



UNIVERSIDAD DE GUAYAQUIL
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA SISTEMAS DE INFORMACIÓN

TRABAJO DE TITULACIÓN
PREVIO A LA OBTENCIÓN DEL TÍTULO DE
INGENIERA EN SISTEMAS DE INFORMACIÓN

ÁREA
CIENCIAS BÁSICAS, BIOCONOCIMIENTO Y
DESARROLLO INDUSTRIAL

TEMA
ANÁLISIS DE SENTIMIENTOS DE LAS NOTICIAS
COMPROBADAS EN TWITTER POR LAS
VERIFICADORAS ACREDITADAS EN ECUADOR
UTILIZANDO PROCESAMIENTO DE
LENGUAJE NATURAL.

AUTORA
JIMÉNEZ OLIVO KIMBERLY ANTONELLA

DIRECTORA DEL TRABAJO
LSI. TOAPANTA BERNABÉ MARIUXI DEL CARMEN, MSIG

GUAYAQUIL, SEPTIEMBRE 2023

ANEXO X.- FICHA DE REGISTRO DE INTEGRACIÓN CURRICULAR
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA SISTEMAS DE INFORMACIÓN
MODALIDAD SEMESTRAL

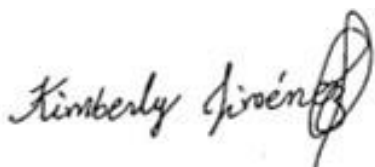
REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA			
FICHA DE REGISTRO DE TRABAJO DE INTEGRACIÓN CURRICULAR			
TÍTULO Y SUBTÍTULO:	ANÁLISIS DE SENTIMIENTOS DE LAS NOTICIAS COMPROBADAS EN TWITTER POR LAS VERIFICADORAS ACREDITADAS EN ECUADOR UTILIZANDO PROCESAMIENTO DE LENGUAJE NATURAL.		
AUTORA:	JIMÉNEZ OLIVO KIMBERLY ANTONELLA		
REVISOR(ES)/TUTOR(ES) (apellidos/nombres):	LSI. TOAPANTA BERNABÉ MARIUXI DEL CARMEN, MSIG. ING. TEJADA CASTRO MARIUXI ILEANA, MGW.		
INSTITUCIÓN:	UNIVERSIDAD DE GUAYAQUIL		
UNIDAD/FACULTAD:	INGENIERÍA INDUSTRIAL		
MAESTRÍA/ESPECIALIDAD:			
GRADO OBTENIDO:	INGENIERA EN SISTEMAS DE INFORMACIÓN		
FECHA DE PUBLICACIÓN:	SEPTIEMBRE 2023	No. DE PÁGINAS:	107
ÁREAS TEMÁTICAS:	CIENCIAS BÁSICAS, BIOCONOCIMIENTO Y DESARROLLO INDUSTRIAL		
PALABRAS CLAVES/ KEYWORDS:	TWITTER , MODELO, PROCESAMIENTO, ANÁLISIS, SENTIMIENTOS.		
RESUMEN/ABSTRACT (150): El presente proyecto tiene como propósito desarrollar un componente de análisis de sentimientos de las noticias en Twitter seleccionadas por las verificadoras de hechos en Ecuador, utilizando Procesamiento de Lenguaje Natural, que sirva de aporte a una investigación a nivel macro que busca ayudar al fact-checker a realizar el proceso de verificación de noticias. Para realizarlo, se emplearon técnicas de recolección de datos como la entrevista, técnicas de análisis de sentimientos y modelos de aprendizaje automático. En su implementación, se utilizó la base de datos en la nube MongoDB Atlas para obtener los datos requeridos, el lenguaje de programación Python utilizando el cuaderno colaborativo Google Colab y el modelo preentrenado de BERT en español como base para entrenar un modelo que pueda realizar predicciones de sentimientos y emociones. Además, se emplearon métricas para evaluar el rendimiento del modelo entrenado. Estas herramientas mencionadas contribuyeron para que se cumplan los objetivos propuestos.			
ADJUNTO PDF:	SI (X)	NO	
CONTACTO CON AUTOR/ES:	Teléfono: 0968165498	E-mail: kimberly.jimenez@ug.edu.ec	
CONTACTO CON LA INSTITUCIÓN:	Nombre: ING. HURTADO PASPUEL JIMMY FERNANDO, MSC		
	Teléfono: 042-658128		
	E-mail: titulación.sistemas.industrial@ug.edu.ec		

**ANEXO XI.- DECLARACIÓN DE AUTORÍA Y DE AUTORIZACIÓN DE
LICENCIA GRATUITA INTRANSFERIBLE Y NO EXCLUSIVA PARA EL USO
NO COMERCIAL DE LA OBRA CON FINES NO ACADÉMICOS.**

**FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA SISTEMAS DE INFORMACIÓN
MODALIDAD SEMESTRAL**

**LICENCIA GRATUITA INTRANSFERIBLE Y NO COMERCIAL DE LA OBRA
CON FINES NO ACADÉMICOS**

Yo, **JIMÉNEZ OLIVO KIMBERLY ANTONELLA**, con C.I. No. **0931077416**, certifico que los contenidos desarrollados en este trabajo de integración curricular, cuyo título es **ANÁLISIS DE SENTIMIENTOS DE LAS NOTICIAS COMPROBADAS EN TWITTER POR LAS VERIFICADORAS ACREDITADAS EN ECUADOR UTILIZANDO PROCESAMIENTO DE LENGUAJE NATURAL.**, son de mi absoluta propiedad y responsabilidad, en conformidad al Artículo 114 del CÓDIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN, autorizo la utilización de una licencia gratuita intransferible, para el uso no comercial de la presente obra a favor de la Universidad de Guayaquil.

