Article

Identifying and Analyzing Toxic Behavior in Ecommerce Industry through Reddit Discussions Using BERT and HateBERT

ANOOP KUMAR , ARPITA SONI , Rajeev Arora * , Dheerendra Panwar

Posted Date: 26 August 2024

doi: 10.20944/preprints202408.1797.v1

Keywords: BERT; Games; HateBERT; Mean Behavior; Reddit

Article

Identifying and Analyzing Toxic Behavior in Ecommerce Industry through Reddit Discussions Using BERT and HateBERT

Anoop Kumar [1], Arpita Soni [2], Rajeev Arora [3,*] and Dheerendra Panwar [4]

[1] Royston Ct, Dublin, USA; anoop.kumar.2612@gmail.com
[2] White Chapel Dr NW, Concord, USA; soni.arpita@gmail.com
[3] Nicklaus Children s Hospital
[4] IEEE; dheerendra.panwar@ieee.org
* Correspondence: rajeev04.study@gmail.com

Abstract: This study begins by looking at how toxic behaviour in video games is talked about and identified. We set up a clear idea of what "toxic behaviour" means and suggested ways to figure out when it's happening. To test these ideas, we scrape a huge collection of posts from Reddit where people talk about games. We cleaned up the data—making sure it was all good for checking. We used Bidirectional Encoder Representations from Transformers (BERT) to look at toxic behaviour. We also used HateBERT to compare how well it works at spotting toxic stuff. To make sure our models were doing a good job, we looked at over 5500 posts and decided if they were toxic or not. This helped us see if the models were right or wrong. We also used measures—to compare how much toxic stuff is in Reddit—especially for games. This helped us understand how common and different toxic behaviour is in different places.

Keywords: Index Terms—BERT; games; HateBERT; mean behavior; reddit

1. Introduction

Online toxicity is a big problem on the internet nowadays. But it is hard to say exactly what it is about because people see it differently. Some experts say it is like cyberbullying—when someone keeps doing mean things to hurt others online. The Oxford dictionary says toxicity is when something is harmful or not nice—especially when it is used to control or trick others [1]. Some researchers such as [2]—say toxicity includes cyberbullying, threats, harassment, hate speech, and abuse online. This matches what [3] thinks—which is that—toxicity is any behaviour that ruins the fun for other players online. So, if we put these ideas together then we can say online toxicity is when someone purposely acts mean to make others feel bad in digital spaces. But it's not easy because sometimes people use "toxicity" and "abuse" in different ways, which can make things confusing. It seems like any mean stuff online can be called toxic, and vice versa. This shows that agreeing on what words mean is tough, which makes it harder to deal with online toxicity. Online toxicity covers a lot of bad things people do online. From cyberbullying to hate speech [4–8], these actions can really hurt people and make online spaces less enjoyable. Even though we don't have one clear definition—the main goal is still the same—to stop these behaviours and make the internet a safer and nicer place for everyone.Past studies such as [9] have looked a lot at how to classify toxic behaviour online but haven't agreed on how to do it. That means different studies have different ways of sorting toxic stuff online, even if they mean the same thing. [2] compared a bunch of these studies that used computers to learn how to sort toxic stuff online. They first looked at the training sets, which are the groups of examples used to teach the computer what's toxic and what's not. Then, made a simple list of toxic stuff categories like sexism, racism, saying hateful things, using offensive language, being mean to women, being aggressive, insults, threats, hating on someone's identity, and just generally being toxic. To make this list—they used a tool called FastText to teach the computer about words and sentences. But found out that even with all this—couldn't really connect the different ways people sorted toxic stuff in the past. They realized that the way the computer learned depended a lot on the examples it was given and how often they showed up. So, they said it is super important to have lots of good examples to teach computers and to be clear about what's toxic and what's not. Another study by [10] talked about different kinds of toxic stuff. They said there's the toxic stuff that's obvious, like bad words and threats. Then there's things that is sneakier, like sarcasm and, some toxic matter is aimed at certain groups, like being racist. Meanwhile, [11] had their own way of sorting toxic stuff. They looked at whether it was obvious or sneaky and whether it was aimed at someone specific or just anyone. They found that knowing who the toxic stuff was aimed at—helped understand it better but wasn't necessary to know for sure if something was mean. So, all these studies show that they need to agree more on how to sort toxic stuff online. Therefore, this study uses Bidirectional Encoder Representations from Transformers (BERT) to solve these challenges.

