Classificador de Textos Otimizado Utilizando Lei de Potência para Palavras Raras

André Dieb Martins , Bruno B. Albert, E. C. Gurjão

***Resumo*— Nesse artigo é descrito o desenvolvimento de um sistema para classificação automática de textos, baseado no classificador Bayes Ingênuo Multinomial. É apresentada uma versão modificada do algoritmo, denominado NB+, na qual a informação proveniente de palavras mais raras é melhor aproveitada. Para avaliar o sistema, um procedimento experimental é realizado utilizando o corpus 20Newsgroups-18828, pre-processado utilizando pesos TF-IDF e seleção de características sob o critério *k-melhores*. A versão modificada apresenta melhorias de 10% a 20% em todas as métricas (F_1, acurácia, precisão e cobertura) para a região entre 50 e 200 características. Ambas versões convergem para o mesmo desempenho nas demais regiões. A melhoria de desempenho na região de baixa densidade de características trás a tona novas oportunidades na construção de classificadores mais baratos e mais eficientes.**

***Palavras-Chave*— Classificador Bayes Ingênuo Multinomial, Lei de Potência para palavras raras, Aprendizado de Máquina.**

***Abstract*— This paper describes the construction of an automatic text classification system based on the classic Multinomial Naive Bayes classifier. It presents a modified algorithm, named NB+, in which the information of rare words is best regarded as important. To evaluate the system, an experimental procedure is performed on the 20Newsgroups-18828 corpus, treated with TF-IDf weighting and K-best feature selection. The modified version shows improvements from 10% to 20% on all metrics (F_1, accuracy, precision and recall) for 50 to 200 features. Both versions converge to the same performance on remaining quantity of features. By increasing the performance on the low feature density area, new opportunities are raised in order to achieve cheaper and better performing classifiers.**

***Keywords*— Multinomial Naive Bayes classifier, Rare words weight rule, Machine learning.**

I. Introdução

É interessante observar que, especialmente nas últimas duas décadas, os avanços tecnológicos tem permitido a criação e produção em massa de computadores cada vez mais velozes, com maior capacidade de armazenamento e ao mesmo tempo mais portáveis e utilizados cada vez mais conectados à Internet. Essa revolução introduz, de forma crescente e progressiva, níveis nunca vistos nos volumes de dados que são trafegados nas redes que interconectam esses computadores, ou por eles armazenados.

Acompanhando essa tendência, a área do Aprendizado de Máquina tem se mostrado um objeto de estudo de bastante interesse por pesquisadores ao redor do mundo. Essa área, como o nome sugere, trata da construção de sistemas capazes de aprender a partir de dados. Tais sistemas são úteis em inúmeros nichos, variando da classificação de mensagens instantâneas em redes sociais [1] até a predição de tendências de mercado para o varejo [2], dentre inúmeras outras aplicações.

André Dieb Martins, Pesquisador independente, Bruno Barbosa Albert, Edmar Candeia Gurjão¸ Departamento de Engenharia Elétrica, Universidade Federal de Campina Grande, Campina Grande-PB, Brasil, E-mails:andre.dieb@gmail.com,{albert, ecandeia}@dee.ufcg.edu.br

No contexto do Aprendizado de Máquina existe um problema denominado *classificação*, que consiste em determinar a quais categorias um determinado objeto se assemelha. Neste artigo, focamos no problema de classificação de texto [3], no qual deseja-se determinar as categorias de um texto ou de um trecho de texto. Por exemplo, tais sistemas podem ser utilizados na classificação de textos de noticiários [4]. A motivação por trás desse problema é a extração de conhecimento a partir de dados, possibilitando novas decisões de negócio, previsões de mercado, previsões de tendências, construção de sistemas inteligentes, dentre inúmeras outras implicações.

De uma maneira geral a classificação de textos apresenta vários desafios. Primeiro, é difícil obter conceitos abstratos das linguagens naturais usando poucas palavras chaves, pois existem várias maneiras de se representar conceitos similares, e uma mesma palavra pode ter vários significados. Além disso, a análise semântica ainda não está bem entendida, embora algumas técnicas tenham sido aplicadas com sucesso em domínios limitados. Segundo, por sua natureza inerentemente complexa, o problema da classificação apresenta uma alta dimensionalidade, acabando por tornar proibitiva a complexidade de certos algoritmos [5].

