# Avaliando contribuições na substituição de termos informais em classificação de texto de redes sociais com NetSpeak-BR

**Rodolpho da Silva Nascimento**[1], **Gabriel dos Santos**[1], **Flavio Carvalho**[1], **Gustavo Guedes**[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

{rodolpho.nascimento, gabriel.santos, flavio.carvalho}@eic.cefet-rj.br,
gustavo.guedes@cefet-rj.br

***Abstract.*** *NetSpeak (NS) is a set of informal words characterized by creative variation in writing, such as using abbreviations, phonetic spelling and other neologisms. NS terms are very common in written communications on online social networking platforms, thus being treated in pre-processing tasks to obtain better results in classification tasks. This work analyses improving data normalization considering the exchange of informal terms (NS) to formal terms, using the NetSpeak-BR lexicon. Although the results indicated a performance increase using different machine learning classification algorithms, this improvement was not consistent for different social networks.*

***Resumo.*** *NetSpeak (NS) é um conjunto de palavras informais caracterizadas por uma variação criativa na escrita, com o uso abreviaturas, grafia fonética e outros neologismos. Os termos NS são muito comuns em comunicações escritas em plataformas de redes sociais online, sendo assim tratados em tarefas de pré-processamento para obter melhores resultados nas tarefas de classificação. Este trabalho analisa a contribuição da normalização de dados considerando a troca de termos informais (NS) por termos formais, utilizando o léxico NetSpeak-BR. Embora os resultados tenham indicado um aumento de desempenho usando diferentes algoritmos de classificação de aprendizado de máquina, essa melhoria não foi consistente para diferentes redes sociais.*

## 1. Introdução

Atualmente, os aplicativos e plataformas *online* de Redes Sociais (RS) se destacam no processo de comunicação entre indivíduos. Observa-se na utilização das RS, em grande parte mediada por *smartphones*, uma proporção crescente de uma comunicação escrita informal. Esta comunicação é caracterizada pela variação criativa utilizando *emoticons*, abreviações, ortografia fonética e outros neologismos, gerando uma neografia [Danet and Herring, 2007] que por muitos é conhecida como NetSpeak (NS).

A utilização do NS é disseminada na Internet porque facilita a comunicação *online* escrita [Liu and Liu, 2014]. Em trabalhos de classificação de textos, especialmente na área de Análise de Sentimento, os termos do NS podem sofrer pré-processamento, visando obter melhores resultados [Oyong et al., 2018]. A relevância da utilização destes termos vem estimulando o desenvolvimento de léxicos de NS com o propósito de servir como apoio em tarefas de pré-processamento [Nascimento et al., 2019].

Em consulta à literatura da área, não foram encontrados trabalhos comparando resultados de classificadores mantendo os termos NS ou substituindo por termos formais da língua portuguesa, checando a relevância do tratamento de NS na etapa de pré-processamento. Assim, o objetivo deste trabalho é responder a seguinte pergunta: é possível observar variação significativa nos resultados de diferentes classificadores, substituindo, na etapa de pré-processamento, os termos informais (NS) por termos formais, em tarefas de classificação de documentos extraídos de redes sociais?

O presente trabalho discute os trabalhos relacionados na Seção 2. Na Seção 3, é apresentada a metodologia de desenvolvimento. Os resultados dos experimentos são discutidos na Seção 4. Por fim, a Seção 5 apresenta considerações do presente trabalho e recomendações para continuidade de trabalhos nesse tema.

## 2. Trabalhos Relacionados

Um léxico de NS foi utilizado como apoio no pré-processamento dos dados no estudo proposto por Mamgain et al. [2016]. O objetivo do estudo consistia em utilizar comentários da RS Twitter para avaliar instituições de ensino, observando os pontos negativos e também os aspectos considerados satisfatórios, buscando assim obter uma referência para auxiliar a tomada de decisão de futuros estudantes. O léxico utilizado, em inglês, foi composto por aproximadamente $5.200$ termos informais e as suas respectivas formalizações (e.g. *"asap"="as soon as possible"; "b4"="before"; "brb"="be right back"*).

O trabalho proposto por Aleksić-Maslać et al. [2019] coletou informações de sistemas de aprendizado eletrônico (*e-learning*) de instituições educacionais da Sérvia e da Croácia. Teve como objetivo mensurar e comparar, entre as instituições de ensino, a frequência de termos NS utilizados pelos alunos em discussões assíncronas. Observaram que alunos de ambas as instituições usaram termos NS e que a área de Tecnologia da Informação e Comunicação representou 75% dos estudantes que mais os utilizam.

