Department of Computer Science

Exploring Toxicity within Online Gaming Forums using BERT

Harrison Grace

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Masters of Engineering in the faculty of Engineering

17th September 2021

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: HARRISON GRACE

DATE: 17th Sept 2021

Ethics Approval

An ethics application has been reviewed and approved by the ethics committee with a review reference of 2021-0034-85.

Table of Contents

List of Figures

List of Tables

Abbreviations

Natural Language Processing & Data Capture Related

NLP – Natural Language Processing
TF-IDF - Term frequency-inverse document frequency
SVM – Support Vector Machines
CNN - Convolutional Neural Network
RNN - Recurrent Neural Network
LSTM – Long Short-Term Memory
BiLSRM – Bidirectional Long Short-Term Memory
BERT - Bidirectional Encoder Representations from Transformers
OLID – Offensive Language Identification Dataset
SOLID – Semi-supervised Offensive Language Identification Dataset
GPT – Generative Pre-trained Transformer
GLUE – General Language Understanding Evaluation
SQuAD -Stanford Question Answering Dataset
SWAG – Situations with Adversarial Generations
CBOW – Continuous Bag-of-Words
GLoVe - Global Vectors
ELMo – Embeddings for Language Model
PMAW – Pushshift Multithread API Wrapper
MLM – Masked Language Model
NER - Named Entity Recognition
MNLI - Multi-Genre Natural Language Inference
NSP – Next Sentence Prediction

Gaming Related

MOBA – Multiplayer online battle arena
RPG – Role-playing Game
LoL – League of Legends
GTA – Grand Theft Auto
DOTA – Defence of the Ancients
PvE – Player vs Enemy
PvP – Player vs Player
ESRB - Entertainment Software Ratings Board

Executive Summary

In recent years, there has been rising concern with widespread toxicity in the gaming industry. The term ‘toxicity’ refers to a broad range of negative behaviours, including but not limited to forms of harassment, hate-speech and insults. With the popularity of gaming increasing the instances of toxicity are rising alongside it. This can have damaging effects to gamers experiencing such occurrences and results in some gamers choosing to avoid certain games completely.

This thesis investigates the toxicity currently present across gaming communities. To accomplish this a literature review was performed focusing on toxicity in the gaming industry. It also explores natural language processing (NLP) techniques for detecting online messages likely to be considered toxic. From this research an actionable definition of toxicity is derived and is then represented with classifiers based around the Offensive Language Identification Dataset (OLID) classification. This representation is then used to explore the prevalence of toxicity within gaming communities residing on the popular social platform known as Reddit. More than 5 million comments are collected and cleaned from the platform to investigate the differences in rates of detectable toxic behaviour.

The selected model for toxicity classification is the Bidirectional Encoder Representations from Transformers (BERT). This versatile model is now an established tool in the NLP field with outstanding results on multiple renowned benchmark tasks. In this thesis, two models are fine-tuned using the OLID classifications for the chosen definition of toxicity. The models are further improved with a portion of self-annotated Reddit data. A hybrid of the BERT model known as HateBERT was also produced for comparison and potential improvement. The final improved BERT models are applied to the collected Reddit data and used to explore some hypotheses about the distribution of toxicity in different online gaming communities. A critical review is performed on all the models as well as the chosen OLID classifications.

Contributions:

- Performed a thorough literature review of toxicity within the gaming industry and natural language processing techniques which can be used to identify such occurrences.
- Formed a definition of toxicity and expressed this with suitable classifiers for measurement.
- Collected and cleaned over 5 million posts from multiple gaming subreddits for evaluation.
- Tested the most effective hyperparameters for OLID classification for the BERT model.
- Created BERT models based around the toxicity definition within Google Colab.
- Created HateBERT models to compare effectiveness.
- Labelled over 5500 posts from the collected data to test and improve the BERT model.
- Used the model predictions alongside additional metrics to compare differences in offensive and toxic content within multiple gaming subreddits.
- Performed a critical evaluation of the OLID classifiers and the metrics used in the OffensEval competitions.

Supporting Technologies

Third party resources used throughout the thesis are outlined below:

- Google Colab for creation of the machine learning model.
- Google's BERT language model.
- The final BERT models were altered from an original work by Chris McCormack [1].
- Offensive Language Identification (OLID) dataset was used to train and test the BERT models.
- HateBERT model.
- The Reddit PMAW scraper was used to obtain the large quantities of data from Reddit.
- The Langdetect Python package was used to detect non-English languages within the collected Reddit data.

Acknowledgements

I would like to thank my supervisor, Dr Matthew Edwards, for his close support throughout this thesis especially during the circumstances of the COVID pandemic. His advice and expertise over continual weekly meetings ensured successful execution of this project. I am extremely grateful for his insights and have learnt a huge amount from them.

1 Literature Review

1.1 Defining Toxicity

Online toxicity is a very ambiguous term to define and there is no clear and standardised definition presented. Blackburn et al describes it as a 'form of cyberbullying, defined as repetitive intentional behaviour to harm others through electronic channels' [2]. The Oxford dictionary defines toxicity as 'the quality of being harmful or unpleasant' and the definition of toxic as 'very unpleasant, especially in the way somebody likes to control and influence other people in a dishonest way'. Mohan et al. expresses any instance of cyberbullying, cyber threats, on-line harassment, hate speech, and abuse as 'toxic' [3]. Riot Games, the makers of League of Legends, define toxicity as 'any behaviour that negatively impacts other players' experiences' [4]. For our purpose we define online toxicity as 'Intentional unpleasant behaviour used to aggrieve others online'. Furthermore, toxicity and abuse are used interchangeably between studies with no standardised representation [5, 6]. Therefore, it is also acceptable to assume that any form of online abuse can be deemed toxic and vice versa.

Classification of online abusive behaviour is widely explored, however little standardisation between studies occurs. This results in many different classifications many of which have similar definitions. Fortuna et al. compare multiple studies which use machine learning techniques for differing classifications [6]. Firstly, an analysis between the differing labelled training datasets is performed on 6 prior studies. The classification is then standardised to a minimal list including sexism, racism, hate speech, offensive, misogynous, aggression, insults, threats, identity hate and toxicity. The study then uses FastText to train word embeddings and extract the centroid of each message classification by averaging the embeddings of its sentences. A cosine distance metric is then calculated to measure the difference between the words. The results proved inconclusive in establishing links between the differing classifiers of abuse from prior works. The study shows that classifiers are strongly dependent on their training datasets and very sensitive to its initial labelling and frequency. Therefore, the training data for this study needs to be well represented and the classifiers of toxicity clearly defined.

Mishra et al. define abuse as separate categories; explicit, implicit or directed. Explicit having the form of expletives and threats, implicit being more subtle characterized by ambiguous terms such as sarcasm and directed abuse targeting specific groups such as racism. Waseem et al. categorises posts via explicit/implicit as well as generalized/directed [7]. The latter gives context to the target which isn't really needed for the definition of toxicity. However, the implicit and explicit interpretations help improve the detection of more subtle abuse.

1.2 Current Toxicity Detection Techniques

Mishra et al.'s study focuses on the comparison of prior NLP abuse detection techniques and their effectiveness [5]. The paper categorizes these studies into Lexicon & Rules based methods or computational. Lexicon & Rule based methods utilize a dictionary of target words which are run

against the database and return how many occurrences of such words were matched. The Lexicon methods show good effectiveness across differing domains of data however struggle with typing & grammar mistakes or implicit abuse. Computational methods are those which use a mixture of machine learning features such as TF-IDF, n-grams and bag-of-words (BOW). Computational methods showed much more robustness with spelling & grammatical errors and are much faster to model than the rule-based counterparts. Their downfall relates to the difficulty of interpreting results and they can miss deeper semantic meaning. The paper introduces a third technique utilizing the up-and-coming research for NLP processing in the form of neural networks. Neural network techniques include CNNs and RNNs, with the former having models such as semantic parsing and the latter utilizing LSTMs. These methods show mixed improvement over Lexicon/Rule based and Computational methods. The GermEval 2019 competition[8] showed that CNN and RNN models didn't feature in any of the top 3 of any sub-tasks, whereas Lexicon and n-gram methods fared well. However, the deep language model known as BERT was utilized in every instance of the winning submissions across all subtasks. The BERT model is a good technique to explore further given its successfulness at handling obfuscation as opposed to CNNs and RNNs.

OffensEval is a 2019 subtask at the academic NLP competition, SemEval, which has been running since 1998. OffensEval specifically focuses on NLP of abuse detection methods showcasing the best techniques in a published paper thereafter. The baseline for the competition is established from a training dataset built from over 14,000 tweets that are provided by the organisers called OLID [9]. The tweets within the dataset are classified by Level A: Offensive Language detection, Level B: Categorization of offensive language and Level C: Offensive language target identification and given an abbreviation associated with each category. These classifiers were applied to the tweets using crowdsourced annotators who mutually agreed on the classification. Once classified the datasets are split into a training set used to train the machine learning technique and a test set to run the trained technique on. The organisers ran SVM, BiLSTM and CNN models on these datasets to give a baseline to the competitors models with the CNN model producing the best results.

The OffensEval competitions are split as per the classifications of the datasets into categories A, B and C with teams attempting to classify one or more. The 2019 competition had a large range of models being implemented, the most popular being deep learning techniques at 70% of the submissions [10]. Most notably the best model for task A was BERT which was used by 7 of the top 10 teams and also came in at first place. For task B ensemble methods proved to be a strong technique with 5 of the top 10 teams applying some form of it. Interestingly the first-place team for task B came in the form of a rule-based lexicon approach. Finally, task C had 5 of the top 10 teams using ensemble methods, however yet again a BERT model came in first place. These findings highlight some strong techniques of NLP abuse detection.

The 2019 competition noted that the OLID dataset was somewhat limiting in its classifications due to a low count of occurrences, this lowered the accuracy of the machine learning models for classification B and C in particular [11]. To combat this issue a new SOLID dataset was produced containing over 9,000,000 tweets for the 2020 competition. A comparison on the improvement of using OLID and SOLID datasets together to train models were performed with FastText and BERT models. This comparison showed that for BERT models classification A had no improvement over using the both datasets in training the model. Classification B showed a minor improvement and C noted a substantial improvement when trained on the OLID dataset initially and the SOLID dataset thereafter. FastText showed a large improvement for classifications A and B yet failed to show any improvement for C. Overall, the paper showed that use of the SOLID dataset as

well as OLID was likely to improve model classification and in some instances, quite substantially. The creation of such large datasets is out of scope for this project to undertake. However, it is worth exploring if these OLID/SOLID datasets can effectively be used to train a machine learning NLP technique to be used on other domains such as Reddit.

The teams working on OffensEval 2020 were aiming to classify tweets as per both the OLID/SOLID datasets [12]. The 2020 competition also offered multilingual options, for the purposes of this paper only the English section is focused upon. Before attempting these identifications most teams performed some form of pre-processing or text normalization. Teams then need train their models with OLID/SOLID or their own labelled datasets classified as above prior to running on the final dataset. Once again, the BERT and BERT hybrid models dominated the contest with all of the top 10 teams applying some form of it sometimes alongside CNNs or LSTMs. Many BERT hybrid models were used with extremely effective results such as ALBERT and RoBERTa.

Unfortunately, OffensEval is not a subtask to the SemEval 2021 event. There is a new subtask 5 which focuses on detecting 'toxic spans' and extracting the toxic words from sentences. Nonetheless, these papers will not be available in time to evaluate for this literature review.

The OffensEval competitions have a very simple model of categorization and the large adoption of this classifier shows its versatility with differing models. This is a good classifier to pursue given the popular and largely successful BERT modelling that has occurred in the competitions [10, 12]. The process of annotating the dataset involved multiple experienced annotators who must agree on the classification, else a third annotator is involved to finally decide [9]. This should give good quality to the dataset with a low number of questionable labels. Also, the multiple categories of OLID allow us to pinpoint if a post is not only toxic (classifier A) but also targeted (classifier B). Classification C looks to find the target specifically, this is not necessary for the definition of toxicity and would not need to be pursued. The theme of explicit and implicit abuse has also been explored by applying extra labels as a sub-category to category A resulting in a separate dataset called AbuseEval [13]. This dataset was produced by cross-checking for a slur or profanity and marking the target as explicit, else the message is marked as implicit. Unusual online vocabulary was captured by checking against an online slang website known as Urban Dictionary. A pre-trained BERT model was used to test the new annotations in detecting both implicit and explicit labels. The findings showed that implicit abuse is tough to detect using the model although it was highlighted that the limitations of this are due to the quantity of data.

The perspective API is another notable model which was created by Jigsaw and Google and uses CNNs trained with GLoVe embeddings based on Wikipedia training data. The API classifies text via the category's; toxicity, severe toxicity, identity attack, insult, profanity, threat and returns a score between 0 and 1 giving the likelihood that the sentence is toxic. The API is generally successful and was even entered into OffensEval 2019 coming 12th for classification A with no additional training [14]. It was further compared alongside BERT which showed Perspective outperforming for classifier A. However, the model did struggle with categorizing of abuse for section B with a disappointing 0.48 F1 score compared to BERTs score of 0.68.

A competition based on the perspective classification was released to the public through Kaggle and many papers have further explored the model [15-18]. One deprecation has been the inadequate training dataset provided for training the model at the time, out of 223,549 comments only 22,468 are classed within toxic and furthermore only 0.3% of comments fall into the sub-category of threats [17]. This also reflects in a high occurrence of false negatives produced from disagreement in

labelling, toxicity with no swear words, rhetorical questions, sarcasm, metaphors and rare words. False positives also had issues with data labelling, use of swear words in non-toxic posts, quotations and rare words. Whilst the solution to some of the incorrectly flagged implicit abuse is tough and requires more context, other issues such as labelling and rare words could be resolved with higher quantity and quality of the labelled training dataset. Furthermore, spelling mistakes can vastly lower the toxicity score of the model [18]. This highlights once again the importance of the pre-processing step in providing adequate training data to the model.

One of the most successful techniques highlighted so far is the use of the BERT model. This model was produced by Google and is a state-of-the-art bidirectional transformer-based machine learning model used for NLP [19]. It was noted at the time that current transformer models such as GPT relied on left-to-right tokenising giving limitations to the context. BERT uses masked language models (MLM) to enable pretrained deep bidirectional representations. It is first initialized with pre-trained parameters which are then fine-tuned using labelled data from downstream tasks. The development of this model has been a major success and has marked a new chapter in the NLP field. At the time it created new highs for multiple benchmarks including GLUE, SQuAD and SWAG. It is worth noting however that instances of BERT's effectiveness rely on using large training datasets which can be cumbersome, LSTM models fared better when handling smaller datasets [9]. Nonetheless, with enough pre-labelled training data being readily available online, the BERT model is the better suited. The model is open-source and has already been adapted by many researchers producing dozens of hybrids which increasingly dominate the field of NLP.

A notable hybrid BERT model is the HateBERT model, not only does it implement the BERT model for abuse detection but it is also trained on an abusive Reddit dataset called RAL-E [20]. RAL-E is a huge dataset consisting of almost 1.5m posts from banned Reddit communities. The BERT model is trained on this dataset applying the MLM objective and with some further pre-processing the HateBERT model is produced. The model has been tested on 3 differently labelled datasets focusing on offensive language (OffensEval), abusive language (AbusEval) and hate speech (HatEval). For all 3 datasets the BERT model was outperformed by HateBERT achieving an increase in F1 score by 0.06, 0.38 and 0.36 respectively. This increase in accuracy of abuse detection makes it favourable to implement in this study. The paper also highlights a pivotal point that substantial improvement of such models is based on the quality of the pre-processing steps as oppose to an increasing learning rate or training time.

The BERT model has continually shown its effectiveness and seems to be one of the most outstanding models since its development. Exploring the HateBERT model would also be beneficial given it is also pre-trained on abusive Reddit data.

1.3 Toxicity in the Gaming Industry

Toxicity is a common occurrence in the gaming industry and many gamers see this as a normal phenomenon. To attempt to phase out toxicity many games rely on players reporting each other. This solution can be ineffective given that many gamers view toxicity as 'acceptable, typical of games, as banter, or as not their concern' [21]. Furthermore, a player's most common resolution to toxicity involves muting or blocking the perpetrator due to its instantaneousness [22]. These justifications of toxicity and lack of punishment perpetuates the occurrences in the gaming community.

Kou et al. uses data from the subreddit for League of Legends (LoL) and looks specifically into contextualizing toxic behaviours [23]. The paper highlights that competitiveness, in-team disagreements, perceived loss and powerlessness lead to a much greater chance of toxicity. The study's target posts focus on explained incidents of toxicity happening in the game as opposed to the direct toxicity within the subreddit. Nonetheless, it highlights interesting findings and offers further exploratory hypotheses to expand on such as 'Do other MOBA games or team-based competitive games have similar or different toxic types and contexts?'.

