



II CONPESQ Congresso de Pesquisa, Pós-Graduação e Inovação

Os novos rumos da ciência pós-pandemia

12 a 16 de abril de 2021 Universidade Federal do Cariri - UFCA

ESTUDO COMPARATIVO DE FEATURE EVALUATION FUNCTIONS UTILIZANDO O MÉTODO ALOFT

GILVAN GOMES DA SILVA JUNIOR¹

Universidade Federal do Cariri
gilvan.gomes@aluno.ufca.edu.br

ROBERTO HUGO WANDERLEY PINHEIRO²

Universidade Federal do Cariri
roberto.hugo@ufca.edu.br

RESUMO: Escolher uma Função de Avaliação de Características (*Feature Evaluation Function*) é um aspecto relevante na classificação de documentos. A grande quantidade de funções que existem é um fator que dificulta a escolha, além de não existir uma função com boa performance para qualquer base de dados. Neste trabalho, avaliamos a efetividade de algumas *Feature Evaluation Functions* em conjunto da metodologia de seleção de características *At Least One FeaTure*, em bases de dados menos volumosas. Os experimentos são conduzidos usando doze *Feature Evaluation Functions* e cinco bases de dados. Dentre as *Feature Evaluation Functions* usadas, a que teve um maior desempenho foi a *Gain Ratio*.

PALAVRAS-CHAVE: Classificação de Documentos; Seleção de Características; Funções de Avaliação de Características.

ABSTRACT: Choose a Feature Evaluation Function is a relevant aspect of text categorization. The large amount of functions is a factor that difficulties the choice, also there is not a single function with good performance for all datasets. In this work, we evaluated the effectiveness of some Feature Evaluation Functions together with the At Least One FeaTure feature selection methodology, in smaller datasets. The experiments are conducted using twelve Feature Evaluation Functions and five datasets. Among the Feature Evaluation Functions used, the one that had the highest performance was the Gain Ratio.

Keywords: text categorization; feature selection; feature evaluation functions.

¹ Bolsista PIBIC/UFCA no Programa Institucional de Iniciação Científica e Tecnológica da UFCA.
² Orientador.

1 INTRODUÇÃO

Com o estreitamento das relações entre humano e tecnologia, crescimento de informações acessadas e criadas todos os dias e a necessidade de uma gestão de dados, a realização de uma **classificação de documentos** passa a ser indispensável para um melhor arranjo informacional. A classificação de documentos é uma aplicação de aprendizado de máquina que tem por objetivo designar classes que carreguem o significado da fundamentação que um documento possua. Fazendo assim, cada documento pertence a várias, uma ou nenhuma classe (JOACHIMS, 1999). Para realizar a classificação de documentos é necessário adotar um método de representação para que os dados sejam matematicamente computáveis (YAN, 2009).

Neste artigo, utilizaremos a representação *Bag of Words* (BoW), que reúne um aglomerado dos termos relevantes, ou seja, que influenciam na interpretação dos documentos, criando uma lista que indica a presença dos termos em cada documento, que é denominada Vetor de Características. No *Bag of Words* (Bow) se destacam dois problemas: alta dimensionalidade, visto que os documentos apresentam uma extensa quantidade de termos, os quais em grande parte não são relevantes para a categorização (KOBAYASHI *et al.*, 2018); e esparsidade dos dados, visto que uma grande porcentagem de termos não aparece na maioria dos documentos resultando em Vetores de Características com muitas ausências, isto é, valores nulos (SU; SAYYAD-SHIRABAD; MATWIN, 2011). Para remediar esta situação, é necessário realizar uma redução de termos. Este procedimento é chamado de seleção de características.

Os métodos de seleção de características são responsáveis pela remoção de termos menos relevantes a partir de dados estatísticos obtidos dos documentos analisados por meio das *Feature Evaluation Function* (YANG; PEDERSEN, 1997; CAROPRESO; MATWIN; SEBASTIANI, 2001). Essas funções determinam as relevâncias dos termos, sendo possível gerar um subconjunto de termos para representar os documentos sem perda significativa de informação.