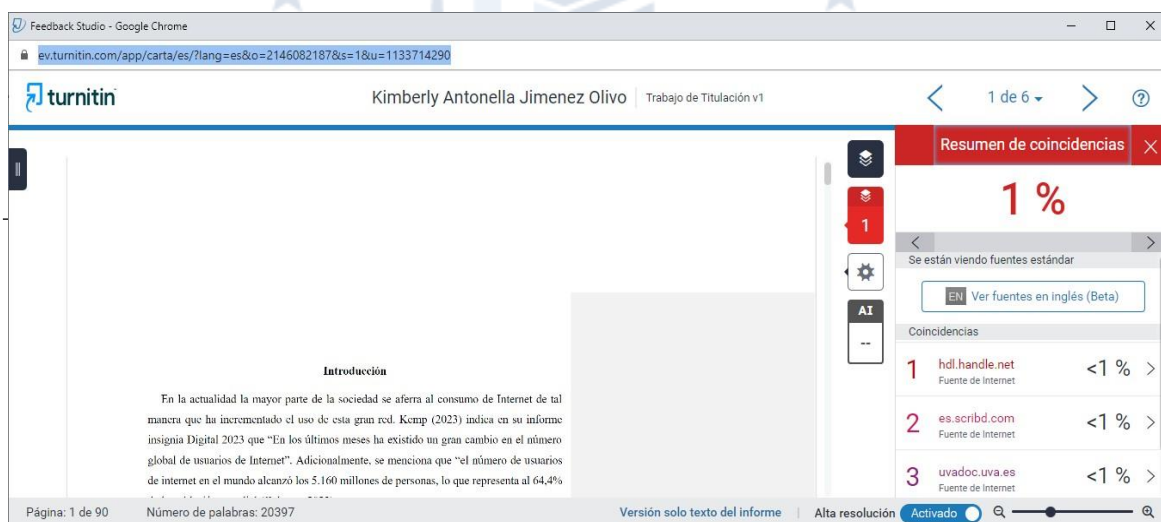


JIMÉNEZ OLIVO KIMBERLY ANTONELLA
C.I. No. 0931077416

**FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA: SISTEMAS DE INFORMACIÓN
MODALIDAD SEMESTRAL**

Habiendo sido nombrado **LSI. TOAPANTA BERNABÉ MARIUXI DEL CARMEN, MSIG.**, tutor del trabajo de integración curricular certifico que el presente trabajo ha sido elaborado por **JIMÉNEZ OLIVO KIMBERLY ANTONELLA**, con mi respectiva supervisión como requerimiento parcial para la obtención del título de **INGENIERA EN SISTEMAS DE INFORMACIÓN**.

Se informa que el trabajo de integración curricular: **ANÁLISIS DE SENTIMIENTOS DE LAS NOTICIAS COMPROBADAS EN TWITTER POR LAS VERIFICADORAS ACREDITADAS EN ECUADOR UTILIZANDO PROCESAMIENTO DE LENGUAJE NATURAL.**, ha sido orientado durante todo el periodo de ejecución en el programa antiplagio **TURNITIN** quedando el **1%** de coincidencia.



The screenshot shows the Turnitin web interface in Google Chrome. The browser address bar displays the URL: <https://ev.turnitin.com/app/carta/es/?lang=es&o=2146082187&s=1&u=1133714290>. The page header identifies the user as "Kimberly Antonella Jimenez Olivo" and the document as "Trabajo de Titulación v1". A red banner at the top right indicates a "Resumen de coincidencias" (Summary of similarities) with a large "1 %" score. Below this, a table lists the sources of similarity:

Se están viendo fuentes estándar		
EN Ver fuentes en inglés (Beta)		
Coincidencias		
1	hdl.handle.net Fuente de Internet	<1 % >
2	es.scribd.com Fuente de Internet	<1 % >
3	uvadoc.uva.es Fuente de Internet	<1 % >

The main content area shows the beginning of a document titled "Introducción". The text reads: "En la actualidad la mayor parte de la sociedad se aferra al consumo de Internet de tal manera que ha incrementado el uso de esta gran red. Kemp (2023) indica en su informe insignia Digital 2023 que "En los últimos meses ha existido un gran cambio en el número global de usuarios de Internet". Adicionalmente, se menciona que "el número de usuarios de internet en el mundo alcanzó los 5.160 millones de personas, lo que representa al 64,4%".

