Project Proposal
Dota 2 Message Classification for Auto-Moderation Purposes

Richard Motorgeanu - motorger - 400436012
Jacquelyn Pohl - pohlj - 400455088
Nicholas Gyorgypal - gyorgypn - 400446785

January 18, 2025

Overview

Dota 2 is a multiplayer online battle arena (MOBA) game where players can communicate in chat logs with other players on their team (and possibly on the other team). Communications between players can be either positive or negative, regardless of team affiliation. We wish to categorize messages into different types: toxic messages, positive messages, casual chat (unrelated to the game), cooperative chat (informative and neutral). We'll break toxicity down further into verbal base, hate speech/offensive language, and negative attitude. Thus, we have a single-label classification task with 6 classes.

Many multiplayer games with chat functionalities have auto-moderation systems put into place to prevent and reduce negativity among players. Many times, looking for profanity in messages may not be a solution for games targeting older audiences and instead need a more sophisticated way of detecting vulgar and ill-intended messages. Additionally, context of messages may turn an initially negative sounding comment positive, and vice-versa.

1 Data

We will use the following Kaggle dataset which contains over 20 million chat messages across over a million Dota 2 matches:

https://www.kaggle.com/datasets/roovpa/gosuai-dota-2-game-chats

The data contains four pieces of information: a match ID that the message corresponds to, the time the message was sent within the game (measured in seconds), the player ID (of the match) where the message came from, and the message itself. We will classify the message, possibly using context of the other columns as well as context from preceding and/or proceeding messages.

We will not use the entire dataset, as this will would be a laborious and time-consuming annotation process. We will take somewhere from 5k-20k messages depending on how laborious the annotation process is.

Additionally, many messages are in Russian. We may attempt to use a translator prior to annotation, as some games include messages from both English and Russian players (resulting in interesting data of how they may communicate), however we may also exclude these messages from the data due to the many game-specific terms that players use that may impact the translation accuracy.

The data was gathered by an individual external to Dota 2 and Valve. Collection methods of the data was not included, nor is a license of the data. We will credit individuals as needed. The data does not have any labels attached to it.

We additionally found the Kaggle notebook below which utilized the same dataset we're using. It will be interesting to compare how our approaches to classifying messages will compare to theirs, and how the accuracy/output will differ.

https://www.kaggle.com/code/adelsondias/introduction-of-zero-shot-classification-w-chats

1.1 Example 1

Message: why all russian are so dirty player
Category: Hate Speech

1.2 Example 2

Message: u cant win
Category: Negative Attitude

1.3 Example 3

Message: gg
Category: Positive Message (if said at the end of a game), Negative Attitude (if said in the start/middle of a game). Good example of ambiguity depending on context.

2 Team Contract

As a team, we hope to create an interesting classification tool that has real-life applications in live-service cooperative/competitive games and learn about NLP throughout the process. We will also ensure that this is an enjoyable and informative process.

2.1 Team Responsibilities

As a team, we will *all*:

- collaborate and participate at *all* stages of this project;
- complete group tasks in a *timely** manner;
- make decisions with unanimous consent*;
- truthfully disclose C02 emissions due to generative AI and attempt to keep it to a minimum;
- abide by McMaster's and this course's academic integrity policy;
- and have fun.

2.2 Team Member Responsibilities

As an individual, we will *all*:

- ensure completion of assigned individual tasks in a *timely** manner;
- *not* deviate from assigned task without consulting other team members;
- truthfully disclose C02 emissions due to generative AI to team members and attempt to keep it to a minimum;
- consult team members when confused or need assistance.

2.3 Leadership and Management of Group Activities

Any decisions related directly to the project will be decided and agreed upon by the entire group. This includes initial project ideation, deciding software stack/NLP methods, and report write-ups.

Any administrative/miscellaneous activities will be handled by Richie Motorgeanu. This includes internal due dates, communication channels/meetings, and task allocation.

2.4 Resolving Disagreements*

In the case that a decision cannot immediately be made unanimously, we will consider arguments to why we should choose choice over another.

If no unanimous decision can be made after an attempt to sort it out as a team, we will consult Dr. Charles Welch to provide guidance and a possible resolution.

In the case that Dr. Charles Welch cannot assist and/or is not available in a timely manner*, the final decision will go to the majority vote. In the case of a stalemate (3-way tie), we do rock paper scissors.

2.5 Communication

Communications will be done in person when convenient. Online communications will be done via Discord primarily, with Teams as a method of communication with Dr. Charles Welch.

2.6 Miscellaneous

Timely may mean different things to different people. In this contract, *timely* defaults to 24 hours if no other deadline is immediately presented/requested.

This project proposal and contract was made without the assistance of generative AI at any point.

3 Signatures

Team Member 1:

Print Name: _____

Date: _____

Signature: _____

Team Member 2:

Print Name: _____

Date: _____

Signature: _____

Team Member 3:

Print Name: _____

Date: _____

Signature: _____

Professor:

Print Name: _____

Date: _____

Signature: _____
