Combinando Fatores de Ponderação para Melhorar a Classificação de Textos

Frederico P. de Souza[1], **Patrick M. Ciarelli**[2], **Elias de Oliveira**[1]

[1] Departamento de Informática – Universidade Federal do Espírito Santo (UFES)
29.075-910 – Vitória – ES – Brasil

[2]Departamento de Tecnologia Industrial - Universidade Federal do Espírito Santo (UFES)

frsouza@ymail.com, {elias,pciarelli}@lcad.inf.ufes.br

Abstract. *Automatic text categorization is an important tool to organize, filter and distribute documents based on their content. To do this, it is important to differentiate the features (words or terms) inside the document that is used to identify the category it belongs, giving to these features different weights. This paper presents an alternative way to weigh terms in a document, that was derived from other two methods, and shows the benefits of the new method comparing its results with those obtained by other methods.*

Resumo. *A classificação automática de textos é uma importante ferramenta que permite organizar, filtrar e distribuir documentos de acordo com o conteúdo. Para isso, é importante diferenciar as características (palavras ou termos) presentes no documento que permitem identificar a categoria à qual ele pertence, dando a estas características um peso maior. O presente trabalho apresenta uma forma alternativa de ponderar os termos de um documento, derivada de outros dois métodos, e mostra as vantagens obtidas com a nova técnica comparando seus resultados com os de outras técnicas disponíveis.*

1. Introdução

Os documentos produzidos atualmente precisam ser organizados de alguma forma. Os seres humanos utilizam um conjunto de critérios para realizar, de modo manual, esta organização, seja ela por assunto, por prioridade, etc. Quando o volume de documentos é muito grande, torna-se dispendioso realizar a classificação manualmente. A classificação automática de textos é uma alternativa menos custosa para resolver este problema.

De modo geral, qualquer problema que envolva um processo de decisão sobre como organizar documentos ou, até mesmo, enviar de maneira seletiva determinado conteúdo, pode fazer uso da classificação automática de textos. No entanto, o problema de classificação não é trivial, principalmente, quando a decisão envolve muitas classes ou categorias.

Recentes estudos publicados, como [Wang e Zhang 2013], [Lan et al. 2009], [Erenel et al. 2011] e [Debole e Sebastiani 2004], trazem abordagens que visam melhorar o resultado das classificações. Os trabalhos mencionados concentram-se, principalmente, em apresentar métodos que aumentam a importância das características (palavras ou termos) que melhor definem uma classe de documentos, atribuindo a elas um peso maior. Em [Wang e Zhang 2013] são apresentadas duas formas de ponderação o icf e o *icf-based*.

O *icf-based* foi desenvolvido a partir de uma combinação entre o icf e o *Relevance Frequency* [Lan et al. 2009]. Em [Debole e Sebastiani 2004], técnicas de filtragem como o χ^2 são empregadas com o mesmo objetivo.

O objetivo do presente trabalho é comparar os resultados obtidos com diferentes formas de ponderação, aplicando-as em três bases de dados distintas, uma com textos escritos em inglês e as outras duas com textos em português. Dentre as formas utilizadas, é apresentada uma nova alternativa, o $rf{*}idf$, derivada da combinação entre o *Relevance Frequency* e o idf [Robertson 2004], que, nos experimentos realizados, apresentou resultados consistentes, e por vezes superiores aos das outras técnicas mencionadas.

Este trabalho está dividido em sete seções. Na segunda seção, são introduzidos conceitos básicos e apresentados detalhes sobre como um documento pode ser representado em um espaço vetorial. Na terceira seção, os fatores de ponderação que serão utilizados no estudo comparativo são explicados. Na quarta, são apresentados os algoritmos de classificação utilizados. Na quinta seção, as métricas utilizadas para avaliar os resultados são detalhadas. Na sexta, foram incluídos os resultados dos experimentos e, na sétima, as conclusões do trabalho.

