

**PUC**

INF3006 – Exame de proposta de dissertação

Utilização de técnicas de aprendizado de

máquina para construção de um sistema de moderação automática de comentários

Silvano Nogueira Buback

Departamento de Informática

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

RUA MARQUÊS DE SÃO VICENTE, 225 - CEP 22453-900

RIO DE JANEIRO - BRASIL

Proposta de dissertação

Utilização de técnicas de aprendizado de

máquina para construção de um sistema de moderação automática de comentários

Silvano Nogueira Buback

snbuback@gmail.com

Proposta de Dissertação apresentada ao Programa de Pós-Graduação em Informática da PUC-Rio como requisito parcial para obtenção do grau de Mestre em Informática

Orientador: Marco Antonio Casanova

Sumário

1 Introdução 1

2 Estado da Arte 3

2.1 Filtro anti-spam 3

2.2 Sentiment Analysis 4

3 Proposta 6

3.1 Sistema de comentários 6

3.2 Sistema especialista 7

3.3 Verificação 7

4 Discussão e contribuições esperadas 9

5 Esboço do documento final 10

6 Plano de ação (esboço da solução) 11

7 Cronograma 13

# 

# Introdução

Com 2 bilhões de usuários (MINIWATTS MARKETING GROUP, 2011) e mais de 120 milhões de websites (DOMAINTOOLS.COM, 2011), a Internet já alcança 28,7% da população mundial. E para atrair cada vez mais usuários, dentre outras estratégias, os sites estão se tornando mais participativos. Isto pode ser conseguido, principalmente, com integração à redes socias, incentivo à contribuição de conteúdo, fóruns, listas de discussão ou comentários. Porém com a enorme quantidade de usuários interagindo, torna-se muito díficil controlar o conteúdo gerado. É provavel ocorrer alguns excessos, como o uso indevido para disseminação de spams, vírus, pornografia e outras mensagens indesejadas.

Para alguns sites a própria comunidade de usuários faz a moderação, através de funcionalidades como denuncie, mecanismos de reputação, dentre outros. O problema desta solução é que ela é pós-moderada, ou seja, a mensagem é sempre exibida e, de acordo com a classificação dos usuários, ela podera ser removida. Porém até que isto ocorra vários usuários já a leram.

Entretanto existem sites em que uma única mensagem indesejada, mesmo que entre muitas, não é sequer admissivel. Imagine um comentário pornográfico um site construido para crianças. Certamente, os mecanismos de pós-moderação não serviriam para este caso. Para alguns sites é necessário que a mensagem seja pré-moderada, ou seja, antes que qualquer usuário do site possa vê-la um moderador, autorizado pelo site, vai validá-la e aprovar sua publicação caso o conteúdo seja compativel. Porém, para sites com grande volume de acessos a quantidade de moderadores necessários para fazer toda a moderação pode inviabilizar o negócio ou muitas contribuições serão deixadas de lado, desmotivando os usuários de participar. Para resolver este problema, o objetivo da dissertação será construir um sistema de moderação de comentários automático, capaz de aprender novos padrões automaticamente através de um moderador humano e auxiliá-lo na tarefa de moderação.

A dificuldade para a realização deste trabalho está na identificação do critério do que deve ser aprovado ou reprovado, pois normalmente ele é subjetivo. As vezes é necessário um entendimento do contexto para avaliar se um determinado comentário poderá ser aprovado ou não. Por exemplo, o comentário feito através do Twitter na página da novela passione durante a campanha eleitoral:

Lá vem o Serra no comercial de #passione

Este comentário foi bloqueado apesar de não possuir nenhuma palavra obscena, pois se tratava de um comentário sobre política e certamente a emissora de TV não iria permitir comentários políticos na página da novela, especialmente durante a campanha, pois precisa manter a imparcialidade.

Apesar da palavra Serra ser um indicativo de que o comentário devesse ser reprovado no exemplo acima, o mesmo não deveria ocorrer se o comentário aparecesse em uma página de debate político. Ou seja, é necessário ainda que o sistema leve em consideração o contexto que o comentário aparece para ser mais preciso na classificação. Porém não é objetivo do trabalho verificar se o comentário pertence ao contexto em que foi inserido, somente ser especializado ao contexto no momento da moderação.

