ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

MODELO DE CLASIFICACIÓN DE RASGOS DE ADICCIÓN A LOS VIDEOJUEGOS EN BASE A POSTS DE REDDIT MEDIANTE TÉCNICAS DE MINERÍA DE TEXTO Y APRENDIZAJE DE MÁQUINA

TESIS PREVIA A LA OBTENCIÓN DEL GRADO DE MAGISTER EN SISTEMAS DE INFORMACIÓN, MENCIÓN INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS MASIVOS

LEONARDO ANDRÉS BENÍTEZ ORELLANA

leonardo.benitez01@epn.edu.ec

DIRECTORA: Dra. Lorena Katherine Recalde Cerda

lorena.recalde@epn.edu.ec

CODIRECTOR: Dr. Edison Fernando Loza Aguirre

edison.loza@epn.edu.ec

Quito, junio 2023

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Leonardo Andrés Benítez Orellana, bajo mi supervisión.

Firmado electrónicamente por:
LORENA KATHERINE
RECALDE CERDA

Dra. Lorena Katherine Recalde Cerda

DIRECTORA DE PROYECTO

Firmado electrónicamente por:
EDISON FERNANDO
LOZA AGUIRRE

Dr. Edison Fernando Loza Aguirre

CODIRECTOR DE PROYECTO

DECLARACIÓN

Yo, Leonardo Andrés Benítez Orellana, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Firmado electrónicamente por:
LEONARDO ANDRES
BENITEZ ORELLANA

_____

Leonardo Andrés Benítez Orellana

DEDICATORIA

A mi madre y a mi esposa

AGRADECIMIENTOS

A mi directora Lorena Recalde, por su paciencia, dedicación y amor a la docencia. A Diana Ramírez por su tiempo y valiosa colaboración.

CONTENIDO

RESUMEN 1
ABSTRACT 2
1. INTRODUCCIÓN 3
1.1 PLANTEAMIENTO DEL PROBLEMA 3
1.2 OBJETIVO GENERAL 4
1.3 OBJETIVOS ESPECÍFICOS 4
1.4 MARCO TEÓRICO 5
1.4.1 Bag of Words 5
1.4.2 Term frequency-inverse document frequency 5
1.4.3 Word2vec 6
1.4.4 Empath 6
1.4.5 Emolex 6
1.4.6 BERT 7
1.5 REVISIÓN DE LA LITERATURA 7
1.5.1 Detección de Trastornos Mentales en Redes Sociales 7
1.5.2 Adicción a los videojuegos 8
2. METODOLOGÍA 10
2.1 Extracción de las publicaciones de Reddit 11
2.2 Filtrado de publicaciones 12
2.3 Proceso de etiquetado por parte de expertos 13
2.4 Análisis y caracterización de las publicaciones etiquetadas 15
2.5 Preprocesamiento de las publicaciones etiquetadas 15
2.6 Modelamiento del texto 18
2.6.1 Representación del texto 18
2.6.2 Modelos de clasificación 21
3. RESULTADOS 23
3.1 Bag of Words (BoW) 23
3.2 Term Frequency Inverse Document Frequency (TF-IDF) 23
3.3 Word2vec 24
3.4 Bag of Words + Empath 25
3.5 Bag of Words + Emolex 26
3.6 Bag of Words + Empath + Emolex 26

3.7 BERT 27
3.8 Limitaciones 27
4. CONCLUSIONES 28
REFERENCIAS 29

ÍNDICE DE FIGURAS

2.1 Etapas de la metodología Design Science Research 10
2.2 Longitud texto Adicto Videojuegos vs No Adicto Videojuegos 12
2.3 Arquitectura SAPRAV 13
2.4 Sistema de análisis y predicción de riesgo de adicción a videojuegos 14
2.5 Adicto Videojuegos vs No Adicto Videojuegos 15
2.6 Word Cloud Adicto Videojuegos 16
2.7 Word Cloud No Adicto Videojuegos 16
2.8 Trigrama Adicto Videojuegos 17
2.9 Trigrama No Adicto Videojuegos 17
2.10 KNN número de vecinos 21
2.11 Decision Tree Profundidad Máxima del Árbol 22
3.1 Visualización modelo Word2vec 25

ÍNDICE DE TABLAS

2.1 Bag of words 18
2.2 TF-IDF 18
2.3 Word2vec 19
2.4 Empath 19
2.5 Emolex 20
2.6 BERT 20
3.1 Modelo de clasificación supervisado BoW 23
3.2 Modelo de clasificación supervisado TF-IDF 24
3.3 Modelo de clasificación supervisado Word2vec 24
3.4 Modelo de clasificación supervisado BoW + Empath 25
3.5 Modelo de clasificación supervisado BoW + Emolex 26
3.6 Modelo de clasificación supervisado BoW + Empath + Emolex 26
3.7 Modelo de clasificación supervisado BERT 27

RESUMEN

En los últimos años, una gran cantidad de estudios se han centrado en la identificación de una amplia gama de enfermedades mentales mediante el uso y la aplicación de minería de datos, análisis de datos y aprendizaje automático. Actualmente, la adicción a los videojuegos se considera un problema mundial que se ha vuelto cada vez más frecuente y que afecta directamente la calidad de vida de las personas que la padecen y también a su entorno familiar más cercano. En el presente trabajo, nos propusimos desarrollar un marco novedoso para la detección de posibles rasgos de adicción a los videojuegos enfocado en un grupo de usuarios dentro de la red social Reddit. Para ello se extrajeron alrededor de 987 publicaciones escritas por usuarios de habla inglesa con el fin de analizar y procesar el texto obtenido para generar seis modelos de texto utilizando BoW, TF-IDF, Word2vec, Empath, Emolex y BERT. Esos modelos también fueron evaluados y probados usando cuatro algoritmos de clasificación supervisados: Logistic Regression, KNN, Decision Tree y AdaBoost.

Los resultados obtenidos demostraron que es posible identificar efectivamente los rasgos de adicción a los videojuegos en un grupo de usuarios en riesgo usando Word2vec mediante la generación de word embeddings, utilizando un conjunto de datos previamente entrenado.

Palabras clave. Adición a los videojuegos, Reddit, BoW, TF-IDF, Empath, Emolex, BERT, Modelamiento de texto, Machine Learning

ABSTRACT

In recent years, a large number of studies have been focusing on the identification of a wide range of mental illnesses through the use and application of data mining, data analysis and machine learning. Currently, addiction to video games is considered a global issue that has become increasingly frequent and that directly affects the life's quality of the people who suffer from it and also their closest family environment. In the present work, we proposed to develop a novel framework for the detection of possible traits of addiction to video games focused on a group of users inside the Reddit social network. We extracted about 987 posts written by English-speaking users with the purpose of analyzing and processing the text obtained to generate six text models using BoW, TF-IDF, Word2vec, Empath, Emolex and BERT. Those models were also evaluated and tested using four supervised classification algorithms: Logistic Regression, KNN, Decision Tree and AdaBoost.

The results obtained showed that it is possible to effectively identify video game addiction traits in a group of users at risk using Word2vec by generating word embeddings over a previously trained dataset.

Keywords. A Video game addiction, mental disorders Reddit, BoW, TF-IDF, Empath, Emolex, BERT, Text modeling, Machine Learning

1. INTRODUCCIÓN

1.1 PLANTEAMIENTO DEL PROBLEMA

La adicción a los videojuegos es un trastorno mental que afecta a una gran cantidad de personas en todo el mundo. Se estima que más de 214 millones de personas en los Estados Unidos consumen habitualmente algún tipo de videojuego, ya sea usando una computadora, teléfono celular o consola de videojuegos. Aproximadamente un 3% de estas personas sufrirá algún tipo de adicción relacionada con los videojuegos a lo largo de su vida [1]. Este trastorno a menudo se caracteriza por el uso compulsivo, problemático o excesivo de videojuegos que puede causar una amplia variedad de perturbaciones emocionales, desde una profunda tristeza hasta la pérdida de interés en las actividades diarias [2]. La dependencia de los videojuegos altera y modifica la forma normal en que los individuos interactúan durante su vida con amigos, familiares y la sociedad que les rodea. De hecho, esta adicción puede producir cambios en la estructura del cerebro similares a los causados por el uso de drogas o estupefacientes [3].

Estudios a largo plazo han demostrado que un gran número de personas que sufren adicción a los videojuegos pueden desarrollar ideas o tendencias suicidas que pueden conducir a la muerte o autolesiones [4]. Esto es alarmante si se considera que alrededor de 800.000 personas mueren por suicidio o causas relacionadas cada año, lo que convierte al suicidio en la segunda causa de muerte entre la población de 15 a 29 años en todo el mundo [5].

