ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

MODELO DE ESTIMACIÓN DEL IMPACTO DE LOS TITULARES DE NOTICIAS PUBLICADAS EN FACEBOOK EN LA AUDIENCIA ECUATORIANA BASADO EN MINERÍA DE TEXTO

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

ROBERTO CARLOS LEMA VINLASACA
roberto.lema@epn.edu.ec

Director: Dra. Lorena Katherine Recalde Cerda
lorena.recalde@epn.edu.ec
Codirector: Dr. Edison Fernando Loza Aguirre
edison.loza@epn.edu.ec

Quito, julio 2023

APROBACIÓN DEL DIRECTOR

Como director del trabajo de titulación MODELO DE ESTIMACIÓN DEL IMPACTO DE LOS TITULARES DE NOTICIAS PUBLICADAS EN FACEBOOK EN LA AUDIENCIA ECUATORIANA BASADO EN MINERÍA DE TEXTO desarrollado por Roberto Carlos Lema Vinlasaca, estudiante de la Maestría en Sistemas de Información, con mención en Inteligencia de Negocios y Analítica de Datos Masivos, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la Defensa oral.

Firmado electrónicamente por:
LORENA KATHERINE RECALDE CERDA

Lorena Katherine Recalde Cerda

DIRECTOR

Firmado electrónicamente por:
EDISON FERNANDO LOZA AGUIRRE

Edison Fernando Loza Aguirre

CODIRECTOR

DECLARACIÓN DE AUTORÍA

Yo, Roberto Carlos Lema Vinlasaca, 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.

Roberto Carlos Lema Vinlasaca

DEDICATORIA

A Dios y a mis padres.

AGRADECIMIENTO

A Dios, por darme la inteligencia y sabiduría para continuar con mis estudios y cumplir mis metas propuestas. A mis padres por su apoyo incondicional en cada etapa de mi vida. A mi director y codirector de proyecto por su guía durante este proceso.

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

El surgimiento de las redes sociales y su acceso público ha permitido que los criterios emitidos por grupos u organizaciones sean recibidos e interpretados por diferentes tipos de audiencias, lo que puede afectar su percepción de la realidad. Dependiendo del interés, ética y profesionalismo del autor, el mensaje puede tener un impacto en el entorno social al redefinir hechos, verdades o creencias y más aún si se trata de noticias de interés social. En este trabajo se propone un modelo de aprendizaje automático supervisado para analizar e identificar el sentimiento transmitido tanto en los titulares de noticias publicados en Facebook por los principales diarios ecuatorianos, como en sus correspondientes artículos y luego, estimar su impacto sobre la audiencia local. Los resultados muestran que los principales diarios del Ecuador mantienen el principio de neutralidad en la publicación de los titulares en Facebook a diferencia de sus artículos. Los artículos expresan sentimientos positivos y negativos definidos. Este comportamiento evidencia que la contextualización de las palabras utilizadas en la publicación influye en el sentimiento que transmite dado los diversos significados que estos puedan tener.

Palabras clave. Titulares de noticias, redes sociales, Facebook, análisis de sentimiento, BERT, modelo de clasificación supervisada basado en léxico.

ABSTRACT

The emergence of social networks and their public access have allowed the criteria issued by groups or organizations to be received and interpreted by different types of audiences, which may affect their perception of reality. Depending on the interest, ethics and professionalism of the author, the message can have an impact on the social environment by redefining facts, truths, or beliefs and even more if it is about news of social interest. In this work, a supervised machine learning model is proposed to analyze and identify the sentiment transmitted in the news headlines published on Facebook by the main Ecuadorian newspapers, as well as their corresponding articles, and then estimate their impact on the local audience. The results show that the main newspapers in Ecuador meet the principle of neutrality in the publication of headlines on Facebook. However, their articles express defined positive and negative sentiments, which results in the fact that the contextualization of the words used in the publication influences the sentiments that it transmits due to the different meanings that these words may have.

Keywords. News Headlines, Social Network, Facebook, Sentiment Analysis, BERT, Lexicon-Based, Supervised Classification Model.

1 INTRODUCCIÓN

1.1 PLANTEAMIENTO DEL PROBLEMA

La evolución tecnológica suscitada en las últimas décadas ha cambiado drásticamente la forma de comunicación de los seres humanos y su interacción en sociedad. Hoy en día, tras el surgimiento de las redes sociales, la comunicación entre individuos se ha tornado global y de acceso público; donde, el criterio emitido de una persona, grupo u organización es receptado e interpretado por diversos tipos de audiencias, cuya apreciación dependerá de cada individuo, afectando su percepción de la realidad. Así, el contenido e intención de un mensaje enviado puede repercutir en el entorno social [1]. Por ejemplo, se ha demostrado que las publicaciones en Facebook han afectado la intención de voto en elecciones locales [2]. Por tanto, en particular, el análisis de la intención de los mensajes transmitidos con sentido informativo y su impacto en la audiencia se ha vuelto un campo de estudio necesario que permitirá identificar el objeto de este.

La gran variedad, velocidad y volumen en la generación y divulgación de publicaciones en redes sociales online, incitan a cuestionar su veracidad y valor. Por lo cual, el hecho de asegurar la procedencia de estos datos se convierte en un reto en este entorno [3]. Es así que el mensaje transmitido en las noticias debe cumplir con el objetivo del periodismo que se define por la imparcialidad, la neutralidad, la equidad y la precisión en sus publicaciones [4]. Los mensajes inadecuados en la comunicación repercuten en la opinión pública, que, de hecho, hoy en día cambia con rapidez en el entorno digital [5]. Por lo tanto, el análisis del trasfondo de una noticia publicada en redes sociales hará posible identificar la imparcialidad de esta hacia una realidad.

Las publicaciones realizadas por diversos grupos de personas transmiten opiniones, actitudes y emociones [6] frente a un tema de interés particular o público, el cual es calificado de acuerdo con el criterio de cada persona. Así, las redes sociales presentan, a más de la publicación, una calificación general de la audiencia, dando una idea del trasfondo emocional del mensaje. Sin embargo, esta primera percepción de la audiencia, en cierta manera, no refleja a fondo la intención del mensaje. Esto debido a que la mayoría de las personas no ahondan en analizar el objeto como tal y se dejan llevar por otros criterios, dando como resultado un efecto dominó; en el cual, el criterio de un individuo es impuesto sobre otro afectando la precisión de la comunicación [7].

El análisis del impacto de las publicaciones en redes sociales ha sido sujeto de estudio en diversos casos y escenarios como en [2], [7], [8], y [9], en los cuales se analizó el texto propio de la publicación e inclusive los comentarios asociados de varios usuarios, con la finalidad de extraer sentimientos o emociones que se transmiten. Este análisis parte de la definición de análisis de sentimientos de una opinión y considera el tipo, polaridad e intensidad del sentimiento trasmitido. De esta manera, el texto puede ser clasificado por su posición positiva, negativa o neutral usando mecanismos de Análisis de Sentimientos como técnica de Procesamiento de Lenguaje Natural (NLP). Este tipo de análisis presenta ciertos retos por superar, los cuales dependerán del método utilizado, el lenguaje, las estructuras semánticas, las ambigüedades, los errores de escritura, entre otros [10]. Una vez aplicados distintos métodos para diversos casos, el resultado es la evaluación de la intensión del texto desde una perspectiva imparcial para el observador.

Los estudios enunciados previamente, han enfocado su análisis en el texto de la publicación como tal y su contraste con las calificaciones o reacciones que los usuarios les dan, para comparar su resultado e impacto en el comportamiento de la audiencia. Sin embargo, no se analizó el trasfondo o contexto de esa publicación. En el caso de las noticias, la mayor parte de los titulares publicados en redes sociales se acompañan de sus artículos a través de un enlace. El ingresar a este enlace y analizar a más profundidad la intención de la publicación, permitirá evaluar de mejor manera su impacto en los usuarios e identificar si la intención del titular de la noticia corresponde a su artículo y cuál es esa intención desde una perspectiva imparcial.

De esta forma, este trabajo se enfoca en el análisis de sentimientos tanto de los titulares de noticias publicados en Facebook por los principales diarios ecuatorianos como el de sus noticias, para así poder estimar su impacto sobre la audiencia local. Para lograrlo, se han definido las siguientes preguntas de investigación: ¿Cuál es el sentimiento que transmiten el titular de una noticia publicada en Facebook y sus noticias? y ¿Cuál es su impacto en el criterio de la audiencia? Para abordar estas interrogantes, se propone un modelo supervisado de clasificación de texto que permita *i)* evaluar el sentimiento transmitido y *ii)* estimar el impacto positivo, negativo o neutral de los titulares de noticias. Así, dentro de este trabajo se analizarán y procesarán los titulares de noticias publicados en Facebook por los cinco diarios más representativos del país: El Comercio, El Universo, El Telégrafo, El Diario y Diario Expreso, al igual que los artículos correspondientes publicados en sus respectivos portales Web para posteriormente contrastar sus resultados y evidenciar su impacto.

1.2 OBJETIVO GENERAL

Desarrollar un modelo de estimación del impacto de los titulares de noticias publicadas en Facebook en la audiencia ecuatoriana a través de la aplicación de minería de texto.

1.3 OBJETIVOS ESPECÍFICOS

- Revisar literatura referente a temas de algoritmos de procesamiento de texto, técnicas de modelos semánticos, análisis de sentimientos y minería de datos en el contexto de comunicación digital y su impacto.
- Analizar y caracterizar los titulares de noticias obtenidos.
- Construir un modelo supervisado de clasificación de texto para luego poder medir el impacto del titular de la noticia en el lector en relación con la noticia.
- Aplicar métricas de evaluación en el modelo propuesto y validar su aplicabilidad.

1.4 MARCO TEÓRICO

1.4.1 Minería de texto

La minería de texto es un proceso que utiliza técnicas de NLP, aprendizaje automático y estadísticas para extraer información y conocimiento útil a partir de grandes cantidades de datos textuales no estructurados como: documentos, correos electrónicos, mensajes de redes sociales, entre otros. Esta técnica permite analizar el contenido de un texto para identificar patrones, temas, relaciones y sentimientos; lo que puede ser utilizado para tomar decisiones informadas en diversos campos, como el empresarial, académico o gubernamental [11].

1.4.2 Análisis de sentimientos

El análisis de sentimientos es una técnica de NLP que tiene como objetivo identificar el sentimiento, emoción o polaridad expresada en un texto, ya sea positiva, negativa o neutral [12]. Se utiliza en la minería de texto y en la inteligencia artificial para comprender la opinión expresada por un usuario sobre un tema, producto o servicio. Esta técnica utiliza algoritmos de aprendizaje automático para clasificar automáticamente el texto en diferentes categorías de sentimiento y proporcionar un análisis cuantitativo de los resultados [10]. El análisis de sentimientos es útil para tomar decisiones informadas en diversos campos,

como la investigación de mercado, el análisis de opiniones políticas, el seguimiento de la reputación de una marca o el análisis de la satisfacción del cliente.

1.4.3 CRISP-DM

La metodología *Cross-Industry Standard Process for Data Mining* (CRISP-DM) es un enfoque estándar para la minería de datos que se utiliza en proyectos de análisis de datos. Esta metodología se compone de seis fases interconectadas y cíclicas: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue. Cada una de estas fases es iterativa y se retroalimenta de las fases anteriores y posteriores para asegurar que el proyecto cumpla los objetivos del negocio y se tomen decisiones informadas en cada etapa del proceso [13].

1.4.4 Word Embeddings

Word Embeddings es una técnica utilizada en NLP para representar palabras como vectores numéricos de alta dimensión en un espacio vectorial continuo. Esta técnica utiliza algoritmos de aprendizaje automático para capturar la semántica y las relaciones entre las palabras a partir de grandes cantidades de texto sin etiquetar (e.g. noticias, libros o artículos). La representación de las palabras como vectores permite a las máquinas comprender el significado de las palabras y encontrar relaciones entre ellas, lo que facilita tareas como la traducción automática, la clasificación de texto, la búsqueda de información y el análisis de sentimientos [14]. Esta técnica ha sido ampliamente utilizada en la industria y la investigación de NLP debido a su capacidad para mejorar la precisión y eficiencia de los modelos de aprendizaje automático.

