Analysis Report

COMP-472

Mini Project 1

Instructor: Dr.Leila Kosseim

Qiantong Zhou (40081938)

Hao Mei (40074373)

Mingyu Gao (40049618)

1. Analysis of the Dataset

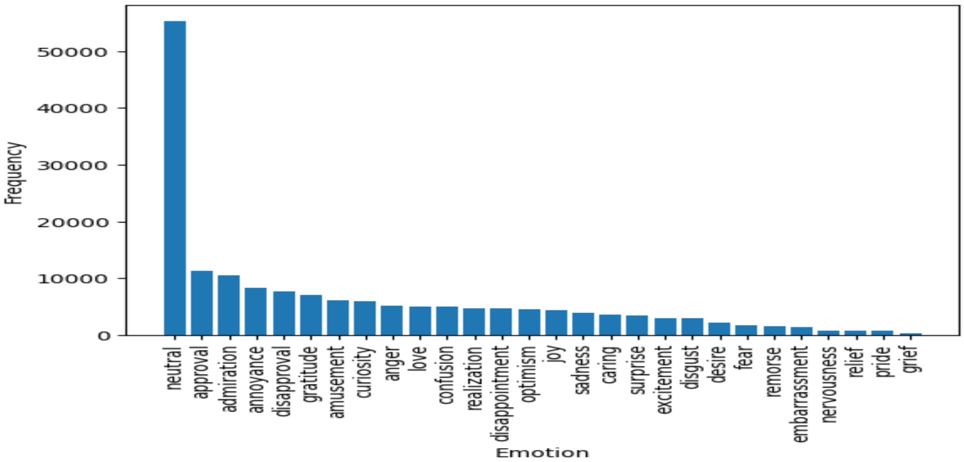
Based on the collection of emotions and sentiment of reddits posts, we got 28 emotions and 4 sentiment. First let's analyze the emotions.

Emotion:

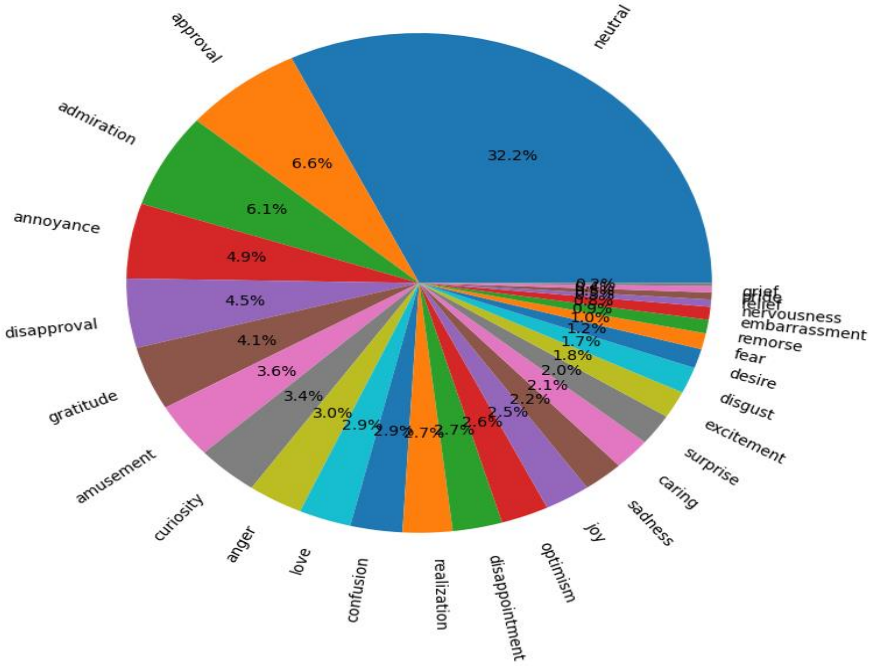
For the emotion category, the dataset is unevenly distributed. According to the following two graphs: histogram and pie chart, the posts with neutral emotion stand at about 32%. Also, he is the most represented of these emotion categories.

