SELEÇÃO DE CARACTERÍSTICAS APLICADA À MODERAÇÃO AUTOMÁTICA DE COMENTÁRIOS DE USUÁRIOS

FEATURE SELECTION APPLIED TO AUTOMATIC MODERATION OF COMMENTS FROM USERS

SAUDE, Marcos Rodrigues

Coord. dos Cursos de Sistemas de Informação e de Engenharia Elétrica da Faculdade Pitágoras - Câmpus Linhares (ES)

mrsaude@oi.com.br

Resumo

Com a expansão das mídias sociais e o advento da Web 2.0, tem havido uma participação cada vez maior de pessoas interessadas em expor suas opiniões e dispostas a discutir em um ambiente coletivo sobre algum assunto divulgado na Internet. No entanto, muitos dos comentários realizados pelos usuários podem ser de caráter ofensivo ou pejorativo, o que pode causar demandas judiciais aos provedores desses ambientes. Neste trabalho é explorada uma abordagem automática para classificação desses comentários, identificando aqueles que poderiam ser divulgados sem causar inconvenientes para os provedores. Para alcançar este objetivo, foram usadas técnicas para tratamento dos dados, tais como redução de dimensionalidade e ponderação de termos (palavras), de forma a obter um modelo computacional capaz de auxiliar as decisões humanas para moderação de comentários. Uma das combinações de técnicas de seleção de características utilizada no trabalho foi capaz de imitar as decisões dos especialistas humanos em 85,56% dos comentários testados.a.

Palavras-chave: classificação de comentários; seleção de características; recuperação de informação.

Abstract

With the expansion of social media and the advent of Web 2.0, there has been an increasing participation of people interested in making their views and willing to discuss in a collective environment about something posted on the Internet. However, many of the comments made by users can be offensive or derogatory character, which can cause lawsuits providers of these environments. In this paper is explored an automatic classification approach to these comments, identifying those that could be disclosed without causing inconvenience to providers. To achieve this goal, techniques were used for data processing, such as dimensionality reduction and weighting of terms (words), to obtain a computational model capable of assisting human decisions for comment moderation. One of the combinations of feature selection techniques used in this work was able to mimic the decisions of human experts in 85.56% of the comments tested.

Key-words: comments classification, feature selection, information retrieval.

INTRODUÇÃO

O advento da Web 2.0 permitiu o aumento da interação do usuário com os sites da Internet, possibilitando-o expressar as suas opiniões e ideias sobre um determinado assunto. No entanto, muitos desses comentários podem ser ofensivos, e os sites tornam-se responsáveis pela sua divulgação, podendo até mesmo resultar em demandas judiciais aos mesmos.

Torna-se, portanto, necessário que tais comentários sejam submetidos a um processo de seleção para posterior divulgação, o qual é extremamente custoso ao ser humano, dado o grande número de notícias constantemente divulgadas e a massiva participação dos usuários em contribuir com suas ideias.

O objetivo deste trabalho é utilizar técnicas de recuperação de informação (RI) para realizar a moderação de comentários, indicando quais comentários devem ou não ser publicados no meio digital. Sendo assim, foram exploradas técnicas de redução de dimensionalidade para melhor caracterização dos comentários contidos em cada categoria, obtendo-se melhorias consideráveis nos índices de classificação em relação ao uso da base original.

Este artigo está organizado da seguinte forma: na seção 2 são mencionados outros trabalhos relacionados ao desenvolvido neste artigo. O modelo adotado para representação de documentos é apresentado na seção 3. As técnicas utilizadas para tratamento dos comentários que compõem a base são descritas na seção 4. Na seção 5 são apresentados os métodos utilizados para seleção de características. Os algoritmos de classificação adotados neste trabalho são explicados na seção 6.

A comparação e discussão dos resultados obtidos nos experimentos encontram-se na seção 7, e; por fim, as conclusões e pesquisas futuras são descritas na seção 8.

TRABALHOS RELACIONADOS

Trabalhos similares ao desenvolvido neste artigo tratam de filtros anti-spam para classificação de conjuntos de mensagens que devam ou não ser consideradas como de conteúdo malicioso. O trabalho desenvolvido em (Shrivastava e Bindu, 2014) utilizou algoritmo genético para extração de termos que melhor identificam o conteúdo spam, obtendo-se alta eficiência nos resultados alcançados, com acurácia acima de 82%.

