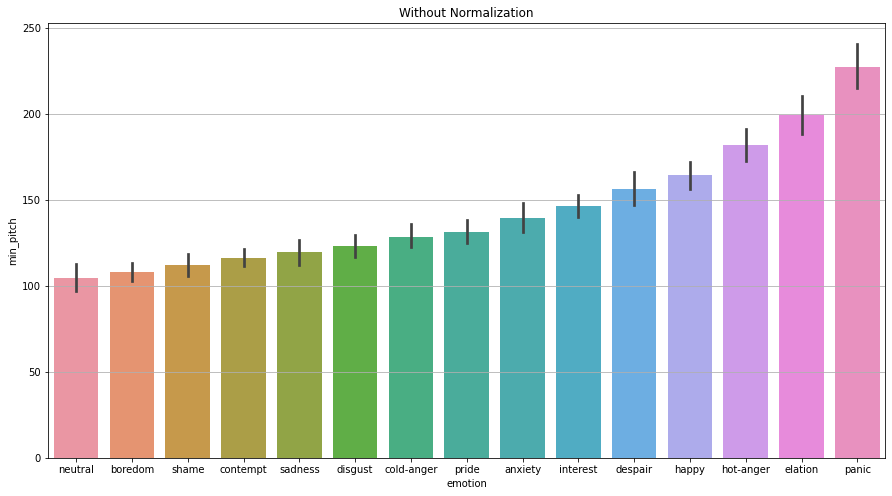
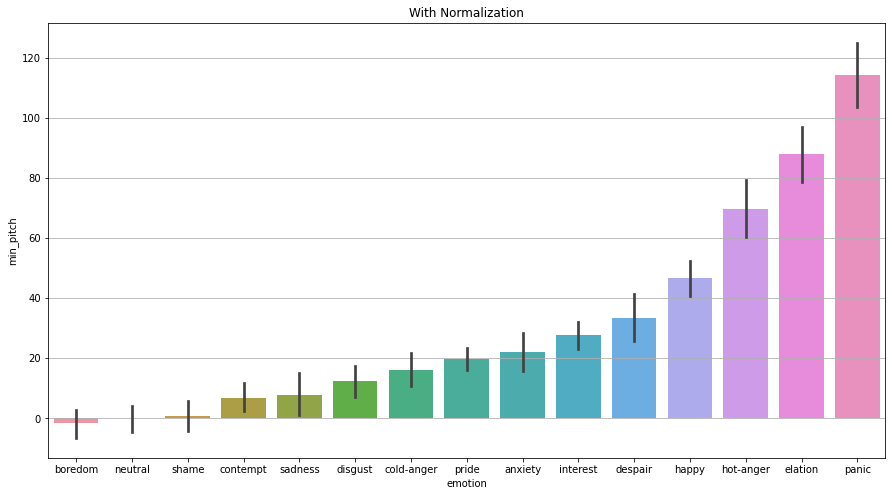
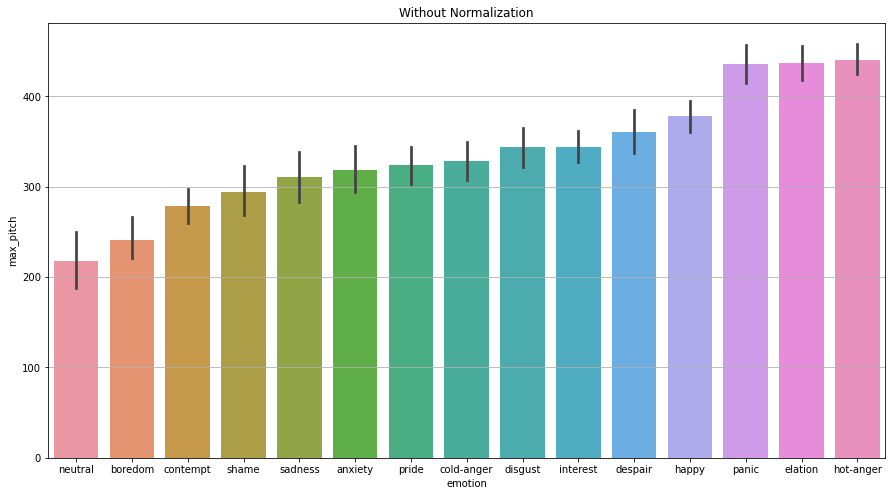
Homework 3: Emotion Recognition

