Classifying Food Reviews on the Amazon Fine Food Reviews dataset

Tiago Vitorino Seixas

I. Dataset

THIS report was made using the Amazon Fine Foods Reviews dataset [3] [5] from Kaggle, with the purpose of aiding the eCommerce industry, following the work of Xinyue Zhao and Yuandong Sun. [6]

II. Classification pipeline

TO preprocess the model, the dataset went through lemmatization [2], using the WordNet as a database. following that, with scikit-learn, the train and test samples were split and the pipeline, which used CountVectorizer [1] to tokenize the words and a Logistic Regression [4] model to train the classifier, was applied to the samples.

III. Evaluation

THE most important words would be the ones that give an accurate image of what are the strengths and weaknesses of the eCommerce operations, although it was noticed that most of the words detected by the classifier are nonsensical, which means that the preprocessing needs improvement, which could be achieved by reducing the scores to 2 (positive and negative) and excluding the words that are nonsensical, for example, or that a Logistic Regression model itself is not an ideal choice for this dataset.

IV. Dataset size

THE accuracy of the model increases with the sample size [Fig. 1], the blue curve being the model and the red curve the actual tests, with both of them tending towards 0.6. Should a bigger sample be used, the results would be better, but still less accurate than the results obtained by Xinyue Zhao and Yuandong Sun [6] who managed to achieve an accuracy of 0.7982 (represented by the green line) with a BERT model, so to achieve better results, it would be better to change the model of the classifier, or review the preprocessing steps.

V. Topic analysis

THE Topic Analysis makes another problem evident, which is that common english words, which were not removed during the preprocessing, have now become an issue, and the accuracy of the model has greatly decreased, from 0.56 to 0.22, which seems to indicate that topic modelling by itself is not a good approach.

T. V. Seixas was an Undergraduate Student of Computer Engineering, Research and Teaching Institute Insper, São Paulo, SP, 04546-042 Brazil. organisation e-mail: tiagovs1@al.insper.edu.br

Project submitted October 3, 2024.

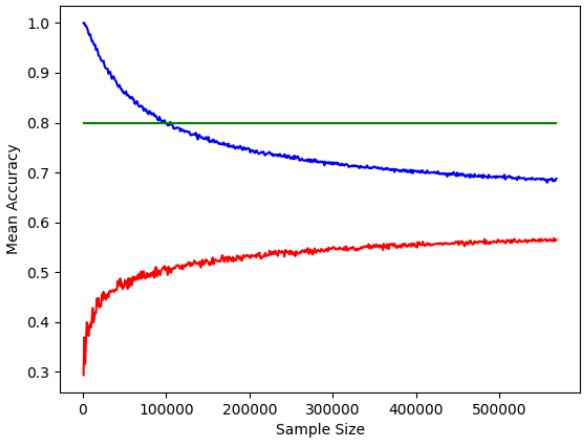

Fig. 1: Mean accuracy of classifier by the sample size

References

[1] P. Jain. “Basics of CountVectorizer”. In: *WWW* (2021). URL: https://towardsdatascience.com/basics-of-countvectorizer-e26677900f9c.

[2] C. D. Manning, P. Raghavan, and H. Schütze. “Stemming and lemmatization”. In: *Cambridge University Press* (2008). URL: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

[3] J. McAuley and J. Leskovec. “From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews.” In: *WWW* (2013).

[4] n.a. “What is logistic regression”. In: *IBM* (2024). URL: https://www.ibm.com/topics/logistic-regression.

[5] Stanford Network Analysis Project and Kaggle Team. “Amazon Fine Food Reviews”. In: *WWW* (2017). URL: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews.

[6] Xinyue Zhao and Yuandong Sun. “Amazon Fine Food Reviews with BERT Model”. In: *Procedia Computer Science* 208 (2022). 7th International Conference on Intelligent, Interactive Systems and Applications, pp. 401–406. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2022.10.056. URL: https://www.sciencedirect.com/science/article/pii/S1877050922014971.