Another study also focusing on LoL explored the factors which predict whether a player has a higher toxicity index [4]. This index was created using an add-on built by Duowan, a Chinese forum website. The add-on gives a strong representation of players' toxicity allowing players to rate each other with a thumbs up or down after a match. The study finds that experienced players expressed more toxicity than newer ones. It also investigated whether a player's choice of a more aggressive in-game character type could predict whether the player was more toxic, this however proved inconclusive. The study also investigated factors relating to the retention defined by the time a player continually spends on the game. It showed that experienced players had higher retention when involved with higher toxicity teams than those who are newer to the game. This highlights the gaming companies need to focus on addressing toxicity in order to keep their communities thriving.

A very recent 2021 study explores the toxicity between different gaming communities. The study uses BoW, sentiment analysis and word embeddings to compare 13 gaming communities from Reddit and Twitter [24]. The findings show that negativity is very equal across the gaming communities and 12 of 13 games had similar levels of negativity at around 20%. Specifically, the study looked at underlying rates of racism, sexism and Trump-hate between the gaming communities. The BOW model term frequencies showed the Fifa community using the most racist and trump-hate words, World of Warcraft also came second for both counts. Minecraft and The Sims had most occurrences of Sexist words. However, once other features were implemented in the model neither were comparatively sexist compared to the other games. After applying all individual model features and weighting coefficients for the most important features a final score was produced. This score lacked some methodological detail and had no specified unit within the research. Nonetheless, Fifa still came out as the most racist community with a score of 0.0472. Both MOBA games in the study, Dota 2 and League of Legends had the highest values of Sexism the highest being 0.0577. Fortnite showed the highest scores for trump-hate at 0.007, however all values were not very substantial in the evaluation bracket of [0,1]. This shows us that gaming communities aren't particularly interested in political abuse and whilst racism and sexism is more common it is still a minority.

Another interesting element to investigate is the triggers of toxicity. Shen et al. focuses on this analysis at a team-based level within the game called 'World of Tanks' [25]. The study ran a logistic regression model on occurrences of toxicity flagged reports from the game servers. The model supported dominant skilled teams showing a lack of toxicity reports compared to their inadequate opponents. Furthermore, skill disparity within the same teams contributes to higher toxicity as well as a higher likelihood in teams who are not associated with each other. Any instance of toxicity had a much greater likelihood to spread to others throughout the duration of a match. The study also showed no difference that a perpetrator's toxicity spreads within their own team compared to their opponents. The study reflects also that higher skilled players exerted far greater toxicity than beginners and this correlated well as the skill level increases. It is worth noting that this research relied upon reporting stories of abuse as oppose to direct abuse.

1.4 Summary

This literature review has successfully drawn a definition of toxicity from surrounding papers and explored classifiers to represent this definition. It has also investigated the relationship between gaming communities and their underlying toxicity, including predictors of toxicity occurrences or toxic players themselves. The review has also shown the dominance of the BERT model and its hybrids within the NLP field, being popularised throughout the OffensEval competitions. Its success in classifying is largely drawn from the use of large high-quality pre-labelled datasets and significant pre-processing. A notable hybrid is the HateBERT model for its improvement of the F1 score and its use of Reddit training data.

2 Contextual Background

2.1 The Gaming Industry

The gaming industry is a multi-billion dollar industry which in 2021 stands at around $138.4 billion [26], this is expected to increase at least two-fold by 2025 to around $300 billion [27]. Furthermore, the number of gamers is ever increasing at around a 5% increase year on year, for 2021 this amounted to 2.81 billion worldwide gamers [28]. This is partly down to the increased accessibility of gaming as technology progresses. Cloud platforms such as Stadia or Amazon Luna allow players to stream games onto their devices without needing the powerful requirements to run them directly. The accessibility is further extended by new developments such as cross-play, whereby people can also play the same games as each other without the need of owning the same consoles. Furthermore, many new games are also being released as 'free-to-play' or included within cheap subscription services.

The COVID-19 pandemic has also increased the popularity of gaming as people look for ways to combat their boredom while in lockdowns. A recent study shows that 10% of respondents played games multiple times a day pre-pandemic, this increased to 40% during the pandemic [29]. This has been reflected in the worldwide increase in sales of consoles and games [30]. During the pandemic the World Health Organisation released a promotion called '#PlayApartTogether', this encouraged and promoted online gaming [31]. The pandemic has also changed gamers' playing habits. One study shows a shift away from strategy, puzzle and sports genres, with players turning their attention towards battle royales, MOBA's and fighting games [32]. The frequency of players playing online multiplayer games in the pandemic has also increased by 60% [33].

The positive effects that gaming brings to individuals has been widely explored. It shows notable improvements to problem-solving skills [34] and helps to elevate interests on topics such as history [35]. One study showed a strong link in gaming and learning the English language whilst those who did not game were actually at a disadvantage [36]. Gaming further incentivises children to open up socially [37] and can be used to improve social illnesses such as Autism [38]. The benefits can also extend to the elderly with more physical consoles such as the Wii and virtual reality improving balance and reducing occurrence of falls [39] as well as improving social and mental health [40]. A Qutee survey in 2018 finds the most selected benefit of gaming as 'an improvement to emotional well-being' at as the highest factor at 43%. The 'forming of new friends' came second at 17% [41], while only 3% stated 'no benefit to wider society'. ADL, an anti-hate organization, grouped together with a games market analyst company Newzoo to investigate toxicity within 18-45 aged US gamers. The 2020 survey recorded 95% of respondents who played multiplayer had positive social experiences [42].

2.2 Toxicity in the Gaming Industry

Alongside the positivity that gaming brings it also is unfortunately renowned for harbouring toxicity. The 2020 ADL report shows that 81% of the respondents who played online multiplayers games experienced a form of harassment increasing by 7% since the 2019 report [42]. Furthermore,

68% experienced severe abuse in the form of physical threats, stalking and sustained harassment. The worst games to be noted were DOTA 2, Valorant, Rocket League, Grand Theft Auto, Call of Duty and Counter Strike: Global Offensive. Another study finds that 76% of respondents believed prejudiced comments in online gaming should be confronted, however only 18.5% actually would respond with confrontation [43] with the majority opting for ignoring the occurrences at 35%.

Toxic behaviours are also extending into the professionalism of Esports. Esports are organized gaming competitions for multiplayer games hosted with spectators. They typically involve professional gamers who make a living from the cash prizes available. Through research it has been found that Esports gamers normalize toxicity [44]. Esports gamers have a higher prolonged exposure to toxicity given the higher play rate. This can further lead to them becoming perpetrators themselves after perceiving abuse as acceptable.

Toxicity in gaming has many forms. It includes yet isn't limited to harassment, discrimination, hate speech, trolling, griefing, intentionally losing, quitting, refusing to play. Trolling is the deliberate act of provoking people online with the aim to cause a reaction. Greifing is similar in that the perpetrator takes pleasure from creating grief for others. More serious cases of toxicity include stalking, threatening, doxing and swatting. Doxing involves publishing private information with malicious intent. Swatting involves false reports of life-threatening emergencies in order to cause armed police or S.W.A.T teams to be deployed to the targets home. In some cases, this has resulted in injuries [45] or even death [46]. It is paramount that games companies try to reduce these toxic occurrences from happening and make the gaming community a welcoming place for all individuals.

2.3 Moderation Techniques & Effectiveness

Toxicity is a well-known aspect of gaming and its effect on turning away newcomers can be devastating with ADLs survey showing 22 percent of gamers receiving harassment to stop playing certain games indefinitely [42]. It is important that gaming companies acknowledge and address the problem to keep their communities flourishing.

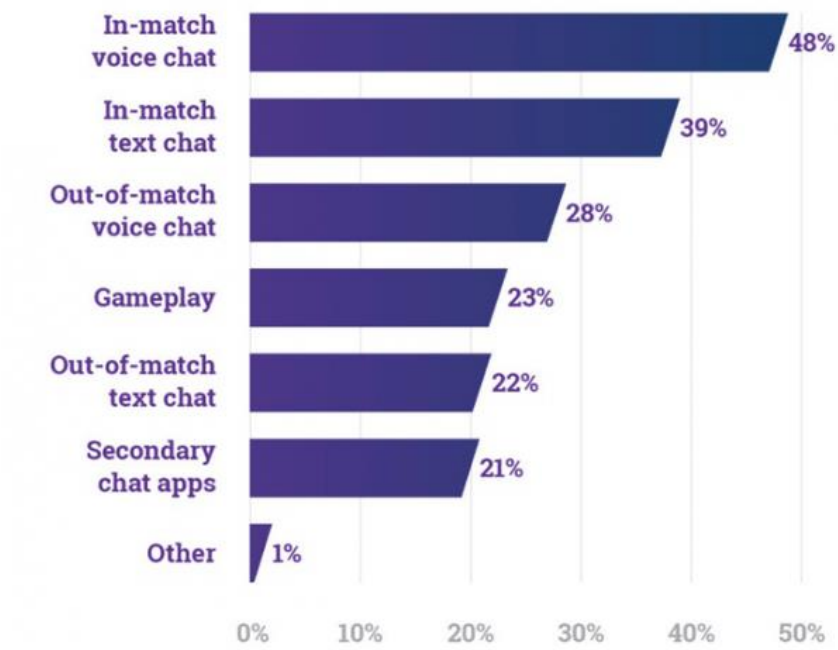
A common feature of multiplayer games is a reporting system whereby an aggrieved player can submit a complaint against another from a set of options. If a player receives enough complaints they will be warned of their behaviour, given chat restrictions, rank restrictions, lose cosmetic in-game items and could be banned for certain periods or even indefinitely. However, this feature can be manipulated by players who file false reports even if players exerted no toxicity. The ADL survey shows that 25% of those who experience harassment don't report it which is mostly due to the occurrence not being disruptive enough [42]. The success of reporting is arguable with 16% of those not reporting harassment stating that previous reports do not result in meaningful action and therefore not a good use of their time [42]. Games companies such as Blizzard are trying to counter this by notifying the reportee if the reported player received a form of punishment [47].

Another solution which companies such as Amazon Luna are patenting is to matchmake toxic players together in the same online lobbies [48]. A form of this has already been attempted by Valve in their game DOTA 2 whereby punished players get assigned to 'Low priority' matchmaking [49]. This involves restricting the game modes players can play and no rewards can be earned from the games. The only resolution for this is to win a certain number of games. Riot games is extending this solution in LoL by patenting a matchmaking system based on positive behaviour known as 'good

behaviour island' [50]. These techniques can be effective for frequent matchmaking multiplayer games such as shooters, battle royales or MOBAs it is arguable how possible it will be to implement in MMORPGs or online survival games where players can select their servers.

Communication between online players tends to be in the form of voice or text chat and can be delivered in the current played game or through the game system being played. The ADL survey showed in-game voice and text chats to have the most frequent harassment as shown in figure 1 [42]. The producers of LoL, Riot games, initially allowed communication with opposing teams, however this decision was reversed defaulting the chat to only be for the current team. Within a week LoL reported a reduction in negativity by 32% [51]. Text communication in gaming tends to be moderated such that no profanity or insults are shown. The characters are blocked out by symbols such as asterisks' however they are still displayed to others and in some games the entire word is not blocked out and thus the words can be guessable. Other games use quick pre-set chat which can be read out in-game or displayed to teammates. Such chats have timeout restrictions for repeated use by the same player to stop spam, although players can still abuse the system in between timeouts. Games such as Apex Legends have introduced text to chat whereby a typed message is read out by an automated bot. However, this is not moderated and can be used to send abusive messages which is common. Games allow muting and blocking of such chats at any point during such games. The ADL survey noted this as the most effective means of combating online hate [42].

Figure 1: Graphical representation of how most players receive toxicity as per the ADL/Newzoo survey. [42]



Rewarding good behaviour is another explored solution for the rife toxicity in gaming. Riot, the creators of LoL, announced in 2014 the rewarding of players who had committed no offences in the year with free skins [52]. The creators of Overwatch, Blizzard, also introduced an endorsement system for positivity. Furthermore, they introduced 'Looking for Group' which allows players to form teams which can filter on a minimum endorsement level. This introduction noted a 40% reduction in toxicity in the Overwatch community [53].

A psychological study noted how the subtlety of coloured messages can lead to higher levels of aggressive behaviour [54]. Riot games put this theory into further practice within LoL [51]. The company displayed positive behavioural statistics in blue or white such as "Players who co-operate with their teammates win X% more games" and negative behavioural statistics in red such as "Teammates perform worse if you harass them after a mistake". This improved the levels of negative attitude, abuse and offensive language by between 5-11% in-game. Displaying neutral messages in red had the opposite effect with an increase of 8-15%.

2.4 Reddit

Reddit is an online forum which primarily allows users to form communities and within them create posts, write comments and upload media. All of which can then be upvoted or downvoted by the community. This affects the visibility of the posts themselves, with the most popular being presented nearer the tops of pages. Signing up to Reddit is free and users can choose a personal username which cannot be changed and alongside it a display name which can be altered multiple times. Reddit promotes a positive atmosphere and users can generate ‘karma’ from quality posts and comments which reflects their standing in the community. Users are also able to create pages dedicated to specific topics or communities which are known as subreddits. The creators of such subreddits can become moderators to the pages and promote others to the status. A moderator manages the pages they are in charge of, they can set and enforce rules including removal of posts which violate such rules. Reddit Admins can also enforce content removal but are paid workers for the company. There are also automated moderators who remove posts automatically through certain violation criteria.

The site is extremely popular and as of Jan 2020 boasts 52 million daily users and over 100,000 communities [55]. The site is dominantly used by the US population boasting 222.63 million users in 2020, more than 12 times the country with the second most number of users [56]. As of December 2020, the site accounted for 49 percent of desktop traffic in the US [57]. A survey conducted on around 1500 US adults (18+) at the start of 2021 showed 36 percent of 18–29-year-olds use the platform as well as 22 percent between 30-49 [58]. This tails off to 10 percent for 50–64-year-olds and 3 percent for 65+. This decline in use as age increases is a typical reflection of internet use, although it is increasing year by year [59]. Whilst the users are weighted more towards the younger generation, there is also a notable difference in gender. The survey showed that 23 percent of Male respondents used Reddit as oppose to 12 percent Female.

Reddit promotes transparency and a released 2020 report showed that 6 percent of posts were removed from the website that year, standing at 233 million in total [60]. It was further outlined that out of these posts the majority were flagged as spam at 99.76 percent, with the remainder including harassment and hate speech. Users can also have their account suspended temporarily or permanently for repeat offences on the platform. The report showed around 135,000 permanent bans in 2020 with just under half due to hateful content.

3 Technical Background

3.1 Transformer Architecture

A transformer is a deep learning neural network model which focuses on attention mechanisms to perform well at NLP tasks. The transformer architecture consists of an equal stack of encoders and decoders which require a list of vectors as an input [61]. These encoders and decoders are broken down into further sub-layers. The encoder sub-layers consist of a self-attention layer which allows the model to grasp the association of words within a sentence. The result of this self-attention mechanism is a numerical score which associates a specific word with every other word in the sentence. As an example, take the sentence 'the dog didn't want to walk because it was tired'. When the model processes the word 'it', a high attention score will be calculated for the association with 'dog' as opposed to other words. This feature improves the encoding step within the transformer architecture. Once the self-attention layer is applied the output is then passed to a feed-forward neural network, a network whereby information is passed strictly forward through nodes with no cycles or loops. Each step has a layer-normalization step which surrounds it which helps reduce the computational expense and speeds up the training process [62].

The decoder contains the same layers with an extra attention layer that exists between the two, this focuses the model's attention to relevant parts of the input sentence. The self-attention layers differ from the encoder in that they only analyse the words preceding the current word. This is done via masking. Once all the attention and feed-forward layers are completed the vectors are passed to a linear layer which converts the input into a much larger logits vector represented by floats. Finally, the logits vector reaches a SoftMax layer which converts the floats into a collection of probabilities which sum to 1. The highest value is selected and the word associated with it presented as an output.

This transformer architecture is utilized in machine learning to focus on the context of words within a sentence. The results of applying this technique allow models to distinguish between the same words that have different meaning. This development in architecture marked an important step in the NLP field and creates a basis of the BERT model.

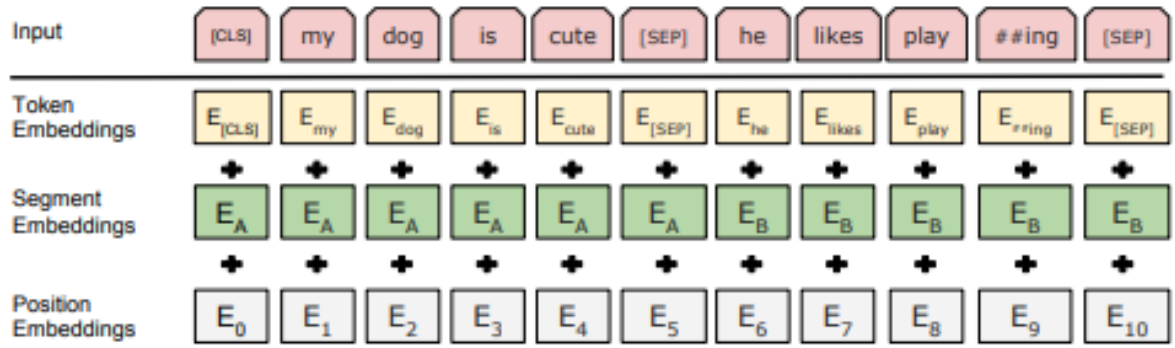
3.2 BERT Model

The Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art NLP model [19]. BERT takes the transformer architecture and throws away the decoder side resulting in a pure encoder transformer model. It builds on the semi-supervised sequence learning which first introduced a "pre-training" step [63]. Its use is incredibly versatile and can be applied to question answering, sentence pair classification, single sentence classification and single sentence tagging tasks. The model is showcased in two different sizes $BERT_{BASE}$ consisting of 12 layers and 110 million parameters and $BERT_{LARGE}$ of 24 layers giving 340 million parameters. $BERT_{BASE}$ was chosen to reflect the OpenAI models transformer size whilst $BERT_{LARGE}$ is used to achieve the best results for NLP tasks at the expense of computational power.