Métodos de seleção de características que utilizam essas *Feature Evaluation Functions* (FEFs) foram amplamente estudados. Yang e Pedersen (1997) relataram que uma remoção agressiva de termos em uma boa seleção de características pode prover uma melhor classificação, revelando a importância de termos comuns na classificação de documentos. Forman (2003) compara 12 FEFs avaliando-as em 229 casos de problemas de

classificação de documentos e propõe uma nova FEF chamada de *Bi-Normal Separation* (BNS) que apresenta bons resultados. Li *et al.* (2009) realizaram análises teóricas de 6 FEFs e ainda introduzem uma nova FEF a WFO (*Weighed Frequency and Odds*). Diferentemente das propostas anteriores, Pinheiro *et al.* (2012) passa a utilizar as FEF em um método próprio no qual ocorre uma busca por um conjunto de características de modo que seja assegurado que cada documento possua pelo menos uma característica no conjunto final de características, tal método é nomeada de ALOFT (*At Least One FeaTure*). Em seus estudos foram realizados experimentos onde o método ALOFT foi submetido a três bases de dados e utilizando 5 FEFs como métricas de pontuações de termos.

Este artigo se propõe a realizar uma análise quantitativa superior aos experimentos de Pinheiro *et al.* (2012), verificando agora doze (12) *Feature Evaluation Functions* (FEFs) aplicadas no método de seleção de características *At Least One FeaTure* (ALOFT) utilizando os classificadores *Support Vector Machine* (SVM) e *Random Forest* para avaliação de desempenho em bancos de dados textuais.

O restante deste artigo está organizado da seguinte forma. A Seção 2 descreve detalhadamente a metodologia estudada. A Seção 3 relata os resultados e análises experimentais. A Seção 4 conclui este trabalho.

2 METODOLOGIA ESTUDADA

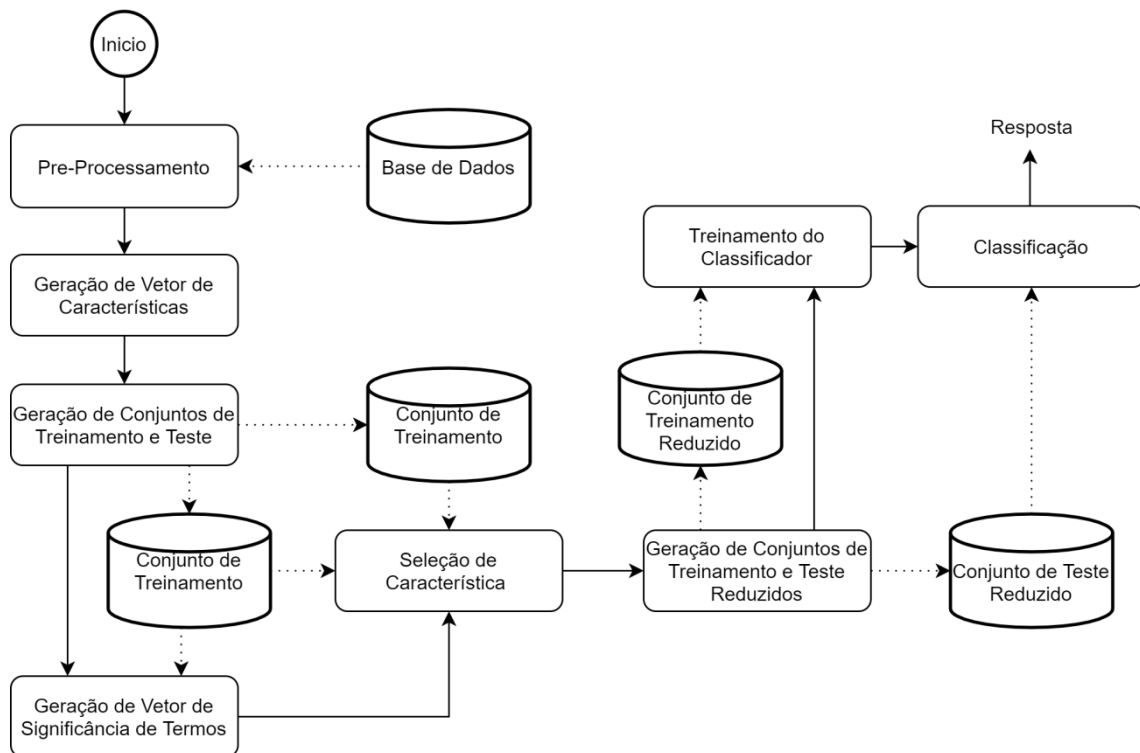
Nesta seção apresentamos a metodologia utilizada para avaliar a efetividade de cada *Feature Evaluation Function* (FEF) aplicada utilizando o método de seleção de características ALOFT. A Figura 1 mostra todas essas etapas que compõem a arquitetura do sistema de Classificação de Documentos, que serão detalhados a seguir.

