6
Conclusões

Nesta seção, será discutido de forma global os resultados obtidos no capítulo anterior, combinando os resultados dos diferentes experimentos em algumas visões que permitem uma melhor comparação dos conjuntos de atributos, métodos de classificação e outros fatores que impactam a tarefa de classificação.

Serão apresentadas também as principais limitações dos resultados obtidos, e direções futuras de exploração do tema abordado.

6.1.
Seleção do Conjunto de Atributos

Abaixo, é apresentada uma tabela com a acurácia média para cada um dos diferentes conjuntos de informação, separadas pelos diferentes conjuntos de atributos. A acurácia média é a média das acurácias apresentadas por cada um dos 3 classificadores treinados para cada conjunto de atributos.

	Treino	Teste1	Posts+Comments	Teste2_Gr	Teste2_Ing
Estruturais	95.08%	91.54%	69%	**65.62%**	56.11%
Texto 80%	78.97%	69.98%	61.67%	30.08%	51.48%
Texto 50%	71.07%	63.65%	42.33%	31.45%	41.83%
Melhores	89.15%	85.78%	56.50%	61.42%	47.88%
Não Tecnológicos	92.68%	86.99%	64.67%	64.99%	**58.05%**
Refinados	93.86%	89.41%	65.83%	64.68%	54.18%
Estruturais + Texto	**96.37%**	**92.59%**	**77.83%**	57.97%	55.99%

Tabela 48 – Acurácia média por conjunto de atributos, segmentada em cada conjunto de informações. Em negrito estão sinalizados os melhores resultados para cada conjunto

Como neste trabalho foram utilizadas 4 classes, para os três primeiros conjuntos (que têm uma quantidade de páginas equilibrada de cada classe), a acurácia mínima esperada seria de 25%. Como todas as médias obtidas se apresentaram acima desta acurácia, é possível concluir que realmente ocorreu aprendizado.

Para os conjuntos “Teste1” e “Posts+Comments”, os atributos que apresentaram os melhores resultados, em termos de acurácia, foram os

combinados estruturais e de texto (palavras presentes em pelo menos 80% das páginas de cada classe). Para esses conjuntos de páginas, estes atributos resultaram nos melhores classificadores individuais e nas melhores médias de acurácia. Para o conjunto "Teste2_Gr", estes atributos resultaram em classificadores com acurácia reduzida, uma vez que os atributos vêm de páginas em inglês, e ele não possui nenhuma página em inglês. Finalmente, para o conjunto "Teste2_Ing", estes atributos resultaram em classificadores com a terceira maior média de acurácia, um resultado sólido.

A desvantagem deste conjunto de atributos é que, por fazer uso de atributos de texto, ele se torna dependente da língua na qual as páginas sendo classificadas foram escritas. Para trabalhos de classificação mais amplos, com páginas em múltiplas línguas, seria necessário algum tipo de tratamento especial ou descarte de páginas para que este conjunto de atributos fosse utilizado na classificação.

O conjunto de atributos estruturais original, no entanto, apresentou bons resultados e é independente de idioma, sendo assim mais indicado para iniciativas de classificação mais genéricas.

Todos os conjuntos que possuem atributos estruturais apresentaram resultados razoavelmente similares em termos de acurácia. Isso mostra que, mesmo sem a utilização de atributos relacionados com tecnologias específicas, é possível se obter bons resultados de classificação.

A vantagem de se ignorar estes atributos é que os classificadores desenvolvidos são mais robustos, uma vez que a disseminação ou o desaparecimento de tecnologias não afetarão a distribuição dos atributos na população.

O conjunto que apresentou os piores resultados de classificação foi o de atributos de texto presentes em pelo menos 50% das páginas de uma determinada classe. Os baixos resultados deste conjunto se devem especialmente a rede neural, que não foi treinada corretamente devido à quantidade de atributos no conjunto, e apresentou índices de acurácia extremamente baixos. No entanto, este conjunto de atributos resultou no melhor classificador para o conjunto de páginas "Teste2_Ing", uma SVM que atingiu 60, 23% de acurácia. A SVM treinada sobre este conjunto de atributos apresentou bons resultados em todos os conjuntos de páginas, exceto o conjunto "Teste2_Gr", devido às razões apresentadas na seção 5.3.

6.2. Classificadores

A seguir é apresentada uma tabela com a acurácia média de cada classificador para os diferentes conjuntos de atributos utilizados. Esta média é a soma das acurácias em cada um dos conjuntos de informação, dividida por 4 (“Teste1”, “Posts+Comments”, “Teste2_Gr”, “Teste2_Ing”).