3.3 BERT Tokenizer

The BERT tokenizer takes input sentences with a maximum length of 510 and splits the words into tokens. It then adds a [CLS] token at the front which represents the classifier of the input and a [SEP] token to the end which is used for next sentence prediction. The tokenizer then converts the tokens into word embedding representations, positional embeddings of the words and a segment embedding which is an embedding representing which sentence the word comes from. A sum of the produced embeddings gives a final embedding of the input sentence ready to be passed into the pre-training or fine-tuning of the BERT model. A graphical representation of outputs is shown in figure 2.

Figure 2: A diagram showing the production of the embeddings from an input within the BERT tokenizer [19]



3.4 BERT Pre-training Steps

The pre-training of BERT has been applied to BookCorpus, a collection of free novel books, consisting of 800 million words as well as text passages from Wikipedia consisting of 2,500 million words. The training was completed separately for $BERT_{BASE}$ and $BERT_{LARGE}$ with both taking 4 days to fully train.

3.4.1 Masked Language Model (MLM)

Due to the bidirectionality of the BERT model each word is indirectly visible thus defeating the purpose of prediction. In order to combat this issue BERT adopts a masked language model (MLM) within the pre-training. The MLM technique masks 15% of the input tokens randomly. If a token is selected to be masked then the word is replaced with either a [MASK] token 80% of the time, a random token (RND) from the input sentence 10% of the time or the token is left unchanged (SAME) 10% of the time. This was the most effective balance as shown in figure 3. The technique notably improves the model for the Named Entity Recognition (NER) and Multi-Genre Natural Language Inference (MNLI) tasks.

Figure 3: A table showing the best balance of the MLM technique [19]

Masking Rates			Dev Set Results		
MASK	SAME	RND	MNLI	NER	
			Fine-tune	Fine-tune	Feature-based
80%	10%	10%	84.2	95.4	94.9
100%	0%	0%	84.3	94.9	94.0
80%	0%	20%	84.1	95.2	94.6
80%	20%	0%	84.4	95.2	94.7
0%	20%	80%	83.7	94.8	94.6
0%	0%	100%	83.6	94.9	94.6

3.4.2 Next Sentence Prediction (NSP)

The BERT model is also trained on next sentence prediction (NSP) task. This involves the prediction that a sentence follows one another, labelling the input sentences as either ‘IsNext’ or ‘NotNext’. A sentence following the preceding has a 50% chance of being the correct one or 50% chance to be another from the corpus. This further training improves the model for question answering and natural language inference as oppose to just using the MLM technique.

3.5 BERT Fine-Tuning

The fine-tuning of the BERT model is incredibly simple. Adding an extra layer on top of the core layers is enough to adequately tailor the model to different tasks. The self-attention mechanism in the transformer allows for this and only the task-specific inputs need to be fed into the model for it to be effective. The optimal hyperparameters were found to differ dependent on the tasks with the learning rate set at 5e-5, 3e-5, 2e-5 the batch size of 16 or 32 and the number of epochs between 2-4. This is reflective of further fine-tune testing that the learning rate should be above 2e-5 to avoid catastrophic forgetting, whereby a model abruptly forgets all previously learned information [64]. A learning rate of 4e-4 was notably too aggressive with the model failing to converge.

3.6 Machine-Learning Measurements

3.6.1 Log Loss Function

The log loss of a machine learning model is a numerical representation of how well the model predicts. A model having the minimum loss of 0 equates to a perfectly correct model whilst the maximum value of 1 reflects the worst. It is denoted by the following:

$$Log\ Loss = \sum_{(x,y)\in D} -y\log(p) - (1-y)\log(1-p)$$

Here, D represents the labelled dataset of (x, y) pairs, y is the labelled binary indicator being either 1 or 0 in value and p is the predicted probability of the observation. Reducing this for a machine learning model is paramount and referred to as empirical risk minimization. For BERT this is automated via the built in 'compute_loss' method.

3.6.2 Precision

Precision is the proportion of positive identifications which were truly correct. Its measurement is given by the number of true positives divided by the total positives.

$$Precision = \frac{TP}{TP + FP}$$

3.6.3 Recall

Recall is the proportion that actual positives are identified correctly. Its measurement is calculated by the number of true positives over the sum of the true positives and the false negatives the model predicted.

$$Recall = \frac{TP}{TP + FN}$$

3.6.4 F1 Score

The F1 score looks to find a balance between precision and recall collating the harmonic mean of them both. It indicates a machine learning model's classification effectiveness with the value between 0 and 1, the former being the worst model possible and the latter being the a perfectly correct model.

$$F1\ Score = 2 \times \frac{Precision\ \times Recall}{Precision + Recall}$$

4 Methodology

To pursue the comparison of toxicity between gaming communities a clear definition of toxicity needs to be decided upon. Further representation of this definition via a classification will allow automatic detection of occurrences using the chosen machine learning model, BERT. The model will be applied to gaming communities hosted by the platform, Reddit. Data from multiple gaming subreddits will be extracted and then thoroughly cleaned. The BERT model will require fine-tuning with the chosen classification and can then be further improved with extra self-annotated data from the extracted subreddits. The model can then be applied to the cleaned Reddit data to investigate some specified hypotheses.

4.1 Definition of Toxicity

The definition chosen for toxicity in this thesis is 'Intentional unpleasant behaviour used to aggrieve others online'. The 'unpleasant behaviour' part of this definition can be represented via the OLID level A classifier that a post is offensive (OFF). The 'use' of this unpleasant behaviour to 'aggrieve others online' can be expressed via the OLID level B classifier that a post is also targeted (TIN). This combination allows for automatic implementation to fully represent the outlined toxicity definition. The clear definitions of each OLID level are expressed in figure 4.

Figure 4: Definitions of NOT/OFF and TIN/UNT classifiers from the OLID paper [9]

Level A:

Not Offensive (NOT): Posts that do not contain offense or profanity;

Offensive (OFF): Posts containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words.

Level B:

Targeted Insult (TIN): Posts containing insult/threat to an individual, a group, or others;

Untargeted (UNT): Posts containing non-targeted profanity and swearing. Posts with general profanity are not targeted, but they contain non-acceptable language.

In order for a post to be marked as Level B it must first be marked as offensive (OFF) at Level A. Truly toxic posts will require classification of OFF as well as TIN whilst non-toxic posts will be marked as either not offensive (NOT) or as offensive yet untargeted (OFF and UNT) as per table 1.

Toxic	OFF + TIN
Non-Toxic	NOT/ OFF + UNT

Table 1: Shows the toxicity classifiers chosen for this thesis.

4.2 Coding & Storage Platform

In order to implement the BERT model a high-performance machine is required given the tough workload it takes to train the models. Initially, Bristol University's high-performance computer

known as BlueCrystal was going to be the option. However, this was reconsidered due to the long waiting times and the added complexity of having to submit jobs. Another option was to find an online IDE which allow users to run code on other systems through a browser. Such websites include Paperspace Gradient, Kaggle, Google Colab, Amazon SageMaker and FloydHub.

For this thesis Google Colab was chosen given the amount of BERT models available and their walkthroughs. It also easily links up to Google Drive which is a safe place to store the collected Reddit data. The platform allows access to Google's powerful GPUs such as the Nvidia K80s, T4s, P4s and P100s or Google's very own TPUs [65]. Unfortunately, the processors cannot be chosen but all available GPUs offer enough power to complete the modelling process. Another notable aspect of the platform is that the interface uses cells, this helps split up the code and can be ran individually aiding the debugging effort.

4.3 Data Collection

The first focus of this project will be to collect Reddit data to apply BERT model classification to. Specific subreddits are the primary source of data and a range of gaming pages have been chosen in order to cross compare toxicity. Initially, 46 gaming subreddits were investigated as potential use to run final predictions on. Given the amount of time it takes to extract and clean the data as well as run predictions on the final list was reduced to a range of 14 popular subreddits from the initial amount. These were chosen based on popularity whilst maintaining a good balance between multiplayer and single player games as well as having a broad range of genres. The final chosen games are outlined in table 2.

For the purposes of this research any references of the following definitions are as follows:

Multiplayer – A game that is dominantly online player vs online player (PvP). The game may include single-player aspects but the assumption is that the majority are playing the multiplayer modes.

Single Player – A game that is dominantly player vs computer-controlled AI also known as player vs enemy (PvE). The game may include multiplayer aspects but the assumption is that the majority are playing the single player modes.

Subreddit Name	**Game Type**	**Description**	**Subreddit Community Members**
League of Legends (LoL)	Multiplayer	An extremely popular multiplayer online battle arena (MOBA). It ranked 3rd in the world for highest eSports earnings in 2020. [66]	5.3m
Rocket League	Multiplayer	A popular vehicular football video game. It ranked 10th in the world for highest eSports earnings in 2020. [66]	1m
Among Us	Multiplayer	A recently popularized social deduction game.	701k
Rainbow 6 Siege	Multiplayer	An online tactical shooter game. It ranked 6th in the world for highest eSports earnings in 2020. [66]	1.4m
Apex Legends	Multiplayer	A free-to-play online battle royale shooter, supporting cross-play across different platforms.	1.6m
Grand Theft	Multiplayer	An open world action-adventure game loosely	1m

Auto 5 (GTA) Online		depicting life in Southern California.	
Fifa	Multiplayer	One of the most celebrated football games.	518k
Minecraft	Single Player	An open-world sandbox focusing on survival and building.	5.4m
Dark Souls III	Single Player	An action role playing game (RPG) which is well known for its extreme difficulty.	427k
Assassins Creed	Single Player	A historical depicting action-adventure series focusing primarily on stealth.	411k
No Mans Sky	Single Player	A science fiction exploratory survival game. Notable for its disappointing and deceiving release.	568k
Total War	Single Player	A strategy game series which has focused mostly on historical scenarios but also includes sci-fi Warhammer spin-offs.	312k
The Sims	Single Player	A life simulation game.	429k
The Elder Scrolls V: Skyrim	Single Player	An immensely popular open-world fictitious action RPG.	1.1m

Table 2: Final chosen gaming communities for analysis

For data collection the data scraping tool called Pushshift Multithread API Wrapper (PMAW) was used [67]. The Pushshift Reddit dataset is a huge collection of posts gathered since Reddit's creation. The API accesses the data and through use of multithreading and asynchronously extracts the needed data. This speeds up the time it takes to extract large quantities of posts compared to other API scrapers. The posts are extracted on each subreddit since 1st Jan 2021 and a limit of 500k is applied to avoid extremely large datasets. The .csv files produced contain only one column containing the extracted subreddit posts and comments.

4.4 Data Cleaning & Conversion

One issue in extracting such comments from forums is that many contain non-ASCII characters. For the sake of standardisation, the prediction datasets will only contain ASCII characters. A script was produced which takes any non-ASCII character and removes the instance from the extracted comments. Removing the letters was the best option so that the number of posts was not affected. This did lead to some lines being completely blank given the entire post was made up of non-ASCII characters and thus also were removed. This had very little impact with between 0.01-0.4% of posts being removed for this reason.

The initial collected data contained instances of '[removed]'. These are posts removed by either a human moderator, an auto-moderator or the spam filter. It is interesting to keep track of the amount of these given that they are often linked to toxic behaviour. There are also '[deleted]' posts which are either deleted by the user themselves or that the users account has been deleted. A users account can be deleted by the admins on Reddit if the user is a repeatedly engages in behaviour violating their contents policy. This includes inciting violence, promoting hate based on identity and vulnerability as well as threatening behaviour. It is safe to assume that some of the '[removed]' and '[deleted]' posts will be due to toxic based occurrences.

Another notable issue with some posts is that they are ‘bot generated’. Such posts were dominantly an automated removal of user posts either due to constraints set up by the subreddits themselves, due to spam or that the user has violated Reddit’s contents policy. These posts mostly contained one or more of the following strings ‘I am a bot’, ‘I’m a bot’ or ‘Your post has been removed’. The amount of these posts varied wildly and seemed dependent on whether the moderators had set up such bots. In some instances, there were substantial amounts. Furthermore, it was difficult to detect if such posts were removed due to toxicity and depended on the notation of the bots. Some bot removal instances described the nature of the posts which clearly noted the reasoning of removal whilst others did not. Nonetheless, the vast majority did not constitute as ‘toxic’ and thus were also removed entirely from the datasets. Table 3 outlines the amount of cleaning per subreddit.

Subreddit Name	**Posts available since 1st Jan 2021**	**Original Posts Obtained**	**Final Data Amount after Cleaning**
League of Legends (LoL)	2,059,782	500,000	467,646
Rocket League	564,796	500,000	465,140
Among Us	233,185	233,185	199,388
Rainbow 6 Siege	474,723	474,573	448,217
Apex Legends	1,944,480	500,000	474,477
Grand Theft Auto 5 (GTA) Online	843,196	500,000	466,751
Fifa	871,893	500,000	473,808
Minecraft	2,012,702	500,000	459,234
Dark Souls III	260,574	260,574	253,855
Assassins Creed	185,226	185,226	173,122
No Mans Sky	341,221	341,139	325,924
Total War	429,278	429,278	419,734
The Sims	127,630	127,630	123,493
The Elder Scrolls V: Skyrim	396,260	396,260	388,307
Total:			5,139,096

Table 3: Data Cleaning for each chosen gaming Subreddit

Another potential issue within the collected data is the length of the posts themselves. The BERT model has a limit of input tokens at 512, with 2 tokens reserved for the SEP and CLS parts this is actually 510 for the maximum input text. This was relatively common and much of the data exceeded this value after pre-processing. The BERT model will handle this by default and truncate the sentences that exceed the max value. A study investigating the fine-tuning of BERT trialled the handling of such instances [64]. It found out that the most effective way to handle long strings is to take a section of the start and the end of the string and concatenate the two. This had the least error rate at 5.42% as opposed to only taking the beginning at 5.63% or the end at 5.44%. Whilst this finding is interesting it is not a substantial increase and would take a long time to implement. Therefore, it was decided to allow the BERT model to apply its default truncation in taking just the beginning of the sentences.

4.4.1 Non-English Word Removal

The removal of non-English characters was explored using the langdetect Python package. The langdetect package is Google's language-detection library ported from Java to Python. The tool detects up to 55 languages and has 99% precision on 53 of these [68]. A script was created to feed in the collected data from .csv files and compile the output sentences with the langdetect results. The langdetect method had issues dealing with comments which started with non-alphanumeric characters as well as comments which began with URLs. Such comments were omitted from the langdetect method and marked as 'NA' to allow the script to continue.

The script was tested on the Among Us Reddit data given it is smaller size of around 200k and has some noticeable occurrences of foreign words. The runtime of the script was 1 hour 59 minutes and showed very poor results. The program detected 32 languages and in total flagged 77% as English, 17% as non-English and 6% as 'NA'. Upon inspection of the non-English comments it was apparent the majority were truly English. This was further confirmed by manually checking all of the Spanish marked comments.

The script marked 872 comments as Spanish, manually checking these results showed only 94 comments to be truly Spanish yielding accuracy of only 11%. The langdetect method notably struggled with short comments. The distribution of < 50 characters and >= 50 within English marked text was 43% and 57% respectively. However, non-English flagged text was extremely skewed, < 50 character comments accounted for 97% of this dataset. This was also largely reflected in Spanish marked comments which were not truly Spanish and < 50 characters, making up 87.5% of the data.

Further observation of the comments showed that langdetect clearly struggles with typos, abbreviations and online slang. Given the length of time to run the script, the poor accuracy and small percentage of truly foreign comments it was decided to avoid all removal of non-English comments.

4.5 Creation & Implementation of the BERT Models

The chosen model for implementation is the BERT model, specifically the BERT-base-uncased. This model is derived from an initial walkthrough presented by Chris McCormick for one of the GLUE benchmark classification problems, the Corpus of Linguistic Acceptability (CoLA) [1]. From here the walkthrough was tweaked to handle OLID classifiers and split into two models, one for handling NOT/OFF and the other for dealing with UNT/TIN classification.

4.5.1 Setup

The models start by monitoring the GPUs available using the torch package from PyTorch. PyTorch is an open-source machine learning library mostly developed by the Facebook AI team. It's primarily an optimized tensor library used for deep learning on GPUs. The tensors are used throughout the BERT models but the package is also used to set the device to utilize the available GPU.

The Hugging Face transformers package is then installed. This package contains high-level APIs which allow easy utilization of transformer-based models. This includes the BERT pre-trained model as well as the BERT tokenizer. Simplified methods from the package allow for easy manipulation without requiring strong technical understanding of the underlying processes.

