Revisión sistemática sobre el análisis de sentimientos en interacciones por chat en videojuegos

John Edison Lozano González

Asesor

Jorge Luis Quintero López

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería - ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Resumen

El análisis de sentimiento en la detección de comportamiento inadecuado en las interacciones de chat es un campo en crecimiento con diversas aplicaciones en la comprensión del comportamiento humano en entornos virtuales, y la importancia de este enfoque radica en la necesidad de herramientas automatizadas capaces de predecir la polaridad del sentimiento expresado en texto en tiempo real, siendo útil para detectar toxicidad y amenazas potenciales para la experiencia de juego. La revisión bibliográfica se enfoca en explorar las últimas metodologías y técnicas empleadas en este análisis aplicado a interacciones de chat en videojuegos en línea y streaming. Se examinaron diversas estrategias de procesamiento de texto y modelos de aprendizaje automático para identificar desafíos y limitaciones en la implementación de sistemas de análisis de sentimientos en entornos de chat en línea, así como posibles áreas de investigación futura para abordar estos desafíos.

***Palabras clave*:** análisis de sentimiento, videojuegos, toxicidad, red social.

Abstract

Sentiment analysis in detecting inappropriate behavior in chat interactions is a growing field with various applications for understanding human behavior in virtual environments. The importance of this approach lies in the need for automated tools capable of predicting the sentiment polarity expressed in text in real-time, which is useful for detecting toxicity and potential threats to the gaming experience. This literature review focused on exploring the latest methodologies and techniques used in this analysis, applied to chat interactions in online video games and streaming platforms. Various text processing strategies and machine learning models were examined to identify challenges and limitations in the implementation of sentiment analysis systems in online chat environments, as well as potential research areas to address these challenges.

***Keywords*:** sentiment analysis, video games, toxicity, social network.

Tabla de contenido

Lista de Tablas

Introducción

En la última década, el auge de los videojuegos en línea y las plataformas de streaming ha revolucionado la manera en que los jugadores se relacionan, creando comunidades dinámicas y vibrantes, sin embargo, esta evolución también ha dado lugar a comportamientos inadecuados, como el ciberacoso, la toxicidad y el uso de lenguaje ofensivo, que pueden perjudicar la experiencia de juego y la seguridad de los usuarios, especialmente de los más jóvenes. Ante esta problemática, el análisis de sentimientos se ha consolidado como una herramienta esencial para identificar y abordar estos comportamientos problemáticos en las interacciones de chat.

El presente trabajo, titulado "Revisión sistemática sobre el análisis de sentimientos en interacciones por chat en videojuegos", busca analizar la literatura existente sobre el análisis de sentimientos aplicado a los chats en videojuegos en línea y streaming. Este análisis es fundamental para entender cómo estas técnicas pueden ayudar a detectar comportamientos inapropiados y mejorar la calidad de la experiencia de los jugadores.

El estudio examinará diversos enfoques y metodologías utilizados en este campo emergente, proporcionando una visión integral de cómo el análisis de sentimientos puede contribuir a promover un entorno de juego más seguro y respetuoso. Al comprender mejor el impacto del análisis de sentimientos en la detección de comportamientos inadecuados, se espera fomentar interacciones más positivas en las comunidades de jugadores.

Planteamiento del Problema

El problema identificado es el aumento del odio y acoso en juegos multijugador en línea y según datos recopilados por la Liga Anti-Difamación (2023), aproximadamente el 75% de los jugadores de juegos en línea en los Estados Unidos han experimentado odio o acoso en los últimos seis meses, en comparación con el 67% del año anterior. Aunque el acoso y el odio dirigidos a adultos han disminuido, ha habido un aumento en casi todos los aspectos para los jóvenes de 10 a 17 años, independientemente del género del juego, destacando la urgencia de abordar este problema.

La investigación sobre comportamientos inadecuados en los chats de videojuegos es esencial, pues revela interacciones cruciales y refleja el estado de ánimo de la comunidad, mostrando cómo el incremento del comportamiento antisocial, como trolling y ciberacoso, afecta negativamente la salud mental de los jugadores, especialmente jóvenes. Sin embargo, a pesar de las medidas implementadas por los desarrolladores, la falta de control del comportamiento en el chat persiste, lo que subraya la necesidad de una investigación más profunda para proponer soluciones efectivas y automatizadas. Esto se debe a que las prohibiciones de chat y de cuentas, basadas en comentarios de usuarios, son costosas y no son viables para todos los desarrolladores. (Kumar, 2022).

La literatura presenta diversos modelos de inteligencia artificial para identificar comportamientos indebidos en las conversaciones de videojuegos, que incluyen tanto chatbots como clasificadores de texto. Negobot fue uno de los primeros en este campo, entrenado con datos del sitio PJ. Otros enfoques incluyen BotHook y modelos de clasificación como SVM y CNN. A pesar de las propuestas, aún no se han evaluado en grandes conjuntos de datos balanceados, ni se han aplicado directamente a plataformas de juegos para mitigar riesgos en

tiempo real. Se necesitan soluciones más efectivas para integrar estos modelos de detección y abordar el problema de conductas inapropiadas en los chats de videojuegos (Faraz et al., 2022).

Las investigaciones en este campo se enfocan en muestras específicas, limitando la generalización y replicación del fenómeno, especialmente en grandes volúmenes de datos en tiempo real. Abordar este desafío podría emplear el análisis de sentimientos para detectar comportamientos inadecuados en los juegos en línea, ofreciendo una guía adaptable para futuras investigaciones y la industria de los videojuegos. Como sugiere Chouhan et al. (2021) al abordar las particularidades de los mensajes, mejora la comprensión de las opiniones y emociones de los jugadores durante las interacciones. Considerando este contexto, la pregunta de investigación es determinar ¿Cómo puede el análisis de sentimientos en las conversaciones de videojuegos ayudar en la detección de comportamientos inapropiados? ¿Y cuál sería una estrategia de implementación que pueda servir como guía o punto de partida para desarrolladores de videojuegos y futuras investigaciones?

Justificación

El problema abordado se centra en el comportamiento inadecuado en los chats de videojuegos en línea, siendo fundamental investigar este fenómeno debido a la creciente accesibilidad de los videojuegos para niños, quienes pueden verse expuestos a conductas agresivas, discursos de odio y ciberacoso por parte de otros jugadores, lo cual conlleva consecuencias negativas para su experiencia y salud mental (Faraz, 2022). Es esencial abordar este desafío desde una perspectiva de análisis de sentimientos para comprender mejor estas interacciones y proponer medidas preventivas efectivas.

La literatura existente destaca deficiencias significativas, incluida la falta de datos reales sobre los chats en videojuegos, la limitación de análisis en idiomas distintos al inglés y la escasez de datos demográficos de los jugadores. Estas limitaciones dificultan la comprensión completa de los comportamientos inadecuados en los juegos en línea a nivel global, subrayando la importancia de investigaciones futuras para abordar estas lagunas (Murnion et al., 2018). Además, la falta de atención de los investigadores en el ámbito de los videojuegos, en comparación con otras redes sociales, resalta la necesidad de una mayor exploración en este campo.

El proyecto propuesto busca abordar las deficiencias actuales mediante la revisión exhaustiva de la literatura disponible en el campo de los videojuegos en línea, con un enfoque específico en el análisis de sentimientos, el cual es una herramienta potencial para mejorar la moderación de los entornos virtuales y mejorar la experiencia de los jugadores. Esta alternativa promete cambiar la gestión de estos entornos, pasando de una moderación basada en denuncias y restricciones a una identificación más rápida y efectiva de los infractores. Se espera que esta investigación contribuya a hacer que la industria de los videojuegos sea más segura, permitiendo

su implementación en diversas plataformas y sirviendo como punto de partida para futuras investigaciones en el campo (Borj et al., 2020).

La relevancia de esta investigación se fundamenta en el crecimiento exponencial de los juegos en línea y la consiguiente interacción entre los jugadores, lo cual incrementa los riesgos asociados con el comportamiento inapropiado (Woo et al., 2022). Este estudio busca impactar la mejora en los sistemas de filtrado y promover interacciones más positivas en los chats, con el objetivo de obtener entornos en línea más controlados. Esto reduciría la exposición a un ambiente negativo para los usuarios, y, además, los padres y tutores se beneficiarían al tener la certeza de que sus hijos están en entornos en línea más confiables.

Objetivos

Objetivo General

Analizar la literatura existente sobre el análisis de sentimientos en interacciones de chat de videojuegos en línea y en streaming, para comparar su eficacia en la identificación de comportamientos inadecuados.

Objetivos Específicos

Identificar los diferentes artículos bibliográficos sobre modelos de análisis de sentimiento para detectar comportamientos inadecuados en los videojuegos en línea y en streaming.

Explorar los métodos de preprocesamiento y análisis de sentimientos utilizados en la literatura, con el propósito de comprender su funcionamiento en el contexto de los videojuegos.

Examinar ventajas y limitaciones de los modelos de análisis de sentimientos en la detección de comportamientos inapropiados en chats de videojuegos y en streaming, comparando el F1-Score en investigaciones previas para evaluar su idoneidad y eficacia.

Marco Conceptual y Teórico

Videojuegos: Rivera-Arteaga y Torres-Cosío (2018) describen un videojuego como una forma de entretenimiento electrónica en constante evolución, donde uno o múltiples jugadores interactúan a través de una pantalla, aprovechando los avances tecnológicos y siendo adaptable a diversas plataformas. El videojuego también se puede definir como un medio visual digital interactivo en el que la experiencia de juego está regida por reglas o narrativas interactivas, con el propósito de entretener a los usuarios (Tavinor, 2008).

Análisis de sentimiento: Mäntylä et al. (2018) señala que el análisis de sentimientos "es una serie de métodos, técnicas y herramientas para detectar y extraer información subjetiva, como opiniones y actitudes del lenguaje, y centrándose en la polaridad de la opinión, es decir, si alguien tiene una opinión positiva, neutral o negativa hacia algo" (p. 1). El análisis de sentimientos, conocido también como minería de opiniones, se concentra en avaluar las emociones, juicios, evaluaciones y actitudes manifestadas en texto hacia entidades y sus características, y estas entidades pueden abarcar productos, servicios, organizaciones, individuos, eventos, asuntos o temas (Liu, 2020).

Los videojuegos han evolucionado significativamente desde su nacimiento como fuente de entretenimiento, atrayendo a personas de todas las edades y géneros, convirtiéndose en parte de la rutina diaria de muchas personas, ya sea como hobbies o como una carrera profesional en competiciones remuneradas, marcando así un avance tecnológico distintivo del siglo XXI. Esta industria cuenta con numerosos profesionales dedicados al diseño y desarrollo de juegos, comprometidos en satisfacer las crecientes expectativas de los usuarios en constante evolución.

El avance de Internet ha permitido que los videojuegos expandan su alcance, incorporando una mayor competencia y cooperación a través de partidas en línea, esta evolución

ha llegado al punto en que la comunicación es posible desde cualquier lugar del mundo, lo que ha llevado a los juegos multijugador a adoptar la característica de jugar en línea. Faraz (2022) sugiere que esta modalidad otorga una dimensión humana adicional al juego, dado que se necesita comunicarse para elaborar estrategias o simplemente conversar durante la partida, y esta comunicación puede ser por voz o texto a través de una interfaz, lo que permite mantener conversaciones tanto informales como relacionadas con aspectos específicos del juego a lo largo de la partida.

