Avaliação estatística sobre seleção de características para categorização de texto

Rogério C. P. Fragoso*, Lucas F. Melo*, Saulo C. R. P. Sobrinho*
Universidade Federal de Pernambuco (UFPE), Centro de Informática (CIn)
Av. Jornalista Anibal Fernandes s/n, Cidade Universitária 50740-560, Recife, PE, Brazil
rcpf@cin.ufpe.br, lfm2@cin.ufpe.br, scrps@cin.ufpe.br

***Abstract*—ESCREVER RESUMO**

I. Introdução

Algoritmos são sequências de instruções bem definidas, porém suas implementações podem apresentar comportamentos difíceis de serem previstos. Seja pelo uso de geração de números pseudo-aleatórios que geram um comportamento não determinístico inerente ao código, ou pelo uso de linguagens de alto nível cuja tradução para linguagem de máquina passa por otimizações e diferentes interações com a arquitetura de destino na qual o algoritmo é executado. Sendo assim, a performance de algoritmos de computação em geral é não-determinística e, ao analisar comparativamente o desempenho destes algoritmos, é necessário levar em consideração que estamos diante de uma amostra aleatória que representa estas performances.

Este trabalho realiza uma comparação entre algoritmos de categorização de texto (mais especificamente sobre a etapa de extração de características), mostrando como determinar e aplicar testes estatísticos adequados para este cenário.

A. Fundamentação teórica

Na abordagem de aprendizagem de máquina, uma instância é representada como um vetor composto por pares de característica e valor. Uma abordagem comum para a representação de textos na forma de vetores de características é a técnica conhecida como *Bag of Words* (BoW) [1]. Nela, um texto é tratado como um conjunto de palavras, sem considerar gramática ou ordem de ocorrência das palavras no texto. Cada palavra do vocabulário da base de dados é considerada uma característica e é associada à frequência desta palavra no documento. Ou seja, o tamanho do vocabulário da base de dados define a dimensionalidade dos vetores. Desta forma, em uma base de dados de tamanho médio, é comum que os vetores de características contenham dezenas de milhares de dimensões [2]. Entretanto, a maior parte destas características é irrelevante ou redundante. A alta dimensionalidade pode tornar a categorização de textos muito dispendiosa em termos de memória e tempo de execução. Adicionalmente, este grande número de características pode impactar negativamente no desempenho de classificação, especialmente em bases de dados com um número pequeno de instâncias em relação ao número de características, fenômeno conhecido como "praga da dimensionalidade". Como muitas das características são irrelevantes para a categorização, estes problemas podem ser tratados através da restrição da quantidade de características do conjunto de dados. Esta abordagem é conhecida como Redução de Dimensionalidade (DR, do inglês *Dimensionality Reduction*).

Uma técnica de DR muito utilizada é a seleção de características. Nesta abordagem, o conjunto final é formado por parte das características do conjunto original. A utilização de métodos de filtragem é a técnica de FS considerada mais adequada para problemas de TC, devido ao custo computacional ser bem mais baixo que o de outras técnicas, como métodos *wrapper*. Métodos de filtragem realizam um ordenamento das características através do uso de algoritmos determinísticos e métricas estatísticas, conhecidas com funções de avaliação de características (FEF, do inglês *Feature Evaluation Function*). Após o ordenamento, uma quantidade, estabelecida pelo usuário, de características é selecionada para a formação do novo subconjunto.

B. Métodos de seleção de características

Maximum f features per Document (MFD) [3], *Maximum f features per Document-Reduced* (MFDR) [3] e *Class-dependent Maximum f features per Document-Reduced* (cMFDR) [4] são métodos de seleção de características para categorização de texto. Estes métodos requerem um valor para o parâmetro f, que indica a quantidade de características a serem selecionadas por instância. A escolha de um bom valor para este parâmetro pode ser um trabalho demorado e exaustivo. Neste contexto, o método *Automatic Feature Subsets Analyzer* (ASFA) [5] foi proposto para ser usado conjuntamente com um dos três métodos: MFD, MFDR ou cMFDR. O objetivo do AFSA é prover, para estes métodos, um valor para o parâmetro f de forma automática.

