Análise de Sentimentos em Tweets: Pré-processamento, Comparação de Algoritmos e Métricas de Avaliação

Luciano D. S Pacífico [1], Lidiane Monteiro [1], Renan Tomazini [1]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco(UFRPE)
Recife – PE – Brazil

luciano.pacifico@ufrpe.br, lidian.monteiro@ufrpe.br,
renan.tomazini@ufrpe.br

Abstract. *Employing a Python script to contrast the efficacy of three classification algorithms, namely Naïve Bayes, K-Nearest Neighbors, and Decision Tree, in sentiment analysis of tweets. The script employs the k-fold cross-validation methodology to assess the accuracy of the algorithms across diverse datasets. Two critical evaluation metrics, accuracy and execution time, are deployed in the study. The primary objective of the experiment is to ascertain the algorithm that exhibits the most exceptional performance in accurately categorizing sentiment in microblogs.*

Resumo. *Utilizando um script em Python para contrastar a eficácia de três algoritmos de classificação, nomeadamente Naïve Bayes, K-Nearest Neighbors e Decision Tree, na análise de sentimentos em tweets. O script utiliza a metodologia de validação cruzada k-fold para avaliar a precisão dos algoritmos em diferentes conjuntos de dados. Dois métricas críticas de avaliação, precisão e tempo de execução, são empregadas no estudo. O objetivo principal do experimento é determinar o algoritmo que exibe a melhor performance na categorização precisa de sentimentos em microblogs.*

1. Introdução

All full papers and posters (short papers) submitted to some SBC conference, including any supporting documents, should be written in English or in Portuguese. The format paper should be A4 with single column, 3.5 cm for upper margin, 2.5 cm for bottom margin and 3.0 cm for lateral margins, without headers or footers. The main font must be Times, 12 point nominal size, with 6 points of space before each paragraph. Page numbers must be suppressed.

Full papers must respect the page limits defined by the conference. Conferences that publish just abstracts ask for **one**-page texts.

2. Base de dados

The first page must display the paper title, the name and address of the authors, the abstract in English and “resumo” in Portuguese (“resumos” are required only for papers written in Portuguese). The title must be centered over the whole page, in 16 point boldface font and with 12 points of space before itself. Author names must be centered in 12 point font, bold, all of them disposed in the same line, separated by commas and with 12 points of space after the title. Addresses must be centered in 12 point font, also with 12 points of space after the authors’ names. E-mail addresses should be written using font Courier New, 10 point nominal size, with 6 points of space before and 6 points of space after.

The abstract and “resumo” (if is the case) must be in 12 point Times font, indented 0.8cm on both sides. The word **Abstract** and **Resumo**, should be written in boldface and must precede the text.

3. Sobre a API usada

In some conferences, the papers are published on CD-ROM while only the abstract is published in the printed Proceedings. In this case, authors are invited to prepare two final versions of the paper. One, complete, to be published on the CD and the other, containing only the first page, with abstract and “resumo” (for papers in Portuguese).

4. Metodologia

Section titles must be in boldface, 13pt, flush left. There should be an extra 12 pt of space before each title. Section numbering is optional. The first paragraph of each section should not be indented, while the first lines of subsequent paragraphs should be indented by 1.27 cm.

5. Modelos apresentados

Figure and table captions should be centered if less than one line (Figure 1), otherwise justified and indented by 0.8cm on both margins, as shown in Figure 2. The caption font must be Helvetica, 10 point, boldface, with 6 points of space before and after each caption.

