Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task

Emily Öhman[1][0000−0003−1363−7361]

University of Helsinki, Finland
emily.ohman@helsinki.fi

Abstract. With the prevalence of machine learning in natural language processing and other fields, an increasing number of crowd-sourced data sets are created and published. However, very little has been written about the annotation process from the point of view of the annotators. This pilot study aims to help fill the gap and provide insights into how to maximize the quality of the annotation output of crowd-sourced annotations with a focus on fine-grained sentence-level sentiment and emotion annotation from the annotators point of view.

Keywords: Emotion Annotation, Sentiment Analysis, Annotator Experience.

1 Introduction

Often annotation tasks are viewed as being somewhat monotonous, repetitive, and even plain boring [1], yet very little has been written about annotator motivation, and even less about how to increase it. Motivation is typically achieved by offering annotators some kind of reward in exchange for their effort. This reward can be extrinsic (monetary profit, course credits, or author credits) or intrinsic (e.g. by using gamification to make the annotation task more 'fun', altruism, or a desire to learn) [7, 9–11, 19, 23]. This paper aims to provide guidelines to help minimize the repetitiveness, maximize the quality of annotations, and reduce the need for expensive extrinsic rewards.

In the particular project discussed in this paper the reward was course credits and the annotation task was sentiment and emotion annotation using a minimally gamified online platform (see section 3). Students were asked to annotate a number of sentences into different sentiment and emotion categories and then provide feedback on their thoughts on the annotation process. The student motivation for the task was thought to be mainly extrinsic, but many mentioned intrinsic motivations in their feedback (fun, interest, learning opportunity, gaining understanding of machine learning and sentiment analysis).

Annotator motivation is different for different types of tasks requiring different levels and types of expertise. The different motivations also affect annotation quality and availability of a worker base [23]. As for the setup, the motivation of the annotator affects the cost of the annotation, and the setup effort. Gamification is a fairly simple method of achieving low extrinsic cost annotations with high motivation.

The next section introduces previous work, especially with regards to gamification. Section 3 briefly explains the data, method, and tools used in both the annotation project, and the feedback collection from annotators. In section 4 the annotators' feedback comments are summarized and the most common topics and themes are examined more carefully. This section is followed by a concluding discussion.

2 Background & Previous Work

Difficulties in the annotation task are often discussed in academic papers related to annotation or a project where annotation has been used [2, 13, 21]. Annotation quality as well as inter-annotator agreement (e.g. kappa scores [6]) are often discussed as an evaluation metric of annotations and annotators. However, it is rare to see anyone write about annotator motivation (exceptions are [18] and [23]). However, annotator motivation is one of the key aspects determining the cost of annotations [23]. Gamification (e.g. Zooniverse [9], Games with a purpose [22]) is often a good way to increase annotator motivation and reduce annotation cost.

Gamification is the use of game elements in environments that are not typically games [7, 11]. Previous studies show that one can achieve a high number of quality annotations by non-experts by using carefully considered gamified aspects such as (1) Relatedness (connected to other players), (2) Competence (mastering the game problems), and (3) Autonomy (control of own life) [14].

Robson et al. [17] posit that gamification can change behavior by tapping into motivational drivers of human behavior: reinforcements and emotions. The emotion we are looking to elicit are of course enjoyment, but negative emotions such as disappointment can also increase commitment and a desire to increase one's competence [15]. Similarly, both positive and negative reinforcements increase repetitive behavior in players [17], and annotation is generally a repetitive task, even if gamified, unless the gameplay style changes regularly in-game.

Another commonly used crowdsourcing tool is Amazon's Mechanical Turk [5]. This tool allows for many different types of crowdsourced help, where the motivation is in the form of monetary compensation. Sometimes and for certain types of tasks, this might be the best option, however, it requires funding that can be difficult to obtain for some researchers, particularly junior researchers not affiliated with projects and PhD students.

Many different emotion annotation schemes and procedures have been used on different crowdsourcing platforms. The different schemes have different strengths and weaknesses [4], however, most use some variation of Ekman's [8] 6 basic emotion categories.

3 Data, Method, and Tools

The data that was to be annotated came from four random sections of the OPUS Movie Subtitle parallel corpus [12]. This enabled the annotators to annotate the same texts, but in their chosen language (Finnish, Swedish, and English were given as options).