Esta seção apresentou conceitos básicos de categorização de textos e seleção de características e introduziu os métodos de seleção de características que são avaliados no trabalho. O restante do trabalho é organizado como segue: A Seção II apresenta o objetivo do presente trabalho. Na Seção III são detalhadas as configurações dos experimentos, incluindo descrição da base de dados, os algoritmos de interesse e as hipóteses a serem verificadas sobre os dados. A Seção IV demonstra os procedimentos estatísticos realizados no trabalho. Finalmente, a Seção V apresenta as conclusões do trabalho.

II. Objetivo

O objetivo desta pesquisa é verificar se o método AFSA é capaz de prover uma configuração para MFD, MFDR e cMFDR de modo que o desempenho de classificação não seja prejudicado. Ou seja, desejamos averiguar se o desempenho dos métodos MFD, MFDR e cMFDR é alterado quando estes métodos são usados conjuntamente com AFSA. Esta análise comparativa visa determinar se os algoritmos possuem desempenhos significativamente diferentes e, em caso positivo, determinar qual apresenta desempenho superior.

III. Experimentos

Esta seção descreve as configurações dos experimentos realizados para gerar o conjunto de dados sobre o qual a análise será realizada.

A. Base de dados

Para a categorização de texto, foi utilizada a base de dados *Reuters 10*. Esta base de dados é um subconjunto da coleção *Reuters-21578* [1], que é uma das bases mais utilizadas em trabalhos de categorização de texto. A base é composta por documentos coletados do *Reuters newswire* de 1987 e apresenta 135 categorias. Entretanto, o subconjunto adotado neste trabalho é composto pelas 10 maiores categorias da base. O base de dados *Reuters 10* contém 9.980 documentos e seu vocabulário abarca 10.987 termos. A base de dados *Reuters 10* também é muito utilizada em trabalhos de categorização de texto [6]–[8].

A distribuição dos documentos é bastante desbalanceada, apresentando categorias representando desde 2,3% até 39% do tamanho total da base. Nesta base foram aplicados os seguintes procedimentos de pré-processamento: *stemming*, com o algoritmo *Iterated Lovins Stemmer* [9], remoção de termos com duas ou menos letras e remoção de *stopwords*.

Vale salientar que a análise comparativa deste trabalho é realizada sobre o desempenho dos algoritmos, tendo a base citada como entrada, e não sobre características da base em si. O processo de geração das amostras utilizadas na análise estatística é detalhado na Seção III-B.

B. Metodologia

Conforme mencionado na Seção II, esta pesquisa visa realizar uma comparação de desempenho de algoritmos de seleção de características para categorização de texto. Neste trabalho, é feita uma avaliação do desempenho do método AFSA. Para tanto, AFSA é usado em conjunto com cada um dos métodos MFD, MFDR e cMFDR (conforme descrito na Seção I-B. O desempenho de cada um dos métodos configurado manualmente é comparado com o desempenho destes métodos configurados automaticamente por AFSA.

O desempenho de um método de seleção de características pode ser auferido em termos de redução de dimensionalidade (tamanho do vetor final de características), tempo de execução e do desempenho de classificação atingido com o vetor de características resultante do processo de seleção. Neste trabalho, a avaliação dos desempenhos dos métodos levou em conta o desempenho de classificação sobre a base de dados resultante do processo de seleção de características. O algoritmo de aprendizagem de máquina empregado para a avaliação dos métodos foi classificador *Naïve Bayes Multinomial* [10].

A base de dados *Reuters 10* foi pré-processada utilizando cada um dos seis algoritmos de seleção de carcaterísticas (MFD, MFDR cMFDR e a combinação de cada um destes com AFSA), gerando, assim, seis versões da base original. Em seguida, o classificador *Naïve Bayes Multinomial* foi treinado e testado com cada uma destas seis versões.

A validação cruzada estratificada foi utilizada como método para estimativa de desempenho. Esta técnica é adotada para avaliar a capacidade de generalização de um modelo a partir de um conjunto de dados. Neste trabalho utilizou-se a variação validação cruzada estratificada com *10 folds*, na qual a base de dados $\mathcal{D}$ é particionada em 10 subconjuntos (*folds*), de tamanhos semelhantes, mantendo a proporção de documentos por categorias equivalente à proporção encontrada no conjunto original. Então, são construídos 10 classificadores, cada um utilizando uma parcela dos *folds* para treinamento e outra parcela para realizar o teste do mesmo, de modo a gerar diferentes combinações dos *folds* [11].