At the bottom of the interface, it shows "Página: 1 de 90" and "Número de palabras: 20397". There are also links for "Versión solo texto del informe" and "Alta resolución", and a search bar with the word "Activado".

<https://ev.turnitin.com/app/carta/es/?lang=es&o=2146082187&s=1&u=1133714290>



Firmado electrónicamente por:
**MARIUXI DEL CARMEN
TOAPANTA BERNABE**

**LSI. TOAPANTA BERNABÉ MARIUXI DEL CARMEN, MSIG.
DOCENTE TUTOR
C.C. 0916653447
FECHA: 15 DE AGOSTO DEL 2023**



ANEXO V. - CERTIFICADO DEL DOCENTE-TUTOR DEL TRABAJO DE INTEGRACIÓN CURRICULAR

**FACULTAD DE INGENIERIA INDUSTRIAL
CARRERA SISTEMAS DE INFORMACIÓN
MODALIDAD SEMESTRAL**

Guayaquil, 15 de agosto de 2023

Sr.

**ING. CABEZAS GALARZA FRANKLIN AUGUSTO, MAE.
DIRECTOR (A) DE LA CARRERA DE SISTEMAS DE INFORMACIÓN
FACULTAD DE INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE GUAYAQUIL
Ciudad. -**

De mis consideraciones:

Envío a Ud. el Informe correspondiente a la tutoría realizada al Trabajo de integración curricular **ANÁLISIS DE SENTIMIENTOS DE LAS NOTICIAS COMPROBADAS EN TWITTER POR LAS VERIFICADORAS ACREDITADAS EN ECUADOR UTILIZANDO PROCESAMIENTO DE LENGUAJE NATURAL.** de la estudiante **JIMÉNEZ OLIVO KIMBERLY ANTONELLA**, indicando que ha cumplido con todos los parámetros establecidos en la normativa vigente:

- El trabajo es el resultado de una investigación.
- El estudiante demuestra conocimiento profesional integral.
- El trabajo presenta una propuesta en el área de conocimiento.
- El nivel de argumentación es coherente con el campo de conocimiento.

Adicionalmente, se adjunta el certificado de porcentaje de similitud y la valoración del trabajo de integración curricular con la respectiva calificación.

Dando por concluida esta tutoría de trabajo de integración curricular, **CERTIFICO**, para los fines pertinentes, que la estudiante **JIMÉNEZ OLIVO KIMBERLY ANTONELLA** está apta para continuar con el proceso de revisión final.

Atentamente,



Firmado electrónicamente por:
**MARIUXI DEL CARMEN
TOAPANTA BERNABÉ**

**LSI. TOAPANTA BERNABÉ MARIUXI DEL CARMEN, MSIG.
TUTOR DE TRABAJO DE INTEGRACIÓN CURRICULAR
C.I. 0916653447
FECHA: 15 DE AGOSTO DEL 2023**



ANEXO VII.- INFORME DEL DOCENTE REVISOR



FACULTAD DE INGENIERÍA INDUSTRIAL CARRERA SISTEMAS DE INFORMACIÓN MODALIDAD SEMESTRAL

Guayaquil, 18 de agosto de 2023

SR.

**ING. CABEZAS GALARZA FRANKLIN AUGUSTO, MAE.
DIRECTOR DE LA CARRERA DE SISTEMAS DE INFORMACIÓN
FACULTAD DE INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE GUAYAQUIL
Ciudad. -**

De mis consideraciones:

Envío a Ud. el Informe correspondiente a la REVISIÓN FINAL del trabajo de integración curricular **ANÁLISIS DE SENTIMIENTOS DE LAS NOTICIAS COMPROBADAS EN TWITTER POR LAS VERIFICADORAS ACREDITADAS EN ECUADOR UTILIZANDO PROCESAMIENTO DE LENGUAJE NATURAL**, de la estudiante **JIMÉNEZ OLIVO KIMBERLY ANTONELLA**. Las gestiones realizadas me permiten indicar que el trabajo fue revisado considerando todos los parámetros establecidos en las normativas vigentes, en el cumplimiento de los siguientes aspectos:

Cumplimiento de requisitos de forma:

El título tiene un máximo de **20** palabras.

La memoria escrita se ajusta a la estructura establecida.

El documento se ajusta a las normas de escritura científica seleccionadas por la Facultad. La investigación es pertinente con la línea y sublíneas de investigación de la carrera.

Los soportes teóricos son de máximo **5** años.

La propuesta presentada es pertinente.

Cumplimiento con el Reglamento de Régimen Académico:

El trabajo es el resultado de una investigación.

El estudiante demuestra conocimiento profesional integral. El trabajo presenta una propuesta en el área de conocimiento.

El nivel de argumentación es coherente con el campo de conocimiento.

Adicionalmente, se indica que fue revisado el certificado de porcentaje de similitud, la valoración del tutor, así como de las páginas preliminares solicitadas, lo cual indica el que el trabajo de investigación cumple con los requisitos exigidos.

Una vez concluida esta revisión, considero que el estudiante está apto para continuar el proceso de integración curricular.