Las interacciones en ciertos juegos se asemejan a una red social, ofreciendo funciones de mensajería y llamadas de audio y vídeo, y además de la experiencia directa del juego, se están popularizando las transmisiones en vivo o streaming, el cual un streamer interactúa con los espectadores mientras juega, comparte consejos, reseñas o simplemente permite que otros observen. Twitch es una plataforma destacada para esto, similar a Facebook o YouTube, pero centrada en contenido de videojuegos, permitiendo la interacción entre el streamer y los espectadores. Los espectadores pueden comentar y participar activamente durante la emisión, creando así una comunidad interactiva en torno al contenido transmitido, y esta combinación de juego, entretenimiento y comunidad en línea ha transformado significativamente la experiencia de los videojuegos y la forma en que las personas se conectan en línea (Yildiz, 2022).

En algunas plataformas de juego en línea o streaming, los chats están abiertos y sin restricciones, lo que puede exponer a los niños a riesgos al interactuar con desconocidos. Con alrededor de 4.950 millones de usuarios de Internet, aproximadamente 1 de cada 3 son menores de 18 años, quienes a menudo navegan sin supervisión, enfrentando así un mayor riesgo de participar en actividades sociales en línea, incluido el ciberacoso (Faraz, 2022). Algunas plataformas, como las videoconsolas, ofrecen acceso parental para limitar el acceso a juegos

específicos y controlar las horas de juego, así como también monitorear su uso, y además, restringen ciertos tipos de juegos clasificados por etiquetas como PEGI, que incluyen contenido no apto para menores, aunque los padres restrinjan estos juegos, los niños pueden acceder a otros catalogados como aptos para su edad, pero que permiten la interacción en línea, lo que queda fuera del alcance de varios controles parentales.

El chat en los juegos en línea y en las transmisiones en streaming es una fuente valiosa de información, ya que registra las interacciones entre jugadores y espectadores, ya que proporciona una visión amplia de las preferencias y emociones de los usuarios, permitiendo analizar el comportamiento y el ánimo de la comunidad. Este ánimo puede manifestarse de forma positiva, promoviendo el compañerismo y una competencia saludable, sin embargo, también puede dar lugar a un ambiente malicioso, con comportamientos antisociales o problemáticos, como el griefing, el spam en el chat, el abuso de bugs y el ciberacoso, son comúnmente conocidos como "tóxicos" dentro de la comunidad de jugadores. Un griefer es aquel jugador que disfruta afectar negativamente la experiencia de otros, mientras que el spam en el chat puede ser irritante al saturar el canal con mensajes repetitivos, perturbando la comunicación y distrayendo a los jugadores. Esto puede impactar negativamente la salud mental, especialmente en jóvenes, al igual que el ciberacoso, que se extiende causando daño mediante abuso verbal y lenguaje inapropiado (Kumar, 2022).

Los jugadores, espectadores e investigadores han observado que algunos miembros de la comunidad muestran comportamientos en línea que difieren de su comportamiento en la vida real, se sienten más relajados en el ciberespacio, lo que les permite expresarse de manera más abierta y sin tantas restricciones. En relación con este fenómeno de estos usuarios, Sule (2004) empleó un término para designarlo:

> El efecto de desinhibición en línea: Esta desinhibición puede funcionar en dos direcciones aparentemente opuestas, a veces las personas comparten cosas muy personales sobre sí mismas y revelan emociones secretas, miedos, deseos. Muestran actos inusuales de amabilidad y generosidad, a veces desviviéndose por ayudar a los demás, podemos llamarlo desinhibición benigna. Sin embargo, la desinhibición no siempre es tan saludable, somos testigos de lenguaje grosero, duras críticas, ira, odio, incluso amenazas (p. 1).

De esta manera, la sensación de impunidad en línea lleva a las personas a comportarse de manera negativa, incluyendo la agresión verbal, amenazas y acoso, sin considerar las repercusiones. Este fenómeno se conoce como desinhibición tóxica, donde los límites sociales se debilitan en el entorno digital.

A medida que la industria de los juegos ha experimentado un crecimiento significativo, impulsado por avances tecnológicos y un mayor nivel de interacción social en su interior, también hemos presenciado un aumento alarmante en el comportamiento antisocial, lo que representa una oscura realidad dentro de los juegos sociales. Este comportamiento negativo ha desencadenado una serie de efectos adversos, incluyendo depresión, estrés, bajo rendimiento académico y laboral, e incluso casos extremos de suicidio. la mitad de los jugadores han declarado haber sido víctimas de acoso cibernético en algún momento (Murnion, 2018), y este problema no solo plantea preocupaciones de índole social, sino también económica, ya que las actividades nocivas asociadas con el acoso cibernético representan un coste considerable para los proveedores de juegos en línea , y según la Liga Anti-Difamación (2023), el 20% de los jugadores gastan menos dinero en espacios de juego online debido al odio y el acoso que encuentran.

La ausencia de regulación en el ciberespacio fomenta una desinhibición tóxica por parte de unos cuantos usuarios, quienes actúan sin considerar el impacto de sus acciones en otros

jugadores, lo que afecta negativamente su salud y bienestar. Este fenómeno es especialmente perjudicial para grupos vulnerables, como niños y mujeres, quienes enfrentan un mayor riesgo de sufrir abuso verbal y acoso en los juegos en línea. Como señalan Wohn y Lee (2013) los mensajes de voz femenina, en particular, reciben tres veces más comentarios negativos. Esta problemática se agrava en juegos con comunicación audiovisual. La falta de consecuencias para comportamientos dañinos crea un entorno poco acogedor, donde el miedo al acoso inhibe el disfrute del juego y la participación.

La influencia negativa en la atmósfera de los entornos de videojuegos se manifiesta a través de comentarios inadecuados o negativos, como el "trolling" (provocación de conflictos en línea), ciberacoso, "griefing" (sabotaje de la experiencia de juego) y "spam". Aunque se investiga la detección de estas formas de hostilidad en interacciones virtuales, la atención se centra en redes sociales como Facebook y Twitter, con menos enfoque en los juegos en línea, a pesar de su gran base de usuarios y el alcance del streaming en directo también es limitado, lo que dificulta abordar problemas en entornos de interacción específicos como el chat de videojuegos y streaming. Específicamente en el contexto de los videojuegos, los desarrolladores han adoptado medidas proactivas para abordar la hostilidad que puede surgir en sus plataformas, como la implementación de una serie de herramientas y políticas diseñadas para mitigar los comportamientos negativos dentro de los entornos de juego, el cual incluye la integración de sistemas de denuncia que permiten a los usuarios reportar de manera rápida y sencilla cualquier actividad inapropiada que encuentren mientras juegan. Con esto se comprende que la toxicidad en línea es un problema grave que ha llevado a las empresas de juegos a tomar medidas serias al respecto.

Un ejemplo notable es el caso de Riot Games, la mente maestra detrás del exitoso juego League of Legends. En 2011, Riot introdujo "The Tribunal", una innovadora plataforma en línea que reunía a jugadores expertos para colaborar en la evaluación de informes de mal comportamiento y determinar las sanciones apropiadas. A pesar de su lanzamiento con gran expectación y el entusiasmo inicial, la eficacia del Tribunal comenzó a verse comprometida por el rápido crecimiento de la comunidad de jugadores y la dificultad de mantener el ritmo de las denuncias. Tras su cancelación en 2018, se reveló que, durante su primer año de funcionamiento, el Tribunal había procesado un asombroso total de más de 47 millones de votos relacionados con comportamientos tóxicos, aunque la iniciativa tuvo sus limitaciones y desafíos, su breve pero impactante existencia marcó un punto de inflexión en la manera en que las empresas de videojuegos abordan la problemática de la toxicidad en línea. Blizzard, conocida por su popular juego Overwatch, adoptó una estrategia innovadora para abordar la conducta de los jugadores, en lugar de enfocarse exclusivamente en los comportamientos negativos, implementaron un sistema que destacaba y recompensaba el comportamiento positivo mediante insignias junto al nombre de los jugadores, con esta táctica resultó efectiva para reducir los comportamientos inapropiados en un 40%. Por otro lado, Valve, creadores de Counter-Strike: Global Offensive, tomaron medidas más drásticas al detectar conductas inadecuadas, los jugadores que recibían múltiples informes por su mal comportamiento eran silenciados, evitando así que influyeran negativamente en la experiencia de juego (Ekiciler et al., 2022).

A pesar de que algunos juegos en línea cuentan con sistemas para monitorear las quejas de los jugadores, los moderadores se enfrentan al desafío de analizar cada caso individualmente, y en entornos donde varios jugadores se conectan simultáneamente, la mensajería en el chat genera una gran cantidad de mensajes que requieren una detección rápida de comportamientos

inadecuados. La presencia constante de moderadores humanos en cada sesión de juego es impracticable, lo que destaca la importancia de la automatización para esta tarea (Williams, 2023). Aunque los desarrolladores de videojuegos han implementado iniciativas para abordar las denuncias, esto requiere una gran cantidad de trabajo, algo que los desarrolladores no pueden sostener debido al esfuerzo que conlleva, y se vuelve especialmente difícil con el crecimiento constante de la comunidad.

La exploración de modelos automatizados con inteligencia artificial (IA) para analizar mensajes de texto en entornos de juegos en línea ha visto un aumento significativo. En un estudio pionero, Thompson et al. (2017) examinaron el análisis de sentimientos en el chat del juego StarCraft online, utilizando técnicas de ponderación para capturar matices específicos asociados con la experiencia de juego, en el que destacaron la importancia de abordar la calidad del chat en el juego, ya que una baja calidad puede influir negativamente en la participación de los jugadores y, por ende, en su experiencia. Sin embargo, la implementación de estos modelos automatizados se enfrenta a desafíos importantes debido a la falta de datos de comunicación del juego disponibles, lo que se considera un obstáculo significativo en el desarrollo de soluciones efectivas para detectar comportamientos hostiles, aunque el sistema de "Tribunal" de League of Legends ha demostrado cómo los datos del chat del juego pueden ser útiles para fines de investigación, pero su cierre temporal desde 2014 ha dejado una brecha en la disponibilidad de datos para investigaciones futuras (Williams, 2023).

Para almacenar grandes volúmenes de datos sobre conversaciones que sigan un hilo, se requiere una inversión considerable, esto es especialmente relevante en pequeñas desarrolladoras de videojuegos, que suelen destinar sus recursos a la expansión de servidores, funciones relacionadas con el juego y su mantenimiento. Además, es importante comprender que las

empresas utilizan el tema de las conversaciones tanto para su beneficio como para resguardar la privacidad de sus usuarios. Sin embargo, es una limitación recurrente que se menciona en muchos artículos sobre el análisis de las interacciones entre jugadores en videojuegos es la falta de datos disponibles, lo cual ha dificultado que futuras investigaciones puedan acceder a la información y utilizarla con fines educativos, principalmente para mejorar el entorno de los videojuegos y su ambiente.