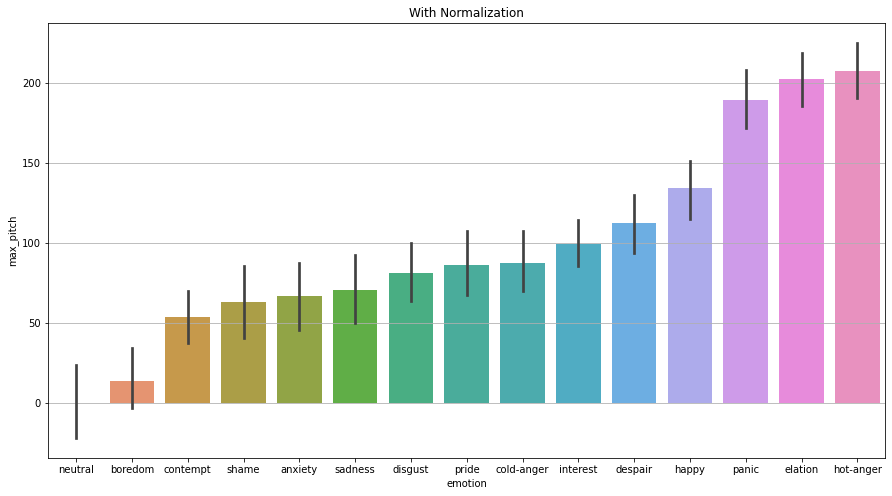
Kyuran Kang – kk3583

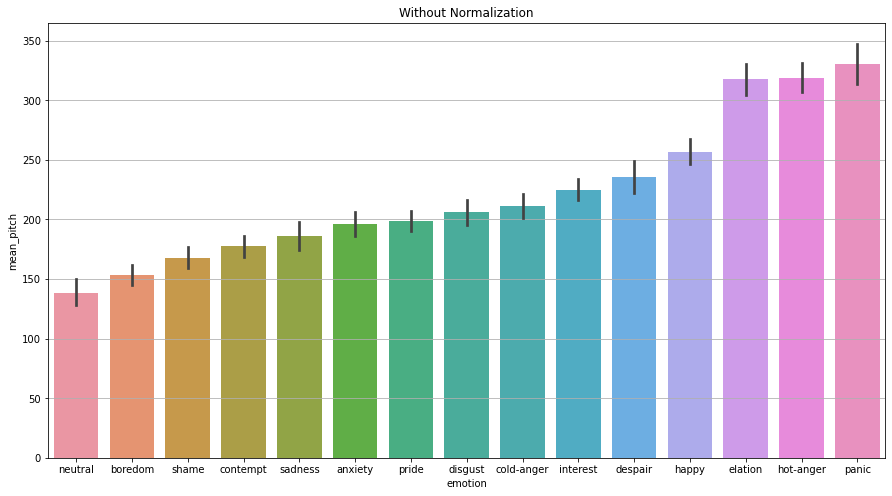
1. Feature Analysis

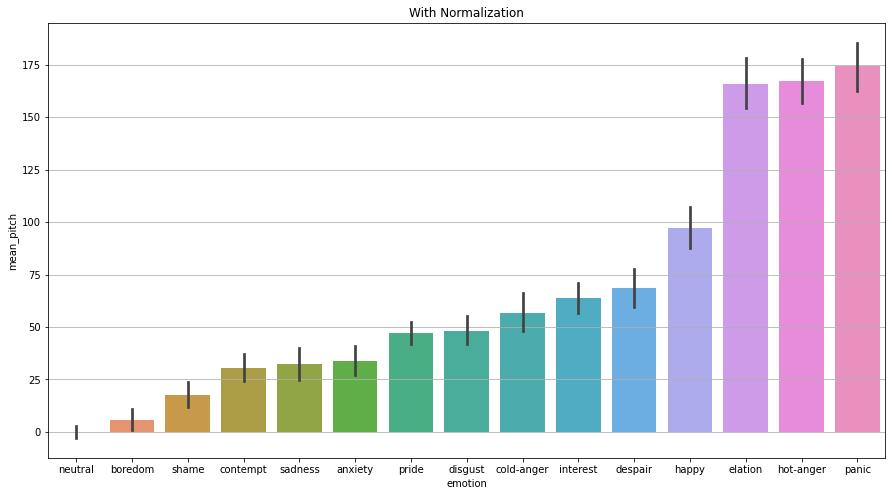


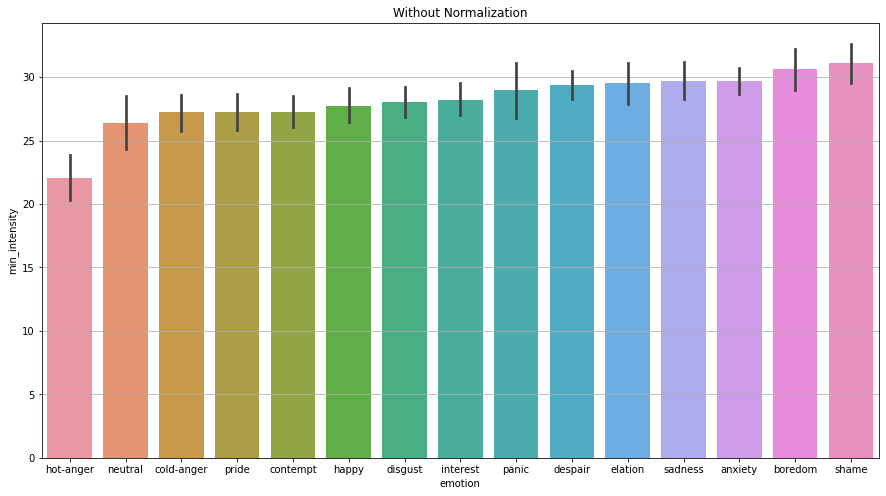


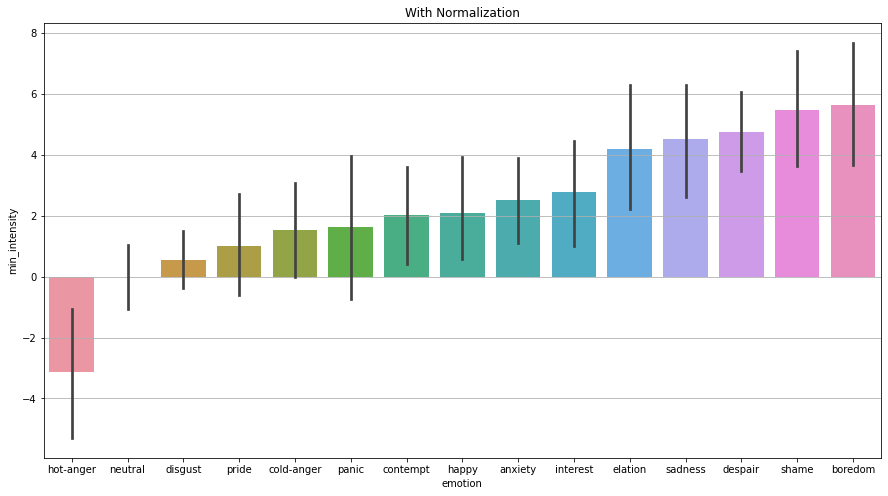