The individual data components the annotators were asked to annotate, were subtitle lines, which generally means sentences, but can be single words or 2-3 short phrases as well. The data was purposely presented context free in order to maximize the usability of the annotated data set for different applications outside the movie subtitle genre and to avoid weighting any emotions doubly [2].

We wanted to create a simple, easy to use platform with little to no need for tutorials and other guidance. Therefore, very few annotation guidelines were provided, however, the students were told to annotate the expressed emotion, i.e. from the perspective of the speaker/writer. The platform works so that the annotator logs in with their individual username and password (in this case their student email address and student id number). They then choose the language with which they wish to annotate, and are shown a short introduction to Plutchik's wheel [16]. The annotations start once they click "Begin Annotation". After this point they get a sentence to annotate using an interactive interface (see figure 1). Once they click save, their points are updated and they are presented with the next sentences. They can at any point view a more detailed outline of their own scoring.

Several aspects have been found to influence inter-annotator-agreement, chief among them "annotation domain, number of categories in a coding scheme, number of annotators in a project, whether annotators received training, the intensity of annotator the method used for the calculation of percentage agreements" [3]. Although annotator training increases agreement, with trained annotators reaching average agreements of 81%, and untrained annotators of only 70%, it has also been shown that high-intensity training leads to significantly better agreement rates [3]. It is not surprising that the more effort you put into training, the more uniform the annotations are, but the more training that is put in, the higher the cost of annotations and the more unfeasible the procedure becomes to implement on crowdsourcing tasks, especially on gamified platforms. This is why *Sentimentator* uses an automated filtering of noisy annotations and noise-creating annotators with the help of a self-perpetuating gold standard based on seed annotations and rank assigned based on what is essentially a confidence score [15]. Theoretically this should enable similar results to highly trained annotators, without the added time and cost required for training. Further comparison of different annotation procedures will be conducted in the future.

The platform used in this project, *Sentimentator*, was, as mentioned, lightly gamified. What this means in practice is that the user interface was designed with a game-like feel, the user can always see their score which reflects the number of annotations they've completed, and the user can see an overview of their annotations. Future enhancements to the interface, include a more complex scoring system, leaderboards, intermissions between annotations, ranking of players, as well as simple cosmetic improvements such as avatars. The physical annotation method is also changing from clicking on buttons and dragging sliders, to directly clicking on Plutchik's wheel.

All the students who annotated were enrolled in the course *Introduction to Language Technology*. The sentiment and emotion annotation task was the final assignment before the final paper. They were to report that they had done the assignment by leaving a short comment on *Moodle*. Their contribution was not graded in any way; the task was strictly pass/fail regardless of quality or number of annotations. In the end the annotators

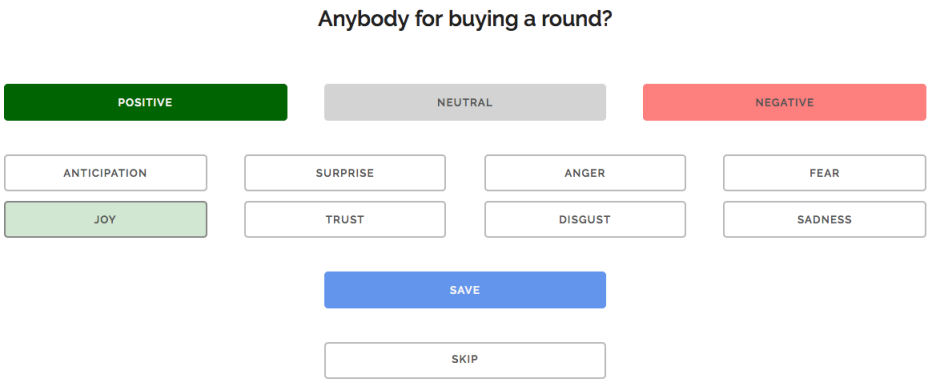

Fig. 1. Screenshot of annotation interface.

managed 56 247 annotated sentences, which is roughly 500 sentences per annotator. This data will be made available on the project's GitHub page[1] once evaluated.

The student feedback was evaluated by downloading and anonymizing the comments on *Moodle* and then manually reading through the comments and dividing them into loose categories based on how the annotator described the task. This approach was at this stage of qualitative investigation mainly to recognize repeating themes, more than to acquire quantitative data. For quantitative analysis, a stricter categorization scheme would be necessary [20].