Nos experimentos realizados com os métodos MFD, MFDR e cMFDR, nove partições foram utilizadas para treinamento e uma partição foi utilizada para teste. O método AFSA requer uma porção dos dados para configuração de seus parâmetreos. Assim, os experimentos executados com AFSA utilizaram oito partições para treinamento, uma para configuração de parâmetros/validação e uma para teste. Deste modo, ao final deste processo, temos dez medidas de desempenho para cada um dos seis métodos de seleção de características avaliados. Estes dados de desempenho correspondem às amostras que são entradas para as análises estatísticas realizadas neste trabalho.

A medida de desempenho utilizada nos experimentos foi *Micro-F1*. Seu cálculo é dado pela Eq. 1.

$$\mathcal{F}1 = \frac{2\text{x}\mathcal{P}\text{x}\mathcal{R}}{\mathcal{P}+\mathcal{R}}, \tag{1}$$

onde $\mathcal{P}$ é uma medida de precisão e $\mathcal{R}$ é uma medida de cobertura [6]. As fórmulas para calcular a precisão $\mathcal{P}$ e a cobertura $\mathcal{R}$ são exibidas a seguir.

$$\mathcal{P} = \frac{\sum_{j=1}^{C} TP_j}{\sum_{j=1}^{C}(TP_j + FP_j)} \tag{2}$$

$$\mathcal{R} = \frac{\sum_{j=1}^{C} TP_j}{\sum_{j=1}^{C}(TP_j + FN_j)} \tag{3}$$

TP_j é a quantidade de instâncias corretamente rotuladas como pertencentes à categoria c_j; FP_j é a quantidade de instâncias incorretamente rotuladas como pertencentes à categoria c_j; e FN_j é a quantidade de instâncias incorretamente rotuladas como não pertencentes à categoria c_j.

[1] Disponível em http://disi.unitn.it/moschitti/corpora.htm.

IV. Análise estatística

A. *Estatística descritiva*

Uma boa prática ao iniciar uma análise de conjunto de dados, a qual é sugerida por muitos autores, é o uso de técnicas de estatística descritiva para se obter intuições iniciais acerca do conjunto de interesse [12].

Para se ter uma indicação sobre os tipos de testes de hipótese que podem ser executados sobre os dados, é interessante verificar se as distribuições que geram as amostras apresentam normalidade. A suposição de normalidade é útil pois, se esta for plausível, podemos aplicar testes paramétricos sobre os dados. Testes paramétricos possuem maior poder estatístico do que seus equivalentes não-paramétricos, o que nos permite extrair conclusões mais fortes.

Para verificar a normalidade, começamos utilizando recursos visuais. Histogramas permitem visualizar a disposição das amostras e consequentemente intuir sobre a distribuição da população geradora. As Figuras 6 a 1 apresentam os histogramas com os desempenhos de classificação de cada algoritmo nas suas versões combinadas com o ASFA e sem o mesmo.

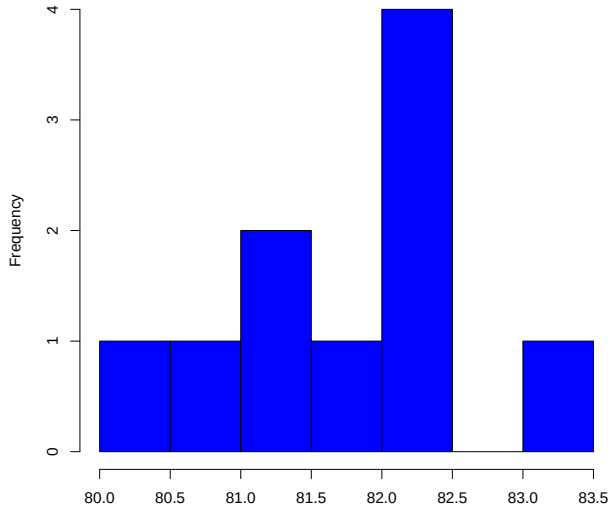

Fig. 1. Histograma do desempenho de classificação do MFD (10 amostras).

Observando os histogramas, não observamos indícios de normalidade em nenhuma das amostras. Para juntar mais evidências, continuaremos a análise.

