SFWRENG 4NL3
Project Step 3 - Annotation Task Report

Sumanya Gulati

March 2025

1 Description of Task

The task consisted of labelling the messages sent by users right before or while playing the DOTA game. The conversations were broken down into games which in my understanding corresponded to a single group session within DOTA. The annotator was asked to label each message in a game's conversation with a number from the range 0-7 categorizing the messages based on their content and apparent sentiment.

2 Interesting Aspects

Despite the short length of each message (usually 2-4 words each), given the context of the entire conversation, I could infer a lot more information about the players and their relationships with each other than I expected. It was easy to guess which group of players in a session were acquaintances based on the way they interacted with each other or by the number of *casual* messages exchanged between them.

3 Insights and Challenges

I am unsure about what the other datasets look like but if they are similar to Part 4, I believe the model might not be trained over diversified content. This is mostly because out of all the messages I annotated, a majority belonged to this set of words - 'ez', 'gg', 'hahaha', 'lol' or their variations. For the model to generate actual, meaningful results, it must be trained on more expressive test messages that are representative of all the 7 labels.

Additionally, I think the preprocessing of the data will require a lot more nuance because a majority of the text messages had typos in them. This presents two options - either disregard any message with a typo in it which would reduce the dataset by a considerable amount or, label them based on their inferred context but have a robust preprocessing system set up. Even if the team chooses the latter option, I am not sure if a sufficiently broad enough system can be developed to accurately identify the typos and correct them unless done manually.

4 Mental Model

Very quickly into the task, I came to the realization that most messages were either 'ez', 'gg' or some variation of that which would be labelled as *positive*. I also put anything with a curse word directly into the *verbal abuse* category which further simplified the task. I had chosen to label test-based emojis as *non-English* and all variations of 'lol', 'lmao', 'haha' as *miscellaneous*. Since about three-quarters of the messages fell neatly into one of these categories, it was pretty easy to annotate them and I made my way through the dataset in half the allocated time.

5 Ambiguity in Instructions

The instructions did not specify the order of priority for messages that could be labelled as two or more of the given categories. For example, if an otherwise positive message also contains the word 'haha' or some other random sequence of characters, should it be labelled as *positive* or *non-English*? Leaving the order of precedence up to the annotators might cause discrepancy in the annotator agreement ratios.

Furthermore, for a lot of the messages, longer sentences were broken down across multiple messages, it would have been helpful to have clear instructions about whether these should be labelled based on the messages that follow and provide additional context or without taking those into consideration. For example, in one of the games the sequence of messages was as follows - 'apparently' which could either be *casual* or *cooperative* based on what the following messages are, and then a message describing the game strategy.