La escasez de investigaciones en el análisis del comportamiento, especialmente en el ámbito de los videojuegos, subraya la importancia de seguir explorando. Este enfoque no solo crea conciencia entre los desarrolladores y los usuarios de juegos, sino que también abre nuevas oportunidades para analizar los streams relacionados con los videojuegos mediante el uso de APIs, lo cual permite la extracción de datos valiosos, y además, existe la posibilidad de que, en el futuro, los desarrolladores compartan información de los chats, incluso si se anonimiza, con el fin de recibir asistencia y beneficiar a la gran comunidad de jugadores. Thompson et al. (2017) señalan que los juegos proporcionan una huella digital con un gran potencial para analizar tanto el rendimiento en el juego como la comunicación entre jugadores, como el chat o la mensajería dentro de los juegos y los streams. Estos textos son minas de oro para la minería de texto, una de las aplicaciones del análisis automatizado de sentimientos en la investigación de los usuarios de juegos es comprender el contexto social en el que se produce el aprendizaje.

La tarea de identificar sentimientos a partir de mensajes de texto plantea un desafío significativo, y varios investigadores han señalado el aprendizaje automático como una herramienta potencial para abordar comportamientos inapropiados. Por ejemplo, Balci y Salah (2015) investigó la detección de comportamientos agresivos o abusivos en el juego en línea Okey mediante técnicas de aprendizaje automático, como Bayes Point Machine, obteniendo resultados

prometedores. Otros estudios, como el de Cheong et al. (2015) en el juego MovieStarPlanet, aplicaron una variedad de modelos de clasificación, como Naïve-Bayes, árboles de decisión, regresión logística, k vecinos más cercanos, clasificadores de vectores de soporte y perceptrón multicapa (un modelo de red neuronal), con resultados interesantes en cada caso.

Esto resalta que el análisis de sentimientos puede emplear una amplia gama de modelos de aprendizaje automático, desde los tradicionales como el SVC hasta redes neuronales y transformadores, e incluso métodos menos convencionales, como la herramienta de análisis de sentimientos de Microsoft Azure, como se describe en Murnion et al. (2018). Algunos estudios utilizaron conjuntos de datos para entrenar modelos, mientras que otros simplemente aplicaron soluciones preexistentes, como el software Semantria for Excel utilizado por Ángeles-Gómez, D. y Quintana-López, M. (2019).

Revisión Sistemática

Esta revisión sistemática se centró en el análisis de sentimientos dentro del ecosistema de videojuegos, incluyendo tanto jugadores activos como espectadores en plataformas de streaming como Twitch, que tienen influencia sobre quienes aprenden o descubren nuevos juegos. Se excluyeron otras redes sociales debido a sus dinámicas distintas y su enfoque no relacionado directamente con los videojuegos, un campo aún poco explorado en la literatura.

Debido a la prevalencia de estudios en inglés y la limitada disponibilidad de artículos, no se aplicaron restricciones de idioma ni región. Los estudios seleccionados se limitaron estrictamente a aquellos relacionados con videojuegos, que abordan la problemática de comportamientos inapropiados y que emplean técnicas de ciencia de datos, como el análisis de sentimientos, para la detección de dichos comportamientos, lo que permitió encontrar artículos coherentes con el propósito de la monografía.

El desarrollo de esta monografía sigue el enfoque PRISMA (Page et al., 2021), adaptado para la búsqueda y evaluación sistemática de la literatura. Aunque PRISMA proporciona una estructura sólida, se permite cierta flexibilidad para ajustar la metodología según las necesidades específicas de la monografía.

Fuentes de Información

La búsqueda de estudios se llevó a cabo en bases de datos académicas reconocidas, incluyendo Google Scholar, IEEE Xplore, ScienceDirect, Semantic Scholar, Springer Link y Dialnet. La investigación se limitará a documentos publicados en los últimos diez años para asegurar la relevancia y actualidad de la información.

Estrategia de Búsqueda

Se utilizaron los siguientes términos de búsqueda combinados para identificar estudios pertinentes:

(ALL=("Multiplayer games" OR "Multiplayer game" OR "Multiplayer gaming" OR "Online game" OR "Online games" OR "Online gaming" OR "Online play" OR "esports" OR "e-sports" OR "video game" OR "video games" OR "video gaming" OR "MOBA" OR "FPS" OR "competitive game")) AND (ALL=("Cyberbullying" OR "Profanity" OR "sexual predator" OR "Censorship" OR "child protection" OR "predatory threats" OR "blasphemy" OR "toxic" OR "insult" OR "grief" OR "troll" OR "offensive" OR "inappropriate" OR "abuse" OR "flaming")) AND (ALL=("Sentiment analysis" OR "Semantic analysis" OR "Linguistic analysis" OR "sentiment text analysis")).

Se aplicarán filtros para restringir la búsqueda a los estudios más relevantes para el objetivo de la monografía.

Proceso de Extracción de Datos

Los datos se extrajeron utilizando una plantilla estructurada que incluye información general sobre cada documento, tales como técnicas empleadas, limitaciones, hallazgos clave, entre otros. Esta plantilla permitirá una recopilación uniforme y sistemática de la información.

Métodos de Síntesis

Los estudios seleccionados se tabularon y compararon para identificar los métodos más efectivos. La síntesis de los resultados se realizó comparando los métodos y enfoques utilizados en los estudios incluidos, justificando las elecciones basadas en la calidad y relevancia de estos, y cada estudio se describió con detalle sobre sus características metodológicas y hallazgos.

Recursos

Para el desarrollo de esta monografía, se emplearon bases de datos académicas como Google Scholar, IEEE Xplore, ScienceDirect, Semantic Scholar, Springer Link y Dialnet, lo que permitió una búsqueda exhaustiva y relevante de la literatura existente. Se utilizaron herramientas como Zotero o Mendeley para gestionar las referencias y mantener un control adecuado de las fuentes, y además, en Excel se empleó para el análisis de datos, asegurando una correcta organización y síntesis de la información con el apoyo de plantillas estructuradas para la extracción y registro de datos clave.

Desarrollo

El análisis de sentimientos aplicado a videojuegos representa una intersección clave entre la ciencia de datos y las ciencias comportamentales, destinada a interpretar las emociones y actitudes expresadas por los jugadores en diversas plataformas digitales. Este campo de estudio cobra relevancia no solo por su potencial para mejorar la experiencia de los usuarios, sino también por su capacidad de identificar comportamientos inadecuados y fomentar comunidades virtuales más saludables. En este contexto, el desarrollo de modelos eficaces requiere abordar desafíos técnicos y éticos, desde la recolección y preprocesamiento de datos hasta la validación e interpretación de los resultados, empleando herramientas avanzadas y técnicas de aprendizaje automático.

El presente capítulo introduce los aspectos fundamentales que constituyen el desarrollo de esta monografía, estructurados en torno a cuatro ejes principales: las fuentes de datos, el preprocesamiento del lenguaje natural, la implementación de modelos analíticos y la evaluación de ventajas y limitaciones. Al final, este trabajo tiene como propósito proporcionar una guía para que los desarrolladores de videojuegos y otros interesados en el ecosistema puedan aplicar modelos de análisis de sentimientos orientados a construir comunidades virtuales más inclusivas, seguras y respetuosas.

Origen y Características de los Datos

Un elemento fundamental en los estudios sobre el análisis de sentimiento en videojuegos ha sido la revisión de las fuentes de datos utilizadas, que provienen de diversos contextos, principalmente de los chats dentro de los videojuegos y plataformas de terceros como Twitch, donde los jugadores y espectadores se comunican a través de mensajes. Tales interacciones en tiempo real ofrecen una visión directa de las emociones y opiniones que emergen durante la experiencia de juego, y que algunas APIs de las plataformas de juego facilitan el acceso a grandes volúmenes de datos sobre las interacciones de los jugadores, lo que permite realizar un análisis masivo de comportamientos y sentimientos en estos entornos virtuales.

La calidad, representatividad y cantidad de los datos son elementos esenciales que influyen en la efectividad de estos estudios, por lo que las fuentes elegidas deben ser sólidas y variadas, capaces de capturar un amplio espectro de interacciones que reflejen con precisión la experiencia emocional de los jugadores. Asimismo, es vital que estas fuentes permitan identificar de manera confiable comportamientos disruptivos como la toxicidad y el acoso, los cuales son frecuentes en los entornos competitivos de juego.

Esta revisión resalta la importancia de emplear fuentes de datos adecuadas para el análisis de sentimientos en videojuegos, ya que los estudios analizados demuestran que la combinación de diversas herramientas y plataformas permite una comprensión más profunda de las dinámicas emocionales y sociales presentes en los juegos en línea, por lo que se seleccionaron 20 estudios iniciales centrados en el análisis de sentimientos en conversaciones de videojuegos y transmisiones en plataformas de streaming, entre otros, que aportan al entendimiento del comportamiento inapropiado.

Cada estudio revisado emplea diversas fuentes de datos, que abarcan una amplia gama de videojuegos, volúmenes de mensajes y características contextuales que complementan la naturaleza del mensaje en sí. Estas fuentes incluyen tanto interacciones en videojuegos como en plataformas de streaming relacionadas con videojuegos, además de comunidades de jugadores y otros entornos no necesariamente vinculados a los videojuegos, pero que aplican análisis de sentimientos para la detección de comportamientos inapropiados. Es relevante destacar que la mayoría de los datasets analizados están en inglés, idioma predominante en los estudios, seguido por una pequeña proporción de conjuntos de datos en idiomas no especificados, que podrían incluir inglés o ser multilingües.

La globalización de los videojuegos en línea implica que no se especifica una región particular en los estudios, dado que cualquier persona en el mundo puede acceder a estos entornos. El inglés predomina posiblemente por ser el idioma predeterminado en muchos desarrollos de videojuegos, uno de los más hablados a nivel mundial y por la fuerte presencia de jugadores en regiones angloparlantes. En el caso de las plataformas de streaming, Twitch como la más popular, facilita el acceso a grandes volúmenes de datos a través de su API, lo que permite la recolección y análisis masivo de mensajes; en contraste, los datos provenientes de los chats internos de videojuegos presentan mayores restricciones debido a que su acceso está controlado por los desarrolladores, lo que limita su disponibilidad a pequeñas porciones en plataformas como Kaggle, donde solo una fracción de estos datos es accesible para investigación, además varios estudios omiten especificar los videojuegos analizados por cuestiones de confidencialidad, lo que dificulta la identificación precisa de las fuentes utilizadas.

Preprocesamiento de Texto

El preprocesamiento es un paso crucial en la minería de texto, ya que transforma los datos en un formato adecuado para su análisis y organiza la información, facilitando la exploración de las relaciones textuales y permitiendo abordar el contenido no estructurado, lo que resulta esencial para preparar la información para su posterior análisis de sentimiento. Dado que los mensajes de chat suelen contener este tipo de texto, implementar un preprocesamiento efectivo es vital, ya que busca convertir estos mensajes en un formato uniforme que los algoritmos de aprendizaje automático puedan interpretar fácilmente, de este modo, el preprocesamiento se convierte en un elemento clave para garantizar la calidad y precisión de los resultados en el análisis de datos textuales (Wirawan et al., 2023).