1.4.5 BERT

BERT (*Bidirectional Encoder Representations from Transformers*) es un modelo de lenguaje basado en redes neuronales transformadoras (*Transformers*) utilizado en el procesamiento del lenguaje natural. BERT es capaz de entender el contexto de las palabras en una oración al procesarlas en ambas direcciones (bidireccional) y utilizar la información de contexto para predecir la siguiente palabra en una oración [15]. Esto permite que el modelo tenga una comprensión más profunda del lenguaje natural y sea capaz de realizar, con mayor precisión, tareas como la traducción, el resumen de texto, la generación de texto y el análisis de sentimientos. BERT es una técnica de aprendizaje profundo que

ha sido pre-entrenada con grandes cantidades de datos no etiquetados y luego afinada (*fine-tuning*) en conjuntos de datos etiquetados específicos para diferentes tareas de NLP.

1.5 REVISIÓN DE LA LITERATURA

1.5.1 Análisis del Impacto de Noticias

El análisis de impacto de noticias publicadas en redes sociales ha sido objeto de estudio en diversos casos y escenarios como [2], [5], [7] y [9] donde se analizan los titulares incluyendo las reacciones y comentarios de los usuarios para extraer sentimientos o emociones. Varios investigadores han centrado sus estudios en comprender cómo se utilizan diferentes estrategias de comunicación para redactar y publicar un titular en las redes sociales. Por ejemplo, en [16], [17], [18] y [19] analizan titulares de noticias comparando los publicados en redes sociales con los artículos publicados en los portales web de cada medio de comunicación, concluyendo que se privilegia la subjetividad y la informalidad en la publicación de titulares en redes sociales con el objetivo transmitir un mensaje emotivo que capte la atención de la audiencia. [20] y [21] analizan cómo los medios transmiten emociones a través de sus titulares para atraer la atención de los usuarios, siendo aquellos que expresan con fuerza sentimientos positivos y negativos los más populares y aceptados. De esta manera, en estos estudios, se analiza el impacto de los titulares de las noticias en función del estilo de redacción y el objetivo sobre la audiencia, evidenciando la forma de hacer periodismo en las redes sociales.

1.5.2 Análisis de Sentimientos

El análisis de sentimiento, o minería de opiniones, ha sido una de las herramientas utilizadas para extraer emociones, actitudes y opiniones de una expresión textual, donde el análisis semántico, sintáctico y lingüístico influye mucho en el procesamiento del texto y la interpretación de los resultados [10]. Estos resultados se han utilizado para identificar patrones y tomar decisiones en diferentes campos como el político, comercial, económico, social y de salud ([2], [22], [23], [24], [25]). Sin embargo, el procesamiento y análisis de este tipo de datos se convierte en una tarea compleja debido a que se presentan de manera no estructurada y diversa. Los trabajos relacionados en esta temática se han centrado en el estudio teórico y práctico de las técnicas y herramientas de análisis de sentimientos basadas en enfoques de *machine learning, lexicon-based, hybrids approach, knowledge transfer, and aspect-based* ([26], [10], [27], [28], [29]). Estudios recientes han implementado

modelos de aprendizaje automático a través de algoritmos de clasificación como *decision trees, linear, probabilistic,* y *k-Nearest Neighbor* ([30], [31], [32], [33]) para analizar el sentimiento implícito en textos cuya efectividad ha dependido del preprocesamiento, extracción de características y evaluación del texto. Así, técnicas como *TF-IDF*, *N-Grams*, *Bag of Words*, BERT y *word2vec* ([34], [35], [29], [36], y [37], respectivamente), han sido implementados para este fin.

La mayoría de estos estudios se han realizado en idioma inglés; sin embargo, efectuar este análisis en otros idiomas, como el español, requerirá un esfuerzo adicional para la normalización y evaluación del texto. Para ello, existen herramientas (p. ej., *TextBlob, Standford CoreNLP, RapidMiner*, etc.), que permiten analizar texto en español a partir de traducciones al inglés, lo que ha demostrado una efectividad imprecisa en la medición de la polaridad del texto [29]. En este sentido, [12], [38] y [39] analizan la aplicación de modelos pre-entrenados, transferencia de conocimiento y técnicas de ajuste para clasificar el texto en español basados en grandes conjuntos de datos obtenidos de documentos, artículos, y recursos digitales a través de Internet, dando resultados favorables para esta tarea. En consecuencia, el proceso y resultado del análisis de sentimiento depende de la forma en que se procese el texto según el idioma. Por tanto, el diseño de un modelo de análisis de sentimiento de texto en lengua española es todavía un campo de estudio por explorar.

Como se detalló, las contribuciones hechas hasta la fecha se centran en analizar los titulares de las noticias con sus calificaciones y reacciones dadas por los usuarios de las redes sociales para estimar el impacto en la audiencia. Sin embargo, no se analizan los antecedentes o contexto de dicha publicación. El trabajo propuesto tiene como objetivo desarrollar un modelo de clasificación de texto supervisado para medir el impacto positivo, negativo o neutral del titular de la noticia contrastándolo con el sentimiento transmitido en sus artículos.

2 METODOLOGÍA

La metodología CRISP-DM es considerada como un estándar "de facto" para el desarrollo de diversos proyectos de minería de datos dado que su aplicabilidad es independiente de los diferentes dominios de la industria y tecnología utilizada [13]. Es así como, el uso de esta metodología permite planificar, ejecutar y administrar las actividades y tareas de una manera flexible y efectiva desde la comprensión del problema hasta el despliegue de la solución considerando la heterogeneidad de las herramientas, tipos de datos y técnicas de minería de datos empleados para el desarrollo del proyecto.

Así, las fases y actividades ejecutadas a lo largo del proyecto se describen a continuación:

Tabla 1: Fases de la Metodología CRISP-DM

FASE	ACTIVIDADES	DESCRIPCIÓN
1	Comprensión del negocio	Se define los objetivos del negocio, la situación actual y los objetivos de minería de datos.
2	Comprensión de los datos	Se realiza la recolección, descripción, exploración, y verificación de la calidad de los datos de acuerdo con su relevancia para el proyecto.
3	Preparación de los datos	Se realiza la selección, limpieza, construcción y representación numérica de los datos para prepararlos para el análisis y la modelización.
4	Modelado	Se realiza la selección del algoritmo de clasificación del texto y se entrena los modelos.
5	Evaluación	Se realiza la evaluación del modelo y se verifica que se cumplen los objetivos del negocio y los criterios de éxito establecidos en la fase de comprensión del negocio.
6	Despliegue	Se despliega el modelo desarrollado, y se presenta los aspectos del problema y los resultados obtenidos.

2.1 COMPRENSIÓN DEL NEGOCIO

Cada minuto se generan grandes cantidades de publicaciones en Facebook y las reacciones de los usuarios son inmediatas al expresar su percepción a través de comentarios y botones de reacciones. Este comportamiento se observa en publicaciones referidas a noticias ya que son de interés nacional y su interpretación es de importancia para la sociedad en general. Estas publicaciones son de acceso público para ser vistos por la mayor audiencia posible y es por ello que el estudio de las emociones implícitas en el texto permite identificar las intenciones de cada autor y su impacto en la audiencia en diferentes campos.

2.1.1 Objetivos del Negocio

Este trabajo propone un modelo de aprendizaje automático para estimar el impacto de la publicación de titulares de noticias en Facebook, por los principales diarios ecuatorianos, identificando el sentimiento transmitido a la audiencia y contrastándolo con el sentimiento transmitido en el artículo de la noticia. El modelo propuesto permitirá comprender los sesgos entre la ética y el objetivo de la industria periodística en el Ecuador permitiendo a la audiencia discernir la fuente de la noticia.

2.1.2 Situación Actual

Actualmente, existen diversos estudios del impacto de las publicaciones de Facebook sobre diferentes tipos de audiencia en diversos campos como el político, turístico, o comercial. De hecho, los resultados obtenidos han mostrado que la forma en que las publicaciones son realizadas repercute en el criterio del usuario alterando una realidad de acuerdo con el objetivo del autor. Es así que, fuentes de información como las originadas en el periodismo, han sido sujeto de estudio ya que el mensaje que transmiten es considerado como un hecho por la sociedad y su impacto puede alterar las decisiones que esta toma para su entorno. Sin embargo, estos estudios se han enfocado en su mayoría en analizar publicaciones en idioma inglés dada su universalidad y disponibilidad de fuentes de datos y herramientas. Por lo cual, el estudio de este aspecto en idioma español es un campo por explorar considerando las particularidades del idioma y otros aspectos que pueden influir en el impacto que estas publicaciones tengan sobre la audiencia local.

2.1.3 Objetivos de minería de texto.

Para la consecución del objetivo de este proyecto, se han establecido los siguientes objetivos de minería de texto:

- Extraer las publicaciones de los principales diarios ecuatorianos de páginas públicas de Facebook y sus correspondientes noticias de los portales web de cada diario.
- Identificar algoritmos de clasificación de texto para el análisis de sentimientos.
- Analizar los sentimientos transmitidos en el texto de las publicaciones y clasificarlos como positivo, negativo o neutro.

2.2 COMPRENSIÓN DE LOS DATOS

Los datos utilizados para la estimación del impacto de los titulares de noticias han sido obtenidos de páginas públicas de Facebook y de los portales webs de cada diario. Una vez almacenados, se realizó la descripción, exploración, y verificación de la calidad de los datos de acuerdo con su relevancia para el proyecto.

2.2.1 Recolección de Datos

Los titulares de noticias se recopilaron de las páginas públicas de Facebook de los diarios El Comercio, El Universo, El Telégrafo, El Diario y Diario Expreso durante el período comprendido entre octubre 2022 y marzo 2023, mediante el uso de bibliotecas Python. Estos datos se procesaron y almacenaron en una base de datos no relacional para su análisis. Posteriormente, los artículos de las noticias se recolectaron de los portales web de cada diario a los cuales se accedió por medio de los enlaces contenidos en las publicaciones realizadas en Facebook. Estos datos se obtuvieron a través de librerías programadas en Python que permiten la exploración, extracción y almacenamiento de texto en páginas web. En la Figura 1 se observa el proceso utilizado para la recolección de los datos.

Finalmente, el corpus obtenido está compuesto por dos conjuntos de datos almacenados en MongoDB y comprende 11000 titulares de noticias junto con sus reacciones, comentarios y artículos que serán utilizados en el análisis propuesto en este trabajo.

2.2.2 Descripción de Datos

Los datos se almacenaron en formato JSON y se dividen en dos colecciones. La primera colección contiene documentos recopilados de las publicaciones de Facebook y la segunda contiene los documentos recopilados de las páginas web de los medios de comunicación, como se muestra en la Figura 2 y en la Figura 3. Los documentos de cada colección tienen atributos con diferentes tipos de valores que dependiendo de su relevancia fueron seleccionados para este trabajo. Esta estructura facilita la manipulación de datos y su análisis para obtener el corpus final.