Además de estos factores, cabe mencionar que la adicción a los videojuegos se ha asociado en gran medida con un número considerable de problemas de salud que involucran obesidad, foto sensibilidad, calambres musculares, tendinitis, entre otros [6]. La Organización Mundial de la Salud (OMS) clasificó la adicción a los videojuegos como un trastorno de salud mental por primera vez en 2021. Debido a esto, la adicción a los videojuegos ha llamado la atención de diversas comunidades de investigación, que intentan comprender sus posibles causas, efectos y tratamientos viables.

Actualmente, los médicos tienen a su disposición algunos cuestionarios que tratan de identificar si una persona muestra o no signos o rasgos de adicción a los videojuegos. La mayoría de ellos se centran en cuestiones referentes al estado de ánimo de la persona y su relación cotidiana con la tecnología. Algunas de las pruebas más aplicadas son el Video Game Addiction Test (VAT) [7] y el Cuestionario de Experiencias Asociadas a los Videojuegos (CERV) [8].

Por otro lado, el rápido y continuo aumento del acceso a Internet y el crecimiento vertiginoso de las redes sociales han hecho que millones de usuarios puedan expresar sus emociones, sentimientos o estados de ánimo en la red. De hecho, los usuarios han encontrado un uso particular de plataformas como Facebook, Twitter o Reddit para este fin [9]. Las redes sociales representan el medio perfecto para captar, no solo el estado anímico y mental de una persona, sino también emociones relacionadas con estados depresivos como soledad, tristeza, culpa, inutilidad, entre otros [10]. Debido a este fenómeno, los datos generados por estas plataformas sociales han captado el interés de un gran número de investigadores en los últimos años. En particular, su objetivo es identificar nuevos métodos para detectar signos de adicción a partir del análisis de los datos disponibles en las redes en línea [11,12].

En este contexto, este trabajo tiene como objetivo evaluar el poder predictivo de cuatro algoritmos de aprendizaje automático cuya tarea es clasificar la publicación de un usuario extraída de Reddit como "riesgo de adicción al juego" o "sin riesgo". Para ello, se desarrollará un marco teórico-práctico para identificar adecuadamente si un usuario de habla hispana presenta signos o rasgos de adicción a los videojuegos, por lo que este trabajo busca contribuir con un método alternativo de detección de este tipo de enfermedad mental de forma que pueda ser utilizado como una herramienta de apoyo por psicólogos para la toma de decisiones.

Este documento está organizado en el siguiente orden: el Capítulo1 analiza el trabajo relacionado y resume el contexto de la presente investigación; el Capítulo2 presenta la metodología que se construyó en este contexto y cómo se entendieron los datos para definir el marco de investigación; el Capítulo 3 detalla la configuración preliminar y recopila los resultados; finalmente, el Capítulo 4 muestra las conclusiones y discute el trabajo futuro.

1.2 OBJETIVO GENERAL

Desarrollar un modelo de clasificación para la detección de rasgos de adicción a los videojuegos basado en Machine Learning, Minería de Datos y Reddit.

1.3 OBJETIVOS ESPECÍFICOS

- Realizar la revisión de literatura enfocada en los datos generados por las redes sociales y estudios relacionados con la adicción a los videojuegos para identificar el estado actual del arte.
- Aplicar minería de texto dentro de la plataforma Reddit con el fin de obtener y procesar las publicaciones necesarias que han sido generadas por los usuarios de esa comunidad.

- Diseñar una estrategia de etiquetado para las publicaciones obtenidas de la plataforma Reddit que contengan información relacionada a usuarios que presentan rasgos de adicción a los videojuegos.
- Realizar el modelamiento de texto de los datos obtenidos utilizando varias técnicas de modelamiento.
- Entrenar y evaluar varios modelos de aprendizaje de máquina usando algoritmos de clasificación supervisados.

1.4 MARCO TEÓRICO

1.4.1 Bag of Words

Bag of Words (BoW) es un modelo de texto aplicado en el campo del procesamiento del lenguaje natural para representar un documento de texto como una bolsa de palabras continuas ignorando la gramática o el orden de las palabras [13]. Para lograr esto, se genera una matriz de términos de frecuencia en la que cada una de las columnas representa una palabra dentro del corpus y cada fila representa cuántas veces aparece esa palabra en cada oración.

Como resultado de este proceso se obtendrá una matriz en la que cada columna representará una palabra y su correspondiente número de apariciones, y cada fila representará una oración dentro del set de datos.

1.4.2 Term frequency-inverse document frequency

Term frequency-inverse document frequency (TF-IDF) es una medida estadística que nos permite determinar qué palabras o conjunto de palabras son más relevantes dentro de un corpus de texto. La idea principal de TF-IDF se basa en la premisa de que, si una palabra se reconoce como rara o poco frecuente, pero aparece varias veces en el documento, es probable que esta palabra refleje las características de ese texto [14].

TD-IDF se obtiene al multiplicar dos métricas: la primera, referente a cuántas veces una palabra aparece en un texto; y la segunda, referente a la frecuencia de documento inversa de una palabra. De esta manera, si una palabra aparece muchas veces en el documento su valor se acercará a 0, caso contrario su valor se aproximará a 1.

1.4.3 Word2vec

Word2vec fue creado e introducido por Mikolov et al. [15] y consiste en generar una representación vectorial de cada palabra disponible en el corpus. El vector producido por esta representación almacena información semántica que permitirá asociar o no ese vector en particular con otros vectores. La información semántica generada es de vital importancia para el funcionamiento de Word2vec debido a que incluye las relaciones semánticas de cada palabra, sus definiciones, el contexto, entre otros.

Gracias a Word2vec se puede identificar vectores de palabras similares o palabras opuestas dentro del corpus y también se pueden aplicar operaciones algébricas a un grupo de vectores para encontrar similitudes o diferencias dentro de los distintos vectores generados por este modelo.

1.4.4 Empath

Empath permite a los investigadores generar y validar nuevas categorías léxicas según sea necesario mediante una combinación de aprendizaje profundo y colaboración colectiva [16]. Empath define 194 categorías de texto que pueden ser utilizadas de acuerdo con las necesidades del usuario. Así, al generar nuevas categorías léxicas también se generarán nuevas características dentro del set de datos que pueden ayudar a los investigadores a entender de mejor manera los tópicos o temas generales que contiene un texto específico dentro del corpus.

1.4.5 Emolex

Emolex es un analizador léxico basado en el National Research Council of Canada Emotion Lexicon (NRCLex) y se emplea para identificar emociones dentro del texto. Emolex define una lista de 10 emociones que pueden ser utilizadas de acuerdo a las necesidades del usuario. Gracias a este tipo de analizadores léxicos, se puede inferir de manera rápida y precisa los estados de ánimo que pueden estar presentes dentro de un texto específico.

De manera similar a Empath, Emolex agregará nuevas características al set de datos, incrementando el tamaño de la matriz generada.

1.4.6 BERT

BERT es un nuevo modelo de representación de datos desarrollado por Google. Utiliza una red bidireccional de transformadores de texto que es capaz de representar una oración o un grupo de oraciones como matrices consecutivas de tokens [17]. Esta bidireccionalidad implica que las oraciones se analizarán de izquierda a derecha y viceversa, proporcionando así una mejor comprensión del texto y su contexto.

Adicionalmente, gracias a la direccionalidad que aporta BERT, se puede entender de mejor manera el contexto al que una palabra hace referencia debido a que cada palabra puede tomar un sentido diferente dependiendo del lugar en el que se encuentre dicha palabra dentro de una oración o texto.

En el caso puntual de BERT y dentro del presente trabajo, se utilizó la implementación DistilBERT porque se considera más rápida y más pequeña que la implementación BERT original [18].

1.5 REVISIÓN DE LA LITERATURA

La minería de datos, el análisis de datos y el aprendizaje automático han evolucionado constantemente para buscar nuevas soluciones a los problemas contemporáneos relacionados con la clasificación de datos, la regresión de datos y la agrupación de datos. En este apartado se describen diferentes estudios que se han llevado a cabo relacionados con las redes sociales, los videojuegos y las enfermedades mentales en las últimas décadas.

1.5.1 Detección de Trastornos Mentales en Redes Sociales

El uso de las redes sociales como un lugar común en el que las personas pueden expresar voluntariamente sus emociones, estado de ánimo u opinión ha cobrado mayor relevancia y popularidad en los últimos años. Debido a esta circunstancia, varios estudios se han centrado en tratar de detectar emociones, trastornos mentales y adicciones utilizando la enorme cantidad de datos que existen en estas redes sociales.

En su trabajo, Ramírez-Cifuentes et al. [19], analizaron y exploraron el comportamiento relacional y multimodal de datos de diferentes usuarios extraídos de la red social Twitter. En su estudio, generaron y evaluaron múltiples modelos de aprendizaje automático con el objetivo de detectar usuarios en riesgo de desarrollar tendencias suicidas o su idealización. Su trabajo

determinó que los métodos aplicados que se utilizaron para identificar esas tendencias suicidas también pueden adoptarse, extenderse y aplicarse a diferentes casos de uso para detectar otras enfermedades mentales, además de la ideación suicida.

