Analysis

1. In total, there 171820 contents and in terms of the sentiment, there are 38545 negative, 58968 positive, 55298 neutral and 19009 ambiguous. So the confusion matrix should be 4x4.Accuracy is good because we want to know how accuracy the model perform and the total number of sample is enough so the accuracy can be evaluated fairly accurately for this dataset. Precision is also good for this dataset. We need to know the ratio that how many predictions have been estimated as positive or negative. Recall is available because we also want to know how many positive or negative cases in the sample were correctly predicted. In terms of emotions, we got 3827 sadness, 55298 neutral, 4957 love, 7075 gratitude, 7686 disapproval, 6130 amusement, 4706 disappointment, 4714 realization, 10531 admiration, 8342 annoyance, 4938 confusion, 4519 optimism, 3020 excitement, 3523 caring, 796 nervousness, 2147 desire, 1510 remorse, 11259 approval, 4329 joy, 1433 embarrassment, 3472 surprise, 5885 curiosity, 5202 anger, 351 grief, 2914 disgust, 690 pride, 788 relief, 1778 fear. Accuracy is not good for emotions because there are some emotions that have small sample size. For example grief has 351 samples but remorse has 11259 samples. Recall ,precision and f-measure are good enough for emotions.

Words as feature:

表格

描述已自动生成

表格

描述已自动生成表格

描述已自动生成

DT MultinomialNB MLP

**For all MLP model, we set max\_iter =1, because if we don’t do this the training time would be more than 6 hours. It makes some data lose.**

The prediction of these 3 model for emotion are similar. For accuracy, MLP which is 0.44 is better than other 2 models which are 0.38 and 0.36. It’s worth noting that in DT model, the precision, recall and f-measure of grief, pride and relief are 0. This is probably because the amount of data in these three emotion is so small. That means DT predicted wrongly for all grief, pride and relief. So, we can make a conclusion that MLP is good at accuracy. DT is not good at precision and recall.

表格

描述已自动生成

表格

描述已自动生成

表格

描述已自动生成

Prediction of these 3 model for sentiment. For accuracy, like we said before, MLP is better than other 2 models. If we focus on ambiguous and negative, DT has a more accurate recall but has a worse performing precision on ambiguous and negative. However, there is a converse result which are better performance in terms of precision and a worse performance in terms of recall on neutral and positive. It’s not difficult to see that ambiguous and negative have a smaller sample size compared with neutral and positive, so we can make a conclusion that DT is good at recall on small sample size dataset and good at precision on big sample size dataset.

表格

描述已自动生成 DT in GridSearchCV

表格

描述已自动生成 MultinomialNB in GridSearchCV

表格

描述已自动生成 MLP in GridSearchCV

Prediction of those 3 models in gridSearchCV for emotions. MLP has a worse accuracy which is 0.36 than DT and MultinomialNB. Also, there are many 0 in DT, as we talked before, this is probably because the amount of data in these emotion is so small. however if we look at neutral, we could find that DT has the most accurate recall which is 0.94. That means DT has a better performance on large sample size dataset.

表格

描述已自动生成 DT in gridSearchCV

表格

描述已自动生成 MultinomialNB in GridSearchCV

表格

描述已自动生成MLP in GridSearchCV

For sentiment, MLP has a better accuracy. DT still has a best recall on neutral, however it doesn’t has a better recall on positive. DT is not good at small size of data.

Embeddings as features:

表格

描述已自动生成

This is the MLP for emotions but use embeddings as features, compared with words as features, they have some similarities and differences. MLP in both embedding as features and words as features have some very low recall, for example annoyance, confusion and excitement, but both of them have a high recall on neutral. The difference between embeddings as features and words as features embeddings’ macro and weighted are greater than words’, this makes the f1-score of embedding and words similar.

表格

描述已自动生成

When we look at sentiment, we can find that the 4 sentiments’ precision and recall are similar with words as features. Words’ accuracy which is 0.56 is little higher than embeddings’. It’s not like emotions, the reason for this may be that there are only four types of sentiment and a lot of data in each sentiment, whereas there are many types in emotion with very little data. In addition to this, embeddings’ f1-score is higher than words’.

表格

描述已自动生成

In terms of using hyperparameters for emotion, we find that words as features’ performance of both precision and recall are better than embeddings.

文本

低可信度描述已自动生成

When we focus on sentiment, the conclusions still apply. Words as features’ precision, recall f1-score and accuracy, all of them are perform better than embeddings as features.

表格

描述已自动生成

表格

描述已自动生成

We use another two English pretrained embedding models which are glove-twitter-200 and conceptnet-numberbatch-17-06-300. The hit rate of word2vec-google-news-300 looks better. The 4 metrics do not vary much.

All group members are contributed equally.