Outra ferramenta visual útil é o plot de probabilidade normal ou *normplots*, que permite verificar o quão bem os dados representam uma distribuição normal. A escala do gráfico é elaborada para que dados provenientes de uma distribuição normal ideal formem um reta que é usada como referência na figura. Num cenário real, devido à presença de ruídos, não se espera que os dados se adequem perfeitamente à reta, porém se espera que, se os dados forem provenientes de uma distribuição normal, estes sejam bem aproximados

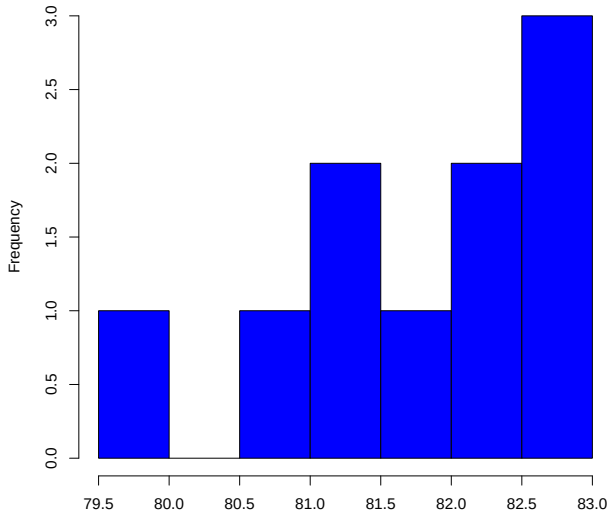

Fig. 2. Histograma do desempenho de classificação do MFD com AFSA (10 amostras).

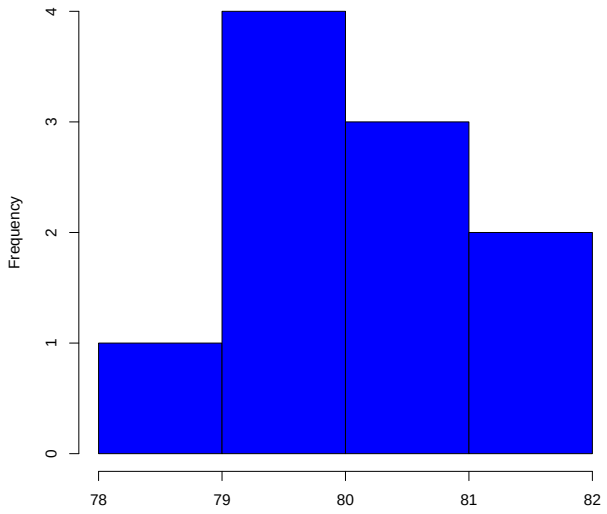

Fig. 3. Histograma do desempenho de classificação do MFDR (10 amostras).

pela reta de referência e estejam próximos a ela. As Figuras 7 a 12 mostram os normplots de cada distribuição.

Podemos observar que o fitting em geral não foi tão próximo quanto se poderia esperar. Além disso, observa-se que nos valores mais extremos da distribuição, a aproximação degrada ainda mais (salvo pelo exemplo na Figura 10). Isso pode indicar que estes pontos talvez sejam outliers, ou, o que consideramos mais provável, que as caudas da distribuição não são parecidas com a de uma

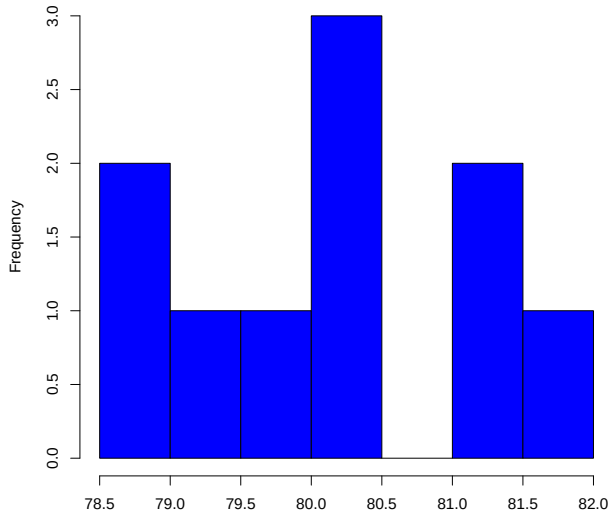

Fig. 4. Histograma do desempenho de classificação do MFDR com AFSA (10 amostras).