	Árvore de Decisão	Rede Neural	SVM
Estruturais	69.66%	**72.69%**	69.36%
Texto 80%	**56.15%**	50.46%	53.29%
Texto 50%	**58.63%**	18.18%	57.63%
Melhores	**65.80%**	61.79%	61.10%
Não Tecnológicos	**72.06%**	68.94%	65.02%
Refinados	**70.64%**	70.00%	64.94%
Estruturais + Texto	66.23%	**74.47%**	72.59%

Tabela 49 – Média de acurácia por classificador para cada um dos conjuntos de páginas utilizados nos experimentos. Em negrito estão sinalizados os melhores resultados para cada conjunto

A rede neural apresentou os melhores resultados globais, quando aplicada no conjunto de atributos estruturais combinado com atributos de texto. A rede neural construída em cima apenas de atributos estruturais atingiu também uma acurácia de aproximadamente 75%, bastante significativa. Para os conjuntos apenas de atributos de texto, no entanto, as redes neurais desenvolvidas não apresentaram os melhores resultados. Para o conjunto de atributos de texto 50%, a rede neural construída teve o pior resultado de todos os classificadores, não conseguindo ultrapassar nem mesmo os 25% de acurácia de um classificador simples.

Através da tabela 49, é possível observar que tanto a árvore de decisão quanto a SVM apresentam pequena variação nos resultados para os diferentes conjuntos de atributos. As árvores de decisão apresentaram ainda as melhores acurácias médias para a maioria dos conjuntos de atributos. Associando isso a maior facilidade de explicação do processo decisório das árvores de decisão, é possível concluir que árvores de decisão são os classificadores mais indicados para esta tarefa de classificação. São também o classificador com processo de treinamento mais rápido.

A seguir, é apresentada uma tabela com a acurácia global de cada um dos classificadores para os diferentes conjuntos de atributos. Esta acurácia é o percentual de páginas classificadas corretamente, considerando todos os conjuntos de páginas como um único grande conjunto.

	Árvore de Decisão	Rede Neural	SVM
Estruturais	83.33%	**85.38%**	81.45%
Texto 80%	**66.33%**	58.62%	59.13%
Texto 50%	71.84%	26.94%	**72.42%**
Melhores	**82.36%**	79.08%	72.13%
Não Tecnológicos	81.49%	**82.28%**	77.06%
Refinados	83.22%	**83.87%**	78.00%
Estruturais + Texto	80.95%	**86.39%**	84.59%

Tabela 50 – Acurácia global por classificador para cada um dos conjuntos de atributos utilizados. Em negrito estão sinalizados os melhores resultados para cada conjunto

A tabela 50 mostra que o classificador com maior acurácia global foi a rede neural construída sobre o conjunto de atributos estruturais e de texto, seguida da rede neural construída sobre apenas atributos estruturais. 14 dos 21 classificadores apresentaram acurácia global acima de 75%, e 11 deles apresentaram acurácia acima de 80%.

Através destes números, é possível observar a qualidade dos classificadores construídos. Embora as acurácias médias sofram alguma distorção devido a conjuntos de teste construídos especificamente para dificultar a classificação, os classificadores apresentaram bons resultados globais.

6.3. Classes de segmentação

Ao longo dos experimentos realizados, algumas das classes selecionadas apresentaram características peculiares que dificultaram a sua classificação. Tanto os classificadores baseados em atributos estruturais quanto os classificadores baseados em atributos de texto mostraram dificuldades em separar blog posts das outras classes.

No caso dos baseados em atributos estruturais, a confusão ocorre entre os blog posts e as notícias. Essa dificuldade está relacionada com a similaridade estrutural destas duas classes, especialmente para posts mais longos ou com muitos comentários, que passam a ter grandes fatias de texto e muitos links em

seu corpo, de forma similar as notícias tradicionais. Outro fator que aumenta esta dificuldade é de que cada vez mais blogs são utilizados ou substituem veículos de informação tradicionais, e os seus posts se tornam equivalentes a notícias em jornais tradicionais, utilizando apenas um estilo de escrita distinto.

Para os classificadores baseados em atributos de texto, a confusão ocorre entre posts e blogs (e, de forma menos acentuada, entre notícias e portais de notícias). Aqui, a explicação é mais simples. Todo página principal de um blog possui pedaços de posts ou até mesmo posts completos. Quando é realizada a análise do texto da página, o mesmo texto é analisado para os blogs e para os posts, resultando em conflito para os classificadores.

