Felix Bühler
Max Wegge
Carlotta Quensel

Corpus Creation

Emotion Analysis – Assignment 1

Data Selection

For the task of annotating emotions, we considered different amazon review corpora and several reddit threads. The amazon reviews were discarded as they mostly contained sentiments rather than emotions. A possible solution of filtering out 'neutral' reviews with a medium number of stars would distort the observable emotions. Thus, the distribution in the corpus would not contain all possible emotions, which makes the corpus unsuitable for an exemplary emotion annotation. This also led to the elimination of most reddit threads, as their domain often primed the users for a specific subset of emotions. We selected the reddit thread *Redditors who live with their SOs, what changed?* for our annotation. The topic of this thread prompts users to describe emotional experiences and the question for *change* also includes negative experiences. The data was scraped from `https://www.reddit.com/r/AskReddit/comments/5nrh7b/redditors_who_live_with_their_sos_what_changed/` and processed further using a python script. We removed deleted comments and users, subsequent conversations and bot answers, thus leaving the corpus with only comments directly answering to the question and containing around 500 documents. From those, the first 100 instances where used for the annotation task.

Annotation

The annotation categories are modelled after Plutchik's wheel for its eight base categories and intensity ratings are suitable for our specific data set. For the topic of *change*, the emotions of surprise and anticipation were needed and the feature of forming 'emotion-combinations' (like love) from neighboring categories open up possibilities to further study the corpus. After annotating the first 10 instances, we chose to disregard Plutchik's intensity categories in favor of unnamed intensity scores. We also encountered the problem of latent emotions implied by the corpus domain, which is why we had to specify annotation rules for those specific categories. The resulting annotation guidelines contain the following information:

Annotation Guidelines

The annotation categories are Plutchik's eight base emotions of joy, sadness, trust, disgust, fear, anger, surprise and anticipation. The annotated value between 0 and 3 measures the intensity of the emotion label. The annotation environment has a document per line and eight columns for the emotion categories, thus every cell should contain a value. The categories are sorted as above with opposing emotions next to each other. Contrary to Plutchik's wheel is the annotation of opposing emotions in one document. This is possible due to the multi-label annotation and the perspective of the annotation.
As the data is related to the event of 'moving in together', most documents will likely contain event descriptions or event appraisals. Therefore, the annotator should assume the emotion felt in the described situation rather than at time of writing. This also includes comical exaggerations building to a punchline, which are annotated without regards to their intended readings. As the documents address living together with a romantic partner, joy and trust are likely to be implicit in many texts. Therefore, trust is in this context defined as an explicit referral to a positive relation to another person. This definition includes the negation, i.e. negative emotions due to the absence of another person but not statements that do not refer to a second person in any way.

Evaluation

To evaluate our annotations, we used Fleiss' Kappa to account for more than two annotators. We initially used three different splits for our evaluation with differing results (fig 1). The most natural split for emotion annotation might be the 'on/off-split', where annotations between 1 and 3 denote the intensity levels of an annotation and 0 denotes the lack of the emotion. This split resulted in an overall lower score than the split between intensities 1 and 2, hereafter referred to as 'half-split'. This means, that our agreement on high or low intensities was stronger than the agreement on the presence or absence of an emotion. While this seems counter-intuitive in the context of emotion annotation, it is a mathematically likely outcome as the half-split allows for more intra-intensity deviation. The κ-scores for comparing all four intensity values are the lowest, as we expected.