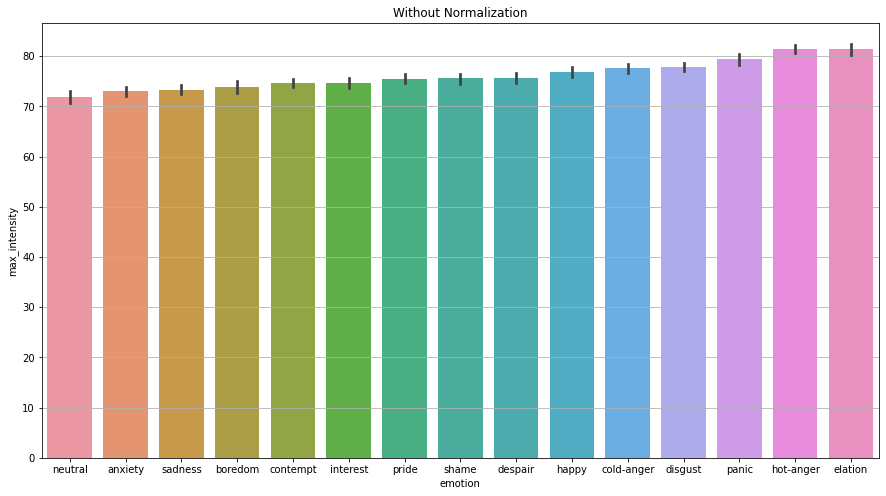


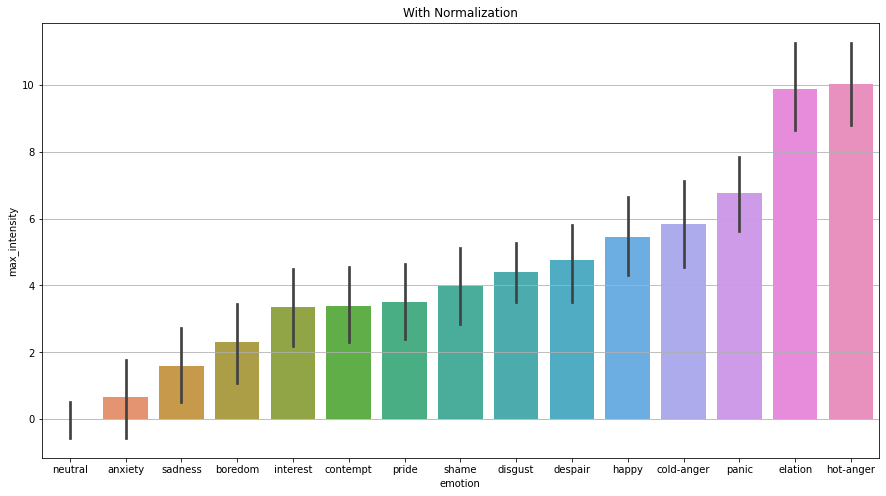


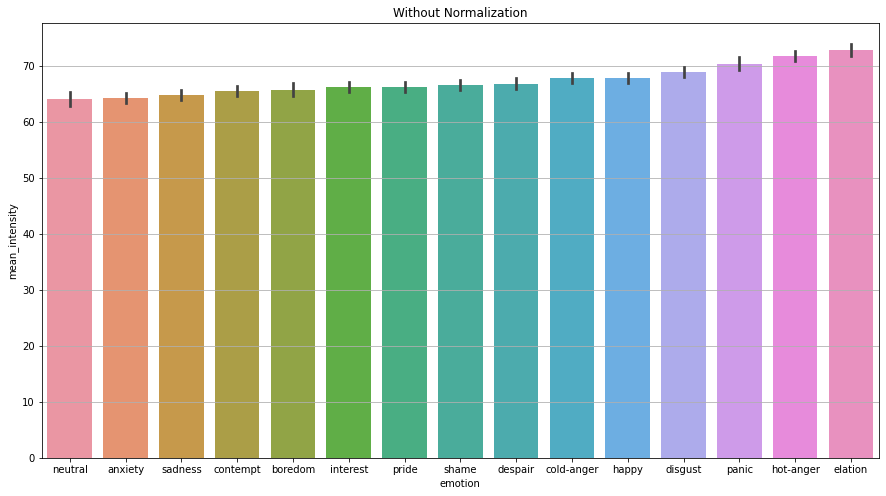


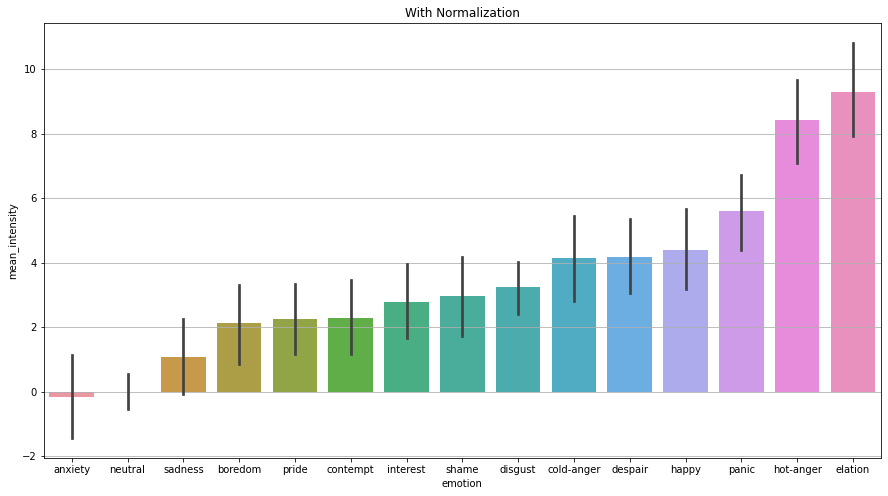






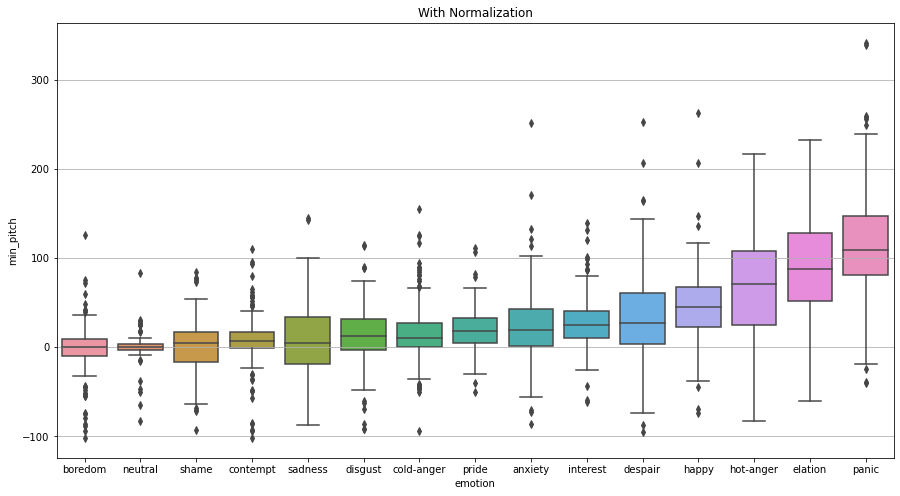
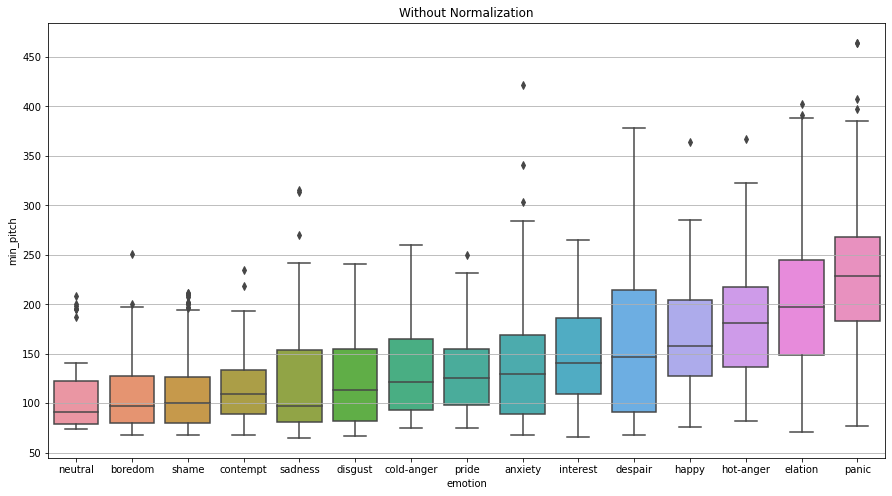