Características Distintivas de los Mensajes

Antes de limpiar los datos es fundamental explorar las características de los mensajes en el chat de videojuegos y plataformas de streaming, ya que estos mensajes presentan particularidades que reflejan la forma en que se comunican los jugadores, incluyendo el uso de jerga, abreviaciones, errores ortográficos y emojis. Comprender estas características ayuda a identificar patrones y problemas, lo que hace que la limpieza de datos sea más efectiva y asegura que el análisis posterior sea relevante para la investigación.

En primer lugar, la brevedad es una característica común en los mensajes de chat, donde los jugadores tienden a enviar textos cortos y concisos debido a la presión de tiempo que implica la dinámica del juego. Esto se observa en estudios como los de Yildiz (2022) y Thompson et al. (2017), los cuales analizan mensajes generalmente breves que suelen incluir errores ortográficos y abreviaciones, lo que puede dificultar la comprensión del mensaje completo, especialmente para quienes no están familiarizados con la jerga utilizada en el juego.

La jerga y el uso de terminología especializada son características distintivas en la comunicación entre jugadores, quienes emplean un vocabulario que incluye acrónimos y términos con significados específicos dentro del contexto del juego, como "gg" (good game), "kill", "beam" y "destroy" (Yildiz, 2022). Esta terminología no solo refleja la cultura inherente a los videojuegos, sino que también puede representar una barrera para los jugadores novatos o aquellos que no forman parte de la comunidad, dado que muchas de estas expresiones carecen de un significado claro fuera del ámbito de los videojuegos.

Un aspecto importante para considerar en el análisis de comportamientos inapropiados, como el acoso, es que los mensajes en el chat de los videojuegos a menudo contienen lenguaje que puede asemejarse al utilizado por depredadores, incluyendo conversaciones sobre relaciones, citas y juegos de roles familiares como la frase "Pretend I am your dad/mom/sister/brother", lo que puede generar un alto número de falsos positivos al detectar comportamientos depredadores (Cheong et al., 2015). La complejidad del lenguaje utilizado en estos entornos dificulta la identificación de conductas inapropiadas, lo que resalta la necesidad de entender la dinámica específica de cada videojuego.

Los emotes y emojis son elementos esenciales en la comunicación de los chats de videojuegos y plataformas de streaming como Twitch, donde los emotes permiten expresar emociones de manera rápida y efectiva, además de que estas representaciones gráficas transmiten sentimientos u opiniones que suelen tener un significado particular dentro de la comunidad, aunque pueden no ser tan reconocidos como los emojis (Chouhan et al., 2021). En cambio, en el videojuego League of Legends, en lugar de usar emojis, los jugadores emplean combinaciones de letras para expresar emociones, como "BRB" (be right back) indica que un jugador se ausentará brevemente(Ángeles-Gómez & Quintana-López, 2019). La capacidad de estos

símbolos para encapsular emociones refleja las tendencias culturales dentro de la comunidad de jugadores y destaca su relevancia en las interacciones sociales y emocionales en línea.

Además de los emojis y emotes, las interacciones en el chat pueden incluir símbolos especiales y caracteres que alteran la expresión del mensaje, como *, &, $, %, -, _, y ><, los cuales, en ciertas combinaciones, pueden transmitir emociones, mientras que en otros contextos pueden no tener significado, lo que genera confusiones durante el análisis posterior y complica la interpretación de la intención del mensaje (Atoum, 2020). Por ello, es fundamental analizar si los caracteres expresan alguna emoción o no, para llevar a cabo su limpieza adecuada.

En resumen, estudiar los mensajes en el chat de videojuegos y plataformas de streaming muestra la complejidad de cómo se comunican los jugadores, por lo que es importante entender las características únicas de esta comunicación, como la brevedad de los mensajes, el uso de jerga, los emojis y los caracteres especiales, para llevar a cabo un análisis efectivo y significativo. A medida que el ámbito de los videojuegos y las plataformas de streaming sigue cambiando, es fundamental desarrollar métodos de análisis que tengan en cuenta estas particularidades, ya que facilitará la identificación de comportamientos y contribuirá a crear análisis de sentimiento más adaptables.

Técnicas de Preprocesamiento de Texto

Después de identificar las características distintivas de los mensajes, se implementan técnicas para preparar los datos para análisis posteriores con el objetivo de limpiar los mensajes de factores de ruido que puedan sesgar los resultados y garantizar la obtención de resultados precisos. Los estudios revisados indican que, para realizar análisis de sentimiento, es fundamental llevar a cabo un preprocesamiento que asegure la calidad de los mensajes y elimine posibles errores, aunque algunos estudios no especifican exactamente qué técnicas utilizaron, es posible inferir su elección a partir del contexto en el que se desarrollaron. A continuación, se detallan los métodos de preprocesamiento empleados en cada estudio.

Tabla 1

Técnicas de Preprocesamiento Implementadas

Estudio	Método de preprocesamiento utilizado
(Ángeles-Gómez & Quintana-López, 2019)	Tokenización, lematización, eliminación de stop words, eliminar palabras con menos de tres caracteres
(Murnion et al., 2018)	Tokenización, normalización a minúsculas, eliminación de URLs
(Cheong et al., 2015)	Eliminación de stop words, lematización, corrección ortográfica basada en diccionarios
(Borj et al., 2020)	Tokenización, normalización de caracteres, remoción de caracteres especiales, eliminación de stop words
(Faraz et al., 2022)	Corrección ortográfica basada en diccionarios, tokenización, eliminación de stop words
(Balci & Salah, 2015)	Tokenización, remoción de caracteres especiales
(Faraz, 2022)	Tokenización, lematización, eliminación de stop words, remoción de URLs y caracteres especiales, corrección ortográfica basada en diccionarios

(Thompson et al., 2017)	Stemming, tokenización, diccionario de emoticones, y corrección ortográfica basada en otros diccionarios
(Kumar, 2022)	Eliminación de stop words, tokenización, normalización a minúsculas, corrección ortográfica basada en diccionarios:
(Yildiz, 2022)	Tokenización, entre otros no especificados
(Chouhan et al., 2021)	Tokenización, eliminación de stop words, lematización
(Kobs et al., 2020)	Remoción de caracteres especiales, tokenización, Corrección ortográfica basada en diccionarios, normalización a minúsculas
(Stoop et al., 2019)	Tokenización, eliminación de stop words, normalización
(Williams, 2023)	Detección de idioma, anonimización de usuarios
(Ekiciler et al., 2022)	Stemming, normalización a minúsculas, tokenización
(Ghosh, 2021)	Tokenización, remoción de caracteres especiales, eliminación de stop words
(Atoum, 2020)	Tokenización, stemming, eliminación de stop words, eliminar espacios extras
(Dreier et al, 2023)	Tokenización, normalización a minúsculas
(Kwak & Blackburn, 2015)	Lematización, eliminación de stop words, uni-gramas y bi-gramas
(Neto et al, 2017)	Tokenización, remoción de URLs, lematización, eliminar textos vacios

Nota. Detalle de las técnicas utilizadas para la preparación de datos en cada estudio.

En el análisis de sentimientos y comportamiento en los chats de videojuegos, se utilizan diversas técnicas de preprocesamiento de texto que optimizan la calidad de los datos analizados. A continuación, se organizan los métodos más utilizados.

Tokenización

La tokenización es el proceso mediante el cual un texto se segmenta en sus distintos elementos significativos, eliminando espacios en blanco y saltos de línea, donde un token puede ser una palabra o un signo de puntuación que tiene relevancia dentro del contexto del texto, por ejemplo, en la frase “El juego es bueno”, se pueden identificar tres tokens: {'El', ‘juego’, ‘es’, ‘bueno’} (Talamé et al., 2019). Esta técnica resulta fundamental en la mayoría de los estudios revisados, ya que los chats de videojuegos contienen una gran cantidad de expresiones informales, emoticonos y jerga específica, y al tokenizar, se facilita que los análisis se concentren en los aspectos más relevantes del lenguaje utilizado por los jugadores.

Lematización

La lematización es un proceso lingüístico que consiste en identificar la forma base o lema de una palabra a partir de sus variaciones flexionadas, por ejemplo, términos como "victorias", "victorioso" y "victoriosos" se reducirían a "victoria" (Obando-Roldán et al., 2020). En los chats de videojuegos, la lematización es fundamental porque reduce la variabilidad del lenguaje de los jugadores, facilitando el análisis de contenido.

Stemming

La lematización, similar al stemming, presenta diferencias clave en su enfoque, mientras que la lematización transforma palabras a su forma base o lema, el stemming reduce diversas palabras a una raíz léxica común, por ejemplo, las palabras "jugar", "jugando" y "jugador" se pueden simplificar a la raíz "jug" (Ordoñez-Eraso & Cobos-Lozada, 2011). Esta técnica es especialmente relevante en los chats de videojuegos, ya que agrupa distintas formas de una palabra en un único término, facilitando el análisis del lenguaje de los jugadores y la

identificación de patrones de comportamiento o sentimientos, especialmente en entornos con variaciones de jerga.

Eliminación de Stop Words

La eliminación de Stop Words se refiere al proceso de filtrar palabras que carecen de significado relevante para el análisis, como artículos, pronombres y preposiciones, por ejemplo, palabras como "un", "desde" y "a", no aportan información significativa al clasificador (Elgueta-Morales et al., 2017). En los chats de videojuegos, esta técnica es esencial, ya que permite enfocar el análisis en términos significativos, ayudando a identificar patrones en la comunicación de los jugadores. Por ejemplo, en lugar de considerar "yo estoy jugando a un juego", se podría centrarse en "jugando" y "juego", facilitando la comprensión de su comportamiento y preferencias.

Normalización

La normalización es un proceso que busca estandarizar los mensajes, como convertir todo el texto a minúsculas o eliminar caracteres especiales y URLs. Este procedimiento es particularmente útil en los chats de videojuegos, donde los jugadores suelen emplear diversos estilos de escritura, incluyendo mayúsculas y caracteres especiales, lo que puede complicar el análisis, y que, al normalizar el texto, se facilitan las interpretaciones y se obtienen resultados más claros.

Corrección Ortográfica

La corrección ortográfica basada en diccionarios es un método que utiliza una lista de palabras correctas para detectar y corregir errores de escritura, por ejemplo, si un jugador intenta escribir "vida" pero lo escribe como "bida", el corrector podrá identificar el error y sugerir la palabra correcta (Cravero, Audano, & De Croce, 2012). Este enfoque es especialmente relevante

en los chats de videojuegos, donde una escritura errónea puede dificultar la comunicación y la coordinación entre los jugadores.

En resumen, las técnicas de preprocesamiento de texto son fundamentales para la organización y preparación de datos en el análisis de sentimientos en los chats de videojuegos, ya que permiten limpiar los mensajes de elementos que podrían sesgar los resultados. Entre las técnicas más utilizadas se destacan la tokenización, que segmenta el texto en unidades significativas; la lematización y el stemming, que reducen las palabras a sus formas base; la eliminación de stop words, que filtra palabras irrelevantes para el análisis; la normalización, que busca estandarizar el texto mediante la conversión a un formato uniforme; y la corrección ortográfica, que corrige errores de escritura para mejorar la calidad de los mensajes. Aunque existen otras técnicas que podrían ser consideradas, las mencionadas anteriormente son las más comunes y efectivas en el contexto analizado.