O processo de classificação de documentos começa com a obtenção da base de dados em linguagem natural (ou *Corpus*) que são transformados pelo pré-processamento. O pré-processamento é a etapa onde são removidos termos de menor significado semântico presentes nos documentos. Além disso, são organizadas as informações para geração do vetor de características. As principais operações de pré-processamento são análise léxica, *stopwords*, *stemming*, tesauro, *lemmatization*, grupos nominais e corretor ortográfico. Recomendamos a leitura de Ceska e Fox (2009) para detalhamento.

Após o pré-processamento é construído o vetor de características, que é uma lista

com todas as palavras presentes no *Corpus*. Ou seja, definindo w_j como a j -ésima palavra do *Corpus*; n o número total de palavras do *Corpus* depois do pré-processamento; m o número de documentos da base de dados. Temos, para cada documento d_i (onde i vai de 1 até m) um vetor de características definido pelo conjunto de termos $\{w_1, w_2, w_3, w_4, w_5, \dots, w_n\}$. Os valores das características w_j de cada documento d_i são calculadas a partir de uma métrica chamada de TF-IDF (*Term Frequency – Inverse Document Frequency*) (RAMOS, 2003).

Figura 1 - Fluxograma que representa a classificação de documentos



Fonte: O autor.

Em seguida é realizada uma validação cruzada estratificada com *10-folds*, ou seja, os documentos são divididos em 10 conjuntos (*folds*) do tamanho mais próximo possível. Após esta divisão os conjuntos são associados com o intuito de gerar diferentes conjuntos de treinamento e teste. Deste modo, tem-se que os experimentos para cada base de dados são realizados 10 vezes, onde em cada execução o conjunto de teste será um dos *folds* e o conjunto de treinamento a combinação dos demais que restaram.

Depois dos conjuntos serem definidos, são realizados os cálculos das probabilidades as quais são necessárias para a obtenção das FEFs. Com esse objetivo foi construído uma função que obtém as informações estatísticas do conjunto de treinamento,

conforme estudadas por Li *et al.* (2009). Assim, em seguida as FEFs são calculadas utilizando diretamente os dados obtidos. Com as FEFs calculadas, o método ALOFT (PINHEIRO *et al.*, 2012) é aplicado, com base nos valores retornados pela função FEF, criando dois novos conjuntos de treinamento e teste, com base nos maiores valores fornecidos pela FEF.

Por fim, para classificar os documentos, utilizamos os classificadores *Support Vector Machine* (SVM) (JOACHIMS, 1998) e *Random Forest* (BREIMAN, 2001) usando o conjunto de treinamento obtido pós-método ALOFT, com intuito de treinar o classificador. Em seguida, cada classificador treinado usa o conjunto de teste com intuito de dar início a classificação de cada documento.

Ao final de tudo, avaliam-se os resultados. Para tal, foram utilizados Macro-F1 e Micro-F1 por serem as métricas mais utilizadas na literatura em se tratando de Classificação de Documentos (BRANK *et al.*, 2002; FORMAN, 2003; PINHEIRO *et al.*, 2012).

3 RESULTADOS E DISCUSSÃO

Nesta seção serão apresentados os experimentos realizados para avaliar o desempenho de cada FEF em conjunto ao método ALOFT. Nas subseções seguintes serão apresentadas as bases de dados utilizadas (Subseção 3.1), as FEFs utilizadas (Subseção 3.2) e por fim os resultados obtidos com considerações sobre os experimentos (Subseção 3.3).

