.69%	52.50%	**66.98%**	57.92%

Tabela 29 – Acurácia por classificador por conjunto de informações. Em negrito estão sinalizados os melhores resultados para cada conjunto

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	242	1	3	1
Notícia	1	246	1	2
Portal de Notícia	11	3	231	0
Blog Posts	1	1	0	239
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	226	3	15	3
Notícia	0	244	1	5
Portal de Notícia	6	4	233	2
Blog Posts	0	6	0	235
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	216	6	25	0
Notícia	3	204	16	27
Portal de Notícia	27	11	207	0
Blog Posts	12	19	0	210

Tabela 30 - Matriz de confusão por classificador para o conjunto “Treino”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	217	4	27	5
Notícia	6	231	8	5
Portal de Notícia	43	12	195	5
Blog Posts	1	9	1	248
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	223	3	22	5
Notícia	1	232	4	13

Portal de Notícia	21	9	218	7
Blog Posts	0	10	0	249
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	207	10	36	0
Notícia	5	197	22	26
Portal de Notícia	24	15	214	2
Blog Posts	17	19	0	223

Tabela 31 - Matriz de confusão por classificador para o conjunto “Teste1”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	76	2	21	1
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	14	2	1	83
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	68	10	20	2
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	24	19	1	56
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	65	3	32	0
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	24	34	2	40

Tabela 32 - Matriz de confusão por classificador para o conjunto “Posts + Comments”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	66	2	18	1
Notícia	15	51	10	19
Portal de Notícia	12	5	74	0
Blog Posts	7	19	9	10
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	59	6	13	9
Notícia	16	47	10	22
Portal de Notícia	4	1	86	0
Blog Posts	6	15	10	14
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	67	2	15	3
Notícia	12	45	16	22
Portal de Notícia	3	6	82	0
Blog Posts	3	11	12	19

Tabela 33 – Matriz confusão por classificador para o conjunto “Teste2_Gr”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	50	2	4	1
Notícia	1	31	3	7

Portal de Notícia	5	5	40	0
Blog Posts	8	66	7	29
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	43	2	7	5
Notícia	0	36	3	3
Portal de Notícia	2	3	43	2
Blog Posts	2	75	4	29
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	39	2	15	1
Notícia	0	38	3	1
Portal de Notícia	2	8	40	0
Blog Posts	0	72	5	33

Tabela 34 - Matriz de confusão por classificador para o conjunto “Teste2_Ing”

A proximidade destes resultados aos obtidos na seção 5.1, com o conjunto original de atributos estruturais, mostra que bons resultados na classificação são possíveis mesmo sem a utilização de atributos relacionados com tecnologias específicas. Os classificadores obtidos neste experimento têm ainda uma vantagem em sua longevidade, uma vez que não perderão eficácia conforme o padrão de uso destas tecnologias se alterar.

Outro ponto interessante foi a elevada acurácia para o conjunto de “Posts+Comments”. A Árvore de Decisão, por exemplo, atingiu acurácia próxima a 80% neste conjunto, cerca de 10% superior a obtida anteriormente. Este resultado, em conjunto com as matrizes de confusão dos diferentes classificadores, permite perceber que a remoção dos atributos relacionados com tecnologia reduziu o grau de confusão entre as classes, melhorando especialmente a capacidade de diferenciação de posts e blogs.

Observando a distribuição dos atributos entre as classes, é possível perceber que os atributos tecnológicos são capazes de distinguir facilmente blogs e blog posts de notícias e portais de notícias, mas não são capazes de distinguir tão bem blogs e blog posts. Isto ocorre pois blogs que oferecem feeds RSS ou ATOM geralmente incluem nos seus blog posts links para estes feeds.

No caso dos conjuntos “Teste2_Gr” e “Teste2_Ing”, os resultados obtidos foram similares aos anteriores, sendo um pouco inferiores no primeiro e ligeiramente superiores no segundo. Para o primeiro conjunto, a principal dificuldade de classificação continua sendo a distinção de notícias e de posts das outras classes. No segundo, a dificuldade de distinção de notícias desaparece,

ficando apenas a confusão de posts. Essas duas dificuldades de classificação também aparecem nos outros conjuntos de atributos.

5.6. Classificação Funcional com Atributos Estruturais Refinados

Avaliando os classificadores gerados sobre os conjuntos de atributos estruturais utilizados até então, é possível observar dois pontos interessantes. O primeiro é que diversos dos atributos são redundantes, no sentido de que transmitem a mesma informação com formatos diferentes. O segundo é que alguns atributos poderiam ter uma representação mais significativa, que transmitisse informações mais concretas sobre a estrutura da página.

