UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO
FACULTAD DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA, INFORMÁTICA Y MECÁNICA
ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA Y DE SISTEMAS

PLAN DE TESIS

Detección de toxicidad en chats de videojuegos multijugador considerando estilos lingüísticos generacionales

Para optar al título profesional de:
INGENIERO INFORMÁTICO Y DE SISTEMAS

Presentado por:
Pumachoque Choquenaira Jhon Esau

Código de alumno:
210940

Asesor:
Mgt. Harley Vera Olivera

Co-asesor:
Mgt. Harley Vera Olivera

Perú, Mayo de 2025

Resumen

Este trabajo de investigación propone mejorar la detección de toxicidad en chats de videojuegos multijugador mediante modelos Transformer (como BERT) adaptados a variaciones lingüísticas generacionales, abordando el sesgo actual que lleva a clasificar erróneamente expresiones propias de ciertos grupos generacionales. Se busca aumentar la precisión y equidad del sistema, beneficiando tanto a la industria con moderación más justa y retención de jugadores como a las comunidades gamer reduciendo falsas sanciones y toxicidad real.

Índice general

1 | Antecedentes

Se presenta los antecedentes revisados que se utilizaron como base inicial del trabajo de investigación, estos son antecedentes directamente relevantes para el proyecto.

Cuadro 1.1: Comparación de artículos sobre comunicación y toxicidad en videojuegos y redes sociales

Artículo	**Contribución**	**Limitaciones**	**Data Sets**	**Enfoque generacional**	**Metodología**
Courtesy Under Fire: A Structural and Contextual Analysis of Toxic Language in Online Gaming	Patrones lingüísticos de toxicidad (voz/texto, género de juegos)	Sesgo geográfico (Polonia), muestra pequeña	300 chats (Valorant, LoL, Minecraft)	No aplica	Observación participante y análisis cualitativo
Conversing in MMO Games: A Discourse Analysis of Chat Interactions in WoW and LoL	Comparación sociolingüística WoW vs LoL (cooperación/competencia)	Encuestas no estandarizadas, solo angloparlantes	183 encuestas y chats	No explícito	Encuestas y análisis estadístico/discursivo
Generation grouping according to Beresfod research and intensity of use of language-games in global communications using English	Juegos de lenguaje aplicados a generaciones (Baby Boomers a Gen Alpha)	Muestra no representativa, solo inglés	100 encuestas (Tegal City)	Sí, 6 generaciones	Chi-cuadrado y encuestas
How Do Different Generations Communicate on Social Media? A Comparative Analysis of Language Styles, Emoji Usage, and Visual Elements	Análisis multigeneracional (texto/emojis/memes)	Solo angloparlantes, sesgo de plataformas	4000 posts (Twitter, FB, IG)	Sí, 4 generaciones	ANOVA + NVivo

1.1. Lenguaje toxico en los videojuegos online

El estudio “Courtesy Under Fire: A Structural and Contextual Analysis of Toxic Language in Online Gaming” (Stojak, 2024) analiza la prevalencia del lenguaje tóxico en videojuegos multijugador como Valorant, League of Legends y Minecraft, exami-

nando su impacto en la experiencia de los jugadores a través de estructuras lingüísticas y contextos de agresión. Mediante observación participante (2017-2024) y el análisis de 300 casos de chats de texto y voz, el estudio categoriza la toxicidad en tres niveles (denigración, discriminación y provocación emocional) e identifica patrones como errores ortográficos y repetición de insultos, destacando diferencias entre comunicación escrita y oral, así como entre géneros de juegos. Aunque es pionero en detallar estos aspectos lingüísticos, presenta limitaciones como sesgo geográfico (datos principalmente de Polonia), muestra reducida y falta de enfoque generacional.

1.2. La comunicacion en juegos masivos online

El estudio "Conversing in Massive Multiplayer Online (MMO) Games: A Discourse Analysis of Chat Interactions in World of Warcraft and League of Legends" Bogdanov (2022) realiza un análisis comparativo de las interacciones lingüísticas en los chats de World of Warcraft (WoW) y League of Legends (LoL), basado en encuestas a 183 jugadores (68 de WoW y 115 de LoL) y análisis de capturas de chat. Los resultados revelan diferencias clave entre ambos juegos: en WoW predomina un discurso cooperativo, con mayor formalidad, uso de vocabulario especializado y construcción de identidades sociales en guilds, mientras que en LoL el lenguaje es más agresivo e informal, con abundantes abreviaciones y un enfoque competitivo. Aunque el estudio destaca por ser pionero en el análisis sociolingüístico comparativo de estos géneros (MMORPG vs. MOBA), presenta limitaciones metodológicas, como el reducido tamaño muestral y el enfoque exclusivo en angloparlantes.