Por su parte, Shuai et al. [20], han demostrado que la creciente popularidad de las redes sociales y su tendencia de uso están asociadas con un número creciente de enfermedades mentales como la adicción a las relaciones cibernéticas, la sobrecarga de información y el uso compulsivo de Internet. En su investigación, los autores propusieron y desarrollaron un marco de aprendizaje automático que intenta explotar los datos generados dentro de las redes sociales para identificar adecuadamente los casos potenciales de esas enfermedades mentales.

1.5.2 Adicción a los videojuegos

En la última década ha habido un incremento considerable de estudios que intentan examinar múltiples aspectos de la adicción a los videojuegos [2]. Como consecuencia, la adicción a los videojuegos se ha convertido en un tema de interés general en las sociedades modernas. Autores como Griffiths et al. [2] han estudiado la prevalencia de la adicción problemática a los videojuegos y sus consecuencias negativas, así como la evolución que ha sufrido esta adicción a lo largo del tiempo. Como parte de su trabajo, los autores analizaron el riesgo que pueden desarrollar los jóvenes al migrar de los juegos en línea a los sitios de apuestas en línea. Aquí se concluye que además de perder el tiempo, también pueden perder su dinero o su trabajo.

De la misma manera, Wang et al. [21] han examinado de cerca la relación de la adicción a los videojuegos móviles con la ansiedad, la depresión y la soledad entre los adolescentes. Los autores concluyeron que este tipo de adicción estaba asociada con emociones negativas y que particularmente los hombres tienden a sufrir más que las mujeres al desarrollar una relación tóxica con la tecnología.

Además, Weinstein [3] revisó múltiples estudios realizados en el cerebro. De hecho, se estudiaron numerosas imágenes de TC para medir los niveles de dopamina producidos mientras una persona jugaba videojuegos. La evidencia de este estudio sugirió que los usuarios de videojuegos podrían mostrar signos reducidos de respuesta de dopamina, presumiblemente debido a cambios en el sistema de recompensa del cerebro similares a los causados por el uso de drogas o sustancias.

Tras la revisión bibliográfica correspondiente, pudimos identificar que, además del test VAT y CERT, no se han utilizado nuevos métodos de detección para identificar el riesgo de adicción

a los videojuegos o los signos relacionados con el mismo. Como resultado, nuestro trabajo presenta un enfoque novedoso para la posible identificación de este problema de salud mental mediante el uso de minería de datos, análisis de datos y aprendizaje automático. Para lograr este objetivo, hemos definido una aproximación basada en etapas sobre el marco metodológico de Design Science Research (DSR). Los detalles de cada etapa se describen en profundidad en el siguiente capítulo.

2. METODOLOGÍA

En el presente trabajo se utilizará Design Science Research (DSR) como metodología de investigación (Figura 2.1). DSR, es un paradigma de investigación que tiene sus orígenes en la ingeniería, el desarrollo de software y la inteligencia artificial [22]. DSR fue escogida para este trabajo debido a que su enfoque se centra en la entrega de un prototipo o de un modelo probado y validado para la solución de un problema específico; el cual, en marco del presente trabajo, corresponde a un modelo para la detección de usuarios en riesgo de adicción a los videojuegos.

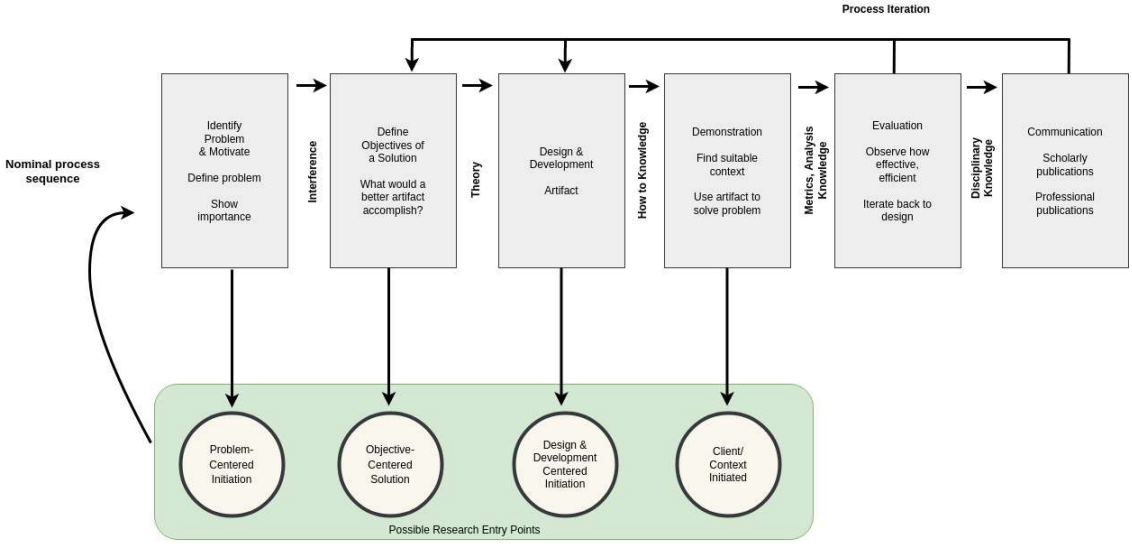

Figura 2.1 Etapas de la metodología Design Science Research [22]

El trabajo se concentró entonces en siete etapas, las cuales son:

1) **Extracción de las publicaciones de Reddit.** El texto de varias publicaciones de usuarios de Reddit se extrajo dentro de entradas relacionadas con videojuegos, adicción a los videojuegos y automotivación.

2) **Filtrado de publicaciones.** El texto de las publicaciones extraídas previamente se inspeccionó y filtró manualmente para obtener publicaciones con una consistencia de texto aceptable y una longitud de vocabulario adecuada.

3) **Proceso de etiquetado por parte de expertos.** Una aplicación web que contiene las publicaciones filtradas fue desarrollada completamente por nosotros como contribución

a este documento. Se otorgó acceso a varios psicólogos a la plataforma para que pudieran etiquetar cada publicación filtrada. Los posibles valores de etiquetado fueron: 1 para “usuario de riesgo” y 0 para “usuario sin riesgo”.

4) **Análisis y caracterización de las publicaciones etiquetadas.** El análisis exploratorio de las publicaciones etiquetadas se realizó para comprender adecuadamente las características únicas del conjunto de datos.

5) **Preprocesamiento de las publicaciones etiquetadas.** Cada una de las publicaciones etiquetadas se procesó para limpiarlas y tokenizarlas. Como parte de este proceso, se eliminaron algunos caracteres del texto original porque podrían interferir con el proceso de modelado del texto. Después de esto, todo el texto útil fue identificado y tokenizado.

6) **Modelamiento del texto.** El texto procesado y tokenizado se transformó para desarrollar múltiples modelos de texto: BoW, TF-IDF, Word2vec, BoW + Empath, BoW + Emolex y BERT. En este paso, se llevaron a cabo diferentes procesos de transformación del texto debido a que la naturaleza de cada modelo de clasificación es distinta para cada caso.

7) **Evaluación de los modelos.** El paso final fue evaluar y probar varios algoritmos de clasificación para verificar cuán efectivo, en términos de predicción, era cada uno de ellos.

2.1 Extracción de las publicaciones de Reddit

La primera etapa del proceso fue extraer texto de la red social Reddit. Con eso en mente, se tuvo que abrir una nueva cuenta de usuario para ese propósito. Reddit es una plataforma social de Internet ampliamente conocida por la amplia variedad de temas que son agregados por la gran comunidad de usuarios. Reddit brinda a los usuarios y desarrolladores acceso a varias interfaces de programación de aplicaciones (APIs) para tareas que involucran extracción de texto, análisis de datos y generación de datos. Una vez que se completó el proceso de registro en Reddit, se recuperaron la ID de usuario y el token de acceso para obtener acceso a la API de extracción de texto de Reddit. El ID de usuario y el token de acceso se utilizaron para desarrollar un script de Python que extrajo la información de las publicaciones escritas por los usuarios registrados dentro de Reddit entre varios temas relevantes: videojuegos, adicción a los videojuegos y automotivación en general.

Para garantizar la privacidad y el anonimato de cada usuario, se eliminó el nombre de usuario original previamente almacenado en la publicación inicial y se generó un nuevo nombre de usuario aleatorio. Se realizaron varias solicitudes de API durante cinco semanas. Como consecuencia, se extrajeron 987 publicaciones. Esas publicaciones contenían 11.276 líneas de texto y constaban de 1.187.375 millones de caracteres. Todas las 987 publicaciones se almacenaron en un archivo de texto plano y continuamos con la siguiente etapa relacionada con el filtrado de texto.