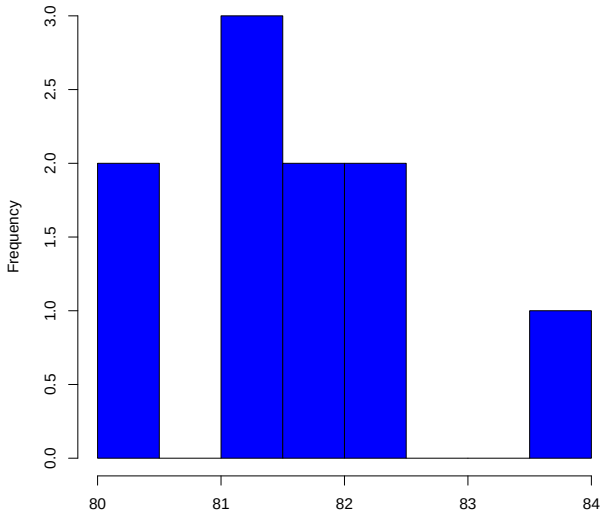

Fig. 6. Histograma do desempenho de classificação do cMFDR (10 amostras).

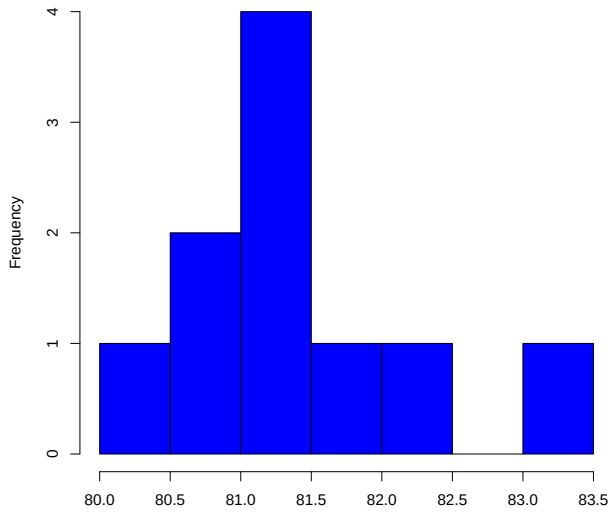

Fig. 5. Histograma do desempenho de classificação do cMFDR (10 amostras).

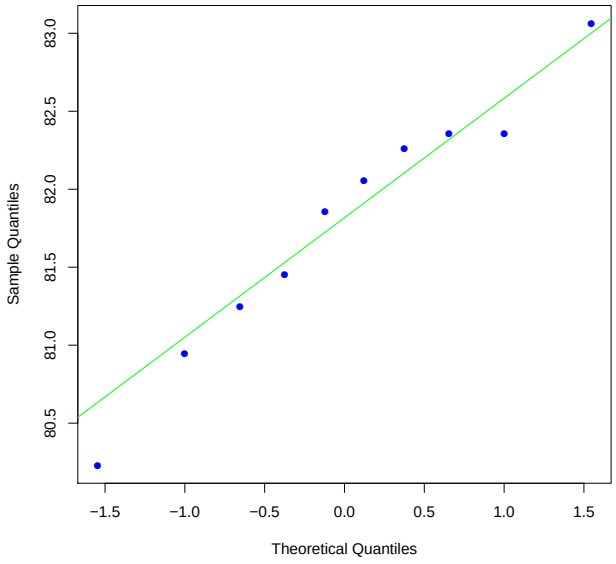

Fig. 7. Normplot desempenho de classificação do MFD (10 amostras).

normal.

Não vejo necessidade de apresentar este aqui. Acho que tem muita informação já. Opinem Ainda recorremos aos diagramas de caixa (Figura 13) como um último recurso visual. Nestes podemos verificar sobretudo a simetria das amostras.

Observando os diagramas, vemos que quase todos são assimétricos, alguns inclusive apresentando assimetria acentuada.

As amostras pequenas se mostraram um fator limitante da análise...

Além dos recursos gráficos, podemos analisar diretamente os sumários numéricos dos dados. Ao comparar médias e medianas, podemos ter uma idéia da simetria do conjunto em questão. A Tabela **??** mostra os valores de média, mediana e devio padrão de cada conjunto. O problema é que esse desvio padrão tá meio solto aí. Não vai ser referenciado pra nada.. Acho que ou apresenta essa tabela, ou o boxplot, talvez até nenhum dos dois. Comentem.