The next step is to import the training and testing data to be implemented by the model. Initially this involves 'mounting' the Google drive which contains all of the train and test data. Once mounted the Google drive directories can easily be referenced to pull in the files needed by the model. To capture the incoming data the pandas dataframes package is imported. This package allows the creation of data structures as either a 1-dimensional array known as a series or a dataframe, a 2-dimensional tabular structure. The data is imported using the pandas dataframes and split into a tabular training and test set.

4.5.2 BERT Tokenizer

Once the setup is complete the *BERTtokenizer* object is loaded from the Hugging Face transformers package. The BERT tokenizer is loaded specifically using the '*frompretrained()'* method which loads the model as specified. For this thesis the model being applied is the BERT-base-uncased model which ignores the casing of the inputs. The *to_lower_case* option is set to 'true' which converts the inputs all to lower case during tokenization. All other options are left as default.

Once the tokenizer is initialized it is then applied by calling the *encode_plus()* method on all of the input data. This method allows multiple settings to be applied during the tokenization. Firstly, the *add_special_tokens* setting is set to 'True' so that the tokenizer adds [SEP] and [CLS] tokens to the inputs. The BERT model only accepts all inputs at the same length hence the *max_length* is required and *pad_to_max_length* is set to 'True' so that the inputs are also padded with [PAD] tokens if too short. The tokenizer will also truncate to the *max_length* if the input exceeds it. The BERT model also requires an 'attention mask' for the padding, this mask is the form of a Boolean array which is set to 1 if the token is a [PAD] token, else 0. This tells BERT where the padding tokens are and that they should not be analysed. The *return_attention_mask* option is set to 'True' so that it can be used later in training the model. The tokenizer will also convert words into token, positional and segment embeddings and return the final summed embedding ID. Finally, the *return_tensors* option is set to 'True' such that PyTorch tensors are returned from the tokenizer which is the format the BERT model accepts. The returned values from the tokenizer are compiled into final tensors to be fed to the BERT model using the TensorDataset method from PyTorch.

Now that the tensors are ready for the model these need to be split up into batches. To do this the Pytorch Dataloaders object is called. The Dataloader is a class which allows an iterable over a chosen dataset. This helps save RAM as an iterator stops the entire dataset being loaded into memory. The Dataloader object is split by a specific batch size given and the data is loaded using the PyTorch SequentialSampler for the test data and the RandomSampler for the training data. The data is now ready for the training steps.

4.5.3 Training the Model

Firstly, the chosen model is loaded from the transformers package. For this thesis the *BertForSequenceClassification* model is used and imported. Options for the model can be specified within the *from_pretrained* method. Here the specific BERT model is chosen which is the BERT-base-uncased model, the number of labels is also required which for this classification is 2. The parameters *output_attentions* and *output_hidden_states* is set to 'false' as these are not required and save on memory.

Next the *AdamW* optimizer is loaded from the HuggingFace transformers package for use within the training steps. The Adam optimizer is a stochastic gradient descent optimization algorithm for training deep learning models [69]. Its use is to adapt the learning rate in between epochs. *AdamW* optimizer differs from Adam such that the weight is applied after controlling the step size making the algorithm more efficient. Here the alpha learning rate is specified as well as the epsilon value which is used to avoid a potential division by zero. Now the learning rate scheduler is imported from the HuggingFace library and given the optimizer. This allows for the application of the optimizer in between the epoch steps. Here the total amount of steps is specified which are calculated from the amount of epochs times the length of the dataloader. The scheduler allows for warm up steps if required but is not utilized in this model and set to zero.

The training loop is now setup. Initially the *random* package is imported and a random seed value assigned to the PyTorch *manual_seed* method. The method generates random numbers and given the seed these random numbers will be the same in between runs which allows for replication of the trained model. Measurements are then initialised to capture the time taken and the loss values. Next the loop within each epoch begins and starts by setting the model into training mode. Then the training loop begins which loops per batch from the training dataloader. Here each batch is unpacked into individual variables containing the tokenized IDs, attention masks and the labels for each comment. The model then clears any previous calculated gradients using the *zero_grad()* method. The next step trains the model by feeding in unpacked variables from the batch, this leaves a tensor object which contains the models resulting loss and predicted logits. The loss is then accumulated into a total average and a backward pass is applied to this result using PyTorch *.backward()* such that the model gradients can be calculated. The PyTorch *clip_grient_norm()* method is then called which 'clips' the gradient to a specified value if it exceeds it. The value is set to 1.0 which safeguards any optimization of the model if the gradients were to explode. Next the PyTorch *step()* method is used on the optimizer to alter the hyperparameters based on the computed gradients, it is applied again on the scheduler to apply the newly acquired values. Finally, the cumulative loss and time taken for the model to run is calculated and printed to the console.

Once the training is complete the model is set to 'evaluation' mode and ran on the test data in a similar loop to the training. The main difference is the removal of the loss calculation and the optimizer. The model performs predictions on the test data such that machine learning metrics can be calculated. Originally the walkthrough contained a calculation of accuracy, however this was removed and replaced with the sklit learn classification report. The method is loaded from the sklearn package and is simply given the true labels as well as the predictions. This results in a table containing the precision, recall, accuracy and F1 scores. This proved useful for the ON/OFF model however was not truly reflective of the TIN/OFF model. Since the TIN/OFF model was passed the already correctly labelled OFF comments the calculation was not accurate as a combination of both models. This had to be hand calculated by running predictions after being given the classified OFF posts by the previous

model. It was still useful to get a separate view of both models however to see where individual model improvements could be focused upon.

Once the metrics are calculated the final trained model is saved to the Google drive which can be loaded for future use. The final model can easily be loaded from the drive to be used for the prediction steps in between Google Colab sessions. This involves replacing the *frompretrained()* parameters with the directory containing the saved model. This alteration was made for the predictions so that the same model was used in between large prediction runs.

4.5.4 Running Predictions

To run predictions a function was created which takes in a string and then applies the pre-loaded tokenizer to it. It then applies the trained model to the tokenized string and returns the classifier value. The Reddit data is imported into a pandas dataframe and then looped through the function pulling out a post each loop and thus obtaining a returned classifier for it. The result is compiled alongside the evaluated post into a dataframe. It is then converted into a .csv and then posted to the Google drive for further analysis.

One interesting development within the collected data has instances of posts which are a single word ‘NA’. The BERT model fails when running these posts through the model. It appears that one of the layers of BERT is expressing this value as the null character, forcing an error out of the model and the entire process to stop. Occurrences of this were extremely minimal with only 3 of 12 games having them, 12 within LoL data, 3 in Rainbow 6 and 1 in Rocket League. It was decided to remove all instances to allow the modelling to proceed on the data.

The prediction runs for ON/OFF were computationally tough and in total took a week to run for all the datasets. One notable issue was the occurrence of disconnections during such lengthy runs. It was decided to split the datasets into smaller quantities to be able to have smaller more frequent runs with less chance of a disconnection. This allowed for less disconnections and also reduced the predicting time quite substantially from 5 hour runs to multiple 1 hour runs. This was not necessary for the UNT/TIN model which was only given the OFF labelled comments from the previous model considerably reducing the number of posts to predict.

4.6 Improving the Model

4.6.1 Annotating Reddit Data

Due to the OLID training data being based on tweets and the application of the BERT model applied on Reddit data it was concerning how well the model would translate. Furthermore, the model requires test data made up from the Reddit data to give accurate metrics. Due to this it was decided to annotate a portion of the Reddit data for further testing and investigate improvement to the model.

In order to create Reddit labelled datasets the OLID annotation process need to be accurately followed so that the data reflects well with the current labelling. The labelling of OLID was thoroughly investigated so that the annotations closely followed suit. The ON/OFF labels represent if

the text contains offensive material which can either be in the form of a profanity, insult or swearing. The TIN/UNT labels are given to only those which are already marked as OFF. The TIN specifies if the offense is targeted towards an individual or a group whereas the UNT is a non-targeted offense.

Firstly, a program was created to extract 100 random samples from each of the 14 subreddit gaming datasets. Taking samples from each of the subreddits will help account for some game specific wording. Initially this was performed twice, once for training data and the other for a test set. It was further rerun another two times to add to the quantity of training and test data. The data was checked between runs such that no repeat posts had been grabbed between the runs. In total 5 comments were removed due to them being a different language and 6 which were bot posts. Due to this the bot cleaning was revisited for some extra removal. Table 4 shows the final amount of OLID classified data.

Level A	Level B	OLID Training Data	Reddit Batch 1 Training Data	Reddit Batch 2 Training Data	Reddit Test Data
OFF	TIN	3,876	38	55	76
OFF	UNT	524	97	168	169
NOT	-	8,841	1261	2573	2547
Total		13,241	1397	2797	2792

Table 4: A table outlining the split of classifiers for OLID and Reddit annotated data

A notable issue was the use of offensive language towards gameplay aspects, game developers or towards the commentor themselves. Given that the sentences contained swearwords or profanity they were all marked with the offensive tag OFF. For the Level B classifiers given that gameplay aspects are not people they were labelled as untargeted (UNT) similarly, self-addressed abuse was also labelled untargeted. However, hate directed at game developers was included as targeted (TIN) given the definition of directing hate towards a group.

Another issue in annotating was the use of specific slang or acronyms. Some of the terms are online slang or abbreviations whilst others are specific to certain games. Some terms required further explanation so the website *UrbanDictionary.com* was used for such definitions. Labelling some of these proved tricky such as ‘LMFAO’ which stands for ‘laughing my fucking ass off’. Whilst this contains a swear word its general use is for positivity. Due to this and its common occurrence, such abbreviations were marked as not-offensive if used to express positivity. Some examples are outlined in table 5.

Abbreviation/ Term	Definition	Level A Classifier
STFU	Shut the fuck up	OFF
LMFAO	Laughing my fucking ass off	NOT
MF	Motherfucker	OFF
AFK	Away from keyboard	NOT
OP	Overpowered	NOT
NPC	Non-playing character	NOT
That’s sick	Something really good, cool or very impressive.	NOT
Simp/ Simping	Being an obsequious person.	OFF
Clutch	When a player is victorious with odds largely against them.	NOT

Noob	A derogatory term for an inexperienced player.	OFF
Git gud (Get good)	A term used to heckle inadequate players to get better at a game.	OFF
Tryhard	Used to describe someone whose effort and emotional investment is excessively high.	OFF

Table 5: Showing some common abbreviations and terms within the collected Reddit data

4.6.2 HateBERT

Further improvement to the toxic model was investigated with implementation of two HateBERT models. These were created to compare any improvements against the BERT model at individual level A and B OLID classifiers. The reproduction of the model is incredibly simple given its open-source availability online. HateBERT involves introduction of a large Reddit dataset known as RAL-E containing banned posts from the platform. This RAL-E dataset contains 43 million tokens and is integrated into the original BERT pre-training data to shift the model towards better social media language detection and polarity. Hence the fine-tuning of the model is still required and thus the model is exactly the same as the previously outlined BERT models. The settings of the hyperparameters were also kept the same as the BERT models. The only difference is that the *frompretrained()* method loads in the HateBERT model by specifying the saved model directory.

4.7 Prediction Comparisons

With the predictions complete some exploratory hypotheses were of interest. The literature review showed that toxicity is exerted within highly competitive multiplayer games. However, there is a lack of research into whether toxicity occurs on forums at a similar level for single player games. Therefore, the first alternative hypothesis is:

H_1: Multiplayer subreddits have a higher occurrence of toxic posts.

Given the large amount of data cleaning for each subreddit it is also interesting to investigate if there is a link between the residing levels of toxicity. For this hypothesis the subreddits are grouped by those which had greater than 5% of its posts excluded and less than or equal to 5%. If a subreddit has more posts excluded then it is more moderated. A subreddit which is more moderated could be expected to show less levels of toxicity. However, it could also mean that there are much higher rates of toxicity and that including moderation maintains a similar level. Therefore, the second hypothesis test will be:

H_2: There is a difference of toxicity between more moderated pages (>5% excluded) than less moderated pages (<=5% excluded).

The age rating of the chosen subreddits will be another factor to explore within toxicity. Given that Reddit is predominantly used by US users [56] the US age rating system will be used. This is the Entertainment Software Ratings Board (ESRB) developed for the United States and Canada. The age ratings are rated either; ‘E’ for everyone, ‘E10’ for ages 10+, ‘Teen’ for 13+, ‘Mature’ for 17+ audience and ‘Adults Only’ for 18+. The chosen groupings are games for those who are of a mature age or greater (>= 17) against any audience aged lower (< 17). Officially sold games in these countries cannot be bought by those younger than the age rating and thus the audience is expected to be older. However, it cannot be assumed that posts within the subreddits are made by children as

young as the rating since Reddit has an age rating itself of 13+. Nonetheless, a lower age rated game is expected to be less frustrating and thus should reflect a less toxic community. Therefore, a final alternative hypothesis is:

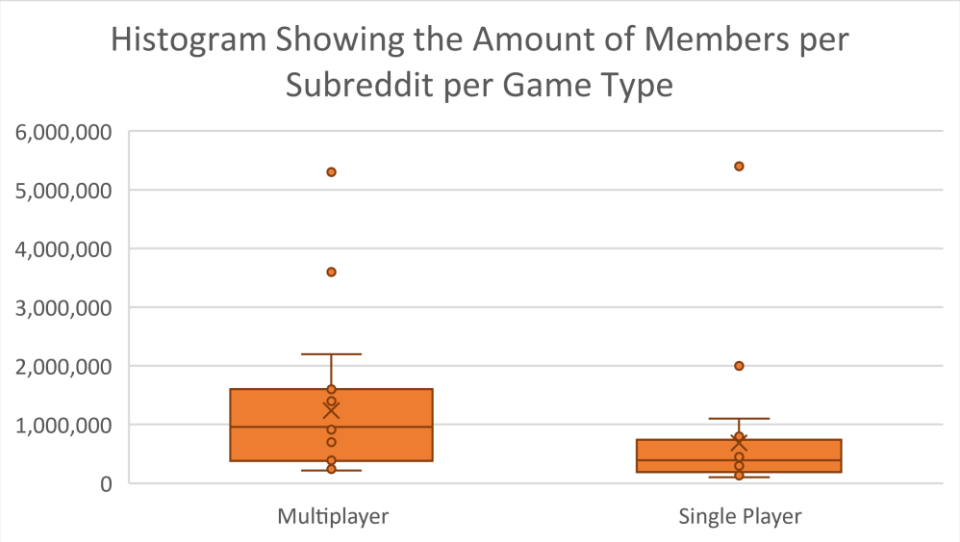
H_3: Mature rated games will have higher toxicity than those with a lower age rating.

5 Analysis & Critical Evaluation

5.1 Data Collection

To obtain a good balance between popular subreddit communities for use in this thesis, 46 different gaming communities were initially investigated. These spanned across differing genres and game types providing a wide variety of communities to analyse.

Figure 5: Histogram Showing the Amount of Members per Subreddit Game Type



The number of followers within the 46 gaming communities differed greatly with the least popular at 99k to the most at 6.4 million as shown in figure 5. Multiplayer games typically had much more followers than single player games with almost double the average amount. However, the most followed game was the single player game Minecraft. Whilst these gave an indication of popularity it did not necessarily mean the pages were very active. Therefore, it was better to review the number of posts made on the pages within a specific timeframe.