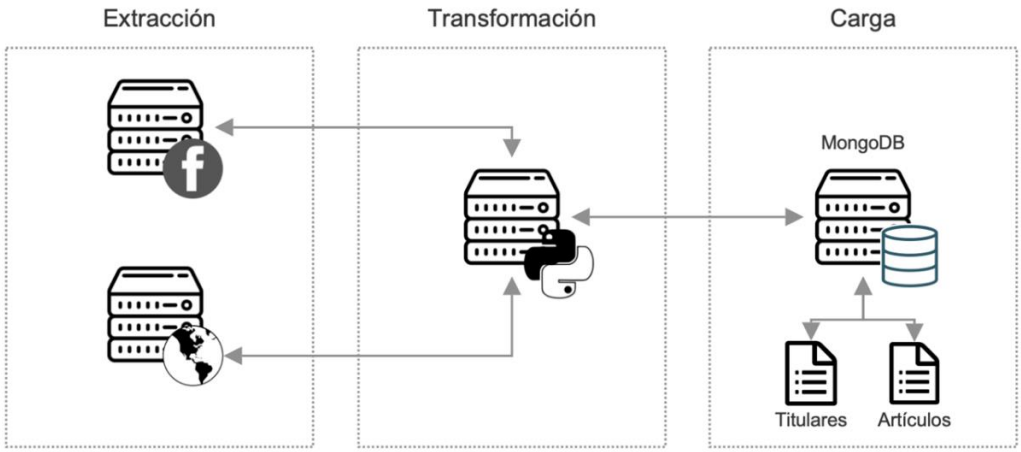

Figura 1: Proceso de Recolección de Datos.

Elaborado por: El Autor.


```
{
  "_id": {...},
  "post_id": "10160464737409560",
  "text": "Walter Chalá anotó en el complemento el único gol del partido, válido por la jornada 14 de la primera ron
  "post_text": "Walter Chalá anotó en el complemento el único gol del partido, válido por la jornada 14 de la primer
  "shared_text": "ELUNIVERSO.COM\nEmelec cae ante Universidad Católica en el George Capwell",
  "original_text": "Walter Chalá anotó en el complemento el único gol del partido, válido por la jornada 14 de la pr
  "time": {...},
  "timestamp": 1653179728,
  "image": null,
  "image_lowquality": "https://external-yyz1-1.xx.fbcdn.net/emg1/v/t13/14146341216891660996?url=https%3A%2F%2Fwww.el
  "images": [],
  "images_description": [],
  "images_lowquality": [...],
  "images_lowquality_description": [...],
  "video": null,
  "video_duration_seconds": null,
  "video_height": null,
  "video_id": null,
  "video_quality": null,
  "video_size_MB": null,
  "video_thumbnail": null,
  "video_watches": null,
  "video_width": null,
  "likes": 9,
  "comments": 8,
  "shares": 1,
  "post_url": "https://facebook.com/eluniversoec/posts/10160464737409560",
  "link": "http://ow.ly/IFAm50JeQhp?fbclid=IwAR1gzjDEitoVveGJknZfToufIx0xYW31pY-amtHJ47Ou2eWBobs5cu_pA8s",
  "links": [...],
  "user_id": "61449504559",
```


Figura 2: Documento JSON de titulares en Facebook.

Elaborado por: El Autor.

```
_id: ObjectId('62ba37f7657e666c027450f5')
title: "Emelec, inofensivo y franqueable, cae ante Universidad Católica en el …"
text: "Emelec, con pobre rendimiento, sobre todo en ataque, cayó la noche de …"
url: "http://ow.ly/IFAm50JeQhp?fbclid=IwAR1gzjDEitoVveGJknZfToufIx0xYW31pY-a…"
username: "El Universo"
fetched_date: 2022-06-27T18:06:31.296+00:00
```


Figura 3: Documento JSON de artículos de notícias.

Elaborado por: El Autor.

2.2.3 Exploración de Datos

La colección de Facebook tiene 11887 documentos; sin embrago, no todos son titulares. Hay publicaciones relacionadas con publicidad y videos que no forman parte de este estudio como se muestra en la Figura 4. Esto se refleja en la colección de noticias, donde hay menos documentos que en la colección de Facebook, como se muestra en la Figura 5. Estos documentos que no correspondían con el objeto del estudio se eliminaron para evitar el ruido en el proceso de modelado. Por otro lado, la Figura 5 muestra cómo se distribuyen los documentos según cada medio de comunicación. Se puede observar que El Comercio lidera el proceso de difusión de noticias en Facebook dando una idea de cuál es la preferencia de consumo de noticias de los usuarios en esta red social. Finalmente, cabe señalar que este corpus no está clasificado por el sentimiento que transmite, por lo que no es posible saber si este corpus contiene datos equilibrados en este punto del estudio.

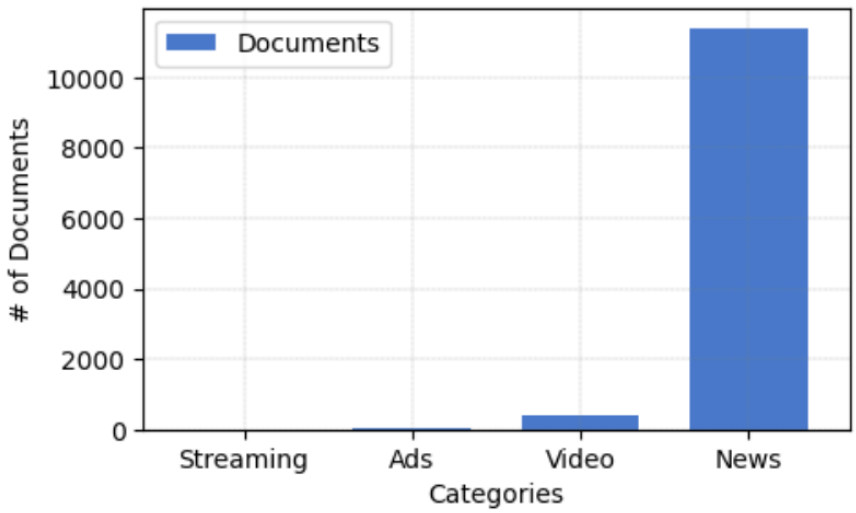

Figura 4: Distribución de documentos por categoría.

Elaborado por: El Autor.

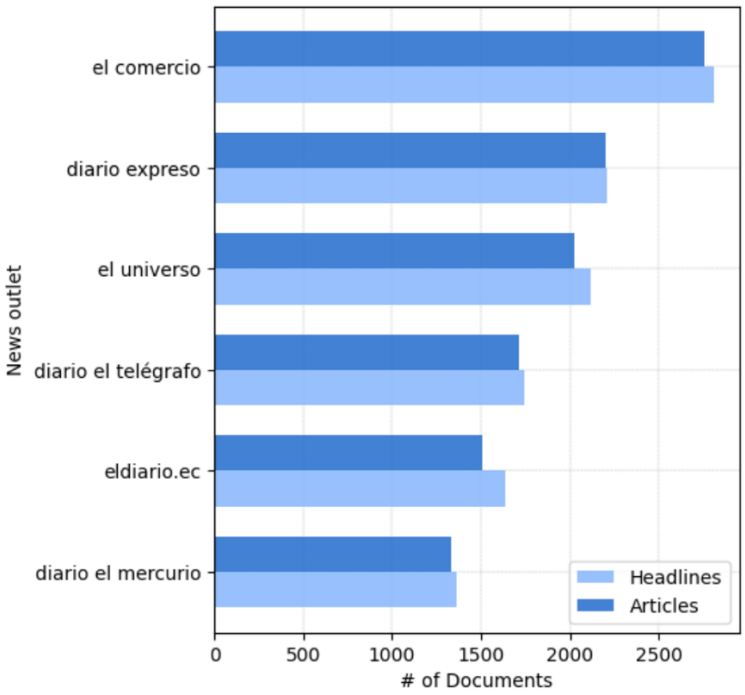

Figura 5: Distribución de documentos por medio de comunicación.

Elaborado por: El Autor.

2.2.4 Verificación de Calidad de Datos

Tras la exploración del corpus, se observó la falta de algunos atributos y valores como enlaces, comentarios y reacciones. Los valores faltantes estaban presentes en documentos relacionados con publicaciones que no contienen titulares de noticias. Dado que el objetivo de este trabajo es estimar el impacto de los titulares de noticias en contraste con su artículo, el atributo *"links"* juega un papel importante. Este atributo permite obtener el artículo relacionado de la página web. Por eso, dependiendo de cómo el medio de comunicación gestione la publicación, el enlace se obtiene de otros atributos del documento. Sin embargo, no siempre tener un enlace significa que hay un artículo detrás. Algunos enlaces conducen a páginas con contenido de entretenimiento o publicidad que no forman parte de este trabajo. Por lo tanto, se requirió un esfuerzo adicional para discriminar estos documentos y obtener un conjunto de datos consistente para un análisis posterior.

2.3 PREPARACIÓN DE LOS DATOS

La preparación de datos recopilados en fases anteriores influye en la efectividad de un modelo de aprendizaje automático. Como se muestra en la Figura 2 y la Figura 3, los datos

poseen varios atributos, pero no todos serán seleccionados para entrenar el modelo. Además, hay campos a partir de los cuales se construirán nuevos atributos derivados. En particular, hay datos que deben limpiarse para garantizar la calidad de los datos y etiquetarse para el propósito de este trabajo.

2.3.1 Selección de Datos

Debido a que los datos recopilados se dividen en dos corpus, se seleccionaron los atributos relevantes de cada uno considerando el objetivo de este trabajo. Del corpus de Facebook se seleccionaron los atributos *"post-id", "text", "username", "comments", "links"* y *"reactions"*. Del corpus de noticias se seleccionaron los atributos *"text"* y *"url"*. Estos atributos se unificaron en un corpus y se redujeron a los atributos: *"post-id", "username", "headline", "comments", "reactions"* y *"article".*

2.3.2 Limpieza de Datos

Los datos textuales se procesaron mediante la eliminación de números, puntuación, caracteres especiales, e hipervínculos para evitar introducir sesgos o ruido en los modelos de aprendizaje automático. Los *"stopwords"* como artículos, preposiciones, conjunciones y otros términos muy utilizados en el lenguaje cotidiano, no se eliminaron teniendo en cuenta su importancia para las tareas que implican la comprensión del texto en su contexto, como la clasificación de textos o el análisis de sentimientos. En estos casos, los *"stopwords"* pueden contribuir a la estructura general y la coherencia del texto, y eliminarlas puede provocar una pérdida de información y escasez de datos [40].

2.3.3 Construcción de Datos

En esta fase se extrajeron los titulares y artículos del corpus final para ser procesados y utilizados para entrenar y probar el modelo de clasificación. Dado que estos datos no están clasificados por su sentimiento, se han utilizado modelos previamente entrenados para el análisis de sentimiento. La selección del modelo pre entrenado se ha centrado en aquellos formados y probados con corpus en español y afines a la región. Así, se han seleccionado los modelos RoBERTuito y SAET para el etiquetado del sentimiento.

- **RoBERTuito** [39] es un modelo de lenguaje pre entrenado basado en la arquitectura de Transformers para el análisis de texto en español. Este modelo ha sido entrenado siguiendo las pautas de RoBERTa en 500 millones de *tweets* y

supera a otros modelos previamente entrenados para el idioma español como BETO, Roberta y BERTin [33]. El modelo ha sido utilizado para el análisis de sentimientos, emociones, ironía y discurso de odio.

- **SAET** [29] es un modelo de lenguaje pre entrenado basado en un enfoque de léxico para el análisis de sentimientos de textos escritos en español que incluyen dialectos o modismos ecuatorianos. Este modelo muestra un rendimiento y una eficacia comparables con la herramienta comercial IBM Watson NLU para identificar la polaridad de los sentimientos, especialmente en textos que incluyen dialectos, palabras coloquiales y expresiones negativas [29].

Como se muestra en las Figuras 6 y 7, después de la clasificación del texto, existen diferencias en cuanto a las clases mayoritarias. En el primer caso (clasificación hecha con RoBERTuito), la clase mayoritaria es Neutral. Este resultado muestra que los titulares y las noticias cumplen con el objetivo de neutralidad en el periodismo [4]. Por otro lado, en el segundo caso (clasificación hecha con SAET), la clase mayoritaria es la clase Positiva. Esta clasificación refleja el resultado obtenido en el estudio realizado por [20], en donde se concluye que diferentes medios de comunicación orientan sus esfuerzos para aumentar y mantener su audiencia en medios digitales utilizando titulares con polaridades positivas y negativas. En ambos casos, la clasificación depende de los datos utilizados para entrenar los modelos. A efectos de evaluación, se han considerado ambos *datasets* para entrenar y evaluar los modelos.