Particular que comunicamos a usted para los fines pertinentes.

Atentamente,



Firmado electrónicamente por:
**MARIUXI ILEANA
TEJADA CASTRO**

**ING. TEJADA CASTRO MARIUXI ILEANA, MGW.
C.I. 0920540259
FECHA: 18 DE AGOSTO DEL 2023**

Dedicatoria

Dedico mi trabajo de titulación en primer lugar a Dios, por darme la fuerza para seguir adelante en cada meta que me propongo en la vida. A mi pequeño hijo, Daniel Ayluardo Jiménez, quien con su llegada ha iluminado mi vida y ha sido un motivo de inspiración para seguir luchando por un futuro mejor, a mis padres, Nérída María Olivo Decimavilla y Luis Arol Jiménez Bajaña, por su amor y apoyo incondicional en todo lo que me propongo a nivel personal y profesional, a mis hermanos, Arlette Jiménez y Ángel Manzano quienes no han dejado de creer en mí. A mi compañero de vida Roberto Ayluardo quien me ha brindado su amor y apoyo cuando lo he requerido.

Agradecimiento

Agradezco a Dios, por guiarme y tener el privilegio de finalizar mis estudios de tercer nivel. A mis padres, Nérida María Olivo Decimavilla y Luis Arol Jiménez Bajaña que se esforzaron para que pudiera tener lo necesario para prepararme profesionalmente y me brindaron sus sabios consejos para no rendirme en los momentos difíciles de la carrera.

A mi familia, por apoyarme emocionalmente y motivarme a no rendirme por más difícil que sea la situación y a mis mejores amigos quienes me han motivado y acompañado a lo largo de este proceso.

A mi tutora, LSI. TOAPANTA BERNABÉ MARIUXI DEL CARMEN, MSIG. por aportar con sus conocimientos en mi trabajo de titulación.

Índice General

N°	Descripción	Pág.
	Introducción	1

Capítulo I Marco Teórico

N°	Descripción	Pág.
1.1.	Planteamiento del Problema	2
1.2.	Formulación del Problema	4
1.3.	Objeto de Estudio	4
1.4.	Delimitación del objeto de investigación	4
1.4.1.	Delimitación Geográfica.	4
1.4.2.	Delimitación en Tiempo – Espacio.	5
1.4.3.	Delimitación Semántica.	5
1.5.	Justificación	7
1.6.	Alcance	10
1.7.	Objetivos	11
1.7.1.	Objetivo General.	11
1.7.2.	Objetivos Específicos.	11
1.8.	Marco Teórico	12
1.8.1.	Desinformación en las redes sociales.	12
1.8.1.1.	Noticias falsas como parte de la desinformación en la red social Twitter.	12
1.8.1.2.	Factores asociados a las noticias falsas.	13
1.8.2.	Fact-Checking para la corroboración de contenido.	13
1.8.2.1.	Métodos para la verificación de hechos.	15
1.8.2.2.	Plataformas de fact-checking.	17
1.8.3.	Verificadores de hechos acreditados en Ecuador.	18
1.8.3.1.	Verificador pionero a nivel nacional – Ecuador Chequea.	18
1.8.3.2.	Verificador de hechos Ecuador Verifica.	18
1.8.4.	Metodologías para la comprobación de veracidad de noticias.	19

N°	Descripción	Pág.
1.8.4.1.	Metodología de Ecuador Chequea para la comprobación de veracidad de las noticias.	19
1.8.4.2.	Metodología de Ecuador Verifica para la detección de noticias falsas.	20
1.8.5.	PLN y aplicación de su campo de desarrollo para la detección de sentimientos y emociones.	21
1.9.	Marco Conceptual	22
1.9.1.	La Inteligencia Artificial.	22
1.9.1.1.	Capacidades de la Inteligencia Artificial.	23
1.9.2.	Machine learning.	23
1.9.3.	Procesamiento del lenguaje natural (PLN).	24
1.9.3.1.	Campos del procesamiento de lenguaje natural.	24
1.9.4.	Análisis de sentimientos.	24
1.9.4.1.	Niveles del análisis de sentimiento.	25
1.9.4.2.	Tareas del análisis de sentimiento.	25
1.9.4.3.	Técnicas para la clasificación de los sentimientos.	25
1.9.5.	Preprocesamiento de los datos.	26
1.9.5.1.	Lematización.	27
1.9.5.2.	Tokenización.	28
1.9.6.	Metodologías en el desarrollo de proyectos de ciencia de datos.	28
1.9.6.1.	Metodología KDD (Knowledge Discovery in Databases).	28
1.9.6.2.	Metodología CRISP-DM.	29
1.9.6.3.	Metodología SEMMA (Sample, Explore, Modify, Model and Access)	30
1.9.7.	Algoritmos para análisis de sentimientos.	31
1.9.8.	Cuadernos colaborativos para ciencia de datos.	31
1.9.8.1.	Deepnote.	32
1.9.8.2.	Jupyter.	32
1.9.8.3.	Google Colaboratory.	32
1.9.9.	IDEs de programación.	32
1.9.9.1.	Visual Studio Code.	33
1.9.9.2.	RStudio.	33

