ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

MODELO DE DETECCIÓN DE DISCURSO DE ODIO EN ECUADOR MEDIANTE CLASIFICACIÓN SUPERVISADA DE TWEETS Y TÉCNICAS DE NLP

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL GRADO DE MAGISTER EN SISTEMAS DE INFORMACIÓN

JOSÉ GABRIEL SUNTAXI RECALDE
jose.suntaxi@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 2022

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por José Gabriel Suntaxi Recalde, bajo mi supervisión.

Dra. Lorena Katherine Recalde Cerda
DIRECTORA DE PROYECTO

Dr. Edison Fernando Loza Aguirre
CODIRECTOR DE PROYECTO

DECLARACIÓN

Yo, José Gabriel Suntaxi Recalde, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada 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.

José Gabriel Suntaxi Recalde

DEDICATORIA

A Dios,

A mis padres,

A mis hermanos,

y a mis sobrinos

AGRADECIMIENTOS

A Dios por darme la fortaleza para culminar este proyecto. A mis padres y mis hermanos. A mi Directora y Codirector por inculcar la dedicación y excelencia en esta etapa académica.

CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

RESUMEN

Hate Speech en redes sociales se refiere al posteo de lenguaje ofensivo generalmente dirigido a grupos vulnerables de la sociedad. Frente a este problema, algunos gobiernos han planteado acciones legislativas para exigir a las plataformas de redes sociales acciones de control de contenido que puedan incitar al odio y la violencia. En Ecuador, no se han planteado acciones gubernamentales de regulación de contenido en redes sociales. Sobre esta base, en el presente estudio se desarrolla una primera aproximación al desarrollo de un framework teórico-practico para identificar el discurso de odio de los usuarios de la plataforma Twitter en Ecuador. Además, esta investigación propone generar evidencia empírica para el debate y discusión de proyectos de Ley que regulen el contenido ofensivo de post generados en Ecuador. Para esto, se propone la especificación de tres modelos de aprendizaje automático basados en técnicas de NLP. Los resultados mostraron que las técnicas de word embeddings se adaptan de mejor manera a la idiosincrasia lingüística de Ecuador, y por lo tanto, son más precisas en identificar Hate Speech con semántica propia de esta zona geográfica, mientras que la técnica de aprendizaje por transferencia del modelo preentrenado BERT (Bidirectional Encoder Representations from Transformers) se adapta de mejor manera a identificar Hate Speech en contextos lingüísticos generales.

Palabras Clave: Discurso de Odio, WordEmbeddings, Word2Vect, BERT, Aprendizaje por Transferencia, Modelo de Clasificación Supervisada.

ABSTRACT

Hate speech on social networks refers to the posting of offensive language, generally, directed at vulnerable groups in society. In this context, some governments have proposed legislative actions to require social media platforms to control content that may incite hatred and violence. In Ecuador, no governmental actions have been proposed to regulate content on social networks. On this basis, the present study develops a first approach to the development of a theoretical-practical framework to identify the hate speech of users of the Twitter platform in Ecuador. In addition, this research aims to generate empirical evidence for the debate and discussion of legal artifacts that would regulate the offensive content of posts generated in Ecuador. For this, three machine learning models based on NLP techniques are used. The results showed that the word embedding techniques are better adapted to the linguistic idiosyncrasy of Ecuador, and therefore, are more accurate in identifying Hate Speech with semantics typical of this geographical area. The transfer learning with the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model is better suited to identify Hate Speech in general linguistic contexts.

Palabras Clave: Hate Speech, WordEmbeddings, Word2Vect, BERT, Transfer Learning, Supervised Classification Model.

1 INTRODUCCIÓN

1.1 PLANTEAMIENTO DEL PROBLEMA

En un mundo globalizado, la integración de los entornos económico, político, social y tecnológico han promovido cambios importantes en el desarrollo de las sociedades. En un contexto económico y político, Flores [1] menciona que la globalización ha fomentado principalmente el crecimiento de los flujos de comercio internacional y del capital financiero entre naciones mientras que, por otro lado, en relación con el componente social y tecnológico, se ha promovido una nueva estructura social basada en sociedades informáticas. García [2] a su vez señala que estas sociedades se caracterizan por la convergencia e interrelación de telecomunicaciones y computadoras que plantean nuevas configuraciones en las interacciones sociales. Dichas interacciones se caracterizan por un proceso de comunicación multidireccional en un espacio digital conocido como redes sociales [3].

Las redes sociales —coadyuvadas por el avance de internet— han impulsado un cambio importante en la cotidianidad de las personas, definiendo un nuevo espacio virtual para la difusión de información a gran escala y la comunicación de diversidad de opiniones (Web 2.0). Además, de los beneficios que ofrece, este paradigma de interacción social plantea también nuevos desafíos relacionados principalmente con la regulación de datos y control del contendido [4], [5].

En los últimos años, la literatura ha centrado su atención en identificar los componentes esenciales que subyacen al desafío del control de contenido. De esta manera, estudios como [6], [7] y [8] coinciden en mencionar que el entorno de anonimato de las redes sociales promueve la desinhibición y, en consecuencia, la divulgación de información falsa y comunicación de opiniones con lenguajes ofensivos. Con frecuencia, en el contexto de las redes sociales, la utilización de lenguaje ofensivo conduce —entre otras cosas— a una discusión despectiva que se denomina “Discurso de Odio” (Hate Speech, HS).

Dependiendo del contexto, el HS puede variar en su definición. No obstante, [9] sugiere considerar tres aspectos esenciales en su conceptualización. En primer lugar, el HS se dirige contra un individuo/grupo específico o fácilmente identificable basado en una característica arbitraria o irrelevante. Por otro lado, el HS estigmatiza al individuo/grupo destinatario atribuyéndole de manera explícita o implícita cualidades consideras no deseables; y, por último, el HS presenta al individuo/grupo destinatario como un objeto legítimo de hostilidad. Estas tres características se enmarcan en la definición considerada por las Naciones Unidas en su Estrategia y Plan de Acción sobre el HS [10]:

> "... se entiende como cualquier tipo de comunicación en el habla, la escritura o el comportamiento, que ataque o utilice un lenguaje peyorativo o discriminatorio con referencia a una persona o grupo en función de quiénes son, en otras palabras, según su religión, etnia, nacionalidad, raza, color, ascendencia, género u otro factor de identidad" [10].

La literatura que analiza el HS bajo enfoques sociales, legales y analíticos concuerda en que la característica de anonimato de las redes sociales en conjunción con el auge que han experimentado en los últimos años ha generado un impacto negativo en el desarrollo de las sociedades, profundizando la violación de derechos humanos en cuanto a la integridad, igualdad, seguridad y prohibición de discriminación. El HS lesiona la posición social y dignidad de las personas que forman parte de un grupo racial, religioso u otro vulnerable [11]. En este sentido, una sociedad en desarrollo debe garantizar como bien público el sentido de inclusión enfrentando de manera eficaz y eficiente la violación de derechos humanos en todos los canales en donde estos se manifiesten [11].

Por lo mencionado, la detección de HS se vuelve una tarea imperativa, más aún en un mundo donde cada día crecen las interacciones de personas en plataformas de redes sociales. Frente a esto, algunas de las principales plataformas como Facebook, YouTube y Twitter han determinado reglas y políticas enfocadas en prevenir y controlar las conductas de odio [12]. Sin embargo, existen dos limitaciones inherentes a este control. En primera instancia, el crecimiento exponencial de la cantidad de contenido hace imposible que las denuncias de HS en publicaciones, comentarios y mensajes sean revisadas por moderadores de contenido. Por otro lado, desde una perspectiva lingüística, el contenido en diferentes lenguajes puede tener distintas bases estructurales semánticas y pragmáticas lo que dificulta la detección de HS.

En este contexto, en el presente trabajo se propone el desarrollo de un modelo de clasificación supervisado de texto utilizando recursos de alto nivel del Procesamiento del Lenguaje Natural (NLP) a fin de identificar el HS en posts generados en Twitter en el Ecuador.

Para tal propósito, se plantea el diseño de un mecanismo integral de recolección de datos, preprocesamiento de texto, aplicación de técnicas de NLP y ejecución de algoritmos de clasificación supervisada, llegando hasta una fase de evaluación del esquema metodológico propuesto.

1.2 OBJETIVO GENERAL

Desarrollar un modelo de detección de HS mediante clasificación supervisada de tweets y utilización de recursos de alto nivel del Procesamiento del Lenguaje Natural.

1.3 OBJETIVOS ESPECÍFICOS

- ❒ Realizar una revisión integral de la literatura relacionada sobre al HS.
- ❒ Construir un corpus de tweets en español.
- ❒ Aplicar distintas técnicas de Word Embedding para la representación numérica de corpus recolectado.
- ❒ Aplicar distintos algoritmos de aprendizaje automático para entrenar y testear modelos de clasificación de texto.

1.4 MARCO TEÓRICO

1.4.1 Word Embeddings

Las técnicas de incrustación de palabras (Word Embeddings, WE) proporcionan representaciones vectoriales que preservan la similitud y semántica de las palabras [13]. De esta manera, WE asocia las palabras a un punto, definido por su vector de características, en un espacio vectorial [14]. Matemáticamente, WE es un mapeo de tipo:

$$V \to \mathbb{R}^D : w \mapsto \vec{w} \tag{1}$$

En la Ecuación 1 se observa que las palabras w de un vocabulario V se asignan a un vector de valor real $\overrightarrow{w}$ en el espacio dimensional D [15].

1.4.2 Word2Vec

Esta técnica se caracteriza por representar palabras en un espacio continuo de baja dimensión[1] llevando información semántica de las palabras [16]. Esta técnica utiliza dos arquitecturas distintas: CBOW (*continuous bag of words*) y Skip-Gram.

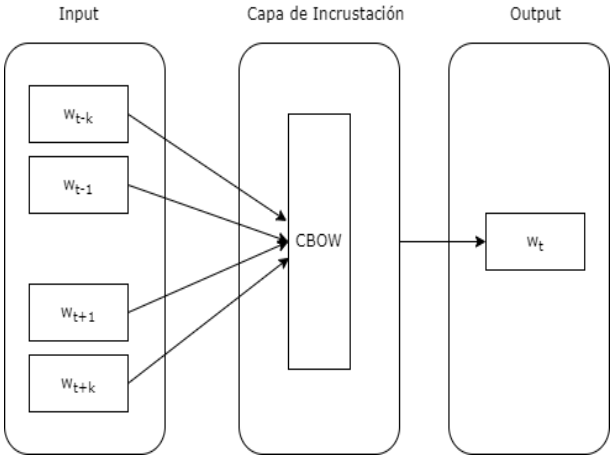
El objetivo de CBOW es predecir una palabra, utilizando las palabras a su alrededor [16] y [17] (Ver Figura 1.1a). Para esto, el modelo maximiza la probabilidad de que una palabra esté en un contexto específico. Por el contrario, Skip-Gram tiene por objetivo predecir el contexto utilizando una palabra central (Ver Figura 1.1b), esto se logra maximizando la probabilidad del contexto dado una palabra.

