Animesh Deb

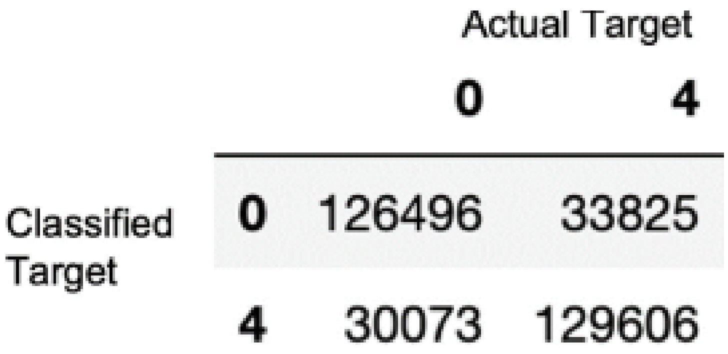
December 17, 2023

Logistic Regression Model

Logistic Regression is the first model I tested. Logistic Regression is a binary classification, and this model is a supervised learning model originally used in the field of statistics. I chose it mainly because I have experience with it using it in class and because it’s one of the most used classification models. I first imported my dataset, specifically choosing the text and sentiment value fields. I then split the dataset such that 80% is to be used in training and the remaining 20% on testing. The dataset was now read, and I started to generate the models.

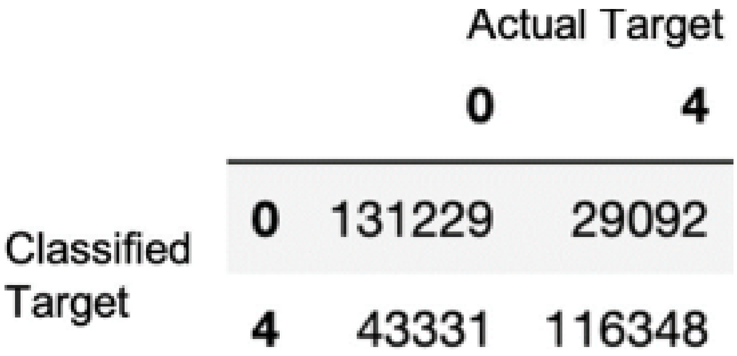
I generated the model where Count Vectorization was used. After training and testing, the resulting accuracy was 0.7983125. The resulting classifications of this model are shown in Figure 6. As shown below, there are 126,496 true positive classifications and 129,606 true negative classifications. With these results and the accuracy score, I believe it is safe to say that the Logistic Regression model did very well on classifying whether the tweets were positive or negative in sentiment.

I also tested the TF-IDF vectorization technique with the Logistic Regression model, and after training and testing, it performed similarly to Count Vectorization, albeit somewhat better. The resulting accuracy score was 0.802515625. As shown in Figure 7, there are 131,229 true positive classifications and 116,348 true negative classifications. Since this accuracy is already higher than the well performing Count Vectorization, this model has done very well itself, and it can be said that TF-IDF reduced the effectiveness of “stopping words” and this benefited the Logistic Regression model.



		Actual Target	
		0	4
Classified Target	0	126496	33825
	4	30073	129606

Figure 1: True positive and true negative classifications of the Logistic Regression model using Count Vectorization



		Actual Target	
		0	4
Classified Target	0	131229	29092
	4	43331	116348

Figure 2: True positive and true negative classifications of the Logistic Regression model using TF-IDF Vectorization.