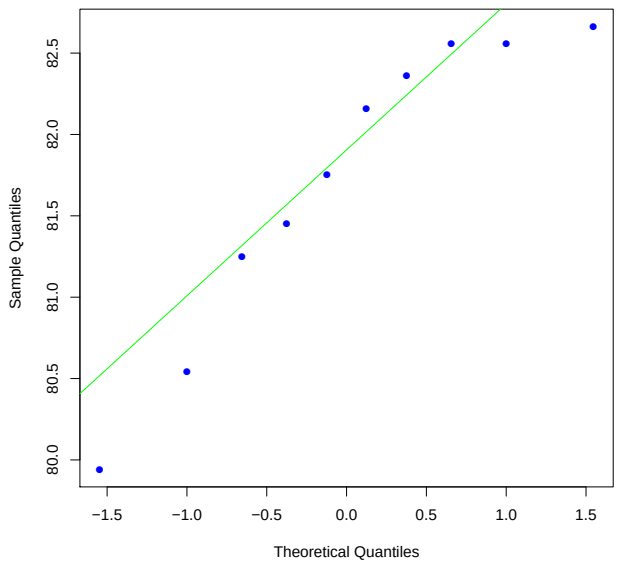

Fig. 8. Normplot desempenho de classificação do MFD com AFSA (10 amostras).

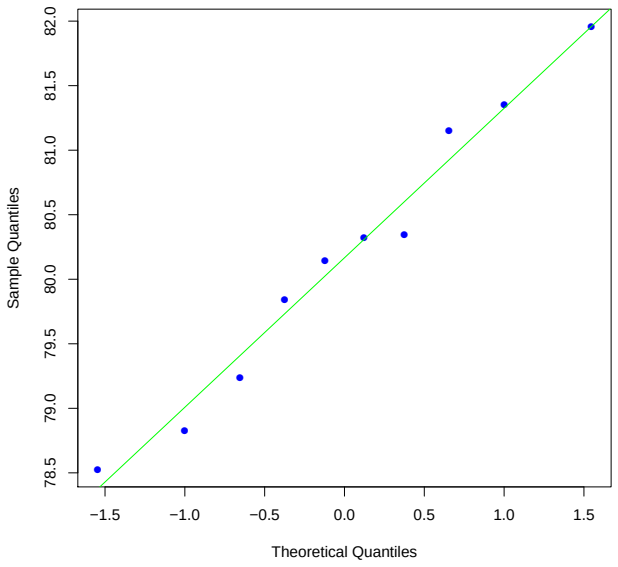

Fig. 10. Normplot desempenho de classificação do MFDR com AFSA (10 amostras).

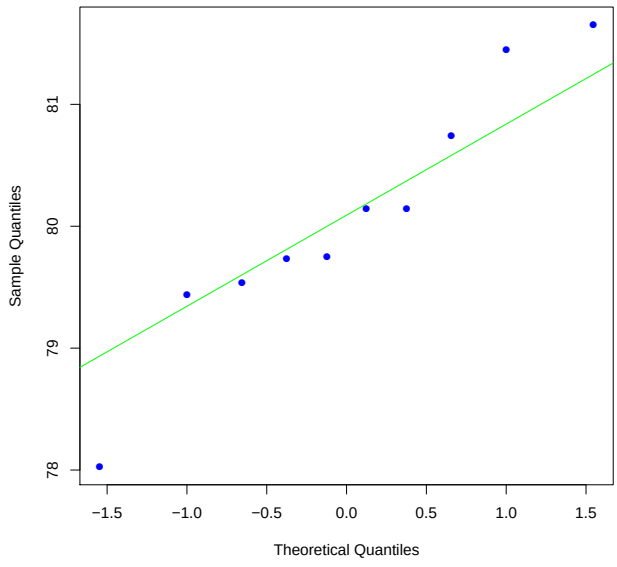

Fig. 9. Normplot desempenho de classificação do MFDR (10 amostras).