N°	Descripción	Pág.
1.9.9.3.	Spyder.	33
1.9.10.	Lenguajes de programación.	33
1.9.10.1.	Python.	33
1.9.10.2.	R.	33
1.9.11.	APIs y librerías.	33
1.9.11.1.	Natural language toolkit (NLTK).	33
1.9.11.2.	Textblob.	34
1.9.11.3.	Scikit-Learn.	34
1.9.12.	Arquitectura de la propuesta.	34
1.10.	Marco Legal	34
1.10.1.	Ley orgánica de comunicación (LOC).	35
1.10.2.	Código de principios de la IFCN.	35
1.10.2.1.	Política de rectificaciones de Ecuador Chequea y Ecuador Verifica.	36
1.10.3.	Política de Twitter para desarrolladores.	36

Capítulo II Metodología

N°	Descripción	Pág.
2.1.	Tipo de investigación	38
2.1.1.	Investigación descriptiva.	38
2.1.2.	Investigación exploratoria.	38
2.2.	Pregunta de investigación	38
2.2.1.	Enfoque de la investigación.	38
2.2.1.1.	Enfoque cualitativo.	38
2.3.	Técnicas de recolección de datos	39
2.3.1.	Entrevista.	39
2.3.1.1.	Aplicación de entrevista.	39
2.2.1.1	Resumen de la entrevista realizada a la periodista de Ecuador Chequea.	40

Nº	Descripción	Pág.
2.3.2.	Análisis documental.	41
2.2.1.2	Aplicación del análisis documental.	41
2.4.	Recopilación del conjunto de datos	42
2.5.	Materialización de variables	42
2.6.	Descripción del conjunto de datos	43
2.7.	Análisis exploratorio	45
2.7.1.	Visualización del conjunto de datos	45
2.7.2.	Preprocesamiento del conjunto de datos	45
2.7.3.	Desarrollo del análisis exploratorio del conjunto de datos.	48

Capítulo III

Desarrollo de la Propuesta

Nº	Descripción	Pág.
3.1.	Selección de Variables	53
3.1.1.	Preprocesamiento del texto de las noticias.	54
3.2.	Modelos Aplicados	58
3.2.1.	Modelo Base.	60
3.2.1.1.	Tokenización en el modelo base.	60
3.2.1.2.	Padding en el modelo base.	61
3.2.1.3.	Codificación en el modelo base	62
3.2.1.4.	Análisis de sentimientos con el Modelo Preentrenado de BERT en español.	63
3.2.2.	Modelo Propuesto	65
3.2.3.	Fase de Entrenamiento.	66
3.2.3.1.	Tokenización.	66
3.2.3.2.	Padding y Codificación.	66
3.2.3.3.	Creación de tensores de PyTorch.	67

N°	Descripción	Pág.
3.2.3.4.	División de datos en conjuntos de entrenamiento y prueba.	67
3.2.3.5.	Creación de conjuntos de datos de entrenamiento y prueba.	68
3.2.3.6.	Definición del Modelo Preentrenado de BERT.	68
3.2.4.	Selección de parámetros e hiperparámetros.	69
3.2.4.1.	Hiperparámetros del modelo.	69
3.2.4.2.	Creación de cargadores de datos.	69
3.2.4.3.	Definición de la función de pérdida y optimizador.	70
3.2.4.4.	Entrenamiento del modelo.	70
3.3.	Evaluación	71
3.3.1.	Métricas.	72
3.3.1.1.	Precisión (Accuracy).	72
3.3.1.2.	Precisión, recall y F1-score para sentimientos.	73
3.3.1.3.	Precisión, recall y F1-score para emociones	74
3.3.1.4.	Matriz de Confusión.	75
3.4.	Resultados	76
3.4.1.	Predicciones de sentimientos y emociones.	77
3.5.	Análisis de Resultados	78
3.6.	Conclusiones	79
3.7.	Recomendaciones	79
	ANEXOS	81
	Bibliografía	83

Índice de Tablas

N°	Descripción	Pág.
1.	Tácticas del proceso de verificación de noticias.	15
2.	Comparación de las metodologías en ciencia de datos.	30
3.	Algoritmos de análisis de sentimientos.	31
4.	Entrevista dirigida a la periodista de Ecuador Chequea.	39
5.	Entrevista dirigida a la periodista de Ecuador Chequea.	40

N°	Descripción	Pág.
6.	Materialización de variables.	43
7.	Descripción del conjunto de datos Sentimientos.	44

