Categorização de Textos da Língua Portuguesa com Árvores de Decisão, SVM e Informações Lingüísticas

Cassiana Fagundes da Silva[1], Renata Vieira[2]

[1]Faculdade Seama
Av. Nações Unidas, 1201 – Jesus de Nazaré – 68.900-000 – Macapá – AP – Brasil

[2] Universidade do Vale do Rio dos Sinos
Av. Unisinos, 950 – 93.022-000 – São Leopoldo – RS - Brasil

cassiana@gmail.com, renatav@unisinos.br

Abstract. *This paper compares Decision Trees (DT) and Support Vector Machines (SVM) for text categorization tasks using linguistic information. We show that linguistic knowledge is useful for term selection for both learning techniques. We also show that DTs perform better than SVM when a reduced number of terms is considered, and they stop improving at a certain point whle SVM continually improves the results when the number of terms increase.*

Resumo. *Este artigo compara Árvores de Decisão (AD) e Support Vector Machines (SVM) para tarefas de categorização de textos baseada em informação lingüística. Mostramos que o uso de conhecimento lingüístico é útil na seleção de termos relevantes nas duas técnicas de aprendizado. Além disso, os experimentos mostram que árvores de decisão possuem um desempenho melhor do que SVM para um número de termos reduzido, e estabilizam-se a partir de um certo ponto, enquanto que o SVM atinge melhores resultados consistentemente com o aumento do número de termos utilizados no aprendizado.*

1. Introdução

O grande volume de dados e informações disponíveis em diversos meios e para diferentes domínios, tem desafiado a habilidade dos seres humanos em interpretá-los e compreendê-los. Assim, para atender ao desafio está sendo desenvolvida uma nova geração de métodos e ferramentas computacionais que automática e inteligentemente processam e analisam grandes volumes de dados e informações. Sob esta perspectiva, um dos métodos consiste na Categorização de Textos, onde os textos (ou documentos) são organizados em categorias pré-definidas[1], de acordo com os conteúdos que os compõem [Sebastiani, 2002].

[1] Por exemplo, turismo, economia, lazer, esportes, etc.

Usualmente, nos processos de categorização de textos, a representação de documentos é baseada na abordagem *bag-of-words*, que ignora a ordem dos termos assim como qualquer informação de pontuação ou estrutural, mas retém o número de vezes que um termo aparece [Gonçalves; Quaresma, 2003]. Esta representação é considerada uma simplificação de toda a abundância de informações expressada por um documento, não fornecendo uma descrição fiel do conteúdo. Para o desenvolvimento de modelos de Recuperação de Informação (RI) mais ricos e que possuam uma melhor descrição do conteúdo, técnicas de Processamento de Linguagem Natural (PLN) podem ser utilizadas. É nesse contexto que o presente trabalho está inserido, e propõe o uso de análise morfossintática na seleção de características com o objetivo de incrementar o modelo de representação de documentos.

O objetivo do estudo consiste em avaliar a utilização de informações lingüísticas para seleção de características na etapa de pré-processamento de categorização de textos da língua portuguesa e comparar seus efeitos em relação a dois métodos de aprendizado distintos. Para isto, foi utilizado um *corpus* de pequeno porte e extraído dessa base características lingüísticas para o treinamento dos algoritmos de aprendizado. Consideramos Árvores de Decisão (AD) e Support Vector Machines (SVM), de forma a avaliar a performance e a praticidade destes métodos. Diversos experimentos foram realizados e resultados comparativos são apresentados.

Este artigo está organizado conforme segue. Na seção 2 são discutidos os trabalhos relacionados. A seção 3 apresenta os materiais e métodos utilizados na realização dos experimentos. Os experimentos e resultados obtidos são descritos nas seções 4 e 5. Por fim, as conclusões e trabalhos futuros são apresentados na seção 6.

2. Trabalhos Relacionados

A união de áreas como Mineração de Textos (MT) e PLN não é recente, muitos trabalhos têm sido realizados nessas áreas, entretanto ainda existe muito a pesquisar. A escolha do conhecimento lingüístico a ser utilizado é difícil devido à crescente e grande quantidade de conhecimentos disponíveis, bem como ao aumento de ferramentas e recursos que possibilitam adquirir tal conhecimento.