1.3. El lenguaje generacional: juegos de lenguaje

El estudio "Generation grouping according to Beresfod research and intensity of use of language-games in global communications using English" Huda and Susdarwono (2023) aborda las diferencias generacionales (desde Baby Boomers hasta Alpha) en el empleo de juegos de lenguaje, expresiones creativas y coloquiales en la comunicación global en inglés, los autores buscan correlacionar la generación con la frecuencia de uso de estos recursos, aunque la crítica señala limitaciones en la representatividad de la muestra y el enfoque exclusivo en inglés, proponiendo futuras réplicas en entornos multiculturales y el análisis de plataformas específicas como TikTok para Gen Z. Este trabajo aporta una perspectiva innovadora al vincular teoría filosófica con dinámicas comunicativas generacionales.

1.4. Comunicacion entre las diferentes generaciones en las redes sociales

El estudio “How Do Different Generations Communicate on Social Media? A Comparative Analysis of Language Styles, Emoji Usage, and Visual Elements” Azad et al. (2023) examina las diferencias generacionales (desde Baby Boomers hasta Gen Z) en estilos lingüísticos, uso de emojis y elementos visuales en redes sociales, mediante el análisis de 4,000 publicaciones en Twitter, Facebook e Instagram. Utilizando herramientas como NVivo y pruebas ANOVA, los autores identificaron patrones clave, revelando que las generaciones más jóvenes (Millennials y Gen Z) emplean un lenguaje más informal, mayor frecuencia de emojis y una integración más dinámica de memes, mientras que generaciones anteriores (Baby Boomers y Gen X) mantienen un estilo más textual y formal. Aunque el estudio se limita a angloparlantes, su enfoque multigeneracional y multimodal (texto, emojis, imágenes) aporta una visión integral de la evolución comunicativa en entornos digitales.

2 | Marco Teórico

(En este apartado el estudiante debe presentar en forma breve los temas que se abordarán en el marco teórico. El marco teórico se desprende de las palabras clave y del titulo. Debe ser totalmente acotado. Ésta sección describe conceptos relacionados a procesamiento de trayectorias y detección de trayectorias anómalas .

2.1. Lenguaje y Comunicación en Videojuegos Multijugador

La comunicación en videojuegos multijugador se refiere a los sistemas, estrategias y métodos que permiten a los jugadores interactuar, coordinarse y socializar dentro de entornos virtuales colaborativos o competitivos. Esta comunicación puede adoptar diferentes modalidades: verbal (interacción en tiempo real mediante micrófonos), textual (mensajes escritos de forma sincrónica o asincrónica) y no verbal (acciones rápidas mediante menús o botones, animaciones de personajes, señales visuales, entre otros) Spyridonis et al. (2018).

En particular la comunicación textual en videojuegos multijugador constituye un eje fundamental para la coordinación estratégica y la interacción social entre jugadores. Sin embargo, este medio también se erige como el principal canal para la manifestación de conductas tóxicas, las cuales impactan negativamente en la experiencia de juego, el rendimiento individual y colectivo, y la sostenibilidad de las comunidades virtuales Neto et al. (2017).

2.2. Toxicidad en lo videojuegos

No existe una definición universalmente aceptada de comportamiento tóxico en los entornos digitales y videojuegos, debido a que diversos investigadores manejan conceptos relacionados como comportamiento desviado, ciberacoso, troleo o antisocialidad Canossa et al. (2021). En términos generales, se entiende como cualquier comportamiento que interfiere intencionalmente en la experiencia y bienestar de otros jugadores,

incluyendo acciones como insultos ("flaming"), trampas o acoso.

Organizaciones como la Anti-Defamation League (ADL) definen la toxicidad como "comportamiento disruptivo", abarcando desde insultos y amenazas hasta acoso sostenido y divulgación no autorizada de información personal. La Fair Play Alliance refuerza esta idea proponiendo el uso del término "comportamiento disruptivo", enfatizando que la interrupción de la experiencia de juego puede ser tanto deliberada como consecuencia del diseño del juego o emparejamientos inadecuados.

Con el objetivo de sistematizar la comprensión de la toxicidad, Kou (2017) y Kowert (2020) como se cita en Canossa et al. (2021) proponen taxonomías que clasifican los tipos de comportamientos y factores contextuales que los provocan. Kou destaca que la toxicidad es una secuencia situada de emociones o acciones que afectan negativamente la cooperación del equipo, mientras que Kowert introduce el concepto de "participación oscura", que agrupa comportamientos online que, al dañar la salud o el bienestar de otros, son considerados tóxicos.

En el videojuego For Honor, Canossa et al. (2021) definen la toxicidad a partir de las sanciones aplicadas a jugadores que infringieron el Código de Conducta. Los jugadores se clasifican como sancionados (amonestados o expulsados) por conductas ofensivas o por obtener ventajas injustas. Estas categorías constituyen la base para etiquetar y predecir comportamientos tóxicos mediante técnicas automatizadas.