No exemplo acima o sistema de moderação precisa ser treinado para identificar os padrões do que deve ser aprovado ou reprovado. Para isto é necessário ter uma quantidade razoável de comentarios (esta quantidade será determinada no trabalho). Enquanto isto, comentários com palavras obscenas já podem ser reprovadas (blacklist), inclusive com a variações de escrita normalmente utilizadas para burlar o sistema. Por exemplo, merda e m#E#r#d#a deveriam ser igualmente identificadas.

O restante deste documento é organizado da seguinte maneira. A seção 2 descreve trabalhos relacionados. Na seção 3 é detalhado o escopo do trabalho. Na seção 4 está listado as contribuições esperadas ao estado da arte. Na seção 5 será apresentado uma proposta da organização da dissertação. Na seção 6 será descrito os passos para a construção do sistema de moderação automático, descrevendo inclusive um esboço da arquitetura. Por fim, a seção 7 será apresentado um cronograma para o desenvolvimento do trabalho.

# Estado da Arte

Não foram encontrados trabalhos diretamente relacionados a auto-moderação de comentários. Contudo este trabalho possui uma semelhança com trabalhos que tratam de filtros anti-spam. Como muito deles estão ligados a linguagem natural utilizando técnicas de aprendizado de maquina, também serão estudados trabalhos sobre analise sentimental. Outro motivo para estudar analise sentimental é por que, empiricamente, os trabalhos mais recentes de aprendizado de máquina são sobre este tema e mais trabalhos comparando algoritmos diferentes parecem existir. Estes trabalhos servirão como guia para o desenvolvimento da dissertação e também ajudarão a construir um base-line.

## Filtro anti-spam

Dentro os artigos pesquisados, (ANDROUTSOPOULOS, KOUTSIAS, *et al.*, 2000) e (ANDROUTSOPOULOS, KOUTSIAS, *et al.*, 2000) são literaturas muito citadas. Nestes trabalhos os autores mostram como utilizar Naive Bayes (LEWIS, 1998) para construção de um filtro anti-spam, qual o impacto que a seleção dos atributos (ou features como chamaremos aqui) possui no resultado final. Como métrica é utilizado uma acurácia com peso λ para mensagens legitimas que são classificadas incorretamente como spam.

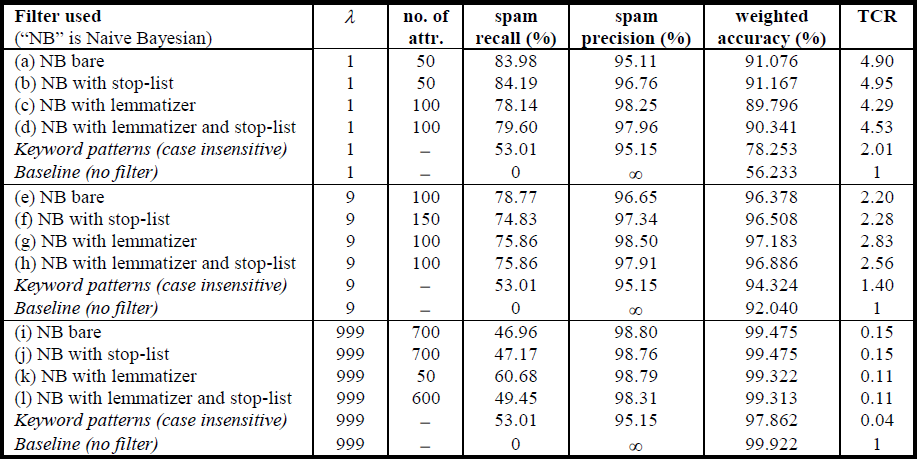


Tabela 1 - Resultados obtidos em (ANDROUTSOPOULOS, KOUTSIAS, *et al.*, 2000) e (ANDROUTSOPOULOS, KOUTSIAS, *et al.*, 2000).