2. O Modelo Vetorial

Uma das formas de realizar a classificação automática é dispor os documentos em um espaço vetorial. Neste espaço, cada documento d_j é representado por um vetor de n dimensões não binário e não-negativo (índice); cada posição no vetor representa um termo k_i; cada termo é associado também a um peso $w_{i,j}$, que pode ser obtido, por exemplo, por meio da contagem das ocorrências do respectivo termo dentro do documento representado (*Term Frequency* - tf)[Salton et al. 1975]. Desta forma, o documento é representado conforme expresso na Equação 1.

$$d_j = (w_{1,j}, w_{2,j}, ..., w_{i,j}, ..., w_{n,j}) \tag{1}$$

Por meio dessa representação é possível utilizar métricas para medir as distâncias entre os diversos pontos no espaço vetorial. A Similaridade Cosseno é uma das medidas utilizadas, sendo calculada conforme a Equação 2.

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

A coleção de documentos que compõe o espaço vetorial pode estar organizada em $|C|$ categorias. A classificação de textos consiste em atribuir um valor booleano a cada par $[d_j, C_p]$, p = 1,..., $|C|$, onde 0 significa "não pertence a C_p", e 1 significa "pertence a C_p". Quando d_j só puder pertencer a uma categoria C_p, a classificação é de apenas um rótulo. Todas as bases de dados deste trabalho são de apenas um rótulo.

Do ponto de vista da classificação, é interessante que a similaridade entre os documentos que pertencem a uma mesma categoria aumente, tornando-a mais densa. Ao mesmo tempo, é desejável que a distância entre estas categorias também aumente, evitando sobreposição entre elas [Salton et al. 1975].

O efeito disso pode ser melhor compreendido pela análise da Figura 1, onde são apresentadas duas configurações de agrupamentos de textos. Considere que os quadrados representam locais fictícios criados para indicar o centro de cada categoria (centroides), enquanto que os documentos são representados pelos pontos pretos e as categorias delimitadas pelas circunferências. Na Figura 1.a, os documentos estão mais espalhados em relação ao centro da categoria e há uma maior sobreposição das categorias, pois muitos documentos podem ser enquadrados em mais de uma categoria. Já na Figura 1.b, por serem mais semelhantes, os documentos de uma mesma categoria estão mais próximos ao centro, tornando as classes mais compactas e diminuindo o efeito da sobreposição.

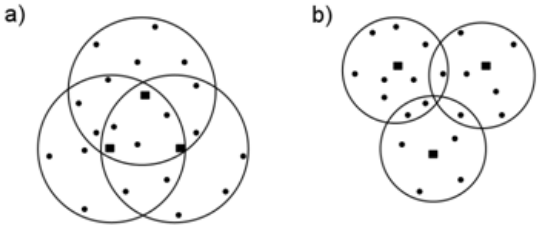

Figura 1. Representação gráfica da densidade dos conjuntos.

Para obter o efeito representado na Figura 1.b, o peso dos termos $w_{i,j}$ pode ser ajustado utilizando-se, para isso, uma propriedade relacionada à semântica dos termos, pois os mais genéricos costumam indexar uma quantidade maior de documentos. Por exemplo, o termo "veículo" costuma indexar mais documentos do que o termo "automóvel" [Jones 2002]. Sendo assim, adotando uma fórmula adequada para calcular o peso, a importância dos termos que ocorrem em uma quantidade maior de documentos pode diminuir, uma vez que estes são menos relevantes para a classificação. Os algoritmos de classificação têm seu desempenho influenciado pelo uso destas fórmulas. A fórmula mais comum é denominada "*Inverse Document Frequency*" (idf) [Robertson 2004], sendo calculada conforme Equação 3:

$$idf_i = \log \frac{|N|}{|n_i|} \tag{3}$$

Nela, $|N|$ é a quantidade total de documentos e $|n_i|$ o número de documentos em que um determinado termo k_i aparece. O peso final do termo pode ser obtido através da multiplicação do número de ocorrências do termo no documento pelo idf correspondente ($tf{*}idf$).