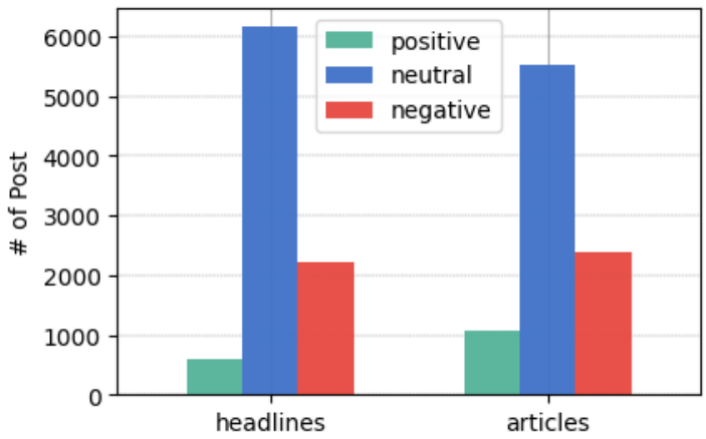

Figura 6: Texto clasificado por RoBERTuito.

Elaborado por: El Autor.

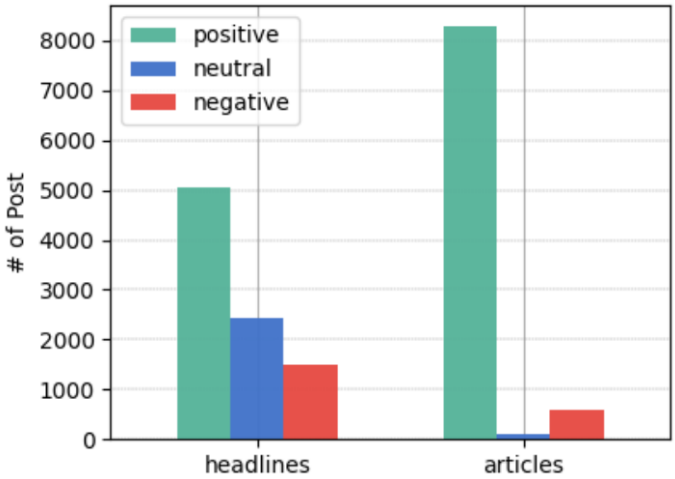

Figura 7: Texto clasificado por SAET.

Elaborado por: El Autor.

2.3.4 Representación de Datos

Los datos textuales requieren una representación numérica para ser procesados por un algoritmo informático. Para este paso, se utilizó *BERT Embeddings* que permiten representar texto en un vector de alta dimensión capturando relaciones semánticas y sistemáticas entre palabras, lo cual es importante para analizar el sentimiento con precisión. Estas incrustaciones (*embeddings*), capturan el significado y el contexto de cada palabra en la oración y se pueden usar como entrada para tareas posteriores de procesamiento del lenguaje natural, como la clasificación de texto.

BERT es un modelo de lenguaje pre entrenado basado en la arquitectura de *Transformers* y está diseñado para comprender el contexto de las palabras en una oración procesándolas de manera bidireccional [15]. Esto significa que la incrustación de una palabra determinada será diferente según el contexto en el que aparezca en la oración. Esto permite que el modelo capture relaciones más matizadas entre las palabras. Como resultado, tenemos un vector de alta dimensión de 768 dimensiones para cada observación del conjunto de datos.

La Figura 8, muestra la representación vectorial reducida a 2 dimensiones donde se puede observar que la distribución de datos es consistente y claramente agrupada.

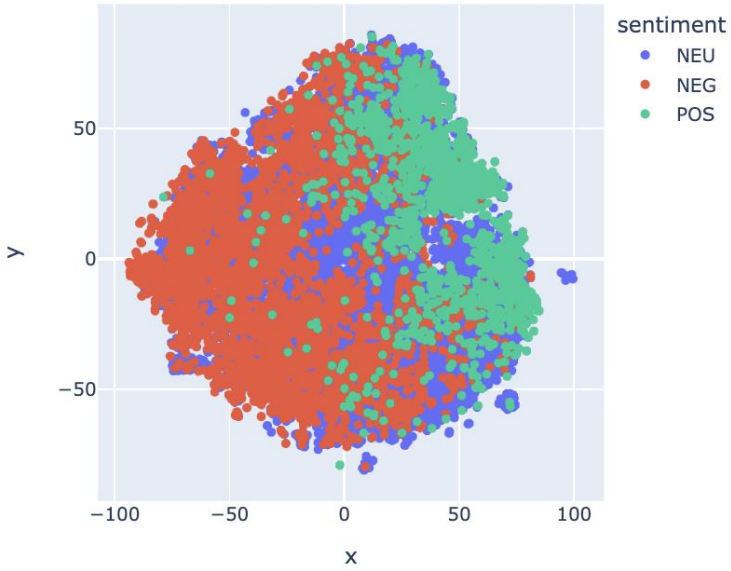

Figura 8: Visualización en 2D de *word embeddings* extraídos con BERT.

Elaborado por: El Autor.

2.4 MODELADO

Una vez transformados los datos textuales a formato numérico, se han realizado algunas tareas necesarias para entrenar y probar los modelos. Primero, el conjunto de datos se dividió en dos partes: el 70% de los datos para entrenamiento y el 30% para pruebas. La división se realizó de forma estratificada para mantener el mismo número de filas en cada clase. En segundo lugar, teniendo en cuenta que el conjunto de datos contiene datos desequilibrados, se seleccionó la técnica de clase de "*weighting class*" para este caso. Esta técnica implica asignar pesos más altos a la clase minoritaria y pesos más bajos a la clase mayoritaria durante el entrenamiento del modelo. Esto asegura que el modelo preste más atención a la clase minoritaria y aprenda a clasificar de mejor manera.

Finalmente, para estimar el impacto de los titulares de noticias, se evaluaron cuatro algoritmos de aprendizaje automático supervisado para la clasificación de texto: Red neuronal convolucional (CNN), Máquina de vectores de soporte (SVM), Regresión logística (LR) y Bosque aleatorio (RF). Estos algoritmos fueron evaluados considerando las particularidades de cada uno. Para el algoritmo CNN, se consideró como capa intermedia a BERT. Para el algoritmo SVM, se utilizaron los siguientes núcleos: sigmoide, lineal, función de base radial y polinomial. Para cada uno de estos, se definió el uso de la función OVO "*One-vs-One*" en el parámetro *decision-function-shape*, que evalúa la pertenencia de las clases en pares. En general, los hiper parámetros de cada algoritmo se configuraron

para clasificación multiclase y los pesos de clase se definieron de acuerdo con el número de observaciones. Posteriormente, se evaluaron los resultados obtenidos en cada uno de estos modelos y se seleccionó el modelo con mayor precisión.

2.5 EVALUACIÓN

La evaluación de los modelos se ha realizado utilizando las siguientes métricas: *accuracy*, *recall*, *precision* y F1-score. La métrica *accuracy* mide la proporción de predicciones correctas sobre el número total de instancias evaluadas. La métrica *precision* mide los patrones positivos que se predicen correctamente a partir del total de patrones predichos en una clase positiva. La métrica *recall* mide la fracción de patrones positivos que se clasifican correctamente. La métrica *F1-score* representa la media armónica entre los valores de *recall* y *precision*.

2.5.1 Evaluación del modelo

La Tabla 2 y la Tabla 3 muestran los resultados de la evaluación del rendimiento obtenidos a partir de algoritmos de clasificación entrenados con conjuntos de datos etiquetados con RoBERTuito y SAET. Como se observa, en ambos casos, el modelo CNN alcanza más del 80% de rendimiento en cada métrica. Por otro lado, los resultados de rendimiento obtenidos en el primer caso difieren de los obtenidos del conjunto de datos de la etiqueta SAET. Como se ve en la Figura 9, las puntuaciones de cada modelo tienen una estrecha relación entre ellas. Mientras que en la Figura 10, las puntuaciones muestran separación entre ellas. Este comportamiento puede comprometer la efectividad de los modelos.

Tabla 2: Resultados de métricas de rendimiento de modelos entrenados con etiquetas de RoBERTuito.

ALGORITMO	***ACCURACY***	***RECALL***	***PRECISION***	***F1-SCORE***
Convolutional Neural Network	0,88	0,87	0,88	0,85
Logistic Regression	0,75	0,75	0,75	0,74
Random Forest	0,74	0,74	0,74	0,72
SVM – Linear	0,71	0,71	0,76	0,72
SVM – Polynomial	0,71	0,71	0,77	0,72
SVM – Sigmoide	0,54	0,55	0,66	0,57
SVM - RBF	0,66	0,66	0,76	0,54

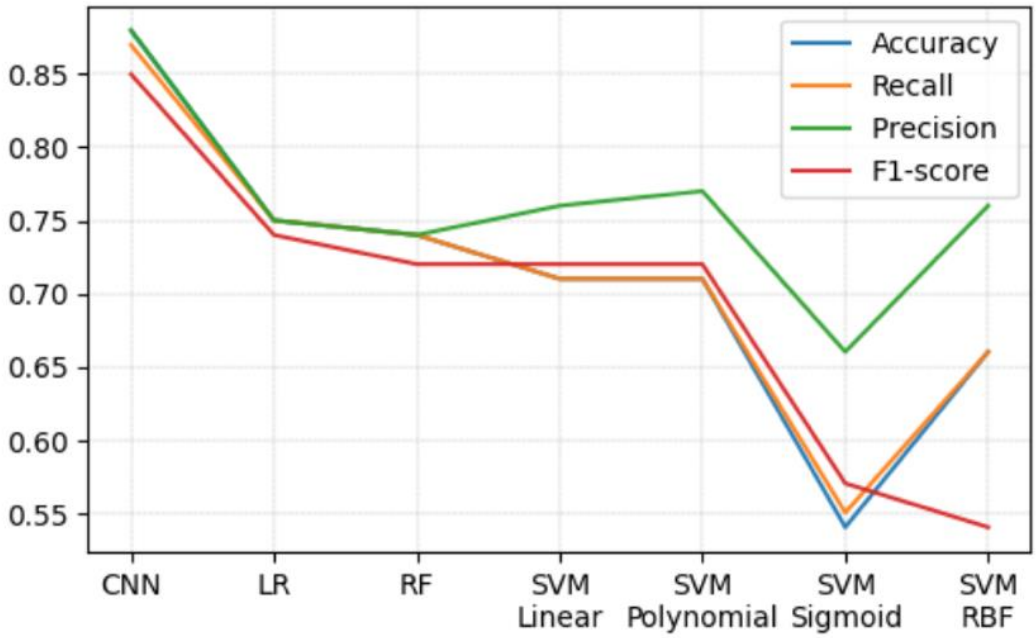

Figura 9: Métricas de rendimiento de modelos entrenados con etiquetas de RoBERTuito.

Elaborado por: El Autor.

Tabla 3: Resultados de métricas de rendimiento de modelos entrenados con etiquetas de SAET.