Ambos modelos están enfocados en el aprendizaje de palabras dado su contexto de uso local, donde el contexto está definido por una ventana de palabras vecinas. Esta ventana es un parámetro configurable del modelo. En este estudio se utilizará la arquitectura skip-gram para la evaluación del modelamiento de texto con Word2Vec.

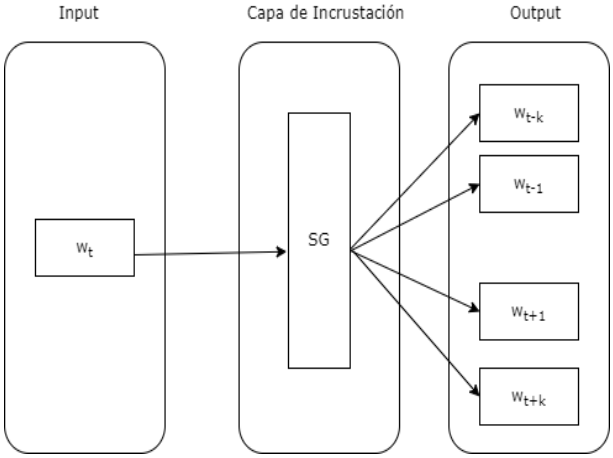
1.4.3 BERT

BERT se ha popularizado en los últimos años por su capacidad de proporcionar incrustaciones vectoriales de representación de lenguaje natural que permiten adaptarse a varias tareas del NLP (Análisis de Sentimientos, Equivalencia Semántica, Preguntas y Respuestas, etc). Es un modelo basado en redes neuronales con transformadores bidireccionales [18], [19]. La arquitectura de transformadores bidireccionales permite analizar el texto sin límite de una ventana como Word2Vec. Sobre esto, Word2Vec no captura la información semántica en el contexto integral de la oración dado que el algoritmo transforma las pala-

[1] En el contexto de NLP, un espacio de baja dimensionalidad codifica las relaciones naturales del lenguaje humano en representaciones matemáticas necesarias para entrenar varias tareas de NLP.



(a) Técnica CBOW



(b) Técnica Skip-Gram

Figura 1.1: Arquitecturas Word2Vec
Elaborado por: El Autor

bras en vectores estables que no se adaptan a contextos diferentes [18]. Por el contrario, BERT utiliza transformadores cuya arquitectura de aprendizaje está basada en procesos de *Self-attention*; proceso que permite emparejar cada palabra de la oración con todas las demás capturando una correlación contextual de cada palabra en toda la oración [20].

1.4.4 Aprendizaje por Transferencia

Este método aprovecha el conocimiento almacenado dentro de un dominio de origen y proporciona un mecanismo para transferir ese conocimiento a un dominio distinto [21]. En el contexto de NLP, el método supone, en primer lugar, pre-entrenar un modelo en un corpus extenso para posteriormente utilizarlo en tareas similares con corpus más pequeños.

1.5 REVISIÓN DE LA LITERATURA

La literatura que existe sobre el HS es extensa, abordando desafíos desde una perspectiva social, legal y analítica. Para [22], los tópicos de investigación de los últimos 30 años con relación al HS se agrupan en temas de: a) Debate general del HS frente a la libertad de expresión y b) Identificación del HS a través de estrategias de aprendizaje automático. El primer grupo de investigaciones tiene un enfoque teórico mientras que el segundo muestra evidencias empíricas asociadas a la aplicación de técnicas de minería de texto con base en NLP.

1.5.1 Evidencia Teórica

Las primeras investigaciones relacionadas al HS abordaron la problemática desde un enfoque teórico. De esta manera, los objetivos iniciales de la literatura relacionada se centraron en definir de manera pragmática su relación con la libertad de expresión. Estudios teóricos más actuales, profundizan en las consecuencias sociales del HS en una sociedad moderna.

Partiendo de lo mencionado, en [9] analizan el HS bajo el derecho moral de la libertad de expresión. Los autores mencionan que el HS debe analizarse a partir de la evaluación de los deberes morales (no amenazar, no ofender, no hostigar, no difamar) y su posible abstención. De la misma manera, en [23] mencionan que el HS no es libertad de expresión, por el contrario, promueve la opresión de grupos débiles (homosexuales, minorías raciales y sociales) que generalmente no tienen acceso a canales adecuados para expresar sus quejas. A fin de promover estrategias de protección para estos grupos en [24] y [25] enfatizan en la necesidad de promover leyes sobre el HS para prevenir consecuencias destructivas en la sociedad. Sobre este mismo debate, en [26] se menciona que los juicios adversos basados en afirmaciones de hecho o valor sobre un grupo racial o religioso no pueden identificarse como HS dado que se enmarcan en la protección de la libertad de expresión. En esta misma línea, [27] afirma que el HS es desagradable, sin embargo, en términos estrictos legales sigue siendo un discurso, no una conducta, y, por lo tanto, se ampara en el derecho de libertad de expresión.

Parte de la literatura teórica más actual examina la transición del HS hacia un nuevo paradigma de interacción social en el marco de la Web 2.0. En este sentido, para [28] y [29]

el HS en redes sociales brinda una aproximación muy precisa del deterioro progresivo de la sociedad en cuanto a la aceptación de la otredad de sus grupos vulnerables. Este argumento es esencial para entender la importancia del análisis del HS dado que sugiere que estos no son discursos aislados, sino que representan manifestaciones reales de conductas antisociales. Con respecto a esto, [30] en su estudio sobre los factores psicológicos del comportamiento de odio, menciona que las personas que declaran expresiones de odio en línea tienen fuertes características psicópatas, es decir, son personas con tendencias a cometer actos delictivos. En esta misma línea, se ha argumentado que existe una relación de causalidad directa entre algunos ataques terroristas y el historial de publicaciones de odio de las personas que materializaron dichos ataques [31].

En los últimos años se ha evidenciado una fuerte motivación para definir políticas, reglas y normas que permitan identificar y monitorear el HS en redes sociales. A tal efecto, en [12] mencionan que la UNESCO ha precisado acciones no legislativas para la regulación del contenido de odio en redes sociales como: i) que la sociedad civil proponga mecanismos de recolección de datos y metodologías de análisis para identificar el HS, y, que ii) los gobiernos y las organizaciones no gubernamentales exijan acciones de control a las compañías de internet que hospedan el contenido específico. En relación con el primer aspecto, en [29] se propone un sistema inteligente denominado “HateNet” que analiza la evolución del HS en Twitter. Los autores desarrollan una herramienta que extrae tweets en tiempo real y determina una probabilidad de que ciertos tweets se categoricen como HS. En la práctica, este sistema es utilizado por la Oficina Nacional contra los Crímenes de Odio de España para identificar tendencias de HS en eventos deportivos, políticos y días festivos. Con respecto a las acciones de control, en el año 2016 la Comisión de la Unión Europea (UE) en colaboración con las principales plataformas como Facebook, Twitter, YouTube y Microsoft elaboraron un código que establece procesos para que estas empresas revisen en el menor tiempo posible las notificaciones de conductas que incitan a la violencia y el odio.

En América Latina, el control de contenido ha tomado mayormente un enfoque legislativo a través de proyectos de ley que sancionan el HS. Por ejemplo, en Colombia, Perú, Venezuela y El Salvador existen proyectos que plantean sanciones pecuniarias y de privación de libertad para los delitos de odio en internet. En Ecuador, se ha propuesto en la Asamblea Nacional el proyecto Ley que Regula los Actos de Odio y Discriminación en Redes Sociales e Internet en el que se propone otorgar toda la responsabilidad del control a las plataformas de redes sociales [12].

1.5.2 Evidencia Empírica

En el marco del Big Data, el crecimiento a nivel mundial de las redes sociales ha potencializado a gran escala la generación de datos. Bajo esta premisa, en los últimos años se han desarrollado un sin número de algoritmos de aprendizaje automático en el campo de la minería de redes sociales que tienen por objetivo representar, analizar y extraer patrones relevantes de las interacciones entre usuarios. Los algoritmos conjugan diferentes herramientas de disciplinas como la informática, minería de datos, minería de texto, estadística, matemática, optimización, entre otras [32].

La literatura que analiza el HS en redes sociales utiliza técnicas de minería de texto en combinación con metodologías de NLP. La evidencia empírica sobre el tema muestra que la mayor parte de las investigaciones sobre el HS utilizan modelos de aprendizaje automático supervisados, tales como Regresión Logística, Naïve Bayes, Árboles de Decisión, Random Forest y Máquinas de Soporte Vectorial ([4], [31], [33], [34], [35]). Estos modelos, consideran representaciones de texto etiquetado para entrenar un clasificador capaz de detectar el HS. Otros trabajos, como [36] y [37], toman en consideración enfoques de algoritmos no supervisados. Estos modelos calculan, a través de medidas estadísticas, la relación y similitud semántica entre palabras relacionadas al HS en una oración [38]. En los últimos años se ha observado que la literatura en cuestión, ha utilizado, además, algoritmos de aprendizaje profundo basados en arquitecturas de Redes Neuronales Convolucionales, Recurrentes ([39], [40], [41], [42], [43]) y Transformadores con mecanismo de auto-atención [44].

Dado que en el HS se analizan datos de tipo texto, los modelos antes mencionados requieren, en primera instancia, representar el texto en forma de números para su posterior entrenamiento. Esta representación, según [45], es el paso más importante del proceso dado que en esta fase se definen las características semánticas propias de cada lenguaje. Existen varias técnicas para la extracción de características como: One-Hot-Encoding, BOW (Bag of Words), N-Gramas, TF-IDF (Term frequency – Inverse document frequency) ([46], [7], [47]) Word2Vec ([48], [49],[50], [51], [52], [53]), BERT ([40], [54], [31], [55], [56]) entre otros. Para estos últimos, la literatura ha prestado particular atención dado que se ha demostrado que presentan un mejor rendimiento en la clasificación de HS puesto que la técnica toma en consideración el contexto sobre el cual se desarrollan las palabras [4].

Se ha observado que la literatura de corte empírico para el idioma español es escasa. Gran

parte de la literatura bajo este contexto utiliza fuentes comunes de corpus etiquetados ([57], [54], [31],[49], [58], [56]). Esto se debe, en gran parte, a la falta de corpus etiquetados con muestras de HS en este idioma [29]. Más particularmente, al investigar sobre el HS en contenido generado por usuarios de Ecuador, se ha evidenciado que no existen contribuciones previas.

2 METODOLOGÍA

Con el avance del Big Data, la analítica de datos ha evolucionado a un dominio de análisis multidisciplinario en donde se plantean estrategias más sofisticadas para el análisis descriptivo, exploratorio y predictivo de distintos tipos de datos [58]. En el dominio de investigación del HS, este paradigma multidisciplinario ha permitido relacionar métodos de aprendizaje automático y profundo de texto con técnicas de NLP proponiendo soluciones innovadoras para analizar su impacto en el desarrollo de las sociedades [59]. Bajo este antecedente, la metodología de Investigación en Ciencias de Diseño (Design Science Research, DSR) brinda un marco conceptual exhaustivo para el diseño y construcción de dichas soluciones denominados en este contexto como artefactos (modelos, métodos, instancias, algoritmos, etc.).