However, there are some categories that account for less than 1% of the dataset, such as "remorse", "grief", "pride", "relief " and "nervousness". Other categories of emotions account for about 5-10% each.

Possible reasons for the large variation in the dataset data are: some posts have low internal consistency due to similarity with other emotions, or the difficulty in detecting the correct emotion from the text.

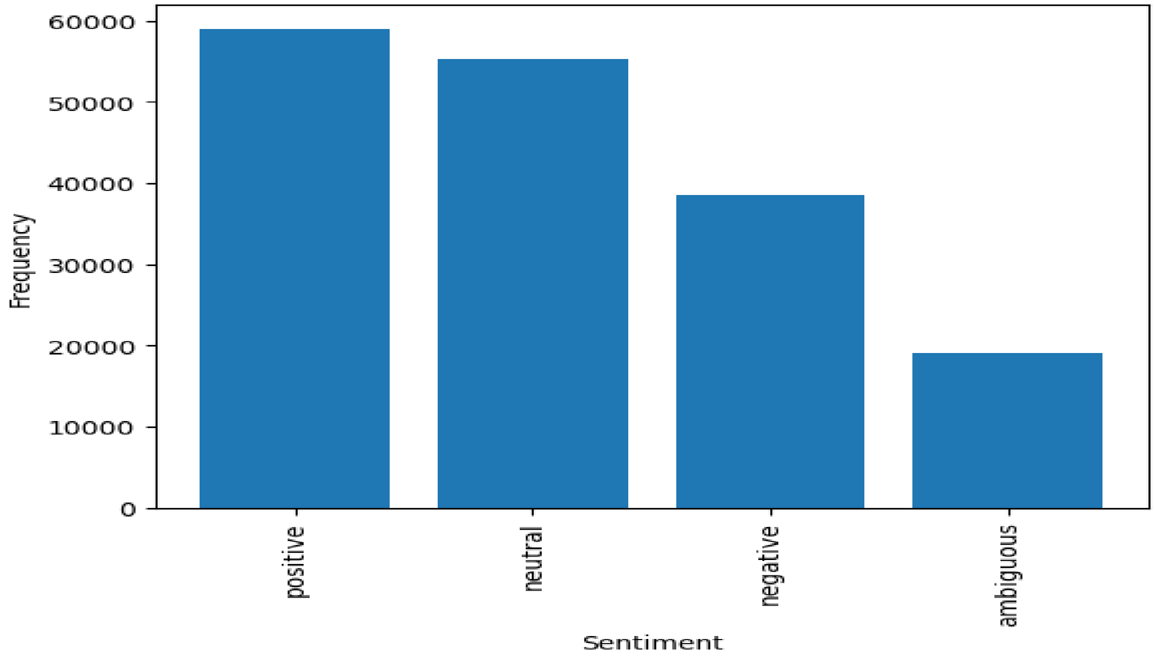


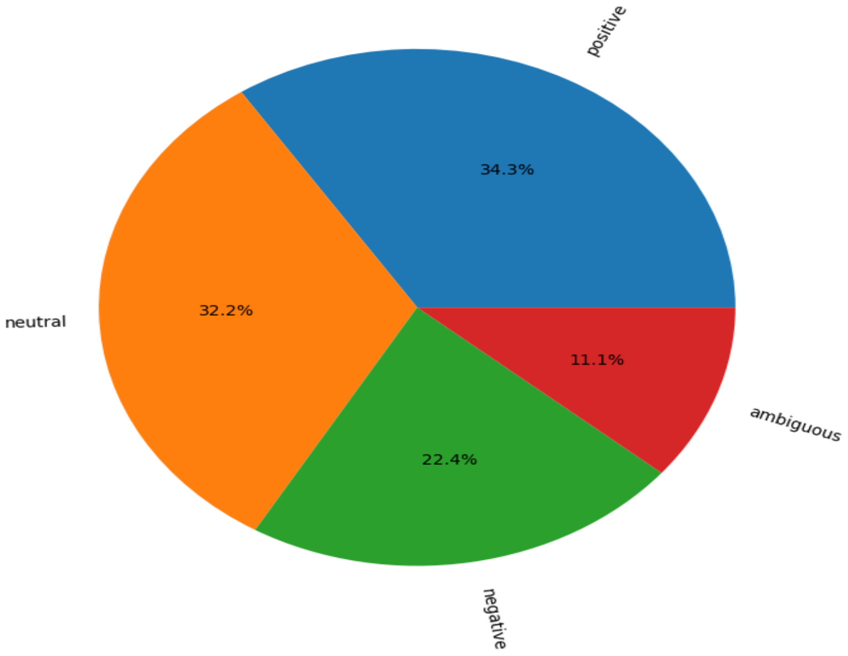
文件中找不到关系 ID 为 rId4 的图像部件。



Sentiment:

There are only 4 categories for the sentiment collection. , the dataset is moderately unbalanced. According to the histogram below, the two most popular categories are "neutral" and "positive", with about 32.2% and 34.3% of posts, respectively. The two most popular categories have about 10% and 20% more posts than the "negative" and "ambiguous"categories.





we have to choose the macro-averaging method. Macro-averaging is probably the most straightforward of the many averaging methods. If you are dealing with an unbalanced data set where all categories are equally important. Using macro-averaging would be a good choice because it treats all categories equally.

2. Model Results Analysis

Emotion(Macro-averaging):

Model	Precision	Recall	f1-score
MultinomialNB	0.59	0.56	0.57
DecisionTreeClassifier	0.70	0.72	0.70
MLPClassifier	0.65	0.60	0.61
MultinomialNB, alpha_0.5	0.59	0.58	0.58
DecisionTreeClassifier,criterion_ginimax_depth_3min_samples_split	0.30	0.29	0.19
MLPClassifier, activation_identityhidden_layer_sizes_(30, 50)	0.62	0.57	0.59

Sentiment(Macro-averaging**):**

Model	Precision	Recall	f1-score
MultinomialNB	0.49	0.18	0.23
DecisionTreeClassifier	0.58	0.51	0.52
MLPClassifier	0.50	0.33	0.37
MultinomialNB, alpha_0.5	0.50	0.28	0.32
DecisionTreeClassifier,criterion_ginimax_depth_3min_samples_split	0.06	0.08	0.07
