Project Proposal
Dota 2 Message Classification for Auto-Moderation Purposes

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

January 22, 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.

Overview

Dota 2 is a multiplayer online battle arena (MOBA) game where players can communicate via chat logs with teammates and, in some cases, players on the opposing team. These communications can range from positive and encouraging to negative and disruptive, regardless of team affiliation. We aim to categorize messages into four types: toxic messages, positive messages, casual chat (unrelated to gameplay), and cooperative chat (informative and neutral). Toxicity is further divided into three subcategories: verbal abuse, hate speech/offensive language, and negative attitudes. This creates a single-label classification task with six distinct classes.

Many multiplayer games with chat functionality implement auto-moderation systems to reduce negativity among players. However, simply detecting profanity is often insufficient for games targeting older audiences, as it fails to address subtler forms of toxicity. Additionally, message context can significantly alter interpretation. An initially negative-sounding comment might be meant positively, and vice versa. A more sophisticated approach is necessary to accurately detect and categorize player communications.

1 Data

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

GOSU.AI Dota 2 Game Chats

The dataset contains four key attributes: a match ID corresponding to the message, the time the message was sent within the game (in seconds), the player ID associated with the message, and the message content itself. Classification will focus on a single message, but contextual information from other attributes and surrounding messages may be used to support the classification of that message.

Due to the labor-intensive nature of annotation, we will work with a subset of the data, selecting between 5,000 and 20,000 messages depending on the effort required.

A significant portion of the messages are in Russian. To address this, we may use a translator before annotation, as many matches feature both English and Russian players. This creates interesting data on cross-language communication. However, we may exclude Russian messages if translation inaccuracies arise due to game-specific terms.

The dataset was collected by an individual unaffiliated with Dota 2 or Valve. It lacks information about the collection process, licensing details, or pre-existing labels. Proper credit will be provided as necessary.

Additionally, we identified a Kaggle notebook that uses the same dataset. Comparing our classification methods and results to theirs will offer insights into differing approaches and their respective outcomes.

Introduction of Zero-Shot Classification with 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). This demonstrates ambiguity depending on the context.

2 Team Contract

This project has real-life applications in live-service cooperative/competitive games and provides an opportunity to deepen our understanding of natural language processing (NLP). We will ensure this process is both enjoyable and informative.

2.1 Team Responsibilities

As a team, we will *all*:

- collaborate and actively participate in *all* stages of the project;
- complete group tasks in a *timely** manner;
- make decisions with unanimous consent*;
- minimize and truthfully disclose CO2 emissions resulting from the use of generative AI;
- abide by McMaster University's and this course's academic integrity policy;
- and have fun.

2.2 Team Member Responsibilities

As an individual, we will *each*:

- ensure assigned tasks are completed in a *timely** manner;
- *not* deviate from assigned task without consulting the team;
- minimize and truthfully disclose CO2 emissions resulting from the use of generative AI;
- seek clarification or assistance from team members when needed.

2.3 Leadership and Management of Group Activities

All project decisions will be made collectively by the group, including initial ideation, software stack and NLP method selection, and report preparation. Administrative and miscellaneous tasks, such as managing internal deadlines, organizing communication channels and meetings, and allocating tasks, will be handled by Richard Motorgeanu.

2.4 Resolving Disagreements*

If a decision cannot be reached unanimously, we will encourage productive dialogue to arrive at a resolution. If no decision is made within a *timely** manner, we will consult Dr. Charles Welch for guidance and a suggested resolution. If Dr. Charles Welch is unavailable or unable to assist in a *timely** manner, the decision will be determined by majority vote. In the case of a three-way tie, a random selection method, such as rock-paper-scissors, will be used.

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 have conflicting definitions. In this contract, *timely* refers to 24 hours if no other deadline is immediately presented.

This project proposal and contract was made with the assistance of generative AI to allow for better clarity and minimization of the ambigious language of this document. 5 prompts to ChatGPT were made, totalling to 21.6g of C02 emmisions.

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: ______________________________