La visión de diseño y construcción de artefactos de DSR es útil para las investigaciones que son intensivas en el uso de NLP [59], como es el caso del HS. Así mismo, en [60] afirman que DSR es una metodología adecuada en aquellas investigaciones en donde se pretende dar una primera solución efectiva a un problema con ausencia de propuestas prácticas manejables. La sugerencia de [60] se relaciona con lo mencionado previamente sobre la falta de contribuciones previas para el análisis del HS en Ecuador. Por lo tanto, dadas las características de este estudio, se utilizará la metodología DSR ya que su *framework* propone una solución efectiva al problema planteado. Por otro lado, dado que el objetivo de este trabajo se centra en la aplicación de múltiples técnicas de Minería de Texto, se propone conjugar DSR con la metodología CRISP-DM (CRoss Industry Standard Process for Data Mining). CRISP-DM es útil en actividades innovadoras, en este sentido, dado que los datos de tipo texto están en constante crecimiento (publicaciones, patentes, datos de redes sociales y diversos tipos de documentos), la innovación para su análisis es cada vez más necesaria [61].

Tabla 2.1: Acoplamiento de Metodologías

Elaborado por: El Autor

CRISP-DM	DSR	Fase
Comprensión del negocio	Identificar el problema y la motivación	Fase 1
Comprensión del negocio	Definir objetivos para la solución	Fase 1
Comprensión de los Datos	Diseño y desarrollo del Artefacto	Fase 2
Preparación de los datos	Diseño y desarrollo del Artefacto	Fase 3
Modelado	Diseño y desarrollo del Artefacto	Fase 4
Evaluación	Evaluación	Fase 4

En la Tabla 2.1 se muestra la estrategia de acoplamiento de las metodologías DSR y CRISP-DM consideradas en este estudio. De esta manera, se proponen 4 fases de desarrollo metodológico, estas son:

Fase 1 Comprensión del negocio: en esta fase se determinarán los objetivos y requerimientos del proyecto desde una perspectiva del negocio.

Fase 2 Comprensión de los datos: en esta fase se recolectarán los datos que se utilizarán en el proyecto y la familiarización con los mismos. En esta etapa es posible el surgimiento de las primeras hipótesis acerca de la información que podría estar oculta.

Fase 3 Preparación de los datos: en esta fase se tratarán y depurarán los datos para construir la vista minable o conjunto de datos final sobre el cual se aplicarán las técnicas de minería.

Fase 4 Modelado y Evaluación: en esta etapa se aplicarán los diversos algoritmos de aprendizaje automático, sobre el conjunto de datos de entrenamiento para evaluar su rendimiento, haciendo uso del conjunto de datos de validación y prueba.

2.1 FASE 1

En esta etapa se especifican los objetivos y requisitos del proyecto desde una perspectiva de negocio [62]. Considerando que este estudio es una propuesta de investigación, la perspectiva de negocio se abordará como un problema social en el marco de la Web 2.0.

2.1.1 Objetivos del Negocio

Con el auge de las redes sociales y la falta de control de contenido [5], el HS se ha establecido como una característica inherente de las nuevas formas de interacción social. Este trabajo propone elaborar un framework teórico-práctico que incluye un modelo de aprendizaje automático para identificar el HS de los usuarios de la plataforma Twitter en Ecuador.

2.1.2 Evaluación de la Situación Actual

En Ecuador no existen trabajos que formulen alguna estrategia de investigación que recopile, procese y analice datos sobre HS en redes sociales. Por lo tanto, el presente trabajo es el primero en caracterizar la semántica particular de esta zona geográfica y contribuir con evidencia empírica confiable que puede ser utilizado para la discusión y elaboración de marcos regulatorios eficientes y eficaces para hacer frente a las posibles consecuencias que genera el HS en la sociedad ecuatoriana.

2.1.3 Objetivos de la minería de datos

Desde la perspectiva del NLP, los objetivos son:

- Identificar el contenido de odio de un tweet generado en la zona geográfica de Ecuador.
- Identificar el mejor esquema de *Word Embedding* para la representación semántica de tweets en esta zona.
- Identificar el modelo de clasificación supervisada con el mejor ajuste.

2.2 FASE 2

En la misma línea de lo mencionado en [62], las tareas que se deben seguir en esta etapa son: a) Recolección Inicial de Datos, b) Descripción de Datos, c) Exploración de Datos, y c) Verificación de Calidad de los Datos.

En esta fase los ejemplos y porciones de análisis que muestran palabras o contenido de odio son parte de la naturaleza propia de la presente investigación. Así se exponen casos reales que han sido extraídos producto del proceso de análisis y recolección de datos utilizados para detección de HS.

2.2.1 Recolección de los Datos

El corpus inicial está compuesto de cuatro fuentes distintas de datos de texto. El primer conjunto de datos será utilizado para identificar características semánticas del corpus, mientras que los otros conjuntos se emplearán para entrenar el modelo de clasificación supervisado adaptado a la semántica de la zona geográfica del Ecuador.

Tabla 2.2: Recolección de Corpus
Elaborado por: El Autor

Corpus	Fuente	N° Tweets
1	API Twitter	166.850
2	HateEval 2019	5.000
3	Pereira-Kohatsu [49]	6.000
4	Charitidis [63]	470
	Total	178.320

Como se observa en la Tabla 2.2, el primer conjunto de datos se compone de 166 mil tweets aleatorios no etiquetados generados en la localidad de Ecuador durante los meses de junio a agosto de 2021. Estos tweets fueron recolectados utilizando la API de Twitter bajo el método de extracción en tiempo real o vía streamnig. En la Figura 2.1, se muestra la arquitectura de solución para la recolección de estos tweets. En el primer paso, se accede a la API de Twitter utilizando el protocolo *Streaming HTTP* para recibir datos a la medida que ocurren [64].

En el segundo paso, se definen funciones ETL desarrolladas en código Python para filtrar los tweets bajo las siguientes condiciones: a) seleccionar los tweets generados en la zona geográfica de Ecuador (Ver Anexo 6.1), b) seleccionar los tweets generados en idioma español y c) eliminar retweets. Además, en este paso, se generan funciones para interactuar con la base de datos PostgreSQL a través de comandos Python.

En el tercer paso, se almacenan los datos transmitidos por la API. Estos datos son tweets

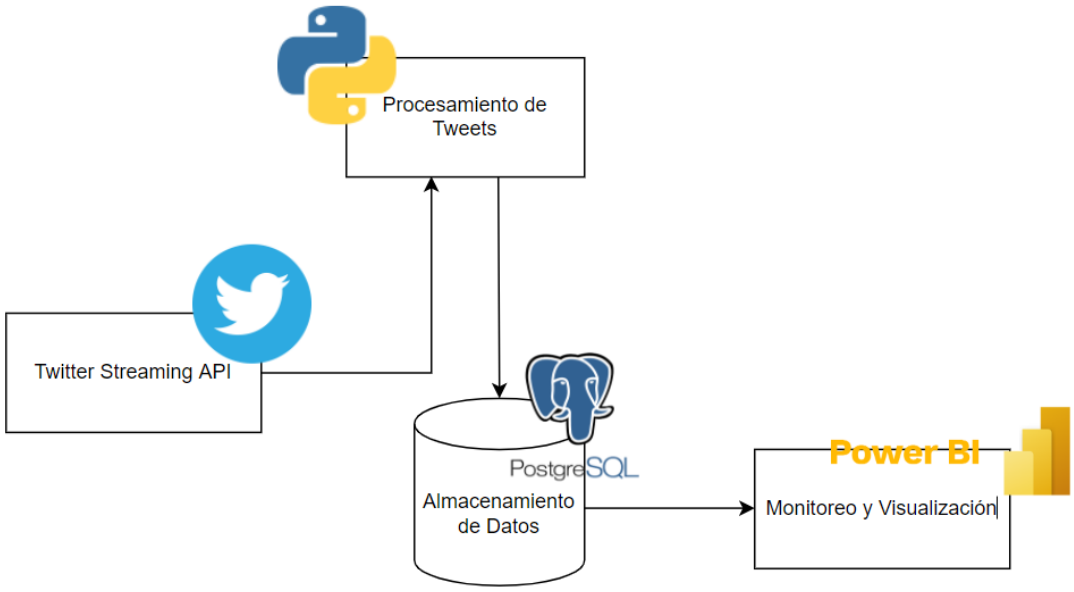


Figura 2.1: Flujo de recolección de Datos API Twitter
Elaborado por: El Autor

codificados en formatos JSON, los cuales son transformados para obtener una representación relacional de base de datos. En la Figura 2.2 se observa la estructura de la tabla "tweets", la cual describe: el identificador único, fecha de creación, lugar de creación, usuario y texto publicado. Las demás tablas especificadas para modelo de base de datos se describen en el Anexo 6.2

tweet_id [PK] text	tweet_created_at text	tweet_place_id text	tweet_user_id text	tweet text
1467934220207206420	2021-12-06 19:09:25	013cb38e7fe501ae	149542235	Primer partido de la final sub 16. @IDV_EC 2 @LDU_Oficial 2.
1467934366324174851	2021-12-06 19:10:00	009924a469d7ace1	1232440491381268485	Que sean 10
1467934405947760644	2021-12-06 19:10:09	009924a469d7ace1	1232440491381268485	Amén
1467935476443230220	2021-12-06 19:14:25	0177bc687a9ac697	317807823	@mcmenesess Que belleza.
1467935502129143819	2021-12-06 19:14:31	01c0324312d210f3	16530311	@iandrade82 Yo hasta ahora no veo, el bicho siempre me dió miedo

Figura 2.2: Almacenamiento en PostgreSQL Tabla "tweets"
Elaborado por: El Autor

Los siguientes conjuntos de datos mostrados en la Tabla 2.2 fueron tomados de la literatura que analiza el HS en redes sociales utilizando técnicas de modelización supervisada. En consecuencia, se trata de tweets que han sido previamente etiquetados bajo algún contexto de HS, como xenofobia, misoginia, racismo, entre otras.

El corpus HateEval 2019, se ha tomado de los talleres de evaluación de sistemas de análisis semántico computacional del año 2019 (SemEval 2019). Los talleres tienen como ob-

jetivo proporcionar soluciones efectivas a los problemas de análisis de texto, como: Análisis Semántico, Análisis de Sentimiento, Respuestas a Preguntas, etc. En particular, en este taller se plantea como tarea la detección multilingüe de HS contra inmigrantes y mujeres en Twitter. La tarea se especifica como un problema de clasificación de dos clases donde los sistemas deben predecir si un tweet en inglés o en español con un objetivo determinado (mujeres o inmigrantes) contiene odio o no [65].

El conjunto de datos se construyó con tweets publicados de julio a septiembre del 2019, considerando los siguientes mecanismos de recolección [66]:

1. Tweets dirigidos de víctimas potenciales de odio.
2. Tweets de haters previamente identificados.
3. Tweets recolectados bajo la estrategia de palabras clave relacionadas al HS.

Una vez recolectados los tweets, se procede a etiquetarlos bajo el mecanismo de *crowdsourcing* con colaboradores no capacitados. De este modo, se identifican tres tipos de etiquetas: Discurso de Odio, Grupo Objetivo y Agresividad [66]. En consideración del alcance de este estudio, se trabajará únicamente con la etiqueta que identifica HS.