De modo geral, em todo trabalho, será utilizada por padrão a fórmula $w_{i,j} = tf * idf$ para atribuir peso aos termos e a função Similaridade Cosseno, Equação 2, para calcular a distância entre os vetores, salvo disposição em contrário.

3. Fatores de Ponderacão

Além do idf, são utilizados os seguintes fatores de ponderação para calcular o peso dos termos:

Inverse-Category-Frequency (icf): Utilizado em [Wang e Zhang 2013], é similar ao idf, com a diferença de que quanto maior o número categorias em que um termo

aparece, menor a sua importância. Nele, $|C|$ é a quantidade total de categorias e $|c_i|$ é o número de categorias em que um determinado termo k_i aparece.

$$icf_i = \log\left(1 + \frac{|C|}{|c_i|}\right) \tag{4}$$

Qui-Quadrado (χ^2): O χ^2 foi utilizado para atribuir peso aos termos em [Debole e Sebastiani 2004]. A Equação 5 extraída de [Yang e Jan 1997] mostra como o χ^2 pode ser calculado para um dado termo k_i em uma classe C_p:

$$\chi^2_{(k_i,C_p)} = \frac{|N| * (a * d - c * b)^2}{(a + c) * (b + d) * (a + b) * (c + d)} \tag{5}$$

Sendo: $|N|$ o número total de documentos, a a quantidade de documentos que possuem o termo k_i e que pertencem a C_p, b a quantidade de documentos que possuem o termo k_i, mas não pertence à C_p, c o número de documentos que pertencem à C_p, mas não possuem o termo k_i e d o número de documentos que não possuem k_i e não pertencem à C_p.

Relevance Frequency (rf): Introduzido por [Lan et al. 2009], gera um fator para o termo em cada classe que ele aparece, sua formula é dada conforme Equação 6. Onde, a e b são os mesmos usados na Equação 5.

$$rf_{(k_i,C_p)} = \log\left(2 + \frac{a}{max(1, b)}\right) \tag{6}$$

ICF-Based: Proposto por [Wang e Zhang 2013], combina as fórmulas de *Relevance Frequency* (Equação 6) e de icf (Equação 4).

$$icf - based_{(k_i,C_p)} = \log\left(2 + \frac{a}{max(1, b)} * \frac{|C|}{|c_i|}\right) \tag{7}$$

Pode-se perceber que o χ^2, o rf e o *icf-based* de um dado termo k_i são calculados para cada classe C_p. Desta forma, é necessário definir qual valor será utilizado para atribuir peso aos termos. Em [Debole e Sebastiani 2004] é usada a Equação 8 para selecionar o maior deles. Nela, f representa o fator de ponderação empregado.

$$f_{max}(k_i) = max_{p=1}^{|C|} f_{(k_i,C_p)} \tag{8}$$

RF*IDF: Neste trabalho é apresentada uma forma alternativa de ponderação que multiplica os valores de *Relevance Frequency* (Equação 6) pelos valores de idf (Equação 3). Desta forma, este novo fator de ponderação considera em sua composição a distribuição dos termos entre as classes e os documentos.

$$rf * idf = max_{p=1}^{|C|} rf_{(k_i,C_p)} * \log\frac{|N|}{|n_i|} \tag{9}$$

4. Algoritmos de Classificação

Para fins de avaliação, os fatores de ponderação foram utilizados com dois algoritmos de classificação: o *k-Nearest Neighbor Algorithm* (k-NN) e o *Centroid-Based Classifier*. Uma breve descrição dos algoritmos pode ser encontrada abaixo:

k-Nearest Neighbor Algorithm: Nele, dado um novo documento d_j, a similaridade entre o novo documento e cada um dos documentos que compõem a base de treinamento pré-classificada é calculada, os k documentos mais próximos são selecionados e o novo documento é classificado de acordo com a classe mais frequente entre os k mais próximos [Masand et al. 1992]. Neste trabalho, será utilizada a Equação 2 no cálculo da similaridade. O valor definido para k será de $k = 10$, mesmo parâmetro utilizado em [Wang e Zhang 2013].