Neste mesmo contexto, em (Pourhashemi e Osareh, 2013) foram utilizados métodos de seleção de características em duas etapas. Na primeira etapa foram extraídos stopwords e termos pouco referenciados. Na segunda etapa de filtragem foi aplicado o método chi-square para extração de termos mais relevantes no corpus considerado.

Comparações foram feitas com uso dos classificadores DMNB (Discriminative Multinomial Nayve Bayes), MNB (Multinomial Nayve Bayes), SVM (Support Vector Model) e

Random Forest, concluindo que a combinação de seleção de características e o uso de um classificador apropriado aumentam os índices da classificação, além de melhorar seu desempenho.

Outra área com intuito similar ao apresentado neste artigo refere-se à análise de sentimentos, que objetiva classificar as opiniões das pessoas, sobre determinado tema, em positiva ou negativa (alguns casos, em neutra também).

O trabalho desenvolvido em (Duarte, 2013) compara resultados de medidas de acurácia média, recall e precision para a classificação de sentimentos, de um total de 300 mil comentários em língua portuguesa, extraídos da rede social Twitter, que continham o verbo sentir e suas diferentes conjugações (presente, passado e futuro do modo indicativo).

Duas classes foram consideradas na classificação: positiva e negativa. Para extração de características foram utilizados o SentiLex, uma ferramenta para identificação de polaridade positiva ou negativa de textos em português, e a estratégia de negação Bigrams encontrado em (Pak e Paroubek, 2010). Os algoritmos usados para classificação foram Naive Bayes, Decision Tree, SVM e kNN.

Os resultados mostraram que, para distinguir comentários positivos e negativos, deve-se utilizar o SentiLex ou uma combinação de SentiLex e negação Bigrams com uso do classificador SVM, com acurácia média em torno de 68,05%.

MODELO VETORIAL

Para que um computador possa operar sobre os comentários, eles são representados utilizando o modelo vetorial (Salton et. al., 1975). Neste modelo, cada comentário dj da base de dados é representado em forma de um vetor no espaço, sendo que cada dimensão do espaço representa um termo (palavra) ki referenciado no comentário e, para cada termo, é associado um peso wi,j, obtido, a princípio, pelo número de vezes que o termo ocorre no comentário (Term Frequency – tf).

Sendo n o total de termos referenciados no documento, a representação do comentário está demonstrada na Equação 1:

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{i,j}, \ldots, w_{n,j}) \quad (01)$$

TRATAMENTO DOS TEXTOS

Os textos que compõem a base de dados devem ser submetidos a um pré-processamento para adequação do conjunto de termos que, se não tratados, podem causar ruídos no resultado final da classificação automática. Serão descritas a seguir as técnicas utilizadas neste trabalho para o tratamento dos textos.

EXTRAÇÃO DE *STOPWORDS*

Stopwords são termos considerados não relevantes na indexação dos documentos por possuírem pouco valor semântico. Na maioria das vezes são utilizados na função de conectivos ou termos auxiliares, como pronomes, artigos, preposições, advérbios ou conjunções.

LEMATIZAÇÃO (STEMMING)

Em Processamento de Linguagem Natural (PLN), o estudo da "normalização de variações linguísticas" busca reduzir os termos a elementos de escrita mais simples. Uma técnica conhecida como stemming consiste num processo de redução de formas variantes de uma palavra a uma representação comum (o radical da palavra, ou stem).

Em (Orengo e Huyck, 2001) foi desenvolvido um algoritmo para extração de radicais de palavras da língua portuguesa, denominado RSLP (Redutor de Sufixos da Língua Portuguesa). O algoritmo considera a extração de radicais de palavras através de 8 (oito) passos, promovendo a retirada da forma plural, feminina, adverbial, aumentativo ou diminutivo, terminações verbais, vogais e remoção de acentos.

Uma grande vantagem desta ferramenta é a utilização de um dicionário externo e editável, contendo cerca de 32 mil palavras, com regras para a correta redução, possibilitando remanejar seu conteúdo ou mesmo aperfeiçoar a extração através das regras de exceção contidas em sua configuração.