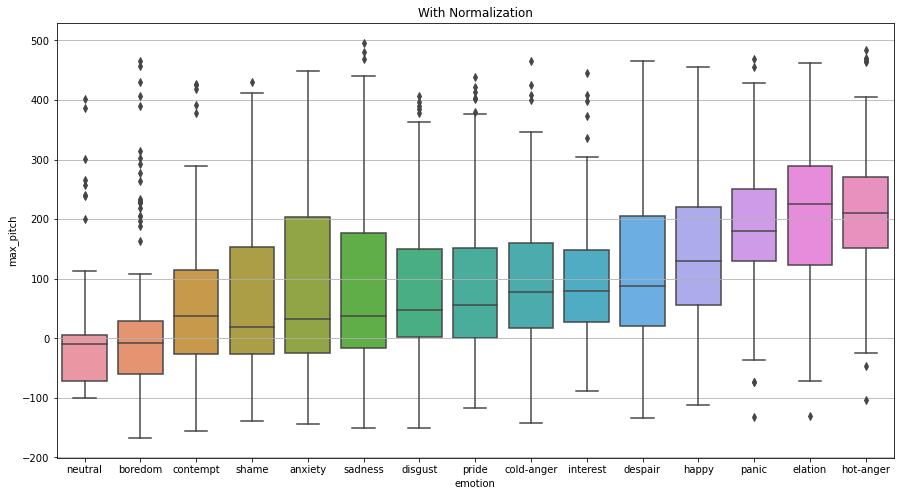
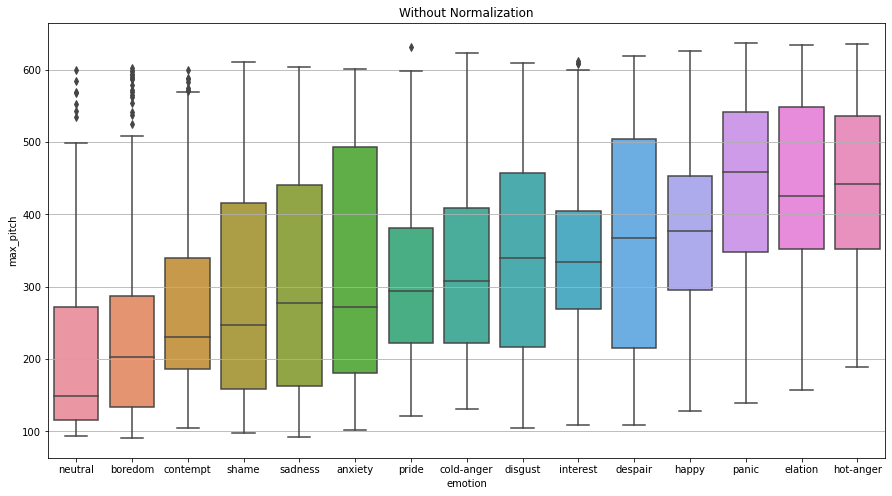


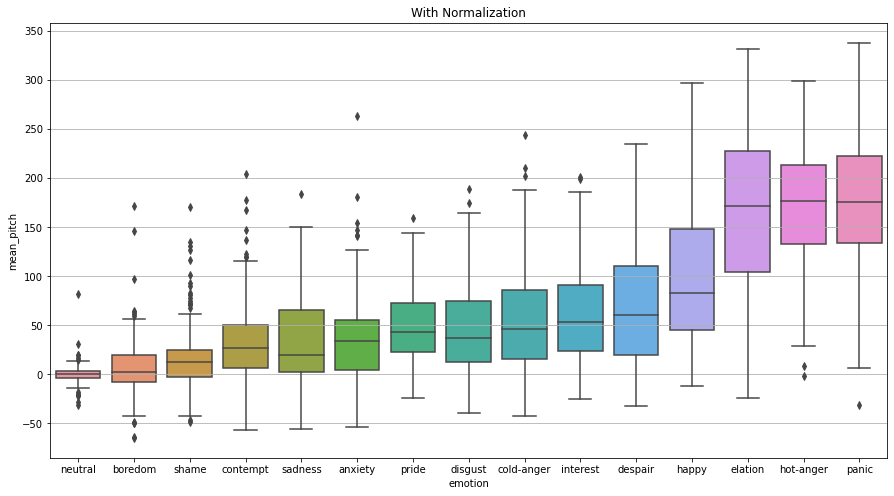
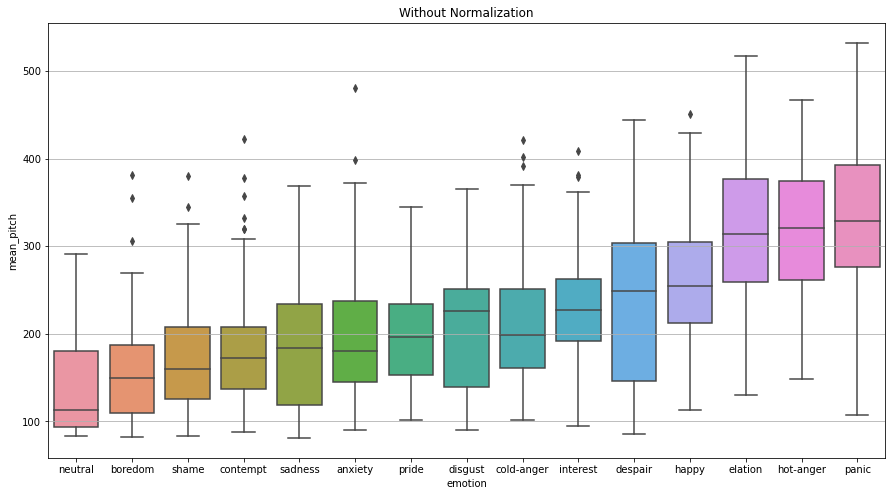