Ideally, the annotations would have been analyzed and compared to find differences in annotation quality among annotators with respect to their output, time spent, and perceived motivation. In future work, the annotations will be analyzed quantitatively as well as qualitatively to pry apart these different factors and understand how motivation affects quality of annotations through these.

4 The Annotator Experience

4.1 The Annotation Task

The annotators were tasked with annotating 1000 sentences (lines), but were told that if it took them more than two hours, they were free to stop and simply write a few words on what it was that made it difficult for them to annotate 1000 sentences in the given time. There were no negative consequences for the students for stopping at a lower number of sentences and in fact, the number of sentences they annotated was not checked.

We were hoping to collect as many comments on the annotation process as possible, and thus were not surprised when only a handful of students managed to annotate 1 000 sentences in one hour, and only a few more students managed to do it in the span of two

[1] https://github.com/Helsinki-NLP/sentimentator

hours. Indeed, this somewhat unreasonable request is probably what netted us so much feedback, and specifically feedback on the time it took the annotators to annotate each sentence. The feedback might express more *frustration* due to this. It should be noted that a handful of students managed to annotate the requested number of sentences within the allotted time without the quality seemingly suffering.

The annotators were provided with access to a tutorial video that showed how the annotation tool worked and how to login, as well as explained how to annotate. We emphasized that there was no right or wrong annotation, but instead to go with their gut feeling. Unfortunately, not all of the students watched the video or read the instructions, which was very clear from some of the feedback.

4.2 Annotator Feedback

The annotator feedback can be difficult to summarize, but a look at the wordcloud in figure 2 helps give a quick overview of the most commonly mentioned topics.

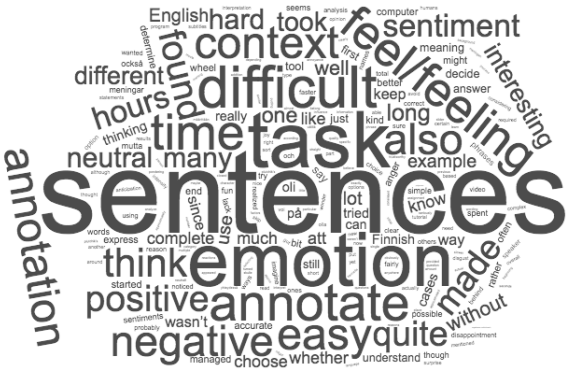

Fig. 2. Wordcloud of annotation feedback.

The student feedback had several common themes. The general consensus seemed to be that the task itself was interesting, even though many felt that the monotonous nature of annotation made it boring after some time. Others, on the other hand, felt that the task was boring at first, but after they go the hang of it, it became more enjoyable. Many of the annotators described the task in a positive way and commented that it made them think about the difficulties of analyzing sentiments and emotions in text.

"It was nice to learn about sentimental annotation and to know about this kind of "game". It was fun to try to think what kind of emotions a sentence could have."

As many annotators commented on the task as a learning experience, but some seemed to have missed the point of the annotation task as a course assignment almost completely, it is clear that (1) the students are very much so driven by the need for a purpose, and (2) it might be wise to more overtly inform the student of the learning goals of the task. Nonetheless, the fact that so many found the task somewhat boring

after a while, suggests that there needs to be more interactive elements or feedback to the annotators during the annotation process to make the task more engaging in the long run as well. This should theoretically increase the quality of the annotations further.

"Because honestly, it's a mindnumbingly boring task."

Interpretation and *perspective* were two other words that were used to describe the difficulty of the task. The annotators wanted to know if they should annotate as a third party observer, as the recipient of the line or the speaker of the line. See figure 2 for a wordcloud based on all the feedback comments. Many annotators explained their reasoning or thoughts in the feedback:

"Jag tycker också att det var enklare att välja att något är neutralt än positivt eller negativt, jag hade också svårt att veta om en mening som var neutralt formulerad, men innehöll t.ex. ett ledsamt budskap, skulle klassificeras som negativ eller neutral."

Translation: *I also think that it was easier to pick that something is neutral than positive or negative, I also had difficulties knowing if a sentence that was neutrally phrased, but, for example, contained a sad message, should be classified as negative or neutral.*

Some comments were quite insightful:

"Negative emotions, especially anger, tend to be less subtle than positive ones as they're not the norm, so they are easier to identify from a contextless sentence than, for example, slight joy or trust."