The study is as follows; the background will be seen in the following section. The related works are presented in Section III. The methods and materials are detailed in Section IV. The experimental analysis is carried out in Section V, and in Section VI, we provide some conclusions and plans for future research.

2. Background

Gaming industry is growing really fast—in 2021—it was worth $138.4 billion. Experts think it will be doubled to $300 billion by 2025. More and more people are playing games all around the world because it is getting easier with the new technology. Platforms like Stadia and Amazon Luna let you play games without needing expensive equipment. During the COVID-19 pandemic, gaming became even more popular. Since people were stuck at home they looked for things to do for fun. Many people started playing games more often. The World Health Organization (WHO) even encouraged gaming with their #PlayApartTogether campaign. During the pandemic, people started playing different kinds of games, like battle royales and games where you play with lots of other people online. Online gaming went up by 60%. Research shows that playing games has lots of good effects. It can help you solve problems better, learn languages, and make friends. This is especially helpful for kids and older people. Playing games can also make you feel better emotionally. Most people who play multiplayer games say they have good social experiences. But gaming isn't all good. There's a lot of toxic behaviour, too. Many gamers say they've been harassed while playing online. Some have even faced serious threats like stalking and violence. Some games, like DOTA 2, Valorant, and Grand Theft Auto, are known for having lots of toxic behaviour. Even though lots of toxic things happen—most people don't do anything about it. Even professional gaming has problems with toxic behaviour. This is because pro gamers are always competing, and they get a lot of attention. After a while, some of them start being toxic to other players. Toxic behaviour in gaming can be lots of things, like bullying, being mean to people, or even doing things that can hurt others in real life. It is important to do something about toxic behaviour in gaming. Game companies need to make sure games are safe for everyone to play. Even though there are problems—the gaming industry keeps getting better. There are lots of chances for the industry to grow and make a positive impact, as long as it deals with toxic behaviour. A study by the [12] found that 22% of gamers have been harassed. Some even quit playing games because of it. Game companies are trying to fix this by letting players report toxic behaviour. But sometimes, these systems do not work well because of false reports or because people think they don't do anything. Different companies try different things to deal with toxic players. For instance, [13] tells players when someone reports them. Amazon Lunaand Valve try putting toxic players together in matches or limiting what they can do until they behave better. [3], which makes League of Legends, even patented a system that puts nice players together in matches. Talking to other players in games can often lead to harassment. This happens a lot in voice and text chats. According to the [12] study, muting is the best way to deal with this. Some game companies reward players for being good. Game companies also use psychology to make games better. For instance, [3] uses colours to encourage players to be nice to each other. Outside of gaming, Reddit, a popular online forum, lets people vote on posts. Reddit has some people and machines that make sure everyone follows the rules. Even though Reddit is really popular, it is still hard to keep things positive. They remove millions of posts every year for reasons like spam or being toxic. Being open about what they do is important for sites like Reddit. In 2020, they showed how much they do to keep things in check. This includes taking down millions of posts and banning thousands of accounts. As the internet keeps changing, dealing with toxicity is a big challenge for gaming and social media sites like Reddit.

3. State of the Art

Several studies have been published such as [10] looked at different ways to find abuse in language. They split these ways into three groups—lexicon and rule-based, computational, and neural network methods. First, lexicon and rule-based use dictionaries of bad words, which work well in different situations but cannot always, catch mistakes or hidden abuse. Second, computational ways use Machine Learning (ML) features like Term Frequency - Inverse Document Frequency (TF-IDF) and *n*-grams, which are good at handling mistakes but sometimes miss deeper meanings. Neural network methods, like Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), are a mix—sometimes they do better than other methods. In the GermEval 2019 contest, [14]—found out that the BERT model was better than CNNs [15–21] and RNNs at finding abuse. OffensEval, a part of SemEval, focuses on finding abuse in language using datasets like Online Hate Speech Detection (OLID). [22] tried different models like Support Vector Machine (SVM), Bidirectional Long Short-Term Memory (BiLSTM), and CNN. Even though OLID was good, [23]—made a new dataset called ETHOS, and it helped a lot, especially for BERT models. They used experienced people to label the data in OffensEval, so it was high quality. They made a simple way to find toxic language and its targets in OffensEval. The perspective Application Programming Interface (API) made by [24] can find different kinds of toxic language in text very well, but sometimes it doesn't do well. It's hard to get enough data to train the models, so sometimes they get things wrong—especially with hidden abuse. The BERT model is one of the best at finding abuse because it looks at words from both sides and is very good with big datasets. Models like HateBERT, trained on bad Reddit posts by [25]—can be even better than BERT. They showed that having a good data and doing things like pre-processing are really important for finding abuse.Some researchers studied such as [26]—why people are toxic in a game called League of Legends (LoL). They found that when people want to win or do not get along with their team—they're more likely to be toxic. They did not just look at online forums—they also looked at stories about toxicness in the game. [27] looked at LoL players in China. They made a tool to see who is toxic in the game. They found that experienced players are meaner than new ones. But they couldn't tell if playing a tough character made someone toxic. They also looked at how long players stick with the game. Experienced players stay longer, even if their team is toxic. In 2021, [28] looked at toxicness in different games. They used tools to see who’s toxic on Reddit and Twitter. They found that almost all games have the same amount of toxicness—about 20%. Some games have more racism or sexism than others. All these studies show that being toxic in games is a big problem. It's not just about some games being worse than others. It's about how people act when they play games together.