2.2 Filtrado de publicaciones

Una vez que se obtuvieron las publicaciones de los usuarios de Reddit a través de un script de Python, todas esas publicaciones se tradujeron automáticamente de inglés a español utilizando la API de Google Translate como otra medida para garantizar el anonimato de los usuarios. En este paso, todas las publicaciones traducidas se escanearon nuevamente en busca de errores gramaticales o sintácticos e inconsistencias dentro del texto. Además, se examinó cada una de las 987 publicaciones para asegurar que el texto poseía información consistente y una longitud adecuada. La longitud mínima de texto obtenida fue de 140 caracteres y la máxima de 11.321 caracteres. La Figura 2.2 muestra la longitud de texto para ambos grupos.

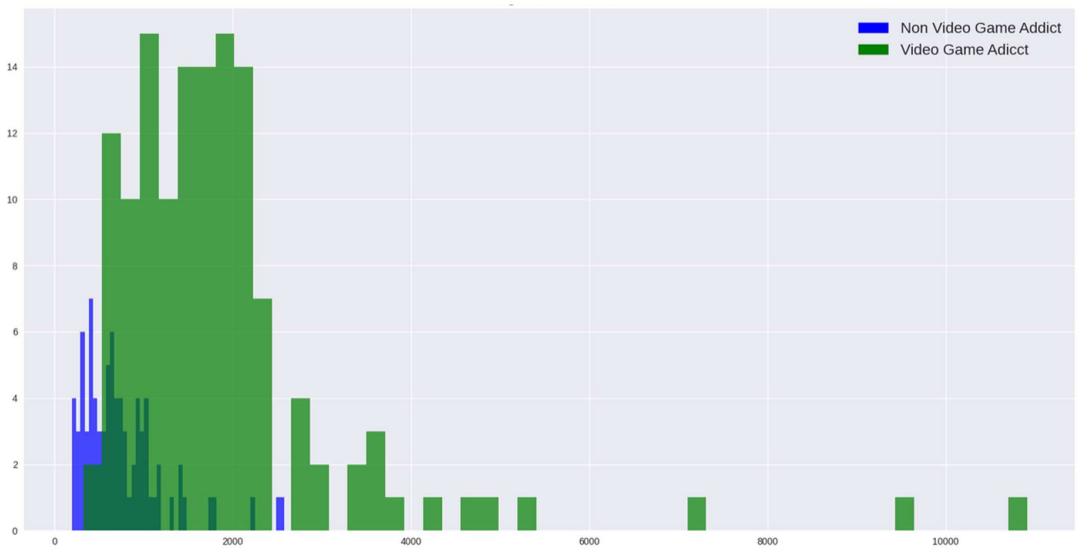

Figura 2.2 Longitud texto Adicto Videojuegos vs No Adicto Videojuegos

Como resultado final de este proceso se obtuvieron 370 publicaciones en español. Cada una de las 370 publicaciones se evaluó nuevamente en busca de posibles errores de traducción, puntuación u ortografía para facilitar el proceso de etiquetado en el siguiente paso.

2.3 Proceso de etiquetado por parte de expertos

En esta etapa desarrollamos una aplicación web para facilitar y acelerar el proceso de etiquetado que tradicionalmente se lograba mediante el uso de documentos de texto u hojas de datos extendidas. Con este propósito, se utilizó un servidor de base de datos MySQL para insertar las 370 publicaciones previamente filtradas y traducidas. El backend de la aplicación web se creó con Flask, un framework escrito en Python, y el frontend se creó con React, una biblioteca de Javascript para crear interfaces web. Se bautizó a esta aplicación como Sistema de análisis y predicción de riesgo de adicción a videojuegos (SAPRAV). La Figura 2.3 describe la arquitectura utilizada para el desarrollo de SAPRAV.

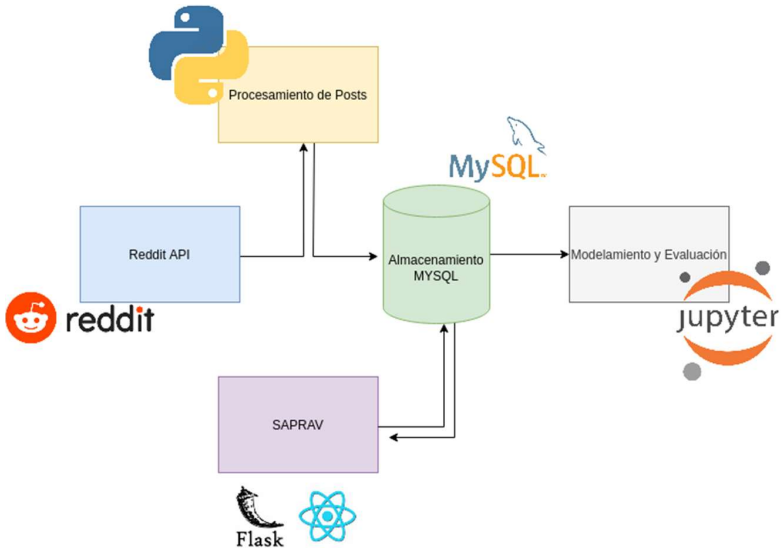

Figura 2.3 Arquitectura SAPRAV

Esta aplicación web facilitó el proceso de etiquetado a través de una interfaz gráfica que permitía a los psicólogos seleccionar cualquier publicación por su ID. Al seleccionar una publicación, se muestra el contenido correspondiente y se le pide al psicólogo que responda la siguiente pregunta**: ¿este usuario es adicto a los videojuegos?**, entonces el psicólogo podrá elegir entre dos opciones:

1) **Sí, el usuario es adicto a los videojuegos**

2) **No, el usuario no es adicto a los videojuegos**

En caso de seleccionar la segunda opción, al psicólogo se le hará una nueva pregunta: **¿en qué categoría encaja el usuario?**, entonces el psicólogo podrá elegir entre cuatro opciones:

1) Usuario o familiar buscando consejo

2) Usuario que habla de su experiencia con los juegos

3) Usuario que brinda consejo

4) Otra categoría

Un grupo de tres psicólogos tuvo acceso a la aplicación web desarrollada como parte del proceso de etiquetado. El grupo de expertos se tomó un tiempo promedio de seis semanas para completar este paso. La Figura 2.4 muestra el aspecto de la aplicación web.

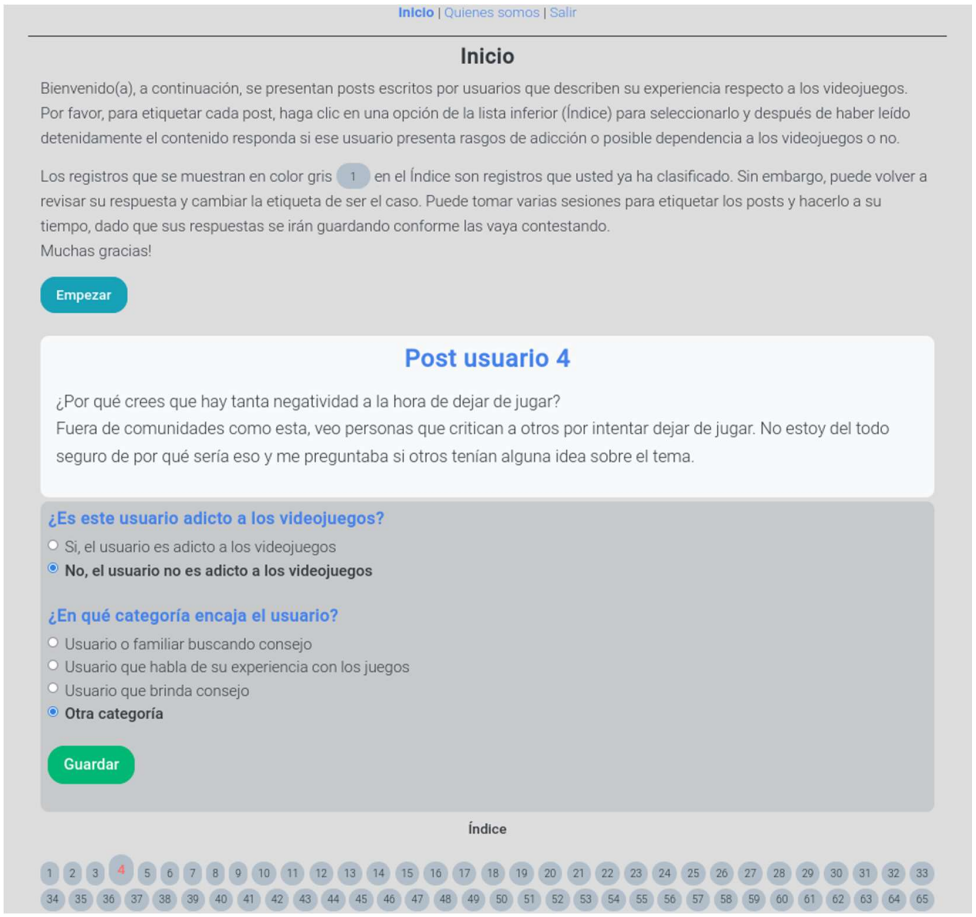

Figura 2.4 Sistema de análisis y predicción de riesgo de adicción a videojuegos (SAPRAV)

