Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 49-58
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos originales
Temática: Inteligencia artificial
Recibido: 21/09/2023 | Aceptado: 05/12/2023 | Publicado: 30/03/2024

Identificadores persistentes:
DOI: 10.48168/innosoft.s15.a124
ARK: ark:/42411/s15/a124
PURL: 42411/s15/a124

Clasificación de comentarios tóxicos en los Videojuegos

Classification of toxic comments in video games

Luis Fernando Luque Nieto [1] [0009-0008-2937-9620]*, **Elmerson Ramith Portugal Carpio** [2]

[1] Universidad La Salle. Arequipa, Perú. lluquen@ulasalle.edu.pe
[2] Universidad La Salle. Arequipa, Perú. eportugalc@ulasall.edu.pe

* Autor para correspondencia: lluquen@ulasalle.edu.pe

Resumen

La toxicidad puede tener un gran impacto en el compromiso y la satisfacción del jugador. Se trata de un fenómeno complejo que tiene causas y consecuencias diversas. Entre las causas más comunes se encuentran la anonimidad, la competitividad y la frustración. Las consecuencias pueden ser graves, como el acoso, el abandono del juego y el daño psicológico. Las empresas de juegos están trabajando para encontrar formas de abordar las formas de toxicidad en sus plataformas. Una de las interacciones más comunes con la toxicidad se produce en las ventanas de chat o en los sistemas de mensajería del juego. El trabajo propuesto es sacar algunos mensajes de chat que se dan en estos “lobby” o sacarlos de internet para así poder clasificarlos y determinar si el jugador que escribió en el chat cometió una infracción y dependiendo de la categoría tomar acciones en el caso.

Palabras clave: Aprendizaje automático, Chat tóxico, Procesamiento de lenguaje natural, Videojuegos.

Abstract

Toxicity can have a major impact on player engagement and satisfaction. It is a complex phenomenon that has diverse causes and consequences. Among the most common causes are anonymity, competitiveness and frustration. The consequences can be serious, such as harassment, abandonment of the game and psychological damage. Gaming companies are working to find ways to address forms of toxicity on their platforms. One of the most common interactions with toxicity occurs in chat windows or in-game messaging systems. The proposed work is to pull some chat messages that occur in these "lobbies" or take them offline so that they can be categorized to determine if the player who wrote in the chat committed an infraction and depending on the category take action on the case.

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 49-58
https://revistas.ulasalle.edu.pe/innosoft

Keywords: *Machine learning, Toxic chat, Natural language processing, Video games.*

Introducción

La toxicidad en los videojuegos, un fenómeno destacado en la era digital, se manifiesta a través de comportamientos agresivos y comunicación perjudicial en comunidades de jugadores en línea. Desde comentarios despectivos hasta amenazas y expresiones discriminatorias, estos comportamientos afectan negativamente la experiencia de juego, generando un ambiente hostil. La toxicidad se caracteriza por la agresión hacia adversarios y compañeros, contribuyendo a emociones desagradables y dificultades comunicativas. Factores como la anonimidad en línea, competitividad extrema y falta de moderación influyen en su prevalencia. Evaluar y abordar esta problemática de manera eficiente es un desafío; sin embargo, la aplicación del Procesamiento del Lenguaje Natural (NLP) ofrece una solución, permitiendo a usuarios y desarrolladores obtener calificaciones que indican el grado de toxicidad en los comentarios, agilizando el proceso y mejorando la experiencia de juego.

Motivación

Nuestro trabajo busca ayudar y fomentar un ambiente de tranquilidad y serenidad a los jugadores de la gran mayoría de juegos que tienen chat de texto incluido, para que les dé una clasificación general si un usuario tiene muchas quejas de comentarios tóxicos podría ser tachado como persona “no grata”. Las preguntas que busca responder este proyecto son:

P1: ¿Cuál es la proporción de cada categoría sobre un comentario de chat?
P2: ¿Cuál es la palabra que más se repite en cada categoría?
P3: ¿Cómo afecta la competitividad a la toxicidad en los videojuegos en términos de mental?
P4: ¿Qué características del lenguaje se pueden utilizar para identificar comentarios tóxicos?
P5: ¿Cómo se puede utilizar la educación para combatir la toxicidad en los videojuegos?
P6: ¿Cómo se puede utilizar el NLP para ayudar a los jugadores a identificar y responder a los comentarios tóxicos en los lobbys?