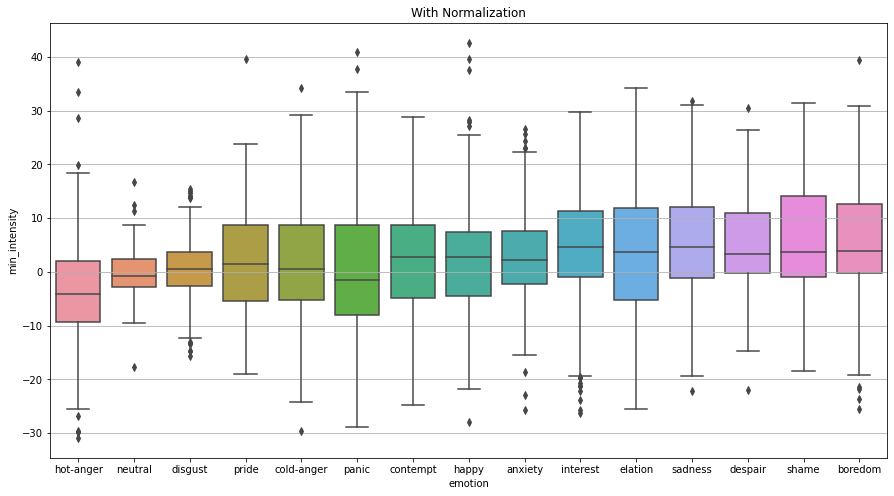
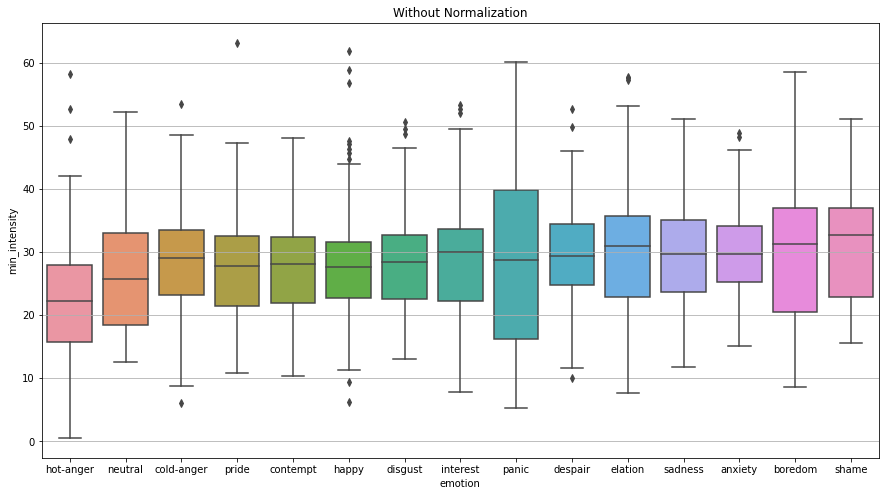
Above are plots of emotions (x-axis) vs. features(y-axis). The titles on top of the plots indicate if a plot is a) without normalization or b) with normalization. The normalization method used here is normalizing by the means of the individual speaker’s neutral utterance. One reason for choosing this method was that a speaker would have its own characteristics in utterance, and it is important to have a standard to normalize these features and neutral utterance can provide this baseline. Furthermore, speaker dependent normalization methods based on the neutral speech is commonly used and proven to be effective in speech emotion recognition tasks. Moreover, with respect to the efficiency, this method is less time consuming and computationally less expensive as we don’t have to extract all pitch segments again and concatenate them to calculate mean and standard deviation values to get normalized features.

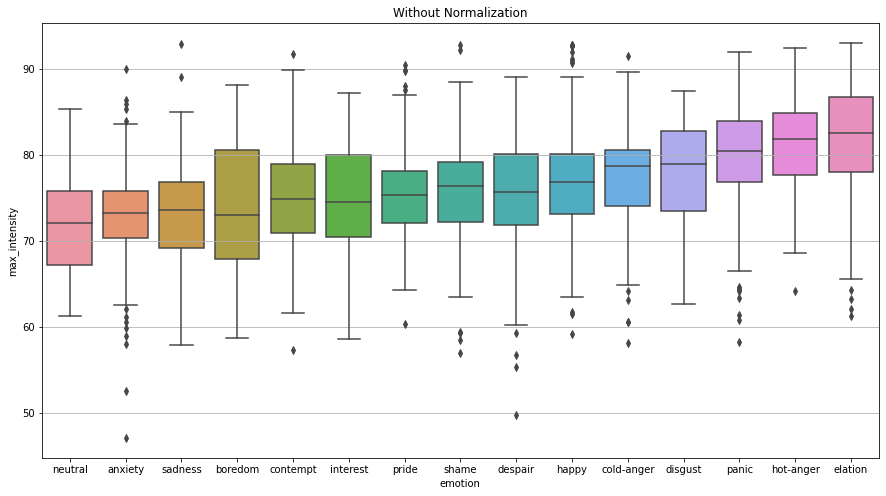
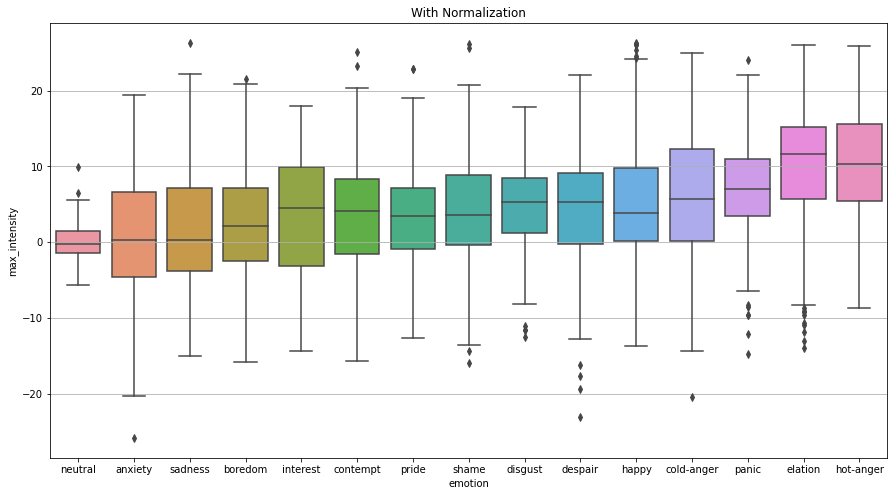
In addition to the above bar plots, I added box plots of the same data for analysis below.