3.1 BASES DE DADOS

Foram utilizadas 5 bases de dados diferentes nos experimentos, vide Tabela 1. As bases de dados passaram pelos pré-processamentos *stopwords* e *stemming*.

Tabela 1 - Informações sobre as bases de dados usadas. Estas informações são: nome da base de dados, número de documentos, número de palavras (termos) presente após o pré-processamento, número de classes e local onde a base foi encontrada

<i>Nome</i>	<i>Documentos</i>	<i>Termos</i>	<i>Classes</i>	<i>URL</i>
BBC	2225	26385	5	http://mlg.ucd.ie/datasets/bbc.html
BBC Sports	737	11964	5	http://mlg.ucd.ie/datasets/bbc.html
Review Polarity	2000	33174	2	http://www.cs.cornell.edu/people/pabo/movie-review-data/
Syskill Webert	334	13804	4	https://kdd.ics.uci.edu/databases/SyskillWebert/SyskillWebert.html

Sentence Classification	900	20169	3	https://archive.ics.uci.edu/ml/datasets/Sentence+Classification
----------------------------	-----	-------	---	---

Fonte: O autor.

3.2 FEATURE EVALUATION FUNCTION

Os cálculos das *Feature Evaluation Function* (FEF) são baseados em probabilidades de um termo (w) relacionados à presença ou ausência deste termo nas classes (c_i) da base de dados. As seguintes probabilidades foram utilizadas para esses cálculos: $P(w)$ é a Probabilidade de um documento conter o termo w ; $P(c_i)$ é a Probabilidade de um documento pertencer à classe c_i ; $P(w|c_i)$ é a probabilidade do termo w estar presente na classe c_i ; $P(w|\bar{c}_i)$ é a probabilidade do termo w não estar presente na classe c_i ; $P(\bar{w}|c_i)$ é a probabilidade dos termos diferentes do termo w de estar presente na classe c_i ; $P(\bar{w}|\bar{c}_i)$ é a probabilidade dos termos diferentes do termo w de não estar presente na classe c_i ; e $P(w, c_i)$, ou $P(w \wedge c_i)$, é a probabilidade de um documento apresentar o termo w e pertencer a classe c_i .

Dentre as FEFs utilizadas nos experimentos temos: a *Bi-Normal Separation*, proposta por Forman a FEF *Bi-Normal Separation* (BNS) (FORMAN, 2003; LI *et al.*, 2009) utiliza a função de probabilidade cumulativa inversa da distribuição Normal (*z-score*), tal função é definida como:

$$BNS(w) = \sum_i \left| F^{-1}(P(w|c_i)) - F^{-1}(P(w|\bar{c}_i)) \right|$$

Chi-Squared Statistics (CHI) (YANG; PEDERSEN, 1997; FORMAN, 2003; LI *et al.*, 2009) é calculada por:

$$CHI(w) = \sum_i \frac{(P(w|c_i) * P(\bar{w}|\bar{c}_i) - P(\bar{w}|c_i) * P(w|\bar{c}_i))^2}{(P(w) * P(\bar{w}) * P(c_i) * P(\bar{c}_i))}$$

Information Gain (IG) (YANG; PEDERSEN, 1997; BRANK *et al.*, 2002; FORMAN, 2003; LI *et al.*, 2009) é calculada por:

$$IG(w) = \sum_i \left(P(w|c_i) * \log \frac{P(w|c_i)}{P(c_i)} \right) + \sum_i \left(P(\bar{w}|\bar{c}_i) * \log \frac{P(\bar{w}|\bar{c}_i)}{P(\bar{c}_i)} \right)$$

Mutual Information (MI) (YANG; PEDERSEN, 1997; LI *et al.*, 2009) é calculada por:

$$MI(w) = \sum_i \log \frac{P(w|c_i)}{P(w)*P(c_i)}$$

A FEF *Multi-class Odds Ratio* (MOR) (CHEN *et al.*, 2009) é uma versão multi-classes da FEF *Odds Ratio*, sua aplicação no lugar da *Odds Ratio* acontece por esta FEF ser uma função de domínio binário. Esta função é definida por:

$$MOR(w) = \sum_i \left| \log \frac{P(c_i)*(1-P(\bar{c}_i))}{P(\bar{c}_i)*(1-P(c_i))} \right|$$

