Classificando sentimentos em reviews no dataset do Yelp

Henrique Martinelli Frezzatti, *Insper*

1 Dataset

O *Yelp Academic Dataset Reviews* [1] contém milhões de avaliações de usuários sobre estabelecimentos, incluindo texto, notas e características dos avaliadores. Esse dataset é amplamente utilizado em pesquisas de *Machine Learning*, como no estudo de Andrii Berko [2], por sua relevância para análises de sentimentos.

A análise automatizada dessas avaliações permite que empresas identifiquem percepções dos consumidores, melhorando a experiência do cliente e fortalecendo a fidelidade. Assim, além de seu valor acadêmico, o dataset é uma ferramenta estratégica para organizações que buscam insights e inovação no mercado.

2 Classification Pipeline

A partir deste conjunto, foi realizado um pré-processamento que incluiu limpeza do texto, remoção de stop words e lematização, utilizando o *Lemmatizer* [3] para reduzir palavras a suas raízes, facilitando a análise semântica.

A pipeline de classificação implementada neste estudo segue uma abordagem de aprendizado de máquina. Inicialmente, foi aplicado a técnica de *Bag of Words* [4] para transformar o texto em um formato numérico que pode ser utilizado pelos algoritmos de classificação. Após a vetorização, foram utilizados alguns modelos, incluindo Regressão Logística [5], Random Forest [6] e KNN [7], para treinar e prever as emoções expressas nas avaliações.

3 Evaluation

A avaliação dos modelos foi realizada através de validação cruzada, utilizando o método *K-fold* para garantir a robustez dos resultados. As métricas de desempenho consideradas incluem a acurácia, precisão e recall. Abaixo, é possível observar os dados obtidos:

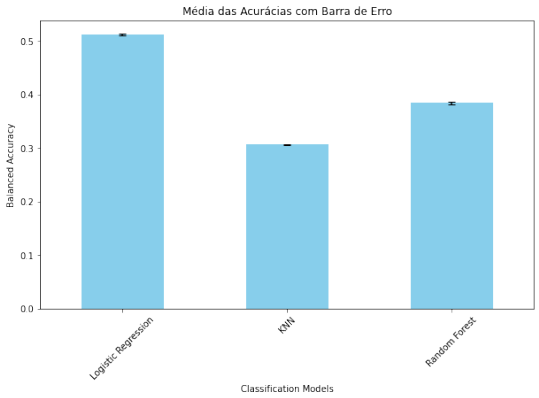

Fig. 1. Média das Acurácias dos Modelos Classificadores com Barra de Erro que demonstram pequenas variações de score.

A partir da análise do gráfico acima, é possível constatar que o melhor modelo registrado, com a maior acurácia, é a Regressão Logística, com um score próximo de 0.51.

4 Dataset Size

Para avaliar o impacto do tamanho do dataset na acurácia dos modelos, foram realizados testes de *downsampling*, onde diferentes tamanhos do *dataset* foram utilizados para treinar os modelos. A análise revelou que, embora o aumento do tamanho do *dataset* tenha potencial para melhorar a acurácia, a eficiência desse aumento depende da diversidade e relevância dos dados. Essa análise é crucial para determinar a viabilidade de expandir o dataset em termos de custo-benefício.

Os resultados obtidos podem ser vistos abaixo:

Downsampling Scores				
Model \ Dataset size	**10%**	**25%**	**50%**	**75%**
Random Forest	0.34	0.36	0.38	0.38
Logistic Regression	0.47	0.49	0.50	0.50
KNN	0.28	0.29	0.29	0.30

Fig. 2. Resultados obtidos com diferentes modelos de classificação em diferentes tamanhos do dataset. É possível observar que, quanto maior o dataset, melhor a precisão de todos os modelos, diminuindo esse crescimento conforme chega perto dos 75%.

5 Topic Analysis

Para entender melhor a estrutura dos dados, foi aplicado o modelo *Latent Dirichlet Allocation (LDA)* [8] para a análise de

Feito em: 1/10/2024

tópicos. O LDA revelou tópicos predominantes que podem influenciar a classificação das avaliações. Além disso, a performance de classificação variou conforme os tópicos, sugerindo que uma abordagem de classificador em duas camadas, onde documentos são inicialmente classificados por tópicos e, em seguida, direcionados a classificadores específicos, pode aumentar a acurácia geral da classificação.

REFERENCES

[1] Yelp Dataset. Disponível em: Kaggle

[2] A. Berko, "Sentiment Classification Using Machine Learning Approaches," in *Scientific Papers of Lviv Polytechnic National University*, vol. 2, 2019. [Online]. Disponível em: A. Berko

[3] B. S. Siddhartha, "An Interpretation of Lemmatization and Stemming in Natural Language Processing," 2021. [Online]. Disponível em: ResearchGate

[4] W. A. Qader, "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges," 2019. [Online]. Disponível em: ResearchGate

[5] C. Starbuck, "The Fundamentals of People Analytics," 2023. [Online]. Disponível em: Springer

[6] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, 2001. [Online]. Disponível em: SpringerLink

[7] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, 1967. [Online]. Disponível em: IEEE Xplore

[8] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, 2003. [Online]. Disponível em: ACM Digital Library