Índice de Figuras

N°	Descripción	Pág.
1.	Localización de los medios de verificación	2
2.	Proceso de verificación de información	3
3.	Ubicación del verificador de hechos Ecuador Chequea	5
4.	Ubicación del verificador de hechos Ecuador Verifica	5
5.	Pasos para verificar la veracidad de una noticia	8
6.	Etapas del proceso de verificación de Ecuador Verifica	10
7.	Factores asociados a fake news	13
8.	Metodología de trabajo general en Fact-Checking	16
9.	Códigos de Principios Deontológicos	16
10.	Métodos de verificación	17
11.	Calificaciones o categorías de Ecuador Chequea	20
12.	Calificaciones o categorías de Ecuador Verifica	21
13.	Etapas del proceso de verificación de Ecuador Verifica	21
14.	Capacidad de una máquina inteligente según el test de Turing	23
15.	Tareas del análisis del sentimiento	25
16.	Técnicas para la clasificación de sentimientos	26
17.	Proceso de análisis de sentimiento	26
18.	Ejemplo de texto preprocesado	27
19.	Ciclo de vida de minería de datos	29
20.	EDA - Visualización general del conjunto de datos	45
21.	EDA – Número de columnas y filas en el conjunto de datos sin preprocesar	45
22.	EDA – Nombres de columnas en el conjunto de datos sin preprocesar	46
23.	EDA – Eliminación de columnas	46
24.	EDA – Renombramiento de columnas	46
25.	EDA – Visualización del tipo de variables	47
26.	EDA – Corroboración de valores faltantes o perdidos	47
27.	EDA – Corroboración de valores faltantes o perdidos	48

Nº	Descripción	Pág.
28.	EDA – Gráfico de barras de Screen name de usuarios	48
29.	EDA – Verificación de user_name	49
30.	EDA – Gráfico de barras de nombres de usuarios	49
31.	EDA – Sentencia de boxplot para la comparativa	50
32.	EDA – Boxplot comparativo de la cantidad de retweets entre un usuario fact-checker y un usuario común	50
33.	EDA – Tweet_id con mayor número de likes	51
34.	EDA – Gráfico Pastel Tweet_id con mayor número de likes	51
35.	EDA – Información Tweet_id con mayor número de likes	51
36.	EDA – Información Tweet_id con mayor número de likes	52
37.	Concatenación de DataFrames	53
38.	Creación del DataFrame df_nuevo con variables seleccionadas	54
39.	Texto limpio de expresiones regulares	55
40.	Descarga de stopwords	55
41.	Eliminación de stopwords	55
42.	Textos sin stopwords	55
43.	Función para verificar existencia de emoticones	56
44.	Verificación de existencia de emoticones	56
45.	Verificación de existencia de emoticones referentes a sentimientos tratados	57
46.	Texto ejemplo aplicando la función detectar_emoticones_sentimientos	57
47.	Texto sin emoticones	58
48.	Función para limpiar caracteres especiales en los textos	58
49.	Texto sin caracteres especiales	58
50.	Análisis de sentimientos con Bert en español – categoría binaria multilingual	59
51.	Análisis de sentimientos con Bert en español – categoría sentimiento multilingual	59
52.	Texto identificado con sentimiento negativo con Bert multilingual	60
53.	Carga de BertTokenizer en español – Modelo Base	60
54.	Función para tokenizar texto con Bert Tokenizer – Modelo Base	61
55.	Texto tokenizado con Bert Tokenizer – Modelo Base	61
56.	Texto con más tokens – Modelo Base	61
57.	Texto con más tokens – Modelo Base	62

N°	Descripción	Pág.
58.	Textos padeados – Modelo Base	62
59.	Función para codificar con Bert – Modelo Base	62
60.	Codificación de los textos – Modelo Base	63
61.	Visualización de los textos codificados – Modelo Base	63
62.	Función de análisis de sentimientos con Bert – Modelo Base	64
63.	Análisis de sentimientos con Bert en español categoría binaria – Modelo Base	64
64.	Análisis de sentimientos con Bert en español categoría spanish.–Modelo Base	65
65.	Texto con sentimiento negativo con Bert en español – Modelo Base	65
66.	Subconjunto de datos etiquetados	66
67.	Tokenización de textos en Modelo Propuesto	66
68.	Codificación de textos en Modelo Propuesto	67
69.	Conversión de etiquetas a números	67
70.	Creación de tensores de PyTorch	67
71.	División en conjunto de entrenamiento y prueba	68
72.	Creación de conjuntos de datos de entrenamiento y prueba	68
73.	Definición del modelo preentrenado de BERT	68
74.	Hiperparámetros de entrenamiento	69
75.	Creación de cargadores de datos	69
76.	Hiperparámetros de entrenamiento	70
77.	Función de pérdida y optimizador	71
78.	Pérdida promedio del modelo	71
79.	Evaluación del modelo	72
80.	Precisión general del modelo	73
81.	Calcular métricas en sentimientos	73
82.	Reporte de clasificación para sentimientos	74
83.	Calcular métricas en emociones	74
84.	Reporte de clasificación para emociones	74
85.	Matriz de confusión para sentimiento	75
86.	Matriz de confusión para emoción	76
87.	Almacenamiento del modelo entrenado	76
88.	Carga del modelo entrenado	76

N°	Descripción	Pág.
89.	Tokenización y codificación de textos de entrada para las predicciones	77
90.	Tokenización y codificación de textos de entrada para las predicciones	77
91.	Predicciones de sentimientos y emociones	77
92.	Almacenamiento de predicciones en archivo JSON	78
N°	Descripción	Pág.
93.	Documento almacenado en formato JSON	78