Examinando, por exemplo, para os atributos que medem o número absoluto de tags, é fácil perceber que estes atributos representariam melhor a página se fossem apresentados como um percentual do número total de tags contidas na página. Da mesma forma, os quantificadores de partes de texto e de scripts poderiam ser relativizados ao tamanho da página.

Por outro lado, olhando para os atributos relacionados com scripts, por exemplo, tem-se 3 atributos (número total de partes de script, tamanho total dos scripts e tamanho médio das partes de script) que informam essencialmente a mesma coisa, ou seja, o grau de utilização de scripts dentro da página sendo analisada.

O objetivo deste experimento, então é de realizar um refino e redução no conjunto de atributos estruturais, de forma a remover as redundâncias e otimizar a representação da estrutura das páginas, seguindo as linhas descritas acima.

Abaixo é apresentada uma tabela com o novo conjunto de atributos estruturais.

	Atributos Estruturais Refinados
1	URL Length
2	URL Depth
3	Average Tag Depth
4	Meta Tag Percent
5	Anchor Tag Percent
6	Link Tag Percent
7	Style Tag Percent
8	Script Tag Percent
9	Image Tag Percent

10	Text Percentage
11	Anchor Text Percentage
12	Script Percentage
13	Has Javascript
14	Has RSS Feed
15	Has ATOM Feed
16	Uses CSS

Tabela 35 - Atributos estruturais refinados.

Este conjunto de atributos é menor do que o conjunto original, com 16 atributos, ao invés de 23. Assim, ganha-se imediatamente em tempo de processamento para extração e, principalmente, no treinamento dos classificadores.

Os atributos deste conjunto também conseguem representar de forma mais direta a estrutura das diferentes páginas, evitando a duplicidade de informações e a redundância. Alguns atributos também foram explicitados, de forma a se tornarem mais aparentes para os classificadores e mais transparentes para quem estiver realizando uma análise.

A seguir é apresentada a acurácia de cada um dos classificadores para os diferentes conjuntos de informações, e as matrizes de confusão dos mesmos.

	Treino	**Teste1**	**Posts+Comments**	**Teste2_Gr**	**Teste2_Ing**
Árvore de Decisão	**97.86%**	91.05%	69.50%	64.47%	**57.53%**
Rede Neural	96.95%	**92.53%**	**71.00%**	61.64%	54.83%
SVM	86.78%	84.66%	57.00%	**67.92%**	50.19%

Tabela 36 – Acurácia por classificador por conjunto de informações. Em negrito estão sinalizados os melhores resultados para cada conjunto

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	238	2	5	2
Notícia	0	249	1	0
Portal de Notícia	4	3	238	0
Blog Posts	3	1	0	237
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	244	0	2	1
Notícia	2	233	3	12
Portal de Notícia	2	2	241	0
Blog Posts	4	2	0	235
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	225	3	18	1
Notícia	17	217	5	11
Portal de Notícia	11	11	223	0

Blog Posts	10	43	0	188

Tabela 37 - Matriz de confusão por classificador para o conjunto “Treino”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	227	3	19	4
Notícia	1	240	3	6
Portal de Notícia	19	13	220	3
Blog Posts	10	9	1	239
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	236	1	11	5
Notícia	0	220	9	21
Portal de Notícia	17	6	230	2
Blog Posts	1	3	0	255
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	235	6	11	1
Notícia	18	205	9	18
Portal de Notícia	18	15	222	0
Blog Posts	17	43	0	199

Tabela 38 - Matriz de confusão por classificador para o conjunto “Teste1”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	75	2	21	2
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	10	23	3	64
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	81	2	14	3
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	13	18	8	61
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	79	5	16	0
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	24	39	2	35

Tabela 39 - Matriz de confusão por classificador para o conjunto “Posts + Comments”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	70	4	11	2
Notícia	1	51	27	16
Portal de Notícia	3	5	83	0
Blog Posts	4	19	21	1
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	66	2	12	7
Notícia	7	44	38	6
Portal de Notícia	2	3	86	0
Blog Posts	5	24	16	0

SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	77	3	6	1
Notícia	4	53	35	3
Portal de Notícia	1	4	86	0
Blog Posts	5	29	11	0

Tabela 40 – Matriz de confusão por classificador para o conjunto “Teste2_Gr”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	46	3	5	3
Notícia	1	33	7	1
Portal de Notícia	1	4	45	0
Blog Posts	3	76	6	25
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	50	1	3	3
Notícia	0	35	5	2
Portal de Notícia	2	5	43	0
Blog Posts	6	86	4	14
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	52	0	3	2
Notícia	0	38	4	0
Portal de Notícia	2	8	40	0
Blog Posts	5	102	3	0

Tabela 41 - Matriz de confusão por classificador para o conjunto “Teste2_Ing”

Com este conjunto de atributos, não houveram ganhos significativos em termos de acurácia em comparação com o conjunto original de classificadores. Alguns apresentaram resultados melhores para determinados conjuntos de páginas, outros apresentaram resultados piores, mas, em linhas gerais, as mesmas dificuldades de classificação se apresentaram.