Figure 6: Histogram Showing the Amount of Posts Between Jan-July 2021 per Subreddit Game Type

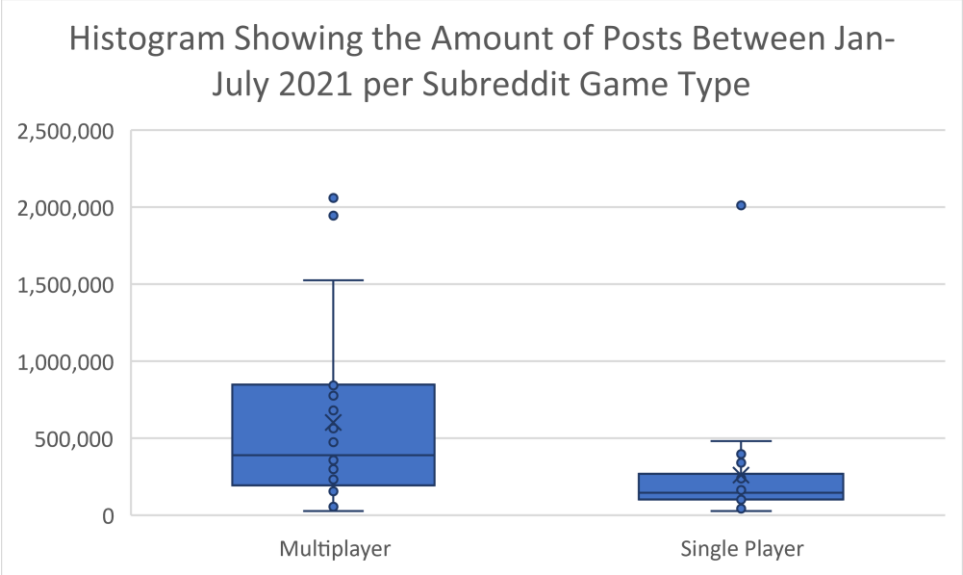


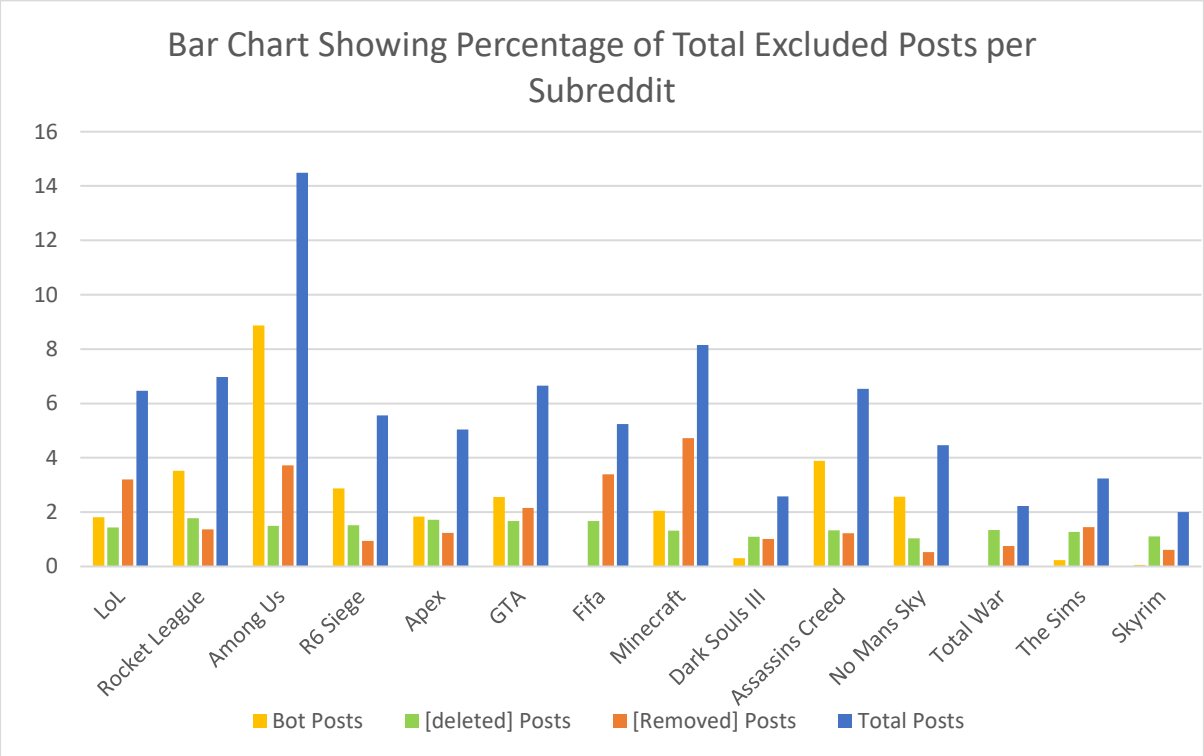
Figure 6 shows between Jan-July 2021, multiplayer games had almost double the average amount of total posts at 1.2 million compared to single player games. 9 out of 10 of the top 10 posts were multiplayer games with Minecraft as the most popular single player game with more than 4 times as many posts as the second most active single player page. For selecting the games for the BERT model, a strict limit of no less than 100,000 comments was chosen. Out of the 8 games under 100,000 posts, 5 were single player. One notable low comment count was PlayerUnknown's Battlegrounds (PUBG) which only had 55k posts in this period yet has amassed 1.7 million followers. This shows that the member count is not a good measurement for how active the subreddit pages are.

The initial 46 gaming subreddit pages were reduced to a final 14 given the length of time to extract, clean and run predictions. The 14 pages were taken forth into a thorough data cleaning process.

5.2 Data Cleaning

The cleaning process adopted for this study removed a large number of posts based on Reddit moderation. Figure 7 outlines the spread of this in between the chosen Reddit communities.

Figure 7: Bar Chart Showing Percentage of Total Excluded Posts per Subreddit



The removal of posts varied wildly between the 14 chosen Subreddits. Most notably is the difference in 'bot' posts with Among Us removing the most at nearly 9% which more than double the second highest, the Assassins Creed Subreddit. This is entirely dependent on the setup of the Subreddit by the moderators with Among Us having very strict posting criteria. In order to post on the Among Us Subreddit a user needs to amass a certain amount of 'karma' in order to be able to post. The Assassin's Creed bot posts were high mostly removed due to the posts asking a 'too common' question, being focused on tech support or that it contains spoilers for the latest released titles. Some Subreddits stated their reasoning of removal clearly, Rocket League specifically categorized posts and it could be easily seen that around 5% of the bot posts were due to toxicity and hate speech. Unfortunately, no other Subreddits offered the specifics and so couldn't be easily compared.

Another varying number of discarded posts were [removed] posts. These are such posts which are deemed offensive by the moderators or break the rules of the sub-reddit. The Minecraft subreddit showed the highest amount of [removed] posts at 4.7% with Among Us second at 3.7% and Fifa third at around 3.4%. [deleted] comments were a lot more consistent accounting between 1-1.8% of the Subreddits posts. Such posts are either deleted by the user which had posted or if the users account has been deleted.

Figure 8: Box and Whisker of Cleaned Data per Game type

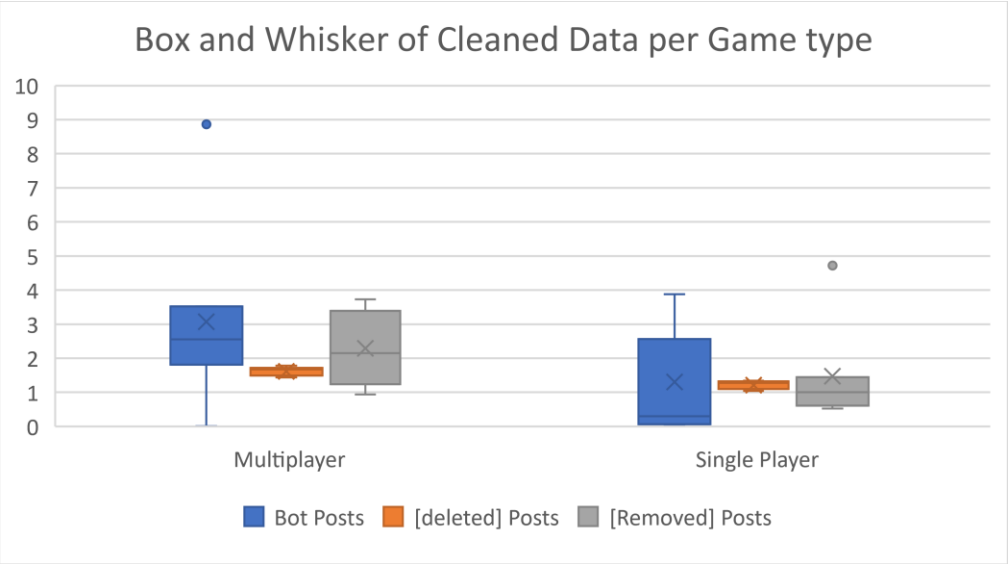


Figure 8 shows that single player games require slightly less cleaning compared to multiplayer games. The multiplayer 'bot' removed posts were more than double the single players at 3% compared to 1.3%. These also had a much bigger spread between the games for both types which shows the difference in moderation. The [removed] posts showed a slight increase of 0.7% for the multiplayer games, Minecraft within the single player was an obvious outlier which significantly impacted the difference. Comparing medians showed a slightly bigger difference in [removed] posts at 1.15%. The [deleted] posts showed a minor difference at 0.4%.

These findings reflect the higher popularity of multiplayer games on Reddit and hence require more moderation. It will be interesting to see if single player games contain more instances of toxicity given that they seem less moderated. However, single player games may not require as much moderation since players have less frustration from such games. This is explored later via hypothesis H_2.

5.3 BERT Optimization

In order to obtain the best model possible for the predictions a number of test optimizations were ran. The BERT paper outlines that fine-tuning a model with; learning rates of 5e-5, 3e-5, 2e-5; epochs of 2, 3, 4 and batches of 16 or 32 yields good results. Hence, it proved useful to test these options on the OLID classifiers A and B which will be used for the toxicity definition. Upon testing the different settings, it became apparent that epochs of 2 and 3 were too small and the loss was not converging enough. Due to this only an epoch value of 4 was compared for recall, precision and F1 scores. Table 6 outlines the testing of these different hyperparameters.

		NOT			OFF			Macro Average		
Batch	**Learning Rate**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
32	5.00E-05	0.88	0.88	0.88	0.69	0.68	0.69	0.78	0.78	0.78
32	3.00E-05	0.88	0.9	0.89	0.72	0.68	0.7	0.8	0.79	0.79
32	2.00E-05	0.88	0.9	0.89	0.72	0.68	0.7	0.8	0.79	0.79
16	5.00E-05	0.87	0.89	0.88	0.69	0.66	0.68	0.78	0.77	0.78
16	3.00E-05	0.88	0.89	0.89	0.71	0.7	0.7	0.8	0.79	0.8
16	2.00E-05	0.89	0.9	0.89	0.73	0.7	0.72	0.81	0.8	0.81
		TIN			**UNT**			**Macro Average**		
Batch	**Learning Rate**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
32	5.00E-05	0.91	0.97	0.94	0.58	0.25	0.35	0.74	0.61	0.65
32	3.00E-05	0.92	0.95	0.94	0.52	0.4	0.45	0.72	0.68	0.69
32	2.00E-05	0.91	0.97	0.94	0.57	0.29	0.39	0.74	0.63	0.66
16	5.00E-05	0.93	0.97	0.95	0.7	0.44	0.54	0.81	0.71	0.75
16	3.00E-05	0.91	0.96	0.93	0.5	0.29	0.37	0.7	0.62	0.65
16	2.00E-05	0.93	0.96	0.94	0.63	0.44	0.52	0.78	0.7	0.73

Table 6: Table of the differing hyperparameters for optimization performed on OLID test data.

Generally, the BERT model performed better than the SVM, BiLSTM and CNN models presented in the original OLID paper as presented in figure 9. The BERT model is generally consistent with the Level A classifier showing a minor improvement between hyperparameter settings. The TIN/UNT varied a lot more between the settings mostly due to the very low count of UNT within the test data. The recall was low for the UNT classifier and actually fared worse in all instances to the CNN. However, it showed a much higher precision showing that the model is selecting the majority correctly from those the model thinks is UNT. This is likely due to the heavily weighted TIN occurrences in both sets of OLID training and test data accounting for around 88% and 89% respectively.

Figure 9: Metric measurement for NOT/OFF and TIN/UNT for SVM, BiLSTM and CNN models from the OLID paper [9].

	NOT			OFF			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
SVM	0.80	0.92	0.86	0.66	0.43	0.52	0.76	0.78	0.76	0.69
BiLSTM	0.83	0.95	0.89	0.81	0.48	0.60	0.82	0.82	0.81	0.75
CNN	0.87	0.93	0.90	0.78	0.63	0.70	0.82	0.82	0.81	**0.80**
All NOT	-	0.00	0.00	0.72	1.00	0.84	0.52	0.72	0.	0.42
All OFF	0.28	1.00	0.44	-	0.00	0.00	0.08	0.28	0.12	0.22

	TIN			UNT			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
SVM	0.91	0.99	0.95	0.67	0.22	0.33	0.88	0.90	0.88	0.64
BiLSTM	0.95	0.83	0.88	0.32	0.63	0.42	0.88	0.81	0.83	0.66
CNN	0.94	0.90	0.92	0.32	0.63	0.42	0.88	0.86	0.87	**0.69**
All TIN	0.89	1.00	0.94	-	0.00	0.00	0.79	0.89	0.83	0.47
All UNT	-	0.00	0.00	0.11	1.00	0.20	0.01	0.11	0.02	0.10

The Level A NOT/OFF classifier shows the best classification metric with a batch of 16 and learning rate set to 2.00E-05. Whilst the Level B TIN/UNT has the best metrics with a batch of 16 and learning rate 5.00E-05. These hyperparameter settings were used for both models for the Reddit toxicity detection models.

		OFF/ NOT Loss			
Batch	**Learning Rate**	**Epoch 1**	**Epoch 2**	**Epoch 3**	**Epoch 4**
16	2.00E-05	0.47	0.35	0.24	0.16
		TIN/ UNT Loss			
Batch	**Learning Rate**	**Epoch 1**	**Epoch 2**	**Epoch 3**	**Epoch 4**
16	5.00E-05	0.34	0.28	0.18	0.1

Table 7: Table showing the loss of the selected hyperparameter models.

The loss of the chosen models converged relatively well more so for the UNT/TIN model. It would have been good to test some further epochs greater than 4 to check that the loss was optimum in its convergence. Nonetheless, the chosen models had a good balance of minimal loss alongside a higher F1, recall and precision as shown in table 7.

5.4 Modelling Toxicity

5.4.1 Level A Model

The initial model was setup to categorize firstly the level A classifier for offensive (OFF) and non-offensive posts. The metrics of such are outlined in table 8.

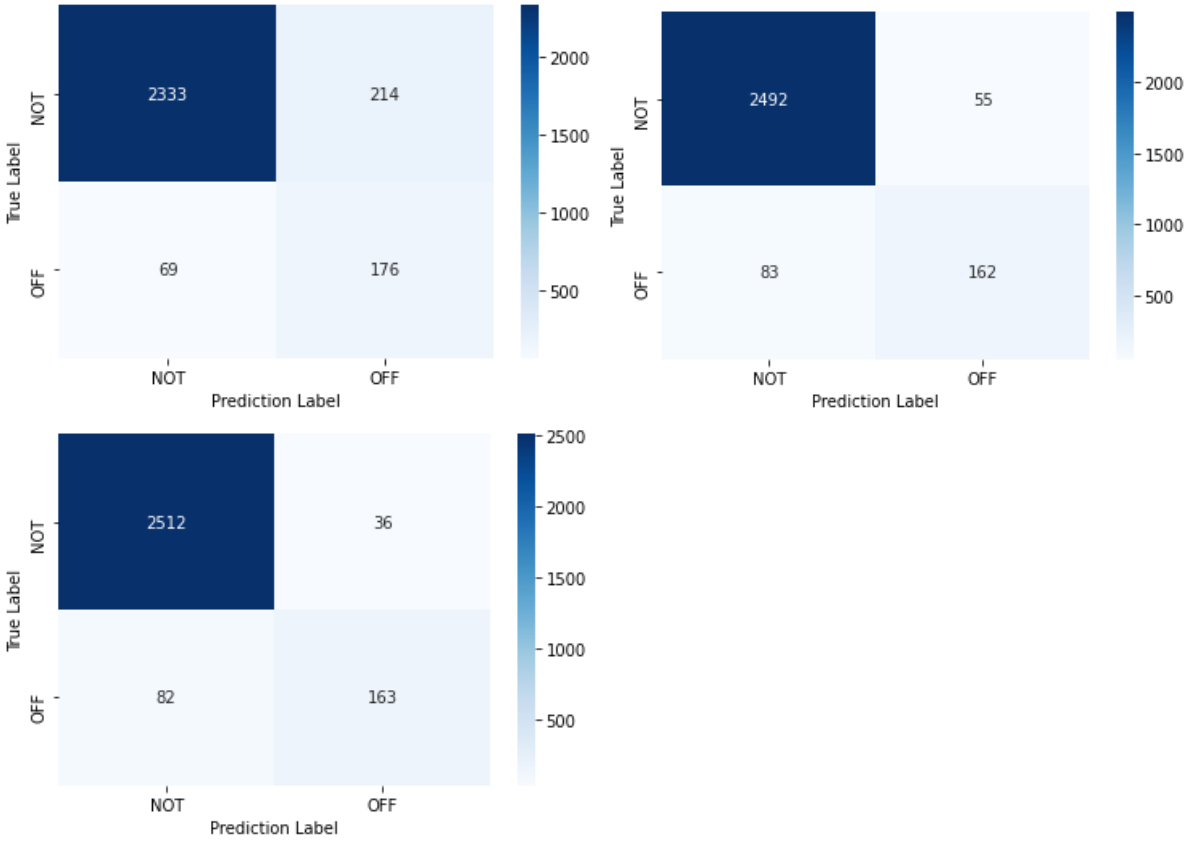
Model	**Non-offensive (NOT)**			**Offensive (OFF)**			**Macro Average**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
OLID	0.971	0.915	0.942	0.451	0.718	0.554	0.711	0.817	0.748
OLID + Reddit 1	0.967	0.978	0.973	0.746	0.661	0.701	0.857	0.819	0.837
OLID + Reddit 2	0.968	0.985	0.977	0.819	0.665	0.734	0.893	0.825	0.855

Table 8: Metrics as per the classification report for level A classifier model applied to Reddit test data

The pure OLID model had the highest recall for offensive posts but was not truly reflective of a better model. The model was classifying a much larger quantity of posts as OFF but had majority of them incorrect at around 55% as per the poor precision. The model also appeared to have the better precision for non-offensive posts yet was due to large misclassification of NOT posts as OFF.