El tercer corpus se ha tomado del trabajo de [49], quienes desarrollan un *framework* para la detección y monitoreo de HS en Twitter. Este corpus se construyó con tweets publicados de febrero a diciembre del 2017. El mecanismo de recolección es semejante al propuesto en este estudio, es decir, se utiliza el método de extracción en tiempo real definiendo la zona geográfica de España. Para este corpus se propone un esquema de etiquetado en dos pasos. En primer lugar, se definen diccionarios con palabras de odio absoluto y relativo[1], de esta manera, si un tweet contiene al menos una palabra de odio absoluto o relativo, se clasifica como tal, caso contrario se elimina del corpus. En el segundo paso, se validan las etiquetas por medio del voto mayoritario considerando el criterio experto de profesionales del campo de la psicología, derecho y criminología.

El cuarto corpus, proviene del trabajo de [58], en donde se analiza el HS contra periodistas en Twitter en varios lenguajes que incluyen al español. Este corpus se recolectó entre

[1] Se dice que una palabra contiene odio absoluto si la palabra expresa inequívocamente odio, independientemente de su contexto. Por otro lado, si el odio depende del contexto, se dice que la palabra contiene odio relativo, por ejemplo, la palabra negro [49].

octubre del 2018 y mayo de 2019. En el proceso de recolección se consideró una lista de 200 cuentas relacionadas al periodismo, de esta manera, se recopilan tweets de los feeds de cada una de estas cuentas. Al igual que en [49], en este trabajo el etiquetado se realiza en dos pasos. En primer lugar, se utiliza el enfoque de etiquetado basado en palabras clave que pueden expresar odio. En el segundo paso, [58] utiliza anotadores para cada lenguaje en análisis. Para asegurar la calidad de etiquetado se dispone de supervisores que validan la correcta asignación de etiquetas de HS.

2.2.2 Descripción de Datos

El corpus final para este trabajo esta compuesto por 178 mil tweets. De estos, 166 mil son tweets no etiquetados, fueron utilizados en la fase de representación semántica, y, alrededor de 11 mil tweets a ser utilizados en la etapa de entrenamiento del modelo supervisado. En la Tabla 2.3, se observa la distribución de datos, dependiendo de la clase, en cada uno de los corpus recolectados. En el corpus de HateEval 2019, se evidencia una distribución balanceada entre ambas clases. Por el contrario, para el corpus de [49], es evidente una mayor concentración en tweets que no contienen HS.

Tabla 2.3: Distribución de etiquetas de HS
Elaborado por: El Autor

Corpus	Fuente	Tweets de HS	Tweets de no HS
2	HateEval 2019	41,58 %	58,42 %
3	Pereira-Kohatsu [49]	26,11 %	73,88 %
4	Charitidis [63]	100,00 %	0,00 %

Por último, en la Tabla 2.3 se observa que el corpus de [63] se compone únicamente de tweets que contienen HS. Esto es debido a que la base proporcionada por los autores contiene únicamente los identificadores únicos de los tweets recolectados, por tanto, al momento de extraer el texto algunos de estos tweets han sido eliminados. Dada esta dificultad, se toma como estrategia rescatar la mayor parte de tweets que identifiquen HS dada esta fuente.

2.2.3 Exploración y Verificación de Calidad de Datos

En la Figura 2.3 se muestra un esquema de conteo de palabras con visualización de *word cloud*. Bajo la definición de [49], se observa que el conjunto de tweets de HS contiene en su mayoría palabras de odio absoluto, mientras que las palabras para el conjunto de tweets de no HS contienen palabras de odio relativo. Algunas palabras muestran altas frecuencias en ambos conjuntos, debido a que las estrategias de recolección de datos de [66] y [49] tienen un sesgo inicial de selección de tweets al utilizar palabras clave de odio o grupos que capturan odio. Por ejemplo, en la Tabla 2.4, se observa que algunas palabras clave de HS pueden presentarse en distintos contextos.

(a) Word Cloud Etiquetas Hate

(b) Word Cloud Etiquetas No Hate

Figura 2.3: Conteo de Palabras en el Corpus Etiquetado
Elaborado por: El Autor

Tabla 2.4: Ejemplos de tweets en distintos contextos
Elaborado por: El Autor

Etiqueta	Texto
No Hate	Intentar hacer humor negro absurdo en USA, que son muy mirados para eso.
Hate	Negros y árabes aquí en UK son una desgracia
No Hate	Stop imponer roles de género que no son ni roles de género, son putos juguetes
Hate	CNNEE Muéranse putos

Lo mencionado resalta el rol fundamental de contar con mecanismos de etiquetado validados por expertos o supervisores que analicen de manera integral el contexto de cada tweet.

Con la finalidad de asegurar la calidad de datos se realizó un análisis exhaustivo de las palabras que pueden capturar odio absoluto y que se encuentran etiquetadas como no HS. De este análisis, se evidenció que en la mayoría de los casos estas casuísticas respondían al contexto del tweet. Sin embargo, se identificaron tweets que bajo el contexto analizado debieron asignarse a tweets de HS (Ver Tabla 2.5)

Tabla 2.5: Ejemplos de tweets mal categorizados en HateEval 2019
Elaborado por: El Autor

Etiqueta	Texto
No Hate	Cállate la boca! Subnormal, que no tienes ni pXXX idea de fútbol
No Hate	eres un hijo de la gran pXXX subnormal cuando te vea te voy a meter tantos puñetazos...

Las palabras que describen odio absoluto se han enmascarado.

La mayoría de estos casos se observaron en el corpus de [66]. Esta inconsistencia podría estar sujeta al mecanismo de validación de etiquetas por *crowdsourcing* dado que utiliza colaboradores no capacitados. Este resultado puede generar un bajo rendimiento de los modelos propuestos de clasificación supervisada. Para atender a esta dificultad, en el corpus de [66] se descartan los tweets de No HS. Esta estrategia permitirá resolver dos aspectos importantes relacionados al conjunto de datos: a) Inconsistencias de etiquetado y b) balanceo de las clases de HS.

Por lo expuesto, en la Tabla 2.6 se muestra la distribución final de corpus.

Tabla 2.6: Distribución de etiquetas de HS
Elaborado por: El Autor

Corpus	N° Tweets	Tweets de HS	Tweets de no HS
API Twitter	166.850	0	0
HateEval 2019	2.079	2.079	0
Pereira-Kohatsu [49]	6.000	1.567	4.433
Charitidis [63]	470	470	0
Total	175.399	4.116	4.433

2.3 FASE 3

En los proyectos de *machine learning* la preparación de datos es una de las etapas más importantes ya que permite mejorar la calidad de datos y en consecuencia potenciará el rendimiento de los modelos matemáticos [67] . Para [68], en los proyectos de NLP en redes sociales esta tarea tiene una mayor relevancia dado que a menudo los datos de tipo texto son incompletos e inconsistentes. En particular, los tweets suelen ser datos que presentan mucho ruido dado que carecen de normas estrictas ortográficas, gramaticales y de sintaxis que limitan la interpretación adecuada el lenguaje humano por parte de las máquinas [58]. Con la finalidad de atender a estas características inherentes de los datos, en esta etapa se realizarán las siguientes tareas: a) Selección de Datos, b) Limpieza de Datos y c) Construcción de Datos.

2.3.1 Selección de los datos

Partiendo de los expuesto anteriormente sobre la calidad de datos (inconsistencia de etiquetado y balanceo de clases), los tweets seleccionados son en total 175.399, de estos 166.850 corresponden a tweets no etiquetados, y 8.549 etiquetados. De estos últimos, el 48 % son tweets con características de HS, el 52 % de No HS. De esta manera, se asegura una estabilidad de clases que evitará el posible sesgo en la predicción del modelo propuesto de clasificación. En el corpus construido se encontraron 30 tweets duplicados cuyos registros fueron eliminados, llegando a un total de 175.369 datos.

Como se mencionó en el Apartado 2.2.1, la extracción de tweets a través de la API se realizó considerando un filtro de localización, en este caso, para la zona geográfica de Ecuador. No obstante, el uso de este filtro captura algunas zonas de países como Colombia y Perú (Ver Anexo 6.1).

En la Tabla 2.7 se observa que el filtro aplicado captura 18.278 tweets de Colombia, Perú y otros que no muestran información de su localización. Por otro lado, los datos de HateEval 2019 y Charitidis. no especifican la zona geográfica de extracción, al contrario de Pereira-Kohatsu. quienes mencionan que su extracción corresponde a datos generados en España.

Tabla 2.7: Distribución de Datos por País
Elaborado por: El Autor

Corpus	Ecuador	Perú	Colombia	España	Sin Información	Total
API Twitter	148.542	8.925	9.320		33	166.820
HateEval 2019					2.079	2.079
Pereira-Kohatsu et al.				6.000		6.000
Charitidis et al.					470	470
Total	148.542	8.925	9.320	6.000	2.582	175.369

En consecuencia, y como nuestro interés reside en el discurso observado en Ecuador, se descartan los tweets generados en Colombia y Perú (9.320 y 8.925, respectivamente). De esta manera, el número de tweets para el análisis es 157.124. Además, se consideran para el análisis únicamente las columnas: id tweet, texto y etiqueta, como se muestra en la Tabla 2.8.

Tabla 2.8: Columnas seleccionadas para el Análisis
Elaborado por: El Autor

id tweet	Texto	Etiqueta
1137231598720602112	JuandeL17964009 CNNEE Que te mueras enfermo	Hate
827888748113383000	Pero es que tú eres la oveja negra y tu solito te lías	No Hate
1401748215305840000	earth cee_explorer Qué belleza!E8	NULL
1401747967149752323	A veces es de sabios callar, aceptar y sonreir.	NULL

Los tweets etiquetados como: *"NULL"* corresponden a los datos del corpus 1, que –según lo mencionado– serán utilizados en la etapa de caracterización vectorial del corpus agregado, sin embargo, dado que carecen de una etiqueta, se eliminarán en la etapa de construcción del modelo de clasificación supervisada.

2.3.2 Limpieza de Datos

En esta etapa se seguirá el proceso descrito por [69], [70] y [71], compuesto por los siguientes pasos: a) Remover Caracteres Especiales y b) Remover *Stop Words*.

1. *Remover Caracteres Especiales*: En este paso se eliminan caracteres especiales, signos de puntuación, símbolos matemáticos, números, componentes alfanuméricos,

URLs y todos aquellos elementos que no son parte de la gramática del idioma español y que pueden afectar el rendimiento de los Modelos de Clasificación. Adicional, en este paso todo el texto se convierte en letras minúsculas y se eliminan emojis. Para esto se utilizó el servicio de procesamiento de texto *re*[2] de Python especializado en operaciones de expresiones regulares.

2. *Remover Stop Words*: Las *Stop Words* son palabras que no contribuyen de manera importante al significado de una oración, sino que contribuyen a mantener su estructura gramatical. Este paso ayuda a reducir la dimensionalidad del vector de características [69] –etapa que se abordará más adelante–. Las *Stop Words* para español se obtuvieron de la librería NLTK de Python.

