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

A análise de sentimentos em mídias sociais, como o Twitter, tem se tornado uma área de pesquisa cada vez mais relevante. Com a crescente quantidade de dados gerados nas redes sociais, entender a opinião do público pode ser uma vantagem competitiva para empresas e governos. Neste estudo, nosso objetivo é comparar a eficácia de três algoritmos de classificação - KNN, NaiveBayes e Árvore de decisão - na classificação de tweets como positivos, negativos ou neutros. A avaliação dos resultados obtidos pelos algoritmos pode auxiliar na escolha da melhor opção de análise de sentimentos para um determinado conjunto de dados e objetivo.

2. Base de dados

No dia 24/03/2022, às 6 da manhã, foram coletados dados via API do Twitter utilizando palavras-chave relevantes para o contexto regional, tais como "CLT", "Recife", "Camaragibe", "Olinda", "UFRPE", "UFPE", "@prefrecife", "APAC", "@SantaCruzFC", "@sportrecife", "@nauticope", "@recifeordinario", "@nautiluslink" e "@mimimidias". O objetivo da coleta foi selecionar assuntos populares e relacionados à mídia em Pernambuco, garantindo, assim, uma base de dados regional.

No total, foram coletados 1082 tweets que passaram por duas etapas de pré-processamento e classificação. Inicialmente, foram classificados manualmente por

humanos para garantir a qualidade e relevância dos dados. Em seguida, foram submetidos ao processamento de linguagem natural utilizando a ferramenta NLKT para melhorar a eficiência da classificação. O objetivo dessa análise é comparar a eficácia dos algoritmos KNN, NaiveBayes e Árvore de Decisão para classificar os tweets em positivo, negativo ou neutro em relação aos assuntos selecionados.

3. Sobre a API usada

A análise de sentimentos é uma técnica de processamento de linguagem natural que permite identificar e extrair opiniões e emoções presentes em um texto. No presente estudo, foram utilizados dois métodos de classificação: o processamento de linguagem natural com a ferramenta NLTK para Python e a classificação manual realizada por humanos.

Os resultados obtidos a partir desses métodos apresentaram diferenças expressivas. Enquanto a ferramenta NLTK obteve uma precisão de 88,1% na classificação dos tweets, a classificação humana obteve uma precisão de menos de 60%. Isso sugere que a utilização de ferramentas de processamento de linguagem natural pode ser uma alternativa mais eficiente para a classificação de grandes volumes de dados. No entanto, é importante considerar que a escolha do melhor método de classificação dependerá dos objetivos e das características do projeto em questão.

4. Metodologia

A metodologia proposta para a comparação de algoritmos consiste na execução de 10-fold cross-validation em 5 iterações para cada uma das duas bases de dados. Cada uma das 50 execuções parte de uma distribuição aleatória dos dados entre os folds, garantindo uma avaliação justa dos resultados. Os conjuntos de treinamento e teste são mantidos iguais para cada algoritmo testado em cada experimento. Os três algoritmos testados são Naïve Bayes, K-Vizinhos Mais Próximos (K-NN) - com variação em 3 valores de k - e Árvore de decisão. Pelo menos duas métricas de avaliação são empregadas em cada análise experimental, juntamente com o tempo médio de execução de cada algoritmo.

5. Modelos apresentados

O KNN (K-Nearest Neighbors) é um algoritmo de aprendizado de máquina supervisionado que se baseia na proximidade entre os dados para classificá-los. Ele funciona encontrando os K pontos mais próximos de um novo ponto e atribuindo a ele a classe mais frequente entre esses K pontos.

O Naive Bayes é um algoritmo de aprendizado de máquina supervisionado que utiliza o teorema de Bayes para realizar a classificação. Ele assume que as variáveis de entrada

são independentes entre si e calcula a probabilidade condicional de uma classe dada as variáveis de entrada.

A Árvore de Decisão é um modelo de aprendizado de máquina que se baseia em uma estrutura hierárquica de decisão. Ela divide o conjunto de dados em subconjuntos menores e constrói uma árvore onde cada nó representa uma decisão ou uma condição. A árvore é percorrida a partir do topo até o nó folha correspondente à classificação final.

Figure 1. A typical figure

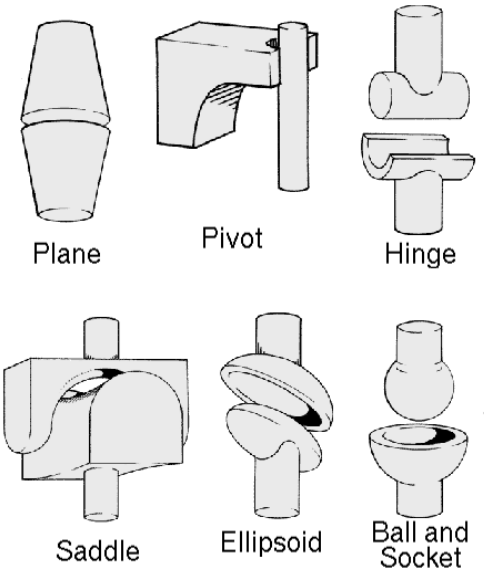

Figure 2. This figure is an example of a figure caption taking more than one line and justified considering margins mentioned in Section 5.

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

Os resultados obtidos indicam que o algoritmo NaiveBayes apresentou acurácia média de 66,6% na base de dados de curadoria humana e de 88,1% na base da NLTK. Já o algoritmo KNN apresentou acurácias médias mais baixas, variando de 59,4% a 60,4% na base de curadoria humana e de 86,9% a 88,9% na base da NLTK, dependendo do valor escolhido para o parâmetro K.

A árvore de decisão treinada com o NLTK obteve uma acurácia muito maior do que a árvore de decisão treinada com a curadoria humana. A acurácia média das repetições para a árvore de decisão do NLTK foi superior a 0,97, enquanto a árvore de decisão com curadoria humana teve uma acurácia média inferior a 0,60. É importante lembrar que a escolha da melhor opção de treinamento depende dos dados utilizados e dos objetivos do projeto. Em geral, o treinamento com dados curados por humanos é considerado mais preciso, mas pode exigir muito mais tempo e recursos. Já o treinamento com ferramentas como o NLTK pode ser mais rápido e escalável, mas pode exigir um esforço maior para a preparação dos dados e seleção dos parâmetros adequados.

(botar a tabela aqui?)

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.