Centroid-Based Classifier: Para cada categoria C_p contendo m documentos, um centroide (c_p) é calculado, sendo cada elemento do centroide definido como a média dos pesos de cada termo (i) pertencente aos documentos da categoria C_p, Equação 10. O novo documento é classificado de acordo com a categoria do centroide mais próximo [Han e Karypis 2000].

$$c_{i,p} = \frac{1}{m} \sum_{j \in C_p} w_{i,j} \tag{10}$$

5. Métricas para Avaliação dos Resultados

Um conjunto de métricas será utilizado para avaliar os resultados dos fatores de ponderação utilizados. As métricas são calculadas conforme abaixo [Sebastiani 2002] e [Yang e Liu 1999]:

F_1: Realiza a média harmônica entre o *Recall* (R) e o *Precision* (P) (Equações 11 e 12, respectivamente) encontrados para cada classe C_p dando o mesmo peso a ambos, conforme Equação 13:

$$R(C_p) = \frac{TP(C_p)}{TP(C_p) + FN(C_p)} \tag{11}$$

$$P(C_p) = \frac{TP(C_p)}{TP(C_p) + FP(C_p)} \tag{12}$$

$$F_1(C_p) = \frac{2P(C_p)R(C_p)}{P(C_p) + R(C_p)} \tag{13}$$

Na equação de *Recall* e *Precision*, TP é a quantidade de documentos atribuídos à classe C_p pelo especialista e pelo classificador automático, FP é a quantidade de documentos que não foram atribuídos à classe C_p pelo especialista, mas que foram atribuídos pelo classificador automático e FN é a quantidade de documentos que foram atribuídos à classe C_p pelo especialista, mas que não foram atribuídos pelo classificador automático.

Macro Average F_1: É a média simples do F_1 de cada classe C_p, e é calculada conforme Equação 14:

$$macF_1 = \frac{\sum_{p=1}^{|C|} F_1(C_p)}{|C|} \tag{14}$$

Micro Average F_1: Baseado no *Recall* e *Precision* calculados para o conjunto de todas as classes, ao invés de para cada classe individual (Equações 15 e 16, respectivamente), é dado pela Equação 17:

$$R = \frac{\sum_{p=1}^{|C|} TP(C_p)}{\sum_{p=1}^{|C|} TP(C_p) + FN(C_p)} \tag{15}$$

$$P = \frac{\sum_{p=1}^{|C|} TP(C_p)}{\sum_{p=1}^{|C|} TP(C_p) + FP(C_p)} \tag{16}$$

$$micF_1 = \frac{2PR}{P+R} \tag{17}$$

6. Experimentos

6.1. Bases de Dados

Para realizar os testes, foram utilizadas três bases de dados[1]. A primeira contém 45.907 textos jornalísticos em português publicados pelo jornal *A Tribuna*. Os documentos estão organizados em 21 categorias que representam as seções do jornal onde foram publicados. As categorias e o número de documentos em cada uma delas são informados na Tabela 1.

Tabela 1. Número de documento de *A Tribuna* organizados por categoria.

Classe	N.Docs	Classe	N.Docs	Classe	N.Docs
AT2	5618	Especial	1470	Opinião	1634
Qual a Bronca?	346	Família	442	Polícia	4671
Cidades	5234	Imóveis	124	Política	5918
Ciência e Tecnologia	470	Informática	1506	Regional	1802
Concursos	309	Internacional	2187	Sobre Rodas	352
Economia	6557	Minha Casa	37	Tudo a Ver	30
Esportes	6657	Mulher	103	TV Tudo	440

A segunda, é a base de dados *20 NewsGroups*, que contém 19.997 textos escritos em língua inglesa, publicados em 20 (vinte) *newsgroups* diferentes. O cabeçalho dos documentos desta base foram filtrados, permanecendo apenas o assunto, as palavras chaves e o conteúdo.

A terceira é uma base de dados que contém pequenos textos extraídos do resumo de patentes. A base está dividida em 08 (oito) categorias e possui um total de 2.249 documentos.