Métodos de Análisis de Sentimientos

El análisis de sentimientos se enfoca en identificar la polaridad de las opiniones y emociones de un grupo sobre un tema específico, incluyendo el resumen de opiniones y la extracción de emociones (Saberi & Saad, 2017). En el ámbito de los videojuegos en línea, esta técnica resulta particularmente significativa debido a la naturaleza dinámica de las interacciones que tienen lugar en los chats de juego, donde las interacciones pueden incluir desde comentarios positivos que contribuyen a un ambiente acogedor, hasta conductas inapropiadas que pueden perjudicar la experiencia de los jugadores. Con el aumento en el uso de plataformas de juego en línea y la interacción entre los usuarios, se vuelve esencial monitorear y analizar estos textos para fomentar un entorno de juego saludable y respetuoso.

A continuación, se presenta una tabla que resume los modelos de aprendizaje automático utilizados en investigaciones previas para el análisis de sentimiento, junto con sus respectivos resultados en términos accuracy, recall y F1-Score. Los resultados han sido recopilados de los estudios que reportaron estas métricas, proporcionando una visión comparativa del rendimiento de cada modelo en el contexto del análisis de sentimiento en videojuegos en línea.

Tabla 2

Desempeño de Modelos de Aprendizaje Automático en Estudios Previos

Fuente	Modelo	Accuracy	Recall	F1-Score
(Faraz, 2022)	XGBoost	0,98	0,98	0,98
	K-Nearest Neighbors (KNN)	0,98	0,98	0,98
	Support Vector Machine (SVM)	0,99	0,99	0,99
	Random Forest	0,98	0,97	0,97
(Ángeles-Gómez & Quintana-López, 2019)	Semantria for Excel	0,63		
	C4.5 (J-48)	0,61	0,61	0,66

(Cheong et al., 2015)	Naïve Bayes (NB)	0,79	0,72	0,57
	Decision Tree (J48)	0,91	0,73	0,74
	Multilayer Perceptron (MLP)	0,93	0,70	0,78
	Logistic Regression (LR)	0,89	0,66	0,70
	K-Nearest Neighbors (IBk)	0,85	0,40	0,49
	Support Vector Machine (SVM)	0,92	0,68	0,77
(Balci & Salah, 2015)	Binomial Probabilistic Models (BPMs)		0,73	0,80
(Kumar, 2022)	Bing Lexicon	0,82	0,72	0,71
	Vanilla Bing Lexicon	0,77	0,71	0,71
(Chouhan et al., 2021)	Support Vector Classifier (SVC)	0,70	0,67	0,67
	Logistic Regression	0,69	0,66	0,66
	Decision Tree Classifier	0,67	0,65	0,65
	Random Forest Classifier	0,66	0,63	0,63
	Multinomial Naive Bayes	0,66	0,60	0,61
(Kobs et al., 2020)	Average-based lexicon approach	0,62	0,59	0,61
	Distribution-based lexicon approach	0,63	0,61	0,62
	Sentence Convolutional Neural Network (CNN)	0,64	0,61	0,63
(Williams, 2023)	VADER	0,82	0,50	0,27
	Support Vector Classifier (SVC)	0,93	0,54	0,64
	SVC + SMOTE	0,93	0,56	0,66
	Multilayer Perceptron (MLP)	0,92	0,52	0,61

Nota. Esta tabla muestra los resultados de accuracy, recall y F1-Score de varios modelos de aprendizaje automático utilizados en estudios previos sobre análisis de sentimientos y detección de toxicidad en chats de videojuegos. Los datos reflejan el rendimiento de cada modelo en tareas específicas de clasificación de comportamientos inapropiados y toxicidad en entornos de juego online.

La tabla presenta un comparativo del rendimiento de diversos modelos de aprendizaje automático en tareas de análisis de sentimiento, enfocándose en el F1-Score, que es una métrica crucial, ya que equilibra la precisión y el recall, lo que resulta fundamental en contextos donde es importante minimizar tanto los falsos positivos como los falsos negativos. Los modelos que destacan por su alto F1-Score provienen principalmente del estudio de Faraz (2022). En este estudio, el modelo Support Vector Machine (SVM) obtiene un F1-Score de 0,99, seguido de cerca por K-Nearest Neighbors (KNN) y XGBoost, que alcanzan un F1-Score de 0,98, lo que sugiere una buena capacidad para clasificar adecuadamente las emociones en los textos analizados.

En contraste, los demás modelos, con la excepción de los estudios de Cheong et al. (2015), Balci y Salah (2015) y Kumar (2022), presentan un F1-Score inferior a 0,7. Esta disminución en el rendimiento puede atribuirse a factores contextuales, como el tipo de videojuego, la plataforma de los chats, el idioma de los textos y la representatividad de los datos utilizados en cada investigación. Por lo tanto, es esencial considerar estos elementos al evaluar la eficacia de los modelos en el análisis de sentimientos en entornos de juegos en línea.

Modelos Basados en Léxicos

Los estudios analizados sobre análisis de sentimientos en videojuegos han empleado diversos modelos léxicos para la detección de comportamientos inapropiados en chats y transmisiones en vivo, con dos enfoques principales se destacan: los métodos basados en léxicos simples y los métodos basados en la distribución.

Dentro de los enfoques simples, se encuentran métodos como el Average-based lexicon approach, que asigna una puntuación de sentimiento promedio a cada palabra en una oración y luego las combina para obtener una puntuación general del sentimiento de la oración. El Bing

Lexicon, otro método simple, utiliza un diccionario predefinido de palabras con polaridad de sentimiento (positiva o negativa) para clasificar el texto. El Vanilla Bing Lexicon es una versión básica de este enfoque, que no considera factores adicionales como la intensidad o el contexto.

Por otro lado, los métodos basados en la distribución, como el Distribution-based lexicon approach, consideran la distribución de las puntuaciones de sentimiento asignadas a una palabra por varios anotadores, el cual captura la ambigüedad de ciertas palabras que pueden tener connotaciones positivas o negativas según el contexto. VADER (Valence Aware Dictionary for Sentiment Reasoning) es un modelo distribucional que incorpora reglas lingüísticas y ponderaciones para analizar el sentimiento del texto, incluyendo emoticones, jerga y mayúsculas. Finalmente, Semantria for Excel, una API del software Lexalytics, permite la creación de diccionarios personalizados con pesos informativos para cada palabra, además de ofrecer opciones para ajustar el umbral de clasificación. Las clases de clasificación comunes en estos estudios son positiva, negativa y neutral, aunque algunos modelos incluyen clases adicionales como "muy positivo" o "muy negativo".

Desempeño

El rendimiento de los modelos léxicos se evalúa utilizando métricas como la accuracy, el recall y F1-score, para esta revisión se centra en el F1-score, ya que este evalúa el equilibrio entre los resultados positivos y negativos, lo cual es crucial para la tarea en cuestión.

En el estudio de Kobs et al. (2020), el Average-based lexicon approach, que emplea Bing Lexicon, el léxico de emojis y el de emotes, alcanzó un F1-score de 61%. Por su parte, el Distribution-based lexicon approach, al incorporar la información distribuida de las palabras, logró un leve aumento en el rendimiento. En otros estudios, como el de Ángeles Gómez &

Quintana López (2019), Semantria for Excel obtuvo un accuracy del 63% al clasificar conversaciones de videojuegos en categorías positiva, negativa y neutral.

En el estudio de Williams (2023), el modelo VADER presentó el rendimiento más bajo entre los evaluados, con un F1-score de 27%, el cual se atribuye a que VADER no fue diseñado específicamente para identificar hostilidad, sino para analizar el sentimiento general de un texto. En contraste, el estudio de Kumar (2022) reportó resultados significativamente mejores utilizando modelos léxicos, en el que se empleó Bing Lexicon y Vanilla Bing Lexicon, alcanzando un F1-score del 71% para ambos modelos y una precisión del 82% con Bing Lexicon, lo cual se atribuye a la actualización del diccionario léxico con términos específicos del videojuego Dota 2 , lo que mejoró la precisión, recall y F1-score, destacando la importancia de incluir términos del juego para detectar comportamientos inapropiados en videojuegos.

Ventajas y Limitaciones

Los modelos léxicos tienen ventajas significativas en la detección de comportamientos inapropiados en videojuegos, principalmente por su simplicidad y facilidad de implementación, lo que los hacen una opción atractiva para la moderación de contenido, permitiendo además la personalización de los diccionarios léxicos para adaptarse a la jerga específica de cada juego y a las diversas formas en que los jugadores expresan comportamientos negativos.

No obstante, también presentan limitaciones importantes, una de ellas es su dificultad para detectar ironía, sarcasmo y otros tipos de lenguaje figurado, comunes en la comunicación en línea, además, los modelos léxicos básicos tienden a interpretar las palabras de manera literal, lo que puede llevar a clasificaciones erróneas del sentimiento (Kumar, 2022). Otra limitación es la necesidad de actualizar los diccionarios constantemente para incluir nuevas palabras, expresiones y formas de evasión de censura, debido a que los jugadores suelen ser creativos en el uso del

lenguaje para eludir los filtros, lo que requiere un esfuerzo continuo para mantener la eficacia de los modelos léxicos. Otra limitación de estos modelos es su enfoque general en el análisis de sentimiento, lo que no los hace adecuados para la detección específica de hostilidad en el contexto de los videojuegos, como se evidenció con el modelo VADER, que mostró un bajo desempeño al etiquetar mensajes como hostiles (Williams, 2023).

Modelos Probabilísticos

Los modelos probabilísticos, como Naïve Bayes, Multinomial Naïve Bayes y los Binomial Probabilistic Models (BPMs), son herramientas ampliamente utilizadas para el análisis de sentimientos en videojuegos, especialmente en chats de transmisiones en vivo y mensajes dentro del juego. Basados en el teorema de Bayes, estos modelos estiman la probabilidad de que un mensaje pertenezca a una categoría específica analizando cada palabra de forma independiente; la variante Multinomial Naïve Bayes incorpora la frecuencia de cada palabra, lo cual es ideal para el análisis de sentimientos donde el conteo de términos es relevante; y por su parte, los BPMs se enfocan en clasificaciones binarias y calculan la probabilidad de categorías específicas según la presencia de características clave, siendo útiles para identificar patrones de comportamiento.

Estos modelos funcionan asignando probabilidades a diferentes clases de salida (como positivo, negativo, y neutral) basadas en la frecuencia y combinación de palabras o características observadas. Su popularidad se debe a su eficiencia en manejar grandes volúmenes de datos textuales, algo esencial en plataformas con alto tráfico de usuarios, como Twitch y MovieStarPlanet.