Aizawa [2001] apresenta um método para incorporar a técnica de PLN no processo de categorização de textos. Para isto, utiliza como seleção de termos um modelo de linguagem probabilístico que possibilita a extração automática de termos compostos no *corpus* anotado. Como resultado dessa abordagem é possível perceber que a utilização dos termos compostos para o aprendizado do categorizador com o uso do algoritmo SVM, em relação à abordagem tradicional *bag of words*, melhora a performance do classificador.

Gonçalves e Quaresma [2003] tendo como referência de análise um conjunto de documentos do português europeu comparam e analisam a performance do SVM com outros algoritmos de Aprendizado de Máquina (AM). O estudo revela por meio da associação entre a freqüência do termo do conteúdo referenciado e a abordagem *bag of words,* que não há variações significativas na performance dos algoritmos.

Moschitti e Basili [2004] apresentam um estudo detalhado de representações de documentos avançadas baseadas no processamento de linguagem natural (complexidade nominal, nomes próprios e sentidos das palavras). Nesse estudo foram

utilizadas quatro diferentes coleções de documentos em duas línguas (inglês e italiano). Os autores concluem que a união de mineração de textos e processamento de linguagem natural não apresenta bons resultados, porém o fazem, com base em um limitado uso de informação lingüística. Silva et al. [2004] apresentam um estudo sobre o impacto do uso de informações lingüísticas na etapa de pré-processamento para coleções de documentos da língua portuguesa do Brasil nas tarefas de mineração de textos.

Neste presente trabalho, apresentamos a comparação entre duas técnicas de aprendizado de máquina (AD e SVM) para categorização de textos com seleção de características baseada em informações lingüísticas. Realizamos experimentos com uma coleção de textos jornalísticos da língua portuguesa.

Na próxima seção, são descritos o *corpus*, as técnicas de aprendizado de máquina e as ferramentas utilizadas para extrair informações lingüísticas.

3. Materiais e Métodos

O processo de mineração de textos é divido em cinco grandes etapas. A primeira refere-se à coleta de documentos relevantes ao domínio do conhecimento a ser trabalhado. A segunda etapa compreende no pré-processamento[2] desses documentos preparando-os para serem representados em um formato adequado e assim serem submetidos aos algoritmos de processamento estatístico. Na terceira etapa, denominada de seleção dos dados, são identificados e selecionados os termos mais relevantes nos dados pré-processados. De posse desses termos, a quarta etapa possibilita a aplicação de técnicas de aprendizado de máquina, com a finalidade de descobrir padrões úteis e desconhecidos nos documentos. Finalmente, a última fase de avaliação e interpretação dos resultados é necessária para verificar se o objetivo proposto foi alcançado.

Nas subseções 3.1, 3.2 e 3.3 seguintes são detalhadas as técnicas de extração de conhecimento AD e SVM, e a coleção de documentos.

3.1. Árvores de Decisão (AD)

AD são amplamente utilizadas na área de Inteligência Artificial (IA). Sua construção é realizada a partir de um conjunto de exemplos utilizando um aprendizado não incremental. Geralmente, um conjunto de exemplos de treinamento é apresentado ao sistema de indução da árvore, esta por sua vez, baseia-se na divisão recursiva [Quinlan, 1986; 1993] e [Steinberg; Colla, 1995] do conjunto de exemplos de treinamento em subconjuntos mais representativos, utilizando a métrica de ganho de informação[3]. Após a construção da árvore, esta poderá ser utilizada para a classificação de novos exemplos, seguindo os mesmos atributos usados na sua representação. A classificação é feita percorrendo-se a árvore, até chegar à folha que determina a classe a que o exemplo pertence ou sua probabilidade de pertencer àquela classe.

[2] Responsável por obter uma estrutura, geralmente, no formato de uma tabela atributo-valor, que represente o conjunto de documentos.

[3] O ganho de informação é uma medida que índica a redução esperada na entropia de um conjunto de dados, causada pelo particionamento dos exemplos em relação a um dado atributo.