Overall, the inclusion of labelled Reddit data showed good improvement with most metrics improving. The first Reddit data inclusion of batch 1 showed a strong shift in metrics with a large improvement to the F1 score for offensive posts. A second batch of Reddit data further improved the model slightly over its initial inclusion. The final model shows a reliable F1 macro score of 0.855. This does hide the slightly lacking recall of the offensive posts with 34% of offensive posts missed by the model. This could have strong ramifications when the further level B classification is applied. Nonetheless, the model's precision of offensive posts is good with nearly 82% of offensive flagged posts being truly offensive. These results strongly support that the inclusion of training data from the platform being analysed is effective in improving BERT model predictions. Figure 10 expresses the results of the NOT/OFF classifier models.

Figure 10: Confusion matrices for the labelling of level A posts on Reddit test data. OLID (top left), OLID + Reddit 1 (top right), OLID + Reddit 2 (Bottom left)



5.4.2 Level B Model

The second model was trained on the Level B classifiers for targeted (TIN) and non-targeted (UNT) model. It was questionable how effective inclusion of the Reddit data was going to be considering OLID training data had 4,400 offensive posts with Reddit batch 1 having only 135 and batch 2 at 223. Furthermore, expectation of the initial OLID model was poor given that the labelled twitter data is heavily weighted towards targeted offences at 88% of posts whilst the Reddit data clearly shows a majority of untargeted offences. The final UNT/TIN model metrics are expressed in table 9.

Model	Targeted (TIN)			Untargeted (UNT)			Macro Average		
	P	R	F1	P	R	F1	P	R	F1
OLID	0.385	0.907	0.541	0.893	0.349	0.502	0.639	0.628	0.521
OLID + Reddit 1	0.536	0.684	0.601	0.837	0.737	0.782	0.687	0.709	0.691
OLID + Reddit 2	0.577	0.539	0.557	0.798	0.822	0.81	0.688	0.681	0.684

Table 9: Metrics as per the classification report for level B classifier model if given perfectly true OFF labels on Reddit test data.

Within level B the OLID model performed very poorly. The recall appears quite positive within the targeted classifier yet has incredibly poor precision. This is because the model is heavily weighted to marking a post as targeted over untargeted which is due to the weighting in the training data. The OLID model was missing a huge number of truly untargeted posts with around 65% misclassified.

The inclusion of Reddit data showed strong improvement over the pure OLID model. The untargeted F1 score showed a substantial improvement of 0.28 owed to the much better recall value which was over double its predecessors. The initial Reddit data inclusion actually was the better model owed to the stronger targeted offense classification. The second batch of Reddit data balanced the metrics for the untargeted offences with reliable scores around 0.8-0.82 for all three measurements. The inclusion also improved the precision of the targeted offences but worsened the recall missing 46% of such offences. This brought some concerns given the definition of toxicity within this study is based on the targeted posts.

Further concerns were noted whereby the metrics gained from the classification report in the script were not a true reflection of the combination of both the models for the definition of toxicity decided upon in the thesis. This is because the model metrics in the classification report are calculated from being fed only the true OFF classifiers from model A. This is not a true reflection of the effectiveness given that a completely unlabelled data would first need to be categorized at level A. Therefore, the metrics were re-calculated such that the model receives all OFF classifiers from the level A model regardless if they were correct. This is shown in table 10.

Model	Targeted (TIN)			Untargeted (UNT)			Macro Average		
	P	R	F1	P	R	F1	P	R	F1
OLID	0.158	0.618	0.252	0.425	0.236	0.304	0.292	0.427	0.278
OLID + Reddit 1	0.337	0.394	0.363	0.687	0.52	0.592	0.512	0.457	0.478
OLID + Reddit 2	0.433	0.342	0.382	0.719	0.591	0.649	0.576	0.466	0.515

Table 10: Metrics calculated after receiving labels from previous level A classifier model even if incorrect for Reddit test data.

With the model now being given all instances of offensive posts from the initial level A model and now missing the misclassified OFF the metrics have been drastically affected. The recall is reduced heavily since the model is not classifying UNT/TIN which the level A model classed as NOT and the precision is lowered since the model is now labelling UNT/TIN which are truthfully NOT. Table 11 shows the incorrect labels being received and missed by the model thanks to NOT misclassification.

Model	Number of incorrectly classified OFF now being received	Number of incorrectly classified NOT now being missed
OLID	214	69
OLID + Reddit 1	55	83
OLID + Reddit 2	36	82

Table 11: Table showing the number of incoming and missed comments from model A into model B.

The OLID only model is still showing its extreme weighting towards the targeted offences with the recall being the highest of the three models. This does nothing to help the model's poor performance elsewhere. The second Reddit model is now the strongest of the three with the best metrics for all besides the targeted recall. However, the model itself is still poor overall. Part of the reasoning behind this is the very small instances of targeted and untargeted offences within the Reddit data at 2.7% and 6% of the test data total respectively. If the initial level A model misclassifies even a small number of posts, then the metrics for the second model is wildly affected.

5.4.3 Combined Models

The combined model metrics were calculated based on the definition of toxicity presented in this thesis. The toxic metrics are exactly the same as the 'True' level B model above since they are the resulting targeted offences from being fed the offensive labelled data from the level A model. The non-toxic offences have been calculated such that if a post was incorrectly classified as NOT when the true classification is OFF+UNT then the verdict is that the model classed correctly and vice versa. This was done since non-toxic posts are a collection of either classification and even if the classifier is incorrect then either post is labelled as 'Non-toxic'. Table 12 expresses the metrics of the final combined toxicity model.

Model	Non-Toxic (NOT/ OFF+UNT)			Toxic (OFF + TIN)			Macro Average		
	P	R	F1	P	R	F1	P	R	F1
OLID	0.988	0.908	0.946	0.158	0.618	0.252	0.573	0.763	0.599
OLID + Reddit 1	0.983	0.978	0.98	0.337	0.394	0.363	0.66	0.686	0.672
OLID + Reddit 2	0.981	0.987	0.984	0.433	0.342	0.382	0.707	0.664	0.683

Table 12: Metrics calculated from combined level A and B models on Reddit test data as per the definition of toxicity

The OLID model performs relatively poorly as expressed by the macro average F1 score of 0.599. The model is skewed in that it predicts non-toxic defined posts very well yet struggles to detect toxic posts correctly. The recall of toxic post detection is relatively high with over 61% being labelled correctly yet this is countered by an incredibly poor precision score of 0.158. This shows that the model is heavily weighted towards classifying an offensive post as targeted (TIN) regardless of content. The poor precision is largely impacted giving the second model all OFF marked posts from the previous model and thus allowing the categorization of incorrect OFF posts into TIN/UNT. These are collectively included as an immediate false negative greatly reducing the precision.

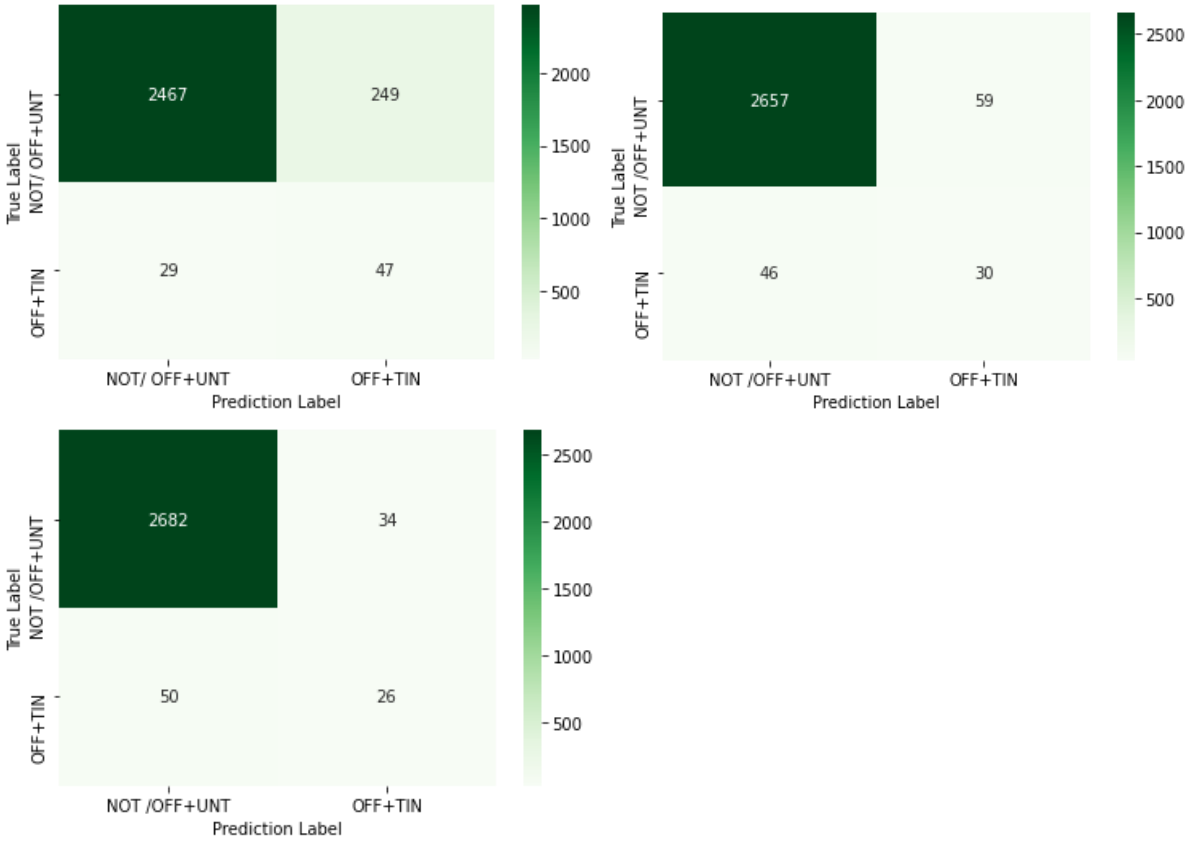
The inclusion of the smaller Reddit data 1 shows some good improvement to the F1 macro score. This is due to an improvement in the recall of 0.07 for the non-toxic posts as well as more

balanced metrics for the non-toxic. The recall of the model is quite disappointing with around 60% of toxic posts classified wrongly. There is a notable improvement in the precision of more than double the OLID model yet still performs poorly with around a third of posts the model thinks is toxic as actually being truly toxic.

The bigger batch of Reddit data shows some minor improvement over the initial Reddit data inclusion. Unfortunately, the recall of toxic posts is further reduced with 66% of truly toxic posts being missed by the model. The precision shows a welcome improvement of 0.1 but is still poor overall. A slight increase in all the non-toxic posts and increase in precision of the toxic post detection helps improve the final F1 macro score over its predecessor. This minor increase didn't support that any further Reddit data annotation would have substantially improved the model.

5.4.4 False Positives

Figure 11: Confusion matrices for the labelling of toxic posts on Reddit test data. OLID (top left), OLID + Reddit 1 (top right), OLID + Reddit 2 (Bottom left)



The false positives showed improvement with the introduction of Reddit data as shown via the confusion matrices in figure 11. Initially the model was labelling a large portion of the test data as toxic at 9.4% but which was substantially reduced to 2.1% which is nearer the true proportion of toxic posts in the training and test data. Whilst this improvement was welcome all three models had a larger proportion of incorrectly labelled toxic posts oppose to correct ones. See table 13 for some examples.

Comment	True Label	OLID Label	OLID + Reddit 1 Label	OLID + Reddit 2 Label
Thats sick	Non-toxic	Toxic	Non-toxic	Non-toxic
SHOOT THE LIMBS!	Non-toxic	Toxic	Non-toxic	Non-toxic
Your brain is on another level. I salute you!	Non-toxic	Toxic	Non-toxic	Non-toxic
And if you manage to win as Imposter, kicked or banned from the lobby.	Non-toxic	Toxic	Toxic	Non-toxic
rip my guy	Non-toxic	Toxic	Toxic	Non-toxic
I can't be the only one that struggled with Me Devil at the start	Non-toxic	Toxic	Toxic	Non-toxic
Fuck ya great idea. Definitely some interesting thinkers in that period like Mr Nihilism himself lmao.	Non-toxic	Toxic	Non-toxic	Toxic
turn on cheats and blow the whole thing up.	Non-toxic	Toxic	Toxic	Toxic
I'm a random guy!!....but I'm useless, sorry.	Non-toxic	Toxic	Toxic	Toxic

Table 13: Table summarising a collection of some false positives between all models. Orange boxes reflect an incorrect classification.

The OLID only model marked 298 comments as toxic which was well over the 76 which were truly toxic. The model had 249 incorrectly labelled as a false positive OFF+TIN, with 35% of those truly OFF+UNT and 65% as truly NOT. This is likely owed to the much higher proportion of TIN comments at 29.2% oppose to UNT at 3.9% in the OLID dataset with the remainder NOT. The model is clearly struggling with the terms of online slang with the mentioning of ‘sick’ being labelled toxic while it’s actual use is positive. Gameplay aspects are also largely labelled ‘Toxic’ such as ‘shooting’, ‘killing’ or ‘death’. Strangely the model labels quite a large number of positive posts as toxic and doesn’t follow a very clear or obvious pattern. The model at the moment cannot distinguish between unusual online or gaming slang which the addition of Reddit data can somewhat alleviate.

The Reddit 1 model showed a notable improvement to the false negatives reducing from 249 to 59 false positives. Breaking down these showed 61% as truly OFF+UNT and the remainder as truly NOT. This shows the model is now failing more at the level B classifier. Many instances of violent gameplay descriptions were now being flagged correctly as non-toxic. The model did still have trouble with some gaming slang especially homographs such as the word, ‘kicked’, which means to be disconnected from a game. The instance of ‘rip’ was also flagging as toxic when it’s use was to describe an in-game death. Some remaining instances weren’t obvious as to why they were flagged as toxic by the model.

The Reddit 2 model improved on the false positives over its predecessors with only 34 cases, 62% of the such were truly OFF+UNT while the rest were NOT. The model does well to flag a post as OFF correctly but then struggles with the level B classification similarly to the Reddit 1 model. In total there were 7 posts which were classified correctly by the Reddit 1 model and then reclassified incorrectly by the Reddit 2 batch, 5 of which at the level B level.

All three models struggled with labelling false positives involving strong worded self-hate which as previously mentioned in the thesis is marked as non-targeted. Some gameplay actions were still being flagged as toxic such as being ‘blown up’ or gameplay descriptions with residing swearwords and profanity. This could be alleviated by more Reddit data incorporation but was out of scope for completing the final thesis predictions.

5.4.5 False Negatives

As the Reddit data was introduced into the models the number of false negatives increased resulting in lower recall as more batches were added. It is unusual as to why this is happening, this could be due to the inconsistency between the OLID annotations and the Reddit data. If scope allowed then running a large batch of Reddit data without OLID annotations would have been useful to test. It would be interesting to see at what level of Reddit data inclusion that the recall starts improving if at all. Table 14 outlines some occurrences of false negatives between the models.

Comment	**True Label**	**OLID Label**	**OLID + Reddit 1 Label**	**OLID + Reddit 2 Label**
Psycho	Toxic	Non-toxic	Non-toxic	Non-toxic
take your meds.	Toxic	Non-toxic	Non-toxic	Non-toxic
Hide and seek settings dumbass	Toxic	Non-toxic	Non-toxic	Non-toxic
Fuck off then and play something else, the Animus is a massive part of Assassin's Creed's identity. It needs to stay.	Toxic	Toxic	Non-toxic	Non-toxic
its a game u weirdo	Toxic	Toxic	Non-toxic	Non-toxic
Motherfucker	Toxic	Toxic	Non-toxic	Non-toxic
You may not be a pickle but you need to shove one down your throat and stfu	Toxic	Toxic	Toxic	Non-toxic
Just dont be dumb	Toxic	Toxic	Toxic	Non-toxic
I just want to say this subreddit sucks not the people though that's actually why I'm mad right now	Toxic	Toxic	Toxic	Non-toxic

Table 14: Table summarising a collection of some false negatives between all models. Orange boxes reflect an incorrect classification.

The OLID model flagged 29 false negatives with 26 of which mislabelled as NOT and 3 as OFF+UNT. The low OFF+UNT false negatives are due to the heavy weighting of the model, classifying an OFF post being much more likely to be TIN than UNT. Which again is reflective of the balance of the OLID posts. Most instances of incorrect NOTs were also classified similarly by the following two models. The Reddit batch 1 model altered only 2 incorrect NOT flags with the batch 2 only altering 1 instance. These incorrect NOTs included some hard to detect implicit abuse yet some comments containing obvious profanity were also flagged as NOT as seen in table 14.

The Reddit 1 model flagged 46 false negatives with 38 labelled NOT and 8 labelled OFF+UNT. The Reddit 2 model slightly increased its false negative count classifying 50 toxic posts incorrectly. 34 of these were labelled as NOT whilst 16 were labelled as OFF+UNT. The reduction of

incorrect NOTs shows a slight improvement over the Reddit 1 model yet the incorrect level B classifiers counteracted this improvement. Both models show more profanity being falsely labelled as non-toxic which is likely due to differences between the OLID and Reddit annotations.