En el caso de Twitch, el estudio analizado utilizó el modelo Multinomial Naïve Bayes para clasificar los mensajes como positivos, negativos o neutrales, entrenándolo mediante la

frecuencia de palabras en un enfoque de Bag of Words y la presencia de emoticones específicos de la plataforma, elementos importantes para captar la expresividad particular de estos contextos (Chouhan et al., 2021). En contraste, en MovieStarPlanet, se utilizó el modelo Naïve Bayes estándar para detectar comportamientos predatorios en los chats, el cual se entrenó con la frecuencia de palabras, y elementos conductuales diseñados para identificar intentos de evadir las normas del juego (Cheong et al., 2015).

Desempeño

El desempeño de los modelos probabilísticos, en particular Naïve Bayes en sus variantes Multinomial y estándar, ha mostrado resultados prometedores, aunque con ciertas limitaciones en la detección de comportamientos inapropiados en videojuegos en línea. En el estudio de Chouhan et al. (2021) sobre análisis de sentimientos en Twitch, el modelo Multinomial Naïve Bayes alcanzó un F1-score de 61%. Aunque estos resultados no superaron los de otros modelos, es importante señalar que el conjunto de datos estaba desbalanceado entre las clases, lo cual pudo afectar negativamente el rendimiento.

Por otro lado, en el estudio de Cheong et al. (2015) sobre detección de depredadores en MovieStarPlanet, el modelo Naïve Bayes, combinado con características como listas negras de palabras y el enfoque Bag of Words, alcanzó un F1-score de 57%, lo que indica desafíos para equilibrar la precisión y la recuperación en la detección de comportamientos predatorios.

En el estudio de Balci y Salah (2015), aunque no se registra directamente un F1-score, se estima que el modelo alcanzó un 80%, el mejor desempeño entre los estudios revisados. Esto se atribuye a su enfoque en el juego específico Okey y al uso de características adicionales como información demográfica, avatar y número de amigos, que ayudaron a mejorar la precisión en la detección de patrones de comportamiento.

Ventajas y Limitaciones

Una de las principales ventajas de los modelos Naïve Bayes es su simplicidad y eficiencia computacional; estas características los hacen ideales para analizar grandes volúmenes de datos generados en entornos de alto tráfico, como los chats de videojuegos y las transmisiones en vivo (Chouhan et al., 2021); además, estos modelos son fáciles de implementar, lo que facilita su uso en la moderación de contenido en plataformas de videojuegos. Otra ventaja significativa es que pueden ofrecer un rendimiento aceptable incluso con conjuntos de datos de entrenamiento limitados, algo útil en contextos donde obtener datos etiquetados puede ser costoso y laborioso (Balci & Salah, 2015)

Sin embargo, estos modelos también presentan limitaciones importantes; la principal es su suposición de independencia entre las características, una premisa que no siempre se cumple en el lenguaje natural, ya que las palabras y expresiones suelen estar relacionadas, esto puede afectar la precisión del modelo en tareas de análisis de texto (Cheong et al., 2015). Otra limitación es su sensibilidad al desbalance de clases en los datos de entrada, lo cual puede disminuir su rendimiento en escenarios donde una categoría tiene significativamente más muestras que otra, requiriendo estrategias adicionales de manejo de datos para mejorar su precisión (Chouhan et al., 2021).

Modelos de Aprendizaje Automático Supervisado Tradicionales

Los modelos de aprendizaje automático supervisado tradicional son ampliamente utilizados en la detección de comportamiento inadecuado en videojuegos, los cuales contemplan los modelos derivados de árboles de decisión, vecinos más cercanos, regresión logística y máquinas de soporte vectorial.

Los arboles de decisión son modelos utilizados para clasificar datos al dividirlos en subconjuntos basados en características clave mediante nodos y ramas, donde cada nodo representa una característica y cada rama, una decisión basada en el valor de esa característica. En el análisis de sentimientos en videojuegos, los árboles de decisión pueden ser útiles para identificar palabras clave o frases que indiquen un comportamiento inadecuado

Los vecinos más cercanos son modelos para clasificar datos en función de su similitud con otros datos, donde calcula la distancia entre un nuevo dato y todos los datos del conjunto de entrenamiento, luego, selecciona los k vecinos más cercanos y asigna al nuevo dato la clase mayoritaria entre esos vecinos. Este modelo es útil en el contexto porque identifica patrones de comportamiento en los mensajes basados en la similitud con comportamientos previamente clasificados.

La regresión logística, son modelos utilizados para predecir la probabilidad de que un dato pertenezca a una clase específica, utiliza una función sigmoide para transformar una combinación lineal de variables predictoras en una probabilidad entre 0 y 1. En la detección de comportamiento inadecuado, la regresión logística puede estimar la probabilidad de que un mensaje sea ofensivo o inapropiado.

Los modelos de máquinas de soporte vectorial se usan para clasificar datos al encontrar el mejor hiperplano que separa las diferentes clases, a su vez buscan el hiperplano que maximiza el margen entre las clases, lo que mejora su capacidad de generalización. En el análisis de sentimientos, las SVM pueden ser efectivas para clasificar mensajes de chat como positivos, negativos o neutrales al encontrar el hiperplano que mejor separa estas categorías.

Desempeño

El análisis de desempeño de los modelos de aprendizaje automático tradicionales en la detección de comportamientos inapropiados en videojuegos en línea revela resultados prometedores. Estos modelos destacan por su capacidad de manejar datos complejos, aunque su efectividad varía según el tipo de modelo, el conjunto de datos y las técnicas de preprocesamiento aplicadas.

Entre los modelos derivados de árboles de decisión, el Random Forest destacó con un desempeño sobresaliente, como lo demuestra el estudio de Faraz et al. (2022), donde alcanzó un F1-score del 97%, atribuible a su capacidad para reducir el sobreajuste mediante la combinación de predicciones de múltiples árboles; en contraste, el estudio de Cheong et al. (2015) reportó un F1-score de 74% al emplear árboles de decisión simples, evidenciando un desbalance entre las clases, ya que el modelo presentó un accuracy del 91%, lo cual sugiere que clasifica con mayor precisión la clase mayoritaria, pero enfrenta dificultades con la clase minoritaria. Otros estudios también informaron resultados más modestos, con F1-scores que oscilaron entre el 63% y el 66%, lo que resalta la variabilidad en el rendimiento de estos modelos según el contexto y las características de los datos empleados.

El modelo KNN mostró un desempeño sobresaliente en el estudio de Faraz et al. (2022), donde alcanzó un F1-score del 98%, un resultado que se atribuye a su eficacia para manejar datos de alta dimensionalidad, como los embeddings de palabras fastText, y al uso del conjunto de datos PAN12, reconocido por su calidad y balance; en contraste, el estudio de Cheong et al. (2015) reportó un F1-score significativamente menor, de solo el 49%, al utilizar IBk, una variante de KNN, y esta diferencia de desempeño podría explicarse por el desbalance de clases y

las características específicas del conjunto de datos de MovieStarPlanet, además de posibles decisiones subóptimas en la configuración de parámetros clave como el número de vecinos (k).

La regresión logística mostró un desempeño moderado, con un F1-score del 70% reportado en el estudio de Cheong et al. (2015) y del 66% en el de Chouhan et al. (2021), aunque el desempeño en accuracy presentó una diferencia notable, alcanzando un 89% en el caso del primero y solo un 69% para el segundo. Lo cual sugiere posibles disparidades en el balance de clases en los conjuntos de datos utilizados; la simplicidad inherente de este modelo, junto con la complejidad y la no linealidad del lenguaje asociado a la detección de depredadores, podría ser un factor que explique estas variaciones en los resultados.

Los modelos de máquinas de soporte vectorial (SVM) tuvieron un desempeño variable, con F1-scores que fluctuaron entre el 64% y el 77%, destacándose en el estudio de Cheong et al. (2015), donde uno de estos modelos alcanzó un F1-score del 77% y un accuracy del 92%, aunque el desbalance de clases impactó negativamente la precisión en la identificación de la clase minoritaria. En el caso del estudio de Williams (2023), se evaluaron dos SVM y se mejoró el rendimiento de uno mediante la aplicación de SMOTE, una técnica que genera muestras sintéticas para equilibrar las clases, logrando incrementar el F1-score de 64% a 66%, y aunque este avance parezca limitado, resulta crucial en el ámbito de la detección de comportamientos no deseados, pues refuerza la capacidad del modelo para reconocer correctamente casos minoritarios, lo que contribuye directamente a la protección y seguridad de los usuarios.

En conjunto, los modelos tradicionales ofrecen resultados prometedores para la detección de comportamientos inapropiados en videojuegos, aunque su desempeño depende en gran medida de la calidad y el balance de los datos, así como de las técnicas adicionales empleadas para optimizar su rendimiento.

Ventajas y Limitaciones

Los modelos de aprendizaje automático supervisado tradicional, como árboles de decisión (C4.5, DT), random forest, vecinos más cercanos (KNN), regresión logística y máquinas de soporte vectorial, presentan tanto ventajas como limitaciones en su aplicación para la moderación de contenido en videojuegos. Los árboles de decisión, destacan por su estructura interpretable y comprensible, lo que facilita la identificación de características clave para clasificar contenido inapropiado, además, estos modelos tienen la capacidad de manejar datos con ruido y son relativamente eficientes durante el proceso de entrenamiento, aunque su principal limitación radica en su susceptibilidad al sobreajuste, un problema que se intensifica en conjuntos de datos complejos y de alta dimensionalidad (Ángeles-Gómez & Quintana-López, 2019). Para abordar esta limitación, los modelos de random forest ofrecen una solución efectiva al combinar múltiples árboles de decisión y promediar sus predicciones, lo que mejora la robustez y la capacidad de generalización del modelo, haciéndolos más adecuados para tareas que involucran datos complejos y heterogéneos.

Los modelos KNN se destacan por su simplicidad en la implementación y su eficacia al trabajar con datos de alta dimensionalidad, como los embeddings de palabras, lo que los convierte en una opción adecuada para la moderación de contenido en plataformas de streaming con un alto volumen de mensajes de (Cheong et al., 2015). Su robustez y capacidad para manejar grandes conjuntos de datos fortalecen su utilidad en estos contextos, aunque su desempeño está condicionado por la correcta elección del número de vecinos (k) y puede verse comprometido cuando el conjunto de datos presenta un desbalance significativo entre clases (Chouhan et al. , 2021). Por otro lado, la regresión logística, conocida por su naturaleza lineal, proporciona una interpretación clara de los resultados y destaca por su eficiencia en el proceso de entrenamiento,

pero su simplicidad puede ser una limitación al enfrentar relaciones complejas y no lineales, como las que suelen presentarse entre el lenguaje y los comportamientos inapropiados.

Las máquinas de soporte vectorial (SVM) son efectivas para manejar datos de alta dimensionalidad y detectar patrones complejos, lo que las hace útiles para identificar comportamientos sutiles en el lenguaje, y son ideales en contextos donde se requiere precisión y robustez, especialmente en datos con ruido o características intrincadas (Williams, 2023). Sin embargo, su entrenamiento puede ser computacionalmente costoso, limitando su uso en escenarios con recursos restringidos, además, su rendimiento puede verse afectado por el desbalance de clases, común en problemas como la detección de comportamientos inapropiados, donde los casos positivos son menos frecuentes. Técnicas como SMOTE pueden ayudar a mitigar este problema al generar muestras sintéticas de la clase minoritaria, mejorando la capacidad del modelo para identificar correctamente los casos positivos y, en consecuencia, su rendimiento general (Williams, 2023).