"Punctuation marks: period, question and exclamation marks defined by 30% in my opinion the classification by sentiments."

Although we were aware of most of the issues the annotators mentioned, we had not yet considered the impact of punctuation. It might be interesting to do a comparison annotation test where all punctuation marks have been removed. However, this of course further removes the sentence from being bound to any context and makes annotation even more difficult. Many annotators found it difficult to not imagine a context within which the line was from. Indeed, context and perspective, as well as whether the annotators found the task interesting or tedious, were some of the most common comments in the feedback.

"I think it was difficult for me to think about them outside of any context, and I often found myself thinking about the contexts that one would say these sentences in, and I'd have to consciously try to not think about the contexts. It was difficult for me to determine whether a sentence was neutral or not because I wasn't quite sure if I was still thinking about it in some sort of context even though I tried not to."

"I found the task rather difficult because it was hard to focus only on the sentence itself and not create any kind of context behind it."

"I kept thinking whether these emotions should be felt from the perspective of the person or the character saying the sentence."

Although it was not checked via *Sentimentator*, nor required in the feedback, many students included information on how long they took to complete the assignment and/or how many lines they annotated. The amount of time the students took per annotation varied wildly from up to a minute to only a few seconds. Some annotators felt they got better and faster annotating after a slow start, whereas others felt like the task got harder

the more they had annotated as they said they started to second-guess themselves and their interpretation more.

"I found the task reasonably easy. It took me about 1,5 hours."

"The 1000 sentences sentiment annotation took me about 2 hours."

"After having spent 2.5 hours annotating the sentences with the result of only 282 sentences, I realized that I had put way too much effort to annotating each sentence."

"I annotated 950 sentences, which took me quite a long time (almost 3 hours, but I still wanted to keep going, because I found the task interesting)."

5 Concluding Discussion

It seems from the number of annotator comments that the first step in making the annotation task less repetitive and easier for annotators would be to add context to the sentence being annotated. Although it was a conscious choice to disregard context, at the very least a comparison between classification accuracy between context-free and context-dependent annotations should be performed to assess whether the increased accuracy of context-dependent annotations indeed does increase the overall accuracy as well, or if it makes the data less generally usable by tying sentences to a given context and thus making it more genre-specific.

It is difficult to address annotator complaints that stem from them not reading the instructions. Although in this case the annotators were students motivated by the fact that this was a compulsory part of their coursework, it seems likely that other annotators too might give up the task if they do not understand it from the get go. Here we touch upon one of the core issues with annotation: how much should annotators be trained? The scientific literature seems to suggest that this is something that changes with the type of annotation task. Some types of annotation tasks require more training, whereas quality suffers in other types when annotators are over-trained.

As the annotators in this case were students of language technology, the way in which they viewed the task might differ from the general public, and again from someone paid to annotate. In short, the specific extrinsic reward received from completing the task, might have an impact of varying degrees on how the annotator views the task of annotation and how motivated they are to annotate consistently to produce high-quality annotations. In any case, further improvements to the gamification aspects of the platform should generate more motivated annotators and annotations of a higher quality. For other types of annotation where gamification is perhaps not an option adding some other type of intrinsic reward system might be a solution (Wang [23] suggests *language learning* as one possible intrinsic reward).

Further comparisons between different annotation types and changes to the platform should be performed in the future to determine the most beneficial setup for (1) increased long-term interest and thus number of annotations, and (2) high quality annotations. This type of setup would also act as a control group for the gamified platform and better help reveal what factors affect motivation. Furthermore, this qualitative-only pilot study was conducted on a very specific subset, i.e. class assignments, where the motivation is quite different from usual crowdsourcing tasks. Therefore, it would be

beneficial to compare output and feedback from the students and more typical annotators (crowdsource).

Another aspect that is crucial in a more in-depth examination of annotator experience, is a quantitative evaluation of the annotation output. That is: Who produces the highest quality output? What is their motivation? Do they take longer to annotate one unit or are they perhaps faster? Does time of day matter? Do annotators get better after, e.g., 20 minutes of annotating, or do they perhaps get worse? Do they get better after 1 000 annotations or does everything start to seem like the same emotion or no emotion? These and many other questions will have an answer when the results are evaluated quantitatively and subsequently published for public use.