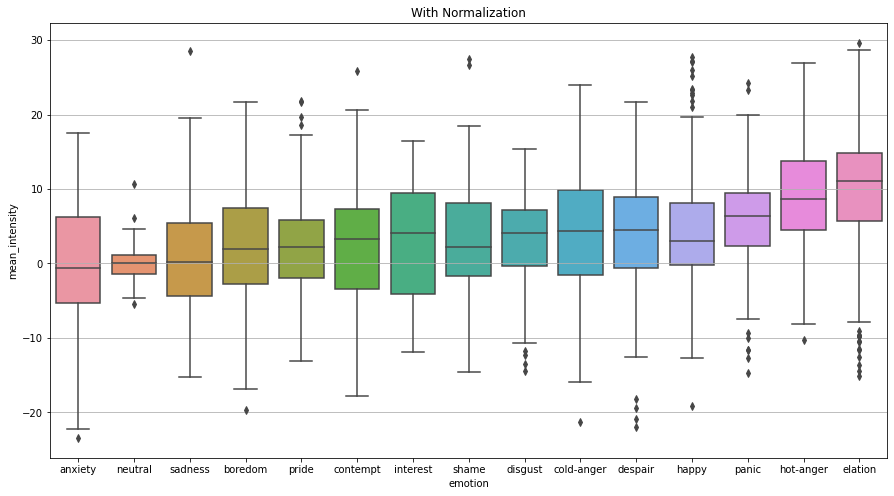
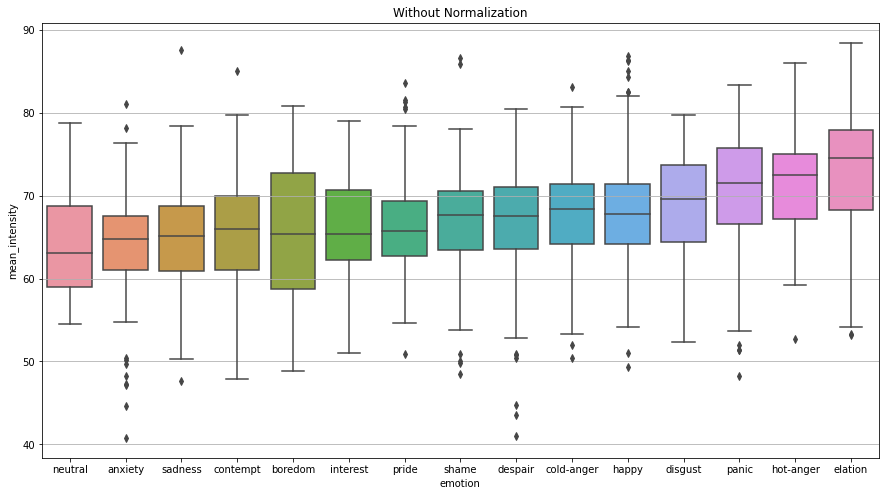










What I learned from the plots & interesting observations

* Panic, hot-anger, and elation have the highest average values in all pitch features (min, max, and mean pitch) in both with normalization and without normalization plots. These patterns can be expected as these emotions could be high-pitched to convey arousal. For instance, hot-anger utterances might be yelling-like which would make them high-pitched and panicking or elated voices could be heightened sounds.
* However, it is interesting that above emotions, which one would associate with intense voices, are not in the higher side of the average minimum intensity feature plot. Hot-anger has the lowest average minimum intensity in both with normalization and without normalization plots but has one of the highest average values in maximum intensity and mean intensity. This may indicate that the hot-anger speeches have huge variations in intensity throughout the utterance; possibly a hot-anger segment could start with lower intensity and end powerfully.
* Furthermore, we can see that the pitch features are similar between neutral and boredom emotions, but the boredom showed the highest average minimum intensity, which is very interesting as I imagined boredom speeches to have similar or even lower minimum intensity compared to that of neutral speeches. This distinction in minimum feature values may help the model (that will be built in the next section) differentiate boredom from neutral speeches.
* With normalization, it is easy to see how neutral, boredom and shame speeches have little variation in mean pitch values. On the other hand, disgust and neutral showed low variability in mean intensity.
* Another interesting point in mean intensity is that anxiety has the lowest average mean intensity and very high deviation. This may indicate that participants would express anxiety in different ways; some people might express anxiety with less power while others illustrate it intense tones.
* It might be challenging to identify and classify some emotions because we can see there are much overlap in the range of overall intensity features and pitch features. For instance, disgust, pride, cold-anger, interest, and despair seem to have many overlapping data points in mean pitch feature and mean intensity feature. Thus, utilizing other features to distinguish them may help to improve recognition task performance.

1. Classification Experiments

I used SVM Classifier as below:

*sklearn.svm.SVC(\*, C=1.0, kernel='rbf', degree=3, gamma=0.005, coef0=0.0, shrinking=True, probability=False, tol=0.001, cache\_size=200, class\_weight=None, verbose=False, max\_iter=-1, decision\_function\_shape='ovr', break\_ties=False, random\_state=None)*

In order to prevent overfitting, I only tried changing the gamma value (0.005 and 0.006) and got 0.005 to be the best one. I reason I chose SVM Classifier is that it works well with dataset with the number of samples lower than 100K but generally show better performance than Neighbor Classifiers. The results are as below.