Figure 1. A typical figure

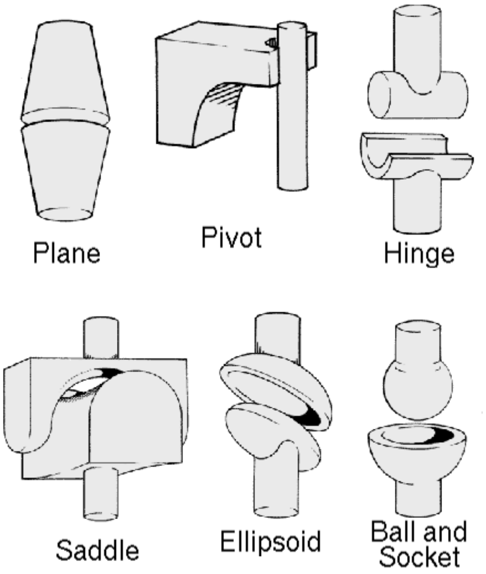
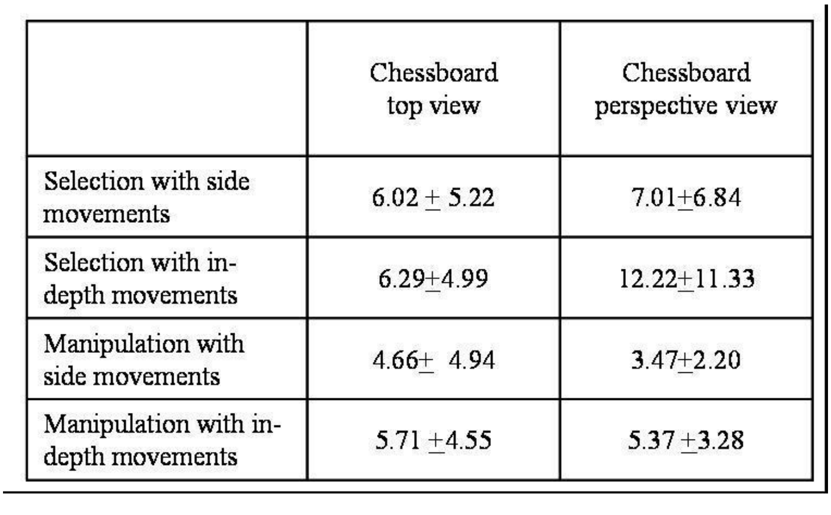


Figure 2. This figure is an example of a figure caption taking more than one line and justified considering margins mentioned in Section 5.

In tables, try to avoid the use of colored or shaded backgrounds, and avoid thick, doubled, or unnecessary framing lines. When reporting empirical data, do not use more decimal digits than warranted by their precision and reproducibility. Table caption must be placed before the table (see Table 1) and the font used must also be Helvetica, 10 point, boldface, with 6 points of space before and after each caption.

Table 1. Variables to be considered on the evaluation of interaction techniques



	Chessboard top view	Chessboard perspective view
Selection with side movements	6.02 ± 5.22	7.01±6.84
Selection with in-depth movements	6.29±4.99	12.22±11.33
Manipulation with side movements	4.66± 4.94	3.47±2.20
Manipulation with in-depth movements	5.71 ±4.55	5.37 ±3.28

6. Análise experimental

All images and illustrations should be in black-and-white, or gray tones, excepting for the papers that will be electronically available (on CD-ROMs, internet, etc.). The image resolution on paper should be about 600 dpi for black-and-white images, and 150-300 dpi for grayscale images. Do not include images with excessive resolution, as they may take hours to print, without any visible difference in the result.

7. Conclusões

Em conclusão, os resultados do estudo indicam que o algoritmo NaiveBayes exibiu uma acurácia média de 66,6% no conjunto de dados curado por humanos e 88,1% no conjunto de dados NLTK. Por outro lado, o algoritmo KNN apresentou acurácias médias mais baixas, variando de 59,4% a 60,4% no conjunto de dados curado por humanos e de 86,9% a 88,9% no conjunto de dados NLTK, dependendo do valor do parâmetro K escolhido.

Além disso, a árvore de decisão treinada com o conjunto de dados NLTK demonstrou uma precisão significativamente maior do que aquela treinada com o conjunto de dados curado por humanos, com uma acurácia média superior a 0,97 em comparação a menos de 0,60, respectivamente. No entanto, a escolha da melhor opção de treinamento depende dos dados disponíveis e dos objetivos do projeto. Embora os dados de treinamento curados por humanos sejam geralmente considerados mais precisos, eles podem exigir mais tempo e recursos, enquanto o treinamento com ferramentas como o NLTK pode ser mais rápido e escalável, mas pode exigir mais esforço na preparação dos dados e na seleção dos parâmetros.

É importante notar que o tempo de execução em segundos foi consistente para todos os testes, sugerindo que não houve variação significativa no tempo de execução entre os classificadores e os conjuntos de dados utilizados. É essencial lembrar que a escolha do algoritmo de classificação e dos parâmetros depende do problema específico e dos dados disponíveis, e os resultados obtidos neste estudo podem não ser aplicáveis a outras aplicações.

References

[1] YADAV, Ashima; VISHWAKARMA, Dinesh Kumar. Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, v. 53, n. 6, p. 4335-4385, 2020.

[2] KATHURIA, Ramandeep Singh et al. Real time sentiment analysis on twitter data using deep learning (Keras). In: 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). IEEE, 2019. p. 47-52.

[3] ALONSO, Miguel A. et al. Sentiment analysis for fake news detection. Electronics, v. 10, n. 11, p. 1348, 2021.

[4] DANG, N. C.; MORENO-GARCÍA, M. N.; DE LA PRIETA, F. Sentiment Analysis Based on Deep Learning: A Comparative Study. Electronics, v. 9, n. 3, p. 483, 2020.