Índice de Anexos

N°	Descripción	Pág.
1.	Preguntas de la entrevista realizada a la periodista Paola Simbaña de Ecuador Chequea.	103

ANEXO XII.- RESUMEN DEL TRABAJO DE INTEGRACIÓN CURRICULAR (ESPAÑOL)

FACULTAD DE INGENIERIA INDUSTRIAL CARRERA SISTEMAS DE INFORMACIÓN MODALIDAD SEMESTRAL

ANÁLISIS DE SENTIMIENTOS DE LAS NOTICIAS COMPROBADAS EN TWITTER POR LAS VERIFICADORAS ACREDITADAS EN ECUADOR UTILIZANDO PROCESAMIENTO DE LENGUAJE NATURAL.

Autor: JIMÉNEZ OLIVO KIMBERLY ANTONELLA

Tutor: LSI. TOAPANTA BERNABÉ MARIUXI DEL CARMEN, MSIG.

Resumen

El presente proyecto tiene como propósito desarrollar un componente de análisis de sentimientos de las noticias en Twitter seleccionadas por las verificadoras de hechos en Ecuador, utilizando Procesamiento de Lenguaje Natural, que sirva de aporte a una investigación a nivel macro que busca ayudar al fact-checker a realizar el proceso de verificación de noticias. Para realizarlo, se emplearon técnicas de recolección de datos como la entrevista, técnicas de análisis de sentimientos y modelos de aprendizaje automático. En su implementación, se utilizó la base de datos en la nube MongoDB Atlas para obtener los datos requeridos, el lenguaje de programación Python utilizando el cuaderno colaborativo Google Colab y el modelo preentrenado de BERT en español como base para entrenar un modelo que pueda realizar predicciones de sentimientos y emociones. Además, se emplearon métricas para evaluar el rendimiento del modelo entrenado. Estas herramientas mencionadas contribuyeron para que se cumplan los objetivos propuestos.

Palabras Claves: Twitter , Modelo, Procesamiento, Análisis, Sentimientos.

ANEXO XII.- RESUMEN DEL TRABAJO DE INTEGRACIÓN CURRICULAR (INGLÉS)

FACULTAD DE INGENIERIA INDUSTRIAL CARRERA SISTEMAS DE INFORMACIÓN MODALIDAD SEMESTRAL

SENTIMENT ANALYSIS OF NEWS VERIFIED ON TWITTER BY ACCREDITED FACT CHECKERS IN ECUADOR USING NATURAL LANGUAGE PROCESSING.

Author: JIMÉNEZ OLIVO KIMBERLY ANTONELLA

Advisor: LSI. TOAPANTA BERNABÉ MARIUXI DEL CARMEN, MSIG.

Abstract

The purpose of this project is to develop a sentiment analysis component of Twitter news selected by fact-checkers in Ecuador, using Natural Language Processing, which serves as a contribution to a macro-level research that seeks to help the fact-checker to perform the news verification process. To do so, data collection techniques such as interviews, sentiment analysis techniques and machine learning models were used. In its implementation, the cloud database MongoDB Atlas was used to obtain the required data, the Python programming language using the Google Colab collaborative notebook, and the pre-trained BERT model in Spanish as the basis for training a model that can perform sentiment and emotion predictions. In addition, metrics were used to evaluate the performance of the trained model. These tools contributed to the achievement of the proposed objectives.

Keywords: Twitter, Model, Processing, Analysis, Sentiment.

Introducción

En la actualidad la mayor parte de la sociedad se aferra al consumo de Internet de tal manera que ha incrementado el uso de esta gran red. Kemp (2023) indica en su informe insignia Digital 2023 que “En los últimos meses ha existido un gran cambio en el número global de usuarios de Internet”. Adicionalmente, se menciona que “el número de usuarios de internet en el mundo alcanzó los 5.160 millones de personas, lo que representa al 64,4% de la población mundial (Galeano, 2023).

Es importante recalcar que los internautas tienden a pasar más tiempo en las redes sociales más que en cualquier otro sitio web (Naso et al., 2012). Por lo tanto, se puede indicar que las redes sociales son llamativas para los usuarios, por lo consecuente se conectan por bastante tiempo a estas aplicaciones, lo cual puede traer consigo ventajas y desventajas.

Para hacer hincapié en el rango de tiempo que pasan los usuarios de Internet en las redes sociales, dice:

De hecho, los usuarios de redes sociales dedican al día una media de dos horas y veinticinco minutos y se prevé que para 2024 el porcentaje de internautas a nivel mundial con perfil en redes sociales supere el 82 por cien. (Piñero, 2021, p.317)

Las redes sociales son medios que permiten la difusión masiva, ya que tienen un gran alcance e impacto en la sociedad moderna, adicionalmente, pueden ser utilizadas tanto por personas naturales como por empresas, por lo que permiten lograr una comunicación interactiva y dinámica (Hütt Herrera, 2012). Al ser medios de propagación masiva, el usuario tiene la libertad de compartir cualquier tipo de contenido, como consecuencia en muchas ocasiones se puede proliferar información errónea, desactualizada o que no es del agrado de los usuarios.