Proposta por Li *et al.* (2009) a FEF *Weighed Frequency and Odds* (WFO) utiliza um parâmetro (λ) que ajusta o peso entre frequência e probabilidades. Para os experimentos realizados neste artigo foram utilizados os parâmetros 0, 0.25, 0.75 e 1, para verificar o desempenho da FEF WFO. Tal função é definida como:

$$WFO(w) = \sum_i (P(w|c_i)^\lambda * (\log \frac{P(w|c_i)}{P(w|\bar{c}_i)})^{1-\lambda})$$

A Gini index (GINI) possui diversas variações, utilizamos a abordagem usada por Shang *et al.* (2009). Esta função é calculada por:

$$GINI(w) = \sum_i (P(w|c_i)^2 * P(c_i|w)^2)$$

A FEF *Class Discriminating Measure* (CDM) (CHEN *et al.*, 2009) é uma função criada a partir da simplificação da FEF MOR. Esta função é definida como:

$$CDM(w) = \sum_i \left| \log \frac{P(w|c_i)}{P(w|\bar{c}_i)} \right|$$

A FEF *Geometric Mean* (GM) (KUBAT; HOLTE; MATWIN, 1998) é calculada por:

$$GM(w) = \sum_i \sqrt{P(w|c_i) * P(\bar{w}|c_i)}$$

Proposta por Quinlan a FEF *Gain Ratio* (QUILAN, 1986; IKONOMAKIS; KOTSIANTIS; TAMPAKAS, 2005) como um refinamento da FEF IG. Sua abordagem é maximizar o ganho de informações do termo, minimizando o número de seus valores. Esta função é definida por:

$$GR(w) = \frac{\sum_i (P(w,c_i) * \log \frac{P(w,c_i)}{P(w)*P(c_i)})}{-\sum_i (P(c_i) * \log(P(c_i)))}$$

A FEF *Probability Ratio* (PR) (FORMAN, 2003; IKONOMAKIS; KOTSIANTIS; TAMPAKAS, 2005) é calculada por:

$$PR(w) = \sum_i \frac{P(w|c_i)}{P(w|\bar{c}_i)}$$

Weighted Log Likelihood Ratio (WLLR) (LI *et al.*, 2009) é calculada por:

$$WLLR(w) = \sum_i (P(c_i) * \log \frac{P(c_i)}{P(\bar{c}_i)})$$

3.3 RESULTADOS OBTIDOS

A Tabela 2 mostra uma ordenação das 12 FEFs, apresentadas anteriormente, em relação ao número de características que foram selecionadas em cada base de dados utilizando o método de seleção de características ALOFT.

Tabela 2 – Ordenação das FEFs que reduziram um maior número de características para as que menos reduziram, em cada base de dados

<i>Syskill Weibert</i>	<i>BBC</i>	<i>BBC Sports</i>	<i>Sentence Classification</i>	<i>Review Polarity</i>
BNS	WFO 1	WFO 0	WFO 1	GINI - WFO 1
WFO 0.75	WFO 0.75	WFO 0.75	GINI	(EMPATE)
WFO 1	GINI	BNS	CHI2	WFO 0.75
GINI	GM	GR	WFO 0.75	GM
IG	BNS	GINI	BNS	CHI2
CHI2	CHI2	GM	GM	BNS
GR	GR	CHI2	GR	IG
WLLR	WLLR	IG	IG	GR
GM	CDM	CDM	WLLR	WLLR
WFO 0.25	IG	WLLR	MOR	WFO 0.25
MOR	MOR	MOR	CMD	WFO 0
CMD	WFO 0.25	WFO 0.25	WFO 0.25	MOR
PR	PR	PR	PR	PR
WFO 0	WFO 0	WFO 0	WFO 0	CMD
MI	MI	MI	MI	MI

Fonte: O autor.

A Tabela 3 mostra a média (de Macro-F1 e Micro-F1) das 12 FEFs usando como classificador a *Support Vector Machine* (SVM) e *Random Forest* (RF).