A utilização de um peso para medir a acurácia é útil pois os autores entendem que classificar como spam uma mensagem legítima é mais grave que classificar como mensagem legítima um spam. Este aspecto é muito importante para o trabalho de auto-moderação de comentários, porém o peso λ possui um efeito inverso. Neste caso é mais grave aprovar uma mensagem que deveria ser reprovada, pois tal ação pode ter um efeito muito negativo para a página que ela está sendo exibida. Se uma mensagem aprovada fosse reprovada, ainda poderíamos contar com a ajuda manual do moderador.

Conforme já publicado por (JOACHIMS, 1998), o SVM tem se mostrado eficiente em tarefas de classificação de texto em geral. (LAI, 2007) fez um estudo sobre a acurácia do Naive Bayes, k-NN e SVM para classificação de spam, utilizando 2 corpus. Nos 2 corpus, o SVM mostrou-se mais eficiente que os outros algoritmos, chegando atingir acurácia de 94,59%, contra 81,59% do k-NN e 90,89% do Naive Bayes. Este resultados são os melhores resultados de cada algoritmo utilizando o corpus II, com 30% do corpus para teste. (DRUCKER, WU e VAPNIK, 1999) também publicou um estudo sobre o SVM e também concluindo que ele possui uma maior acurácia para a classificação de spam.

Para tentar melhorar ainda mais a acurácia, alguns trabalhos utilizam técnicas para combinar algoritmos. (CARRERAS e MÀRQUEZ, 2001) implementam uma variante do AdaBoost e consegue precisão de 98,73% com recall de 97,09%, para λ=1. (SAKKIS, ANDROUTSOPOULOS, *et al.*, 2001) combinam vários algoritmos com stacked generalization e consegue melhorar a acurácia de (ANDROUTSOPOULOS, KOUTSIAS, *et al.*, 2000), especialmente quando λ>1.

## Sentiment Analysis

Analisando trabalhos sobre classificação de texto, encontramos um tema de muito interesse atualmente: Sentiment Analysis. Este consiste em uma classificação de texto em que se está interessado em saber se o mesmo tem um sentimento positivo ou negativo (alguns trabalhos consideram também neutro). Este tema ganhou muita evidencia devido a proliferação das redes sociais. Ele permite, em larga escala, saber qual o sentimento dos usuários sobre um contexto definido, que pode ser um produto, pessoa, ou serviço.

O trabalho mais conhecido sobre Sentiment Analysis é (PANG, LEE e VAITHYANATHAN, 2002). Este trabalho parece ser o primeiro a utilizar técnicas de aprendizado de máquina para resolver o problema. Nele os autores mostram que contar palavras, selecionadas manualmente, que intuitivamente fornecem uma identificação sobre o sentimento (como brilhante, excelente ou terrível) não dão uma boa acurácia. A partir daí eles utilizam 3 algoritmos de aprendizado de máquina, Maxima Entropia (BERGER, PIETRA e PIETRA, 1996), Naives Bayes (LEWIS, 1998) e SVM (JOACHIMS, 1998), combinando técnicas diferentes para a seleção das features. Os autores mostram que o SVM utilizando como feature somente a presença ou não das palavras e símbolos de pontuação tem a melhor acurácia. A análise também é feita com Part of Speech (POS), somente de adjetivos e adicionado as palavras sua posição no texto. Os autores ainda argumentam que Sentiment Analysis parecem ser uma tarefa mais desafiadora que a tradicional classificação de texto em tópicos.

Mais recentemente, (ALVIM, VILELA, *et al.*, 2010) propuseram o uso do SVM com Bigrams e POS Tag conseguindo acurácia de 84,80% sob o mesmo corpus de (PANG, LEE e VAITHYANATHAN, 2002), com 3 folds e 86,09% com 10 folds, tornando-se o estado da arte para este corpus. Utilizando a mesma técnica num corpus de notícias da Petrobras (em portugues), os autores conseguiram acurácia de 84,00%, desta vez com 5 folds, porém utilizando somente SVM com presença a acurácia foi para 85,94%. Neste corpus o uso de Bigrams e POS Tag atingiu acurácia somente de 84,00%.

