2.3 SpeechRecognition

Originally we wanted to analyse the interaction between the streamer and the users, so we needed data from the streamer. The audio downloaded was used with the python library SpeechRecognition But the time needed to analyse an audio file was around 1:1, so 1 minute of audio took 1 minute to parse. The text that was returned was completely gibberish and useless. This is expected as speech recognition is still not at the required level to properly gather reliable data, moreover music is usually played on stream and together with game sounds this interferes significantly with the speech recognition. We conclude that speech recognition on a massive dataset is still not a viable option. In the near future when services are cheaper and the recognition technology is more advanced it could become interesting to look into this.

2.4 Training

To analyse the chat data, we had to gather a training set in which the chat is rated based on their toxicity. We converted the compressed pickle file of to a .sql insert statement to store the chat into a MySQL database.

A website implemented in javascript and php showed this chat and we manually rated each chat message with a 1 to 5 (1 being non-toxic, 5 being very toxic).

2.5 Classifier