The annotator comments could very well be further analyzed and the qualitative data (i.e. annotator comments) more closely linked to the quantifiable data, like for example, average time taken to annotate one sentence and subsequent relation to accuracy and how that relates to how the annotator experienced the task. Does annotating mainly negative data result in more negative comments? This data has much further use in improving annotation platforms beyond simply exploring the annotator experience.

References

1. Acero, A., Wang, Y.Y., Wong, L.: Computer-aided natural language annotation (Jul 19 2011), US Patent 7,983,901
2. Andreevskaia, A., Bergler, S.: Clac and clac-nb: Knowledge-based and corpus-based approaches to sentiment tagging. In: Proceedings of the 4th International Workshop on Semantic Evaluations. pp. 117–120. SemEval '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007).
3. Bayerl, P.S., Paul, K.I.: What determines inter-coder agreement in manual annotations? a meta-analytic investigation. Computational Linguistics **37**(4), 699–725 (2011).
4. Bostan, L.A.M., Klinger, R.: An analysis of annotated corpora for emotion classification in text. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 2104–2119 (2018).
5. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? Perspectives on psychological science **6**(1), 3–5 (2011)
6. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement **20**(1), 37 (1960).
7. Deterding, S., Sicart, M., Nacke, L., O'Hara, K., Dixon, D.: Gamification. Using game-design elements in non-gaming contexts. In: CHI'11 extended abstracts on human factors in com-puting systems. pp. 2425–2428. ACM (2011).
8. Ekman, P.: Universals and cultural differences in facial expressions of emotion. In: Nebraska symposium on motivation. University of Nebraska Press (1971).
9. Greenhill, A., Holmes, K., Lintott, C., Simmons, B., Masters, K., Cox, J., Graham, G.: Playing with science: Gamised aspects of gamification found on the online citizen science project-zooniverse. In: GAMEON'2014. EUROSIS (2014).
10. Groh, F.: Gamification: State of the art definition and utilization. Institute of Media Informatics Ulm University **39** (2012).
11. Hamari, J., Koivisto, J.: Social motivations to use gamification: An empirical study of gamifying exercise. In: ECIS. p. 105 (2013).
12. Lison, P., Tiedemann, J.: Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles (2016).

13. Mohammad, S.: A practical guide to sentiment annotation: Challenges and solutions. In: WASSA@ NAACL-HLT. pp. 174–179 (2016)
14. Musat, C.C., Ghasemi, A., Faltings, B.: Sentiment analysis using a novel human computation game. In: Proceedings of the 3rd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources and Their Applications to NLP. pp. 1–9. Association for Computational Linguistics, Stroudsburg, PA, USA (2012)
15. Öhman, E.S., Tiedemann, J., Honkela, T.U., Kajava, K., et al.: Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics (2018)
16. Plutchik, R.: A general psychoevolutionary theory of emotion. Theories of emotion **1**, 3–31 (1980)
17. Robson, K., Plangger, K., Kietzmann, J.H., McCarthy, I., Pitt, L.: Is it all a game? understanding the principles of gamification. Business Horizons **58**(4), 411 – 420 (2015), http://www.sciencedirect.com/ science/article/pii/S000768131500035X
18. Sabou, M., Bontcheva, K., Derczynski, L., Scharl, A.: Corpus annotation through crowdsourcing: Towards best practice guidelines. In: LREC. pp. 859–866 (2014)
19. Schell, J.: The pleasure revolution: Why games will lead the way, googletechtalks std. november 2011 (2015)
20. Srnka, K.J., Koeszegi, S.T.: From words to numbers: how to transform qualitative data into meaningful quantitative results. Schmalenbach Business Review **59**(1), 29–57 (2007)
21. Tokuhisa, R., Inui, K., Matsumoto, Y.: Emotion classification using massive examples extracted from the web. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. pp. 881–888. Association for Computational Linguistics (2008)
22. Von Ahn, L.: Games with a purpose. Computer **39**(6), 92–94 (2006)
23. Wang, A., Hoang, C.D.V., Kan, M.Y.: Perspectives on crowdsourcing annotations for natural language processing. Language resources and evaluation **47**(1), 9–31 (2013)