Outro trabalho proposto sobre sentiment analysis é (TURNEY, 2002). Uma característica deste trabalho que difere bastante dos demais é que ele utiliza um método não-supervisionado, ou seja, não é necessário um corpus para fazer o aprendizado. Em vez disto ele utiliza um algoritmo baseado em três etapas, a saber:

1. Consiste em identificar os adjetivos e advérbios, através de um algoritmo de POS Tag.
2. Calcular a orientação semantica de cada frase através de um algoritmo de PMI-IR (CHURCH e HANKS, 1990), porém fazendo acesso a um search engine para obter informações sobre cada palavra selecionada na primeira etapa e o quão mais aproximada de *excellent* ou *poor* ela estaria. A isto era atribuido um valor numérico
3. Por fim, era atribuido uma classe recomendado ou não-recomendado de acordo com a média da orientação semantica das frases calculadas no item acima.

Observe que este trabalho mostra uma heurística que não depende de um corpus para fazer o aprendizado (não-supervisionado), porém precisa de um acesso a um serviço web de busca para avaliar se o adjetivo ou advérbio em questão define uma boa avaliação ou não, o que poderia tornar o algoritmo inviável. Além disto, a acurácia obtida não é melhor que o método proposto por (PANG, LEE e VAITHYANATHAN, 2002), sendo de apenas 65,83% para filmes, porém a comparação não é feita com o mesmo corpus.

# Proposta

O objetivo da dissertação é mostrar que é possível construir um sistema especialista para auxiliar os moderadores na tarefa de moderação. O sistema de comentários já está desenvolvido e será sobre ele que o sistema especialista deverá atuar. Para entender como o sistema especialista deverá funcionar é necessário entender como o sistema atual organiza os comentários.

## Sistema de comentários

O sistema de comentários existente foi concebido para funcionar em um browser e não está acoplado a nenhum outro sistema. Neste sistema o assunto que está sendo comentado chama-se tópico. Portanto um tópico poderia ser uma matéria do G1, um post de um blog, uma receita culinária, enfim qualquer coisa que se queira comentar. Além dos comentários, o tópico contém a url e o título da página onde ele está inserido. Estas informações são úteis quando o moderador precisa consultar a página antes de aprovar ou reprovar um comentário.

Os tópicos são agrupados em editorias. A estrutura de editorias segue um modelo hieráquico. Os nós são separados por barra (/). A raiz da hierárquia é “/”. Veja o exemplo na figura 1:

Figura 1 - Exemplo da organização hierárquica das editorias e composição dos tópicos

No exemplo acima, “/” é a raiz e contém 2 nós, /g1 e /entretenimento. O primeiro possui mais 2 nós, chamados /g1/economia e /g1/educacao. Os tópicos podem ser associados a qualquer nó da editoria, não estão limitados somente as folhas. /g1/educacao possui 2 tópicos, sendo que o primeiro possui 2 comentários, enquanto que o segundo somente 1.

Somente 2 perfis de usuário são relevantes para o entendimento deste trabalho:

* Usuário comum: qualquer usuário que possa fazer um comentário. O usuário precisa estar logado no sistema para ser identificado;
* Usuário moderador: É responsável por fazer a moderação dos comentários feitos pelos usuários comuns.

## Sistema especialista

Apresentado o sistema atual, a seguir temos as funcionalidades que serão implementadas para a construção do sistema especialista:

* Capacidade de aprender o critério de moderação através do moderador humano (aprendizado supervisionado) e assim auxiliá-lo, moderando automaticamente ou apenas priorizando os comentários que possuem mais chances de serem aprovados.
* Utilizar a hierarquia de editorias para conseguir ter uma especialização sobre o que pode ou não ser aprovado, sendo capaz de entender que a palavra Serra poderia ser uma palavra proibida dentro da editoria /entretenimento/novelas, mas perfeitamente aceitável dentro da editoria do /g1.
* Bloquear palavras impróprias, inclusive tentativas de burlar sua escrita, como merda e m#E#r#d#a.
* Quando os usuários comuns denunciam um comentário ou quando moderador altera seu status, o sistema deverá utilizar este feedback para aprimorar o aprendizado.