2.3.3 Construcción de Datos

Por su cuenta, los datos de tipo texto no pueden ser utilizados directamente en los problemas de NLP. Para esto, los datos deben tener una representación numérica que permita a los modelos matemáticos interpretar de la mejor manera las características semánticas de un idioma. Dicha representación numérica se puede realizar a través de técnicas de incrustación de palabras (Words Embeddings, WE). Existen varias técnicas de WE de representación continua o discreta. En el primer caso, se pueden mencionar, por ejemplo: *TF-IDF (Term frequency – Inverse document frequency)*, *GloVe (Global Vectors)*, *Word2Vec* y *BERT (Bidirectional Encoder Representations from Transformers)*. Estas últimas utilizan distintos enfoques de arquitecturas de modelos de redes neuronales. Por otro lado, en el caso discreto se puede mencionar a *One-Hot-Encoding*. Según [72] la representación continúa es más efectiva para capturar palabras con semántica parecidas. En este trabajo, se evaluará el desempeño de las técnicas de *Word2Vec* y *BERT* las cuales han sido mayormente utilizadas por la literatura del HS.

En la Figura 2.4 se muestra la representación vectorial de la aplicación de WE bajo el método Word2Vec. Como se observa, existe cierta diferenciación para el clúster de tweets categorizados como Hate.

De la misma manera, en la Figura 2.5 se muestra el resultado de aplicar WE bajo el mé-

[2] https://docs.python.org/3/library/re.html

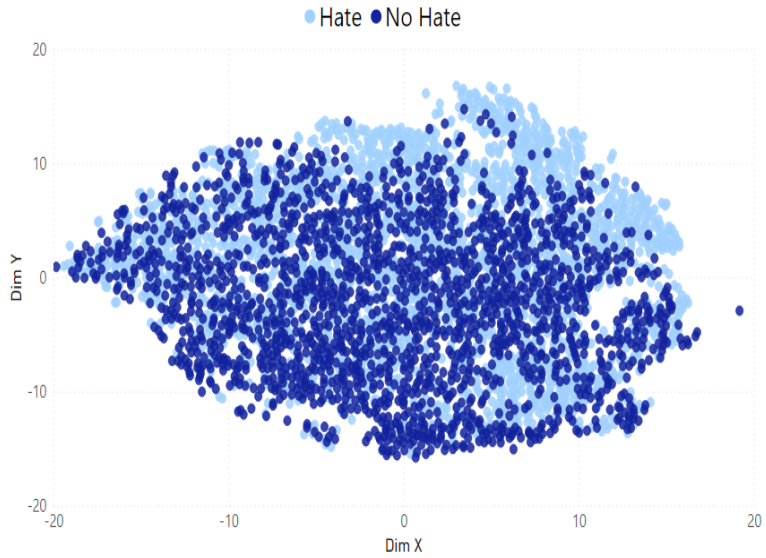


Figura 2.4: Visualización 2D de tweets modelados por Word2Vec
Elaborado por: El Autor

todo BERT. A diferencia del anterior método, las representaciones vectoriales de tweets se muestran menos dispersas y con una mejor discriminación entre clusters, por lo tanto, la representación vectorial sugiere un mejor desempeño de los modelo de clasificación supervisada a ser utilizados. Esta premisa se verificará en el siguiente capítulo.

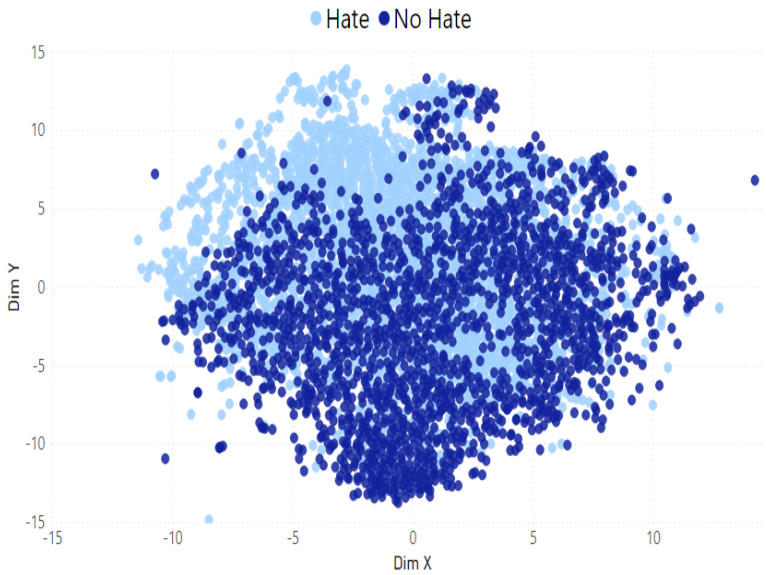


Figura 2.5: Visualización 2D de tweets modelados por BERT
Elaborado por: El Autor

2.4 FASE 4

Una vez identificadas las técnicas de representación vectorial de texto, se procede a esquematizar el proceso de modelamiento. En primer lugar, se describe el proceso de representación vectorial de tweets en el espacio generado por Word2Vec. Teniendo en cuenta que los tweets han sido tratados según lo descrito en 2.3.2, se procede con su representación vectorial. Para esto, en primer lugar, se *tokeniza*[3] cada tweet del corpus final (Ver Figura 2.6). A continuación se entrena el modelo Word2Vec para obtener los vectores de cada palabra en el corpus, como se muestra en la Tabla 2.9.

```
[['historia', 'volveré', 'leer', 'mil', 'veces'],
 ['acaba', 'publicar', 'foto', 'quito', 'capital', 'ecuador'],
 ['perdimos',
  'constitución',
  'orden',
  'social',
  'gobernabilidad',
  'vamos',
  'puertas',
  'comunismo',
  'importa',
  'casen',
  'hombres',
  'mujeres',
  'sólo',
  'espero',
  'ver',
  'años',
  'haber',
  'resulta',
  'dentro',
  'lavadora',
  'social'],
 ['veces', 'sabios', 'callar', 'aceptar', 'sonreir'],
 ['escapar', 'ciudad']]
```

Figura 2.6: Tokenización del Corpus
Elaborado por: El Autor

[3] La tokenización es la técnica de dividir textos grandes en tokens más pequeños. Los fragmentos de texto más grandes se pueden tokenizar en oraciones y las oraciones se pueden tokenizar en palabras. https://blogs.query.ai/natural-language-processing-dictionary

Tabla 2.9: Representación Vectorial de Palabras con Word2Vec

Elaborado por: El Autor

u_n	historia	volveré	leer	mil	veces
1	0,853485	0,5859549	0,11263777	-0,3989093	-0,23999493
2	-1,6030842	0,09246575	0,09172359	-0,08741009	-0,72399604
3	0,3425727	0,26447064	0,09143059	-0,43343273	0,11043115
...	...	...	...	...	...
100	0,6697398	0,02944793	-0,12824333	0,15479325	-0,16361938

La representación vectorial de cada tweet se realiza considerando las siguientes operaciones matriciales:

$$u + v = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ ... \\ u_{100} \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ ... \\ v_{100} \end{bmatrix} = \begin{bmatrix} u_1 + v_1 \\ u_1 + v_2 \\ u_1 + v_3 \\ ... + ... \\ u_{100} + v_{100} \end{bmatrix} \tag{2}$$

Donde u_1 y v_1 es la representación vectorial de los tokens generados para cada tweet del corpus. Para el ejemplo expuesto en la Tabla 2.9, la suma de cada token x_n retorna la representación vectorial del tweet (Ver Tabla 2.10).

Tabla 2.10: Representación Vectorial de Tweets con Word2Vec

Elaborado por: El Autor

u_i	historia	volveré	leer	mil	veces	**Tweet**
u_1	0.853485	0.5859549	0.11263777	-0.3989093	-0.23999493	**0.91317344**
u_2	-1.6030842	0.09246575	0.09172359	-0.08741009	-0.72399604	**-2.23030099**
u_3	0.3425727	0.26447064	0.09143059	-0.43343273	0.11043115	**0.37547235**
...	...	...	...	...	...	...
u_{100}	0.6697398	0.02944793	-0.12824333	0.15479325	-0.16361938	**0.56211827**

Con respecto a BERT, para este estudio se utilizará el modelo pre-entrenado en lenguaje español de [73]. Al contrario de Word2Vec, la representación vectorial de cada tweet está dado por el promedio de cada token. Esto produce un vector para cada tweet de dimensión 768 dado por el tamaño de la capa oculta (Ver Tabla 2.11)

Tabla 2.11: Representación Vectorial de tweets con BERT

Elaborado por: El Autor

u_i	$Tweet_1$	$Tweet_2$	$Tweet_3$	$Tweet_4$	...	$Tweet_j$
u_1	0.23897783	0.13833503	0.735587	0.5979923	0.18574499	-0.206442
u_2	0.16124845	0.04320721	0.52123153	-0.08741009	-0.72399604	-2.23030099
u_3	-0.48453024	0.52123153	-0.22190332	-0.43343273	0.11043115	0.37547235
...	...	...	...	...	...	...
u_{768}	0.02973371	-0.19997412	-0.42814505	-0.10513387	-0.46501443	-0.4491084

Una vez especificados los métodos de WE a ser utilizados, se propone entrenar y evaluar 11 algoritmos de aprendizaje automático. Esta estrategia de modelamiento permitirá identificar qué modelo se ajusta de mejor manera a cada una de las representaciones vectoriales. De los 11 modelos a ser estimados, 4 se basan en algoritmos de *Gradient Boosted Machines*, estos son: GB (Gradient Boosting), XGBoost (eXtreme Gradient Boosting), LGBM (Light Gradient Boosting Machine), AdaBoost(Adaptive Boosting). Estos algoritmos se basan en el desarrollo de múltiples modelos estimados secuencialmente, así cada modelo estimado va corrigiendo el error cometido por los modelos anteriores proporcionando estimaciones más precisas de la variable objetivo [74]. Además, se estiman 3 algoritmos de SVM (Support Vector Machine) con especificación de kernel: lineal, polinomial y radial. Esta especificación permitirá identificar, de ser el caso, esquemas de clasificación no lineal. Además de estos algoritmos, se propone la estimación de técnicas clásicas como: Árboles de Decisión, Regresión logística, Bosques Aleatorios y KNN (K Nearest Neighbor).

Además de utilizar WE, en este estudio se propone utilizar el enfoque de aprendizaje por transferencia para la clasificación de tweets. Esta estrategia se toma en consideración de lo mencionado por [21], donde se enfatiza la importancia del método en escenarios donde no existen datos etiquetados suficientes para entrenar un modelo con precisión. En consecuencia, para esta tarea se utilizará el modelo BERT pre-entrenado de [73] para posteriormente entrenar un modelo de clasificación supervisada con el corpus etiquetado.

3 RESULTADOS

En este apartado se describen los resultados obtenidos una vez especificado y aplicado el diseño metodológico.

3.1 WORD2VEC

Como se mencionó, la primera etapa en la representación vectorial con Word2Vec consiste en entrenar el modelo con el corpus integrado de las 4 fuentes descritas. El modelo se entrenó utilizando el módulo *Word2vec* de la librería *gensim*[1]. Los hiperparámetros configurados se muestran en la Tabla 3.1.

