Part A

For the first 50 tweets we had to annotate in round 1, the inter-annotator agreement (Cohen-Kappa score) was about 0.164. For the second 25 tweets in round 2 it was about 0.269. This is significantly higher than the first batch of annotations, although both scores are not great. We think the score not being very high is because our way of interpreting the text is quite different. The increase we see is because after calculating the golden standard, and revising the guidelines with each other, we got a look at the others' way of thinking. When talking about the golden standard for the first round of annotations we also talked about the annotation guidelines, and we really tried to understand why the other annotated the tweet the way they did. We were quite thorough here. This made it so that when we got to the revision of the guidelines for round 2, we were quite on the same line on the guidelines. This helped when annotating the next batch of tweets, but it did not suddenly make us think about the tweets in the same way. We think this different way of interpreting certain tweets prevented us from obtaining an even higher score for the second round.

Part B

This is the classification report:

	precision	recall	f1-score	support
Anger	0.215	0.979	0.353	
AnnotatorNotes	0.000	0.000	0.000	1
Anticipation	0.000	0.000	0.000	25
Caring	0.000	0.000	0.000	23
Disgust	0.000	0.000	0.000	54
Fear	0.000	0.000	0.000	31
Joy	0.000	0.000	0.000	31
None	0.409	0.087	0.144	103
Other	0.000	0.000	0.000	12
Sadness	0.000	0.000	0.000	25
Surprise	0.000	0.000	0.000	29
Trust	0.000	0.000	0.000	32
accuracy		0.225	463	
macro avg	0.052	0.089	0.041	463
weighted avg	0.136	0.225	0.106	463

This is the confusion matrix:

```
[[95  0  0  0  0  0  0  2  0  0  0  0]
 [ 0  0  0  0  0  0  0  1  0  0  0  0]
 [23  0  0  0  0  0  0  2  0  0  0  0]
 [23  0  0  0  0  0  0  0  0  0  0  0]
 [52  0  0  0  0  0  0  2  0  0  0  0]
 [31  0  0  0  0  0  0  0  0  0  0  0]
 [30  0  0  0  0  0  0  1  0  0  0  0]
 [94  0  0  0  0  0  0  9  0  0  0  0]
 [12  0  0  0  0  0  0  0  0  0  0  0]
 [24  0  0  0  0  0  0  1  0  0  0  0]
 [29  0  0  0  0  0  0  0  0  0  0  0]
 [28  0  0  0  0  0  0  4  0  0  0  0]]
```

From this we can conclude that the trained program was only really familiar with the anger and none tags, and therefore only assigned those to the unknown tweets during the testing phase. This is probably because the annotated tweets contained considerably more angry tweets than other emotions.

Part C

We agree with the 'one truth' myth. In annotation there is not always one truth, one correct answer or one true interpretation. Some text elements can be ambiguous, and the words can mean multiple things or radiate multiple feelings. While annotating we often came across tweets that brought up different emotions in us. These were often small differences like whether someone was annoyed or showed dislike towards a certain subject. Often we would have multiple of these occurrences, and one would stick with the 'Anger', and the other one with 'Disgust'.

We also agree with the 'disagreement is bad' is a myth statement. Because of disagreement it is necessary to talk with the annotators about the difference and the ambiguous text elements. We also often disagreed with the annotations and because of the disagreement we looked at the tweet in more detail and found a better label for the text element.

We think detailed guidelines can help the annotators and prevent a big difference in the annotations. It can be beneficial to make the guidelines more detailed. This way the annotators know when to use a label. For example when we annotated the tweets we could only choose from a certain set of emotions, these emotions were quite well described. This way the difference between emotions that were easily confused were more clear, like 'Anger' and 'Disgust'. Whenever we did not find an emotion in a tweet we did not assign one, because this was in the guidelines. This was very useful. Even though there were a lot of useful tips in there for annotating, remembering them all while annotating was sometimes confusing, and we might forget a thing or two. We helped each other out here while creating the golden standard, which fixed everything.

We agree with myth five: 'experts are better' as well. You do not always need experts to do your annotations, even when you might think you do. In cases mentioned in the paper, it was not really beneficial to use experts. However, if you need to annotate text that contains a lot of language specific to their area of expertise, we think you do need experts to do the annotation. We made annotations for tweets with the topic of Covid-19 in the Netherlands, while we're not experts in this field. So this was an example of a case where this is possible.

We agree that 'One is Enough' statement is a myth. With the annotations for machine learning, two people made annotations for the text elements. This had a lot of benefits. One perspective is not always enough, some cases cannot be properly annotated by only one annotator. Especially when it comes to ambigious tweets.

We also think the 'once done, forever valid' statement is definitely a myth. Annotations are not always valid in different points of time. Just like the meaning of the Covid tweets might change as more discoveries are done about the virus.