Nesse trabalho foi projetado e implementado um sistema de classificação de textos baseado no classificador Bayes Ingênuo (do inglês *Naive Bayes*). Neste classificador, foram introduzidas otimizações nas etapas de treinamento e de classificação. Além disso, foi aplicada uma regra de potência na classificação. A implementação foi testada utilizando o banco de dados *20Newsgroups-18828*, uma coleção de 18828 mensagens de grupos de notícias da rede UseNet (disponível no portal MLcomp.org), e apresentou melhorias consideráveis nos níveis de acurácia, precisão e cobertura do classificador.

O artigo está organizado da seguinte forma, na Seção II é feita uma descrição geral do algoritmo de Bayes Mutlinomial. Na Seção III está descrita a modificação proposta nesse trabalho. Na Seção IV são descritos os procedimentos experimentais para validar a modificação proposta. Na Seção V são apresentados os resultados obtidos e finalmente na Seção VI são apresentadas as conclusões e perspectivas de desenvolvimento para este trabalho.

II. Algoritmo de Aprendizado Bayes Ingênuo Multinomial

O Bayes Ingênuo multinomial é um algoritmo de aprendizado supervisionado bastante comum na literatura de Aprendizado de Máquina. Apesar de sua simples construção, apresenta bons resultados, sendo assim um interessante objeto de estudo. Assim como outros métodos supervisionados, este baseia seu aprendizado num modelo probabilístico.

Pode-se notar que a tarefa de classificação, quando executada por humanos em documentos de texto, se dá pela observação de agrupamentos, sequências, frequências de palavras localizadas no texto. Sendo assim, podemos ver que o processo decisório humano é intrinsecamente complexo, sendo difícil simulá-lo ou até mesmo modelá-lo devido à sua subjetividade.

Este algoritmo, no entanto, supõe uma simplificação do processo de decisão na qual as palavras e suas frequências são o único fator decisivo na determinação da classe de um documento. Essa simplificação, apesar de grosseira, facilita a modelagem matemática do problema e fornece relativamente bons resultados.

O problema em questão se trata em determinar a qual classe c_i, dentre um conjunto de classes $\mathcal{C}$, um certo documento $\mathbf{d}$ pertence. Este evento, convenientemente representado por $c_i|\mathbf{d}$, pode ser estimado através de sua probabilidade.

A. Modelo

Sendo assim, supomos que nosso documento $\mathbf{d}$ possua uma distribuição multinomial de variáveis aleatórias d_i, $i = 1 \ldots n$, doravante denominadas características. Pela regra de Bayes, podemos escrever:

$$P(c_i|\mathbf{d}) = P(c_i|d_1, d_2, \ldots, d_n) = \frac{P(d_1, d_2, \ldots, d_n|c_i)P(c_i)}{P(\mathbf{d})}. \tag{1}$$

Pode-se notar que o numerador do lado direito da equação é equivalente à probabilidade conjunta $P(c_i, d_1, d_2, \ldots, d_n)$. Supondo que as características d_i, d_j são independentes entre si[1] ($i \neq j$) e da classe c_i, a probabilidade conjunta pode ser escrita como

$$P(c_i, d_1, d_2, \ldots, d_n) = P(c_i) \prod_{j=1}^{n} P(d_j|c_i). \tag{2}$$

Substituindo a expressão (2) em (1), obtemos:

$$P(c_i|\mathbf{d}) = \frac{P(c_i) \prod_{j=1}^{n} P(d_j|c_i)}{P(\mathbf{d})}. \tag{3}$$

Finalmente, o processo decisório se dá pela avaliação da expressão

$$c = \underset{c_i \in \mathcal{C}}{\operatorname{argmax}}\ P(c_i|\mathbf{d}) = \underset{c_i \in \mathcal{C}}{\operatorname{argmax}} \frac{P(c_i) \prod_{j=1}^{n} P(d_j|c_i)}{P(\mathbf{d})}. \tag{4}$$

em que c é a classe mais provável para o documento $\mathbf{d}$.