ALGORITMO	***ACCURACY***	***RECALL***	***PRECISION***	***F1-SCORE***
Convolutional Neural Network	0,81	0,8	0,88	0,74
Logistic Regression	0,65	0,65	0,75	0,69
Random Forest	0,75	0,76	0,72	0,68
SVM – Linear	0,64	0,65	0,78	0,69
SVM – Polynomial	0,64	0,64	0,79	0,68
SVM – Sigmoide	0,54	0,54	0,67	0,59
SVM - RBF	0,74	0,75	0,8	0,65

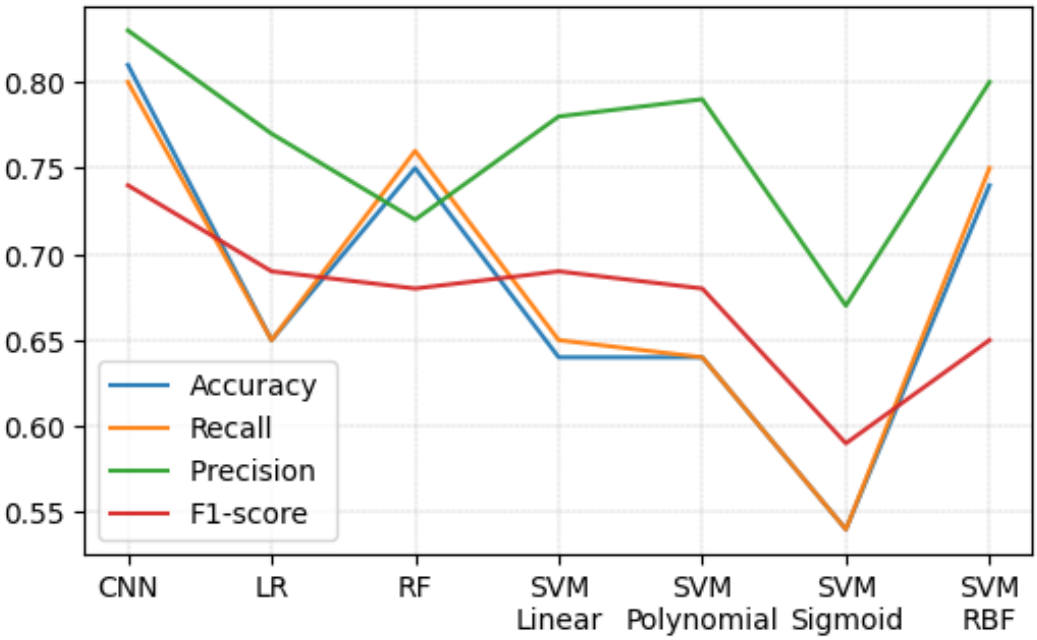

Figura 10: Métricas de rendimiento de modelos entrenados con etiquetas de SAET.

Elaborado por: El Autor.

2.5.2 Análisis y descripción de resultados obtenidos

Los modelos CNN presentan un 88% de precisión, lo que supone un buen comportamiento del modelo. Sin embargo, esta métrica tiene varias debilidades: menos distintividad, menos discriminabilidad, menos información y sesgo hacia los datos de clase mayoritaria [41] en los casos en que los modelos se entrenaron con datos desequilibrados. Como se ve en la Figura 11 y la Figura 12, la matriz de confusión muestra la mejor precisión al clasificar las observaciones en las clases mayoritarias. Un comportamiento esperado en este estudio. Por lo tanto, la métrica *F1-score* se utilizó para evaluar el rendimiento del modelo tomando en cuenta tanto la precisión como la recuperación, lo que proporciona una evaluación más equilibrada del rendimiento del modelo.

F1-score es una métrica de uso común para evaluar el rendimiento de un modelo de clasificación, especialmente cuando se trata de datos de entrenamiento desequilibrados. Combina precisión y recuperación en un solo valor, lo que lo hace adecuado para conjuntos de datos desequilibrados donde la distribución de clases está sesgada. *F1-score* varía de 0 a 1, donde 1 indica una precisión y recuperación perfectas, y 0 indica un rendimiento deficiente.

De esta manera, el modelo de CNN entrenado con un conjunto de datos etiquetado por el modelo RoBERTuito presenta un 85% de *F1-score*, lo que supera las puntuaciones obtenidas de otros modelos y se utilizó para estimar el impacto de los titulares de noticias.

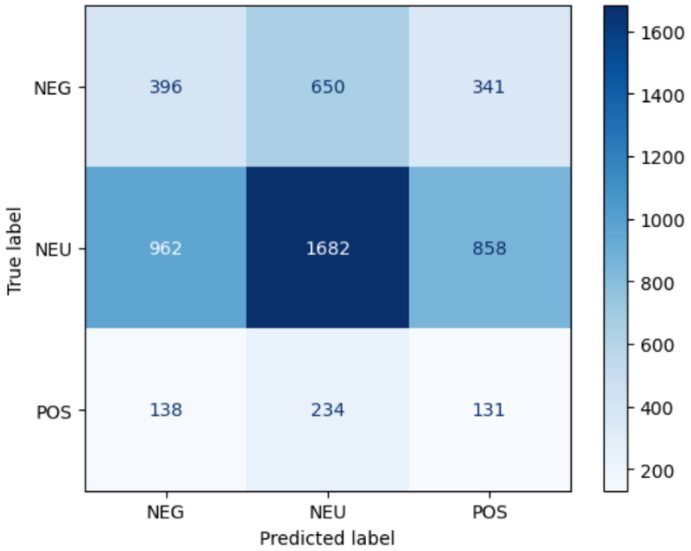

Figura 11: Matriz de confusión del modelo CNN basado en RoBERTuito.

Elaborado por: El Autor.

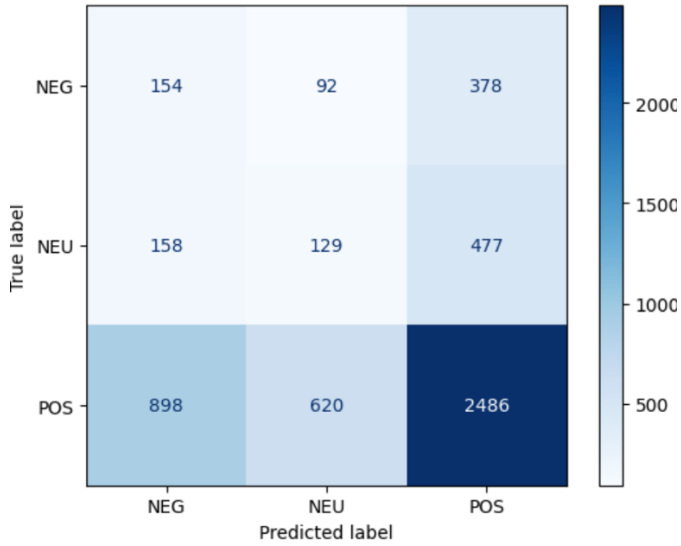

Figura 12: Matriz de confusión del modelo CNN basado en SAET.

Elaborado por: El Autor.

2.6 DESPLIEGUE

El modelo seleccionado se implementó como un RESTful API utilizando el Framework Flask, que toma el texto como entrada, lo procesa y retorna la etiqueta del texto de acuerdo

con el sentimiento transmitido. El modelo permite analizar el impacto del titular de la noticia y su artículo etiquetándolos como positivo, negativo o neutral. Luego, se contrasta la clasificación entre el titular y el artículo, donde se muestra la presencia o ausencia de sesgo entre ellos. De esta manera, se evidencia la forma de escribir en el periodismo digital ecuatoriano. En la Figura 13 y Figura 14, se muestra la clasificación del texto del titular de la noticia y el articulo realizado por el modelo respectivamente.

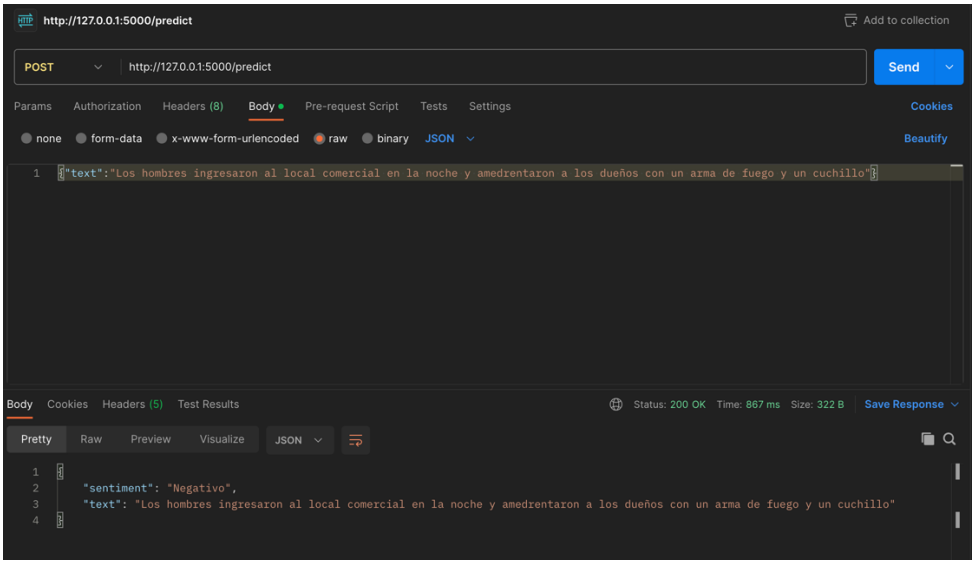

Figura 13: Etiquetado del titular de la noticia por el modelo implementado.

Fuente: El Autor.

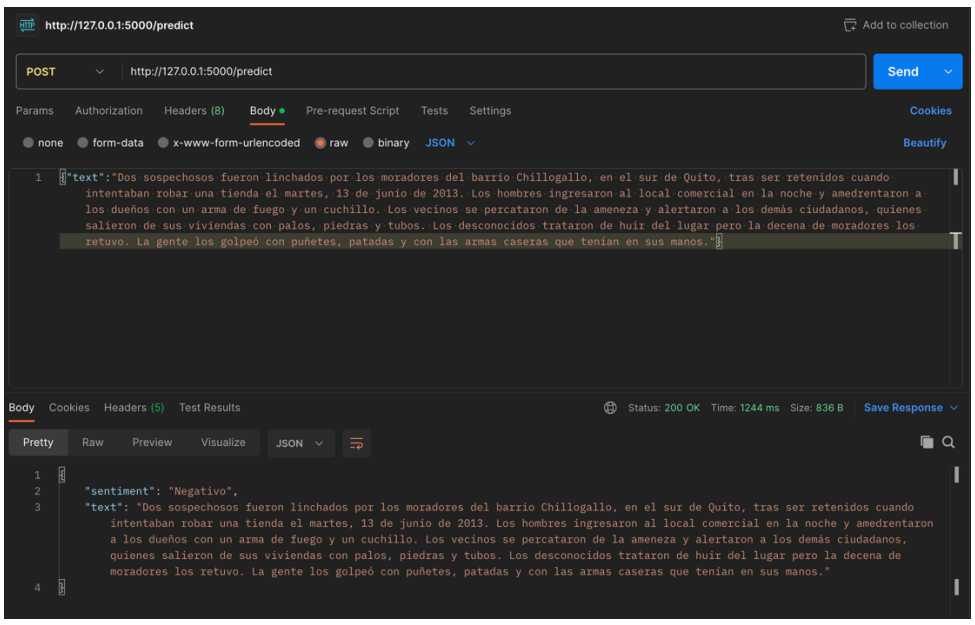

Figura 14: Etiquetado del artículo de la noticia por el modelo implementado.

Fuente: El Autor.

Para verificar la efectividad del modelo, se elaboró un conjunto de datos de 280 titulares y artículos para ser clasificados. Este conjunto de datos fue clasificado por el modelo y por un grupo de 28 estudiantes de ingeniería de la Escuela Politécnica Nacional. Los resultados mostraron que la clasificación realizada por los estudiantes difiere en promedio en un 50% en comparación con el modelo, como se observa en la Figura 13 y la Figura 14. Tomando en cuenta que, el criterio de clasificación de cada estudiante dependerá de varios factores como el social, el económico y el político, el resultado puede no ser imparcial.