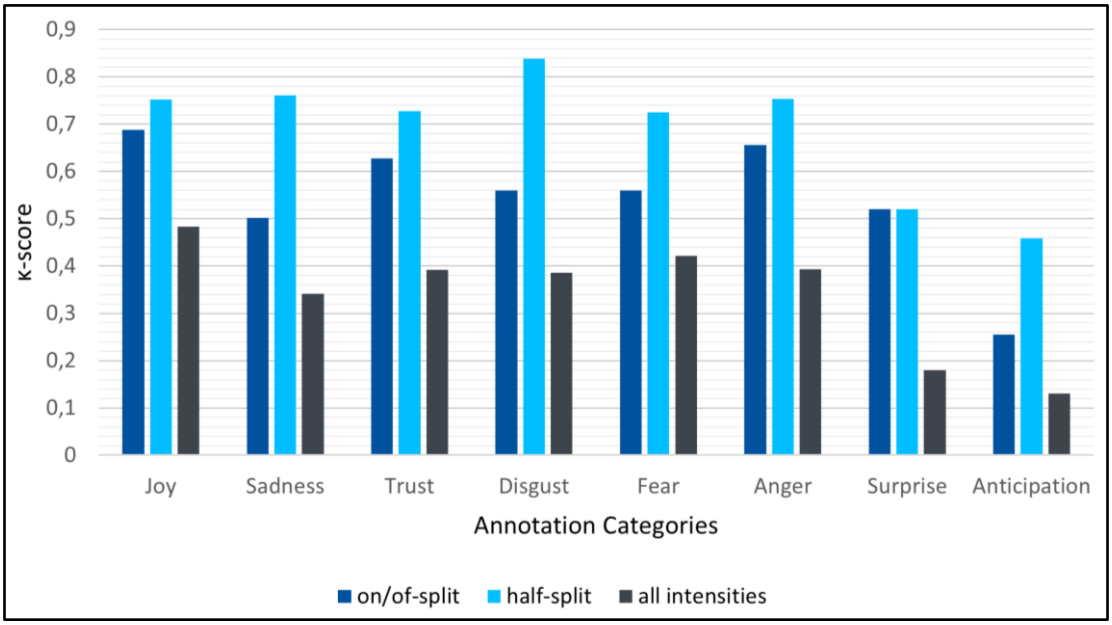

Fig 1 Fleiss' Kappa scores for all emotion categories with three different splits.

As the half-split corresponds to Plutchik (the base emotion matches the intensity score of 2) and results in the highest κ -scores, we chose it as the basis for all further evaluation.

Inter-annotator agreement The highest κ-score of 0,838 belongs to disgust while the lowest score of 0,458 belongs to anticipation. A possible explanation could be the difference in vagueness for these emotions in the data source. While anticipation can often occur in the domain of *change* (*what changed?*) and might therefore be annotated implicitly, disgust is not implicit in the corpus domain. The agreement for joy, trust and anger is also high. These emotions seem to prevail in the domain of our data source (relationships and living together).
The average number of annotations differs greatly between annotators, with an overall average of 1,22 annotations per instance while the lowest average is 0,87 and the highest average is 1,56. All annotators agree on the most frequently used tags of surprise and joy, which matches their above stated prevalence in the data source.

Corpus statistics Through computing an average annotation for each instance, we can observe the number of co-occurrences for all emotions (fig 2). These numbers are obtained by calculating the average intensity for each emotion of the instance and again using the half-split (now with rounded averages) as a binary measure for emotion annotation. The resulting average annotations are then checked for co-occurrences and summed up in the following table.

	Joy	Sadness	Trust	Disgust	Fear	Anger	Surprise	Anticipation
Anticipation	1	0	1	0	0	1	0	**2**
Surprise	7	1	2	1	0	3	**18**	
Anger	1	2	1	1	0	**18**		
Fear	0	0	1	1	**2**			
Disgust	0	0	0	**5**				
Trust	6	1	**16**					
Sadness	1	**5**						
Joy	**24**							

Fig 2 Number of co-annotations given the average annotation for each document instance and using half-split

This table allows us to make use of the combination of neighboring emotions in Plutchik's wheel. We can observe 6 instances of love as the co-annotation of joy and trust. Beyond the combinations named by Plutchik, we can also observe interesting co-occurrences like joy-surprise or surprise-anger.

Qualitative evaluation The low κ-score for the four way split between all intensities implies that annotators rarely agree on intensity scores. This is most prevalent in long texts, as those instances are likely to contain a variety of emotions and intensities. As the intensity score seems highly subjective in our annotation, the longer a document, the more room for interpretation and the lower the agreement between annotators.

Irony and humorous exaggeration pose further difficulties for the annotators. Even though our annotation guidelines demand a 'literal' annotation without regard to intended readings, this requirement leads to annotations that do not accurately portray the emotions depicted in the instances. Lastly, instances with objective event descriptions might not include any emotions at all.

Conclusion

Our chosen data set has a non-normal distribution of emotions due to its domain, which favors annotations of joy, surprise and anger. This also increased the difficulty of distinguishing implicit and explicit emotions and the overall difficulty annotating these emotions. The task showed us that the annotation of emotion categories and intensities is highly subjective. Despite using guidelines, there remain ambiguous cases and debatable annotations.