

Competition Launch Report

Dota 2 Message Classification for Auto-Moderation Purposes

Richard Motorgeanu - motorger - 400436012

Jacquelyn Pohl - pohlj - 400455088

Nicholas Gyorgypal - gyorgypn - 400446785

March 19, 2025

Content Warning: This document deals with the categorization of toxic content and messages found online that contain profane, vulgar, or offensive content. Proceed with caution.

1 Agreement

We used the Percentage Agreement metric to calculate agreement for our data. We chose this metric as the other 3 that were suggested didn't really apply to our situation or was more difficult to calculate. We didn't have more than 2 annotators, so Fleiss' Kappa wouldn't work. All the duplicated data in our datasets were annotated, so there was no need to consider Krippendorff's Alpha. Finally, since we had 8 categories, computing Cohen's Kappa would result in a lot of probability calculations (~ 50).

Using Percentage Agreement, we got an agreement score of 39%.

Since a lot of people mentioned that they struggled to differentiate between the different types of negativity, we also did a separate calculation where we grouped up categories 3, 4, and 5 and considered them equal. With this, we got an agreement score of 45%.

We did another scenario, where we calculated whether people agreed on if a message was toxic (categories 3, 4, and 5) or not toxic (categories 0, 1, 2, 6, 7). This resulted in an agreement score of 68%.

Finally, we did one last scenario, where we calculated whether people agreed on if a message was toxic or not toxic, and disregarded instances where one or both annotators labelled a message as non-English. This resulted in an agreement score of 62%.

2 Ground Truth Labels

For the ground truth, any piece of data that was annotated by only 1 person was determined to be the ground truth (for simplicity).

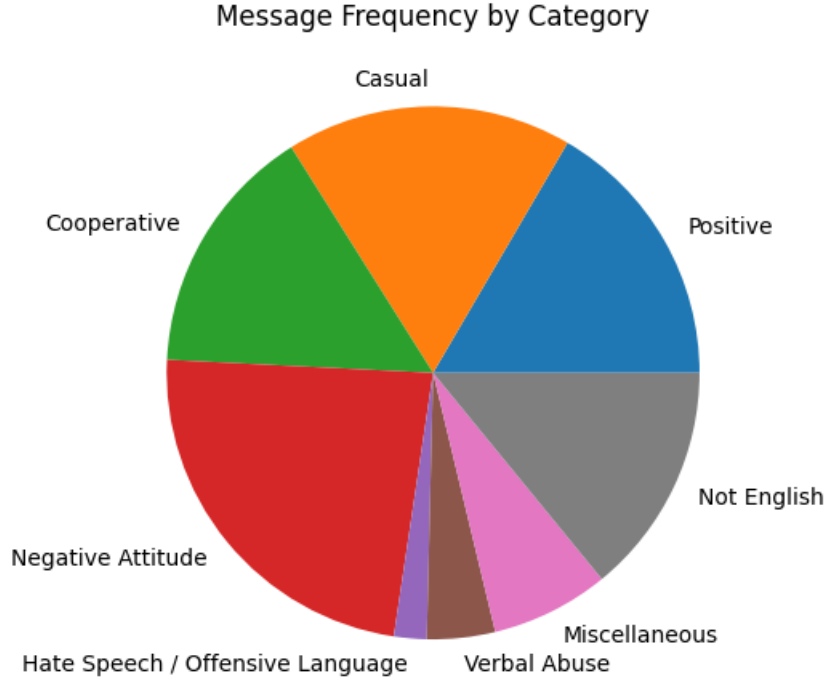
For duplicate pieces of data, if both annotators agreed, we let that be the ground truth. Otherwise, we did one of two things.

If the data came from data segment 1 and 2, we gave priority to data segment 1 as the annotator 1 has more experience with competitive games like Dota.

For data segment 3 and 4, it seemed like both annotators were inexperienced with competitive gaming and chats, so we had a member of our group go through their labels and pick which was more accurate, and let this act as a tie-breaker.

Below we've also attached the distribution of our categories. We notice that the classes are very imbalanced. Despite removing any non-English characters from dataset, we still got a lot of non English text (may be due to annotators classifying misspelt, abbreviated words, or emoticons as non

English).



3 Baseline

For our simple baselines we compared Random Baseline and Majority Baselines and found that Majority was better in terms of accuracy (with an accuracy of 27.6%) but had a false report rate of 1. This means using the Majority model all of our non-negative messages were labeled as negative. Meanwhile Random had a worse accuracy (with an accuracy of 11.3%) but had a better false report rate of 0.376. So with the Random model only 37.6% of non-negative messages were labeled as negative.

Comparing both of these to our chosen trained model, Logistic Regression, we see it outperform both the baselines in terms of accuracy (with an accuracy of 39.1%) but has a slightly worse false report rate compared to the Random model with a rate of 0.468.

4 Codabench and Google Slides

[Codabench](#)

[Google Slides](#)

5 Use of Generative AI

Generative AI (GitHub Copilot) was used for creating template code for parsing through data and creating baseline models. We approximate around 25 prompts, resulting in 108g of C02 emissions.