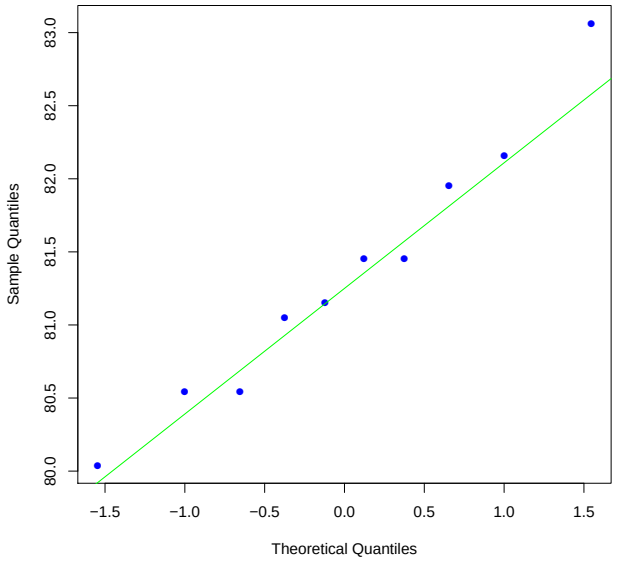

Fig. 11. Normplot desempenho de classificação do cMFDR (10 amostras).

FALAR SOBRE OS RESULTADOS Quais aparentam ser normais (média aprox. igual à mediana)

B. *Testes de aderência*

Uma maneira mais formal de verificar se as amostras seguem uma distribuição normal é com o uso de testes de aderência. Nesta pesquisa, foram executados os testes Shapiro-Wilk [13], Anderson-Darling [14] e Cramer-von Mises [15]. Os resultados dos testes para cada amostra são apresentados na tabela II.

Consideramos que uma amostra não segue uma distribuição normal caso o *p-value* seja menor que 0.05. Deste modo, o resultado dos três testes apontam que, para todas as amostras, não existem evidências suficientes para rejeitar a hipótese nula de que a amostra segue uma distribuição normal, reforçando a impressão obtida da análise dos valores das médias e medianas das amostras IV-A. Porém é importante salientar que a

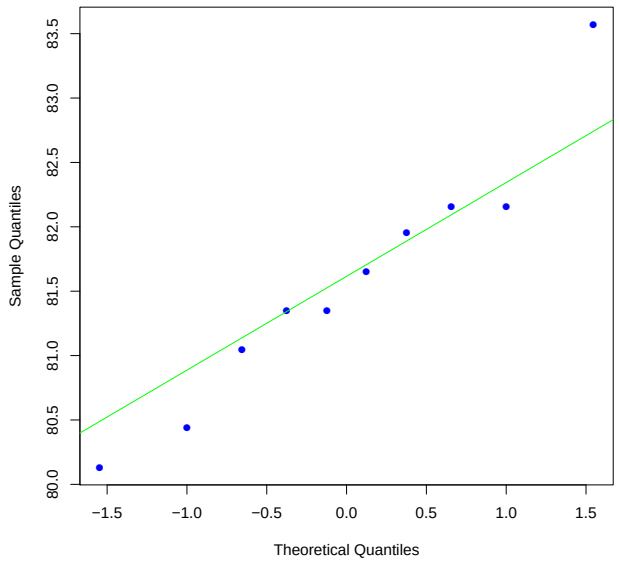

Fig. 12. Normplot desempenho de classificação do cMFDR com AFSA (10 amostras).

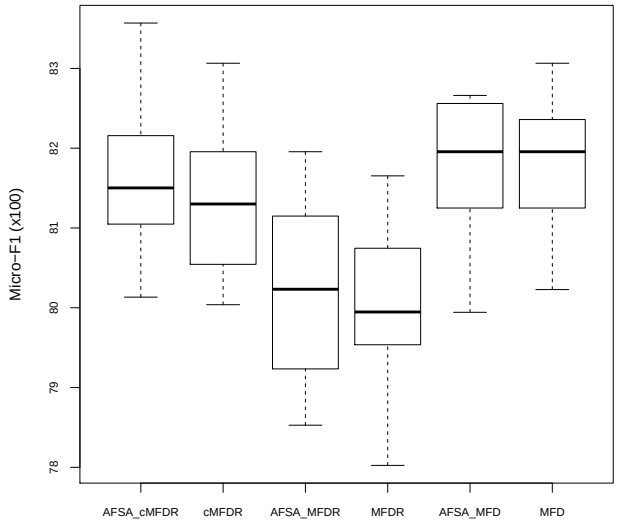

Fig. 13. Escrever uma descrição.

eficiências de testes de aderência (não paramétricos) é muito prejudicada quando as amostras são pequenas, como no nosso caso. Adicionalmente, a análise dos histogramas, normplots e dos boxplots (da maioria das amostras) dão indícios de que as amostras não seguem uma distribuição normal. Por estas razões, concluímos que, com os dados disponíveis, não temos segurança em afirmar que as amostras seguem uma distribuição normal. Diante disto, utilizaremos teste de hipótese não-paramétrico para verificar a existência de diferenças significativas entre as amostras.