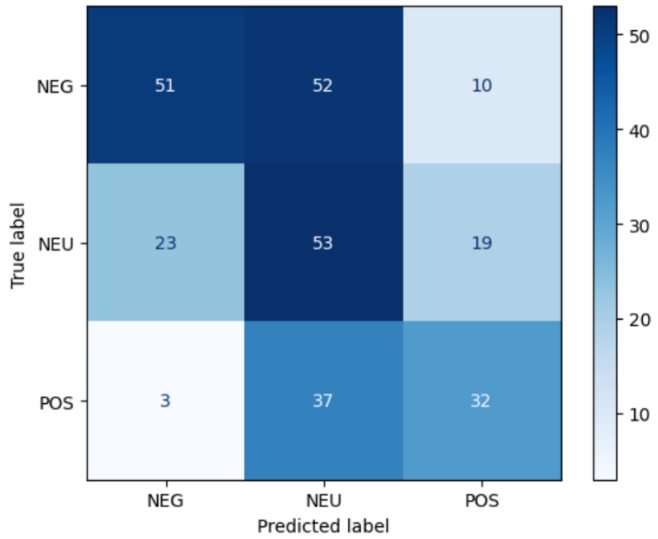

Figura 15: Matriz de confusión de clasificación de titulares.

Elaborado por: El Autor.

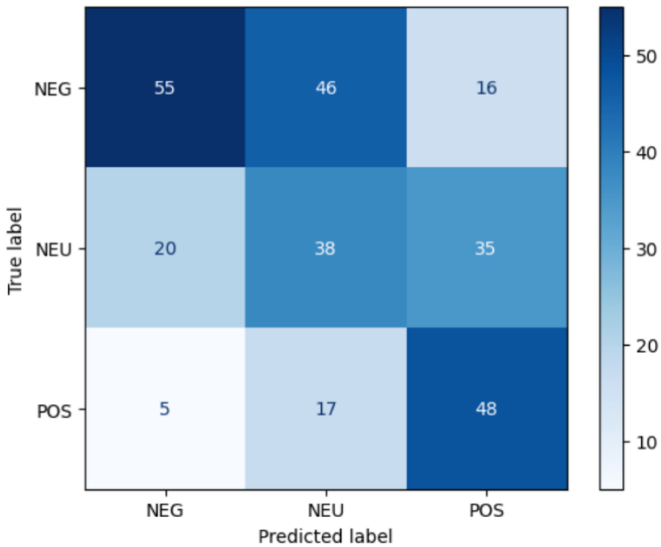

Figura 16: Matriz de confusión de clasificación de artículos.

Elaborado por: El Autor.

La clasificación realizada por el modelo ha mostrado diferencias entre el sentimiento que transmiten los titulares publicados en Facebook y sus respectivos artículos. Como se muestra en la Figura 15, la mayoría de los titulares y artículos han sido clasificados como neutrales. Sin embargo, el sentimiento que transmite el titular no es el mismo que el que transmite el artículo. Asimismo, se puede observar que los titulares de las noticias cumplen en su mayoría con el principio de neutralidad en el periodismo. En la Tabla 4 se muestra los datos analizados donde se puede observar que los artículos, cuyos titulares califican como neutrales, son clasificados por el modelo como positivos o negativos. El contexto de la oración puede cambiar el impacto en la audiencia.

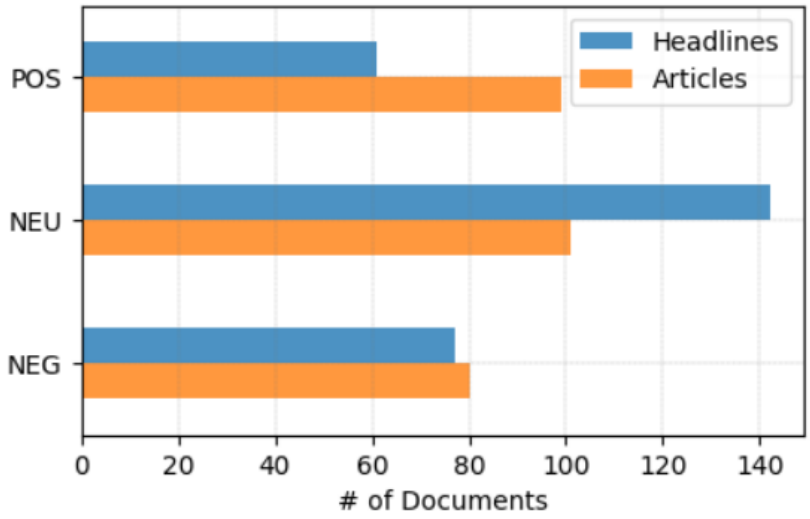

Figura 17: Clasificación realizada por el modelo.

Elaborado por: El Autor.

Tabla 4: Muestra de clasificación de titulares y artículos.

TITULAR	SENTIMIENTO	ARTICULO	SENTIMIENTO
Gerente de comercio internacional de PETROECUADOR explica la diferencia de negociar ahora con petrolera china a como se hacía en el correato.	Neutro	El gerente de comercio internacional de PETROECUADOR, Pablo Noboa, explica las diferencias entre el criticado contrato de largo plazo de PETROCHINA y la última venta por 3,2 millones de barriles que acaba de ganar la misma empresa. la brecha entre …	Negativo
Los supervisores de Guayaquil Siglo XXI realizan rondas para detectar los objetos faltantes en diferentes puntos regenerados de la urbe.	Neutro	Carla García vive desde hace 20 años en un departamento ubicado en un edificio en la avenida 9 de octubre y Rumichaca. ella dice que, a pesar del ajetreo usual del centro de la ciudad, con comerciantes y alto flujo de usuarios, era una zona …	Negativo
Se tiene previsto despegar no antes del último trimestre de 2022.	Neutro	Trajes espaciales diseñados por spacex para su realización. además de probar las comunicaciones basadas en láser STARLINK en el espacio. la tripulación de esta primera misión es completada por Scott Poteet, teniente …	Positivo
Un logo identificará los productos más económicos que parten de una iniciativa del gobierno y la asociación de productores de grasas y aceites.	Neutro	Aceites comestibles con una etiqueta que los identifique como “aceite popular” y que costarán un 20 % menos empezarán a ser distribuidos durante esta semana en las tiendas a nivel nacional. el gobierno y la asociación de productores de grasas y aceites …	Positivo
El dirigente de la CONAIE indicó que van a esperar una respuesta del gobierno.	Neutro	Tras un receso de 20 minutos el diálogo que se desarrollaba en el liceo Matovelle se vio interrumpido debido a que no retornaron los ministros del gobierno de Guillermo Lasso ya que se dirigieron al palacio de Carondelet con las propuestas que dieron a conocer los dirigentes indígenas…	Negativo
A base de autogestión se mantienen entidades regentadas por la curia; pacientes van por tarifas bajas y especialidades.	Neutro	Justina Delgado acudió con su madre, Rita Cusme, para una revisión en tres especialidades que ofrece la sede matriz de la red de dispensarios de la arquidiócesis de Guayaquil (redima), situada en la esquina de las calles Venezuela y Tulcán, en el suroeste…	Positivo
El cable de fibra óptica tendrá una longitud de 17 mil kilómetros con conexión de alta velocidad, de singapur a Francia, pasando por Egipto y el cuerno de África.	Neutro	El presidente de Estados Unidos, Joe Biden, anunció este domingo un proyecto para crear un cable submarino de fibra óptica que conectará Europa Occidental con Asia, como uno de los proyectos estrella del gran plan de infraestructuras lanzado por el G7 frente a china. Biden explicó…	Positivo

3 RESULTADOS

La estimación del impacto de los titulares de noticias publicados en Facebook se ha realizado en base a los datos recopilados de esta red social y de cada portal web del medio de comunicación considerado para este estudio. Los datos recopilados se clasificaron según el sentimiento transmitido mediante el uso de dos modelos previamente entrenados para el análisis del sentimiento. Estos datos se dividieron para entrenar y probar los modelos en una proporción de 70/30. Los datos textuales se representaron en forma numérica a través de la extracción de incrustaciones contextuales utilizando BERT para el idioma español. Una vez preparados los datos, se definieron cuatro algoritmos de aprendizaje automático supervisado para el desarrollo del modelo: CNN, SVM, RF y LR. Para entrenar estos modelos, se ajustaron los pesos de cada clase, donde a los pesos de las clases minoritarias se les asignó un valor más alto en comparación con las clases mayoritarias, lo que permitió que el modelo tuviera mayor sensibilidad hacia la clase minoritaria. Posteriormente, se evaluó el desempeño de los modelos, obteniendo los siguientes resultados:

- Los modelos entrenados con datos etiquetados por SAET mostraron una tendencia a clasificar el texto como POSITIVO, siendo esta la clase mayoritaria. En este caso, hay que tener en cuenta que la herramienta SAET clasifica el texto en función del léxico, que asigna un peso a cada palabra a partir de un diccionario de palabras según el sentimiento que transmite. Además, para ciertas palabras, se ha aumentado su peso teniendo en cuenta los dialectos y modismos. Estas particularidades influyen en la clasificación del texto ya que se considera que la herramienta está enfocada a evaluar el lenguaje informal expresado en los tuits.

- Por otro lado, los modelos entrenados con datos etiquetados por RoBERTuito mostraron una tendencia a clasificar el texto como NEUTRO, siendo esta la clase mayoritaria durante el entrenamiento. Como se describió anteriormente, RoBERTuito es un modelo pre entrenado para el análisis de texto en español basado en BERT, que ha sido entrenado con una cantidad considerable de texto y es capaz de interpretar el contexto de las palabras. Esta característica ha permitido clasificar el texto evaluando su contexto y cuyo resultado se compara con el principio de neutralidad que debe presentar el periodismo.

- El rendimiento de los modelos ha sido evaluado considerando la métrica F1-score, que para ambos casos presentaba una puntuación del 74% y del 85% como máximo. Así, se seleccionó el modelo CNN entrenado con datos etiquetados por RoBERTuito. La eficacia de este modelo se evaluó utilizando un conjunto de datos de 280 titulares y sus artículos etiquetados manualmente por estudiantes, que se compararon con la clasificación realizada por el modelo. Como resultado, se observa que la clasificación difiere en un 50% para cada clase. Este resultado puede considerarse subjetivo ya que el criterio de clasificación de cada estudiante es variable y las personas tienden a tener una percepción polarizada de las noticias [42].

El modelo ha permitido identificar el sentimiento que transmiten los titulares de noticias ecuatorianas publicados en Facebook y sus artículos. En general, el sentimiento que transmiten los titulares es mayoritariamente neutral, mientras que se lo puede considerar como polarizado en los artículos a los cuales estos titulares hacen referencia. Esto ha permitido identificar que la tendencia del periodismo digital en Ecuador es mantener la neutralidad en los titulares que publica. Sin embargo, este resultado difiere de las reacciones que expresan los usuarios a través de la red social. Se ha observado que, dependiendo del tema de la publicación, los usuarios la califican como positiva o negativa en función de una ideología adoptada por ellos. Esto sugiere que, independientemente del sentimiento expresado por el titular, los usuarios se encuentran en una posición que les parece adecuada (raramente esta postura es neutral).

3.1 ESTIMACIÓN DEL IMPACTO

Para estimar el impacto de las noticias se recogieron un total de 508 publicaciones de los seis principales medios de comunicación de Ecuador, las cuales han sido analizadas y clasificadas. La Figura 16 y la Figura 7 muestran cómo se distribuyen los datos según el sentimiento transmitido por cada medio de comunicación. Las publicaciones de cada medio mantienen la tendencia de neutralidad en sus publicaciones, sin embargo, el sentimiento transmitido en los artículos presenta polaridad. De acuerdo con esto, se asume que los artículos presentan esta polaridad con el fin de mantener la atención de los lectores.