Modelos Avanzados

En los estudios previos se emplearon tres modelos avanzados para detectar comportamientos inadecuados en chats y transmisiones de videojuegos: XGBoost, Multilayer Perceptron (MLP) y Sentence Convolutional Neural Network (CNN). Estos modelos han demostrado ser efectivos en tareas de análisis de sentimientos y moderación de contenido, gracias a su capacidad para manejar datos complejos y variados.

El modelo XGBoost es un algoritmo de aprendizaje automático que emplea el método de aumento de gradiente para generar predicciones al combinar múltiples árboles de decisión, siendo útil tanto en problemas de clasificación como de regresión. Este modelo construye los

árboles de manera secuencial, permitiendo que cada uno ajuste los errores cometidos por el árbol anterior mediante un proceso iterativo que busca alcanzar un nivel óptimo de predicción.

El Multilayer Perceptron (MLP) es una red neuronal artificial compuesta por varias capas de neuronas interconectadas, en la que cada neurona recibe una entrada, la procesa a través de una función matemática y luego transmite el resultado a la siguiente capa. Este modelo se entrena mediante un enfoque supervisado, lo que implica ajustar los pesos de las conexiones entre las neuronas para minimizar los errores en las predicciones realizadas durante el entrenamiento.

El modelo Sentence Convolutional Neural Network (CNN) es una red neuronal diseñada para procesar datos de texto mediante capas de convolución. En el análisis de sentimientos, la CNN convierte las oraciones en vectores numéricos utilizando técnicas como word embeddings, luego estos vectores pasan por las capas de convolución para extraer características importantes, y al final, una capa softmax clasifica el sentimiento en categorías como positivo, negativo o neutral.

Desempeño

El modelo XGBoost de Faraz (2022) alcanzó un f1-score de 98%, destacándose por su capacidad para manejar datos desbalanceados y su robustez en la clasificación de textos, y gracias a las técnicas de optimización y regularización implementadas, el modelo mostró un buen rendimiento, además, el balance entre clases en la base de datos ayudó al modelo a aprender tanto de ejemplos positivos como negativos, sin restarle mérito a su avanzada arquitectura y su habilidad para capturar interacciones complejas entre características.

Los estudios de Cheong et al. (2015) y Williams (2023) emplearon el modelo Multilayer Perceptron (MLP), obteniendo buenos resultados en términos de accuracy, sin embargo, el

estudio de Cheong et al. (2015) destacó con un f1-score de 78% frente al 61% de Williams (2023), debido a un mejor preprocesamiento de datos y selección de características, con la normalización de datos y eliminación de ruido, lo que permitió una mayor capacidad para identificar patrones significativos; y el estudio de Williams (2023), por otro lado, podría mejorar con más datos, un mejor balance entre clases y la optimización de hiperparámetros.

El modelo Sentence Convolutional Neural Network (CNN) implementado por Kobs et al. (2020) alcanzó un f1-score de 63%, destacando por su habilidad para capturar las características semánticas y contextuales de las oraciones en los chats de videojuegos, y su arquitectura convolucional facilitó la extracción de características locales de los textos, mejorando la identificación de patrones de toxicidad y comportamientos inapropiados, no obstante, su desempeño podría beneficiarse de un mayor volumen de datos y un enfoque más detallado en el preprocesamiento del texto.

Ventajas y Limitaciones

XGBoost tiene varias ventajas, como su eficiencia en el manejo de datos, incluyendo aquellos con valores faltantes, lo que le permite clasificar de manera efectiva, además, utiliza regularización para penalizar modelos complejos y evitar el sobreajuste, y su técnica de ensemble de múltiples árboles de decisión mejora el rendimiento en comparación con modelos individuales. Sin embargo, también presenta limitaciones, como la complejidad del modelo, ya que la ensemblación de varios árboles puede dificultar la interpretación del mismo, además, los árboles no pueden ser entrenados en paralelo, lo que aumenta los costos de almacenamiento al requerir un orden específico de los datos, y por último, la necesidad de definir la profundidad máxima del árbol antes del entrenamiento puede reducir la flexibilidad del modelo (Faraz, 2022).

El modelo Multilayer Perceptron (MLP) tiene la ventaja de ser capaz de identificar patrones complejos en los datos, como lo mostró el estudio de Cheong et al. (2015), que alcanzó una alto accuracy al detectar comportamientos inapropiados en los chats de juegos, además, el MLP puede mejorar su desempeño en datos desbalanceados mediante el uso de técnicas como SMOTE, tal como se evidenció en la investigación de Williams (2023). No obstante, presenta limitaciones, como la tendencia a priorizar la clase mayoritaria cuando hay pocos ejemplos de la clase minoritaria, lo que afecta tanto la precisión como el recall; y su efectividad podría verse reducida si solo se emplean características simples como "bag of words", por lo que sería más beneficioso integrar una variedad de características para mejorar su capacidad de detección.

El modelo Sentence Convolutional Neural Network (CNN) es especialmente efectivo para analizar texto con un lenguaje característico, como los comentarios en Twitch, ya que puede capturar tanto las características semánticas como contextuales, lo que lo hace adecuado para identificar comportamientos inapropiados. Según el estudio de Kobs et al. (2020), la CNN logró un F1-score aceptable en la detección de sentimientos y conductas en plataformas de streaming, especialmente al incorporar elementos como emoticones y jerga propia de la comunidad, sin embargo, una de sus principales limitaciones es su dificultad para adaptarse a vocabulario nuevo o expresiones no vistas durante el entrenamiento, lo que puede afectar su desempeño cuando los comentarios son más impredecibles. Además, la CNN requiere un entrenamiento intensivo y mucho espacio de almacenamiento, lo que puede ser un desafío en aplicaciones que necesitan tiempos de respuesta rápidos y alta eficiencia computacional.

Resumen de los Modelos de Análisis de Sentimiento

El estudio con los mejores resultados en términos de métricas de desempeño, como precisión, recall y F1-score, es el de Faraz (2022), quien utilizó el conjunto de datos Pan12, un

corpus público creado para la competencia PAN 2012 sobre la detección de depredadores online, que a diferencia de otros estudios, este conjunto incluye conversaciones tanto predatorias como no predatorias, lo que lo convierte en un recurso valioso para entrenar y evaluar modelos de detección de comportamientos inapropiados, y aunque no especifica la cantidad exacta por cada clase, este conjunto de datos ofrece un balance ligeramente mejor que el habitual en fuentes similares. Faraz también implementó un proceso de preprocesamiento similar al de otros estudios, como la eliminación de stop words y la aplicación de stemming, pero se destacó por su eficaz uso de un corrector ortográfico automático para mejorar la calidad del texto y la eficiencia de la representación "bag of words" (BoW). En cuanto al modelado, exploró una variedad de modelos de clasificación, como XGBoost, KNN, SVM y Random Forest, utilizando técnicas de validación cruzada y optimizando los hiperparámetros de cada modelo para lograr los mejores resultados.

Los estudios de Cheong et al. (2015), Kumar (2022) y Williams (2023) señalan un problema común de desbalance en los conjuntos de datos utilizados para entrenar sus modelos, lo cual se refleja en un alto accuracy pero un F1-score relativamente bajo. Este desbalance puede hacer que los modelos se inclinen hacia la clase mayoritaria, dificultando la detección de la clase minoritaria, que en este caso corresponde a las conductas inapropiadas. Un accuracy alto puede resultar engañoso en situaciones de desbalance, ya que el modelo podría simplemente clasificar la mayoría de las instancias como pertenecientes a la clase mayoritaria, y para abordar este desbalance, existen varias técnicas, como la implementación del algoritmo SMOTE (Synthetic Minority Over-sampling Technique), utilizado por Williams (2023) para generar muestras sintéticas de la clase minoritaria (hostil), aunque este método produjo solo una mejora ligera,

resalta la importancia de explorar otras estrategias de balanceo de datos para mejorar la precisión en la detección de conductas inapropiadas.

Los autores de los documentos destacan varios desafíos futuros en la detección de comportamientos inapropiados en videojuegos, señalando la necesidad de desarrollar modelos más robustos capaces de manejar la complejidad del lenguaje utilizado en estos entornos y minimizar tanto los falsos positivos como los falsos negativos, además de requerir estrategias más efectivas para abordar el desbalance de datos y asegurar que los modelos puedan detectar la clase minoritaria de manera confiable, también subrayan la importancia de que los modelos sean capaces de interpretar el lenguaje en su contexto social y emocional para diferenciar entre comportamientos inapropiados genuinos y el uso humorístico o irónico del lenguaje. Además, se enfatiza la necesidad de explorar otros idiomas más allá del inglés, ya que muchos modelos tienden a centrarse en este idioma, limitando su alcance; incorporar otros lenguajes podría enriquecer la experiencia y permitir que los modelos lleguen a más usuarios, mejorando así el ambiente en los videojuegos y promoviendo una moderación más inclusiva.

Conclusiones

El análisis de la literatura sobre el análisis de sentimientos en chats de videojuegos en línea y plataformas de streaming permitió identificar cómo estas herramientas son fundamentales para detectar y gestionar comportamientos inadecuados, como el ciberacoso y el lenguaje ofensivo, que afectan la experiencia del usuario y la salud mental, particularmente en audiencias jóvenes. Los estudios revisados destacan la creciente relevancia de métodos basados en modelos léxicos, probabilísticos, de aprendizaje automático y avanzados, evidenciando que la detección efectiva de conductas inapropiadas requiere modelos adaptados a las particularidades del lenguaje informal, breve y no estructurado de estos entornos virtuales.

El preprocesamiento de datos emerge como una etapa esencial en este campo, ya que la naturaleza breve y fragmentada de los mensajes, junto con el uso de jerga, emojis y caracteres no convencionales, dificulta la extracción de información útil. Técnicas como la tokenización, normalización, lematización y la implementación de correctores ortográficos automáticos han demostrado ser efectivas para mejorar la calidad de los datos y, en consecuencia, el rendimiento de los modelos. Sin un preprocesamiento exhaustivo, la aplicación de técnicas avanzadas de análisis de sentimientos, como las basadas en redes neuronales o modelos probabilísticos, se ve severamente limitada.

La literatura analizada evidenció ventajas y limitaciones en los enfoques utilizados; por un lado, los métodos basados en léxicos ofrecen simplicidad y transparencia, pero enfrentan dificultades para interpretar el sarcasmo, la ironía y el contexto emocional. Por otro lado, los modelos avanzados, como los basados en aprendizaje profundo, presentan un mejor desempeño en métricas como el F1-score, pero requieren grandes volúmenes de datos balanceados y un

procesamiento computacional elevado. En este sentido, la selección del modelo debe ser estratégica, considerando el idioma, las características del videojuego y la estructura de los datos.

Un desafío recurrente identificado es el desbalance de clases en los conjuntos de datos, donde las conductas inapropiadas están subrepresentadas frente a las interacciones normales. Este fenómeno afecta significativamente el rendimiento de los modelos, que tienden a priorizar la clase mayoritaria, reduciendo su capacidad para identificar conductas inapropiadas. Estrategias como el uso de SMOTE, submuestreo y ponderación de clases durante el entrenamiento han mostrado mejoras limitadas, subrayando la necesidad de enfoques más efectivos para abordar este problema y garantizar un rendimiento robusto en métricas como el recall y el F1-score.