[1]Os dados utilizados nos experimentos podem ser encontrados em http://www.inf.ufes.br/∼elias/onTheBeach/2014/data.rar

Os documentos das bases de dados foram submetidos a uma etapa de pré-processamento. Cada palavra contida nos documentos foi extraída e contada, obtendo então a frequência de cada uma delas por documento. Para a base *A Tribuna*, foram encontradas 168.212 palavras diferentes, para a base *20 NewsGroups*, 111.418 palavras diferentes, e para a base *Patentes*, 13.884 palavras diferentes. O conjunto das palavras com suas respectivas frequências deu origem a um índice, que passou a representar seu respectivo documento. Durante o processo de indexação, todas as palavras foram colocadas em maiúsculas e foram excluídos caracteres especiais, números, acentos e *stopwords*.

Visando diminuir a quantidade total de palavras e, consequentemente, o tamanho do espaço vetorial, foi empregada uma técnica para redução da dimensionalidade. Dentre as técnica disponíveis, pode-se destacar as técnicas de filtragem, onde um termo k_i é selecionado, caso tenha obtido um valor superior ao limite em uma função f [Sebastiani 2002].

A função f escolhida foi a do χ^2, Equação 5. Os χ^2 dos termos foram calculados para cada uma das categoria e, então, os resultados foram combinados através da Equação 18.

$$\chi^2_{avg}(k_i) = \sum_{p=1}^{m} P_r(C_p) * \chi^2(k_i, C_p) \tag{18}$$

Onde $P_r(C_p)$ é a probabilidade de um documento d_j pertencer à classe C_p. O termo foi selecionado após obter um valor superior ao limite definido para cada base de dados. Para a base de dados *A Tribuna*, o limite definido foi aproximadamente 3,57; para a base *20 NewsGroup*, 2,99; e para a base *Patentes*, 3,89.

Após a redução da dimensionalidade, restaram 35.494 palavras diferentes na base de dados *A Tribuna*, 24.673 na base de dados *20 NewsGroup* e 865 na base *Patentes*.

Além disso, os pesos calculados para os termos foram normalizados. Este processo visa diminuir os efeitos causados por documentos extensos, que possuem maior chance de serem classificados erroneamente [Jones 2002]. A Equação 19 mostra como é feita a normalização. O idf pode ser substituído por outro fator de ponderação na expressão, conforme o caso.

$$w_{i,j} = \frac{tf * idf(k_i, d_j)}{\sqrt{\sum_{s=1}^{|K|}(tf * idf(k_s, d_j))^2}} \tag{19}$$

Dito isto, a Tabela 2 mostra as informações sobre o espaço vetorial formado pelos índices obtidos após indexação dos documentos, extração de palavras e normalização dos pesos, usando como base o idf. As métricas utilizadas foram extraídas de [Salton et al. 1975]. O centroide principal, que consta das métricas utilizadas, é calculado para o conjunto total de documentos, utilizando também a Equação 10.

Para fins de experimento, os documentos que compõem as bases de dados foram divididos em duas partes de forma aleatória e estratificada. Uma das partes, contendo 33% dos documentos, foi utilizada para os testes e a outra, com os documentos restantes, para o treinamento dos algoritmos.

Tabela 2. Caracterização das bases de dados. Tamanho Médio dos Índices (MI), Média da Similaridade dos Documentos com seus Centroides (SDC), Média da Similaridade entre os Centroides e o Centroide Principal (SCC) e Média da Similaridade entre Pares de Centroide (SPC)

Base de Dados	MI	SDC (x)	SCC	SPC (y)	Razão (y/x)
A Tribuna	188,18	0,1905	0,6142	0,3510	1,8425
20 NewsGroups	72,61	0,1743	0,5447	0,2611	1,4975
Patentes	17,66	0,2355	0,6225	0,3027	1,2851

6.2. Resultados

Nas Tabelas 3 e 4, são apresentados, respectivamente, os resultados de macro-$F1$ e micro-$F1$ obtidos para cada um dos fatores de ponderação aplicados sobre a base de dados *A Tribuna*. Da análise, pode-se observar que a técnica proposta, o $rf{*}idf$, obteve bons resultados com o algoritmo Centroid. O macro-$F1$ obtido foi 1,60% superior ao do idf, segundo melhor resultado. Quanto ao micro-$F1$, pode-se considerar que $rf{*}idf$ e o idf foram equivalentes. Além disso, os resultados obtidos pelo idf, com o algoritmo Centroid, foram superiores aos do icf, rf, *icf-based* e χ^2.