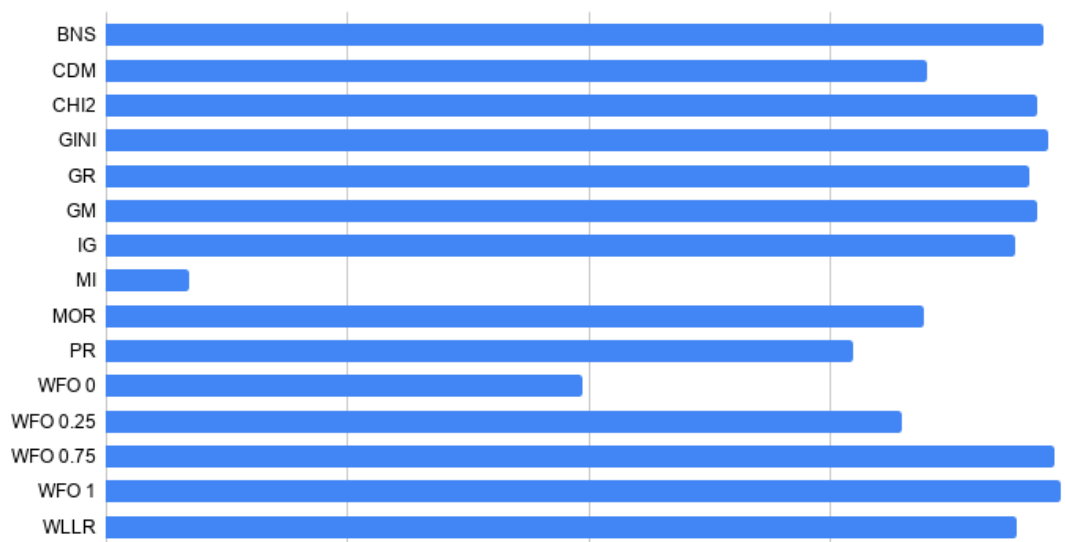
Tabela 3 - Média geral do desempenho da classificação para cada FEF e, na última linha, resultado sem seleção de características

<i>FEF</i>	<i>MACRO-F1 (SVM)</i>	<i>MICRO-F1 (SVM)</i>	<i>MACRO-F1 (RF)</i>	<i>MICRO-F1 (RF)</i>
BNS	0,8703 (0,0709)	0,8720 (0,0713)	0,7738 (0,1392)	0,7941 (0,1229)
CDM	0,8896 (0,0760)	0,8900 (0,0750)	0,7646 (0,0963)	0,7916 (0,0793)
CHI2	0,8856 (0,0884)	0,8864 (0,0885)	0,7751 (0,1295)	0,7944 (0,1063)
GINI	0,8160 (0,1627)	0,8168 (0,1604)	0,7289 (0,1390)	0,7511 (0,1295)

GR	0,9075 (0,0608)	0,9089 (0,0612)	0,8007 (0,1346)	0,8202 (0,1123)
GM	0,7296 (0,0820)	0,7357 (0,0821)	0,6217 (0,0797)	0,6688 (0,0780)
IG	0,9086 (0,0709)	0,9098 (0,0711)	0,7962 (0,1431)	0,8173 (0,1168)
MI	0,6311 (0,0572)	0,6480 (0,0513)	0,3377 (0,2272)	0,4678 (0,1248)
MOR	0,8987 (0,0810)	0,8992 (0,0808)	0,8052 (0,1209)	0,8192 (0,1048)
PR	0,8946 (0,0783)	0,8949 (0,0787)	0,7576 (0,1307)	0,7822 (0,1105)
WFO 0	0,7944 (0,0972)	0,7948 (0,0995)	0,5139 (0,2066)	0,5856 (0,1292)
WFO 0.25	0,8983 (0,0699)	0,8990 (0,0701)	0,7686 (0,1314)	0,7846 (0,1160)
WFO 0.75	0,7124 (0,2036)	0,7314 (0,1782)	0,6313 (0,2833)	0,6838 (0,2137)
WFO 1	0,5002 (0,2790)	0,5596 (0,2268)	0,4607 (0,3017)	0,5395 (0,2366)
WLLR	0,9056 (0,0613)	0,9063 (0,0611)	0,8013 (0,1234)	0,8179 (0,1073)
sem FEF	0,9440 (0,0633)	0,9446 (0,0633)	0,6162 (0,2944)	0,6938 (0,2061)

Fonte: O autor.