PONDERAÇÃO DE TERMOS (TF-IDF)

Os pesos dos termos contidos em cada documento da base podem ser adequados com uso de fatores de ponderação, fazendo com que termos que ocorrem em grande quantidade de documentos tenham seu peso diminuído em virtude de sua menor importância na classificação. Uma medida de ponderação mais comumente utilizada é chamada idf (Inverse Document Frequency) indicada por (Robertson, 2004), que é calculada de acordo com o demonstrado na equação 2.

$$\mathrm{idf}_i = \log \frac{|N|}{|n_i|} \qquad (02)$$

onde |N| é o total de documentos do corpus e |ni| é o total de documentos que contém o termo ki. O novo peso é calculado multiplicando-se este fator aos pesos de cada termo (tf*idf).

ALGORITMOS PARA SELEÇÃO DE CARACTERÍSTICAS

Segundo (Baeza-Yates e Ribeiro-Neto, 2011), um grande espaço de características (ou termos) pode tornar impraticável a classificação de documentos, visto que a classificação de novos documentos consumiria muito tempo e prejudicaria a eficiência dos classificadores. A solução clássica para este problema é reduzir o tamanho do espaço de características, selecionando um subconjunto de todos os termos para a representação dos documentos. Este passo é chamado de seleção de características.

SFS (SEQUENTIAL FORWARD SELECTION)

O algoritmo Sequential Forward Selection (SFS) inicia seu funcionamento a partir do conjunto vazio de termos, e vai adicionando sequencialmente

à base de treino o termo x+ cuja presença implica uma elevação na função objetivo (medida de avaliação) J(Yk + x+), quando combinada com os termos Yk já anteriormente selecionados (Ladha e Deepa, 2011). Os passos do algoritmo são enumerados abaixo:

1. inicia com o conjuto vazio $Y_0 = \{ \}$
2. seleciona o próximo termo

$x^+ = \text{argmax}[J(Y_k + x)]; x \square Y_k$

3. atualiza $Y_{k+1} = Y_k + x^+; k = k + 1$
4. volta para o passo 2

ALGORITMOS GENÉTICOS

Um algoritmo genético pode ser definido como um processo repetitivo que mantém uma população de "indivíduos" representando as possíveis soluções para um determinado problema (Mitchell, 2002). A cada "geração", os indivíduos da população passam por uma avaliação de sua capacidade em oferecer uma solução satisfatória para o problema. Essa avaliação é feita por uma função de adaptação, também chamada função de fitness.

Apesar do algoritmo genético não contemplar todas as possíveis combinações para se atingir um resultado ótimo, o que consumiria muito tempo de processamento, seu uso justifica-se pela possibilidade de encontrar soluções aceitáveis em casos de problemas de elevado grau de complexidade matemática ou com grande número de soluções possíveis.

Para a implementação do algoritmo genético neste trabalho foi utilizada a GAlib (Genetic Algorithm Library), uma biblioteca de funções para a linguagem C++, disponível em ftp://lancet.mit.edu/pub/ga.

Na implementação do algoritmo genético, devem ser especificados os operadores genéticos: tamanho da população, taxa de mutação e número de gerações. Estudos realizados em (Catarina e Bach, 2003) concluíram que o tamanho da população não deve ser muito grande, pois faz o algoritmo trabalhar por um período de tempo maior, por outro lado, não deve ser muito pequeno para não causar diminuição do espaço de busca da solução pretendida.

Para aplicação do algoritmo genético de codificação binária (usada neste trabalho), a taxa de mutação deve ser baixa, entre 1% e 5%, evitando que a busca se torne essencialmente aleatória. O número de gerações não deve ser muito alto, acima de 200, o que provocaria substituição da maior parte da população, tornando a execução do algoritmo muito lenta.

COMBINAÇÃO DE RETIRADA DE TERMOS MAIS RAROS E DE TERMOS MAIS COMUNS

Uma técnica utilizada para seleção de características busca identificar as possíveis combinações de retirada de termos que poderiam aperfeiçoar os resultados da classificação.