A tarefa de classificação se dá pela busca da classe c_i que maximiza o operando em (4). Como o termo $P(\mathbf{d})$ é constante[2], podemos omiti-lo sem ônus para a classificação. Além disso, como a probabilidade é um número positivo e estamos interessados apenas nas distâncias absolutas, podemos aplicar a função logarítmo em ambos os membros de (3). Feitas essas alterações, obtemos:

$$c = \underset{c_i \in \mathcal{C}}{\operatorname{argmax}} \left(\log P(c_i) + \sum_j \log P(d_j|c_i) \right). \tag{5}$$

Note que a tarefa decisória deste classificador é função da probabilidade *a priori* por classe $P(c_i)$ e da probabilidade de ocorrência da característica d_j condicionada à classe c_i, $P(d_j|c_i)$. A estimação dessas probabilidades, dado um conjunto de classes e documentos, é um processo conhecido como treinamento, apresentado na seção seguinte.

B. Treinamento

O processo de treinamento busca obter o modelo do classificador dada uma observação de um conjunto de documentos, para os quais já se conhece a classe. O modelo para o Bayes Ingênuo Multinomial consiste das probabilidades $P(c_i)$ e $P(d_j|c_i)$.

Dentre os vários métodos de estimação, utilizamos o Estimador de Máxima Verossimilhança, que infere um conjunto de suposições sobre os dados para obter fórmulas em função da frequência relativa dos parâmetros. Para este estimador, as probabilidades podem ser escritas como:

$$P(c_i) = \frac{N_{c_i}}{N}, \tag{6}$$

$$P(d_j|c_i) = \frac{N_{d_j,c_i} + 1}{N_{dci} + |\mathbf{d}|}, \tag{7}$$

em que N é o número total de documentos, N_{c_i} é o número de documentos da classe c_i, N_{d_j,c_i} o número de ocorrências da característica d_j na classe c_i, N_{dci} o total de ocorrências das características na classe c_i e $|\mathbf{d}|$ o tamanho do vetor de características.

III. Algoritmo de Aprendizado Bayes Ingênuo Multinomial Modificado (NB+)

Neste trabalho é introduzida uma modificação no algoritmo Bayes Ingênuo Multinomial. A partir de observações nas frequências relativas calculadas nas equações (6) e (7), foi possível notar uma maior relevância (para a tarefa de classificação) das características d_i do que as probabilidades *a*

[1] Essa suposição de independência costuma ser taxada como ingênua, caracterizando o popular nome *Naive Bayes* (em português numa tradução livre **Bayes Ingênuo**) para este classificador.

[2] Pode-se verificar que o termo $P(\mathbf{d})$ da equação (4), também conhecido como evidência, é constante fazendo $P(\mathbf{d}) = \sum_{c \in \mathcal{C}} P(\mathbf{d}|c)P(c)$

priori das classes. Entretanto, fora observada que a presença da probabilidade *a priori* é de fato benéfica para o classificador [1].

Sendo assim, este trabalho buscou introduzir uma variável capaz de controlar, a grosso modo, quanta informação é aproveitada de cada uma das partes (*a priori* e *a posteriori*). Isto foi feito utilizando uma lei de potência inspirada na Lei de Zipf [6], que será melhor explicada na seção seguinte.

A. Modelo

A Lei de Zipf é uma lei empírica que descreve um comportamento estatístico comum à maioria das linguagens naturais. Ela dita que a frequência de ocorrência de palavras de um texto de uma certa língua costuma ser similar aos demais textos da mesma lingua.

Nesse contexto, vale observar que, para propósitos de classificação, palavras de alta ocorrência contribuem com pouca informação sobre os limiares entre as classes (pois ocorrem em todas as classes), visto que os textos obedecem aproximadamente à Lei de Zipf. Sendo assim, tais palavras geralmente são removidas e não são consideradas como características.

Por outro lado, pode-se supor que palavras mais raras, isto é, de baixa ocorrência, tem grande potencial de informar sobre os limiares. Por exemplo, em um *corpus* com documentos de diversas áreas da ciência, a presença de uma palavra rara como "tireóide", exclusiva à área médica, fornece forte evidência de que o texto de fato pertence a esta categoria.

Com isso, neste trabalho buscamos introduzir uma variável de controle para melhor aproveitar as características mais raras. Além de remover as palavras de alta ocorrência, efetuamos uma modificação na probabilidade conjunta com intuito de melhor distribuir a contribuição das características, levando em consideração o potencial informativo das palavras mais raras.

Para isso, supôs-se que a probabilidade condicionada $P(c_i|\mathbf{d})$ é relacionada com suas características através de uma lei de potência da forma $f(x) \propto ax^k$, isto é,

$$P(c_i|\mathbf{d}) \propto P^k(d_j|c_i). \tag{8}$$

Modificações similares foram realizadas em [7]. Como a probabilidade P é um número real $0 \leq P \leq 1$, a introdução da potência efetivamente reduz o valor da probabilidade da característica.

Essa modificação, na verdade, reduz a distância entre características comuns e raras. Por exemplo, a distância entre $P(c_1, d_1) = 0.5$ e $P(c_1, d_2) = 0.1$ é originalmente $0.5 - 0.1 = 0.4$. Para a condicional modificada e $k = 2$, essa distância é reduzida para $0.5^2 - 0.1^2 = 0.24$.

Sob as mesmas suposições de independência do algoritmo original (Seção II-A), podemos reescrever a probabilidade conjunta como

$$P(c_i, d_1, d_2, \ldots, d_n) = P(c_i) \prod_{j=1}^{n} P^k(d_j|c_i). \tag{9}$$

Substituindo a expressão (9) em (1), obtemos:

$$P(c_i|\mathbf{d}) = \frac{P(c_i) \prod_{j=1}^{n} P^k(d_j|c_i)}{P(\mathbf{d})}. \tag{10}$$

Finalmente, o processo decisório modificado se dá pela avaliação da expressão

$$c = \underset{c_i \in \mathcal{C}}{\operatorname{argmax}}\ P(c_i|\mathbf{d}) = \underset{c_i \in \mathcal{C}}{\operatorname{argmax}} \frac{P(c_i) \prod_{j=1}^{n} P^k(d_j|c_i)}{P(\mathbf{d})}. \tag{11}$$

Efetuando simplificações similares as da seção II-A, a equação (11) pode ser reduzida a

$$c = \underset{c_i \in \mathcal{C}}{\operatorname{argmax}} \left(\log P(c_i) + k \sum_j \log P(d_j|c_i) \right). \tag{12}$$

Feitas essa modificações, um procedimento experimental verificou os índices de qualidade do classificador, comparado ao algoritmo original. Tais procedimentos são descritos na seção seguinte.

IV. Procedimento Experimental

Neste trabalho foi realizado um procedimento experimental com objetivo de comparar o desempenho dos classificadores Bayes Ingênuo Multinomial (NB) e de sua versão modificada (NB+). Essa análise foi implementada utilizando a linguagem de programação Python, a biblioteca de aprendizado de máquina *scikit-learn* [8] e a biblioteca de cálculo numérico *numpy*.

A preparação do experimento inicia-se pelo tratamento do *corpus*, etapa na qual são escolhidas as categorias e os documentos utilizados no treinamento e nos testes. Em seguida, esse *corpus* passa por um processo denominado vetorização em que o texto dos documentos é transformado em um vetor de números, forma esta apropriada para manipulação numérica. O resultado da vetorização é um *corpus* de vetores. Por fim, é reduzida a dimensão desses vetores utilizando um processo denominado redução de dimensionalidade. Finalmente, tem-se o processo de treinamento e de avaliação.

Nas seções a seguir descrevemos cada etapa do experimento em mais detalhes.

A. Preparação do Corpus

Neste experimento foi utilizado o *corpus 20Newsgroups-18828*, que consiste numa coleção de 18828 mensagens de grupos de notícias da rede UseNet. Essa coleção é derivada da coleção original *20Newsgroups-19997* - coletada por Ken Lang - em que foram removidas mensagens duplicadas e campos de meta-dados das mensagens. Essa coleção encontra-se disponível tanto no portal MLcomp.org quanto no arcabouço *scikit-learn* [8].

[1]Foi realizado um experimento comparando o desempenho do Bayes Ingênuo Multinomial para dois casos: banco de dados com e sem equidistribuição de classe. Vale notar que, ao induzir uma equidistribuição de classes, a probabilidade $P(c_i)$ é a mesma para qualquer $c_i \in \mathcal{C}$. Os resultados demonstraram que a probabilidade *a priori* é benéfica para o desempenho.

Dentre as vinte categorias disponíveis no *corpus*, foram escolhidas apenas cinco para os testes: `comp.graphics`, `talk.religion.misc`, `talk.politics.misc`, `alt.atheism`, `sci.space`. Dentre estas categorias, 50% dos documentos foram utilizados na etapa de treinamento e os restantes para os testes.

B. Vetorização e Seleção de Características

Uma vez preparado o *corpus*, os documentos de texto foram transformados em vetores de características utilizando a estatística TF-IDF (do inglês *term frequency - inverse document frequency*) de vetorização como definido em [3]. Esse método tem como objetivo pontuar as palavras de um documento de acordo com sua importância para o documento, considerando o contexto do *corpus*. Essa pontuação cresce com a frequência da palavra no documento e recua com a frequência da mesma no *corpus*, compondo uma métrica geral que também considera o *corpus*.

O produto dessa vetorização é um conjunto de vetores cuja dimensão é proporcional ao número de palavras distintas no documento. Note que, caso o *corpus* seja volumoso, essa dimensão pode assumir valores proibitivos à classificação. Na remediação dessa dificuldade, é comum se aplicar alguma técnica de redução de dimensionalidade, conhecida também como seleção de características. Neste trabalho, selecionamos as melhores características seguindo o critério das *k-melhores*, como definido a seguir.

Definição 1: A seleção de características sob o critério K-melhores seleciona K características que melhor pontuem no teste-F [9].

O procedimento de vetorização e seleção foi aplicado para o *corpus* como um todo, incluindo os subconjuntos de treinamento e de teste.

C. Avaliação de Desempenho

Os classificadores em questão foram avaliados utilizando as métricas F_1, acurácia, cobertura (*recall*) e precisão, como definidas em [3]. Para uma completa visualização do comportamento dos classificadores, a avaliação foi repetida para quantidades crescentes de características, obtendo os valores das métricas para cada número.

No controle do número de características, foi regulado o valor de K (e.g. $K = 5, 10, \ldots, 1000$) do critério *K-melhores*, efetivamente escolhendo apenas as melhores características.

V. Resultados

O procedimento experimental descrito na seção IV foi inicialmente realizado utilizando um valor $k = 2$ na equação (12). Como descrito anteriormente, a avaliação de ambos classificadores foi repetida para um crescente número de características, filtradas pelo processo de seleção *K-melhores*. Os resultados são apresentados na Figura 1 a seguir.

Como podemos observar, o classificador Bayes Ingênuo Modificado, denotado no gráfico por *NB+*, supera o original em todas as métricas observadas no intervalo $|c| < 200$ características. Para um número de características superior a

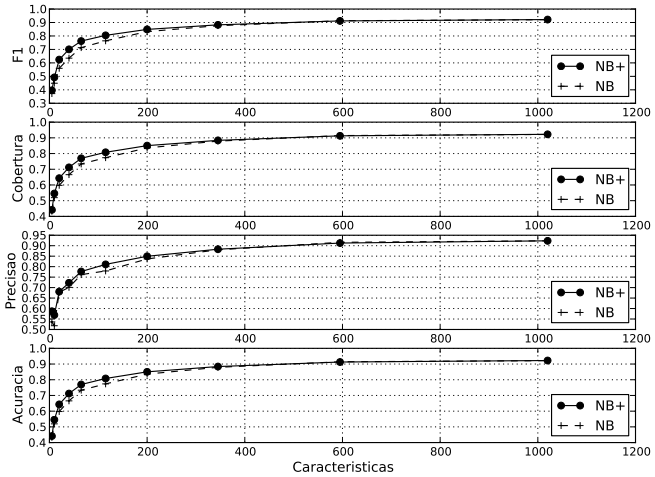

Fig. 1. Métricas F1, Cobertura, Precisão e Acurácia de classificadores *NB* e *NB+* e $k = 2$

200, as curvas convergem para um ponto em comum (92% de acurácia, precisão, cobertura e F_1).

A fim de analisar mais a fundo o comportamento da modificação, o expoente k foi variado para valores ao redor de $k = 2$. Para o valor $k = 3$ (Figura 2), foi obtido um resultado similar, porém com métricas levemente melhores.

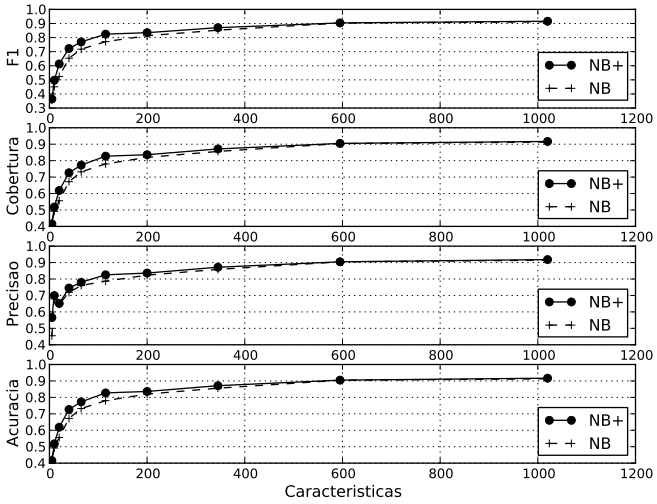

Fig. 2. Métricas F1, Cobertura, Precisão e Acurácia de classificadores *NB* e *NB+* e $k = 3$

Para valores além de $k = 3$ e abaixo de $k = 1$, o classificador apresentou degradação no desempenho, apresentado assim indícios de um possível ponto ótimo para k.

VI. Conclusões e Perspectivas

Neste trabalho foi introduzida uma versão modificada do classificador Bayes Ingênuo. Essa versão considerou uma mudança na distribuição de probabilidades das características

do texto, buscando melhor considerar a informação contribuída por palavras raras. Para tal, modificou-se a suposição da probabilidade condicional a fim de melhor distribuir a probabilidade dentre as características.

Um procedimento experimental foi realizado com objetivo de comparar o classificador resultante, nomeado Bayes Ingênuo Multinomial Modificado (NB+), com o classificador original. Para esse experimento foi utilizado o corpus 20Newsgroups-18828, o qual foi submetido a uma vetorização com pesos TF-IDF e uma seleção de características sob o critério *K-melhores*.

Controlando o número de características através do critério *K-melhores*, os classificadores foram treinados e avaliados utilizando as métricas F_1, precisão, cobertura e acurácia.

Os resultados apresentados demonstraram uma superioridade do algoritmo modificado quando comparado ao original em todas as métricas utilizadas. Mais especificamente, para os parâmetros $k = 2$ e $k = 3$, o algoritmo apresentou entre 10% e 20% de melhoria nas métricas para o intervalo $50 \leq |d| \leq 200$ características. Essas melhorias sugerem que, apesar de sua simplicidade, ainda é possível alcançar níveis maiores de desempenho com o classificador Bayes Ingênuo Multinomial, melhorando ainda mais a relação de custo e benefício já amplamente conhecida pela comunidade.

Analisando os resultados obtidos, é de se esperar que exista um k ótimo para cada problema de classificação textual, isto é, que para cada corpus o classificador NB+ supere o NB. Essa conjectura sugere possíveis trabalhos futuros na busca de um algoritmo de busca para o k.

Por fim, fica sugerida a utilização do algoritmo NB+ em conjunto com a técnica da Maximização da Expectativa (EM). Trabalhos realizados em [10] com o algoritmo original apresentaram uma melhoria de performance ao se utilizar amostras não marcadas, isto é, adentrando à classe de aprendizado semi-supervisionado. Espera-se que, aliando o NB+ com uma técnica de maximização de verossimilhança, seja possível alcançar novos níveis de desempenho.

REFERÊNCIAS

[1] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, ser. CIKM '11. New York, NY, USA: ACM, 2011, pp. 1031–1040. [Online]. Available: http://doi.acm.org/10.1145/2063576.2063726

[2] D. L. Olson and B. Chae, "Direct marketing decision support through predictive customer response modeling," *Decis. Support Syst.*, vol. 54, no. 1, pp. 443–451, Dec. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2012.06.005

[3] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, Mar. 2002. [Online]. Available: http://doi.acm.org/10.1145/505282.505283

[4] Q. Wu, E. Fuller, and C.-Q. Zhang, "Text document classification and pattern recognition," in *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 405–410. [Online]. Available: http://dl.acm.org/citation.cfm?id=1602240.1602729

[5] Y. H. Li and A. K. Jain, "Classification of text documents," *The Computer Journal*, vol. 41, pp. 537–546, 1998.

[6] M. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[7] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the poor assumptions of naive bayes text classifiers," in *In Proceedings of the Twentieth International Conference on Machine Learning*, 2003, pp. 616–623.

[8] scikit learn. (2013, Jan.) scikit-learn: machine learning in python. [Online]. Available: http://scikit-learn.org/

[9] R. Myers, *Classical and modern regression with applications*. Duxbury Press Belmont, CA, 1990, vol. 2.

[10] K. P. Nigam, "Using unlabeled data to improve text classification," Ph.D. dissertation, Pittsburgh, PA, USA, 2001, aAI3040487.