Annotation Guidelines
Dota 2 Message Classification for Auto-Moderation Purposes

Richard Motorgeanu, Jacquelyn Pohl, Nicholas Gyorgypal

February 5, 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 Overview

The annotator's job is to give each message provided a label based on the guidelines presented below. This will be done by running the file *annotator.py* in which the annotator will have 7 labels to choose from before moving on to the next message.

These messages have been split up into individual games, where the annotator can see the current game they're labeling, how many messages are in the game, and the time the last message was sent in the game. Each player in the game is assigned a random name to easily keep track of players and their random name, message and time of message sent is given to the annotator. The annotator is also able to see the 5 preceeding and 4 subsequent messages in order to understand the context in which the message was sent.

After the annotator has put a label for every message in a game, that game along with the labels will be saved to *input_data.json* and the annotator will be able to quit the annotation return to annotating anytime after this without risking losing data. The annotator is expected to annotate around 600 messages (through testing we have found that this is roughly 1 hour of work), however this number might vary.

Sample Output:

```
============================================================
Game 0 - Message (1/16) - Last Message Time: 37:46
--> [20:38] Hettie: carry
    [20:48] Laurel: yes dog
    [20:53] Hettie: lul
    [21:21] Dave: HAHAH
    [25:59] Laurel: yeah

(0) Positive (1) Casual (2) Cooperative (3) Negative Attitude
     (4) Hate Speech/Offensive Language (5) Verbal Abuse (6)
    Miscellaneous (7) Not English.
Your label:
```

2 Annotation Guidelines

2.1 Labels

Positive	Messages that contain compliments, good sportmanship, congratulations, and focuses on the good parts of the game.
Casual	Messages that are not related or off-topic to the game that are neither positive or negative.
Cooperative	Messages that are on-topic about the game that are neither positive or negative.
Negative Attitude	Messages that contain giving up, attempting to hurt their own team, cursing, poor sportmanship, and focuses on the bad parts of the game.
Hate Speech / Offensive Language	Hateful messages towards individuals/groups related to their sexuality, religion, gender, and often containing slurs.
Verbal Abuse	Hateful messages targeted towards an individual in the game which differs from above.
Miscellaneous	Any message that doesn't fall into the above categories
Non-English	Messages that cannot be labeled since they're not in english

When annotating, the annotators should ask questions like "Is this message positive in the given context?", "Is this casual or is this player giving strategy", "Is this non-english or is that phrase gamer slang/misspelled?" in order to narrow down a label. It is recommended to read through the following page on Dota 2 slang before annotating.

2.2 Examples

```
 ================================================
Game 0 - Message (8/16) - Last Message Time: 37:46
    [20:53] Hettie: lul
    [21:21] Dave: HAHAH
    [25:59] Laurel: yeah
    [26:03] Laurel: fast and furious
    [29:17] Laurel: too fas
--> [33:16] Bob: idiot drow
    [33:26] Ivan: no idiot
    [33:29] Ivan: we too pro
    [33:36] Laurel: haha
    [37:42] Laurel: sad
```

Verbal Abuse as we see Bob calling another player an idot and to drown (mispelled) themselves, which is hateful towards that individual, but not hate speech.

```
============================================================
Game 0 - Message (13/16) - Last Message Time: 37:46
        [33:16] Bob: idiot drow
        [33:26] Ivan: no idiot
        [33:29] Ivan: we too pro
        [33:36] Laurel: haha
        [37:42] Laurel: sad
    --> [37:42] Michael: fkasjfoiashnfiplwqjdf80q23iwkm d123q
        [37:43] Ivan: lol
        [37:43] Dave: COMMEND ME TY
        [37:46] Dave: EZ
```

Non-english, a random key-spam of letters that we would not be able to use.

```
1
     ============================================================

2     Game 2 - Message (18/37) - Last Message Time: 42:11
3         [4:07] Laurel: i killed u
4         [6:54] Ursula: WORST HOOK IN HISTORY
5         [7:04] Rupert: ur not even a good hooker kid
6         [10:20] Laurel: almost
7         [10:26] Ursula: YOU THOUGHT
8     --> [10:31] Ursula: IM THE #1 ROAMER NA
9         [14:34] Ursula: STUPIDD PIUDGE
10        [14:36] Ursula: STUPID!
11        [15:15] Hettie: ?
12        [17:52] Kyle: report!!!!!!!!!!!!!!!!@
      !!!!!!!~!!!!!!!!!!!!!!~!
13
```

Negative Attitude as we see Ursula displaying poor sportmanship towards their fellow teamates.

```
1 ============================================================
2 Game 2 - Message (27/37) - Last Message Time: 42:11
3     [17:52] Kyle: report!!!!!!!!!!!!!!!!@
      !!!!!!!~!!!!!!!!!!!!!!~!
4     [18:14] Ursula: nice ult medusa
5     [18:15] Ursula: commended
6     [18:22] Ursula: DOWNYS  GET DUMKED
7     [28:32] Rupert: passive shadow blade?
8 --> [37:59] Kyle: what is the best soup?
9     [38:27] Laurel: gg
10    [41:23] Ursula: ez game
11    [41:24] Ursula: Ty ty
12    [42:05] Rupert: gg
13
```

Casual as we see Kyle attempting to drum up some small talk with the players in their game (however they are very much ignored).

```
============================================================
Game 3 - Message (3/8) - Last Message Time: 29:54



    [-2:42] Bob: nice random ... lol
    [19:59] Oscar: ?
--> [22:29] Xavier: PUSH
    [22:37] Xavier: not defending
    [26:22] Xavier: dodger lc
    [29:07] Xavier: swap commend ty
    [29:52] Oscar: UPS
```

Cooperative as we see Xavier using the slang "push" letting their teamates know to focus on attacking instead of defending in that moment.

```
============================================================
Game 5 - Message (1/1) - Last Message Time: 29:00




--> [29:00] Hettie: GG
```

Positive as Hettie (begins) and ends the game with the slang "gg" meaning "good game".

```
============================================================
Game 9 - Message (2/2) - Last Message Time: 29:29



    [28:27] Alice: GG
--> [29:29] Xavier: WELL PLAYED SHIT CUNTS
```

Hate Speech / Offensive Language as Xavier is refering to others in their game as the derogatory term "cunts".

3 Contact Information

- Richard Motorgeanu - motorger@mcmaster.ca - Can contact on Teams as well.
- Jacquelyn Pohl - pohlj@mcmaster.ca - Can contact on Teams as well.
- Nicholas Gyorgypal - gyorgypn@mcmaster.ca - Can contact on Teams as well.