A dificuldade da distinção de notícias do conjunto de páginas "Teste2_Gr" é interessante pois ocorre para todos os classificadores construídos, tanto os estruturais quanto os baseados em atributos de texto. Ao mesmo tempo, esta dificuldade não aparece nos outros conjuntos de páginas. Como a única diferença entre o conjunto "Teste2_Gr" e os outros é o idioma, e o treinamento de todos os classificadore foi realizado apenas sobre páginas em inglês, esta dificuldade sugere que a segmentação de notícias de idiomas diferentes do inglês é mais complexa.

6.4. Extração de Atributos

A estratégia de extração de atributos de texto utilizada foi uma contribuição original deste trabalho. Não foi realizado nenhum tratamento especial como remoção de stopwords, extração de radicais das palavras e outras técnicas relacionadas. A lógica por trás de não realizar a filtragem e remoção de stopwords é de que palavras que poderiam ser filtradas pelo processo seriam importantes na diferenciação de páginas como blogs e posts.

A avaliação das árvores de decisão construídas para o conjunto de atributos de texto 80% e para o conjunto que combina atributos estruturais com estes mesmos atributos de texto demonstra a validade deste raciocínio. Abaixo, é exibida uma fotografia destas árvores de decisão.

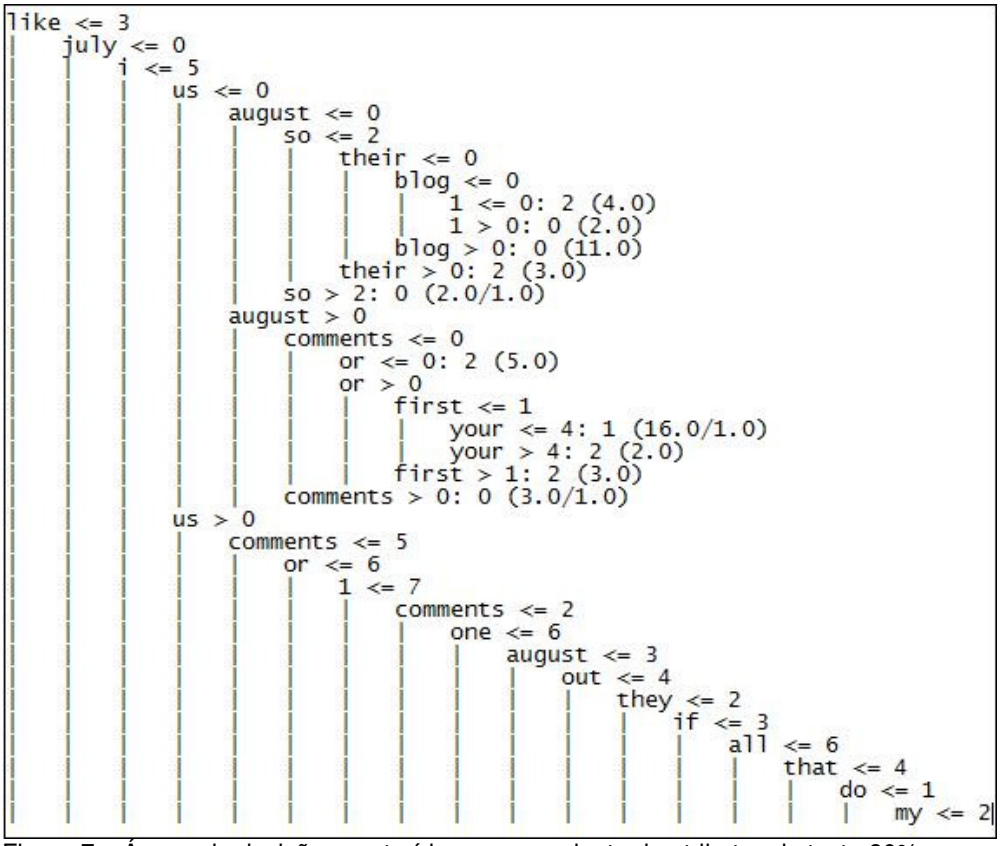

Figura 7 – Árvore de decisão construída para o conjunto de atributos de texto 80%

```
URL Length <= 48
|   i <= 4
|   |   Has ATOM Feed <= -1
|   |   |   Has RSS Feed <= -1
|   |   |   |   Total Script Length <= 744: 3 (4.0/1.0)
|   |   |   |   Total Script Length > 744: 0 (21.0)
|   |   |   Has RSS Feed > -1: 1 (4.0/1.0)
|   |   Has ATOM Feed > -1
|   |   |   URL Depth <= 2
|   |   |   |   comments <= 1
|   |   |   |   |   will <= 9
|   |   |   |   |   |   us <= 0
|   |   |   |   |   |   |   blog <= 3
|   |   |   |   |   |   |   |   comments <= 0: 2 (19.0)
|   |   |   |   |   |   |   |   comments > 0
|   |   |   |   |   |   |   |   |   Uses CSS <= -1: 2 (3.0)
|   |   |   |   |   |   |   |   |   Uses CSS > -1: 0 (2.0)
|   |   |   |   |   |   |   blog > 3: 0 (2.0)
|   |   |   |   |   |   us > 0: 2 (164.0)
|   |   |   |   |   will > 9
|   |   |   |   |   |   5 <= 0: 0 (2.0)
|   |   |   |   |   |   5 > 0: 2 (3.0)
|   |   |   |   comments > 1
|   |   |   |   |   Has RSS Feed <= -1
|   |   |   |   |   |   my <= 1: 0 (12.0/1.0)
|   |   |   |   |   |   my > 1: 2 (2.0)
|   |   |   |   |   Has RSS Feed > -1: 2 (13.0)
|   |   |   URL Depth > 2
|   |   |   |   Avg Tag Depth <= 8.794498: 1 (22.0/1.0)
|   |   |   |   Avg Tag Depth > 8.794498
|   |   |   |   |   blog <= 8: 2 (18.0/2.0)
|   |   |   |   |   blog > 8: 0 (2.0)
|   i > 4
|   |   me <= 1
|   |   |   Has ATOM Feed <= -1: 0 (15.0/1.0)
|   |   |   Has ATOM Feed > -1
|   |   |   |   comments <= 3
|   |   |   |   |   Uses External Components <= -1