4. Materials and Methods

This study focuses at how toxic behaviour compares in gaming groups. We say we need a clear idea of what toxic behaviour means to study it properly. We used BERT [29–32] to help find and understand toxic behaviour. We looked at gaming groups on Reddit— clean up the data—and use BERT to learn from it. We wanted to see if our ideas about toxic behaviourare right. We say toxic behaviour online is when people are mean to others on purpose. We use two levels to understand it better—Level A and Level B. Level A checks if a post is toxic, while Level B checks if it's toxic on purpose to hurt someone.

4.1. Data Analysis

We focused on 14 popular gaming discussion areas on Reddit—including both multiplayer and single-player games. To collect the data—we used the Pushshift Multithread API Wrapper (PMAW) to get posts from January 2021. We set a limit of 500,000 posts per area to keep the data manageable. We also cleaned up the data by removing special characters to make it easier to work with. In the data, we found the posts that said '[removed]' or '[deleted]'—this often means a moderator or someone else took down the post—possibly because it was toxic. We also spotted posts made by bots and took those out too— though figuring out if they were toxic was tricky. One problem we faced was that some posts were too long for BERT to handle all at once. So, we had to cut them down. We also looked at how to take out letters that aren't English. We used a tool called langdetect in Python. It can find 55 languages right and are correct 99% of the time for 53 of them. We made a program to take data from .csv files and make sentences based on what langdetect finds. But,langdetect had trouble with comments starting with symbols or web links. These comments were marked as 'NA' to keep the program running smoothly. To see if the program worked well—we tested it on data from the Among Us Reddit. This data had about 200,000 things written, with some foreign words. The program took almost two hours to run and didn't do great. It found 32 languages, saying 77% of the comments were English, 17% were not English, and 6% were 'NA'. But when we checked, many comments it said were not English were actually in English. We saw this by looking closely at the comments marked as Spanish—of the 872 comments—the program said were Spanish—only 94 were really in Spanish. That's just 11% right. The program had a hard time with short comments, especially those marked as Spanish. Most of these were less than 50 characters long. We also found that langdetect had trouble with mistakes in spelling, short forms, and internet slang. This made it less accurate. Because the program took so long to run, didn't work well, and only a small part of the comments were really in other languages—we decided it wasn't practical to take out non-English comments.