Com o algoritmo de classificação k-NN, o $rf{*}idf$ apresentou valor 0,84% superior para o micro-$F1$, quando comparado ao segundo melhor colocado, que neste caso foi o *icf-based*. No entanto, os resultados de macro-$F1$ obtidos pelo idf foram 0,47% superiores ao do $rf{*}idf$. Além disso, O icf apresentou resultados superiores ao rf, *icf-based* e χ^2, sendo este o que obteve o pior resultado em ambas as tabelas para ambos algoritmos.

Tabela 3. Resultados da base *A Tribuna* - Macro Average F_1 .

Classificador	$tf{*}idf$	$tf{*}\chi^2$	$tf{*}rf$	$tf{*}icf$	$tf{*}$*icf-based*	$tf{*}rf{*}idf$
k-NN	**0,5944**	0,5043	0,5608	0,5904	0,5770	0,5897
Centroid	0,6066	0,4619	0,5671	0,5862	0,5622	**0,6226**

Tabela 4. Resultados da base *A Tribuna* - Micro Average F_1 .

Classificador	$tf{*}idf$	$tf{*}\chi^2$	$tf{*}rf$	$tf{*}icf$	$tf{*}$*icf-based*	$tf{*}rf{*}idf$
k-NN	0,7822	0,7151	0,7769	0,7831	0,7794	**0,7915**
Centroid	0,7347	0,6029	0,6662	0,6718	0,6622	**0,7348**

As Tabelas 5 e 6, apresentam os resultados obtidos na base de dados *20 Newsgroup*. Neste caso, o $rf{*}idf$ obteve desempenho superior quando utilizado com o algoritmo Centroid. O macro-$F1$ e micro-$F1$ obtidos foram, respectivamente, 1,36% e 1,27% superiores ao idf, que, neste caso, foi o segundo melhor resultado, tendo superado o icf, *icf-based*, rf e χ^2.

O $rf{*}idf$ foi aquele que também obteve melhor resultado para macro-$F1$ e micro-$F1$ com o algoritmo k-NN na base de dados *20 Newsgroup*. Os resultados obtidos pelo $rf{*}idf$ foram, respectivamente, 1,59% e 1,62% superiores ao icf. O rf foi o terceiro melhor resultado com o algoritmo k-NN, tendo superado com uma pequena margem o idf.

Tabela 5. Resultados da base *20 NewsGroup* - Macro Average F_1 .

Classificador	$tf*idf$	$tf*\chi^2$	$tf*rf$	$tf*icf$	$tf*$*icf-based*	$tf*rf*idf$
k-NN	0,8148	0,6294	0,8202	0,8227	0,7815	**0,8386**
Centroid	0,8270	0,6170	0,7859	0,8059	0,7669	**0,8406**

Tabela 6. Resultados da base *20 NewsGroup* - Micro Average F_1 .

Classificador	$tf*idf$	$tf*\chi^2$	$tf*rf$	$tf*icf$	$tf*$*icf-based*	$tf*rf*idf$
k-NN	0,8162	0,6295	0,8219	0,8246	0,7852	**0,8408**
Centroid	0,8291	0,6149	0,7835	0,8065	0,7662	**0,8418**

Na base *Patentes*, Tabelas 7 e 8, o $rf*idf$ apresentou resultado superior em todos os casos. Com o algoritmo Centroid, seus valores de macro-$F1$ e micro-$F1$ foram, respectivamente, 2,43% e 2,39% maiores do que o segundo colocado e com o algoritmo k-NN, estes percentuais foram de 2,12% e 2,03%.