One notable comment was 'it's a game u weirdo' in which both Reddit models classified incorrectly as NOT. However, another similar instance of 'You're a weirdo' was classed correctly as toxic by all models. These unusual instances hinder progress in developing an accurate model.

5.4.6 HateBERT

The HateBERT model for level A showed a minor improvement over the BERT model as seen in table 15. The classification of offensive posts showed good improvement with much more balanced metrics around 0.62-0.68 for precision, recall and F1 score. However, the non-offensive (NOT) posts declined in metrics over its predecessor resulting in only a 0.013 increase in the F1 macro average. This is lower than the HateBERT paper which showed an increase of 0.06 in the F1 macro average for level A [20].

The level B HateBERT model fared a lot worse than the BERT model. Nearly all metrics were worse for the HateBERT with only the recall of targeted (TIN) offences fractionally better at 0.001. This is likely due to the use of Reddit test data which has a lack of targeted (TIN) occurrences whilst the OLID data is heavily weighted to untargeted (UNT) classification.

Model	**Non-offensive (NOT)**			**Offensive (OFF)**			**Macro Average**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERT OLID	0.971	0.915	0.942	0.451	0.718	0.554	0.711	0.817	0.748
HateBERT OLID	0.859	0.885	0.872	0.679	0.625	0.651	0.769	0.755	0.761
Model	**Targeted (TIN)**			**Untargeted (UNT)**			**Macro Average**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERT OLID	0.385	0.907	0.541	0.893	0.349	0.502	0.639	0.628	0.521
HateBERT OLID	0.354	0.908	0.509	0.860	0.254	0.393	0.607	0.581	0.451

Table 15: Metrics of BERT against HateBERT for classifying Reddit test data for levels A and B.

The poor performance is surprising and inspection of the RAL-E dataset shows that a lot of the banned Reddit comments aren't particularly offensive as per table 16. This could also be why a lack of improvement is seen. Although, the aim of the model was to improve its use on forum detection by having more exposure to online language which is still achieved. The token count of RAL-E is 43 million, this is relatively small compared to BERTs original token count at 3300 million, which amounts to only 1.28% of the pre-training data. A much larger portion of forum data is needed for a substantial effect.

Non-offensive Banned RAL-E Comments
Proof please
If I ever make a movie, this shall be a scene
Yes! Thank you.
Try refreshing. Looks deleted to me.
You made a wise decision

Either one works.

Table 16: Examples of non-offensive banned posts from RAL-E dataset.

The findings have highlighted that fine-tuning BERT oppose to extra pre-training steps using data from the analysed platform is clearly more effective. It would have been interesting to investigate both additions by applying the Reddit training data on the HateBERT model however the scope of the project could no longer accommodate the extra fine-tuning. Given the poor performance witnessed at level B the pursuit of implementing a full HateBERT toxicity model was halted in favour of the already improved BERT models.

5.5 Evaluation of OLID Competition and Annotations

Figure 12: Figure showing OffensEval 2019 results for subtask B.

Sub-task B	
Team Ranks	**F1 Range**
1	0.755
2	0.739
3	0.719
4	0.716
5	0.708
6	0.706
7	0.700
8	0.695
9	0.692
CNN	**0.690**
10	0.687
11-14	.680-.682
15-24	.660-.671
BiLSTM	**0.660**
25-29	.640-.655
SVM	**0.640**

Overall, the final model had poor performance. The F1 macro score of 0.683 would have placed the model at 11th place within the sub-task B of the 2019 OffensEval competition as per figure 12. However, the Reddit test data is different to the competitions and thus isn't truly comparable.

The metrics of the competition are questionable since the models for sub-task B are passed the already correctly labelled OFF posts. This would not reflect a real-life scenario as a completely un-annotated dataset would firstly need perfectly categorizing into level A NOT/OFF classifiers first. Due to this, a pure OLID training and test data run was performed to test the 'True' values at classifier level B. The hyperparameters were as per the previous BERT models. The model was passed the offensive marked posts from the level A model and the results are shown in table 17. The results show a large decrease in performance by the model having a 0.2 reduction in targeted offence detection and a more minor 0.05 reduction for untargeted. The decrease is substantial and reflects how a model would classify from a completely unlabelled set of data.

Another notable issue with the OLID competition is that it relies solely on the F1 macro metric which can easily hide how poor a model could be. The model for this thesis is a good example of this. If only the F1 macro score was visible then it could be assumed an adequate model but investigating the underlying metrics shows that toxic posts are classified poorly. Other presented models in the competition could be hiding this problem.

Calculation	**Targeted (TIN)**			**Untargeted (UNT)**			**Macro Average**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Given perfectly labelled OFF	0.92	0.93	0.93	0.42	0.37	0.39	0.67	0.65	0.66
Given all OFF from previous model	0.62	0.605	0.737	0.322	0.37	0.344	0.471	0.488	0.541

Table 17: Table comparing an original model for subtask B against a 'True' classification for OLID.

Whilst trying to replicate the OLID annotations for the Reddit data it became apparent that there are also contradictions within the OLID dataset. For OFF classification the definition states that a comment contains ‘profane language’ or ‘swear words’ yet there were multiple instances of this left as non-offensive. There were also instances of OFF+TIN which were clearly non-targeted offences and more so not even offensive. This was the same for some OFF+UNT classifications, which even had some obvious targeted offences too. Some examples are outlined in table 18. This was surprising given that the annotation of such labels had a thorough approach with 2 experienced labellers used for the dataset. Both labellers had to label all posts and the majority agreed label was the resulting label. If the label conflicted then a 3rd labeller would give a deciding label. It would be worth revisiting the labels to alleviate the inconsistencies.

Comment	Incorrect Label
@USER IT'S FUCKING TIME!!	NOT
@USER Best part is if you read the exchange he had in that thread, a year later this dude comes back on his same blm and Antifa caused nazis" bullshit lmao. And by "best" i mean saddest."	NOT
@USER Seriously who's the dick that unfollowed	NOT
@USER Trump kicks dem butt - its so fun.	OFF+TIN
@USER I mean it worked for gun control right? URL	OFF+TIN
@USER Oh my Carmen. He is SO FRICKING CUTE	OFF+TIN
@USER Ouch!	OFF+UNT
@USER FUCK OFF	OFF+UNT
@USER Absolutely NFW.	OFF+UNT

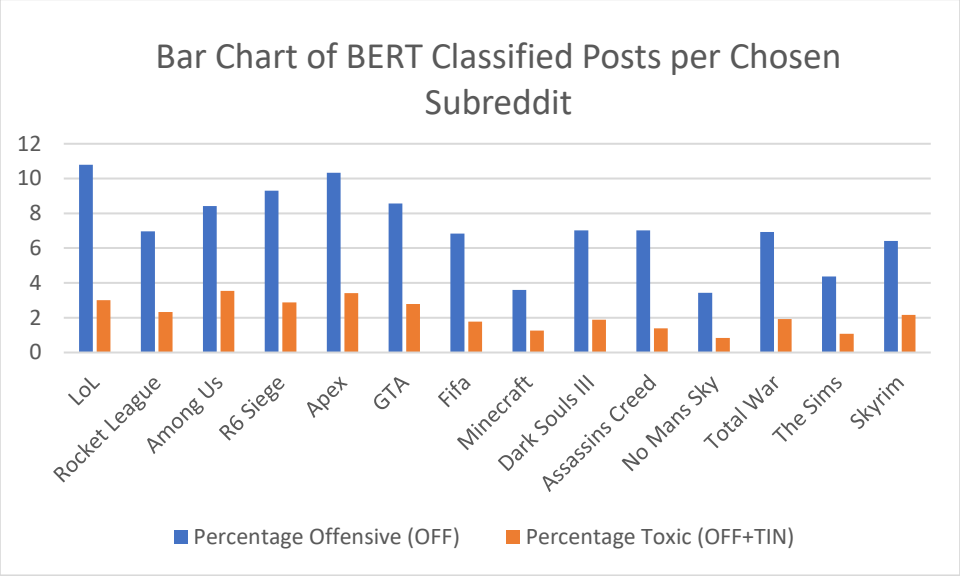
Table 18: Table showing some incorrect labels within the original OLID Twitter annotations.

From the unexpected issues within the OLID labels it became obvious that the best trained model for this thesis would likely be solely trained on text from the subreddit gaming communities. Labels from other platforms or communities introduce contradictions into in the model and thus lower accuracy. If time allowed then a much larger annotated Reddit dataset would have been optimal to train the BERT model.

5.6 Toxicity Results

Whilst the final BERT toxicity model was poor it was decided to run the finalised model on the collected Reddit data and investigate the toxicity within the communities. With the level A OLID classifier label somewhat reliable some hypothesis could still be performed with good confidence of findings for offensive occurrences. However, the final combined level A and B models were poor and hence no conclusions regarding toxic classified posts can be very reliable. The section is written such that the model was effective in its toxic classification.

Figure 13: Bar Chart of BERT Classified Posts per Chosen Subreddit



The final predictions showed a varying rate of between 11-3% of posts being flagged as offensive (OFF) as per figure 13. The most offensive subreddit page was LoL at 10.8%, with Apex and Rainbow 6 Siege following closely. Minecraft and No Mans Sky were the least offensive and the only two games to have under 4% of posts being labelled offensive.

Within such offensive posts 16-30% were marked as targeted (TIN) and thus labelled toxic as per this thesis definition. The toxic posts varied between 0.8-3.6% of the total collected subreddit data. The game with the most toxicity was Among Us at 3.55%, this was surprising given it was the most moderated subreddit page and also was only the 6th highest in the offensive comment classification. Apex followed closely second at 3.41% with LoL third with 3.01% of posts marked toxic. The least toxic pages were also the least offensive being No Mans Sky, The Sims and Minecraft. No Mans Sky was the only subreddit to have under 1% of posts marked as toxic.

Figure 14: Box and Whisker of BERT Classified Toxicity per Game Type

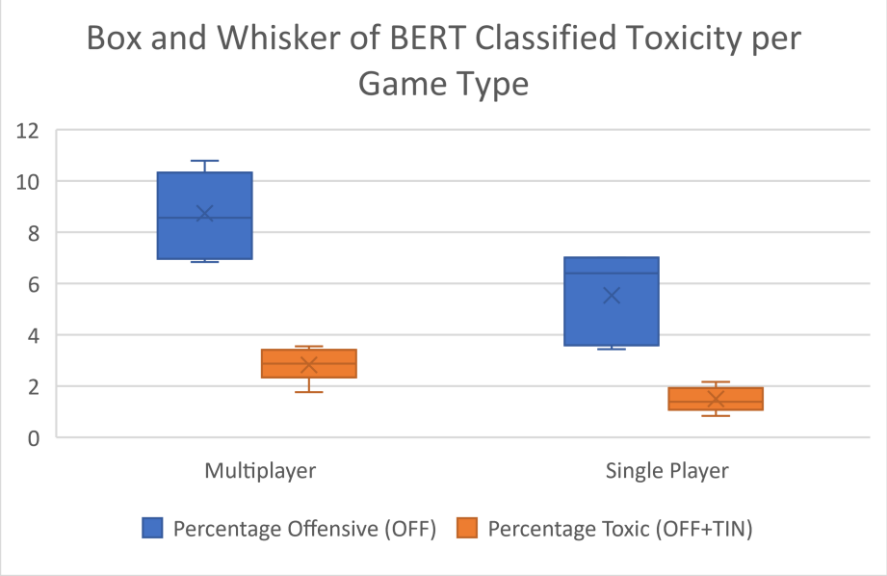
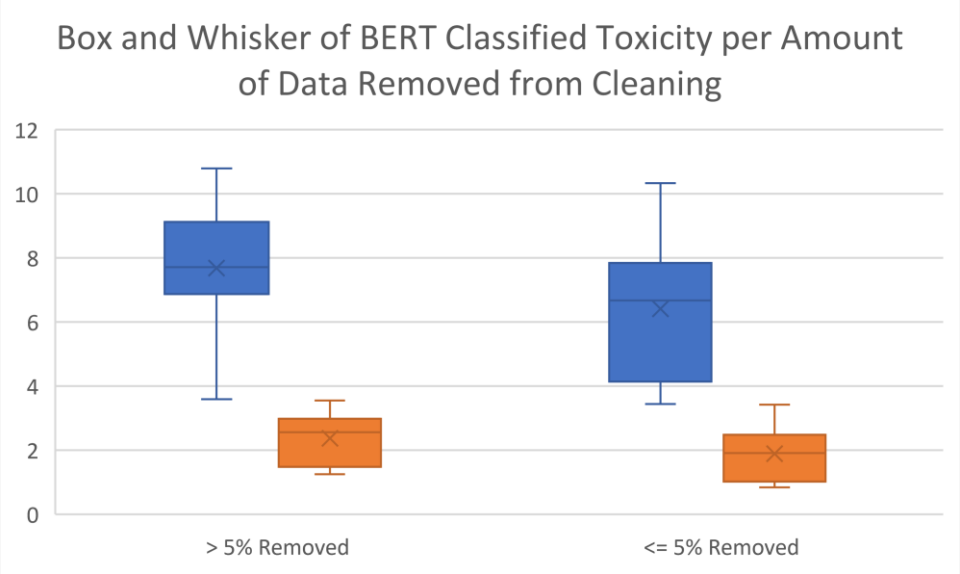


Figure 14 separates the games by game type and shows that offensive posts and toxicity are more common within multiplayer subreddit communities. Multiplayer subreddits had an average of 8.7% of offensive occurrences whilst single player had 5.5%. Multiplayer games also had 2.8% toxic occurrences, nearly double the amount of single player toxic classification posts at 1.5%. To further test the significant difference the H_1 hypothesis test was performed using a one-tailed t-test of α=0.05:

H_1: Multiplayer subreddits have a higher occurrence of toxic posts.

This resulted in a p-value of 8e-4 showing strong rejection of the null hypothesis in favour of H_1. Multiplayer subreddits clearly harbour more toxicity which is likely due to higher frustrations with competitive online play.

Figure 15: Box and Whisker of BERT Classified Toxicity per Amount of Data Removed from Cleaning



The predictions were cross compared with the data cleaning performed within this thesis on each subreddit as expressed in figure 15. Offensive and toxic posts were slightly higher in the data which had more than 5% of its data removed and thus more moderated. This is somewhat surprising as the expectation would be that higher subreddit moderation would result in lower instances of toxicity compared to pages without it. This could be because less moderated don't require much moderation with less toxic communities. A two-tailed t-test was performed to check on the significance between the groups at α=0.05. The hypothesis in question is:

H_2: There is a difference of toxicity between more moderated pages (>5% excluded) than less moderated pages (<=5% excluded).

This resulted in a p-value of 0.31 which is not low enough to reject the null hypothesis. Therefore, there is no statistical significance of differing toxicity between more moderated subreddit pages. This shows the moderation is working well to reduce the toxicity to around similar levels of less moderated pages. This could mean that higher moderated pages originally have more toxicity which is then removed through moderation, however this is arguable given the missing detail of removed posts.

Figure 16: Box and Whisker of BERT Classified Toxicity per ESRB Age Rating

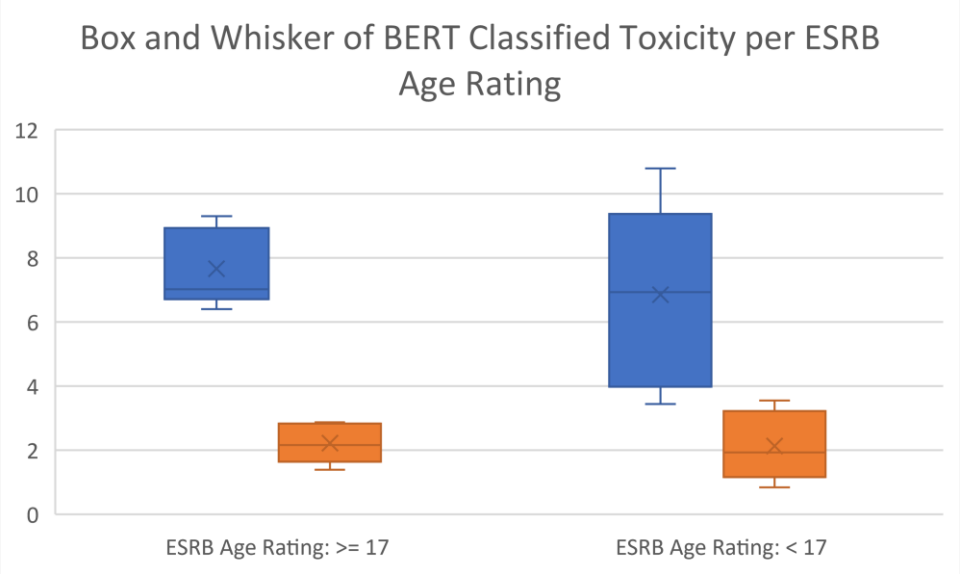


Figure 16 shows that comparing toxicity against the ESRB age rating the occurrences of offensive content was slightly more common in subreddits with a higher rating by 0.81%. The spread of the offensive content was more compact in the mature group between 9.3-6.4% compared to a much wider range of 10.7-3.4% for the younger age rated group. Toxicity within the subreddits was hardly any different at 0.07%. Another hypothesis test was performed to check the significance, as below:

H_3: Mature rated games will have higher toxicity than those with a lower age rating.