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 49-58
https://revistas.ulasalle.edu.pe/innosoft

Trabajos Relacionados

El NLP se utilizará para analizar los mensajes en los chats de los juegos. El NLP se utilizará para identificar palabras y frases que se asocian con la toxicidad, como insultos, amenazas y acoso. En [2], este trabajo aborda el problema de la toxicidad en los juegos multijugador en línea. Se utiliza un novedoso marco de procesamiento del lenguaje natural para detectar malas palabras y clasificar los comentarios tóxicos en función de su gravedad. Los resultados muestran que la toxicidad está relacionada de manera no trivial con el éxito del juego, y cómo es que impacta a la hora de tomar decisiones importantes en la partida.

En el trabajo [3], habla del trolling es un comportamiento tóxico que se observa con frecuencia en las plataformas de comunicación en línea, pero sigue siendo un fenómeno poco claro. Esto se debe a que hay poca investigación empírica sobre él y a que no hay consenso entre los expertos sobre su definición. Este artículo tiene como objetivo proporcionar una visión general del trolling presentando una revisión de la literatura anterior sobre los comportamientos tóxicos que se consideran trolling en las comunidades en línea y en el contexto de los juegos en línea. Además, se realizó un cuestionario a 83 participantes para observar cómo perciben el trolling. El estudio también encontró que la percepción del trolling varía de persona a persona. Algunos participantes perciben el trolling como un comportamiento divertido e inofensivo, mientras que otros lo perciben como un comportamiento dañino y malicioso. El estudio concluye que el trolling es un problema complejo que requiere más investigación. Se necesitan más estudios para comprender mejor las motivaciones de los trolls. En el estudio [4], habla de cómo los juegos en línea son populares entre los jugadores para comunicarse, discutir estrategias y hacer amigos. Sin embargo, a menudo se enfrentan a discursos abusivos y de acoso.

Para abordar este problema, se utilizó un conjunto de datos de ciberbullying recopilados de los foros de World of Warcraft (WoW) y League of Legends (LoL) para entrenar modelos de clasificación, como Toxic-BERT, con el objetivo de detectar automáticamente comentarios abusivos. Los resultados muestran que el modelo Toxic-BERT logra una puntuación macro F1 del 82,69% para el foro LoL y del 83,86% para el foro WoW en el conjunto de datos de ciberbullying. Esto ayuda a mantener los foros de juegos limpios y amigables al identificar y eliminar comentarios ofensivos de manera automática En el trabajo [5] se aborda el crecimiento significativo de la interacción social en línea, que a menudo está plagada de comportamientos hostiles o agresivos, como el ciberacoso. Se desarrolla un clasificador

de lenguaje tóxico basado en audio utilizando Redes Neuronales Convolucionales (CNN) auto-atentas. A diferencia de depender de términos de léxico individuales, se toma un enfoque más general para identificar expresiones tóxicas que considere el contexto acústico completo de las frases cortas o expresiones. La arquitectura propuesta utiliza el mecanismo de autoatención para capturar la dependencia temporal del contenido verbal al resumir toda la información relevante de diferentes partes de la expresión. El modelo de CNN auto-atentas basado en audio se evalúa en un conjunto de datos público y otro interno, logrando un 75% de precisión, un 79% de recall y un 80% de recuperación en la identificación de grabaciones de discursos tóxicos.

Propuesta

Primero se buscó un dataset ya clasificado por gente experta en la materia de clasificación de textos, donde se evaluaron diferentes métricas para seleccionar uno. Antes de realizar el procesamiento propiamente dicho, es necesario llevar a cabo el pre procesamiento de los datos. Esto implica eliminar palabras vacías, puntuación, caracteres especiales y otros elementos innecesarios que no aportan información relevante al análisis. Después de haber preprocesador los datos, se procede al análisis de la toxicidad. En este paso, se calculan las probabilidades de las categorías de toxicidad que tiene el dataset, así como la categoría predominante en cada reseña. Esto permitirá comprender el tipo de texto que es de acorde a la categoría. Una vez clasificadas las reseñas, se exportan los datos en un archivo CSV para su posterior análisis. Esta exportación facilita el procesamiento y la manipulación de los resultados obtenidos. Finalmente, se procede a la visualización gráfica de los datos, donde las reseñas se muestran agrupadas por categorías del producto.

Esto permite identificar patrones y tendencias en las opiniones de los usuarios sobre los productos analizados. (ver Fig. 1).

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 49-58
https://revistas.ulasalle.edu.pe/innosoft

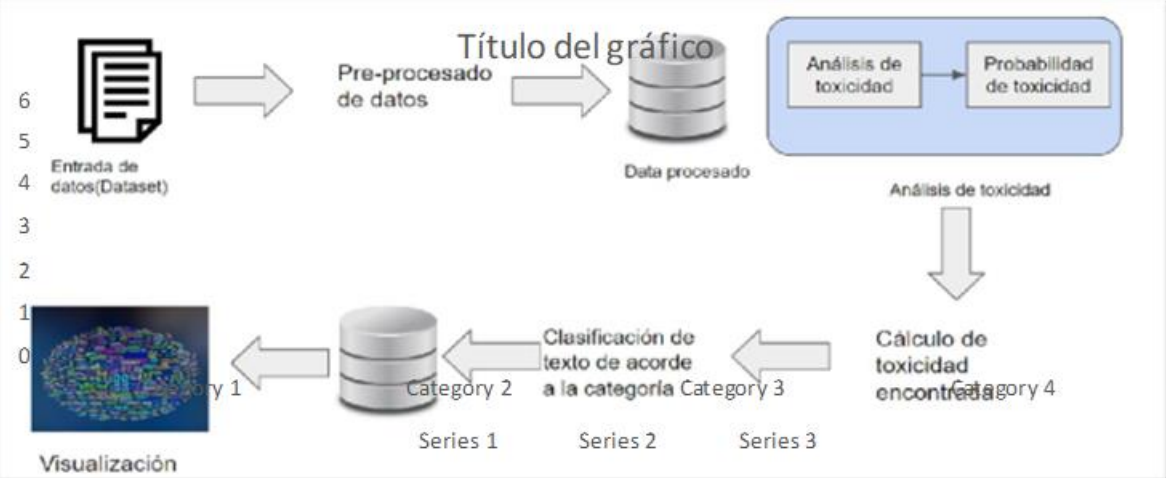

Figura 1. Pipeline Propuesto

Descripción de Data

El datasat consiste en más de 500 mil registros en csv, con su id, el comentario y la clasificación, siendo las clases “toxic”,“severe-toxic”, “obscene”, “threat”, “insult”,"identity_hate".

Tabla 1. Atributos y Descripción

Atributo	**Descripción**
Id	Identificador del comentario
Comment_text	El comentario como tal
classification	Tipo de toxicidad en el cual clasificado

Implementación

Para este trabajo, seguimos los pasos descritos en el algoritmo de la Tabla 2, implementado en Python, para abordar la clasificación de mensajes de chat en entornos de videojuegos con respecto a su toxicidad.

Tabla 2. Algoritmo Entrenamiento

Algorithm 1: Entrenamiento del Modelo

```
def train_model(model, train_loader, val_loader, optimizer, device, max_steps_per_epoch, num_epochs):
1.train_loss_history = [] 5
2.val_loss_history = []
3.val_accuracy_history = []
4.for epoch in range(num_epochs):
5.model.train()
6.total_loss = 0
7.steps_taken = 0
8.for batch in train_loader:
9.if steps_taken >= max_steps_per_epoch:
10.break
11.input_ids, attention_mask, labels = batch
12.input_ids,attention_mask,labels=input_ids.to(device), attention_mask.to(device), labels.to(device)
13.optimizer.zero_grad()
14.outputs = model(input_ids, attention_mask=attention_mask, labels=labels)
15.loss = outputs.loss
16.total_loss += loss.item()
17.loss.backward()
18.optimizer.step()
19.train_loss_history.append(loss.item())
20.steps_taken += 1
21.if (steps_taken + 1) % 10 == 0:
22.print(f"Training Step {steps_taken + 1}: Loss = {loss.item()}")
23.model.eval()
24.val_loss = 0
25.correct_predictions = 0
```

```
26.total_predictions = 0
27.with torch.no_grad():
28.for val_batch in val_loader:
29.val_input_ids, val_attention_mask, val_labels = val_batch
30.val_input_ids,val_attention_mask,val_labels =
val_input_ids.to(device),val_attention_mask.to(device),val_labels.to(device)
31.val_outputs = model(val_input_ids, attention_mask = val_attention_mask, labels = val_labels)
32.val_loss += val_outputs.loss.item()
33.predictions = torch.argmax(val_outputs.logits, dim=1)
34.correct_predictions += torch.sum(predictions == val_labels).item()
35.total_predictions += len(val_labels)
36.val_accuracy = correct_predictions / total_predictions
37.val_loss_history.append(val_loss / len(val_loader))
38.val_accuracy_history.append(val_accuracy)
39.print(f"Epoch {epoch + 1}/{num_epochs}: "
40.f"Train Loss = {total_loss / steps_taken}, "
41.f"Val Loss = {val_loss / len(val_loader)}, "
42.f"Val Accuracy = {val_accuracy}")
```

Debido a que no contamos con los recursos físicos necesarios para poder hacer el decidimos utilizar Google Colab para agilizar el procesamiento de cada archivo CSV en la etapa de búsqueda de cada palabra de un comentario en los diccionarios. Nos centramos en la clasificación de comentarios tóxicos en videojuegos utilizando el modelo BERT. Para reducir el tiempo de procesamiento, implementamos la biblioteca nativa multiprocessing de Python.

Resultados

El entrenamiento fue realizado en google Colab, con una TPU proporcionado por la misma google. El tamaño del dataset era de unos 40 MB aprox. Tuvimos algunas limitaciones. ya que Colab solo nos daba un tiempo límite de 12

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 49-58
https://revistas.ulasalle.edu.pe/innosoft

horas (algunas veces solo eran 8). Generamos un Word-cloud, para cada categoría en la cual se puede apreciar cuáles son las palabras más usadas por cada categoría.