3.2 Support Vector Machine (SVM)

SVM são máquinas de aprendizagem desenvolvidas em 1992 [Boser et. al., 1992], cuja fase de aprendizado é realizada por meio de um treinamento supervisionado. Elas podem ser consideradas máquinas fundamentadas na Teoria de Aprendizagem Estatística e utilizam em sua formulação o Princípio de Minimização do Risco Estrutural [Vapnik, 1995]. Seu treinamento é realizado por intermédio da resolução de um QP (*quadratic programming*), que possui um custo computacional elevado.

A principal característica das SVM's é a determinação automática dos dados de treinamento mais relevantes para o problema abordado, chamados vetores de suporte. Outras referências sobre o assunto podem ser encontradas em Cristiani e Shawe-Taylor, [2000], Haykin [1999] e Burges [1998].

3.3. Descrição do Corpus

Em nosso experimento utilizamos uma coleção de documentos provenientes de um *corpus* composto por artigos do Jornal Folha de São Paulo do ano de 1994[4]. A partir desse corpus foram selecionados e classificados manualmente 855 documentos em 5 categorias[5] tais como: informática, imóveis, esporte, política e turismo. Em média um documento da coleção possui 215 palavras e 124 palavras distintas por textos, totalizando 19.519 palavras distintas.

4. Experimentos

Para a realização dos experimentos além da classificação manual do *corpus* foram necessários o pré-processamento e a preparação deste para aplicação dos algoritmos de AM adotados.

Na fase de pré-processamento dos documentos adotou-se para limpeza dos dados a remoção de termos irrelevantes e redução dos radicais. Para remoção dos termos foi utilizada uma *stop list*, contendo 476 palavras (tais como: artigos, preposições, verbos ser e estar, pronomes, entre outros) enquanto que para a redução de radical dos termos optou-se pelo algoritmo proposto por Martin Porter [1980] desenvolvido para a língua portuguesa[6]. As informações lingüísticas foram adquiridas utilizando um analisador sintático denominado PALAVRAS [Bick, 2000]. Este analisador sintático é bastante robusto, pois possibilita a análise sintática de sentenças incorretas ou até mesmo incompletas. Segundo Bick [2003], PALAVRAS apresenta baixas taxas de erro, ou seja, menos de 1% para classe de palavras e 3 ou 4% na análise sintática. As informações lingüísticas consideradas foram as de categoria gramatical: substantivos (*nn*); substantivos + adjetivos (*nn+adj*); substantivos + adjetivos + nomes próprios (*nn+adj+prp*); substantivos + nomes próprios (*nn+ prp*); adjetivos + nomes próprios (*adj+nn*); verbos + substantivos (*v+nn*); verbos + adjetivos + nomes próprios (*v+adj+prp*); verbos (*v*) e sintagma nominal (*sn*).

[4] Este corpus foi cedido pelo NILC (Núcleo Interinstitucional de Lingüística Computacional), ao grupo de pesquisa em PLN da Unisinos-RS. Disponível em http://www.inf.unisinos.br/~renata/laboratorio/mais_jornal_mt.htm

[5] Estas categorias foram escolhidas devido à consistência dos textos apresentados.

[6] Disponível em http://snowball.sourceforge.net

A seleção e redução dos atributos foram obtidas por meio da freqüência do termo e do método de truncagem[7]. Assim, vetores correspondentes a cada categoria foram construídos e os termos mais relevantes foram selecionados para compor os vetores globais. Para a construção desses vetores que compõem as bases de dados para o aprendizado, utilizamos diferentes números de termos: 30, 60, 90, 120, 150, 300, 500, 1000 e 1500. Classificadores baseados em AD e SVM foram treinados usando a ferramenta WEKA[8] [Witten; Frank, 2000]. Essa ferramenta é constituída por uma coleção de algoritmos de AM. Em nossos experimentos realizamos *10-fold cross-validation*[9].

As medidas de avaliação dos experimentos são baseadas nos erros de classificação observados para cada abordagem adotada, bem como nas medidas de precisão, abrangência e tempo de processamento para as melhores categorias.