MLPClassifier, activation_identityhidden_layer_sizes_(30, 50)	0.52	0.31	0.34

Macro-averaging is probably the most straightforward of the many averaging methods. The macro-average F1 score (or macro-F1 score) is calculated using the arithmetic mean (aka unweighted mean) of the F1 scores for all each class. This method treats all classes equally, regardless of their support values.

In general, if you are working with an imbalanced dataset where all classes are equally important, using the macro average would be a good choice as it treats all classes equally.

According to Macro-averaging, it can be seen that DecisionTree Classifier is performing better than the other two models. Its Macro-averaging f-1score for emotion and sentiment is 0.7 and 0.52, respectively.

Compare with the other two Macro-averaging data. As before, the MultinomialNB model does not perform as well as the other two models, probably because it assumes that the features are independent of each other and ignores unrelated features.

The macro precision, recall, f1-score of the DecisionTreeClassifier with sample split in emotion is 0.06, 0.08, 0.07. The poor results for recall and precision are due to the imbalance of neutral labeled posts in the training dataset. What is happening here is the same problem as the basic MNB emotion model. The common words used in all types of posts were assigned a low entropy neutral value, resulting in a significant increase in information gain when labeling posts as neutral. The accuracy and recall results for many emotions were 0.0. One possible explanation is that posts were first filtered as neutral posts and then extended to only a subset of emotion starting from that node.

3. Team Member Contributions and Responsibilities

Contributions		
Qiantong Zhou	40081938	3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8
Hao Mei	40074373	2.1 2.2 2.3 2.4 2.5

Mingyu Gao	40049618	1.1 1.2 1.3 4.1 4.2 4.3