4.2. Model Analysis

The BERT-base-uncased model—is the basis for this setup. It's used to create two models for different classifications—one for NOT/OFF and another for UNT/TIN. The system checks for available Graphics Processing Unit (GPU) and installs the HuggingFace transformers package using PyTorch [33–38]. This package helps in using transformer-based models like BERT easily. Data is stored in Google Drive and brought in using pandas dataframes. The BERT tokenizer is then loaded with specific settings like ignoring capitalization and adding special tokens. Inputs are divided into tokens using *encode_plus*() and put into PyTorch tensors with attention masks. To handle memory well, the data is divided into batches by using PyTorchDataloaders. Training begins by loading the BertForSequenceClassification model and Adam optimizer. A learning rate scheduler is used—but not for warm-up steps in this case. The training process starts with a random seed for reproducibility. In each cycle, the model is set to training mode, and batches from the training dataloader are processed. Gradients are cleared, and the model is trained. The loss is calculated and gradients are clipped to avoid issues. The optimizer and scheduler adjust hyperparameters accordingly. After training, the model is tested on the test data. Predictions are made, and evaluation metrics are calculated using scikit-learn's classification report. However, for the TIN/OFF model, which gets correctly labeled OFF comments from the previous model, separate calculations are necessary. In short, the setup follows a structured process—importing data, setting up the model, training, and evaluating. We also make a special tool to help with predictions. This tool takes some text as input. It then uses a tool we loaded earlier to break down the text, and finally, it uses the trained model to guess what the text is about. We use this tool on lots of posts from Reddit, one after the other, using a pandas dataframe. Then, we gather all the results into a dataframe, turn it into a .csv file, and store it in Google Drive for more studying. But there's a problem we run into when looking at the data. Some posts are just one word—'NA'. This messes up the BERT model when it tries to read them. It seems like part of the BERT model thinks 'NA' means nothing—which causes errors and stops the process. Even though there aren't many 'NA' posts in different sets of data—we need to get rid of them to keep everything running smoothly. Doing these predictions, especially for figuring out if something is ON or OFF, takes a lot of computer power and time. It can take more than a week to finish all the data sets. Plus, sometimes the computer connection breaks during these long processes. To make things better, we split the data into smaller parts. This way, we can do shorter runs more often, which lower the chances of the connection breaking. This change doesn't just make things smoother; it also makes the predictions faster. Instead of taking 5 hours for one big run, we can do several 1-hour runs. It's interesting that we don't need to do this for the UNT/TIN model. It only has to work with some comments that we already labeled, so there are fewer posts to predict.

5. Experimental Analysis

We test the model with Reddit data. We decided to label part of the Reddit data to test and improve the model. Therefore, we followed the OLID annotation rules to make sure the labeling was consistent—especially when marking offensive content and who it was aimed at. First, we made a program to take random samples from 14 gaming-related subreddits. We took 100 samples from each subreddit to cover different gaming terms. We did this a few times to get more data for training and testing. We made sure to remove non-English comments and posts made by bots. Offensive language, like swearing or insults about gameplay or people—got marked as offensive. But it was tricky to tell if the offense was targeted or not, especially when people hated on game developers. We also had to label slang words and abbreviations used online or in games. We used sites like *UrbanDictionary.com* for help. For example, we found that "LMFAO"—despite having bad words—was not offensive when used positively. We wanted to understand how gamers talk online better. We tried to make the toxicity detection model better by using two HateBERT models. These models looked at banned posts from Reddit to understand online language and feelings better. Even with these improvements, we still needed to fine-tune the model like the original BERT model. We wanted to see if multiplayer game subreddits were more toxic than single-player ones. We thought subreddits with more rules might have less toxic posts. We also wanted to check if games rated for older players had more toxic posts than games for younger players. We considered things like how frustrating a game might be and who uses Reddit [1].

5.1. BERT Optimization

We tried different settings like learning rates, epochs, and batch sizes. We tested these settings on two classifiers—A and B—which are important for finding toxic content. After testing, we found that using 2 or 3 epochs didn't work well because the loss didn't improve enough. Therefore, we decided to use 4 epochs for recall, precision, and F1 scores. Tables I and IIshows what we found from testing these settings. The BERT model performed better than the SVM, BiLSTM, and CNN models—like the OLID. Specifically, the BERT model worked consistently well with the Level A classifier—improving a bit with different settings. But, the TIN/UNT classification varied a lot with different settings—mostly because there weren't many UNT examples in the test data. Interestingly, the recall for the UNT classifier was quite low, always doing worse than the CNN. But, it had much higher precision—meaning it was good at finding most UNT instances. This happened because the training and test data had mostly TIN content—about 88% and 89% each. The Level A—NOT/OFF classifier worked best when we trained it with a batch size of 16 and a learning rate of 2.00e^{-05}. On the other hand, the Level B—TIN/UNT classifier got the best results with a batch size of 16 and a learning rate of 5.00e^{-05}. We used these exact settings when making Reddit toxicity detection models based on these classifiers. The models we chose had good loss convergence, especially the UNT/TIN model. Although it would've been useful to try more than 4 epochs to find the best loss convergence—the models we picked balanced low loss with high F1, recall, and precision scores.

Table 1. Model Performance on not and off Classes.

Batch Size	Learning Rate	Precision (NOT)	Recall (NOT)	F1 Score (NOT)	Precision (OFF)	Recall (OFF)	F1 Score (OFF)	Macro Average Precision	Macro Average Recall	Macro Average F1 Score
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	0.90	0.89	0.72	0.68	0.70	0.80	0.79	0.79
32	2.00e^{-05}	0.88	0.90	0.89	0.72	0.68	0.70	0.80	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	0.70	0.70	0.80	0.79	0.80
16	2.00e^{-05}	0.89	0.90	0.89	0.73	0.70	0.72	0.81	0.80	0.81

Table 2. Model Performance on Tin and Unt Classes.

Batch Size	Learning Rate	Precision (TIN)	Recall (TIN)	F1 Score (TIN)	Precision (UNT)	Recall (UNT)	F1 Score (UNT)	Macro Average Precision	Macro Average Recall	Macro Average F1 Score
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	0.40	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	0.70	0.44	0.54	0.81	0.71	0.75
16	3.00e^{-05}	0.91	0.96	0.93	0.50	0.29	0.37	0.70	0.62	0.65
16	2.00e^{-05}	0.93	0.96	0.94	0.63	0.44	0.52	0.78	0.70	0.73

5.2. Modelling Toxicity

At first, we made a model to decide if posts were toxic or not toxic using data from OLID. But, this model was good at finding toxic posts but bad at saying which ones were really toxic. We added data from Reddit—which made the model better, especially at finding toxic posts. Then, we made another model to tell if posts were meant for someone or not. The first model didn't do so well because it was too focused on posts meant for someone as shown in Table III. But, when we added Reddit data—it got better at finding posts not meant for someone. Still, there were worries about getting the balance right between being accurate and not missing things. There were more worries about if the numbers we got were right because we only looked at the correct mean posts from the first model. When we looked at all the toxic posts—even the ones it got wrong—the numbers changed a lot. When we looked at all the models together—we saw that the first one wasn't so good at finding mean posts but was okay at getting not toxic ones right. Adding Reddit data made things a bit better, but the model still wasn't great at finding toxic posts. Adding Reddit data also helped with making fewer mistakes and saying fewer nice posts were toxic, but all the models still got a lot of toxic posts wrong. The first model was especially bad at this, maybe because it had too many posts meant for someone. It also struggled with understanding internet talk and gaming words, often saying nice things were toxic. The Reddit 1 model got better at finding posts it said weren't toxic but still had trouble with some internet talk. Even with these improvements, the models still need more work to really get toxic and not toxic posts right—especially when people are talking in a special way online or about games. The Reddit 2 model is a big improvement in catching wrong stuff compared to older versions. It only got 34 things wrong. 62% of these were right as OFF+UNT, and the rest were labeled NOT.

Table 3. Level a model performance metrics for offensive and non-offensive detection.

Model	Precision (NOT)	Recall (NOT)	F1 Score (NOT)	Precision (OFF)	Recall (OFF)	F1 Score (OFF)	Macro Average Precision	Macro Average Recall	Macro Average F1 Score
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

However, it was good at finding offensive stuff, but not so good at Level B classification, like Reddit 1 as shown in Table IV. Seven posts that Reddit 1 got right were seen wrong by Reddit 2, with five of these at Level B. All models had trouble labeling things like strong self-hate, which were often marked as not aimed at anyone. Some game actions, like being 'blown up', or descriptions with bad words, were wrongly seen as toxic. Adding more Reddit data might help fix this, but it wasn't part of the study. Using Reddit data made more mistakes happen. This made it harder to remember things when more batches were added. The OLID model found 29 wrong things, mostly labeled as NOT as shown in Table V, because it was more likely to see offensive posts as targeted insults. Reddit batch 1 and 2 models hardly changed wrong NOT flags, showing it's still hard to find hidden abuse. Reddit 1 found 46 wrong things, with Reddit 2 finding slightly more at 50, mostly labeled as NOT. Both models still struggled with Level B classification, often seeing bad words as not harmful because of differences between OLID and Reddit. Some cases, like "it's a game u weirdo", were seen as NOT by both Reddit models, showing problems with accuracy. HateBERT was a bit better than BERT at catching offensive stuff as shown in Tables VI and VII—but not so good at finding non-offensive stuff. HateBERT was worse than BERT at Level B, maybe because Reddit data didn't have enough targeted offenses like OLID. Checking banned Reddit comments showed many weren't really offensive, so there wasn't much improvement. Even though it helped find bad stuff in forums, there wasn't enough Reddit data compared to BERT's original data. Using specific platform data to train BERT was better than adding extra training steps.

Table 4. Level B model performance metrics for targeted and untargeted detection.

Model	Precision (TIN)	Recall (TIN)	F1 Score (TIN)	Precision (UNT)	Recall (UNT)	F1 Score (UNT)	Macro Average Precision	Macro Average Recall	Macro Average F1 Score
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	0.810	0.688	0.681	0.684

Table 5. Metrics calculated from combined Level A and B models on reddit test data.

Model	Non-Toxic (NOT/ OFF+UNT) - Precision	Non-Toxic (NOT/ OFF+UNT) - Recall	Non-Toxic (NOT/ OFF+UNT) - F1 Score	Toxic (OFF + TIN) - Precision	Toxic (OFF + TIN) - Recall	Toxic (OFF + TIN) - F1 Score	Macro Average Precision	Macro Average Recall	Macro Average F1 Score
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	0.980	0.337	0.394	0.363	0.660	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 6. Metrics of bert against hatebert for classifying reddit test data for levels A and B for offensive and non-offensive detection.

Model	Precision (NOT)	Recall (NOT)	F1 Score (NOT)	Precision (OFF)	Recall (OFF)	F1 Score (OFF)	Macro Average Precision	Macro Average Recall	Macro Average F1 Score
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

Table 7. Metrics of bert against hatebert for classifying reddit test data for levels A and B for targeted and untargeted detection.

Model	Precision (TIN)	Recall (TIN)	F1 Score (TIN)	Precision (UNT)	Recall (UNT)	F1 Score (UNT)	Macro Average Precision	Macro Average Recall	Macro Average F1 Score
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

6. Conclusions and Future Works

This study focused at different ways to understand toxicity and how computers can help find it using BERT. It gives a clear idea of what toxicity means using OLID classifiers. When choosing data, it sees that Reddit pages about multiplayer games are busier than those about single-player games. But how well these pages are watched by moderators varies a lot. The tool 'langdetect' does not work well with Reddit's way. We finds that using less than four tries (epochs) doesn't work well. Mixing Reddit's own labeled data with OLID Twitter data helps find toxic stuff on Reddit better. But using HateBERT, which knows a lot about Reddit, doesn't help much. It's better to adjust the model than add more Reddit knowledge. Looking at OLID classifiers and measurements, we finds problems with some of the ways to judge if a model is good. Mistakes in labeling data make the computer less accurate. The models say multiplayer game pages on Reddit have more bad stuff, but age ratings don't make a big difference. The last model doesn't do well because there aren't many toxic posts on Reddit. It is hard to make the model better because the study is small. Maybe comparing different BERT models or adding more info from Reddit could help. Fixing language detection and looking at how bad gaming talk affects kids could be good next steps. To sum up, the study shows that finding and understanding toxic stuff online is hard. We says we need to use data from the same place we are looking and suggests ways to keep online spaces safer for gamers.

References

1. Oxford University Press, “Assessment noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced American Dictionary,” 2024. https://www.oxfordlearnersdictionaries.com/definition/american_english/toxicity (accessed May 24, 2024).
2. P. Fortuna, J. Soler-Company, and L. Wanner, “Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets,” in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020, pp. 6786–6794. Accessed: May 24, 2024. [Online]. Available: https://aclanthology.org/2020.lrec-1.838
3. P. D. Falko, L. Leuphana, and U. Lüneburg, “The Success of the Freemium Business Model. How Riot Games flourishes with a free to play game,” *Manager Journal*, vol. 29, no. 1, pp. 114–124, 2019, Accessed: May 24, 2024. [Online]. Available: https://www.proquest.com/openview/fa79ec9ae04a87cebb761a62c21f4f1a/1?pq-origsite=gscholar&cbl=2032296