|   |   |   |   |   |   i <= 12: 2 (10.0)
|   |   |   |   |   |   i > 12: 0 (2.0)
|   |   |   |   |   Uses External Components > -1: 0 (2.0)
|   |   |   |   comments > 3: 0 (7.0)
|   |   me > 1
|   |   |   with <= 3
|   |   |   |   or <= 2: 3 (2.0)
|   |   |   |   or > 2: 0 (4.0)
|   |   |   with > 3: 0 (170.0)
```

Figura 8 – Ramo de árvore de decisão construída para o conjunto de atributos que combina atributos estruturados com atributos de texto

Olhando para estas duas árvores, é imediatamente perceptível que palavras como “I”, “us” e “me”, que seriam filtradas como stopwords em um processo comum, realizam um papel importante na classificação, ou seja, são atributos fortemente preditivos para as classes selecionadas. Isso valida a estratégia de seleção de atributos adotada.

Indo além, os bons resultados obtidos mesmo com os classificadores de texto demonstram que a estratégia de seleção de atributos não traz nenhum prejuízo para a classificação, além de ser mais eficiente, uma vez que consome uma quantidade menor de recursos computacionais e pode ser facilmente adaptada

para outros conjuntos de treino, sem a necessidade de ajustes em listas de exclusão ou de execução de métodos especiais de definição de stopwords para os conjuntos.

6.5. Limitações

Os resultados obtidos ao longo deste trabalho não são, nem se propõe a ser, definitivos e completos. As classes observadas compõe apenas uma pequena fração de todas as classes e páginas existentes hoje na Web. Portanto, as conclusões alcançadas sobre os melhores métodos de classificação e sobre os melhores conjuntos de atributos podem não se sustentar quando novas classes forem investigadas.

Durante a construção dos classificadores, foram utilizados os parâmetros pré-configurados em todos os classificadores. Desta forma, é possível que um estudo mais profundo dos diferentes métodos de classificação permita a melhor seleção dos parâmetros de treinamento e, portanto, alcançar resultados melhores dos que os aqui apresentados.

Os conjuntos de atributos utilizados, embora abrangentes, não são exaustivos. Uma série de outras possibilidades de combinação de atributos ou variações nos parâmetros de seleção faz com que exista quase uma infinidade de conjuntos de atributos a serem utilizados. Os 7 conjuntos utilizados nos experimentos deste trabalho, no entanto, se mostraram de simples coleta e trouxeram bons resultados.

6.6. Trabalhos Futuros

Uma primeira direção de estudo que se aponta de imediato é a exploração da construção de classificadores, utilizando apenas atributos estruturais, com conjuntos de páginas de múltiplas línguas. Todos os classificadores deste trabalho foram criados em cima de páginas na língua inglesa. Seria interessante observar o comportamento dos classificadores sobre conjuntos de teste mistos caso estes também fossem criados sobre conjuntos mistos.