TABLE I. SUMÁRIOS NUMÉRICOS DOS DADOS

Método	Média	Mediana	Desv. Padrão
AFSA cMFDR	81.58	81.50	0.97
cMFDR	81.34	81.30	0.88
AFSA MFDR	80.16	80.23	1.10
MFDR	80.06	79.94	1.05
AFSA MFD	81.72	81.95	0.92
MFD	81.78	81.95	0.82

TABLE II. RESULTADOS DOS TESTES DE ADERÊNCIA.

	p-value		
	Shapiro-Wilk	Anderson-Darling	Cramer-von Mises
AFSA cMFDR	0.75	0.67	0.74
cMFDR	0.90	0.86	0.86
AFSA MFDR	0.88	0.89	0.87
MFDR	0.61	0.47	0.45
AFSA MFD	0.19	0.25	0.30
MFD	0.88	0.77	0.72

C. *Resultados*

Depois de avaliar as distribuições das amostras, optamos pelo caminho mais seguro que consiste na comparação dos pares de métodos por testes não paramétricos. Como as comparações a serem executadas são entre pares de algoritmos, o teste a ser utilizado é o teste de sinais com postos de Wilcoxon [16].

Dado que temos a média amostral já computada para cada conjunto, optamos por um teste unilateral pois este tende a fornecer mais informação. As hipóteses são estabelecidas de acordo com a média amostral obtida. A hipótese nula é que as médias dos métodos sendo comparados (com e sem AFSA) são iguais, a hipótese alternativa é...

Os resultados dos testes são apresentados na Tabela III

TABLE III. RESULTADOS DOS TESTES DE ADERÊNCIA.

Hypothesis	p-value
H_1 : AFSA MFD $<$ MFD	0.22
H_1 : MFDR $<$ AFSA MFDR	0.27
H_1 : cMFDR $<$ AFSA cMFDR	0.039

V. CONCLUSÃO

Completar. Percebemos que para que a análise seja mais completa, se faz necessário uma amostra maior. Vale salientar que embora o resultado do teste seja inconclusivo para as duas amostras, ele fortalece a intuição de que não há diferenças significativas de fato. Completar.

REFERENCES

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[2] E. Gabrilovich and S. Markovitch, "Text categorization with many redundant features: using aggressive feature selection to make svms competitive with c4. 5," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 41.

[3] R. Pinheiro, G. Cavalcanti, and T. Ren, "Data-driven global-ranking local feature selection methods for text categorization," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1941–1949, 2015.

[4] R. Fragoso, R. Pinheiro, and G. Cavalcanti, "Class-dependent feature selection algorithm for text categorization," in *International Joint Conference on Neural Networks*, 2016.

[5] ——, "A method for automatic determination of the feature vector size for text categorization," in *2016 Brazilian Conference on Intelligent Systems*, 2016.

[6] Y. Chang, S. Chen, and C. Liau, "Multilabel text categorization based on a new linear classifier learning method and a category-sensitive refinement method," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1948–1953, 2008.

[7] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naive Bayes," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5432–5435, 2009.

[8] J. Yang, Y. Liu, Z. Liu, X. Zhu, and X. Zhang, "A new feature selection algorithm based on binomial hypothesis testing for spam filtering," *Knowledge-Based Systems*, vol. 24, no. 6, pp. 904–914, 2011.

[9] J. B. Lovins, *Development of a stemming algorithm*. MIT Information Processing Group, Electronic Systems Laboratory Cambridge, 1968.

[10] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Madison, WI, 1998, pp. 41–48.

[11] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Stanford, CA, 1995, pp. 1137–1145.

[12] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*. John Wiley & Sons, 2010.

[13] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[14] T. W. Anderson and D. A. Darling, "A test of goodness of fit," *Journal of the American statistical association*, vol. 49, no. 268, pp. 765–769, 1954.

[15] J. Durbin and M. Knott, "Components of cramer-von mises statistics. i," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 290–307, 1972.

[16] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.