Tabla 3.1: Representación Vectorial de Palabras con Word2Vec
Elaborado por: El Autor

Hiperparámetro	Descripción	Valor
vector_size	Dimensión del vector de palabras	100
window	Distancia entre la palabra actual y sus cercanas	5
min_count	Criterio de exclusión de palabras de baja frecuencia	10
seed	Semilla para generación de números aleatorios	593

Estos hiperparámetros se determinaron como los más adecuados después de evaluar la métrica *accuracy* del modelo de clasificación generado por su representación vectorial. Con la finalidad de evaluar el desempeño de Word2Vec, en la Figura 3.1 se muestra una representación de palabras en un espacio reducido[2]. Para esta construcción se identificaron como base dos diccionarios. El primero de estos, se obtuvo del sitio *HateBase*[3], el cual provee un conjunto de palabras que se relacionan con el HS en español. El segundo diccionario, contiene palabras no relacionadas al HS, estas se determinaron bajo un criterio

[1] https://pypi.org/project/gensim/

[2] La reducción de dimensionalidad se hizo con t-SNE de la librería *sklearn*

[3] https://hatebase.org/

propio definiendo palabras que pueden expresan sentimientos positivos. Una vez especificados los diccionarios se calculó las palabras con su mayor similaridad.

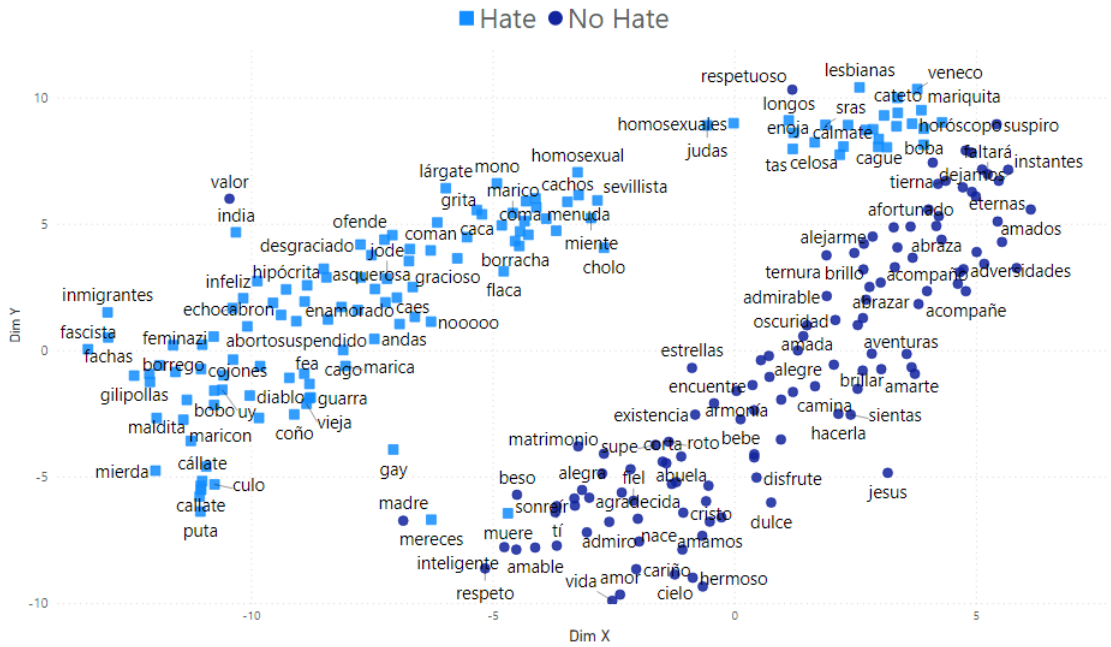


Figura 3.1: Similaridad de Palabras con Word2Vec
Elaborado por: El Autor

Como se observa, el modelo genera representaciones vectoriales mutuamente excluyentes para cada grupo de etiquetas: Hate y No Hate.

Una vez que se ha identificado que la representación vectorial de Word2Vec tiene un buen desempeño, se procede a estimar el modelo de clasificación supervisado. El conjunto de datos de entrada en cada modelo tiene 100 variables y 8.542 filas o tweets. Como se mencionó, en este conjunto la variable objetivo tiene una distribución balanceada.

Tabla 3.2: Modelos de Clasificación Supervisada con WE Word2Vec
Elaborado por: El Autor

Modelo	accuracy	precision	recall	specificity	roc_auc
XGB	0,71	0,69	0,64	0,76	0,70
Bosques Aleatorios	0,71	0,73	0,59	0,81	0,70
KNN	0,68	0,65	0,65	0,70	0,68
LGBM	0,70	0,70	0,61	0,78	0,69
Gradient Boosting	0,70	0,70	0,59	0,79	0,69
SVM (RBF)	0,70	0,71	0,58	0,80	0,69
Regresión Logística	0,69	0,69	0,58	0,78	0,68
Árboles de Decisión	0,65	0,61	0,64	0,65	0,65
SVM (Lineal)	0,69	0,72	0,53	0,83	0,68
AdaBoost	0,66	0,63	0,59	0,71	0,65
SVM (Polinomial)	0,66	0,82	0,32	0,94	0,63

En la tabla 3.2 se muestran los resultados de los algoritmos de clasificación ejecutados con Word2Vec. En cada modelo se realizó una partición del 80 % para entrenamiento y 20 % para validación o testeo. Considerando la métrica *accuracy* se observa que el mejor modelo es *XGB*. El algoritmo de *Bosques Aleatorios* tiene una métrica de *accuracy* similar, no obstante, la métrica *recall* es más baja. Esta métrica identifica la capacidad del modelo en identificar tweet de HS, por lo tanto, es relevante para evaluar el modelo con el mejor desempeño. Por lo expuesto, bajo una representación vectorial con Word2Vec el modelo *XGB* es el adecuado para identificar el HS en la zona geográfica del Ecuador.

3.2 BERT

En la Tabla 3.3 se muestran los resultados de los modelos de clasificación ejecutados con WE BERT. Siguiendo el mismo diseño de partición descrito, se observa que el algoritmo con el mejor desempeño es la *Regresión Logística* con un *accuracy* de $0,77$. Los algoritmos de *SVM* con sus especificaciones de kernel también muestran un buen desempeño. Sin embargo, al momento de evaluar la métrica *recall* carecen de un ajuste idóneo para identificar tweets de HS. Como se observa, BERT muestra un mejor desempeño en la clasificación de tweets que Word2Vec. Este resultado, va en la línea de lo mencionado por [18] y [19], en donde se reporta que BERT captura de manera integral el contexto de la oración. Por lo expuesto, bajo una representación vectorial con BERT el modelo de *Regresión Logística* es el adecuado para identificar el HS en la zona geográfica del Ecuador.

Tabla 3.3: Modelos de Clasificación Supervisada con WE BERT
Elaborado por: El Autor

Modelo	accuracy	precision	recall	specificity	roc_auc
Regresión Logística	0,77	0,76	0,73	0,81	0,77
SVM (Polinomial)	0,78	0,80	0,69	0,86	0,77
SVM (RBF)	0,78	0,79	0,69	0,85	0,77
SVM (Lineal)	0,75	0,74	0,71	0,79	0,75
LGBM	0,76	0,75	0,69	0,81	0,75
XGB	0,75	0,74	0,69	0,80	0,75
Gradient Boosting	0,75	0,74	0,69	0,80	0,74
Bosques Aleatorios	0,74	0,73	0,67	0,80	0,73
KNN	0,69	0,62	0,78	0,61	0,69
AdaBoost	0,71	0,69	0,65	0,75	0,70
Árbol de Decisión	0,61	0,57	0,59	0,63	0,61

3.2.1 Aprendizaje por Transferencia BERT

En la Tabla 3.4 se muestran las métricas utilizando el enfoque de aprendizaje por transferencia. Como se observa, las métricas son superiores a las técnicas propuestas de WE. Este resultado va en línea de lo mencionado por [8] y [21] quienes mencionan que el método tiene una mayor capacidad para adaptarse de manera dinámica al contexto en cada oración.

Tabla 3.4: Aprendizaje por Transferencia BERT
Elaborado por: El Autor

Modelo	accuracy	precision	recall	specificity	roc_auc
Modelo de Clasificación Supervisada	0,88	0,87	0,87	0,88	0,95

3.3 EVALUACIÓN

En la Tabla 3.5 se muestran ejemplos de tweets reales con su respectiva predicción bajo los tres modelos propuestos. Estos tweets se han escogido identificando aquellos que contengan palabras relacionadas a la idiosincrasia lingüística de Ecuador como: cholo, longo, mono y negro. Se observa que, bajo Word2Vec, los tweets que contienen la palabra **cholo** (tweets 1, 2 y 3) se categorizan como Hate, esto dado que el modelo identifica características de odio con dicha palabra independientemente del contexto. Por el contrario, el enfoque de aprendizaje por transferencia identifica que los tweets 1 y 3 no muestran características de odio, no obstante, para el tweet 2 el resultado es congruente al observado con Word2vec. Si bien los tweets 1 y 3 expresan matices ofensivos, el tweet 2 muestra un nivel más acentuado de HS. Este resultado puede sugerir la necesidad de incorporar en la detección del HS distintos niveles o categorías de odio para robustecer la precisión de los modelos de clasificación propuestos.

Los tweets (4 y 5) asociados a la palabra “longo” muestran características de odio absoluto. Estos tweets se combinan palabras de odio propias de la zona geográfica como: **hediondo** y “serrano”. En consecuencia, se observa que Word2vec es más preciso. Un resultado parecido se observa al evaluar los resultados para los tweets (6 y 7) con la palabra “mono”, en donde las técnicas de WE identifican de manera correcta ambos contextos. El enfoque de aprendizaje por transferencia no identifica de manera correcta el carácter ofensivo que

Tabla 3.5: Estimaciones Modelos Propuestos

Tweet	Texto	Word2Vec	BERT	Aprendizaje por Transferencia BERT
1	Jajajaja que cholo oe...	Hate	No Hate	No Hate
2	Quién mXXXXX les dijo que CC significa como cuando? cholos de a vXXXX	Hate	Hate	Hate
3	Más cholo no puede ser	Hate	Hate	No Hate
4	Fue lo que le recomendé a tu madre, longo hXXXXXXX a mote.	Hate	No Hate	No Hate
5	Nada le duele más a un serrano que lo traten de "longo". Es que hXXX, les arde.	Hate	No Hate	Hate
6	Mono batracio	Hate	Hate	No Hate
7	Compré en Mono verde con uber eats $1.90 el envío	No Hate	No Hate	No Hate
8	#Arruinaunacitacon4palabras Odio el chocolate negro;	Hate	No Hate	No Hate
9	Gracias mi Negro del alma..... #GraciasDonRodrigo	Hate	No Hate	No Hate
10	en cuanto se puede vender un riñón en el mercado negro buscar*	Hate	No Hate	No Hate
11	¿Que son muchos negros en una pared blanca? Un código de barras.	No Hate	No Hate	Hate

Las palabras que describen odio absoluto se han enmascarado.

refleja el tweet 6, esto puede estar sujeto a que el tweet recoge en su totalidad términos propios de la semántica ecuatoriana, que no necesariamente pueden ser parte del modelo pre-entrenado de BERT.

