**WoW Preliminary Exploratory Data Analysis**

In our World of Warcraft posts dataset, we are given 1033 rows of 5 features: topic\_id, post\_number, author, html\_message and timestamp. For the purposes of this project, we will ignore timestamp, as it will not be used to train our models. The additional WoW annotations dataset reveals which messages were flagged as cyberbullying, from which the only useful features are the post\_number and topic\_id.

Posts in the main dataset are formatted using HTML, meaning that our model will either have to be trained to recognise patterns such as paragraph breaks or we will have to clean the data and transform it into something more appropriate.

Furthermore, as we are only using a single set of features (post\_number and topic\_id) of the annotations dataset, we may be able to add another column to the posts dataset – a Boolean value representing whether or not a specific post contains cyberbullying. This will eliminate the need for the use of two separate datasets to develop our models.

Here are the data types of each feature:

Posts datasetAnnotations dataset

topic\_id – int topic\_id - int

post\_number – int post\_number – int

author – hex number (string) annotator – int

html\_message – string offender – hex number (string)

timestamp – string victim – hex number (string)

There are 5 unique topic IDs, and they are as follows:

**149** – discussion of a lack of updates for the game

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

There are 179 posts in this category, and 15 of them are marked as cyberbullying. This results in an 8.379% of all posts in this topic being related to cyberbullying.

**10178** – general complaints about the game

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

There are 87 posts in this category, and 20 of them are marked as cyberbullying. This results in an 22.989% of all posts in this topic being related to cyberbullying.

**11554** – posts relating to farming in the game

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

There are 186 posts in this category, and 13 of them are marked as cyberbullying. This results in an 6.989% of all posts in this topic being related to cyberbullying.

**11817** – discussion about LGBT characters

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

There are 471 posts in this category, and 7 of them are marked as cyberbullying. This results in an 1.486% of all posts in this topic being related to cyberbullying.

**13120** – complaints about the game store

Graphical user interface, text, application, email, Teams

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

There are 100 posts in this category, and 18 of them are marked as cyberbullying. This results in an 18% of all posts in this topic being related to cyberbullying.