Citado em trabalhos de mineração de textos em redes sociais [Olagunju et al., 2020; Tadesse et al., 2019], o Linguistic Inquiry and Word Count (LIWC) é um programa que utiliza um léxico composto por palavras organizadas em categorias e subcategorias. A versão de $2015$ do LIWC introduziu a subcategoria "Netspeak", contendo exclusivamente termos utilizados na Internet. Entretanto, o léxico do LIWC não possuiu uma estrutura que aponte para termos equivalentes.

No trabalho apresentando o desenvolvimento do léxico "Netspeak-BR" [Nascimento et al., 2019], são extraídos arquivos de texto em língua portuguesa de três RS distintas. O trabalho detalha o processo de curadoria de termos **informais** (NS). No estudo em questão, o léxico não é utilizado em tarefas típicas da área de mineração de textos.

De maneira geral, nos trabalhos citados previamente nesta seção, não foram apresentadas propostas que observam a variação de resultados ocorrida pela substituição de termos **informais** (NS) por termos de acordo com as regras ortográficas da língua portuguesa (**formais**). Como contribuição do presente trabalho, é destacada a comparação de resultados de tal substituição mediante o uso do "Netspeak-BR" em atividades de pré-processamento dos textos, resultando na alteração dos textos que são submetidos aos classificadores de aprendizado de máquina.

## 3. Metodologia

De forma a comparar os resultados de variados classificadores mantendo os termos NS ou substituindo-os por termos formais, seguiu-se a metodologia apresentada na Figura 1. Primeiro, ocorre a obtenção dos dados das redes. A seguir, é realizado o pré-processamento dos dados utilizando o léxico contendo termos informais para substituição de termos informais por termos formais. São então selecionados os atributos mais relevantes para então ser realizada a classificação e a avaliação das diferenças observadas pelo ajuste dos parâmetros em cada etapa. Cada um destes aspectos está detalhado a seguir.


**Figura 1. Visão geral do modelo utilizado.**

Utilizou-se textos da RS Twitter e da comunidade de conteúdo Youtube. Da RS Twitter, foram extraídos, via *stream*, no período de junho de 2018, textos relacionados a classificação de polaridade emocional negativa / positiva, sendo selecionados 1.500 textos classificados como *entradas positivas* e 1.500 como *entradas negativas*, totalizando 3.000 documentos. Da RS Youtube buscou-se entradas de cunho depressivo, obtidos pela busca de vídeos utilizando o termo "depressao". Foram então extraídos comentários com a *string* "depressao+suicidio", resultando em um total de 1.705 entradas.

O léxico contendo termos informais utilizado neste trabalho é o NetSpeak-BR [Nascimento et al., 2019]. Composto por um total de 429 termos da língua portuguesa, o NetSpeak-BR[1] é um léxico contendo termos informais que são amplamente utilizados em redes sociais brasileiras. Neste léxico, cada termo informal é associado a sua respectiva formalização (e.g. *"pfv=por favor"; "mt=muito"; "plmds=pelo amor de Deus"*).

Utilizando a linguagem Python 3.7, para cada texto foi aplicado o processo de conversão para letras minúsculas (*lower case*), seguido do processo conhecido como "*tokenização*", que transforma o texto em um vetor de termos (*tokens*). Cada *token* foi então submetido ao processo de busca de termos no Léxico NetSpeak-BR, com o objetivo de substituir o termo informal pelo termo formal, caso encontrado.

O *Chi-Squared* (CHI2) é um teste estatístico utilizado para obter uma medida da relação entre resultados experimentais da medida de uma variável e valores conhecidos de média e desvio padrão [Liu and Setiono, 1995]. O CHI2 é aplicado em mineração de textos como técnica de apoio para seleção de atributos (*features*) mais relevantes [Setiyanin-

[1]https://eic.cefet-rj.br/~lacafe/netspeak/

grum et al., 2019], descartando os de menor importância. Neste trabalho, considerou-se as faixas de $50$, $100$ e $200$ melhores atributos para o experimento [Liu and Setiono, 1995].