Olhando para os classificadores originais, a árvore de decisão apresentou melhora significativa em sua acurácia para o conjunto “Posts + Comments”, enquanto a SVM apresentou redução maior em sua acurácia para os conjuntos “Posts + Comments” e “Teste2_Ing”. Para esse último conjunto, a SVM apresentou os piores resultados em termos de acurácia exclusivamente devido a classe “Blog Posts”. Quando retiramos esta classe, a acurácia da SVM fica em 87,25% (130 de 149 páginas classificadas corretamente).

A árvore de decisão, devido à própria natureza do método de classificação, extrai benefícios do refino dos atributos. Quando atributos absolutos são transformados em percentuais, a diferenciação entre páginas é facilitada. Duas páginas da mesma classe podem ter um número completamente distinto de tags de

imagem, por exemplo, devido simplesmente a uma delas ser muito maior do que a outra. Ao mesmo tempo ambas podem ter o mesmo *percentual* de tags de imagem em seu código HTML. Com ambos os tipos de informação, é possível para a árvore de decisão aprender o relacionamento. No entanto, quando os atributos são absolutos, o relacionamento de percentuais de tags é representado por dezenas de nós (do tipo SE a página possui 20 tags E 2 de imagem, ...E 3 de imagem, etc.), e esses nós possuem poucos casos. Assim, um processo de poda da árvore quase certamente remove esses nós da árvore, enquanto um nó que utiliza um valor percentual possui muito mais casos, e portanto uma chance menor de ser podado. Portanto, é natural que o desempenho da árvore de decisão melhore com este refino de atributos.

No caso da rede neural e da SVM, não existe processo de poda, onde relacionamentos que representam poucos casos possam ser removidos do processo de classificação. Logo, não houveram grandes diferenças de desempenho entre a rede neural e a SVM treinada sobre este conjunto de atributos refinados e o conjunto de atributos estruturais original.

As matrizes de confusão mostram que a confusão entre posts com comentários e notícias persiste, e continua sendo a principal causadora da baixa performance dos classificadores no conjunto "Teste2_Ing". Os erros de distinção entre as classes permanecem essencialmente os mesmos apresentados no conjunto original de testes.

5.7. Classificação Funcional Combinando Atributos Estruturais e de Texto (80%)

A análise dos resultados dos classificadores construídos nos experimentos anteriores demonstra que as dificuldades de classificação para os conjuntos de teste são distintas quando são considerados atributos de texto ou atributos estruturais. No caso dos atributos estruturais, os experimentos apontam uma dificuldade de distinção entre os posts e notícias. Por outro lado, no caso dos de texto, a dificuldade de distinção está entre blogs e blog posts e entre notícias e portais de notícias.

Uma vez que as dificuldades de classificação são não-coincidentes, ou seja, que os classificadores baseados em atributos estruturais cometem erros diferentes

dos baseados em atributos de texto, surge a possibilidade de que a combinação destes dois conjuntos de atributos resulte em uma maior acurácia na classificação.

O objetivo deste experimento é justamente validar a hipótese de que é possível se obter classificadores melhores através da combinação de atributos estruturais com atributos de texto.

Para realizar esta combinação, foi utilizado o conjunto de atributos estruturais extraído originalmente (devido aos resultados superiores apresentados por seus classificadores). Do lado dos atributos de texto, existiam duas opções: utilizar o conjunto de palavras que aparecem em pelo menos 80% das páginas de uma determinada classe, ou o conjunto das que aparecem em pelo menos 50%. Devido à dificuldade de treinamento da rede neural para o conjunto de 50%, e a pequena diferença nos resultados apresentados pelos outros classificadores para este conjunto, foi tomada a opção de utilizar o conjunto de 80%.

Apesar do conjunto "Teste2_Gr" não possuir nenhuma página em inglês, sua utilização neste experimento permite uma avaliação do quão dependente cada um dos modelos construídos é dos atributos de texto. Abaixo, é apresentada a acurácia de cada classificador para os diferentes conjuntos de páginas, e as matrizes de confusão dos mesmos.

