

Assignment 8: Sentiment Analyzer

In this assignment you will construct your own sentiment analyzer.

1 Word Based Sentiment Analyzer

Write a function that takes a tweet as input and throws back its sentiment. To this end, the function needs to 1) split the tweet into individual words devoid of all punctuation 2) extract any valence terms from this **list** found among the words, and 3) calculate the valence of the tweet by averaging the corresponding valence values. Step 2) is best achieved using `terms %in% words`. Regarding step 1) I recommend implementing the procedure from Assignment 7 to split sequences of Emojis. This has the positive side-effect that the Emojis reduce to the smaller set of single-byte Emojis.

2 Measure Emoji Valence

Include all or some Emojis in the list of valence terms. Two to ways to achieve this:

2.1a Context Route

The context route shares similarities to the co-occurrence approach applied in past assignments. That is, you can infer the valence of an emoji by assessing the words it co-occurs with. The easiest way to do this is to derive valence using the Sentiment Analyzer from the section above for the tweets the Emoji occurs in and averaging the results. A slightly different and possibly better approach would be to determine the proportion of tweets that are positive for each Emoji. Note, however, that in the latter case the proportions need to be scaled to the range of -4 and 4 to match the range in the valence term list.

2.1b Similarity Route

The similarity route is a sophisticated version of the context route. That is, rather than defining the context as the collection of tweets the Emoji occurs in, we can define the context using the similarities learned from the tweets using, e.g., `word2vec`. The best way to do this, is to train the model again for all occurring Emojis and all occurring valence terms and then compute the valence of the Emojis as the sum of valences of the terms weighted by the cosine similarities between the terms and the respective Emoji. Note, however, that the cosines need to be normalized so that they sum up to 1 to ensure again a valence range of -4 and 4.

2.2 Expand Valence Terms

After completing either the Context or the Similarity Route add the full set of Emojis (or only a subset, in which case the extreme ones should be preferred) to the term list.

3 Evaluate Sentiment Analyzer

Download this **list** of classified tweets and use it to evaluate your sentiment analyzer. Specifically, extract the tweets and apply your sentiment analyzer to each of them. Then compare the predictions of your sentiment analyzer with the existing valence classification of positive or negative valence. One way to do this is to compare summary statistics for the predictions for the positive and the predictions for the negative tweets. For instance, one could calculate the means and standard deviations of the prediction and calculate a Cohen's d as a measure of sensitivity. Alternatively, you can analyze the number of times your sentiment analyzer predicts the correct polarity, i.e., whether a positive tweet is also assigned an above zero number and vice versa. Make sure that the results demonstrate a successful classification, e.g., by checking the sign and value of d .

END