Os algoritmos classificadores escolhidos são: Multinomial Naive Bayes (MNB), máquina de vetores de suporte ou *Support Vector Machine* (SVM), e Floresta Aleatória ou *Random Forest* (RF). Além da conveniência da disponibilização de implementações embutidas e prontas para uso destes algoritmos na biblioteca *scikit-learn*[2], a escolha de tais algoritmos também é motivada pela prévia utilização dos mesmos em outros trabalhos de mineração de textos [Wang and Qu, 2017]. Os hiperparâmetros utilizados nos classificadores são os mesmos valores definidos pela biblioteca como "padrão".

Nesta tarefa de classificação, o conjunto de dados é submetido ao processo conhecido como validação cruzada. Desta forma, o conjunto é particionado em cinco subconjuntos de forma estratificada. Cada subconjunto utiliza a mesma proporção percentual dos documentos e classes de todo o *corpus*, buscando manter a homogeneidade dos dados.

Para cada subconjunto, os textos são representados no espaço vetorial utilizando o esquema de pesos de termos conhecido como TF-IDF (*Term Frequency – Inverse Document Frequency*, ou Frequência dos Termos - Inverso da Frequência nos Documentos). Este esquema considera que a frequência do termo (TF) representa o número de vezes que um determinado termo $t_i$ ocorre em um documento $d_j$ e é proporcional à quantidade de documentos do conjunto em que o termo $t_i$ aparece ao menos uma vez.

Logo em seguida, ocorre a seleção dos *n-atributos* indicados pelo teste CHI2. Então, os vetores compostos pelos atributos são enviados ao classificador. Como método para avaliação do desempenho do classificador, foi adotada a média de cada partição do valor de F1, que é o resultado da medida harmônica entre precisão e revocação. Os cálculos destas medidas, consideram os elementos corretamente identificados (VP), indevidamente identificados (FP) e indevidamente não identificados (FN).

## 4. Resultados

Cada conjunto de texto foi submetido de forma individual ao processo de classificação dos dados. Os valores de F1 obtidos utilizando MNB, SVM e RF para os textos extraídos do Youtube são apresentados na Tabela 1, enquanto na Tabela 2 estão indicados os valores de F1 para os textos do Twitter. Os valores em **negrito** destacam os maiores valores de F1, de forma a facilitar a comparação visual entre a manutenção do termo informal (NS) e a substituição pelo termo de acordo com as regras ortográficas da língua portuguesa (formal). Para uma comparação estatística destes valores, é calculado o Valor $W$ utilizando o teste dos postos sinalizados de Wilcoxon [Gehan, 1965], considerado como uma versão não paramétrica do Teste-T pareado [Kerby, 2014].

Para os textos do Youtube, observa-se que a utilização do NetSpeak-BR para troca dos termos apresentou valores maiores em dois dos três classificadores utilizados para as faixas de $100$ e $200$ atributos de CHI2. Foram realizados testes de Wilcoxon, sendo observado o Valor $W$ e avaliado que o valor não foi significativo para $p < 0,05$.

Isto pode ser notado pelos resultados encontrados, conforme demostrado na Tabela 1. Houve um empate no classificador SVM utilizando $100$ atributos de CHI2. Porém,

[2]https://scikit-learn.org/stable/

na faixa de 200 atributos de CHI2 com o algoritmo classificador RF, foi observado um valor inferior ao realizar a substituição de termos.

Para os textos do Twitter, o ganho obtido utilizando o léxico NetSpeak-BR se observa em todos os classificadores e em todas a faixas dos melhores atributos (i.e., 50, 100 e 500) indicados pelo CHI2. Diferente dos resultados observados para os textos do Youtube, o resultado foi significativo para $p < 0,05$ nos testes de Wilcoxon realizados. Esta diferença se explica pela limitação de caracteres em cada publicação no Twitter, que favorece a utilização de *emoticons* e abreviações [Ott, 2017].