Tabela 7. Resultados da base *Patentes* - Macro Average F_1 .

Classificador	$tf*idf$	$tf*\chi^2$	$tf*rf$	$tf*icf$	$tf*$*icf-based*	$tf*rf*idf$
k-NN	0,7154	0,6418	0,6236	0,6479	0,5936	**0,7366**
Centroid	0,7071	0,6130	0,6341	0,7014	0,6250	**0,7314**

Tabela 8. Resultados da base *Patentes* - Micro Average F_1 .

Classificador	$tf*idf$	$tf*\chi^2$	$tf*rf$	$tf*icf$	$tf*$*icf-based*	$tf*rf*idf$
k-NN	0,7941	0,7091	0,7529	0,7636	0,6972	**0,8154**
Centroid	0,7729	0,6719	0,6852	0,7317	0,6905	**0,7968**

Os experimentos indicam que o método proposto $rf*idf$ apresenta resultados no mesmo patamar ou até melhores que os demais métodos avaliados, principalmente para a base *Patentes*. Nas ocasiões em que ele não obteve o melhor resultados, a diferença foi pequena, mostrando que o método é bastante competitivo em relação aos demais.

7. Conclusão

Este trabalho apresentou diversas formas de atribuir peso aos termos contidos em um documento. Os experimentos foram realizados em três bases de dados distintas contendo textos em inglês e português. Cinco formas diferentes de ponderação, desde as mais clássicas até as publicadas em estudos recentes, foram analisadas e comparadas. Além disso, foi introduzida uma alternativa de ponderação que combina outras duas formas anteriores.

Dos resultados obtidos, percebeu-se que a ponderação dos termos influencia significativamente a qualidade da classificação. Os experimentos mostraram que o método de ponderação proposto $rf*idf$ apresentou resultados superiores aos demais métodos na maioria das vezes, principalmente para o algoritmo Centroid. Para o algoritmo k-NN, o método proposto foi inferior somente a técnica idf em um único caso, e ainda assim a

diferença entre os resultados foi pequena. Esses resultados indicam que o método proposto é uma boa opção para ponderação dos termos. Como pesquisas futuras, mais estudos serão feitos com novas bases de dados e diferentes técnicas de classificação, além de serem pesquisadas outras formas de ponderar a combinação de técnicas utilizadas.

Referências

Debole, F. e Sebastiani, F. (2004). Supervised term weighting for automated text categorization. In *Text mining and its applications*, pages 81–97. Springer Berlin Heidelberg.

Erenel, Z., Altincay, H., e Varoglu, E. (2011). Explicit use of term occurrence probabilities for term weighting in text categorization. In *Journal of Information Science and Engineering*, volume 27 (3), pages 819–834. Institute of Information Science.

Han, E. e Karypis, G. (2000). Centroid-based document classification: Analysis and experimental results. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 424–431. Springer Berlin Heidelberg.

Jones, K. S. (2002). A statistical interpretation of term specificity and its application in retrieval. In *Journal of documentation*, volume 18 (01), pages 11–21. Emerald Group Publishing.

Lan, M., Tan, C. L., Su, J., e Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 721–735. IEE Computer Society.

Masand, B., Linoff, G., e Waltz, D. (1992). Classifying news stories using memory based reasoning. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 59–65. ACM.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for idf. In *Journal of documentation*, volume 60 (05), pages 503–520. Emerald Group Publishing.

Salton, G., Wong, A., e Yang, C. (1975). A vector space model for automatic indexing. In *Communications of the ACM*, volume 18 (11), pages 613–620. ACM.

Sebastiani, F. (2002). Machine learning in automated text categorization. In *ACM computing surveys*, volume 34 (01), pages 1–47. ACM.

Wang, D. e Zhang, H. (2013). Inverse-category-frequency based supervised term weighting schemes for text categorization. In *Journal of Information Science and Engineering*, pages 209–225. Emerald Group Publishing.

Yang, Y. e Jan, O. P. (1997). A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420. International Machine Learning Society.

Yang, Y. e Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM.