Annotator Survey

Comparing Emoji and Textual Tasks

All the annotators found the experience of working on the Hatemoji project very positive (see Fig. 1). All the annotators had previously worked on Dynabench projects with textual hate speech. To assess the relative difficulty of the emoji annotation rounds, they were asked to compare project experiences. 56% of annotators found the level of difficulty was comparable, 33% found the emoji rounds less challenging and 11% found them more challenging (see Fig. 1).

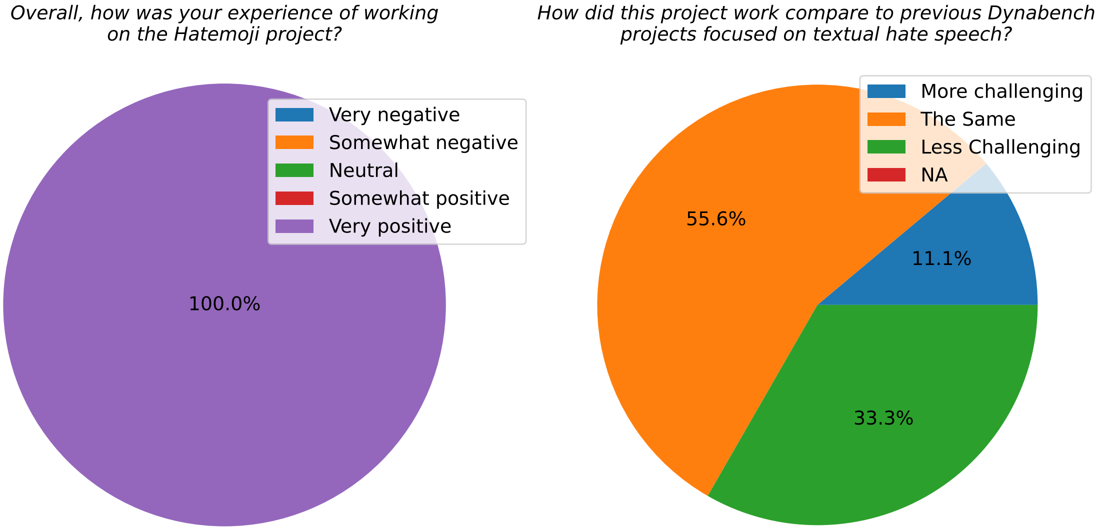

Figure 1: Experiences working on the Hatemoji project.

When asked what challenges the emoji data collection rounds posed, annotators predominately spoke of the ambiguity in interpreting emoji, the creativity required to use emoji which the model wasn't already wise to and some technical difficulties in searching for and inputting emoji. Anonymized excerpts from annotators' comments are provided below.

Ambiguity in Emoji

- **A2**: "The challenge was mainly understanding what everyone meant when choosing an emoji as they can be interpreted in many different ways - e.g. the man with the turban could be Arab, muslim etc."
- **A3**: "It was also often more difficult to ascertain someone's intended meaning behind a statement when validating due to the variety of ways one can interpret the emojis."
- **A4**: "Entries that only contained emojis and no text were a bit confusing sometimes as the exact message the annotator was trying to express was not 100% clear."

- **A5**: “Emoji-based hate was more challenging as you had ensure that people would actually get what was trying to be said and wasn’t too vague.”
- **A6**: “I think emojis are more limited and more ambiguous than words.”

Creativity of Selection

- **A1**: “It was different as you had to look into emojis that you may not use as the common one were being overused.”
- **A2**: “It was different because a statement can be hateful purely by the inclusion of the emoji, so you had to really think creatively when validating and inputting.”
- **A9**: “It became challenging very quick in a short period of time, which is positive as the bot was getting better at spotting hate/non hate.”

Technical Difficulties

- **A3**: “It took longer to construct examples, with having to search for the emojis etc.”
- **A8**: “The copying and pasting of the emojis slowed down the rounds more than previous projects but manageable within the set time limits.”
- **A9**: “A lot of the time you needed to combine the two text and emoji.”

Experiences Per Round

The quality of examples fell in the final round, specifically with regards to being unrealistic and more ambiguous. To understand this pattern, annotators were asked to rank the rounds by level of difficulty with the majority finding the final round was the most challenging (see Fig. 2). They were then asked to explain why some rounds were more challenging than others. Their answers revealed several factors at play. The model improved throughout the rounds so became wise to simple strategies of emoji-based hate. However, this improvement was relative to annotators also gaining more skills and becoming settled into the task. Finally, annotators found the later rounds harder because of errors and mistakes in the inputted examples, which made interpreting and labeling these inputs more challenging. Excerpts from the annotators are provided below.

Model Improving

- **A1**: “No doubt as the rounds went on the model was getting better at figuring out our strategies.”
- **A4**: “Round 5 was the least challenging one because it was really easy to fool the model as DynaBench was not familiar with emojis. In Round 6 the model knew the emojis and it was really difficult to enter content that could fool it.”

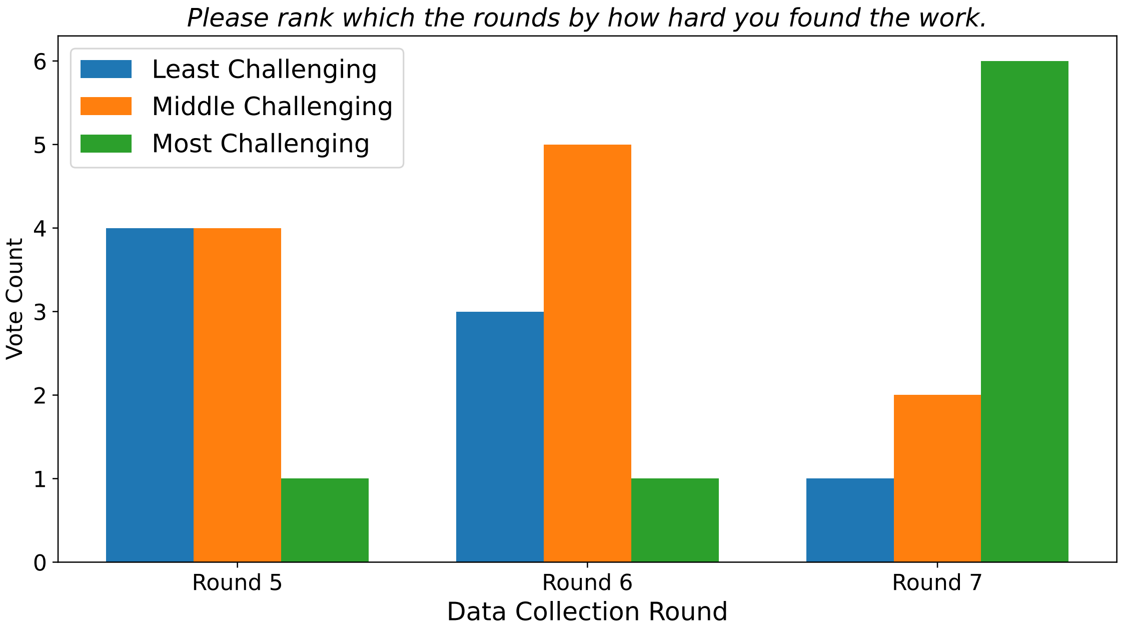

Figure 2: Ranked difficulty of progressive rounds.

- **A5**: "I think the fact that the dynabench system could really interpret examples so well it added pressure to try and trick the system more."
- **A9**: "The bot advanced quite quickly as there's not many emojis, harder to create slurs/not slurs to trick the bot."

Model Improving Relative to Annotator Experience

- **A2**: "Round 6 was the least as I'd got used to the project and the model was still relatively easy to trick/ I still had lots of ideas."
- **A6**: "The learning curve of round 5 made it somewhat of a challenge. By round 6 I had a pretty good grasp of what was expected and so found it the easiest."

Model Improving Relative to Quality of Examples

- **A2**: "Round 7 was the hardest as my originality was wearing thin and I wanted to keep including high quality entries but this was hard."
- **A3**: "The earlier rounds were easier due to the models' naivety, peoples' examples were also much more explicit making validations and pertrubations easier, these both increased in difficulty throughout."
- **A4**: "By the third iteration, most examples had already been exhausted and so unusual words and implausible statements were used to produce novelty. I think this is where mistakes crept in - which were often not corrected until

the final round of validations. The latter stages of the process were time-consuming because it was necessary to correct mistakes produced in the first stage of producing examples."

- **A6**: "Round 7 was the most difficult as it was the round with the most errors from other annotators and the model became harder to trick."

Model Fooling Strategies

One benefit of human-and-model-in-the-loop training is added explainability in which strategies fool the model. Successful strategies included identity swaps, polarity swaps, slur homonyms, sarcasm and complex statements:

Identity Swaps

- **A2**: "After a while the model gets better at knowing if something is a non identity so saying 'kill all [hamster emoji]' would work in the 5th round but maybe not the 7th."

- **A3**: "I had a lot of early success using flags to represent protected national identities. When the model was then trained on these it was then easy to exploit its somewhat oversensitive nature and use these flags to trick it in the other direction."

- **A4**: "Substituting an identity for an emoji that represented that identity."

- **A6**: "The model seemed to be overly sensitive. For example, a statement like 'let's kill [them] all' in reference to a rival sports team would trick the model into thinking it is a hateful statement."

- **A7**: "I realised that the model didn't recognise ethnic food and could be easily tricked into thinking statements relating to it were hateful. e.g 'I could murder a Chinese *takeaway emoji*' 'I f*cking hated that Mexican restaurant'."

Polarity Swaps

- **A3**: "Overall I found tricking it into thinking hate when it was actually not-hate seemed a lot easier, especially as the model became hyper-aware of the commonly used hateful emojis. Simply swapping out the emotive word in a sentence (such that without the word the sentence lacks any explicit positive or negative skew) for an emoji was also a consistently successful tactic."

- **A4**: "Inputing neutral statements discussing identities and adding less-used emojis."

- **A8**: "Neutral or non hateful statements with a black person emoji would fool the model."