A one-tailed t-test of α=0.05 resulted in a p-value of 0.43 which is not low enough to reject the null hypothesis. Therefore, there is no statistical significance that mature rated games have more toxicity. This is not a positive finding and shows that gamers who play even the youngest of age rated games express similar rates of toxicity. This is not ideal given that children may be within the online community and thus experiencing toxicity from a young age.

6 Conclusion

6.1 Contributions

This thesis has presented a thorough literature review investigating the multiple definitions of toxicity. It also has presented a range of NLP techniques and their application to classifying toxicity occurrences. This has led to a clearly presented definition of toxicity which is expressed using OLID classifiers.

The investigation into the data selection has shown that multiplayer game pages on Reddit are far more active than single player. Furthermore, the amount of subreddit followers does not reflect the true activity of the pages. The data cleaning has shown a huge variance in the moderation setup of the subreddit pages. There is no clear difference of moderation between the pages and are dependent on how well the moderators self-manage the pages. The data cleaning has also shown that the language detection package 'Langdetect' is ineffective at classifying languages within Reddit posts due to the use of very short messages, online slang and misspellings within posts.

Two BERT NLP models have been produced for level A and B OLID classifiers and the most effective hyperparameter settings have been tested. This has shown that epochs less than four are ineffective at log loss convergence for OLID classification. Learning rate and batch sizes have slight increases in metrics and are worth testing for improvement. The models have been combined into a final toxicity model as per the outlined definition. Both have shown a substantial improvement in using the self-labelled Reddit training data alongside the original OLID Twitter data for toxicity detection on the Reddit platform. This reflects that training a classification model should include labelled data from the platform being predicted. Further improvement involved a potential change to multiple HateBERT models over BERT. However, HateBERT's extra pre-training on Reddit data showed little improvement at level A and even declined at level B. This shows that fine-tuning a model on platform data is far more effective than adding to the pre-training with the platform data for BERT.

This thesis performed a thorough review of OLID classifiers, annotations and metrics. These have outlined that the F1 macro metrics used to rank the OffensEval competitions are not always reflective of an effective model with underlying classifiers potentially being poor. It has also shown that the hierarchal classifier metrics, level B and C, aren't a true measurement of effectiveness if given an unlabelled set of data. A 'True' measurement for level B has been calculated and shows far less effectiveness if taking into account the preceding classifier results. Further investigation into the labelling of the OLID classification in the test and training data has brought to light inconsistencies within the data, contradicting their own definitions in places. Such issues within the annotations can only reduce the metrics of trained machine learning models.

Comparing the final predictions of the toxicity model showed some insightful findings. Firstly, a statistically significant difference in multiplayer toxicity was witnessed over single player games. The competitive element in multiplayer games clearly raises the frustration of players compared to single player-based games. Higher moderated pages showed no difference in toxicity or offensive language use. Finally, the ESRB age rating of games also showed no statistically significant difference in levels of toxicity. This is a worrying finding and shows that levels of toxicity are just as high for games which include young children within their communities.

6.2 Limitations & Future Work

The final model predictions were relatively ineffective at detecting toxicity correctly. This is partly due to a lack of toxic occurrences on the Reddit platform with only 2.7% percent within the labelled test data. If a large portion of posts were misclassified at the initial level A classifier, then the metrics would already be largely affected due to such low occurrences. A two level subclassifier model doesn't seem to have practical application for such minimal occurrences. It is also ineffective for use on unlabelled raw data as shown by the OLID 'True' metric calculations for level B. This is especially true if the training data differs to the platform being analysed.

Improvements of the models were investigated but hindered by the scope of the project. Further Reddit annotation could have been helpful and it would have been interesting to see when the target classifiers recall started improving over its predecessor. Furthermore, a comparison of pure Reddit training data against the combination of OLID and Reddit data could have shown whether there is use having data from other platforms at all. Further improvement involved applying HateBERT which proved to be initially ineffective. However, the models were never investigated with training Reddit data as well as the original OLID data. This could have been an interesting direction to pursue if time allowed.

Future work could compare the two BERT models of NOT/OFF and UNT/TIN against a single BERT model which had a 3-way classification of NOT/OFF+UNT/OFF+TIN. Another option would be to compare a single BERT model based on self-created labels of TOX which captures OFF+TIN and NTOX which captures OFF+UNT and NOT. However, both would require conversion of the original OLID classifiers and one would assume the metrics to be similar.

The cleaning of the Reddit data was largely successful yet the language detection of the posts was not. This meant all non-English posts had to be left within the data which is not ideal for training a machine learning model. Improvement for such a tool could be beneficial in the NLP field. A tool which focuses on the context as well as the content may prove more effective. Also, incorporation of an online slang tool could classify slang as a separate classifier and separate it from any incorrect language detection.

The toxicity findings from the predictions of the model showed that multiplayer subreddit pages exhibited more instances of toxicity and offensiveness over single player. This could be investigated further by tracking gamers' emotions whilst playing the different types. Investigating the reasoning into the difference of toxicity could prove beneficial and highlight how games can reduce levels of toxicity to single player levels.

Another finding in the thesis showed that there isn't a difference in toxicity between the ESRB age rating of games. This worrying finding could mean that children are experiencing toxicity from a very young age leading to early acceptance that gaming toxicity is expected and accepted. There is limited research surrounding occurrences of gaming toxicity experienced by youngsters and its impact. Any further investigation would be welcome and help to highlight the need to bring an end to toxicity within the gaming community.

Bibliography

1. McCormick, C. *BERT Fine-Tuning Sentence Classification v4*. 2020; Available from: https://colab.research.google.com/drive/1pTuQhug6Dhl9XalKB0zUGf4FIdYFlpcX#scrollTo=jrC9__lXxTJz.
2. Blackburn, J. and H. Kwak, *STFU NOOB! predicting crowdsourced decisions on toxic behavior in online games*, in *Proceedings of the 23rd international conference on World wide web*. 2014, Association for Computing Machinery: Seoul, Korea. p. 877–888.
3. Mohan, S., et al. *The Impact of Toxic Language on the Health of Reddit Communities*. 2017. Cham: Springer International Publishing.
4. Shores, K.B., et al. *The identification of deviance and its impact on retention in a multiplayer game*. in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 2014.
5. Mishra, P., H. Yannakoudakis, and E. Shutova, *Tackling Online Abuse: A Survey of Automated Abuse Detection Methods*. 2019.
6. Fortuna, P., J. Soler, and L. Wanner. *Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets*. 2020. Marseille, France: European Language Resources Association.
7. Waseem, Z., et al. *Understanding Abuse: A Typology of Abusive Language Detection Subtasks*. 2017. Vancouver, BC, Canada: Association for Computational Linguistics.
8. Struß, J.M., et al., *Overview of germeval task 2, 2019 shared task on the identification of offensive language.* 2019.
9. Zampieri, M., et al., *Predicting the Type and Target of Offensive Posts in Social Media*. 2019. 1415-1420.
10. Zampieri, M., et al. *SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)*. 2019. Minneapolis, Minnesota, USA: Association for Computational Linguistics.
11. Rosenthal, S., et al., *A Large-Scale Semi-Supervised Dataset for Offensive Language Identification.* ArXiv, 2020. **abs/2004.14454**.
12. Zampieri, M., et al. *SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)*. 2020. Barcelona (online): International Committee for Computational Linguistics.
13. Caselli, T., et al. *I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language*. 2020. Marseille, France: European Language Resources Association.
14. Pavlopoulos, J., et al. *ConvAI at SemEval-2019 Task 6: Offensive Language Identification and Categorization with Perspective and BERT*. 2019. Minneapolis, Minnesota, USA: Association for Computational Linguistics.
15. Noever, D., *Machine learning suites for online toxicity detection.* arXiv preprint arXiv:1810.01869, 2018.
16. Lees, A., J. Sorensen, and I. Kivlichan. *Jigsaw@ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model*. in *EVALITA*. 2020.
17. van Aken, B., et al. *Challenges for Toxic Comment Classification: An In-Depth Error Analysis*. 2018. Brussels, Belgium: Association for Computational Linguistics.
18. Hosseini, H., et al., *Deceiving Google's Perspective API Built for Detecting Toxic Comments.* ArXiv, 2017. **abs/1702.08138**.
19. Devlin, J., et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. Minneapolis, Minnesota: Association for Computational Linguistics.
20. Caselli, T., et al., *HateBERT: Retraining BERT for Abusive Language Detection in English.* ArXiv, 2020. **abs/2010.12472**.

21. Beres, N.A., et al., *Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming*, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, Association for Computing Machinery: Yokohama, Japan. p. Article 438.
22. Türkay, S., et al. *See no evil, hear no evil, speak no evil: How collegiate players define, experience and cope with toxicity*. in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020.
23. Kou, Y., *Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends*. 2020.
24. Ghosh, A., *Analyzing Toxicity in Online Gaming Communities.* Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021. **12**(10): p. 4448-4455.
25. Shen, C., et al., *Viral vitriol: Predictors and contagion of online toxicity in World of Tanks.* Computers in Human Behavior, 2020. **108**: p. 106343.
26. Statista. *Value of the global video games market from 2012 to 2021*. 2021; Available from: https://www.statista.com/statistics/246888/value-of-the-global-video-game-market/.
27. GlobalData. *Video Games – Thematic Research*. 2019; Available from: https://store.globaldata.com/report/gdtmt-tr-s212--video-games-thematic-research/.
28. FinancesOnline. *Number of Gamers Worldwide 2021/2022*. 2021; Available from: https://www.statista.com/statistics/246888/value-of-the-global-video-game-market/.
29. Barr, M. and A. Copeland-Stewart, *Playing Video Games During the COVID-19 Pandemic and Effects on Players' Well-Being.* Games and Culture, 2021. **0**(0): p. 15554120211017036.
30. Statista. *Increase in sales in the video game industry during the coronavirus (COVID-19) pandemic*. 2020; Available from: https://www.statista.com/statistics/1109979/video-game-console-sales-covid/.
31. VentureBeat. *WHO and game companies launch #PlayApartTogether*. 2020; Available from: https://venturebeat.com/2020/03/28/who-and-game-companies-launch-playaparttogether-to-promote-physical-distancing/.
32. Statista. *Increase in time spent video gaming during the COVID-19 pandemic worldwide*. 2020; Available from: https://www.statista.com/statistics/1188558/gaming-genres-spent-covid/.
33. Statista. *Impact of COVID-19 on the frequency of playing multiplayer video games*. 2020; Available from: https://www.statista.com/statistics/1188549/covid-gaming-multiplayer/.
34. Yang, Y.-T.C., *Building virtual cities, inspiring intelligent citizens: Digital games for developing students' problem solving and learning motivation.* Computers & Education, 2012. **59**(2): p. 365-377.
35. Zirawaga, S., A. Olusanya, and T. Maduku, *Gaming in Education: Using Games as a Support Tool to Teach History.* 2017.
36. Olsén, J., *Computer gaming's facilitation of the English subject : A quantitative research on the influence of computer gaming on students' English performance*. 2017. p. 42.
37. Kotaku. *Therapists Are Using Dungeons & Dragons To Get Kids To Open Up*. 2017; Available from: https://kotaku.com/therapists-are-using-dungeons-dragons-to-get-kids-to-1794806159.
38. Beaumont, R., et al., *Randomized Controlled Trial of a Video Gaming-Based Social Skills Program for Children on the Autism Spectrum.* Journal of Autism and Developmental Disorders, 2021.
39. Claudine J.C. Lamoth, S.R.C., Klaas Postema, *Active Video Gaming to Improve Balance in the Elderly.* Annual Review of Cybertherapy and Telemedicine 2011, 2011. **Volume 167**: p. 159-164.
40. Theng, Y.-L., P.H. Chua, and T.P. Pham, *Wii as entertainment and socialisation aids for mental and social health of the elderly*, in *CHI '12 Extended Abstracts on Human Factors in Computing Systems*. 2012, Association for Computing Machinery. p. 691–702.
41. Qutee. *Gaming and you report*. 2018; Available from: https://www.qutee.com/gaming-and-you-report/.
42. ADL/Newzoo. *Free to Play? Hate, Harassment and Positive Social Experience in Online Games 2020*. 2020; Available from: https://www.adl.org/free-to-play-2020.

43. Lindsey A. Cary, J.A., Alison L. Chasteen, *The interplay of individual differences, norms, and group identification in predicting prejudiced behavior in online video game interactions.* Journal of Applied Social Psychology, 2020. **50**(11): p. 623-637.
44. Adinolf, S. and S. Turkay, *Toxic Behaviors in Esports Games: Player Perceptions and Coping Strategies*, in *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. 2018, Association for Computing Machinery: Melbourne, VIC, Australia. p. 365–372.
45. MSNBC. *No charges for man who shot police chief in Oklahoma*. 2015; Available from: https://www.msnbc.com/msnbc/no-charges-man-who-shot-police-chief-oklahoma-msna508901.
46. PCGamer. *A 60-year-old man died of a heart attack after being swatted over his Twitter handle*. 2020; Available from: https://www.pcgamer.com/uk/a-60-year-old-man-died-of-a-heart-attack-after-being-swatted-over-his-twitter-handle/.
47. DotEsports. *Blizzard will now notify you if the player you reported in Overwatch was punished*. 2017; Available from: https://dotesports.com/overwatch/news/blizzard-ban-system-update-overwatch-16706.
48. gamesindustry.biz. *Amazon patents method of grouping toxic players together online*. 2020; Available from: https://www.gamesindustry.biz/articles/2020-11-03-amazon-patents-method-of-grouping-toxic-players-together-online.
49. Steam. *Dota - Low Priority Matchmaking*. Available from: https://help.steampowered.com/en/faqs/view/0438-BAAC-F9CE-BA22.
50. Jefferey Lin, P.S., *Systems and methods that enable player matching for multi-player online games*. 2012.
51. SpectrumLabs. *How Riot Games Used Science to Curb Toxic Behaviour in League of Legends*. 2012/2013; Available from: https://www.spectrumlabsai.com/the-blog/how-riot-games-is-used-behavior-science-to-curb-league-of-legends-toxicity.
52. GameSkinny, *To Punish Bad Behavior or Reward Good Behavior? That Is the Question.* 2015.
53. Kotaku. *Blizzard Says Overwatch Toxicity Is Down 40 Percent*. 2019; Available from: https://kotaku.com/blizzard-says-overwatch-toxicity-is-down-40-percent-1833506700.
54. Elliot, A.J. and M.A. Maier, *Color and Psychological Functioning.* Current Directions in Psychological Science, 2007. **16**(5): p. 250-254.
55. Reddit. *Reddit by the Numbers*. 2020; Available from: https://www.redditinc.com/press.
56. Statista. *Ranking of the number of Reddit users by country 2020*. 2020; Available from: https://www.statista.com/forecasts/1174696/reddit-user-by-country.
57. Statista. *Regional distribution of desktop traffic to Reddit.com*. 2021; Available from: https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/.
58. MarketingCharts. *US Adults' Social Platform Use, by Demographic Group*. 2021; Available from: https://www.marketingcharts.com/charts/us-adults-social-platform-use-by-demographic-group-in-2021/attachment/pew-social-platform-use-by-demographic-apr2021.
59. AgeUK. *Not like riding a bike: Why some older people stop using the internet*. 2020; Available from: https://www.ageuk.org.uk/globalassets/age-uk/documents/reports-and-publications/lapsed_users_report_march-2020.pdf.
60. Reddit. *Reddit Transparency Report*. 2020; Available from: https://www.redditinc.com/policies/transparency-report-2020.
61. Vaswani, A., et al. *Attention is all you need*. in *Advances in neural information processing systems*. 2017.
62. Ba, J.L., J.R. Kiros, and G.E. Hinton, *Layer normalization.* arXiv preprint arXiv:1607.06450, 2016.
63. Dai, A.M. and Q.V. Le, *Semi-supervised sequence learning.* Advances in neural information processing systems, 2015. **28**: p. 3079-3087.
64. Sun, C., et al. *How to fine-tune bert for text classification?* in *China National Conference on Chinese Computational Linguistics*. 2019. Springer.
65. Google. *Colaboratory*. Available from: https://research.google.com/colaboratory/faq.html.

66. EsportsObserver. *Top 10 Esports Games of 2020 by Total Winnings*. 2021; Available from: https://archive.esportsobserver.com/top10-games-2020-total-winnings/.
67. Baumgartner, J., et al., *The Pushshift Reddit Dataset.* Proceedings of the International AAAI Conference on Web and Social Media, 2020. **14**(1): p. 830-839.
68. Nakatani, S. *Language Detection Library for Java*. 2010; Available from: https://www.slideshare.net/shuyo/language-detection-library-for-java.
69. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

Appendix

All final scripts can be found in the online repository found at https://github.com/HarriG109/Dissertation. The repository contains the following:

- Reddit scraper
- Language detection script
- Final BERT models for OLID classification level A and B
- HateBERT models for OLID classification level A and B
- Prediction scripts for OLID classification level A and B