Finalmente, se destaca la importancia de avanzar en la creación de conjuntos de datos más diversos y representativos, que incluyan una variedad de idiomas, plataformas y géneros de videojuegos, así como la necesidad de desarrollar modelos que no solo detecten comportamientos inapropiados, sino que también comprendan el contexto social y emocional del lenguaje. Esto requiere una colaboración multidisciplinaria entre investigadores, desarrolladores de videojuegos y plataformas de streaming, para garantizar entornos virtuales más inclusivos, seguros y respetuosos, fomentando interacciones más saludables entre los jugadores.

Recomendaciones

Para mejorar la efectividad de los modelos de análisis de sentimientos en la detección de conductas inapropiadas, es esencial adoptar estrategias que mitiguen el desbalance de clases en los datos de entrenamiento, con el uso de técnicas como SMOTE para el sobremuestreo de la clase minoritaria puede incrementar la representatividad sin comprometer la diversidad de los datos, mientras que el submuestreo controlado de la clase mayoritaria puede equilibrar las proporciones sin perder información clave. Asimismo, la implementación de métodos de ponderación de clases durante el entrenamiento permitiría al modelo priorizar correctamente la detección de conductas infrecuentes, mejorando su desempeño en métricas críticas como el recall y el F1-score.

Es importante impulsar la creación de conjuntos de datos más diversos y representativos que incluyan interacciones en una variedad de idiomas, géneros de videojuegos y plataformas, ya que no solo ampliará la aplicabilidad de los modelos, sino que también permitirá desarrollar soluciones más adaptables y robustas frente a las diferencias culturales y lingüísticas. La colaboración entre desarrolladores, investigadores y comunidades de jugadores será clave para construir bases de datos que reflejen la complejidad y riqueza de las interacciones en los videojuegos, facilitando así un progreso significativo hacia entornos virtuales más inclusivos y respetuosos.

Se recomienda avanzar en el desarrollo de modelos de análisis de sentimientos que no solo se centren en el significado literal de las palabras, sino que también puedan comprender el contexto social y emocional detrás de los mensajes, y para lograrlo, es esencial incorporar modelos avanzados capaces de detectar matices complejos como la ironía, el sarcasmo y el humor, comunes en la interacción en línea, especialmente en los videojuegos. Estos aspectos

lingüísticos son difíciles de interpretar para los modelos tradicionales, pero son cruciales para diferenciar entre comportamientos inapropiados y expresiones informales o humorísticas que no deben ser penalizadas, ya que al reducir los falsos positivos y evitar que comentarios humorísticos o irónicos sean erróneamente etiquetados como inapropiados, los sistemas de moderación pueden alcanzar una mayor precisión.

Implementar enfoques híbridos que combinen modelos basados en léxicos, probabilísticos y de aprendizaje automático permite aprovechar las fortalezas de cada uno, ya que, por ejemplo, los modelos basados en léxicos pueden identificar vocabulario ofensivo, mientras que los modelos de aprendizaje automático, como redes neuronales o transformadores, se encargarían de analizar el contexto y la intención detrás de los mensajes, de esta manera, la sinergia entre estos métodos fortalecería la precisión y robustez de los sistemas de detección, permitiendo abordar tanto patrones simples como complejos de comportamiento inapropiado en diferentes escenarios y comunidades.

Un aspecto clave que no debe pasarse por alto es garantizar la privacidad y el anonimato de los jugadores al recopilar y utilizar datos de comunicación en juegos, lo que implica la implementación de medidas de protección de datos para prevenir la identificación de usuarios individuales, además de obtener el consentimiento informado de los jugadores antes de recopilar sus datos, asegurando así que se respeten sus derechos y se mantenga la confianza en los sistemas de moderación.

Dado que el campo de estudio en moderación de contenido está en crecimiento y el esfuerzo en el ámbito de los videojuegos es limitado, se recomienda fomentar la colaboración entre investigadores, desarrolladores de videojuegos y plataformas de streaming para compartir datos, conocimientos y mejores prácticas. Esta colaboración sería clave para desarrollar

soluciones más efectivas para la moderación de contenido y la creación de entornos de juego más seguros y positivos para todos los usuarios, ya que un desafío crítico actualmente es el acceso limitado a las bases de datos, ya que estas son propiedad de los desarrolladores y están sujetas a restricciones de confidencialidad, lo que dificulta la investigación y el avance en este campo.

Referencias Bibliográficas

Ángeles-Gómez, D., & Quintana-López, M. (2019). Análisis de sentimientos en videojuegos. *ReCIBE. Revista electrónica de Computación, Informática, Biomédica y Electrónica, 8*(2).

Anti-Defamation League (2023). *Hate is No Game: Harassment and Positive Social Experiences in Online Games 2023*. https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2023

Balci, K., & Salah, A. (2015). Automatic analysis and identification of verbal aggression and abusive behaviors for online social games. *Computers in Human Behavior, 53,* 517-526. https://doi.org/10.1016/j.chb.2014.10.025

Borj, P. R., Raja, K., & Bours, P. (2020). On preprocessing the data for improving sexual predator detection: Anonymous for review. *In 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA)* (pp. 1-6). https://doi.org/10.1109/SMAP49528.2020.9248461

Cheong, Y.-G., Jensen, A. K., Guðnadóttir, E. R., Bae, B.-C., & Togelius, J. (2015). Detecting predatory behavior in game chats. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(3), 220-232. https://doi.org/10.1109/TCIAIG.2015.2424932

Chouhan, A., Halgekar, A., Rao, A., Khankhoje, D., & Narvekar, M. (2021). Sentiment Analysis of Twitch.tv Livestream Messages using Machine Learning Methods. *En 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-5). doi:10.1109/ICECCT52121.2021.9616932

Cravero, M., Audano, E., & De Croce, M. (2012). Analizador semántico: Una extensión para un corrector ortográfico en español. En XV Concurso de Trabajos Estudiantiles (EST 2012) (XLI JAIIO, La Plata, 27 al 31 de agosto de 2012).

Ekiciler, A., Ahioğlu, İ., Yıldırım, N., Ajas, İ. İ., & Kaya, T. (2022). The bullying game: Sexism based toxic language analysis on online games chat logs by text mining. *Journal of International Women's Studies, 24*(3), 1-16.

Elgueta-Morales, J., Vidal-Castro, C., & Segura-Navarrete, A. (2017). *Comparación de rendimiento de técnicas de aprendizaje automático para análisis de afecto sobre textos en español* [Tesis de Magíster, Universidad del Bío-Bío]. http://repobib.ubiobio.cl/jspui/handle/123456789/1772

Faraz, A. (2022). Protectbot: *A Chatbot to Protect Children on Gaming Platforms* [Master's Thesis, Rochester Institute of Technology]. Digital Institucional Repository. https://repository.rit.edu/theses/11392.

Faraz, A., Mounsef, J., Raza, A., & Willis, S. (2022). Child safety and protection in the online gaming ecosystem. *IEEE Access, 10*, 115895-115913. https://doi.org/10.1109/ACCESS.2022.3218415

Kumar Singh, A. (2022). *Sentiment Analysis of Dota 2 videogame chat in context of Cyber-bullying* [Master's Thesis, National College of Ireland]. Norma eResearch NCI Library. https://norma.ncirl.ie/id/eprint/6310

Liu, B. (2020). Introduction. In Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. *Cambridge University Press*, (pp. 1-17).

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review, 27*, 16-32. https://doi.org/10.1016/j.cosrev.2017.10.002

Murnion, S., Buchanan, W. J., Smales, A., & Russell, G. (2018). Machine learning and semantic analysis of in-game chat for cyberbullying. *Computers & Security, 76,* 197–213. https://doi.org/10.1016/j.cose.2018.02.016.

Obando-Roldán, J. C., Pulido-Díaz, J. A., & Gómez-Ávila, J. A. (2020). Procesamiento del lenguaje natural para reconocer mensajes de textos extorsivos a través del análisis sintáctico y lematización. *Revista Ciencia y Tecnología, 16*(1), 33-42.

Ordoñez-Eraso, H., & Cobos-Lozada, C. (2011). Stemming en español para documentos recuperados de la web. *Revista Unimar, 58,* 107–114.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., Moher, D., Yepes-Nuñez, J. J., Urrútia, G., Romero-García, M., & Alonso-Fernández, S. (2021). Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. *Revista Española de Cardiología, 74*(9), 790-799. https://doi.org/10.1016/j.recesp.2021.06.016

Rivera-Arteaga, E., & Torres-Cosío, V. (2018). Videojuegos y habilidades del pensamiento. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo, 8(16)*, 267-288. https://doi.org/10.23913/ride.v8i16.341

Saberi, B., & Saad, S. (2017). Sentiment analysis or opinion mining: A review. *International Journal of Advanced Science, Engineering and Information Technology*, *7*(5), 1660-1666.

Suler, J. (2004). The Online Disinhibition Effect. *CyberPsychology & Behavior, 7(3),* 321–326. https://doi.org/10.1089/1094931041291295

Talamé, L., Cardoso, A., & Amor, M. (2019). Comparación de herramientas de procesamiento de textos en español extraídos de una red social para Python. En *XX Simposio Argentino de Inteligencia Artificial (ASAI 2019)-JAIIO 48* (Salta).

Tavinor, G. (2008). Definition of Videogames. *Contemporary Aesthetics (Journal Archive), 6(Article 16)*

Thompson, J. J., Leung, B. H. M., Blair, M. R., & Taboada, M. (2017). Sentiment analysis of player chat messaging in the video game StarCraft 2: Extending a lexicon-based model. *Knowledge-Based Systems, 137,* 149-162. https://doi.org/10.1016/j.knosys.2017.09.022

Williams, S. (2023). *Detecting hostility in user-created content of mobile games using machine learning models* [Master's tesis, University of Helsinki]. HELDA. https://helda.helsinki.fi/server/api/core/bitstreams/a45a25a4-e8b0-4aac-918f-7729802c5a29/content

Wirawan, A., Cahyono, H. D., & Winarno. (2023). Easy data augmentation in sentiment analysis of cyberbullying. In *2023 6th International Conference on Information and Communications Technology (ICOIACT)* (pp. 443–447). https://doi.org/10.1109/ICOIACT59844.2023.10455817

Wohn, D., & Lee, Y.-H. (2013). Players of Facebook games and how they play. *Entertainment Computing, 4,* 171–178. https://doi.org/10.1016/j.entcom.2013.05.002

Woo, J., Park, S. H., & Kim, H. K. (2022). Profane or Not: Improving Korean Profane Detection using Deep Learning. *KSII Transactions on Internet and Information Systems, 16(1),* 305-318. https://doi.org/10.3837/tiis.2022.01.017

Yildiz, S. N. (2022). *Comparison of various methods of sentiment analysis: For the case of Twitch* [Master's Thesis, Tilburg University]. https://arno.uvt.nl/show.cgi?fid=161872