	Treino	Teste1	Posts+Comments	Teste2_Gr	Teste2_Ing
Árvore de Decisão	**98.88%**	90.36%	74.5%	45.60%	54.44%
Rede Neural	96.64%	**94.40%**	**81.00%**	**65.72%**	**56.76%**
SVM	93.59%	93.02%	78.00%	62.58%	**56.76%**

Tabela 42 – Acurácia por classificador por conjunto de informações. Em negrito estão sinalizados os melhores resultados para cada conjunto

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	246	0	0	1
Notícia	1	248	1	0
Portal de Notícia	2	2	240	1
Blog Posts	1	1	1	238
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	238	0	8	1
Notícia	1	245	2	2
Portal de Notícia	9	6	229	1
Blog Posts	3	0	0	238
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	226	0	18	3
Notícia	3	245	0	2

Portal de Notícia	4	11	230	0
Blog Posts	11	11	0	219

Tabela 43 - Matriz de confusão por classificador para o conjunto “Treino”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	230	0	12	11
Notícia	10	226	12	2
Portal de Notícia	20	9	222	4
Blog Posts	10	7	1	241
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	238	0	9	6
Notícia	3	240	7	0
Portal de Notícia	12	10	229	4
Blog Posts	5	1	0	253
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	234	0	11	8
Notícia	3	243	3	1
Portal de Notícia	7	17	230	1
Blog Posts	15	5	0	239

Tabela 44 - Matriz de confusão por classificador para o conjunto “Teste1”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	91	2	3	4
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	24	18	0	58
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	85	1	12	2
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	21	2	0	77
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	88	1	11	0
Notícia	0	0	0	0
Portal de Notícia	0	0	0	0
Blog Posts	27	4	1	68

Tabela 45 - Matriz de confusão por classificador para o conjunto “Posts + Comments”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	47	5	21	14
Notícia	29	11	28	27
Portal de Notícia	5	2	84	0
Blog Posts	23	3	16	3
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	67	2	8	10
Notícia	12	41	24	18

Portal de Notícia	4	5	81	1
Blog Posts	10	7	8	20
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	55	3	22	7
Notícia	4	56	29	6
Portal de Notícia	0	6	85	0
Blog Posts	0	24	18	3

Tabela 46 – Matriz de confusão por classificador para o conjunto “Teste2_Gr”

Árvore de Decisão				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	47	3	3	4
Notícia	2	34	3	3
Portal de Notícia	4	7	39	0
Blog Posts	15	55	19	21
Rede Neural				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	43	3	9	2
Notícia	1	38	3	0
Portal de Notícia	2	6	42	0
Blog Posts	14	70	2	24
SVM				
Classificado Como >	**Blog**	**Notícia**	**Portal de Notícia**	**Blog Posts**
Blog	49	2	4	2
Notícia	1	37	4	0
Portal de Notícia	1	8	41	0
Blog Posts	12	77	1	20

Tabela 47 – Matriz de confusão por classificador para o conjunto “Teste2_Ing”

Para o conjunto de páginas “Teste1”, todos os classificadores apresentaram as melhores acurácias. A utilização destes atributos fez com que a acurácia dos classificadores ultrapassasse todos os resultados obtidos anteriormente, ainda que por uma pequena margem.

Sobre o conjunto “Posts+Comments”, a rede neural construída com estes atributos foi o único classificador a ultrapassar o nível de 80% de acurácia de classificação. A SVM também teve o seu melhor resultado de classificação com estes atributos, e a árvore de decisão teve seu segundo melhor resultado (o melhor tendo sido acurácia de 79,5% obtida no experimento exibido na seção 5.5).

A combinação dos atributos de texto com os atributos estruturais serviu para reduzir o nível de confusão entre blog posts e as outras classes, melhorando consideravelmente o desempenho dos classificadores para o conjunto “Posts+Comments”.

É interessante observar que, para o conjunto “Teste2_Gr”, embora a acurácia da árvore de decisão tenha sido muito baixa, a da rede neural e da SVM

foram relativamente altas, indicando que estes classificadores colocaram mais peso nos atributos estruturais do que nos atributos de texto. A pequena queda em sua acurácia pode ser explicada por confusão resultante dos atributos de texto que foram incluídos. A árvore de decisão, por outro lado, colocou mais peso nos atributos de texto, resultando em uma acurácia de classificação inferior.

Sobre o conjunto "Teste2_Inglês", os resultados foram mistos. A rede neural e a SVM atingiram resultados similares aos obtidos com o conjunto inicial de atributos, enquanto a árvore de decisão apresentou uma pequena piora. A matriz de confusão destes classificadores mostra que, embora não tenham ocorrido melhoras na capacidade de distinção de blog posts, houve uma melhora na capacidade de distinção de notícias das outras classes.

Outra observação interessante sobre as matrizes de confusão é de que, no conjunto "Teste1", a maior confusão de classificação ocorreu para os portais de notícias, que foram freqüentemente confundidos com notícias ou até mesmo blogs. Esta dificuldade reflete uma dificuldade semelhante apresentada pelos dois conjuntos de atributos que originaram este experimento, e tende a demonstrar que a confusão na classificação por estrutura dos portais é semelhante, ou ocorre para o mesmo conjunto de documentos, que a confusão na classificação por texto. Do contrário, seria esperada uma melhora significativa nestes erros de classificação.