4. V. Kanaparthi, "Credit Risk Prediction using Ensemble Machine Learning Algorithms," in *6th International Conference on Inventive Computation Technologies, ICICT 2023 - Proceedings*, 2023, pp. 41–47. https://doi.org/10.1109/ICICT57646.2023.10134486.
5. V. K. Kanaparthi, "Examining the Plausible Applications of Artificial Intelligence & Machine Learning in Accounts Payable Improvement," *FinTech*, vol. 2, no. 3, pp. 461–474, Jul. 2023. https://doi.org/10.3390/fintech2030026.
6. V. Kanaparthi, "Robustness Evaluation of LSTM-based Deep Learning Models for Bitcoin Price Prediction in the Presence of Random Disturbances," Jan. 2024. https://doi.org/10.21203/RS.3.RS-3906529/V1.
7. V. Kanaparthi, "Evaluating Financial Risk in the Transition from EONIA to ESTER: A TimeGAN Approach with Enhanced VaR Estimations," Jan. 2024. https://doi.org/10.21203/RS.3.RS-3906541/V1.
8. V. K. Kanaparthi, "Navigating Uncertainty: Enhancing Markowitz Asset Allocation Strategies through Out-of-Sample Analysis," Dec. 2023. https://doi.org/10.20944/PREPRINTS202312.0427.V1.
9. S. Donaldson, "I predict a riot: Making and breaking rules and norms in league of legends," in *Proceedings of the 2017 DiGRA International Conference, DiGRA 2017*, 2017.
10. P. Mishra, H. Yannakoudakis, and E. Shutova, "Tackling Online Abuse: A Survey of Automated Abuse Detection Methods," Aug. 2019, Accessed: May 24, 2024. [Online]. Available: https://arxiv.org/abs/1908.06024v2
11. Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding abuse: A typology of abusive language detection subtasks," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, May 2017, pp. 78–84. https://doi.org/10.18653/v1/w17-3012.
12. Y. W. Jeong, Y. R. Han, S. K. Kim, and H. S. Jeong, "The frequency of impairments in everyday activities due to the overuse of the internet, gaming, or smartphone, and its relationship to health-related quality of life in Korea," *BMC Public Health*, vol. 20, no. 1, pp. 1–16, Jun. 2020. https://doi.org/10.1186/s12889-020-08922-z.
13. A. Grossman, "Nihilistic Software's VAMPIRE: THE MASQUERADE—REDEMPTION . by robert huebner," in *Postmortems from Game Developer*, Routledge, 2021, pp. 62–73. https://doi.org/10.4324/9780080522159-9.
14. A. Paraschiv and D. C. Cercel, "UPB at GermEval-2019 task 2: BERT-based offensive language classification of German tweets," in *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019*, 2020, pp. 398–404. Accessed: May 24, 2024. [Online]. Available: https://www.researchgate.net/publication/337007402
15. S. Wazir, G. S. Kashyap, and P. Saxena, "MLOps: A Review," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.10908v1
16. S. Naz and G. S. Kashyap, "Enhancing the predictive capability of a mathematical model for pseudomonas aeruginosa through artificial neural networks," *International Journal of Information Technology 2024*, pp. 1–10, Feb. 2024. https://doi.org/10.1007/S41870-023-01721-W.
17. G. S. Kashyap, K. Malik, S. Wazir, and R. Khan, "Using Machine Learning to Quantify the Multimedia Risk Due to Fuzzing," *Multimedia Tools and Applications*, vol. 81, no. 25, pp. 36685–36698, Oct. 2022. https://doi.org/10.1007/s11042-021-11558-9.
18. N. Marwah, V. K. Singh, G. S. Kashyap, and S. Wazir, "An analysis of the robustness of UAV agriculture field coverage using multi-agent reinforcement learning," *International Journal of Information Technology (Singapore)*, vol. 15, no. 4, pp. 2317–2327, May 2023. https://doi.org/10.1007/s41870-023-01264-0.
19. S. Wazir, G. S. Kashyap, K. Malik, and A. E. I. Brownlee, "Predicting the Infection Level of COVID-19 Virus Using Normal Distribution-Based Approximation Model and PSO," Springer, Cham, 2023, pp. 75–91. https://doi.org/10.1007/978-3-031-33183-1_5.
20. P. Kaur, G. S. Kashyap, A. Kumar, M. T. Nafis, S. Kumar, and V. Shokeen, "From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility," Feb. 2024, Accessed: Mar. 21, 2024. [Online]. Available: https://arxiv.org/abs/2402.16142v1
21. G. S. Kashyap, A. Siddiqui, R. Siddiqui, K. Malik, S. Wazir, and A. E. I. Brownlee, "Prediction of Suicidal Risk Using Machine Learning Models." Dec. 25, 2021. Accessed: Feb. 04, 2024. [Online]. Available: https://papers.ssrn.com/abstract=4709789
22. M. Ostendorff, P. Bourgonje, M. Berger, J. Moreno-Schneider, G. Rehm, and B. Gipp, "Enriching BERT with knowledge graph embeddings for document classification," in *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019*, Sep. 2020, pp. 307–314. Accessed: May 24, 2024. [Online]. Available: https://arxiv.org/abs/1909.08402v1
23. I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "ETHOS: a multi-label hate speech detection dataset," *Complex and Intelligent Systems*, vol. 8, no. 6, pp. 4663–4678, Jun. 2022. https://doi.org/10.1007/s40747-021-00608-2.
24. H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving Google's Perspective API Built for Detecting Toxic Comments," Feb. 2017, Accessed: May 24, 2024. [Online]. Available: https://arxiv.org/abs/1702.08138v1