Figura 2 - Percentual médio da efetividade na remoção de características de cada FEF



Fonte: O autor.

A partir de uma análise individual de cada FEF (Tabela 2 e 3) foi observado que a FEF MI produz os piores resultados, evidenciados na Figura 2. Com relação à FEF WFO utilizando como parâmetros (λ) os valores 0, 0.25, 0.75 e 1 possui resultados instáveis na maioria das bases de dados. Dentre os parâmetros (λ) usados na FEF WFO a que apresenta resultados mais eficientes é a de 0.25. Ainda foi verificado que a FEF GM apesar de reduzir de maneira efetiva a quantidade de termos da base de dados não teve uma boa performance. E, por fim, a FEF PR apesar de produzir de resultados efetivos (Figura 2), várias outras FEFs conseguiram apresentar melhores resultados. Por fim, no espectro das FEFs de baixo desempenho, temos as FEFs CDM e MOR que não realizam uma redução

de termos adequada comparada com as FEFs remanescentes (Figura 2).

Considerando tais observações, 6 FEFs mais robustas: CHI, BNS, GR, GINI, IG E WLLR. Dentre essas, a de melhor desempenho para as bases de dados estudadas e que produz resultados mais estáveis é a FEF GR. De modo geral, apesar da FEF GR selecionar um número maior de termos que as FEFs BNS, CHI e GINI ela produz resultados mais estáveis e superiores. Comparativamente, GR seleciona um número menor de termos do que a WLLR. Em relação à FEF IG, GR não apresenta uma grande disparidade de valores, pelo fato da FEF GR ser um refinamento da FEF IG. Analisando os experimentos realizados é possível observar que em apenas dois experimentos a FEF IG selecionou menos termos, entretanto a FEF GR apresentou em uma das bases uma seleção de características mais precisa tendo diferença de selecionando 58,8 características a menos que a FEF IG. Apesar de os resultados obtidos por ambas as FEFs podem ser consideradas próximas (com exceção do caso explanado anteriormente) é possível notar uma tendência maior a FEF GR apresentar resultados de maior qualidade na classificação.

4 CONSIDERAÇÕES FINAIS

Neste artigo, foram avaliadas 12 FEFs em conjunto com a metodologia de seleção de características ALOFT. Esse estudo foi realizado com o intuito de indicar FEFs que consigam reduzir de maneira efetiva o número de características usadas para classificação de documentos, de modo que a qualidade da classificação seja mantida.

A partir dos experimentos realizados é indicado que para estudos futuros abordando o uso ou aplicações de um projeto comercial utilizando o método ALOFT é recomendado usar as FEFs: BNS, CHI, GINI, GR, IG e WLLR, com recomendação especial para a FEF GR por possuir uma maior estabilidade nos seus resultados de classificação e redução efetiva na quantidade de termos.

O presente artigo limitou-se à avaliação de base de dados menores comparadas às estudadas por Pinheiro *et al.* (2012), isto é, com menos documentos e características, pela complexidade de lidar com grandes quantidades de termos nos documentos que proporciona um aumento no tempo para realizar a classificação de documentos.

Para Trabalhos futuros é necessário comparar novamente as FEFs BNS, CHI, GINI, GR, IG e WLLR em um maior número de base de dados e de categorias diferentes, com o intuito de avaliar minuciosamente os prós e contras de cada uma. Além disto, é necessário

reproduzir o experimento para bases de dados com maior número de documentos e termos (em comparação as bases usadas neste artigo), para analisar se a qualidade das FEFs BNS, CHI, GINI, GR, IG e WLLR se mantêm independente do volume de dados.

Também pode ser interessante a realização de um estudo que busca encontrar um valor para o parâmetro λ que estabilize o uso da FEF WFO. Visando uma aplicação mais estável desta FEF em conjunto com o método ALOFT.

AGRADECIMENTOS

Agradecemos à PRPI da Universidade Federal do Cariri (UFCA) e à própria UFCA por promover e viabilizar o estudo para a realização deste artigo.