5. Resultados

Primeiramente são mostrados os resultados obtidos no aprendizado com AD, seguidos dos obtidos com SVM e finalmente os dois são comparados. Analisando os resultados obtidos nos grupos com substantivos, conforme ilustra a Figura 4, pode-se observar que os melhores resultados são obtidos ao utilizarmos a combinação *nn+adj* e *nn+adj+prp*. Observou-se também que os resultados variam conforme o número de termos. Para o intervalo de termos entre 60 a 150, a melhor combinação é substantivos e adjetivos (*nn+adj*), com uma significância estatística superior a 95%. Porém, para um número maior de termos, a adição de nomes próprios melhora o resultado (*nn+adj+prp*).

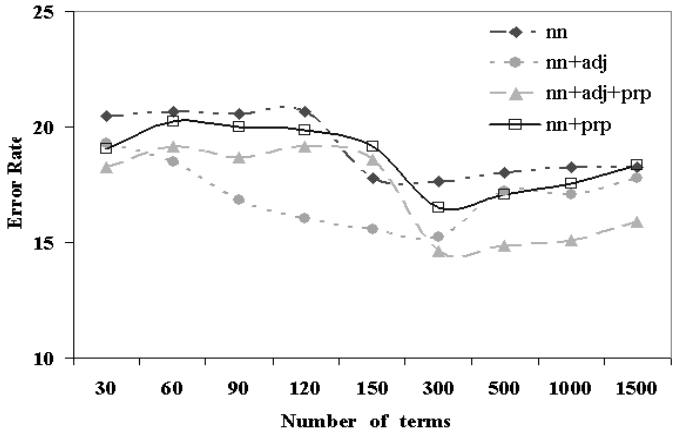

Figura 4. Taxa de Erro para o grupo de Substantivos – AD

[7] Este método permite selecionar os n termos mais relevantes de um conjunto de termos.

[8] Disponível em http://www.cs.waikato.ac.nz/ml/weka/

[9] Os métodos *k-fold cross-validation* consistem em dividir o conjunto de dados em dois conjuntos mutuamente exclusivos: um conjunto de treinamento que é fornecido a um sistema de aprendizado para a extração do conhecimento, e um conjunto de teste utilizado exclusivamente para medir a precisão do conhecimento induzido. Por exemplo, o método mais frequentemente utilizado, *10-fold cross-validation*, exige a criação de 10 pares de conjuntos de treinamento e teste.

Outro grupo gramatical testado foi o verbo com diferentes combinações. Na Figura 5, os resultados da combinação *v+nn* podem ser conferidos. De uma forma geral, o erro nessa combinação é superior ao menor erro obtido no grupo dos substantivos com adjetivos e nomes próprios.

Além das combinações utilizando verbos e substantivos, o sintagma nominal também foi testado. Para estes experimentos duas abordagens foram adotadas: uma incluindo todos os SNs, e outra apenas extraindo os sintagmas de tamanho menor ou igual a cinco termos (*sn-5*). A seleção considerando apenas de sintagmas com tamanho limitado apresenta melhores resultados do que a seleção de termos que considera todos os sintagmas. No entanto, se compararmos os *sn-5* com as menores taxas de erro das outras combinações observamos que a partir dos 90 termos ela é a opção que resulta no maior número de erros de classificação (Figura 5).

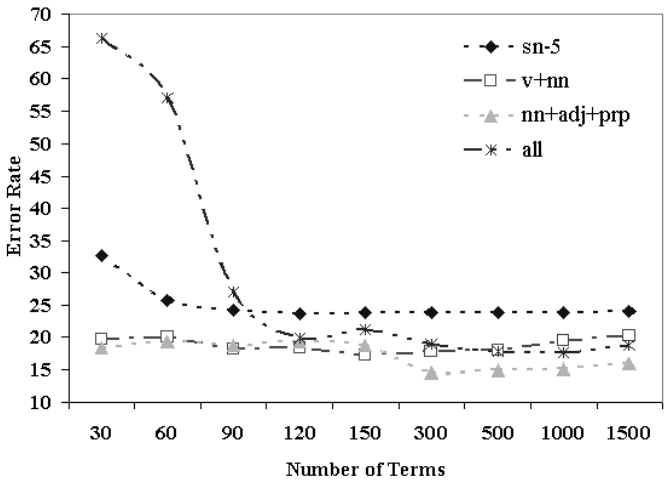

Figura 5. Comparação entre os grupos gramaticais - AD

Nos experimentos realizados com a técnica de aprendizado SVM (Figura 6), a melhor combinação para o grupo dos substantivos foi *nn+prp* apresentando os menores erros de classificação nos experimentos que empregam 500 termos ou mais, com significância estatística superior a 95%. Diferentemente da AD, o aumento do número de termos tem influência na redução dos erros de maneira contínua.

Cabe ressaltar que ambas as técnicas SVM e AD apresentam resultados similares nas classificações com base em 30 termos. Além disso, assim como para a AD, a utilização de sintagmas nominais completos como atributos não contribuiu para a classificação. Apesar dos sintagmas com restrição de tamanho apresentar melhores resultados do que os sintagmas completos. A partir dos 120 termos esse é o grupo com maior taxa de erro entre os demais para o SVM (Figura 6).

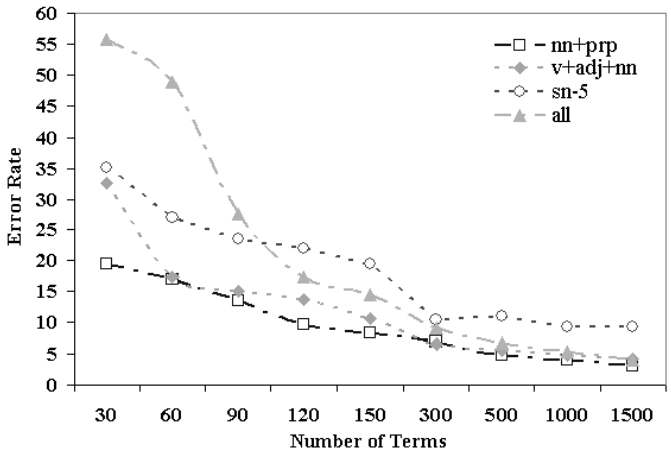

Figure 6. Comparação entre os grupos gramaticais - SVM

Para a classe *all,* os resultados obtidos com SVM apresentaram altas taxas de erro quando o número de termos é inferior a 90. Com o aumento do número de termos (até 1500) o erro diminui consideravelmente. No entanto, a combinação gramatical *nn+prp* é de uma maneira geral, a melhor de ser utilizada (significância de 99,5%). Percebe-se que a técnica de aprendizado SVM apresenta uma maior performance quando o número termos é elevado, enquanto que as AD são melhores no aprendizado com um número reduzido de termos. Gabrilovich e Markovitch [2004] apresentam uma discussão detalhada sobre essas diferenças.

A performance da AD e do SVM foram avaliadas com base no tempo de processamento para os experimentos que apresentaram as menores taxas de erro (combinação gramatical *nn+adj+prp*). Os experimentos foram executados por linha de comando na ferramenta Weka, no sistema operacional Linux em um micro-computador IBM Netvista - Penitum IV 1.8 Ghz, com 256 RAM e 40 GB de HD. Observando os tempos de construção do modelo, pode-se verificar que a técnica SVM apresenta menores tempos do que a técnica de AD. Em relação aos testes, as ADs apresentam uma grande oscilação. As ADs aumentam gradativamente o tempo conforme o aumento do número de atributos testados.

Com o intuito de identificar qual categoria do corpus apresenta maior precisão no processo de categorização foram calculadas as medidas de abrangência, precisão e *F-measure* [10] . Pôde-se constatar que os documentos e termos mais discriminantes pertencem à categoria Política com uma *F-measure* média de 0,982 utilizando a técnica de aprendizado SVM.

5. Conclusões

Normalmente os trabalhos da área de categorização de textos consideram abordagens estatísticas na construção da representação dos documentos, sem fazer uso de

[10] F-measure mede a capacidade de generalização do modelo gerado, permitindo verificar se durante o treinamento este assimilou do conjunto de treinamento características significativas que permitem um bom desempenho em outros conjuntos de dados, ou se concentrou em peculiaridades.

conhecimento lingüístico. Além disso, a maioria dos trabalhos utilizam bases de documentos da língua inglesa. Este trabalho apresentou experimentos em bases textuais de língua portuguesa, considerando diferentes combinações gramaticais para a seleção de termos relevantes.