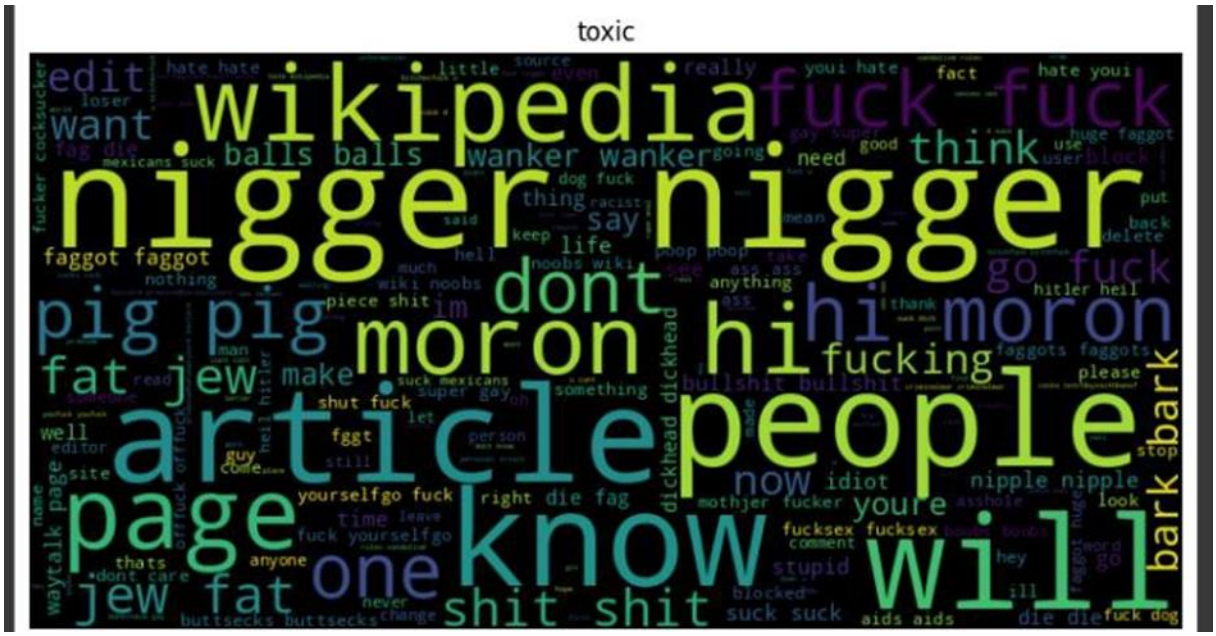

Figura. 2. WordCloud

Dadas las limitaciones de tiempo y al no contar con los recursos físicos necesarios, nuestro entrenamiento no pudo ser el esperado, el entrenamiento fue de 8 épocas y con 50 pasos por épocas. Con una función que evalúa el modelo se logró sacar los siguientes datos: Recall: F1 = Verdadero Positivo Verdadero Positivo + Falso Negativo Donde el F1 score se calculó de la siguiente manera: F1 = 2 * precisión * recall / (precision + recall)

Tabla. 2. Resultados de Bert

Precisión	***Recall***	***F1 Score***
0.8325	*0.8900*	*0.8590*

Conclusiones y trabajos futuros

A lo largo de este proyecto, las restricciones temporales y de recursos han tenido un impacto significativo en el entrenamiento del modelo de Procesamiento del Lenguaje Natural (PLN). Aunque los indicadores de F1 Score y Precisión proporcionan resultados alentadores, es evidente que la duración limitada del entrenamiento ha afectado la capacidad predictiva del modelo.

A pesar de estas limitaciones, hemos adquirido conocimientos sustanciales en el campo del Aprendizaje Automático y el funcionamiento general de los modelos de Procesamiento del Lenguaje Natural. Reconocemos la necesidad de una inversión adicional de tiempo y recursos para lograr un entrenamiento más exhaustivo, lo que, a su vez, mejoraría la capacidad de clasificación del sistema.

Este proceso de aprendizaje ha sido invaluable, proporcionando una comprensión más profunda de cómo, incluso bajo condiciones subestimas, un algoritmo respaldado por un conjunto de datos sólido puede aprender y mejorar sus capacidades. Aunque las circunstancias actuales han impuesto limitaciones, aspiramos a contar con los recursos necesarios para llevar a cabo un entrenamiento más prolongado y detallado en el futuro.

Contribución de Autoría

Luis Fernando Luque Nieto: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Elmerson Ramith Portugal Carpio:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura.

Referencias

[1] https://apdev.org.pe/la-toxicidad-en-los-esports/

[2] Märtens, M., Shen, S., Iosup, A., & Kuipers, F. (2015, December). Toxicity detection in multiplayer online games. In 2015 International Workshop on Network and Systems Support for Games (NetGames) (pp. 1-6). IEEE.

[3] Komaç, G., & Çağıltay, K. (2019, November). An overview of trolling behavior in online spaces and gaming context. In 2019 1st International Informatics and Software Engineering Conference (UBMYK) (pp. 1-4). IEEE.

[4] Vo, H. H. P., Tran, H. T., & Luu, S. T. (2021, August). Automatically detecting cyberbullying comments on online game forums. In 2021 RIVF International Conference on Computing and Communication Technologies (RIVF) (pp. 1-5). IEEE.

[5] Yousefi, M., & Emmanouilidou, D. (2021, August). Audio-based toxic language classification using self-attentive convolutional neural network. In 2021 29th European Signal Processing Conference (EUSIPCO) (pp. 11-15). IEEE