The Type 2 improvement has a simple underlying logic. It involves utilizing texts that have been labeled with multiple emotions, and assuming that when multiple emotions are labeled for the same text, there is a close relationship among them. To construct 28 emotion vectors with 58009 features, we transpose the matrix of the GoEmotion dataset, which contains 58009 valid samples with 28 emotions. We then use correlation to gauge the similarity between any two emotions and perform clustering based on the correlation calculation results. We group them into major categories by setting the distance to 1.05.

Next, we create a multiple-stage classification model that first classifies the text into a major category and then maps it to the corresponding GoEmotions categories. To evaluate the performance of this model, we compare it with the baseline and the Type 1 improvements using F1 scores as the metrics.

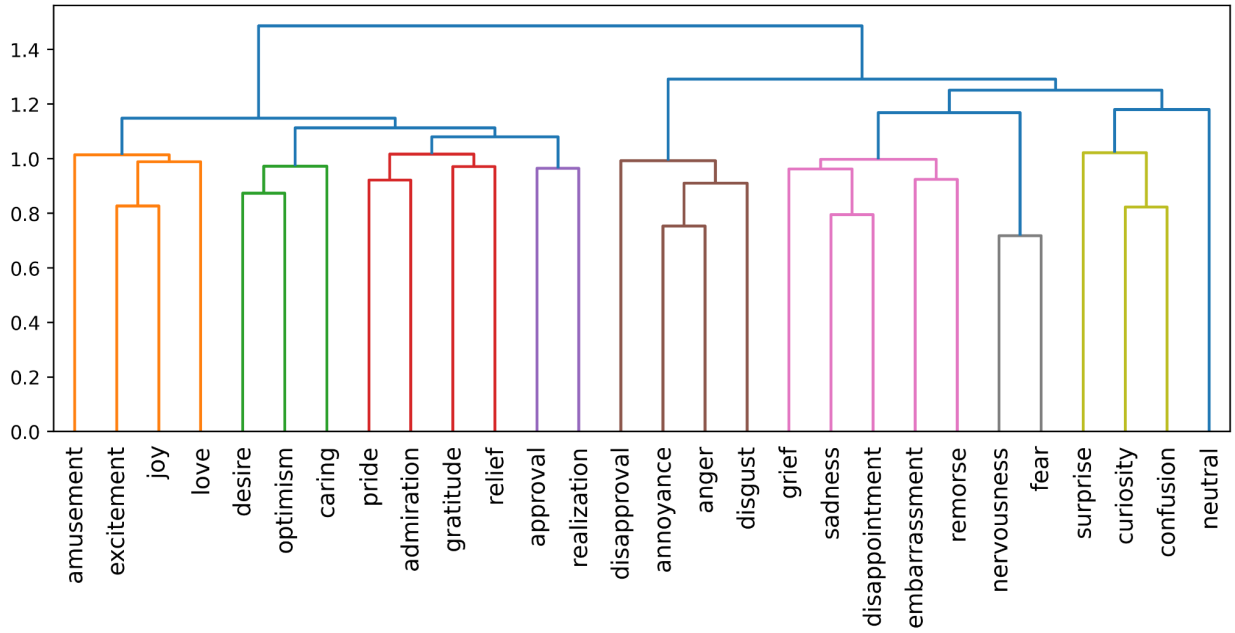

Fig. 4: Data clustering