## Verificação

Para construir um aprendizado e fazer a validação do trabalho é necessário um conjunto de treino (aprendizado) e um conjunto para testes. Tenho disponível para estas tarefas um conjunto com 450.845 comentários, com texto, status (aprovado ou reprovado) e a categoria. Estas categorias deverão ser agrupadas através de um de-para com as editorias existentes no sistema de comentários atual, pois este corpus não foi construido sobre o sistema de comentários e sim sob outro sistema que também faz moderação e busca os comentários do Twitter. A tabela 2 mostra o total de comentários por categoria e status.

|  |  |  |
| --- | --- | --- |
| Categoria | Aprovados | Reprovados |
| Araguaia | 2.581 | 4.106 |
| Bom Dia Brasil | 128 | 416 |
| Caldeirão de Twittadas | 2.344 | 7.722 |
| Fantástico | 1.493 | 44.902 |
| G1 | 2.176 | 8.434 |
| G1 Carros | 56 | 1.051 |
| G1 Eleições | 889 | 2.566 |
| G1 Pop e Arte | 7 | 78 |
| G1 RJ | 31 | 240 |
| G1 SP | 176 | 3.804 |
| Globo News | 152 | 1.563 |
| Globo Repórter | 476 | 1.681 |
| Hipertensão | 4.597 | 21.920 |
| Inscrições BBB 11 | 26.361 | 86.523 |
| Jornal Hoje | 219 | 1.622 |
| Junto & Misturado | 1.414 | 13.288 |
| Passione | 19.110 | 97.774 |
| Ti-Ti-Ti | 39.781 | 49.575 |
| TV Garagem | 799 | 790 |
| Total | 102.790 | 348.055 |

Tabela 2 - Total de comentários por categoria e status

Para o aprendizado, 70% do corpus será utilizado. O restante, 30%, será utilizado para fazer o teste e validação da eficiência da auto-moderação. Para medir esta eficiência vamos utilizar algumas métricas, já conhecidas na área de aprendizado de máquina. Elas serão cálculas sempre sobre o conjunto de teste.

* Acurácia (WIKIMEDIA FOUNDATION, INC., 2011a): Mede a quantidade de resultados corretos na população inteira. No caso deste trabalho, verificaremos qual a porcentagem de mensagens que o sistema conseguiu classificar corretamente, tanto como aprovado e reprovado. Pode ser representado pela fórmula:
* Precisão (WIKIMEDIA FOUNDATION, INC., 2011a): É dado pela fórmula:

Ou seja, quanto menos comentários reprovados o sistema classificar como aprovado maior será esta métrica. Análogamente, quanto mais comentários reprovados o sistema deixar passar menor será esta métrica.

* Recall (WIKIMEDIA FOUNDATION, INC., 2011a): É dado pela fórmula:

Ou seja, quanto menos comentários aprovados forem reprovados pelo sistema, maior será esta métrica. Análogamente, quanto mais comentários aprovados o sistema bloquear menor será esta métrica.

Além das métricas acima outras poderão ser utilizadas durante o trabalho e serão definidas de acordo com seu uso.

# Discussão e contribuições esperadas

Apesar da semelhança que este trabalho possui com filtro anti-spam, existem algumas diferenças importantes:

* Podemos fazer paralelo entre comentários reprovados e mensagens de spam. (ANDROUTSOPOULOS, KOUTSIAS, *et al.*, 2000) adicionou um custo λ para mensagens legitimas que foram classificadas incorretamente como spam, porém para este trabalho a preocupação maior seria das mensagens reprovadas que foram classificadas como aprovadas. Matematicamente seria apenas utilizar valores de λ menores que 1, porém é necessário alterar a seleção de features e talvez os algoritmos para maximar a precisão. Conforme mostrado por (BUBACK, 2010) para o corpus de comentários, o recall foi maior que a precisão. Empiricamente isto pode ser entendido como o sistema ter uma maior tendência de aprovar as mensagens, não o contrário.
* Mostrar que a tarefa que moderação de comentários é, do ponto de vista da aprendizagem de máquina, mais complexa que a classificação de spam. Enquanto a classificação de spam parece ter um certo padrão devido a sua natureza, normalmente para envio de vírus, pornografia ou vendas de produtos, o que facilita a classificação, a moderação de comentários não é tão previsível, pois a decisão depende também do contexto em que ele está inserido.
* Avaliar como e quanto um aprendizado específico pode ajudar na tarefa de classificação, como o aprendizado especializado de acordo com a árvore de editorias dos sistema de comentários.