Pôde-se observar que o uso de informações lingüísticas foram úteis no aprendizado de categorização de textos, em duas técnicas de aprendizado: AD e SVM. As melhores combinações gramaticais para ambos os algoritmos de aprendizado incluiu substantivos e nomes próprios (*nn+adj+prp* e *nn+prp*) e excluíram verbos e sintagmas nominais. Além disso, os experimentos mostram que AD possuem um desempenho melhor do que SVM para um número de termos reduzido, e estabilizam-se a partir de um certo ponto, enquanto que o SVM atinge melhores resultados consistentemente com o aumento do número de termos utilizados no aprendizado.

Como trabalho futuro planeja-se refazer o experimento com um corpus maior. Assim, será possível verificar se os métodos utilizados são eficazes para um grande conjunto de documentos.

Agradecimentos

Este trabalho foi realizado com apoio do CNPq e da Capes.

Referências

Aizawa, A. (2001). *Linguistic Techniques to Improve the Performance of Automatic Text Categorization* in Proceedings of NLPRS-01, 6th Natural Language Processing Pacific Rim Symposium, Tokyo, JP, 2001, pp. 307-314.

Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Gramatical Analysis of Porutugese in a Constraint Grammar Framework*. Århus University. Århus: Århus University Press.

Bick, E. (2003). *A Constraint Grammar Based Question Answering System for Portuguese.* Proceedings of the 11º Portuguese Conference on Artificial Intelligence, pages 414-418. LNAI Springer Verlag.

Boser, B. E.; Guyon, I. M. and Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers*. In D. Haussler, editor. Proceedings of the 5th Annual ACM Workshop on Computacionak Learning Theory, pp. 144-152. ACM Press, 1992.

Burges, C. J. C. (1998). *A tutorial on support vector machines for pattern recognition.* Data Mining and Knowledge Discovery, 2(2), pp. 121-167, 1998.

Christiani, N. and Shawe-Taylor, J. (2000). *An Introduction to Support vector Machines*. Cambridge U. P., 2000.

Gasperin, C.; Vieira, R.; Goulart, R. and Quaresma, P. (2003). *Extracting XML Syntactic Chunks from Portuguese Corpora.* Proc. of the TALN Workshop on Natural Language Processing of Minority Languages and Small Languages, pages 223-232. Batz-sur-Mer France.

Gonçalves, T.; Quaresma, P. (2003). *A preliminary approach to the multilabel classification problem of Portuguese juridical documents*. In F. Moura-Pires and S.

Abreu, editors, 11th Portuguese Conference on Artificial Intelligence, EPIA 2003, LNAI 2902, pages 435–444, Évora, Portugal, December 2003. Springer-Verlag.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. 2nd edition. Pretince-Hall, 1999.

Moschitti, A; Basili, R. (2004). *Complex Linguistic Features for Text Classification: A Comprehensive Study,* Volume 2997/2004 Title: Advances in Information Retrieval: 26th European Conference on IR Research, ECIR 2004, Sunderland, UK, April 5-7, 2004. Proceedings Editors: Sharon McDonald, John Tait

Porter, M. F. (1980). *An Algorithm for Suffix Stripping*. Program, 14(3): 130-137, 1980.

Quinlan, J. R. (1986). *Induction of Decision Trees. In Readings in Knowledge Acquisition and Learning*, Bruce G. Buchnan & David C. Wilkins, Morgan Kaufmann, pp. 349-361, 1986.

Quinlan, J. R. (1993). *C 4.5 : Programs for Machine Learning*. San Mateo: Morgan Kufmann Publishers, 1993.

Sebastiani, F. (2002). *Machine learning in automated text categorization*, ACM Computing Surveys, 34 (2002), 1-47.

Silva, C.F, Vieira, R., Osório, F. e Quaresma, P. (2004). *Mining linguistically interpreted texts*. In Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora, Geneva, Switzerland, August 2004.

Steinberg, D. and Colla, P. (1995). *CART: Tree-Strutured Non_Parametric Data Analysis*. Salford Systems, San Diego, CA. 1995.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

Witten, I. H. (2000). *Data mining: Pratical Machine Learning tools and techniques with Java implementations.* Academic Press, 2000.