1) Test Speaker: {'cc'} | Train Speakers: {'cl', 'mm', 'gg', 'jg', 'mk', 'mf'}

precision recall f1-score support

anxiety 0.10 0.20 0.13 10

boredom 0.00 0.00 0.00 15

cold-anger 0.00 0.00 0.00 15

contempt 0.25 0.14 0.18 22

despair 0.09 0.33 0.14 9

disgust 0.45 0.16 0.24 31

elation 0.25 0.88 0.39 16

happy 0.20 0.13 0.16 23

hot-anger 0.38 0.71 0.50 14

interest 0.07 0.12 0.09 17

neutral 0.94 0.94 0.94 18

panic 0.44 0.39 0.41 18

pride 0.14 0.04 0.07 23

sadness 0.29 0.15 0.20 13

shame 0.00 0.00 0.00 21

accuracy 0.26 265

macro avg 0.24 0.28 0.23 265

weighted avg 0.26 0.26 0.23 265

2) Test Speaker: {'cl'} | Train Speakers: {'mm', 'gg', 'jg', 'mk', 'cc', 'mf'}

precision recall f1-score support

anxiety 0.25 0.10 0.14 21

boredom 0.24 0.55 0.34 29

cold-anger 0.58 0.41 0.48 27

contempt 0.00 0.00 0.00 25

despair 0.00 0.00 0.00 29

disgust 0.21 0.64 0.31 22

elation 0.31 0.48 0.38 27

happy 0.35 0.29 0.32 21

hot-anger 0.48 0.38 0.43 26

interest 0.26 0.38 0.31 26

neutral 0.33 0.88 0.48 17

panic 0.00 0.00 0.00 21

pride 0.27 0.12 0.17 24

sadness 0.38 0.19 0.25 27

shame 0.12 0.04 0.06 26

accuracy 0.29 368

macro avg 0.25 0.30 0.24 368

weighted avg 0.25 0.29 0.24 368

3) Test Speaker: {'gg'} | Train Speakers: {'cl', 'mm', 'jg', 'mk', 'cc', 'mf'}

precision recall f1-score support

anxiety 0.25 0.43 0.32 30

boredom 0.67 0.07 0.12 30

cold-anger 0.20 0.11 0.14 27

contempt 0.37 0.27 0.31 26

despair 0.11 0.04 0.05 28

disgust 0.36 0.16 0.22 51

elation 0.33 0.89 0.48 28

happy 0.16 0.53 0.25 30

hot-anger 0.23 0.23 0.23 22

interest 0.50 0.03 0.06 30

neutral 0.89 0.89 0.89 9

panic 0.43 0.74 0.54 27

pride 0.22 0.08 0.12 25

sadness 0.00 0.00 0.00 33

shame 0.29 0.29 0.29 24

accuracy 0.28 420

macro avg 0.33 0.32 0.27 420

weighted avg 0.31 0.28 0.23 420

4) Test Speaker: {'jg'} | Train Speakers: {'cl', 'mm', 'gg', 'mk', 'cc', 'mf'}

precision recall f1-score support

anxiety 0.23 0.16 0.19 19

boredom 0.12 0.79 0.21 14

cold-anger 0.00 0.00 0.00 22

contempt 0.00 0.00 0.00 23

despair 0.00 0.00 0.00 21

disgust 0.38 0.52 0.44 23

elation 0.25 0.05 0.08 20

happy 0.00 0.00 0.00 20

hot-anger 0.56 0.28 0.37 18

interest 0.00 0.00 0.00 19

neutral 0.73 1.00 0.84 8

panic 0.00 0.00 0.00 14

pride 0.10 0.44 0.17 18

sadness 0.00 0.00 0.00 19

shame 0.11 0.07 0.08 15

accuracy 0.18 273

macro avg 0.17 0.22 0.16 273

weighted avg 0.14 0.18 0.13 273

5) Test Speaker: {'mf'} | Train Speakers: {'cl', 'mm', 'gg', 'jg', 'mk', 'cc'}

precision recall f1-score support

anxiety 0.38 0.36 0.37 22

boredom 0.35 0.52 0.42 27

cold-anger 0.24 0.40 0.30 20

contempt 0.57 0.30 0.39 44

despair 0.00 0.00 0.00 16

disgust 0.05 1.00 0.09 1

elation 0.06 0.04 0.05 26

happy 0.12 0.13 0.12 23

hot-anger 0.34 0.48 0.40 21

interest 0.15 0.37 0.22 19

neutral 1.00 1.00 1.00 10

panic 0.41 0.75 0.53 12

pride 1.00 0.06 0.11 18

sadness 0.50 0.05 0.09 20

shame 0.29 0.10 0.15 20

accuracy 0.29 299

macro avg 0.36 0.37 0.28 299

weighted avg 0.37 0.29 0.27 299

6) Test Speaker: {'mk'} | Train Speakers: {'cl', 'mm', 'gg', 'jg', 'cc', 'mf'}

precision recall f1-score support

anxiety 0.04 0.03 0.04 29

boredom 0.08 0.05 0.06 20

cold-anger 0.09 0.09 0.09 23

contempt 0.07 0.05 0.06 21

despair 0.20 0.25 0.22 53

disgust 0.08 0.29 0.13 21

elation 0.42 0.61 0.50 23

happy 0.31 0.10 0.15 42

hot-anger 0.18 0.09 0.12 22

interest 0.27 0.25 0.26 44

neutral 1.00 1.00 1.00 8

panic 0.00 0.00 0.00 21

pride 0.08 0.04 0.06 23

sadness 0.14 0.27 0.19 22

shame 0.13 0.12 0.12 25

accuracy 0.18 397

macro avg 0.21 0.22 0.20 397

weighted avg 0.18 0.18 0.17 397

7) Test Speaker: {'mm'} | Train Speakers: {'cl', 'gg', 'jg', 'mk', 'cc', 'mf'}

precision recall f1-score support

anxiety 0.56 0.23 0.33 39

boredom 0.21 0.53 0.30 19

cold-anger 0.00 0.00 0.00 20

contempt 0.11 0.42 0.17 19

despair 0.08 0.17 0.11 18

disgust 0.40 0.09 0.14 23

elation 0.08 0.05 0.06 19

happy 0.20 0.39 0.26 18

hot-anger 1.00 0.06 0.12 16

interest 0.33 0.05 0.08 21

neutral 1.00 0.89 0.94 9

panic 0.57 0.14 0.23 28

pride 0.00 0.00 0.00 19

sadness 0.09 0.18 0.12 17

shame 0.28 0.29 0.29 17

accuracy 0.21 302

macro avg 0.33 0.23 0.21 302

weighted avg 0.32 0.21 0.19 302

MODEL: SVC(gamma=0.005)

**Aggregated Average Accuracy: 0.24311531841652323**

**Aggregated Average F1-Score(weighted): 0.21172854646889172**

Aggregated Average F1-Score(macro): 0.22937300292714202

1. Error Analysis

* Best Performing

Model: svm.SVC(gamma=0.005)