Los tweets 8 al 11 se han escogido con la finalidad de evaluar los modelos en contextos relativos. Para esto se identificaron tweets que contengan la palabra “negro”. Los resultados muestran que el modelo de clasificación determinado con Word2vec es menos efectivo al momento de identificar el odio en tweets que contienen palabras de odio en estos escenarios. Por el contrario, con los modelos BERT y aprendizaje por transferencia BERT se identifica de manera correcta la palabra **negro** en cada uno de los contextos. En particular, para el tweet 11 el modelo de aprendizaje por transferencia identifica el contexto racista en expresiones discriminatorias.

4 CONCLUSIONES

En el nuevo contexto de interacción social, a través de redes sociales, el análisis relacionado con la identificación del HS en línea se ha vuelto una tarea fundamental. Esta importancia parte del hecho de considerar al HS como un elemento que profundiza la violación de derechos humanos y propende en actos delictivos. Este último elemento ha tomado una mayor relevancia dado que, según la literatura expuesta, se ha evidenciado una correlación entre las expresiones de odio manifestadas en redes sociales y actos psicópatas y terroristas.

Para abordar esta problemática, en este estudio se planteó como estrategia utilizar DSR y CRISP-DM de manera combinada. Esta estrategia permitió definir un esquema metodológico estructurado en 4 fases, de esta manera, bajo DSR se identificó la base teórica para sustentar la utilización de herramientas y procesos relacionados al NLP, mientras que con CRISP-DM se orientaron y aplicaron las tareas relacionadas al tratamiento de datos de tipo texto y especificación de algoritmos de aprendizaje automático.

En este estudio se propusieron tres enfoques de modelización. De estos, dos toman como base representaciones vectoriales a través de WE como son Word2vec y BERT. La tercera de estas se basa en el enfoque de aprendizaje por transferencia de un modelo preentrenado con BERT. El estudio realizado con representaciones vectoriales muestra que BERT tiene una mayor capacidad para identificar el HS. Este resultado va de la mano con lo observado en la Figura 2.5 en donde se observa que las clases de la variable respuesta se muestran más dispersas permitiendo que los algoritmos especificados, incluso aquellos más simples y eficientes como la Regresión Logística, identifiquen de manera más precisa los tweets con componentes de HS. Por otro lado, con el enfoque de aprendizaje por transferencia se logró una mayor precisión en la clasificación, esto dado que BERT permite que las palabras se adapten de manera independiente al contexto de cada oración, al contrario de Word2Vec en donde las palabras tienen una única representación vectorial en todo el corpus analizado.

Si bien los resultados con BERT muestran tener un mejor rendimiento, es importante mencionar que no necesariamente estos resultados se adaptan a la idiosincrasia lingüística ecuatoriana. Las metodologías basadas en BERT toman como punto de partida un modelo pre-entrenado con un corpus recolectado en varias regiones. Por el contrario, Word2vec se entrena con un corpus generado en la zona geográfica del Ecuador permitiendo caracterizar la semántica de esta zona geográfica. Esto se evidencia en la Figura 3.1 en donde se observa que Word2vec identifica ciertas palabras de odio propias del contexto ecuatoriano como son: “cholo” y “mono”; dos palabras que generalmente capturan el odio en un problema de carácter social en Ecuador denominado “regionalismo”.

El análisis de evaluación en contextos semánticos de la zona geográfica del Ecuador mostró que Word2vec es más efectivo al momento de identificar tweets de odio con palabras propias de la idiosincrasia ecuatoriana. Por otro lado, las metodologías basadas en BERT son más efectivas en contextos que contienen palabras de odio relativo.

Por lo mencionado, próximos estudios pueden enfocarse en entrenar un modelo bajo una especificación BERT utilizando el corpus extraído en este estudio. De esta manera, se puede potencializar tanto el rendimiento del modelo como la caracterización semántica propia de la zona del Ecuador.

5 REFERENCIAS BIBLIOGRÁFICAS

[1] M. Victoria Flores, «Globalization as a political, economic and social phenomenon,» *ORBIS*, vol. 12, n.º 34, págs. 26-41, 2016.

[2] J. E. García, «Surgimiento de la sociedad de la información,» *Biblioteca universitaria*, vol. 4, n.º 2, págs. 77-86, 2001.

[3] A. Pantoja Chaves, «Los nuevos medios de comunicación social: las redes sociales,» 2011.

[4] H. Watanabe, M. Bouazizi y T. Ohtsuki, «Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection,» *IEEE access*, vol. 6, págs. 13 825-13 835, 2018.

[5] N. DePaula, K. J. Fietkiewicz, T. J. Froehlich, A. Million, I. Dorsch y A. Ilhan, «Challenges for social media: Misinformation, free speech, civic engagement, and data regulations,» *Proceedings of the Association for Information Science and Technology*, vol. 55, n.º 1, págs. 665-668, 2018.

[6] P. Burnap y M. L. Williams, «Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making,» *Policy & internet*, vol. 7, n.º 2, págs. 223-242, 2015.

[7] A. Gaydhani, V. Doma, S. Kendre y L. Bhagwat, «Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach,» *arXiv preprint arXiv:1809.08651*, 2018.

[8] M. Mozafari, R. Farahbakhsh y N. Crespi, «A BERT-based transfer learning approach for hate speech detection in online social media,» en *International Conference on Complex Networks and Their Applications*, Springer, 2019, págs. 928-940.

[9] J. W. Howard, «Free speech and hate speech,» *Annual Review of Political Science*, vol. 22, págs. 93-109, 2019.

[10] A. Guterres y col., «United Nations strategy and plan of action on hate speech,» *Tomado de: https://www. un. org/en/genocideprevention/documents/U*, n.º 20Strategy, 2019.

[11] E. Barendt, «What is the harm of hate speech?» *Ethical Theory and Moral Practice*, vol. 22, n.º 3, págs. 539-553, 2019.

[12] M. Díaz, «Discurso de Odio en América Latina,» *Revista Derechos Digitales*, 2020.

[13] S. Ghannay, B. Favre, Y. Esteve y N. Camelin, «Word embedding evaluation and combination,» en *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, págs. 300-305.

[14] Y. Bengio, R. Ducharme y P. Vincent, «A neural probabilistic language model,» *Advances in Neural Information Processing Systems*, vol. 13, 2000.

[15] T. Schnabel, I. Labutov, D. Mimno y T. Joachims, «Evaluation methods for unsupervised word embeddings,» en *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, págs. 298-307.

[16] M. G. Ozsoy, «From word embeddings to item recommendation,» *arXiv preprint arXiv:1601.01356*, 2016.

[17] B. Wang, A. Wang, F. Chen, Y. Wang y C.-C. J. Kuo, «Evaluating word embedding models: Methods and experimental results,» *APSIPA transactions on signal and information processing*, vol. 8, 2019.

[18] Z. Gao, A. Feng, X. Song y X. Wu, «Target-dependent sentiment classification with BERT,» *Ieee Access*, vol. 7, págs. 154 290-154 299, 2019.

[19] M. Moradshahi, H. Palangi, M. S. Lam, P. Smolensky y J. Gao, «HUBERT untangles BERT to improve transfer across NLP tasks,» *arXiv preprint arXiv:1910.12647*, 2019.

[20] A. Vaswani, N. Shazeer, N. Parmar y col., «Attention is all you need,» *Advances in neural information processing systems*, vol. 30, 2017.

[21] T. Semwal, P. Yenigalla, G. Mathur y S. B. Nair, «A practitioners'guide to transfer learning for text classification using convolutional neural networks,» en *Proceedings of the 2018 SIAM international conference on data mining*, SIAM, 2018, págs. 513-521.

[22] A. Tontodimamma, E. Nissi, A. Sarra y L. Fontanella, «Thirty years of research into hate speech: topics of interest and their evolution,» *Scientometrics*, vol. 126, n.º 1, págs. 157-179, 2021.

[23] O. Bakircioglu, «Freedom of expression and hate speech,» *Tulsa J. Comp. & Int'l L.*, vol. 16, pág. 1, 2008.

[24] T. M. Massaro, «Equality and freedom of expression: The hate speech dilemma,» *Wm. & Mary L. Rev.*, vol. 32, pág. 211, 1990.

[25] J. Walker, *Hate speech and freedom of expression: Legal boundaries in Canada*, 2018.

[26] C. Yong, «Does freedom of speech include hate speech?» *Res Publica*, vol. 17, n.º 4, págs. 385-403, 2011.

[27] J. Waldron, «The harm in hate speech,» en *The Harm in Hate Speech*, Harvard University Press, 2012.

[28] N. Chetty y S. Alathur, «Hate speech review in the context of online social networks,» *Aggression and violent behavior*, vol. 40, págs. 108-118, 2018.

[29] C. A. Calderón, D. Blanco-Herrero y M. B. V. Apolo, «Rejection and hate speech in Twitter: Content analysis of Tweets about migrants and refugees in Spanish,» *Revista Española de Investigaciones Sociológicas (REIS)*, vol. 172, n.º 172, págs. 21-56, 2020.

[30] P. Sorokowski, M. Kowal, P. Zdybek y A. Oleszkiewicz, «Are online haters psychopaths? Psychological predictors of online hating behavior,» *Frontiers in psychology*, vol. 11, pág. 553, 2020.

[31] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian y O. Frieder, «Hate speech detection: Challenges and solutions,» *PloS one*, vol. 14, n.º 8, e0221152, 2019.

[32] R. Zafarani, M. A. Abbasi y H. Liu, *Social media mining: an introduction*. Cambridge University Press, 2014.

[33] D. Elisabeth, I. Budi y M. O. Ibrohim, «Hate Code Detection in Indonesian Tweets using Machine Learning Approach: A Dataset and Preliminary Study,» en *2020 8th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 2020, págs. 1-6.

[34] B. Raufi e I. Xhaferri, «Application of machine learning techniques for hate speech detection in mobile applications,» en *2018 International Conference on Information Technologies (InfoTech)*, IEEE, 2018, págs. 1-4.

[35] H. Rathpisey y T. B. Adji, «Handling imbalance issue in hate speech classification using sampling-based methods,» en *2019 5th International Conference on Science in Information Technology (ICSITech)*, IEEE, 2019, págs. 193-198.

[36] N. Ruwandika y A. Weerasinghe, «Identification of hate speech in social media,» en *2018 18th international conference on advances in ICT for emerging regions (ICTer)*, IEEE, 2018, págs. 273-278.

[37] C. Udanor y C. C. Anyanwu, «Combating the challenges of social media hate speech in a polarized society: A Twitter ego lexalytics approach,» *Data Technologies and Applications*, 2019.

[38] M. Hajar y col., «Using YouTube comments for text-based emotion recognition,» *Procedia Computer Science*, vol. 83, págs. 292-299, 2016.

[39] A. Ribeiro y N. Silva, «INF-HatEval at SemEval-2019 Task 5: Convolutional neural networks for hate speech detection against women and immigrants on twitter,» en *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, págs. 420-425.

[40] H. Faris, I. Aljarah, M. Habib y P. A. Castillo, «Hate Speech Detection using Word Embedding and Deep Learning in the Arabic Language Context.,» en *ICPRAM*, 2020, págs. 453-460.

[41] C. Michele, S. Menini, A. Pinar y col., «Inriafbk at germeval 2018: Identifying offensive tweets using recurrent neural networks,» en *GermEval 2018*, 2018, págs. 80-84.