# Esboço do documento final

Esta seção apenas descreve a proposta de organização da dissertação final.

1. Introdução
2. Trabalhos relacionados
3. Resultados dos experimentos
4. Modelagem do sistema
5. Conclusão

Referências

# Plano de ação (esboço da solução)

Já foram feitos alguns experimentos com o corpus que será utilizado (BUBACK, 2010). A implementação inicial com o Naive Bayes e SVM deram os seguintes resultados:



Tabela 3 – Consolidação dos resultados obtidos em (BUBACK, 2010)

Como pode ser visto pela tabela 3 a acurácia levando em consideração o conjunto inteiro é de 80,1% para o Naive Bayes, com unigram e bigrams, e de 81,4% com o SVM, somente com unigrams.

Uma das técnicas que pretendo testar para melhorar o classificador são:

* Uso de stop-words
* Uso de apenas das features mais relevantes, conforme mostrado em (PERKINS, 2010), pode trazer bons resultados
* Filtros baseados em expressões regulares para proibir escrita de palavras impróprias (palavrões)
* Teste do corpus com o AdaBoost pois X relatou melhor acurácia com este algoritmo. Já existe implementação disponível na internet

Como é sabido, as técnicas de classificação não são imperativas na determinação da melhor classe, elas apenas estimam a probabilidade de uma determinada classe, mas normalmente se está interessado apenas naquela de maior probabilidade. Como o classificador é binário, podemos entender que a probabilidade será apenas dele ser reprovado, pois a probabilidade do comentário ser aprovado será a probabilidade complementar. Este conceito é muito importante no desenvolvimento deste trabalho pois no momento em que o sistema estimar a probabilidade de moderação com valores muito próximos de 50%, com uma incerteza α a ser determinada experimentalmente, o classificador não deverá moderar o comentário e deixar neste caso que um operador humano o faça. Naturalmente que esta moderação depois será utilizada para aprimorar o classificador. O valor α deverá ser ajustado para minimizar a chance do sistema aprovar incorretamente um comentário, mas ainda deverá permitir uma margem de erro pois um valor muito alto deixará muito comentários para serem moderador manualmente, tornando o sistema ineficiente.

Para conseguir obter um aprendizado com diferentes níveis de especialização de acordo com a editoria cada nó da árvore de editorias deverá ter seu próprio modelo de aprendizado. Assim, cada vez que um nó for treinar um comentário, além do próprio treino ele deverá submeter o comentário para treino do nó pai (ou editoria pai) e assim sucessivamente, até a raiz, que deverá possuir um modelo genérico de moderação. Quando uma moderação for feita, o nó raiz poderá questionar o resultado do nó pai e assim sucessivamente, e ter como resultado a classificação mais provável. Também, enquanto o nó raiz contrói seu modelo de aprendizado (um critério de número minimo de mensagens deverá ser definido), deverá obter o resultado da classificação do nó pai.

Arquiteturalmente, este sub-sistema deverá atuar como um “robô”, assincronamente, classificando os comentários de acordo com a probabilidade dele ser reprovado. Também poderá atuar juntamente com os comentáristas para fazer a moderação automaticamente, nas editorias assim especificadas, e levando em consideração um nível α de certeza. Isto será utilizado para repriorizar o trabalho para os comentários que possuem mais chances de serem aprovados.

# Cronograma

Nesta seção é apresentado apenas um cronograma especificando as tarefas que deverão ser feitas para a conclusão do trabalho de dissertação.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Jan | Fev | Mar | Abr | Mai | Jun |
| Proposta | X |  |  |  |  |  |
| Treino com AdaBoost |  | X |  |  |  |  |
| Treino com seleções de atributos diferentes |  | X |  |  |  |  |
| Determinação de α |  |  | X |  |  |  |
| Escrita do capítulo sobre Resultados |  |  | X |  |  |  |
| Implementação do modelo de aprendizado específico |  |  |  | X |  |  |
| Finalização da dissertação |  |  |  |  | X | X |

Referências Bibliográficas

ALVIM, L. G. M. et al. **Sentiment of Financial News:** A Natural Language Processing Approach. Proceedings of the 1st Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology. Buenos Aires: [s.n.]. 2010. 10-14 Mai 2010.