Test Speaker: {'mf'} | Train Speakers: {'cl', 'mm', 'gg', 'jg', 'mk', 'cc'}

precision recall f1-score support

anxiety 0.38 0.36 0.37 22

boredom 0.35 0.52 0.42 27

cold-anger 0.24 0.40 0.30 20

contempt 0.57 0.30 0.39 44

despair 0.00 0.00 0.00 16

disgust 0.05 1.00 0.09 1

elation 0.06 0.04 0.05 26

happy 0.12 0.13 0.12 23

hot-anger 0.34 0.48 0.40 21

interest 0.15 0.37 0.22 19

neutral 1.00 1.00 1.00 10

panic 0.41 0.75 0.53 12

pride 1.00 0.06 0.11 18

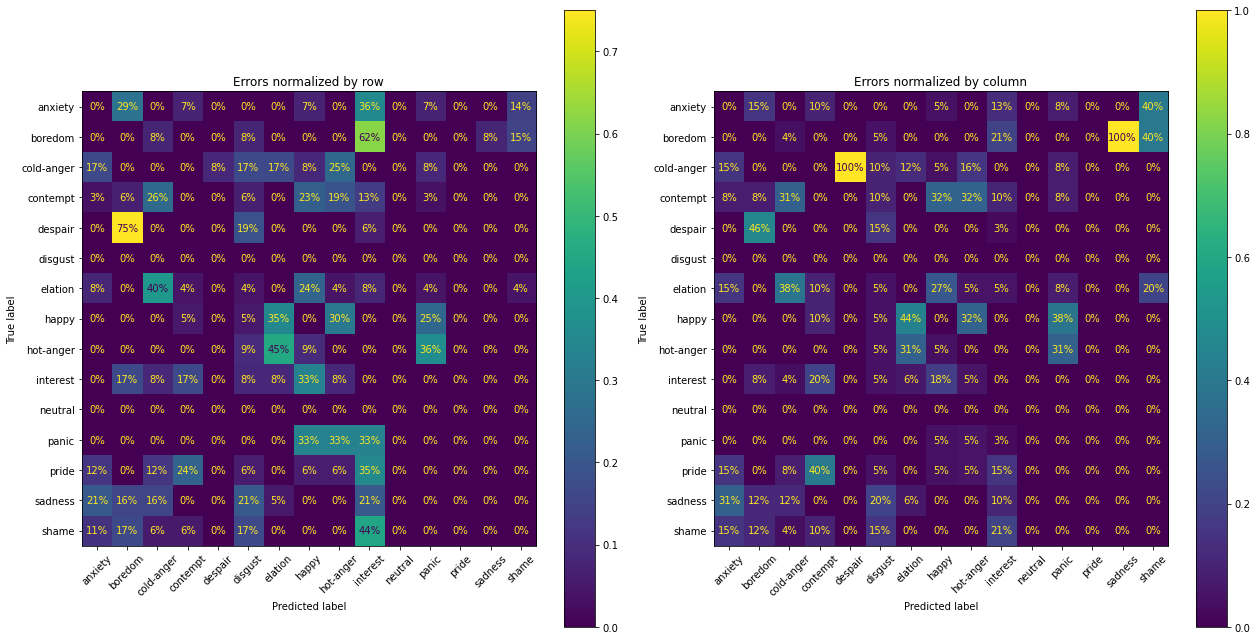
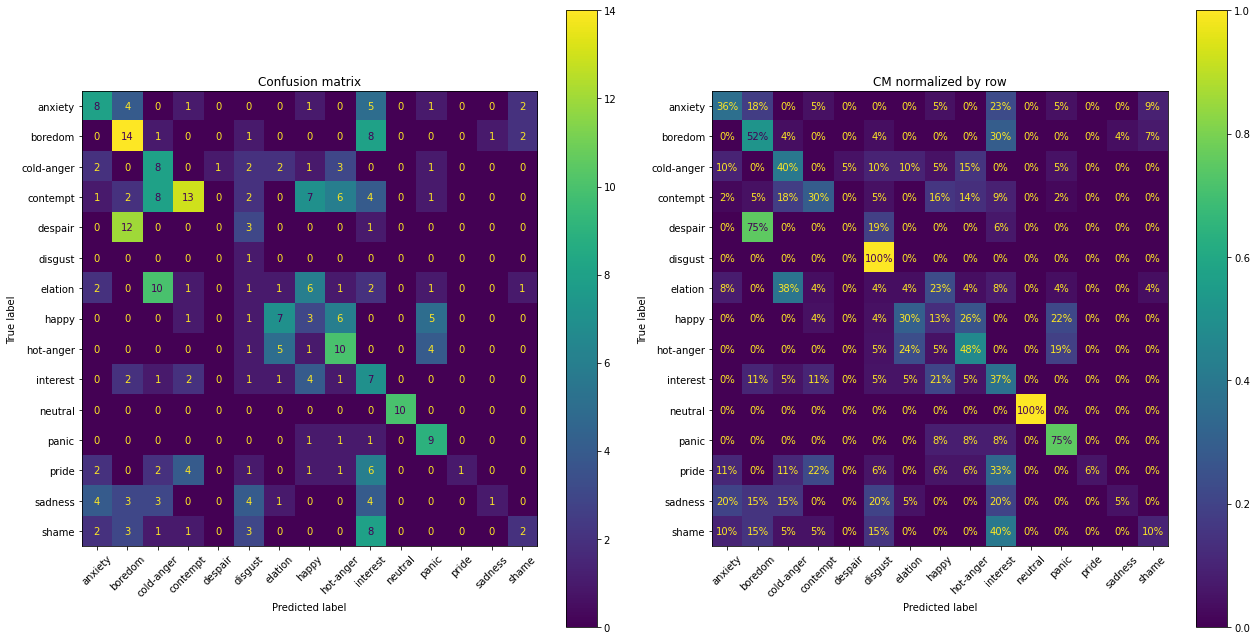
sadness 0.50 0.05 0.09 20

shame 0.29 0.10 0.15 20

accuracy 0.29 299

macro avg 0.36 0.37 0.28 299

weighted avg 0.37 0.29 0.27 299



* Which class(es) were easiest to predict? Why do you think they were easy?

Class with the highest f1-score was neutral, followed by panic, and boredom. This would be because neutral speech segments are usually monotonal have low variation among speakers which would be easily distinguishable from other emotional voices. Also, we can say that the neutral speeches of the speaker "mf" didn't have outlying data points that could have lied outside of the boundary defined by the model trained with rest of the speakers' data. 'boredom' would also be distinguishable from other emotional speeches in a similar sense but have difference in features that made the model to separate 'neutral' from 'boredom'. Strong, aroused emotions would be reflected in the 'panic' speeches, making the model to relatively predict it well. Furthermore, we can assume these emotions were easier to be conveyed by the participants; possibly participants had similar ideas on expressing these emotions in speech.

* Which were the most difficult? Why do you think they were difficult?

The hardest one was despair, followed elation, disgust, sadness, and pride. Elation was misclassified as cold-anger and happy. We can imagine that elation speeches have similar traits with those of cold-anger and happy speeches. 12 out of 16 disgust segments were classified as boredom. It is possible that speaker mf's expression of despair had similar traits with the rest of the speakers' boredom. Possibly mf's expression of despair was low voiced or lacked power in speech. Speaker mf had only one disgust segment, while it was correctly classified, some non-disgust segments were incorrectly predicted as disgust, making the f1-score low for disgust. We can imagine that these emotions are hard to be expressed in such short segments of speech, and possibly people show higher variety in conveying these emotions in speech.

* Based on this analysis, what ideas do you have to further improve your classifier?

First, trying different methods of normalization would help. While normalizing based on the speaker's neutral speech is a commonly used method, experimenting other normalization would help; possibly combine it with standardization. Second, feature selection process would improve the quality of this problem as noisy features may hinder the performance. Moreover, we can try deleting outliers to make the data more general and use over- or under- sampling methods to balance the datasets. Lastly, improving the quality and the size of the dataset would be very helpful. Some emotions seem to be very difficult to express in short segments without any context; making a dataset by providing the participants with appropriate text contexts can be helpful while we would have to be careful in providing these contexts as some texts may induce people to speak in a certain way, making the model trained based on these less general.