2.4 Análisis y caracterización de las publicaciones etiquetadas

La tabla MySQL correspondiente a las publicaciones etiquetadas dentro de la base de datos se exportó a un archivo CSV. Este archivo CSV se cargó en un Jupyter Notebook usando Google Collab y se realizó el análisis de texto inicial. El primer paso fue comprobar el número de usuarios que estaban etiquetados como adictos a los videojuegos y los que no. Del total de 370 publicaciones etiquetadas, solo se tomó en cuenta aquellas en las que los tres psicólogos estuvieron totalmente de acuerdo. Es decir, las publicaciones seleccionadas fueron aquellas con tres etiquetas de riesgo de adicción o aquellas con tres etiquetas de inexistencia de riesgo. Así, se obtuvieron 214 publicaciones que formaron el conjunto de datos final. Como se muestra en la Figura 2.5, 132 publicaciones se marcaron como adictos a los videojuegos y 82 como no adictos a los videojuegos.

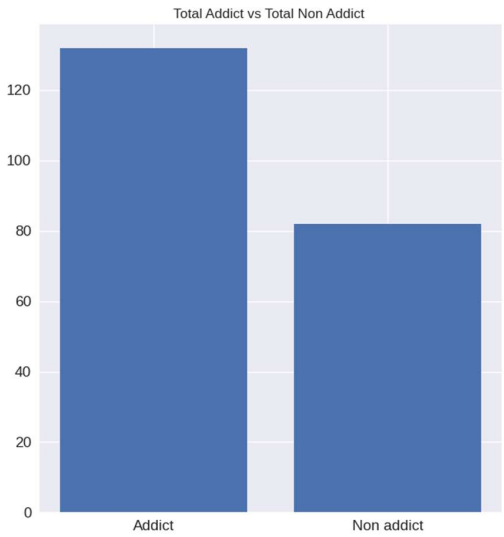

Figura 2.5 Adicto Videojuegos vs No Adicto Videojuegos

Se pudo concluir que el conjunto de datos presentaba un desbalance mínimo, por lo que no se consideró necesario aplicar un proceso de balanceo al conjunto de datos actual.

2.5 Preprocesamiento de las publicaciones etiquetadas

En la siguiente etapa, se transformó el texto filtrado a minúsculas y se eliminó la puntuación y los caracteres numéricos de cada publicación. Además, las palabras vacías en español se

eliminaron de la publicación original utilizando la biblioteca NLTK, el kit de herramientas de lenguaje natural de Python. En este punto, se prestó especial atención a los acentos de las vocales por lo que se sustituyeron por las vocales sin ellos.

Para que las publicaciones etiquetadas sean utilizables por los modelos de aprendizaje automático, es necesario tokenizar el contenido de cada publicación. La tokenización simplemente consiste en segmentar toda la información del texto en una lista de palabras. Procedimos a tokenizar cada una de las publicaciones etiquetadas e introducir una nueva columna con la publicación tokenizada dentro del conjunto de datos.

Con el corpus final, se construyó una nube de palabras para cada uno de los grupos de control, en cada nube de palabras se pudo observar sentimientos que describían estados anímicos opuestos. Las Figuras 2.6 y 2.7 muestran esos hallazgos.

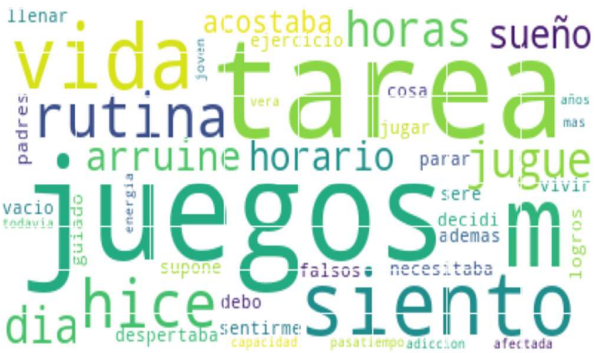

Figura 2.6 Word Cloud Adicto Videojuegos

Figura 2.7 Word Cloud No Adicto Videojuegos

Del mismo modo, se empezó con la búsqueda de n-gramas, obteniendo interesantes resultados dentro del grupo de trigramas. En este grupo de usuarios etiquetados como adictos a los videojuegos, los trigramas mostraban algunos nombres de videojuegos de fama mundial, como: League of Legends (LOL), Call of Duty, World of Warcraft (WOW) y Age of Empires. Por su parte, en el grupo de usuarios etiquetados como no adictos a los videojuegos se detectaron nombres de actividades cotidianas como: disfrutar del aire, dejar de jugar súbitamente, programas de televisión y pasatiempos para reemplazar juegos. Las Figuras 2.8 y 2.9 muestran los primeros 10 valores encontrados.

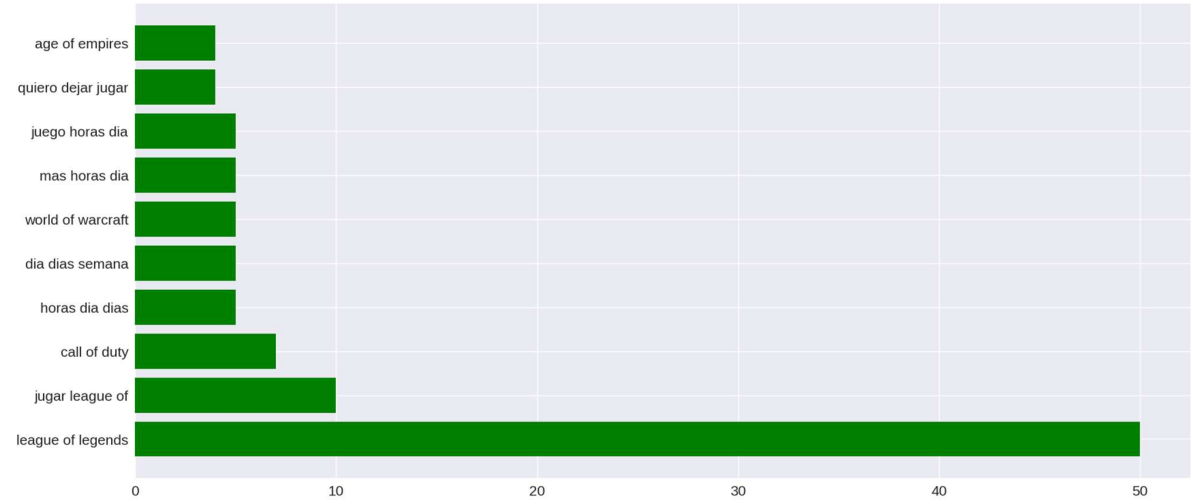

Figura 2.8 Trigrama Adicto Videojuegos

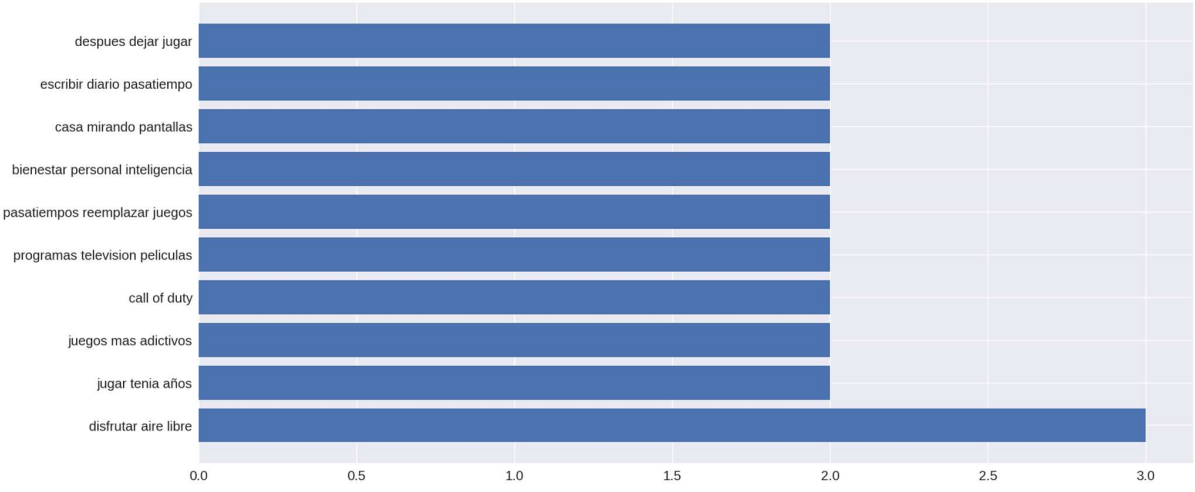

Figura 2.9 Trigrama No Adicto Videojuegos

2.6 Modelamiento del texto

Las publicaciones preprocesadas sirvieron como entrada para los siguientes modelos de clasificación: Logistic Regression, KNN, Decision Tree y Adaboost. Los pasos involucrados en el proceso de modelado de datos y las diferentes técnicas aplicadas al texto se describen a continuación.