**Tabela 1. Valores de média de F1, obtidos pela manutenção do termo informal (NS) e pela substituição pelo termo de acordo com as regras ortográficas (formal), da classificação do conjunto de textos do Youtube utilizando Multinomial Naive Bayes (MNB), Support Vector Machine (SVM) e Random Forest (RF).**

| Classificador | 50 atributos | | 100 atributos | | 200 atributos | |
|---|---|---|---|---|---|---|
| | Formal | NS | Formal | NS | Formal | NS |
| MNB | 0,680 | **0,689** | 0,745 | **0,756** | 0,773 | **0,778** |
| SVM | 0,785 | **0,788** | 0,789 | 0,789 | 0,784 | **0,785** |
| RF | 0,785 | **0,786** | 0,784 | **0,790** | **0,789** | 0,781 |

**Tabela 2. Valores de média de F1, obtidos pela manutenção do termo informal (NS) e pela substituição pelo termo de acordo com as regras ortográficas (formal), da classificação do conjunto de textos do Twitter utilizando Multinomial Naive Bayes (MNB), Support Vector Machine (SVM) e Random Forest (RF).**

| Classificador | 50 atributos | | 100 atributos | | 200 atributos | |
|---|---|---|---|---|---|---|
| | Formal | NS | Formal | NS | Formal | NS |
| MNB | 0,578 | **0,586** | 0,615 | **0,623** | 0,647 | **0,647** |
| SVM | 0,652 | **0,668** | 0,669 | **0,683** | 0,676 | **0,692** |
| RF | 0,633 | **0,649** | 0,648 | **0,660** | 0,659 | **0,671** |

## 5. Conclusões

Este trabalho apresentou o resultado da classificação de polaridade emocional de textos obtidos de plataformas online de RS, comparando o pré-processamento dos dados com manutenção ou a substituição de termos informais por termos formais. Para isso, foi utilizado um léxico contendo termos que são amplamente utilizados nas RS e seus equivalentes, que estão de acordo com as regras ortográficas da língua portuguesa. Foi possível notar uma diferença nos resultados de classificação dos textos de RS, mas que não se mostrou consistente nas diferentes redes sociais analisadas. Isto aponta para uma etapa decisória adicional com relação a adesão ou não desta atividade em tarefas de pré-processamento de textos.

A limitação deste estudo, tanto pela utilização de um único léxico de NS, quanto pela obtenção dos dados de somente duas plataformas de RS, suscita novos questionamentos. Além da comparação de diferentes léxicos de NS, trabalhos futuros podem contemplar a avaliação dos resultados da classificação de textos de outras redes sociais, de forma a poder generalizar os resultados.

## 6. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Os autores agradecem à FAPERJ e ao CNPq pelo financiamento parcial desta pesquisa.

## Referências

Aleksić-Maslać, K., Bulatović, V., and Biočina, Z. (2019). Netspeak in asynchronous student-student discussion among different faculty, gender and language groups. In *2019 IEEE Frontiers in Education Conference (FIE)*, pages 1–6.

Danet, B. and Herring, S. C. (2007). The multilingual internet. *Journal of Computer-Mediated Communication*.

Gehan, E. A. (1965). A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2):203–224.

Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, 3:11–IT.

Liu, H. and Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, pages 388–391. IEEE.

Liu, W. and Liu, W. (2014). Analysis on the word-formation of english netspeak neologism. *Journal of Arts and Humanities*, 3(12):22–30.

Mamgain, N., Pant, B., and Mittal, A. (2016). Categorical data analysis and pattern mining of top colleges in india by using Twitter data. In *2016 8th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 341–345.

Nascimento, R., Santos, L. F., and Guedes, G. P. (2019). Netspeak-br: Um léxico sobre expressões criadas na língua portuguesa brasileira para a internet. *Conference: STIL 2019 - XII Symposium in Information and Human Language Technology At: Salvador, BA, Brazil*.

Olagunju, T., Oyebode, O., and Orji, R. (2020). Exploring key issues affecting african mobile ecommerce applications using sentiment and thematic analysis. *IEEE Access*, 8:114475–114486.

Ott, B. L. (2017). The age of Twitter: Donald J. Trump and the politics of debasement. *Critical studies in media communication*, 34(1):59–68.

Oyong, I., Utami, E., and Luthfi, E. T. (2018). Natural language processing and lexical approach for depression symptoms screening of indonesian twitter user. *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*.

Setiyaningrum, Y. D., Herdajanti, A. F., Supriyanto, C., and Muljono (2019). Classification of Twitter contents using chi-square and k-nearest neighbour algorithm. In *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pages 1–4.

Tadesse, M. M., Lin, H., Xu, B., and Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Wang, Z. and Qu, Z. (2017). Research on web text classification algorithm based on improved CNN and SVM. In *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, pages 1958–1961.