25. T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "HateBERT: Retraining BERT for Abusive Language Detection in English," in *WOAH 2021 - 5th Workshop on Online Abuse and Harms, Proceedings of the Workshop*, Oct. 2021, pp. 17–25. https://doi.org/10.18653/v1/2021.woah-1.3.
26. L. A. Nexø and S. Kristiansen, "Players Don't Die, They Respawn: a Situational Analysis of Toxic Encounters Arising from Death Events in League of Legends," *European Journal on Criminal Policy and Research*, vol. 29, no. 3, pp. 457–476, Sep. 2023. https://doi.org/10.1007/s10610-023-09552-y.
27. J. C. Aguerri, M. Santisteban, and F. Miró-Llinares, "The Enemy Hates Best? Toxicity in League of Legends and Its Content Moderation Implications," *European Journal on Criminal Policy and Research*, vol. 29, no. 3, pp. 437–456, Sep. 2023. https://doi.org/10.1007/s10610-023-09541-1.
28. A. Ghosh, "Analyzing Toxicity in Online Gaming Communities," Apr. 2021. Accessed: May 24, 2024. [Online]. Available: https://www.turcomat.org/index.php/turkbilmat/article/view/5182
29. V. Kanaparthi, "Examining Natural Language Processing Techniques in the Education and Healthcare Fields," *International Journal of Engineering and Advanced Technology*, vol. 12, no. 2, pp. 8–18, Dec. 2022. https://doi.org/10.35940/ijeat.b3861.1212222.
30. V. Kanaparthi, "Exploring the Impact of Blockchain, AI, and ML on Financial Accounting Efficiency and Transformation," Jan. 2024, Accessed: Feb. 04, 2024. [Online]. Available: https://arxiv.org/abs/2401.15715v1
31. V. Kanaparthi, "Transformational application of Artificial Intelligence and Machine learning in Financial Technologies and Financial services: A bibliometric review," Jan. 2024. https://doi.org/10.1016/j.jbusres.2020.10.012.
32. V. Kanaparthi, "AI-based Personalization and Trust in Digital Finance," Jan. 2024, Accessed: Feb. 04, 2024. [Online]. Available: https://arxiv.org/abs/2401.15700v1
33. G. S. Kashyap *et al.*, "Detection of a facemask in real-time using deep learning methods: Prevention of Covid 19," Jan. 2024, Accessed: Feb. 04, 2024. [Online]. Available: https://arxiv.org/abs/2401.15675v1
34. M. Kanojia, P. Kamani, G. S. Kashyap, S. Naz, S. Wazir, and A. Chauhan, "Alternative Agriculture Land-Use Transformation Pathways by Partial-Equilibrium Agricultural Sector Model: A Mathematical Approach," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.11632v1
35. G. S. Kashyap *et al.*, "Revolutionizing Agriculture: A Comprehensive Review of Artificial Intelligence Techniques in Farming," Feb. 2024. https://doi.org/10.21203/RS.3.RS-3984385/V1.
36. G. S. Kashyap, D. Mahajan, O. C. Phukan, A. Kumar, A. E. I. Brownlee, and J. Gao, "From Simulations to Reality: Enhancing Multi-Robot Exploration for Urban Search and Rescue," Nov. 2023, Accessed: Dec. 03, 2023. [Online]. Available: https://arxiv.org/abs/2311.16958v1
37. G. S. Kashyap, A. E. I. Brownlee, O. C. Phukan, K. Malik, and S. Wazir, "Roulette-Wheel Selection-Based PSO Algorithm for Solving the Vehicle Routing Problem with Time Windows," Jun. 2023, Accessed: Jul. 04, 2023. [Online]. Available: https://arxiv.org/abs/2306.02308v1
38. H. Habib, G. S. Kashyap, N. Tabassum, and T. Nafis, "Stock Price Prediction Using Artificial Intelligence Based on LSTM– Deep Learning Model," in *Artificial Intelligence & Blockchain in Cyber Physical Systems: Technologies & Applications*, CRC Press, 2023, pp. 93–99. https://doi.org/10.1201/9781003190301-6.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.