2.6.1 Representación del texto

El primer modelo de texto utilizado fue Bag of Words (BoW), para la creación de la matriz de frecuencia de términos; se requirió la biblioteca Scikit-learn Python. Esta biblioteca nos permitió crear el objeto *CountVectorizer* para este propósito. La matriz de frecuencias de términos obtenida tuvo 5408 columnas y 214 filas. La Tabla 2.1 describe el conteo de palabras resultantes en BOW.

u_n	$word_1$	$word_2$	…	$word_n$
1	0	1	...	1
2	1	1	...	0
3	...	...	...	...
4	1	1	...	1

Tabla 2.1 Bag of words

En el caso del segundo modelo Term frequency-inverse document frequency (**TF-IDF**), para obtener la matriz de frecuencia se utilizó el objeto *TfidfVectorizer* de la librería Scikit-learn de Python que como resultado nos devolvió una matriz de 5408 columnas y 214 filas. La Tabla 2.2 muestra esos resultados.

u_n	$word_1$	$word_2$	…	$word_n$
1	0.001231	0.001231	...	0.001231
2	0.020345	0.005611	...	0.001411
3	...	...	...	...
4	0.021112	0.001231	...	0.000428

Tabla 2.2 TF-IDF

Para el tercer modelo de Word2vec se utilizó la librería Gensim de Python. Este paquete permite la generación de word embeddings utilizando el objeto *models.word2vec*. Además, se utilizó un modelo previamente entrenado por parte de Google con aproximadamente 3 millones de palabras y frases. Para convertir cada publicación a su representación vectorial, fue necesario transformar cada palabra dentro de esa publicación a su representación Word2vec usando la biblioteca Gensim. Luego se creó un nuevo vector con la suma de todos los vectores resultantes para esa publicación específica. Como resultado se generó una matriz de 300 columnas y 214 filas. Los resultados se muestran en la Tabla 2.3.

u_n	$vector_1$	$vector_2$	…	$vector_n$
1	-0.121488	-1.400876	...	2.200485
2	0.880013	1.300930	...	-7.998426
3	...	...	...	...
4	8.516687	-0.036691	...	-6.310077

Tabla 2.3 Word2vec

Para el cuarto modelo basado en Empath, la librería *empath-client* se usó para analizar cada una de las publicaciones para generar un diccionario que contenía un tópico como la clave del diccionario y el valor asociado para dicho tópico. El proceso se realizó para cada una de las 194 categorías léxicas dentro de cada publicación. De esa manera se generó una matriz de 194 columnas y 214 filas. La Tabla 2.4 muestra la matriz resultante.

u_n	$help_1$	$office_2$	…	$musical_n$
1	0.0	2.0	...	3.0
2	1.0	5.0	...	0.0
3	...	...	...	...
4	2.0	2.0	...	4.0

Tabla 2.4 Empath

En el caso del quinto modelo basado en Emolex, el proceso para generar la matriz fue similar a lo realizado con el modelo basado en Empath; pero, en este caso se utilizó la librería *nrclex* con la que género una matriz de 10 columnas y 214 filas. Para el uso de Empath y Emolex se utilizaron las librerías de Python *empath y nrclex* las mismas que nos permiten generar nuevas

características basadas en la detección de categorías y emociones. En la Tabla 2.5 se pueden observar dichos resultados.

u_n	$fear_1$	$anger_2$	…	joy_n
1	0.0667	0.1000	...	0.1500
2	0.0931	0.0489	...	0.1333
3	...	...	...	...
4	0.0821	0.0224	...	0.0881

Tabla 2.5 Emolex

Para el sexto modelo basado en BERT se tomó en cuenta que una de las singularidades que tiene DistilBERT está asociada a la tokenización del texto. DistilBERT considera un máximo de 512 tokens por oración, por lo que se deben crear oraciones con menos de 512 tokens para que ninguna de las oraciones se trunque automáticamente dentro del conjunto de datos. Este paso extra es fundamental para obtener mejores resultados con texto que contiene más de 512 caracteres de longitud.

Para llevar a cabo el proceso de tokenización se utilizó el objeto *DistilBertTokenizer* de la librería Python Transformers. Luego de realizar la tokenización se obtuvo una matriz conformada por 512 columnas y 227 filas. La Tabla 2.6 muestra la matriz que se obtuvo.

u_n	$sentence_1$	$sentence_2$	…	$sentence_n$
1	-0.381695	-0.079512	...	0.274537
2	-0.013423	-0.370102	...	0.185917
3	...	...	...	...
4	0.061567	-0.013278	...	-0.146876

Tabla 2.6 BERT

Después de estos pasos, finalmente podemos proceder a probar varios modelos de clasificación detallados en la siguiente etapa.

2.6.2 Modelos de clasificación

Luego de haber obtenido la representación del texto para cada una de las publicaciones filtradas, se procedió a entrenar cuatro algoritmos de aprendizaje automático supervisado. Debido a la naturaleza del problema identificado como una tarea de clasificación que busca establecer si un usuario presenta o no rasgos de adicción o dependencia a los videojuegos, se decidió entrenar y evaluar los siguientes modelos de clasificación: Logistic Regression, KNN, Decision Tree y Adaboost. Para el caso específico de KNN, el número de vecinos se estableció en 7 luego de haber realizado el respectivo análisis de precisión con el conjunto de datos tokenizado. La Figura 2.10 muestra el resultado.

De igual manera, en el caso del clasificador Decision Tree, la profundidad máxima del árbol se estableció en 7 luego de realizar el respectivo análisis de profundidad. La figura 2.11 muestra el resultado.

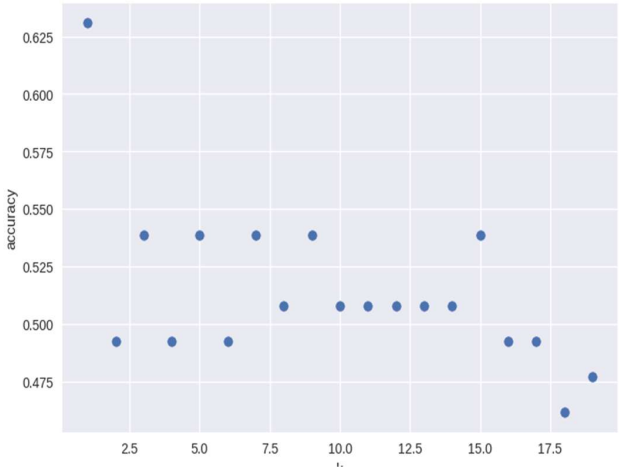

Figura 2.10 KNN número de vecinos

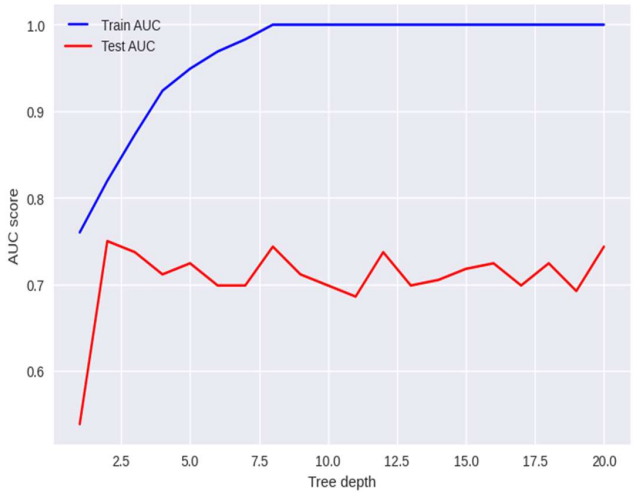

Figura 2.11 Decision Tree Profundidad Máxima del Árbol

En la sección 3 se detallarán los resultados de cada uno de estos modelos.

3. RESULTADOS

En esta etapa, cada uno de los conjuntos de datos obtenidos después del modelado de texto se dividió de la siguiente manera: 70 % de los datos para entrenamiento del modelo y 30 % de los datos restantes para validación y prueba del modelo. Para cada uno de los modelos de clasificación definidos previamente, se utilizaron las mismas particiones del conjunto de datos durante la fase de entrenamiento y prueba.