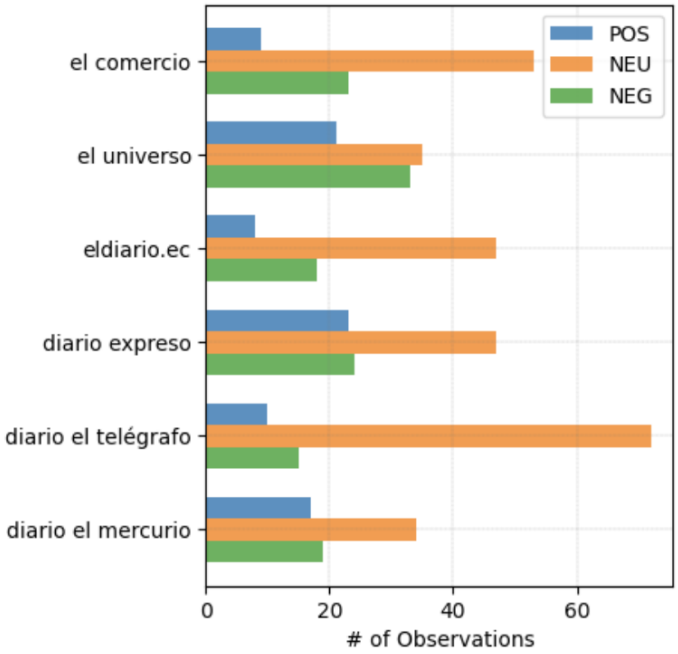

Figura 18: Estimación del impacto de titulares.

Elaborado por: El Autor.

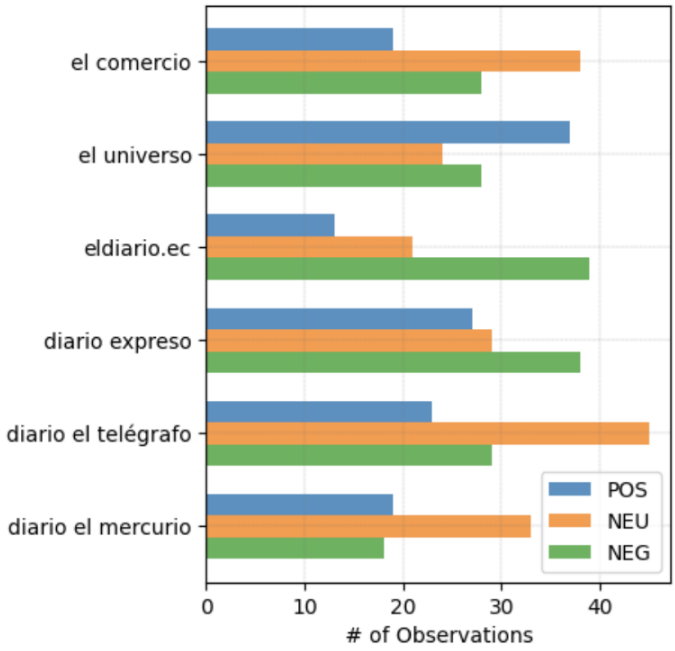

Figura 19: Estimación del impacto de artículos.

Elaborado por: El Autor.

Para analizar este comportamiento, se separaron los titulares cuyos artículos presentan diferentes sentimientos y se compararon sus polaridades. La Tabla 5 muestra el porcentaje de observaciones que no comparten el mismo sentimiento en el titular y en su artículo. El Comercio, El Diario y El Telégrafo son medios cuyas publicaciones superan el 40% de

diferencia. Este valor implica que sus publicaciones tienen un impacto definido en el lector, es decir, sus publicaciones tienen un propósito implícito. La Figura 18 muestra, en el caso de El Diario, la mayoría de sus noticias son negativas mientras que sus titulares son neutrales. En la Figura 19, en el caso de El Telégrafo, las noticias presentan sentimientos negativos y positivos. En la Figura 20, en el caso de El Comercio, las noticias presentan sentimientos negativos y positivos como en el caso anterior. Estos resultados muestran que las noticias, en su redacción, muestran una polaridad definida, siendo estas positivas o negativas, a diferencia de sus titulares que muestran neutralidad. Por tanto, se ha identificado que el impacto de los titulares publicados en Facebook es neutro frente a los artículos que presentan una valencia afectiva definida y mayor trascendencia.

Tabla 5: Distribución de datos.

DIARIO	OBSERVACIONES	DIFERENCIA	%
El Diario	73	42	57
El Comercio	85	35	41
El Telégrafo	97	39	40
Expreso	94	33	35
El Mercurio	70	25	35
El Universo	89	29	32

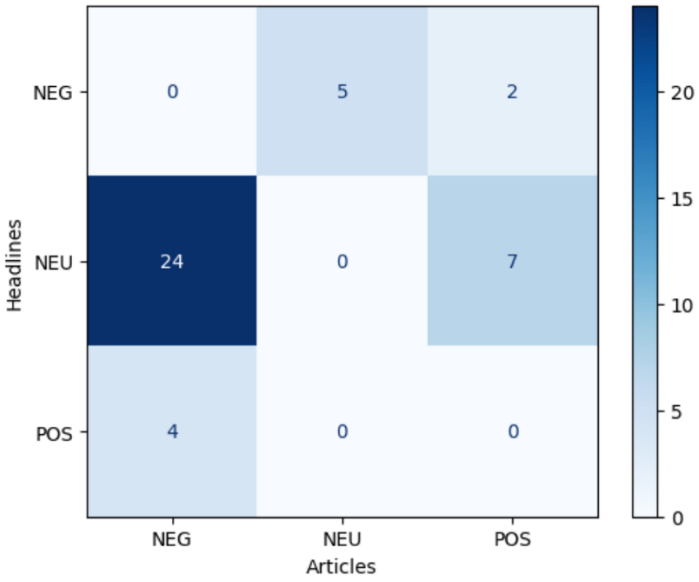

Figura 20: Mapa de calor para estimar el impacto de "El Diario".

Elaborado por: El Autor.

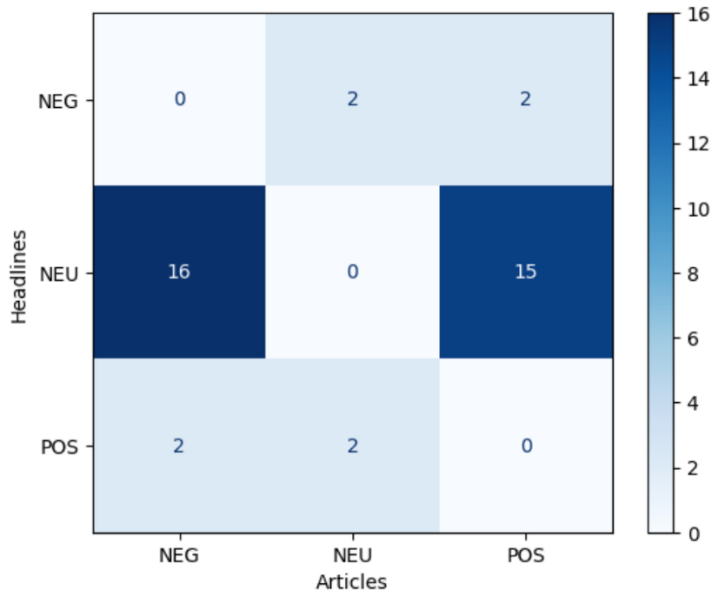

Figura 21: Mapa de calor para estimar el impacto de "El Telégrafo".

Elaborado por: El Autor.

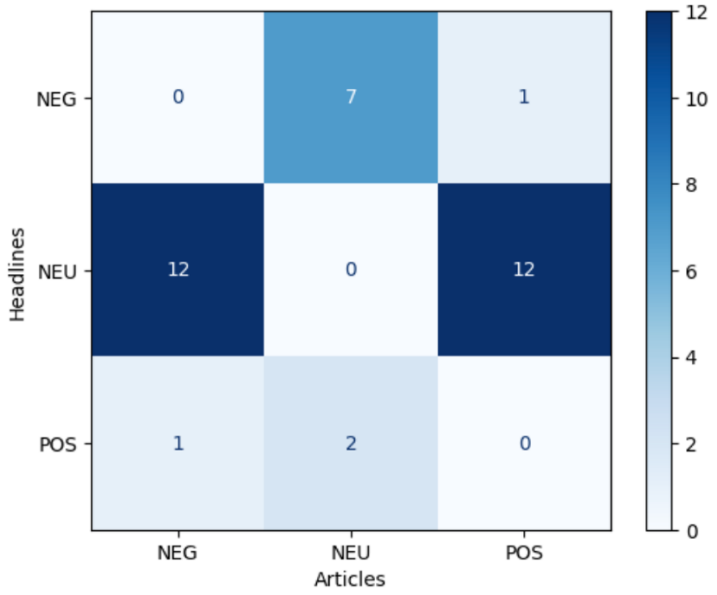

Figura 22: Mapa de calor para estimar el impacto de "El Comercio"

Elaborado por: El Autor.

3.2 DISCUSIÓN

El modelo desarrollado clasifica la mayoría de los titulares de noticias como neutrales, a diferencia de los artículos que se clasifican en positivos y negativos. Este comportamiento se puede interpretar considerando el contexto de las palabras. En el caso de los titulares,

el texto es corto y las palabras pueden no ser relevantes para el modelo y éste las toma como neutras. En cambio, en los artículos, el texto es extenso y se puede definir mejor el contexto de las palabras, dando como resultado la interpretación de negatividad y positividad. En contraste con este resultado, el modelo entrenado con datos etiquetados con la herramienta SAET, clasificó los datos en su mayoría como positivos y negativos. Este comportamiento se debió a que el SAET se basa en el léxico y sus palabras tienen un peso de acuerdo con su significado en el idioma español sin considerar el contexto. Adicionalmente, esta herramienta utiliza un diccionario de intensificadores en el que tienen más peso las palabras utilizadas en el léxico ecuatoriano.

Teniendo en cuenta que los datos analizados son titulares de noticias, consideramos que la forma de redacción es formal y sin modismos de la lengua de la región. Con esto, surgió la interrogante de si los titulares clasificados como neutrales por el modelo pertenecen a una sola categoría, es decir, si se mantiene la neutralidad para temas políticos, económicos, sociales, culturales, religiosos o deportivos. Siendo la neutralidad un principio básico del periodismo, los resultados obtenidos en este estudio demuestran que este principio es cumplido por el periodismo ecuatoriano, sin embargo, sería de interés conocer la temática en la que se han desarrollado.

A diferencia de los titulares, los artículos presentan una polaridad definida en el sentimiento que transmiten. Se observa una mayor distribución en las clases positivas y negativas. Esto demuestra que la redacción de los artículos busca captar la atención del lector y así cautivar a una mayor audiencia, lo cual tiene sentido considerando que esta práctica se ha vuelto común en el medio digital. Por otro lado, según [21], las noticias publicadas en las redes sociales tienden a estar polarizadas, es decir, positivas y negativas. Esto contrasta con el resultado obtenido, por lo que se realizó una evaluación y comparación del texto clasificado por este modelo y un grupo de estudiantes universitarios. En esta práctica se observó que efectivamente los titulares presentan una polaridad definida. Sin embargo, este resultado se vuelve subjetivo ya que el criterio de cada persona está influenciado por aspectos políticos, religiosos, sociales o culturales que, dependiendo de su preferencia, tomarán una decisión preestablecida.

Para evaluar estos resultados, una posible medida comparativa podría ser contrastarlos con las reacciones que dan los usuarios de Facebook a cada publicación. Es decir, se podrían analizar reacciones como "me gusta", "me divierte", "me encanta", "me entristece" y "me enoja" que utilizan los usuarios para calificar una publicación. De esta forma, se

podría validar el sentimiento evaluado por el modelo. Pero evaluar las reacciones para identificar el sentimiento de la publicación también se vuelve subjetivo. Se debe analizar qué tipo de reacción corresponde a un sentimiento positivo, negativo o neutro. Por ejemplo, en [2], los autores agrupan ciertos tipos de reacciones como positivos, negativos, neutros y no definidos, de acuerdo con los criterios establecidos para su estudio. Como puede verse, la evaluación de estas opciones depende enteramente del observador.