ANDROUTSOPOULOS, I. et al. **An evaluation of Naive Bayesian anti-spam filtering**. Proceedings of the workshop on Machine Learning in the New Information Age and 11th European Conference on Machine Learning. Barcelona, Spain: ECML 2000. 2000. p. 9-17.

ANDROUTSOPOULOS, I. et al. **An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages**. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information. New York, NY, USA: ACM. 2000. p. 160-167.

BERGER, A. L.; PIETRA, S. A. D.; PIETRA, V. J. D. A maximum entropy approach to natural language processing. **Computational Linguistics**, Cambridge, MA, USA, v. 22, n. 1, p. 39-71, Mar 1996. ISSN 0891-2017.

BUBACK, S. N. **INF2979 - Auto-moderação**. Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro, RJ, p. 8. 2010. disponível em: https://github.com/snbuback/auto-moderacao-svm.

CARRERAS, X.; MÀRQUEZ, L. **Boosting Trees for Anti-Spam Email Filtering**. Proceedings of RANLP-2001. Bulgaria: [s.n.]. 2001. p. 58-64.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Computational Linguistics**, Cambridge, v. 16, n. 1, p. 22-29, Mar 1990. ISSN 0891-2017.

DOMAINTOOLS.COM. Daily DNS Changes and Web Hosting Activity by DailyChanges.com. **DailyChanges.com**, 2011. Disponivel em: <http://www.dailychanges.com/>. Acesso em: 19 Jan 2011.

DRUCKER, H.; WU, D.; VAPNIK, V. N. Support Vector Machines for Spam Categorization. **Neural Networks, IEEE Transactions on**, v. 10, n. 5, p. 1048 -1054, Set 1999. ISSN 1045-9227.

JOACHIMS, T. Text categorization with Support Vector Machines: Learning with many relevant features. In: JOACHIMS, T. **10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings**. Dortmund: Springer Berlin / Heidelberg, v. 1398, 1998. p. 137-142.

LAI, C.-C. An empirical study of three machine learning methods for spam filtering. **Know.-Based Syst.**, Amsterdam, v. 20, n. 3, p. 249-254, Abr 2007. ISSN 0950-7051.

LEWIS, D. Naive (Bayes) at forty: The independence assumption in information retrieval. In: NÉDELLEC, C. A. R. C. **10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings**. New Jersey: Springer Berlin / Heidelberg, v. 1398, 1998. p. 4-15. ISBN 978-3-540-64417-0.

MINIWATTS MARKETING GROUP. World Internet Usage Statistics News and World Population Stats. **Internet World Stats**, 2011. Disponivel em: <http://www.internetworldstats.com/stats.htm>. Acesso em: 19 Jan 2011.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. **Proceeding of the conference on empirical methods in natural language processing (EMNLP, 2002)**, Philadelphia, PA, USA, v. 10, p. 79-86, 6-7 Jul 2002.

PERKINS, J. Text Classification for Sentiment Analysis – Eliminate Low Information Features. **streamhacker.com**, 2010. Disponivel em: <http://streamhacker.com/2010/06/16/text-classification-sentiment-analysis-eliminate-low-information-features/>. Acesso em: 26 Jan 2011.

SAKKIS, G. et al. **Stacking classifiers for anti-spam filtering of e-mail**. Proceedings of "Empirical Methods in Natural Language Processing". Pittsburgh, PA: EMNLP 2001. 2001. p. 44-50.

TURNEY, P. D. **Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews**. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania: Association for Computational Linguistics. 2002. p. 417-424.

WIKIMEDIA FOUNDATION, INC. Accuracy and precision. **Wikipedia**, 2011a. Disponivel em: <http://en.wikipedia.org/wiki/Accuracy\_and\_precision#Accuracy\_and\_precision\_in\_binary\_classification>. Acesso em: 20 Janeiro 2011.