3.1 Bag of Words (BoW)

Los resultados para BoW nos muestran que el modelo de clasificación Logistic Regression fue el que mejor resultado obtuvo, la métrica de accuracy fue de 0.89 superior a todas las demás. Decision Tree y AdaBoost obtuvieron métricas similares, pero Decision Tree presentó un mejor accuracy de 0.69 sobre los 0.68 de AdaBoost. Finalmente, KNN fue el peor de todos los modelos con la métrica de accuracy de 0.51, similar a lo que elegir un resultado al azar. La Tabla 3.1 muestra el resultado de los modelos de clasificación para BoW.

Classification model	*precision*	*recall*	*f1-score*	*accuracy*
Logistic Regression	0.89	0.88	0.89	0.89
KNN	0.73	0.64	0.53	0.51
Decision Tree	0.67	0.67	0.67	0.69
AdaBoost	0.66	0.66	0.66	0.68

Tabla 3.1 Modelo de clasificación supervisado BoW

3.2 Term Frequency Inverse Document Frequency (TF-IDF)

Similares a los resultados obtenidos para el modelo anterior, Logistic Regression obtuvo los mejores resultados, con un accuracy de 0.84. AdaBoost obtuvo mejores resultados, con un accuracy de 0.70 superando a Decision Tree con un accuracy de 0.66. Del mismo modo KNN fue el que peor resultado obtuvo de los cuatro, con un accuracy de 0.62. La Tabla 3.2 muestra estos resultados.

Classification model	*precision*	*recall*	*f1-score*	*accuracy*
Logistic Regression	0.88	0.81	0.82	0.84
KNN	0.83	0.60	0.57	0.62
Decision Tree	0.64	0.64	0.64	0.66
AdaBoost	0.68	0.69	0.68	0.70

Tabla 3.2 Modelo de clasificación supervisado TF-IDF

3.3 Word2vec

Para el caso de Word2vec, se observa una mejora sustancial con respecto a todas las métricas para cada uno de los modelos. Este particular puede estar dado debido a que se obtuvo un modelo previamente entrenado por parte de Google para obtener los vectores resultantes con el set de datos utilizado dentro del presente estudio. En la Tabla 3.3 se pueden analizar esos resultados.

Se puede observar que nuevamente Logistic Regression es el modelo que mejores resultados obtiene con un accuracy de 0.94, seguido de cerca esta vez por KNN con un accuracy de 0.91. AdaBoost obtuvo un accuracy de 0.87 seguido finalmente por Decision Tree con un accuracy de 0.84. En la Figura 3.1 se observa claramente los dos grupos de usuarios generados mediante el modelo Word2vec.

Classification model	*precision*	*recall*	*f1-score*	*accuracy*
Logistic Regression	0.94	0.94	0.94	0.94
KNN	0.90	0.90	0.90	0.91
Decision Tree	0.90	0.85	0.86	0.84
AdaBoost	0.90	0.85	0.86	0.87

Tabla 3.3 Modelo de clasificación supervisado Word2vec

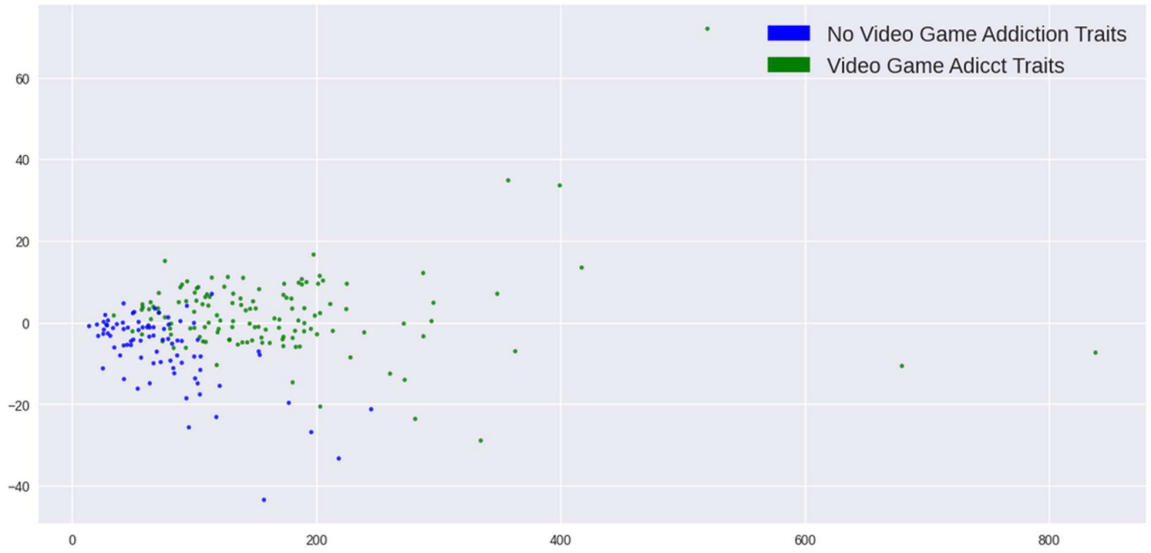

Figura 3.1 Visualización modelo Word2vec

3.4 Bag of Words + Empath

Los resultados obtenidos para el modelo BoW + Empath, en comparación con el primer modelo BoW, mostraron peores métricas para todos los modelos, excepto para Decision Tree, que esta vez logró mejores resultados que los obtenidos inicialmente utilizando BoW solamente. Esto podría deberse al aumento de la dimensionalidad en el conjunto de datos, ya que Empath generó 194 características nuevas. Adicionalmente, se mantuvo el orden de los resultados obtenidos en el primer modelo BoW, siendo Logistic Regression el mejor modelo con un accuracy de 0.79, seguido de Decision Tree, AdaBoost y KNN con 0.70, 0.68 y 0.40 respectivamente. La Tabla 3.4 muestra estos resultados.

Classification model	*precision*	*recall*	*f1-score*	*accuracy*
Logistic Regression	0.80	0.77	0.78	0.79
KNN	0.65	0.57	0.44	0.40
Decision Tree	0.69	0.67	0.68	0.70
AdaBoost	0.66	0.66	0.66	0.68

Tabla 3.4 Modelo de clasificación supervisado BoW + Empath

3.5 Bag of Words + Emolex

En el caso de BoW + Emolex, se puede observar que Logistic Regression se mantuvo como el mejor algoritmo de los cuatro, con un accuracy de 0,89. Decision Tree obtuvo un accuracy de 0,70 y AdaBoost obtuvo 0,68. De nuevo, KNN obtuvo los peores resultados, con un accuracy de 0,51. En la Tabla 3.5 se pueden observar esos resultados.

Se obtuvieron mejores resultados para los modelos de Logistic Regression y KNN que los obtenidos en BoW + Empath. Esto podría estar relacionado con la reducción de la dimensionalidad en el conjunto de datos utilizado en este modelo, ya que Emolex solo genera 11 características nuevas, a diferencia de las 194 generadas por Empath.

Classification model	*precision*	*recall*	*f1-score*	*accuracy*
Logistic Regression	0.89	0.88	0.89	0.89
KNN	0.73	0.64	0.53	0.51
Decision Tree	0.69	0.67	0.68	0.70
AdaBoost	0.66	0.66	0.66	0.68

Tabla 3.5 Modelo de clasificación supervisado BoW + Emolex

3.6 Bag of Words + Empath + Emolex

Al combinar BoW + Empath + Emolex se pudo observar que el mejor modelo fue Logistic Regression con un accuracy de 0.89, seguido de AdaBoost con 0.78, Decision Tree con 0.75 y por último KNN con 0.43. Es interesante notar que, con la extracción de nuevas características, los modelos Decision Tree y AdaBoost mejoraron, a diferencia de KNN que empeoró su rendimiento en comparación con el modelo BoW inicial. La Tabla 3.6 muestra estos resultados.

Classification model	*precision*	*recall*	*f1-score*	*accuracy*
Logistic Regression	0.90	0.88	0.89	0.89
KNN	0.72	0.59	0.46	0.43
Decision Tree	0.74	0.74	0.74	0.75
AdaBoost	0.78	0.78	0.78	0.78

Tabla 3.6 Modelo de clasificación supervisado BoW + Empath + Emolex

3.7 BERT

Finalmente, los resultados obtenidos con BERT muestran que nuevamente el mejor modelo fue Logistic Regression con 0.89. AdaBoost fue el segundo mejor modelo, esta vez con un accuracy de 0,88, seguido de KNN y Decision Tree con 0,87 y 0,82 respectivamente.

Al usar BERT, hay una mejora sustancial en el rendimiento de todos los modelos, especialmente en el caso de KNN. Similar al caso de Word2vec, se utilizó un modelo previamente entrenado por Google para luego ser aplicado sobre el conjunto de datos utilizado en el presente trabajo.