[42] A. Chaudhari, A. Parseja y A. Patyal, «CNN based hate-o-meter: a hate speech detecting tool,» en *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, 2020, págs. 940-944.

[43] G. K. Pitsilis, H. Ramampiaro y H. Langseth, «Effective hate-speech detection in Twitter data using recurrent neural networks,» *Applied Intelligence*, vol. 48, n.º 12, págs. 4730-4742, 2018.

[44] B. Amrutha y K. Bindu, «Detecting hate speech in tweets using different deep neural network architectures,» en *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, 2019, págs. 923-926.

[45] F. E. Ayo, O. Folorunso, F. T. Ibharalu e I. A. Osinuga, «Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions,» *Computer Science Review*, vol. 38, pág. 100 311, 2020.

[46] G. Koushik, K. Rajeswari y S. K. Muthusamy, «Automated hate speech detection on Twitter,» en *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, IEEE, 2019, págs. 1-4.

[47] O. Oriola y E. Kotzé, «Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets,» *IEEE Access*, vol. 8, págs. 21 496-21 509, 2020.

[48] I. G. M. Putra y D. Nurjanah, «Hate Speech Detection In Indonesian Language Instagram,» en *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, IEEE, 2020, págs. 413-420.

[49] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore y M. Camacho-Collados, «Detecting and monitoring hate speech in Twitter,» *Sensors*, vol. 19, n.º 21, pág. 4654, 2019.

[50] N. Sevani, I. A. Soenandi, J. Wijaya y col., «Detection of Hate Speech by Employing Support Vector Machine with Word2Vec Model,» en *2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, IEEE, 2021, págs. 1-5.

[51] Z. Mossie y J.-H. Wang, «Vulnerable community identification using hate speech detection on social media,» *Information Processing & Management*, vol. 57, n.º 3, pág. 102 087, 2020.

[52] M. P. K. Dewi y E. B. Setiawan, «Feature Expansion Using Word2vec for Hate Speech Detection on Indonesian Twitter with Classification Using SVM and Random Forest,» *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, n.º 2, págs. 979-988, 2022.

[53] A. R. Isnain, A. Sihabuddin e Y. Suyanto, «Bidirectional long short term memory method and Word2vec extraction approach for hate speech detection,» *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 14, n.º 2, págs. 169-178, 2020.

[54] J. A. Gonzalez, L.-F. Hurtado y F. Pla, «TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter,» *Neurocomputing*, vol. 426, págs. 58-69, 2021.

[55] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Urena-López y M. T. Martín-Valdivia, «Comparing pre-trained language models for Spanish hate speech detection,» *Expert Systems with Applications*, vol. 166, pág. 114 120, 2021.

[56] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz y L. Plaza, «Automatic classification of sexism in social networks: An empirical study on twitter data,» *IEEE Access*, vol. 8, págs. 219 563-219 576, 2020.

[57] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios y R. Valencia-García, «Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings,» *Future Generation Computer Systems*, vol. 114, págs. 506-518, 2021.

[58] P. Charitidis, S. Doropoulos, S. Vologiannidis, I. Papastergiou y S. Karakeva, «Towards countering hate speech against journalists on social media,» *Online Social Networks and Media*, vol. 17, págs. 1-10, 2020.

[59] S. O'Riain, E. Curry y P. Buitelaar, «Engaging Practitioners within Design Science Research: A Natural Language Processing Case Study,» en *European Design Science Symposium*, Springer, 2012, págs. 155-169.

[60] M. T. Mullarkey, A. R. Hevner, T. Grandon Gill y K. Dutta, «Citizen data scientist: A design science research method for the conduct of data science projects,» en *International conference on design science research in information systems and technology*, Springer, 2019, págs. 191-205.

[61] W. Y. Ayele, «Adapting CRISP-DM for idea mining: a data mining process for generating ideas using a textual dataset,» *International Journal of Advanced Computer Sciences and Applications*, vol. 11, n.º 6, págs. 20-32, 2020.

[62] P. C. NCR, J. Clinton, R. K. NCR y col., «CRISP-DM 1.0,» 1999.

[63] A. Elragal y M. Haddara, «Design science research: Evaluation in the lens of big data analytics,» *Systems*, vol. 7, n.º 2, pág. 27, 2019.

[64] T. D. Documentation, «API Overview,» *URl: https://dev. twitter. com/overview/api (visited on 08/18/2017)*,

[65] Competitions.codalab.org, «SemEval 2019 Task 5 - Shared Task on Multilingual Detection of Hate,» *https://competiciones.codalab.org/competiciones/19935 (visitado el 08/18/2021)*,

[66] V. Basile, «Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection,» en *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, págs. 460-463.

[67] J. Tang, H. Li, Y. Cao y Z. Tang, «Email data cleaning,» en *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, págs. 489-498.

[68] Z. Zainol, M. T. Jaymes y P. N. Nohuddin, «Visualurtext: a text analytics tool for unstructured textual data,» en *Journal of Physics: Conference Series*, IOP Publishing, vol. 1018, 2018, pág. 012 011.

[69] H. Sandaruwan, S. Lorensuhewa y M. Kalyani, «Sinhala hate speech detection in social media using text mining and machine learning,» en *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, IEEE, vol. 250, 2019, págs. 1-8.

[70] G. B. Herwanto, A. M. Ningtyas, K. E. Nugraha e I. N. P. Trisna, «Hate speech and abusive language classification using fastText,» en *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, 2019, págs. 69-72.

[71] B. Pahwa, S. Taruna y N. Kasliwal, «Sentiment analysis-strategy for text preprocessing,» *International Journal of Computers and Applications*, vol. 180, n.º 34, págs. 15-18, 2018.

[72] J. E. Font y M. R. Costa-Jussa, «Equalizing gender biases in neural machine translation with word embeddings techniques,» *arXiv preprint arXiv:1901.03116*, 2019.

[73] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang y J. Pérez, «Spanish Pre-Trained BERT Model and Evaluation Data,» 2020.

[74] Y. Zhang y A. Haghani, «A gradient boosting method to improve travel time prediction,» *Transportation Research Part C: Emerging Technologies*, vol. 58, págs. 308-324, 2015.

6 ANEXOS

6.1 ANEXO: ESPECIFICACIÓN ZONA GEOGRÁFICA ECUADOR.

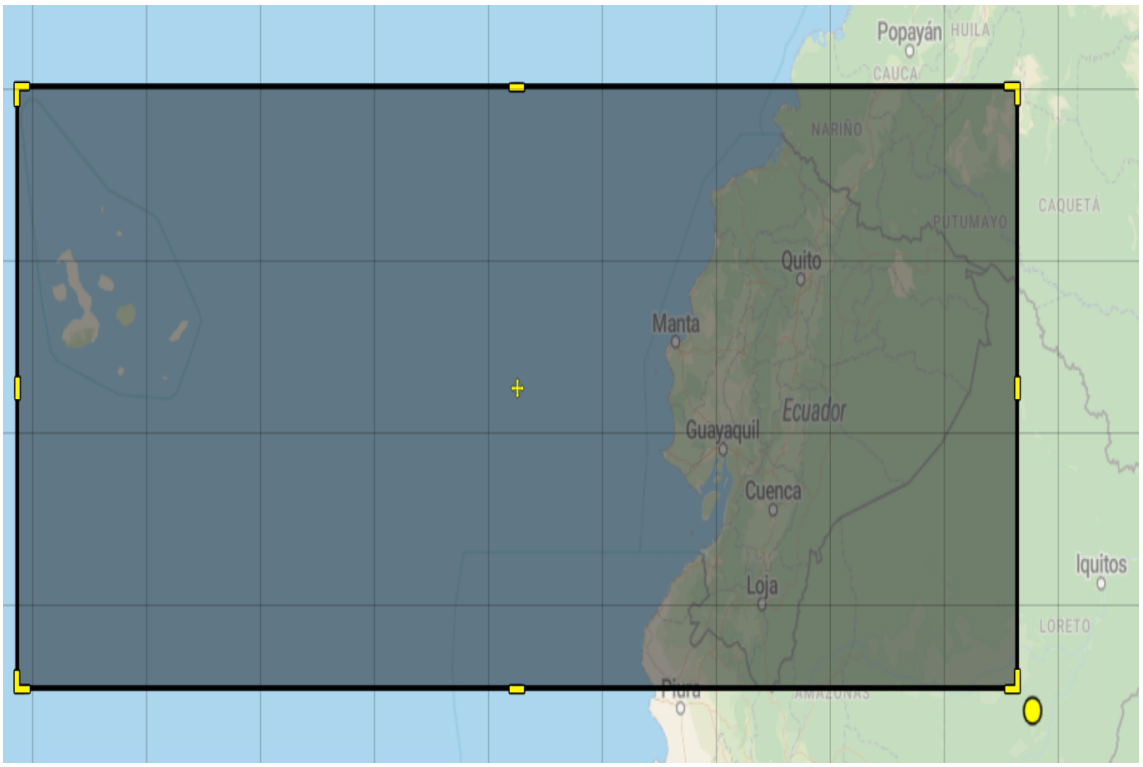


Figura 6.1: Especificación de la Región Geográfica Ecuador
Elaborado por: El Autor

6.2 ESQUEMA DE TABLAS EN EL MODELO DE BASE DE DATOS DE POSTGRESQL.

user_id [PK] text	user_name text	user_screen_name text	user_location text	user_description text
149542235	Fernando (Puma) Balarezo	pumabalarezo	Cuenca	Periodismo deportivo de Radio del Sur 91.7 FM de Cuenca. Amante
1232440491381268485	Od. Melissa Cevallos	Meliyao_Od	Guayaquil, Ecuador	Mom||Dentist||Athlete||AnimalLover||Entrepreneur
537789951	Mariin••'	YorlanD_Mg28	Caracas	Jagaess
317807823	Carmen Lucía Ramón	CaluRamon	Quito, Ecuador	Lojana, comunicadora de la @upsalesiana especialista en #ComPol
16530311	Pablo Robayo	robayo_dev	127.0.0.1	Bahai,

(a) Tabla: users

id_entity [PK] integer	entity_tweet_id text	hashtag text	user_mentions text
1	1467936043471101955	{}	{NoronaCarlos}
2	1467936111884320769	{}	{}
3	1467936121959206922	{}	{}
4	1467936128556847114	{Quito,PonleMás,CaritaDeDios,Fundación}	{}
5	1467936162534805507	{}	{Amjeliux,BryanChavez_12,RoGeRZeTo}

(b) Tabla: entities

place_tweet_id [PK] text	place_id text	place_type text	place_name text	place_full_name text
1467934220207206420	013cb38e7fe501ae	city	Cuenca	Cuenca, Ecuador
1467935502129143819	01c0324312d210f3	city	Atacames	Atacames, Ecuador
1467935530184761344	005b9f350cad74f6	city	Popayán	Popayán, Colombia
1467936249696727056	013cb38e7fe501ae	city	Cuenca	Cuenca, Ecuador
1467936404185493505	00189e3503ff4e54	city	Santo Domingo	Santo Domingo, Ecuador

(c) Tabla: place

Figura 6.2: Almacenamiento en PostgreSQL