2.3. Diferencias Generacionales en Estilos Lingüísticos

La comunicación digital ha evolucionado de manera significativa, marcando contrastes notables entre generaciones. Cada grupo etario desarrolla estilos lingüísticos propios influenciados por su contexto tecnológico, cultural e histórico. Estudios como el de Huda and Susdarwono (2023) demuestran que estas diferencias no son aleatorias, sino que responden a patrones claros vinculados a la adaptación a herramientas digitales y a la apropiación de nuevos formatos comunicativos. Por ejemplo, mientras los Baby Boomers (1946–1964) mantienen un lenguaje formal y estructurado, la Generación Z (1997–2012) emplea códigos abreviados, memes y multimodalidad (texto-imagen) con naturalidad, reflejando su condición de nativos digitales.

2.4. Deteccion y Analisis de toxicidad

La detección de la toxicidad se basa en identificar comportamientos dañinos como el acoso, el discurso de odio, los insultos y el sabotaje intencional ("griefing"). Estos comportamientos suelen analizarse mediante dos enfoques principales: métodos tradicionales y técnicas avanzadas basadas en inteligencia artificial. Los enfoques tradicionales incluyen sistemas basados en reglas, como listas de palabras prohibidas y

algoritmos de coincidencia de patrones, que son simples pero limitados en precisión debido a su incapacidad para captar el contexto o la ironía. Por ejemplo, un modelo de aprendizaje automático como SVM puede clasificar mensajes tóxicos en chats de juegos, pero su rendimiento depende en gran medida de la calidad del conjunto de datos de entrenamiento (Raman et al., 2020, citado en Zhuo et al. (2025) y Wijkstra et al. (2023)).

Los modelos modernos de procesamiento de lenguaje natural (PLN), como BERT y GPT, han mejorado significativamente la detección al analizar el contexto y las sutilezas del lenguaje. Estos modelos pueden identificar no solo insultos directos, sino también formas más encubiertas de toxicidad, como el sarcasmo o el acoso indirecto. Sin embargo, su implementación enfrenta desafíos, como el sesgo algorítmico. Estudios han demostrado que herramientas como Perspective API pueden etiquetar erróneamente dialectos no estándar, como el inglés afroamericano (AAVE), como tóxicos con mayor frecuencia que el inglés estándar Sap et al. (2019). Además, la naturaleza dinámica del lenguaje en los videojuegos —donde términos como ”kill.º ”noob”pueden ser técnicamente neutrales o incluso amistosos— complica aún más la clasificación.

El análisis de la toxicidad no solo implica su detección, sino también la comprensión de sus patrones y efectos. En el estudio de Kordyaka et al. (2020) han identificado que la competitividad, el anonimato y la falta de consecuencias inmediatas fomentan comportamientos tóxicos. Además, el análisis de redes de interacción en juegos multijugador revela que la toxicidad tiende a propagarse, creando entornos hostiles que disuaden a los nuevos jugadores. Para abordar este problema, algunos investigadores proponen el uso de modelos de aprendizaje automático que no solo detecten, sino que también predigan la toxicidad basándose en patrones de comportamiento previos Canossa et al. (2021).

Aun hay desafíos importantes, como la falta de conjuntos de datos equilibrados y la subjetividad en la definición de toxicidad. Mientras que para algunos jugadores ciertos comentarios como inofensivos hechos por otros jugadores, otros los consideran muy ofensivos.

Bibliografía

Azad, I., Chhibber, S., and Tajhizi, A. (2023). How do different generations communicate on social media? a comparative analysis of language styles, emoji usage, and visual elements. *Language, Technology, and Social Media*, 1(2):86–97.

Bogdanov, M. (2022). *Conversing in massive multiplayer online (MMO) games: a discourse analysis of chat interactions in World of Warcraft and League of Legends*. PhD thesis, Doktorska disertacija.

Canossa, A., Salimov, D., Azadvar, A., Harteveld, C., and Yannakakis, G. (2021). For honor, for toxicity: Detecting toxic behavior through gameplay. *Proceedings of the ACM on Human-Computer Interaction*, 5(CHI PLAY):1–29.

Huda, T. and Susdarwono, E. T. (2023). Generation grouping according to beresfod research and intensity of use of language-games in global communications using english.

Kordyaka, B., Jahn, K., and Niehaves, B. (2020). Towards a unified theory of toxic behavior in video games. *Internet Research*, 30(4):1081–1102.

Neto, J. A., Yokoyama, K. M., and Becker, K. (2017). Studying toxic behavior influence and player chat in an online video game. In *Proceedings of the international conference on web intelligence*, pages 26–33.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Spyridonis, F., Daylamani-Zad, D., and O'Brien, M. P. (2018). Efficient in-game communication in collaborative online multiplayer games. In *2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 1–4. IEEE.

Wijkstra, M., Rogers, K., Mandryk, R. L., Veltkamp, R. C., and Frommel, J. (2023). Help, my game is toxic! first insights from a systematic literature review on intervention systems for toxic behaviors in online video games. In *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 3–9.

Zhuo, H., Yang, Y., and Peng, K. (2025). Combating toxic language: A review of llm-based strategies for software engineering. *arXiv preprint arXiv:2504.15439*.