Classification model	*precision*	*recall*	*f1-score*	*accuracy*
Logistic Regression	0.87	0.88	0.87	0.89
KNN	0.92	0.81	0.84	0.87
Decision Tree	0.79	0.82	0.80	0.82
AdaBoost	0.87	0.85	0.86	0.88

Tabla 3.7 Modelo de clasificación supervisado BERT

3.8 Limitaciones

Tras analizar los resultados alcanzados en este trabajo, identificamos varias limitaciones. La primera está relacionada con el proceso de etiquetado de datos, debido a que los psicólogos que participaron en el proceso de etiquetado son especialistas en diversas áreas de la psicología, algunos de ellos pueden sentirse inseguros al momento de etiquetar a un usuario. Esto se debe a que, en gran medida, algunos factores (p. ej. el nivel de uso de los dispositivos tecnológicos, la terminología utilizada en los videojuegos, la antigüedad, el tamaño del conjunto de datos a etiquetar, entre otros); pueden influir negativamente en el proceso de etiquetado. De hecho, no todos los psicólogos estaban familiarizados con la terminología utilizada en redes sociales como Reddit.

La segunda está relacionada con el conjunto de datos recopilado de Reddit. El número de publicaciones obtenidas originalmente se redujo drásticamente después del proceso de filtrado. Además, el conjunto de datos original se escribió en inglés y luego se tradujo al español. Por lo tanto, existe la posibilidad de que algunas de las observaciones dentro del conjunto de datos no se identifiquen correctamente debido a errores sintácticos o gramaticales durante el proceso de traducción.

4. CONCLUSIONES

Las redes sociales han estado en constante evolución desde su creación. Ante este fenómeno, en los últimos años se han generado nuevas formas de comunicación e interacción. Como resultado de su expansión y crecimiento, las redes sociales generan una gran cantidad de datos útiles y valiosos que pueden ser utilizados en la creación de nuevas interfaces de software o hardware que mejoren la vida de otras personas. En el presente trabajo, con el método propuesto, ha sido posible detectar publicaciones que identifiquen a usuarios que presentan rasgos de adicción a los videojuegos. Así, ha sido posible clasificar a los usuarios de Reddit como “riesgo de adicción” o “sin riesgo”. Nuestra propuesta que involucra el uso de modelado de texto, aprendizaje automático y análisis de datos, y ha demostrado ser confiable.

En este trabajo se analizó el desempeño de seis modelos de texto y cuatro algoritmos de clasificación. Debido a la naturaleza del problema planteado inicialmente, Logistic Regression demostró un mejor rendimiento en la mayoría de los algoritmos de clasificación en combinación con Word2vec. Del mismo modo, BERT fue el segundo mejor modelo que obtuvo resultados aceptables en la identificación de rasgos de adicción a los videojuegos.

Con esto en mente, los estudios futuros pueden enfocarse en identificar varios tipos de adiciones relacionadas con la tecnología; por ejemplo, la dependencia de las redes sociales o el juego online. Para extender el presente trabajo, sugerimos no solo evaluar el texto generado por el usuario, sino también procesar y analizar el contenido de audio y video generado en otras redes sociales. Además, se podría usar un conjunto de datos más grande en el futuro, teniendo en cuenta que los datos del usuario deben etiquetarse de manera consistente y precisa para validar esas etiquetas nuevamente.

Para superar las limitaciones del presente trabajo se recomienda buscar la participación de un nuevo grupo de psicólogos expertos enfocados en adicciones tecnológicas y que, gracias a su especialización, estén familiarizados de mejor manera con el vocabulario y las expresiones utilizadas por los usuarios que padecen esa adicción.

REFERENCIAS

1. “2022 video game addiction statistics and Facts,” The Recovery Village Drug and Alcohol Rehab, 11-Oct-2022. [Online]. Available: https://www.therecoveryvillage.com/process-addiction/video-game-addiction/gaming-addiction-statistics/

2. M. D. Griffiths, D. J. Kuss, and D. L. King, “Video game addiction: Past, present and future,” Current Psychiatry Reviews, vol. 8, no. 4, pp. 308–318, 2012.

3. A. M. Weinstein, “Computer and video game addiction—a comparison between game users and non-game users,” The American Journal of Drug and Alcohol Abuse, vol. 36, no. 5, pp. 268–276, 2010. doi:10.3109/00952990.2010.491879

4. E. Messias, J. Castro, A. Saini, M. Usman, and D. Peeples, “Sadness, suicide, and their association with video game and internet overuse among teens: Results from the Youth Risk Behavior Survey 2007 and 2009,” Suicide and Life-Threatening Behavior, vol. 41, no. 3, pp. 307–315, 2011.

5. “Depression,” World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression. [Accessed: 24-Aug-2021].

6. M. D. Griffiths, “Video Game addiction: Further thoughts and observations,” International Journal of Mental Health and Addiction, vol. 6, no. 2, pp. 182–185, 2007.

7. A. J. van Rooij, T. M. Schoenmakers, R. J. J. M. van den Eijnden, A. A. Vermulst, and D. van de Mheen, “Video game addiction test: Validity and Psychometric Characteristics,” Cyberpsychology, Behavior, and Social Networking, vol. 15, no. 9, pp. 507–511, 2012.

8. A. Chamarro, X. Carbonell, J. M. Manresa, R. Munoz-Miralles, R. Ortega-Gonzalez, M. R. Lopez-Morron, C. Batalla-Martinez, and P. Toran-Monserrat, “El Cuestionario de Experiencias relacionadas con Los Videojuegos (CERV): Un Instrumento Para detectar El Uso problemático de videojuegos en adolescentes españoles,” Adicciones, vol. 26, no. 4, p. 303, 2014.

9. A. Holst, “Total data volume worldwide 2010-2025,” Statista, 07-Jun-2021. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/. [Accessed: 24-Aug-2021].

10. M. Nadeem, M. Horn, G.Coppersmith. “Identifying Depression on Twitter - arXiv.” [Online]. Available: https://export.arxiv.org/pdf/1607.07384. [Accessed: 24-Aug-2021].

11. G. Wallner, S. Kriglstein, and A. Drachen, “Tweeting your destiny: Profiling users in the Twitter landscape around an online game,” 2019 IEEE Conference on Games (CoG), 2019.

12. Y. Fan, Y. Zhang, Y. Ye, X. li, and W. Zheng, “Social Media for opioid addiction epidemiology,” Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017.

13. V.-A. Tran, D.-S. Le, H. H. Hung, and D.-Q. Nguyen, “Improving the accuracy of speech recognition models for non-native English speakers using bag-of-words and deep neural networks,” Scientific Review, no. 92, pp. 10–14, 2023. doi:10.32861/sr.91.10.14

14. X. Guan, Y. Li, and H. Gong, “Improved TF-IDF for WE MEDIA ARTICLE keywords extraction,” Journal of Physics: Conference Series, vol. 1302, no. 3, p. 032003, 2019. doi:10.1088/1742-6596/1302/3/032003

15. T. Mikolov and J. Dean, “Distributed representations of words and phrases and their compositionality,” Advances in neural information processing systems, 2013.

16. E. Fast, B. Chen, and M. S. Bernstein, “Empath,” Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016. doi:10.1145/2858036.2858535

17. Z. Gao, A. Feng, X. Song, and X. Wu, “Target-dependent sentiment classification with bert,” IEEE Access, vol. 7, pp. 154290–154299, 2019. doi:10.1109/access.2019.2946594

18. A. Joshy and S. Sundar, “Analyzing the performance of sentiment analysis using Bert, Distilbert, and Roberta,” 2022 IEEE International Power and Renewable Energy Conference (IPRECON), 2022. doi:10.1109/iprecon55716.2022.10059542

19. D. Ramírez-Cifuentes , A. Freire, R. Baeza-Yates, J. Puntí, P. Medina-Bravo, DA Velazquez, JM Gonfaus, J. Gonzàlez, “Detection of suicidal ideation on social media: Multimodal, Relational, and behavioral analysis,” Journal of Medical Internet Research, vol. 22, no. 7, 2020. doi:10.2196/17758

20. S.S. Shuai , CY Shen, DN Yang, YF Lan, WC Lee,PS Yu, MS Chen, “Mining online social data for detecting social network mental disorders,” Proceedings of the 25th International Conference on World Wide Web, 2016. doi:10.1145/2872427.2882996

21. J.-L. Wang, J.-R. Sheng, and H.-Z. Wang, “The association between mobile game addiction and depression, social anxiety, and loneliness,” Frontiers in Public Health, vol. 7, 2019. doi:10.3389/fpubh.2019.00247

22. J. V. Brocke, A. Hevner, and A. Maedche, “Introduction to Design Science Research,” Progress in IS Design Science Research. Cases, pp. 1–13, 2020.