REFERÊNCIAS

BRANK, J. *et al.* Interaction of feature selection methods and linear classification models. *In: Proceedings of the ICML-02 Workshop on Text Learning*, 19., 2002, Sydney, Australia. **International Workshop on Text Learning, in Conjunction with International Conference on Machine Learning**. [S.l.]: Forthcoming, 2002. p. 1-6.

BREIMAN, L. Random forests. **Machine learning**, Boston, v. 45, n. 1, p. 5–32, 2001.

CAROPRESO, M. F.; MATWIN, S.; SEBASTIANI, F. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *In: CHIN, A. G. (ed.). Text Databases and Document Management: Theory and Practice*. Hershey, PA, USA: IGI Publishing, 2001. p. 78-102.

CESKA, Z; FOX, C. The Influence of Text Pre-processing on Plagiarism Detection. *In: International Conference RANLP-2009*, 2009, Borovets, Bulgaria. **International Conference Recent Advances In Natural Language Processing**. USA: Association for Computational Linguistics, 2009. p. 55-59.

CHEN, J. *et al.* Feature selection for text classification with naive bayes. **Expert Systems with Applications**, [Rockville], v. 36, n. 3, p. 5432-5435, 2009.

FORMAN, G. An extensive empirical study of feature selection metrics for text classification. **Journal of Machine Learning Research**, [Cambridge], v. 3, Special Issue on Machine Learning Methods for Text and Images, p. 1289-1305, 2003.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. *In: Proceedings of ECML'98*, 10., 1998, Chemnitz, Germany. **Proceedings of the 10th European Conference on Machine Learning**. Berlin: Springer-Verlag, 1998. p. 137-142.

JOACHIMS, T. Transductive Inference for Text Classification Using Support Vector Machines. *In: Proceedings of ICML'99*, 16., 1999, Bled, Slovenia. **Proceedings of the Sixteenth International Conference on Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1999. p. 200-209.

IKONOMAKIS, M.; KOTSIANTIS, S.; TAMPAKAS, V. Text classification using machine learning techniques. **WSEAS TRANSACTIONS on COMPUTERS**, [S. l.], v. 4, n. 8, p. 966-974, 2005.

KUBAT, M.; HOLTE, R. C.; MATWIN, S. Machine learning for the detection of oil spills in satellite radar images. **Machine Learning**, Boston, v. 30, n. 2, p. 195–215, 1998.

KOBAYASHI, V. B. *et al.* Text Classification for Organizational Researchers: A Tutorial. **Organizational Research Methods**, [S. l.], v. 21, n. 3, p. 766–799, 2018.

LI, S. *et al.* A Framework of Feature Selection Methods for Text Categorization. *In: Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, 47., 2009, Suntec, Singapore. **Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP**. USA: Association for Computational Linguistics, 2009. v. 2, p. 692-700.

PINHEIRO, R. H. W. *et al.* A global-ranking local feature selection method for text categorization. **Expert Systems with Applications**, [Rockville], v. 39, n. 17, p. 12851-12857, 2012.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, Boston, v. 1, n. 1, p. 81-106, 1986.

RAMOS, J. **Using tf-idf to determine word relevance in document queries**. Piscataway, NJ, USA: Department of Computer Science, Rutgers University, 2003

SHANG, W. *et al.* A novel feature selection algorithm for text categorization. **Expert Systems with Applications**, [Rockville], v. 33, n. 1, p. 1-5, 2007.

SU, J.; SAYYAD-SHIRABAD, J.; MATWIN, S. Large scale text classification using semi-supervised multinomial naive bayes. *In: Proceedings of ICML'11*, 28., 2011, Bellevue, Washington, USA. **Proceedings of the 28th International Conference on International Conference on Machine Learning**. Madison, WI, USA: Omnipress, 2011. p. 97-104.

YAN J. Text Representation. *In: LIU L.; ÖZSU M.T. (eds) Encyclopedia of Database Systems*. Boston, MA, USA: Springer, 2009. p. 3069-3072.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. *In: D. H. FISHER (ed.). Proceedings of ICML'97*, 14., 1997, Nashville, Tennessee, USA. **Proceedings of the 14th International Conference on Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1997. p. 412-420.