Dessa forma, a experimentação combinatória verifica as diversas combinações de remoção de

termos presentes nas extremidades do ranking de pesos dos termos da base. Para cada combinação foram removidas uma determinada porcentagem dos termos mais raros (com baixa frequência na base de dados) e uma porcentagem dos termos mais comuns (mais frequentes) contidos na base de dados.

A combinação com melhor resultado de classificação é selecionada. Os percentuais aplicados em cada combinação não podem totalizar ou ultrapassar o valor de 100%, equivalente a remover todos os termos, o que descaracterizaria a base de dados.

TÉCNICAS DE CLASSIFICAÇÃO

Em (Chen et. al, 1996) é afirmado que o processo de classificação envolve a construção de um modelo para que sejam aplicados os dados ainda não classificados, visando categorizá-los numa classe pré-definida, baseando-se nas características comuns entre o conjunto de dados da base de treinamento, produzindo uma identificação para cada dado submetido na classificação.

Os algoritmos de aprendizagem possuem diferentes características. Nesta seção serão descritos os algoritmos de classificação selecionados para compor os experimentos realizados neste artigo.

SVM

O classificador SVM (Support Vector Machine) é uma técnica de aprendizado de máquina, comumente utilizada em problemas de categorização de documentos (Joachims, 1999). De maneira geral, o método constitui uma abordagem geométrica para o problema da classificação. Tomando-se como exemplo um corpus com duas classes Ca e Cb, a técnica busca encontrar uma superfície de decisão (hiperplano) que pode ser usada como separador dos elementos das classes.

O hiperplano é obtido na fase de aprendizagem, através dos dados da base de treino, e divide o espaço em duas regiões, de tal forma que os documentos da classe Ca estejam em uma região e os documentos da classe Cb estejam na outra região. Após a obtenção do hiperplano, um novo documento dj pode ser classificado pela sua posição relativa ao hiperplano (Baeza-Yates;Ribeiro-Neto, 2011).

CBC

O classificador CBC (Centroid-Based Classifier) (Han; Karypis, 2000) é baseado na ideia de interpretar a base de treino de cada classe como um conjunto de informações. Para cada categoria de documentos de treino Cp, contendo m documentos, é calculado um centroide cp através da média dos pesos de cada termo ki pertencente aos documentos da categoria Cp, conforme demonstrado na Equação 3.

$$c_p = \frac{1}{\sum_{j \in C_p} w_{i,j}} \qquad (03)$$

Após os cálculos dos centroides de cada classe, cada documento da base de testes é

classificado de acordo com a maior proximidade ao centroide de uma determinada categoria, sendo atribuída sua classificação a esta categoria.

kNN (k Nearest Neighbors)

O classificador kNN, do inglês k-Nearest Neighbors("k vizinhos mais próximos"), é um método baseado na analogia. O conjunto de treino é formado por vetores n-dimensionais e cada elemento deste conjunto representa um ponto no espaço n-dimensional. Sendo assim, dado um documento de teste dt, o método kNN realiza as seguintes atividades para classificá-lo (SHAKHNAROVICH; INDYK; DARRELL, 2006):

- a distância entre o documento dt e cada um dos documentos de treino é calculada utilizando alguma medida de similaridade entre documentos;
- os k documentos de treino mais próximos, ou seja, mais similares do documento dt são selecionados;
- o documento dt é classificado em determinada categoria de acordo com algum critério de agrupamento das categorias dos k documentos de treino selecionados na etapa anterior.

Em geral, são observadas quais são as classes desses k vizinhos mais próximos, e o documento dt será classificado como pertencente à classe mais frequente.

A escolha deste classificador justifica-se por ser um método amplamente utilizado em experimentos que envolvem recuperação de informação. Apesar de sua simplicidade, seus resultados alcançam bom desempenho em diferentes cenários (Yang e Liu, 1999).

EXPERIMENTOS E DISCUSSÃO DOS RESULTADOS

MATERIAL, MÉTRICAS E FERRAMENTAS

Os experimentos foram realizados sobre uma base de dados com 978 documentos extraídos dos comentários de notícias do site http://g1.globo.com/ (Portal de Notícias Globo), cedidos pela empresa Globo Comunicação e Participações S.A.. Estes documentos foram pré-classificados nas categorias de Aprovados (539 comentários) ou Reprovados (439 comentários) para divulgação, através de um trabalho manual realizado por especialistas da referida empresa.

A base de dados foi submetida a um pré-processamento, com retirada das stopwords, além da extração de radicais das palavras e uso do fator de ponderação tf-idf. Como resultado do pré-processamento, apenas 4434 termos foram mantidos na base.

As medidas de caracterização da base de dados estão demonstradas na Tabela 1, de acordo com as medidas utilizadas em (Salton et. al., 1975), sendo MSDC a média de similaridade entre documentos e seus centroides, MSCCP a média de similaridade entre os centroides de cada classe e o

Tabela 1 – Caracterização da base de dados

MSDC	MSCCP	MSPC	Razão (MSDC/MSPC)
0,0805	0,9138	0,6701	6,2255

centroide principal, MSPC a média de similaridade entre pares de centroides e Razão a razão entre MSDC e MSPC.

As medidas de similaridade entre documentos foram calculadas utilizando o cálculo do cosseno do ângulo entre pares de documentos di e dj (vetores), conforme ilustrado na Equação 4.

$$\cos(d_i, d_j) = \frac{d_i \cdot d_j}{\|d_i\| \|d_j\|} = \frac{\sum_{k=1}^{n} w_{k,i} w_{k,j}}{\sqrt{\sum_{k=1}^{n} w_{k,i}^2} \sqrt{\sum_{k=1}^{n} w_{k,j}^2}} \quad (04)$$

Esta medida indica maior similaridade entre documentos quanto mais próximo de 1 (um) for seu valor. Valores próximos de zero indicam menor similaridade.

As informações contidas na Tabela 1 demonstram que os comentários de uma mesma categoria encontram-se espacialmente bem separados em virtude do baixo valor médio de similaridade entre os documentos das classes com seus centroides (medida MSDC).

Como o valor de MSCCP é alto, próximo do valor máximo 1, podemos concluir que os centroides das classes estão muito próximas do centroide principal. O valor obtido para MSPC indica que as classes encontram-se relativamente próximas, ocasionando um alto índice da Razão.

Dessa forma, no espaço vetorial, os documentos de cada categoria estão bem espalhados, mas existem documentos de categorias diferentes que estão próximos, ocorrendo um alto índice de sobreposição, o que dificulta a classificação de novos comentários.

As métricas Recall, Precision e F1-measure foram adotadas para medir a qualidade da classificação:

$$Recall(C_p) = \frac{TP(C_p)}{TP(C_p) + FN(C_p)} \times 100\% \quad (05)$$

$$Precision(C_p) = \frac{TP(C_p)}{TP(C_p) + FP(C_p)} \times 100\% \quad (06)$$

$$F1-measure(C_p) = \frac{2Recall(C_p)Precision(C_p)}{Recall(C_p) + Precision(C_p)} \times 100\% \quad (07)$$

nas quais TP é a quantidade de documentos atribuídos corretamente à classe Cp pelo classificador automático, FP é a quantidade de documentos atribuídos incorretamente à classe Cp pelo classificador automático e FN é a quantidade de documentos pertencentes à classe Cp classificada incorretamente pelo classificador automático.

ANÁLISE DOS RESULTADOS

Para a classificação dos documentos foi utilizada a técnica do 10-fold cross-validation, onde a base é dividida em 10 (dez) partes e são obtidos 10 resultados diferentes, cada uma sobre uma partição diferente, e as demais partes são usadas para treinamento. Ao final é calculada a média dos índices obtidos, de acordo com o descrito por Kohavi (1995).

Os experimentos foram realizados usando os classificadores SVM (Support Vector Machine), CBC (Centroid-based Classification) e kNN (k-Nearest Neighbors).

Foram realizados experimentos em etapas, cujos resultados obtidos para Recall, Precision e F1-measure estão ilustrados nas Tabelas 2, 3 e 4.

Os valores de desvio padrão estão demonstrados nas colunas DPR (desvio padrão da medida Recall), DPP (desvio padrão da medida Precision) e DPF1 (desvio padrão da medida F1-measure).

A primeira etapa de classificação (etapa 1) foi processada com a base no estado inicial de pré-processamento, com um total de 4434 termos.

A seguir, na etapa 2, aplicou-se a retirada de termos presentes em menos de 2 (dois) documentos, reduzindo o número de termos para 1726 termos.

Apesar de não trazer melhorias significativas nas taxas de classificação dos três classificadores, a retirada de termos pouco referenciados reduziu o tempo de processamento da classificação em razão da expressiva redução de termos de 4434 para 1726.

Em busca de tentar melhorar a classificação, foram usadas três estratégias para seleção de características.

A estratégia de SFS (etapa 3.1), buscando maximizar a métrica F1-measure, foi usada até

Tabela 2 – Índices de classificação com uso do classificador SVM

Etapa	*Recall* (%)	*DPR*	*Precision* (%)	*DPP*	*F1-measure* (%)	*DPF1*
1	60,75	0,16	62,11	0,18	60,00	0,17
2	62,58	0,14	63,27	0,15	62,21	0,15
3.1	76,06	0,13	77,64	0,14	76,84	0,14
3.2	71,31	0,14	75,44	0,13	73,32	0,13
4.1	76,06	0,13	77,64	0,14	76,84	0,14
4.2	**79,86**	**0,12**	**81,64**	**0,13**	**80,74**	**0,12**

Tabela 3 – Índices de classificação com uso do classificador CBC

Etapa	*Recall* (%)	*DPR*	*Precision* (%)	*DPP*	*F1-measure* (%)	*DPF1*
1	62,35	0,16	63,33	0,17	62,84	0,17
2	62,50	0,17	63,52	0,18	63,01	0,18
3.1	82,71	0,13	83,36	0,13	83,03	0,13
3.2	81,84	0,14	82,63	0,13	82,24	0,14
4.1	82,71	0,13	83,36	0,13	83,03	0,13
4.2	**83,85**	**0,12**	**84,48**	**0,12**	**84,17**	**0,12**

Tabela 4 – Índices de classificação com uso do classificador kNN

Etapa	*Recall* (%)	*DPR*	*Precision* (%)	*DPP*	*F1-measure* (%)	*DPF1*
1	61,89	0,13	63,16	0,15	62,52	0,14
2	62,43	0,12	63,26	0,13	62,84	0,12
3.1	76,03	0,04	80,25	0,03	78,08	0,04
3.2	73,02	0,04	80,95	0,04	76,78	0,04
4.1	76,03	0,04	80,25	0,03	78,08	0,04
4.2	**85,56**	**0,09**	**87,91**	**0,08**	**86,72**	**0,08**

selecionar 528 termos. Selecionando mais termos, o valor de F1-measure apresentava um decréscimo.

A estratégia de seleção de características com uso de Algoritmos Genéticos (etapa 3.2) buscou selecionar os termos que produziram aumento na densidade de cada categoria da base, uma vez que a proximidade dos comentários de uma mesma categoria aos seus respectivos centroides favorece a categorização de novos comentários.

A aplicação desta estratégia reduziu a dimensionalidade da base para 710 termos.

Ambas as estratégias das etapas 3.1 (SFS) e 3.2 (Genético) foram aplicadas à base de dados a partir da etapa 2.

Por último, foi aplicada a extração de termos mais raros e termos mais comuns, através da combinação de percentuais de retiradas de termos pertencentes a estes extremos no ranking de frequências dos termos da base, buscando-se a combinação que produzisse aumento na métrica F1-measure.

Assim, a etapa 4.1 realizou combinações de retirada de termos raros e comuns após a seleção de características aplicada com o SFS (etapa 3.1), enquanto a etapa 4.2 realizou a mesma tarefa, porém após a seleção de características aplicada com o Algoritmo Genético (etapa 3.2).

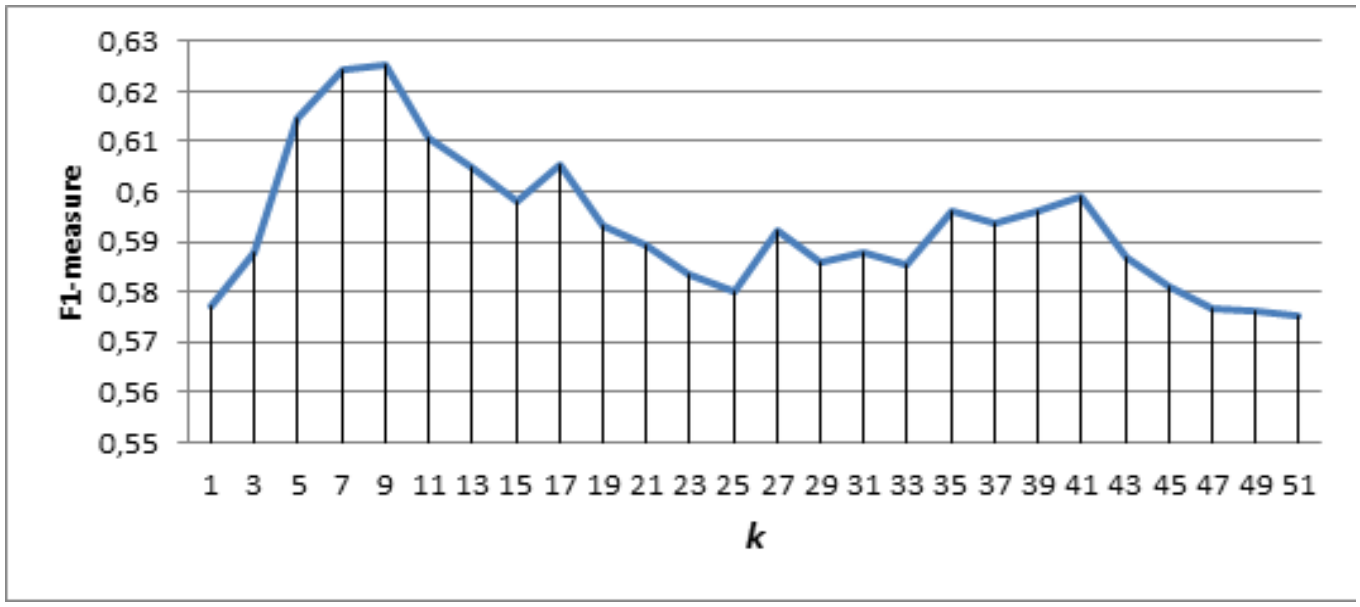
Com o classificador SVM (Tabela 2), a aplicação do Algoritmo Genético conjugado com a extração de 4% dos termos mais comuns produziu um aumento nas medidas de desempenho (etapa 4.2) em relação aos obtidos nas etapas anteriores.

Os dados da classificação da etapa 4.2 foram analisados, sendo observado que, em média, 78,45% dos comentários que deveriam ser reprovados realmente foram reprovados pela técnica proposta.

O uso do classificador CBC (Tabela 3) seguiu o mesmo padrão dos resultados obtidos na Tabela 2.

Os dados da classificação da etapa 4.2 foram

Figura 1 – Escolha do valor de k. O valor de k foi selecionado com o objetivo de aumentar o valor do F1-measure.



analisados, sendo observado que, em média, 83,09% dos comentários que deveriam ser reprovados realmente foram reprovados pela técnica proposta.

Para aplicação do classificador kNN, a base de dados foi submetida a um processo de calibração que consistiu em comparar a média da medida F1-measure para um conjunto variado de valores de k.

Na Figura 1 (pagina 46) é exibido um gráfico comparativo para a calibração e escolha do valor de k.

Para tornar a avaliação mais criteriosa, foi aplicado em cada caso o teste não paramétrico de Wilcoxon pareado com nível de significância de 1% sobre os resultados da medida F1-measure. Na avaliação foi percebido que os resultados obtidos para os três classificadores nas etapas 3.1 e 3.2 foram estatisticamente superiores aos obtidos nas etapas 1 e 2.

Da mesma forma, os resultados obtidos na etapa 4.2 foram estatisticamente superiores aos obtidos na etapa 3.2.

Não houve diferença estatística entre os resultados da etapa 1 e etapa 2, e entre os resultados da etapa 3.1 e 4.1.

CONCLUSÕES

Este artigo apresentou uma metodologia para moderação de comentários de sites de notícias, identificando os comentários que podem ou não ser divulgados na mídia, auxiliando os profissionais da área na sua árdua tarefa.

Os experimentos indicaram que o uso de estratégias de redução de dimensionalidade aliadas às estratégias de seleção de termos com uso de Algoritmos Genéticos seguido da extração de termos raros e de termos comuns retornaram uma melhora nos resultados, além de reduzir significativamente a quantidade de termos usados para classificação.

Pesquisas futuras incluem o uso de uma base de dados com um número maior de amostras e o uso de técnicas de processamento de linguagem natural para melhorar a qualidade na identificação dos comentários que devem ser reprovados para divulgação.

Também serão pesquisadas técnicas de ponderação das classes, de forma a reduzir o risco de que um comentário com conteúdo impróprio seja aprovado pelo classificador.

REFERÊNCIAS

BAEZA-YATES, R., RIBEIRO-NETO, B.. Modern Information Retrieval: the concepts and technology behind search. London, Pearson Education Limited. 2011. ed. 2th, p. 320.

BASAVARAJU, M.; PRABHAKAR, D. R. A novel method of spam mail detection using text based clustering approach. International Journal of Computer Applications, International Journal of Computer Applications, 244 5 th Avenue,# 1526, New York, NY 10001, USA India, v. 5, n. 4, p. 15–25, 2010.

CATARINA, A. S.; BACH, S. L. Estudo do efeito dos parametros genéticos sobre a solução otimizada e sobre o tempo de convergência em algoritmos genéticos com codificações binária e real. Acta Scientiarium, Tecnology, v. 2, n. 2, p. 147–152, 2003.

CHEN, M.-S.; HAN, J.; YU, P. S. Data mining: an overview from a database perspective. Knowledge and data Engineering, IEEE Transactions on, IEEE, v. 8, n. 6, p. 866–883, 1996.

DUARTE, E. S. Sentiment analysis on twitter for the portuguese language. Faculdade de Ciências e Tecnologia, 2013.

HAN, E.-H. S.; KARYPIS, G. Centroid-based document classification: analysis and experimental results. [S.l.]: Springer, 2000.

JOACHIMS, T. Making large scale svm learning practical. Universit¨at Dortmund, 1999.

KOHAVI, R.. A study of cross-validation and bootstrap for accuracy estimation and model selection. v. 14:1137-1145, 1995.

LADHA L., DEEPA T.. Feature Selection Methods and Algorithms, International Journal on Computer Science and Engineering (IJCSE), v.3, n 5 (2011): 1787-1797.

MITCHELL, M. An introduction to genetic algorithms. 8th. ed. [S.l.]: MIT Press Cambridge, 2002.

ORENGO, V.M. AND C.R. HUYCK, A Stemming Algorithm for the Portuguese Language. Laguna de San Raphael, Chile. International joint Conference on artificial intelligence. 2001. p. 183 – 193.

PAK, A.; PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. In: LREC. [S.l.: s.n.], 2010.

POURHASHEMI, S. M.; OSAREH, A.; SHADGAR, B. E-mail spam filtering by a new hybrid feature selection method using Chi2 and CNB Wrapper. Int. J. Emerg. Sci, v. 3, n. 4, p. 410–422, 2013.

ROBERTSON, S. Understanding inverse document frequency: on theoretical arguments for idf. Laguna de San Raphael, Chile. Journal of documentation 2004. v. 60, p. 503 – 520.

SALTON, G., WONG, A., YANG, C. S. A vector space model for automatic indexing. Communications of the ACM. 1975. v. 18. p. 613 – 620.

SHAKHNAROVICH, G.; INDYK, P.; DARRELL, T. Nearest-neighbor methods in learning and vision: theory and practice. [S.l.: s.n.], 2006.

SHRIVASTAVA, J. N.; BINDU, M. H. E-mail spam filtering using

VAPNIK ,V. AND CORTES, C.. Support-vector networks. Machine learning 20.3 (1995): 273-297.

YANG, Y.; LIU, X. A re-examination of text categorization methods. In: ACM. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. [S.l.], 1999. p. 42–49.

Artigo submetido em:29.08.2014

Artigo aceite poara publicação em: 29.12.2014