La clasificación realizada por el modelo presenta una ventaja frente a las consideraciones anteriores. Partiendo de que el modelo está basado en BERT, en su capa intermedia, la contextualización de las palabras se vuelve más precisa debido a su pre-entrenamiento con gran cantidad de datos textuales en español. Esto permite que la extracción de incrustaciones de palabras sea más efectiva teniendo en cuenta el contexto y el significado de las palabras en lugar de influir en sus criterios de clasificación por atributos preestablecidos como los diccionarios de datos. Es decir, los resultados obtenidos por el modelo son imparciales y no subjetivos. Por tanto, este modelo permitirá identificar el impacto que el periodista o editor transmite a través de la noticia a su audiencia, contribuyendo así a poder discernir el tipo de periodismo digital que realiza cada medio ecuatoriano e incluso poder regular esta conducta y preservar el principio de neutralidad e imparcialidad del periodismo evitando efectos negativos en la sociedad.

4 CONCLUSIONES

El análisis de sentimiento es una tarea del campo de NLP que nos ha permitido identificar el sentimiento que un texto transmite a sus lectores. Este sentimiento puede tener un impacto positivo, negativo o neutro, que cambia la perspectiva u opinión de una realidad social. Así, medir este impacto ayuda a comprender cómo es el comportamiento de un medio de comunicación a la hora de transmitir noticias de interés para la sociedad en la que se desarrolla.

El modelo desarrollado ha demostrado que, para este tipo de tareas, la interpretación del texto de forma contextual es importante debido a las relaciones que se generan entre las palabras y los diferentes significados que pueden tener en el contexto en el que se expresan. Así, BERT ha mostrado mejores resultados a la hora de analizar y clasificar el texto en comparación con otros algoritmos. En este estudio, el modelo CNN presentó mejores resultados sobre los otros modelos evaluados con un 85% en *F1-score* y un 88% en *accuracy*. Los resultados obtenidos definieron la selección del modelo para estimar el impacto de los titulares de noticias ecuatorianas publicados en Facebook. Así, el modelo permitió identificar el sentimiento transmitido y analizar el estilo del periodismo digital ecuatoriano.

De esta forma, se observó que el periodismo digital ecuatoriano mantiene el principio de neutralidad en sus publicaciones en Facebook, evitando el sensacionalismo que se ha convertido en una práctica común en las redes sociales. Sin embargo, este resultado no es el mismo para los artículos publicados en los sitios web del medio. Los artículos presentan en su mayoría sentimientos positivos y negativos. Esta diferencia se debe a que el texto es más largo que un titular y su contexto es más amplio (lo que puede evocar una postura definida en la audiencia).

Por otro lado, la clasificación realizada por el modelo puede ser objeto de análisis ya que los datos utilizados para entrenar el modelo no fueron categorizados por su contenido en políticos, religiosos, deportivos, culturales o educativos. Estas categorías podrían incidir en la apreciación de la validez de los resultados obtenidos ya que el ser humano se ve influenciado por factores externos para definir sus criterios. Así, para ciertas personas el impacto identificado por el modelo puede ser erróneo.

En este sentido, se puede validar la efectividad del modelo propuesto teniendo en cuenta los criterios expresados por los usuarios de Facebook. De esta forma, analizar el

sentimiento transmitido en reacciones como “me gusta”, “me divierte”, “me enoja”, “me entristece” y “me encanta” puede dar una pauta de cuál es la impresión del lector. Sin embargo, esta tarea implicaría que el criterio de clasificación del sentimiento podría analizarse considerando el contexto psicológico que la reacción puede representar. Por ejemplo, "me gusta" podría significar neutralidad o positividad, al igual que "jaja" podría significar sarcasmo o negatividad. Esta clasificación permitiría afinar el modelo propuesto y mejorar los resultados obtenidos.

De igual manera, la categorización de los datos por temas agregaría un criterio de evaluación adicional al tener en cuenta el alcance de la publicación. Así, un titular categorizado como deportivo podría tomarse como positivo en la mayoría de los casos y uno categorizado como político podría ser negativo. De esta forma, el impacto estimado para el titular se consideraría más preciso. La categorización previa de los datos para el entrenamiento del modelo ayudaría a aumentar la efectividad del modelo propuesto.

5 REFERENCIAS BIBLIOGRÁFICAS

[1] C. V. Baccarella, T. F. Wagner, J. H. Kietzmann, and I. P. McCarthy, “Social Media? it’s serious! understanding the dark side of social media,” *European Management Journal*, vol. 36, no. 4, pp. 431–438, 2018.

[2] R. Sandoval-Almazan and D. Valle-Cruz, “Facebook impact and sentiment analysis on political campaigns,” *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, pp.1-7,2018.

[3] H. Sebei, M. A. Hadj Taieb, and M. Ben Aouicha, “Review of social media analytics process and Big Data Pipeline,” *Social Network Analysis and Mining*, vol. 8, no. 1, 2018.

[4] C. Porlezza, “Accuracy in journalism,” *Oxford Research Encyclopedia of Communication*, 2019.

[5] C. S. Lee and L. Ma, “News sharing in social media: The effect of gratifications and prior experience,” *Computers in Human Behavior*, vol. 28, no. 2, pp. 331–339, 2012.

[6] Z. Li, Y. Fan, B. Jiang, T. Lei, and W. Liu, “A survey on sentiment analysis and opinion mining for Social Multimedia,” *Multimedia Tools and Applications*, vol. 78, no. 6, pp. 6939–6967, 2019.

[7] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu, “Fake news on social media: The impact on society,” *Information Systems Frontiers*, pp. 1-16, 2022.

[8] M. Guo and F.-S. Sun, “Like, comment, or share? exploring the effects of local television news Facebook posts on news engagement,” *Journal of Broadcasting & Electronic Media*, vol. 64, no. 5, pp. 736–755, 2020.

[9] P. T. Ngoc and M. Yoo, “The lexicon-based sentiment analysis for fan page ranking in Facebook,” *The International Conference on Information Networking 2014 (ICOIN2014)*, pp. 444–448, 2014.

[10] K. Ahmed, N. E. Tazi, and A. H. Hossny, “Sentiment analysis over social networks: An overview,” *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2174–2179, 2015.

[11] T. Jo, “Text mining - Concepts, Implementation, and Big Data Challenge,” *Studies in Big Data*, pp. 3–5, 2019.

[12] J. Vásquez, H. Gómez-Adorno, and, G. Bel-Enguix, “Bert-based Approach for Sentiment Analysis of Spanish Reviews from TripAdvisor” *IberLEF@ SEPLN*, pp. 165-170, 2021.

[13] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM process model,” *Procedia Computer Science*, vol. 181, pp. 526–534, 2021.

[14] S. Selva Birunda and R. Kanniga Devi, “A review on word embedding techniques for text classification,” *Innovative Data Communication Technologies and Application*, pp. 267–281, 2021.

[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4171-4186, 2019.

[16] J. Rieis et al., “Breaking the news: First Impressions matter on online news,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 9, no. 1, pp. 357–366, 2021.

[17] S. Yoon et al., “Detecting incongruity between news headline and body text via a deep hierarchical encoder,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 791–800, 2019.

[18] K. Park, H. Kwak, J. An, and S. Chawla, “How-to present news on social media: A causal analysis of editing news headlines for Boosting User Engagement,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, pp. 491–502, 2021.

[19] K. Welbers and M. Opgenhaffen, “Presenting news on social media,” *Digital Journalism*, vol. 7, no. 1, pp. 45–62, 2018.

[20] N. Kumar, R. Nagalla, T. Marwah, and M. Singh, “Sentiment Dynamics in social media news channels,” *Online Social Networks and Media*, vol. 8, pp. 42–54, 2018.

[21] A. Piotrkowicz, V. Dimitrova, J. Otterbacher, and K. Markert, “Headlines matter: Using headlines to predict the popularity of news articles on Twitter and Facebook,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, pp. 656–659, 2017.

[22] T. H. Nguyen, K. Shirai, and J. Velcin, “Sentiment analysis on social media for stock movement prediction,” *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603–9611, 2015.

[23] H. Ishijima, T. Kazumi, and A. Maeda, “Sentiment analysis for the Japanese stock market,” *Global Business and Economics Review*, vol. 17, no. 3, pp. 237–255, 2015.

[24] A. M. Idrees, F. Kamal, and A. I., “A proposed model for detecting Facebook news’ credibility*,” International Journal of Advanced Computer Science and Applications*, vol. 10, no. 7, 2019.

[25] K. Denecke and Y. Deng, “Sentiment analysis in medical settings: New opportunities and challenges,” *Artificial Intelligence in Medicine*, vol. 64, no. 1, pp. 17–27, 2015.

[26] M. Birjali, M. Kasri, and A. Beni-Hssane, “A comprehensive survey on sentiment analysis: Approaches, challenges and Trends,” *Knowledge-Based Systems*, vol. 226, p. 107134, 2021.

[27] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, “The application of text mining methods in innovation research: Current State, evolution patterns, and development priorities,” *R&D Management*, vol. 50, no. 3, pp. 329–351, 2020.

[28] K. Zvarevashe and O. O. Olugbara, “A framework for sentiment analysis with opinion mining of Hotel Reviews,” 2018 *Conference on Information Communications Technology and Society (ICTAS)*, pp. 1–4, 2018.

[29] I. Utitiaj, P. Morillo, and D. V. Huanga, “Sentiment analysis tool for Spanish tweets in the Ecuadorian context,” 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence, pp. 1–6, 2020.

[30] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[31] S. Biswas, “Scope of sentiment analysis on news articles regarding stock market and GDP in struggling economic condition,” *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 7, pp. 3594–3609, 2020.

[32] Azam, M. et al. "Feature extraction-based text classification using k-nearest neighbor algorithm." *International Journal of Computer Science and Network Security*, vol. 18, no. 12, pp. 95-101, 2018.

[33] F. Rustam et al., “A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis,” *PLOS ONE*, vol. 16, no. 2, 2021.

[34] [R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, “The impact of features extraction on the sentiment analysis,” *Procedia Computer Science*, vol. 152, pp. 341–348, 2019.

[35] H. Ahmed, I. Traore, and S. Saad, “Detection of online fake news using N-gram analysis and Machine Learning Techniques,” *Lecture Notes in Computer Science*, pp. 127–138, 2017.

[36] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, “Transformer models for text-based Emotion Detection: A review of Bert-based approaches,” *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5789–5829, 2021.

[37] D. E. Cahyani and I. Patasik, “Performance comparison of TF-IDF and word2vec models for emotion text classification,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, 2021.

[38] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, “Comparing pre-trained language models for Spanish hate speech detection,” *Expert Systems with Applications*, vol. 166, p. 114120, 2021.

[39] J. Pérez, D. Furman, Laura. Alemany, and F. Luque. “RoBERTuito: a pre-trained language model for social media text in Spanish” *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* pp 7235–7243, 2022.

[40] S. Hassan, M. Fernandez, and A. Harith, “On stopwords, filtering and data sparsity for sentiment analysis of twitter.” *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC'14),* pp. 810-817, 2014.

[41] M. Hossin and S. M.N, “A review on evaluation metrics for Data Classification Evaluations,” *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, 2015.

[42] R. Fletcher, A. Cornia, and R. K. Nielsen, “How polarized are online and offline news audiences? A comparative analysis of twelve countries,” *The International Journal of Press/Politics*, vol. 25, no. 2, pp. 169–195, 2019.

6 ANEXOS

6.1 REPOSITORIO DE DATOS

- https://epnecuador-my.sharepoint.com/:f:/g/personal/roberto_lema_epn_edu_ec/EirIw5xCsTFPuZdxDl38yQkBTNdmT4kOtTZ3XgWprTk4cg?e=VJou9R