Según Gleick (2011) el aumento impresionante que han tenido los contenidos de información en Internet repercute en la posibilidad de entenderlos, contrastarlos y sobre todo de comprobar si es información verídica. De esta forma al navegar en estas plataformas de índole social, los usuarios encuentran distintos tipos de información, visualizando noticias engañosas o falsas conocidas también como fake news.

Resulta insostenible remontarnos a los orígenes de esta problemática, ya que las noticias falsas existen desde tiempos inmemorables; sin embargo, su nivel de impacto y trascendencia es en la actualidad desmedido por varios factores, uno de ellos es el alcance y uso que se le da a la tecnología y la

amplificación de los contenidos de las redes sociales. (Cusot & Palacios, 2019)

Por esta razón, se ha buscado garantizar que los textos informativos estén comprobados para que las personas que naveguen en Internet, especialmente en redes sociales no sean víctimas de fake news. Actualmente, a nivel mundial existen varias entidades que están autorizadas y certificadas por la International Fact-Checking Network (IFCN) para comprobar la veracidad de una noticia y Ecuador no se queda atrás, también cuenta con verificadores de hechos como Ecuador Chequea y Ecuador Verifica.

El presente trabajo de titulación es parte de una investigación a nivel macro que busca ayudar al fact-checker a realizar el proceso de verificación de noticias, aportando con el desarrollo de un componente que realice el análisis de sentimientos en las noticias de Twitter que son seleccionadas por las verificadoras acreditadas en Ecuador y los comentarios de los usuarios en estas noticias.

Se considera importante realizar este componente porque las noticias falsas son creadas para intervenir en las emociones que pueden tener los usuarios, generar miedo, incertidumbre, infamar y desequilibrar (Santamaria, 2017), esto es posible porque se hace uso de la emocionalidad como herramienta para conseguir interés por parte de los internautas en este tipo de contenido.

Frente a lo mencionado anteriormente, se hace énfasis en la importancia que tiene el desarrollo de un análisis de sentimientos en las noticias seleccionadas por las verificadoras de hechos en Ecuador y los comentarios realizados por los usuarios en estas.

Es importante mencionar que esta investigación se centra en realizar un análisis de sentimientos utilizando los datos de los tuits preprocesados y extraídos de la base de datos perteneciente al trabajo de titulación de la señorita Jenny Melissa Cercado Ruiz, “RECOPIACIÓN Y EXTRACCIÓN DE TWEETS REALIZADOS POR LAS VERIFICADORAS ACREDITADAS EN EL ECUADOR POR LA IFCN PARA LA GENERACIÓN DEL CONJUNTO DE DATOS USANDO TWEETPY.”

La presente investigación abarca tres secciones. En el capítulo I se plantea la problemática encontrada, se identifica el objeto de estudio, la delimitación del objeto de investigación, se justifica la investigación y se especifica el alcance, adicionalmente se determinan los objetivos que se pretenden cumplir a lo largo del proyecto, las teorías relacionadas a la investigación, el marco conceptual y el marco legal de la misma.

Por otra parte, el capítulo II comprende la metodología aplicada a la investigación presente, en la cual, con la ayuda de técnicas de investigación, metodologías de minería de

texto, entre otras herramientas permitieron lograr tomar la información de la data para realizar el análisis de sentimientos.

En cuanto al capítulo III, está compuesto por el desarrollo de la propuesta del componente análisis de sentimientos de noticias, el que contribuye con el proyecto macro de investigación “Sistema de ayuda para los verificadores de hechos empleando la metodología utilizada por Ecuador Chequea y Ecuador Verifica”. Adicional, el componente a desarrollar utiliza técnicas de limpieza y preprocesamiento de datos enfocadas en el análisis de sentimientos, traducción del texto al idioma inglés para su respectiva tokenización y herramientas para realizar el análisis de sentimientos a los textos de las noticias y comentarios.

Capítulo I

Marco Teórico

1.1. Planteamiento del Problema

Para determinar la problemática en la que se centra este trabajo de titulación es importante mencionar el problema a nivel general, que busca contrarrestar la propagación de noticias no verificadas conocidas como noticias falsas o fake-news, es importante aclarar que en la actualidad ya existen medios que se dedican a la verificación del discurso público y contenido fraudulento que circula en internet.

En base a la problemática expuesta que persiste en la actualidad respecto a las noticias engañosas, se ha buscado a nivel mundial contrarrestar esta difusión de información inverosímil por lo que se ha incrementado la cantidad de medios que se dedican a verificar cualquier tipo de noticia, los mismos que se localizan en mayor porcentaje en el continente europeo seguido del continente más grande y poblado de la tierra, Asia, como se puede visualizar en la siguiente imagen.

	África	América	Asia	Europa	Oceanía	Total
International Fact-checking Network (n=71)	2,82%	19,72%	29,58%	45,07%	2,82%	100%
Duke University Reporters' Lab (n= 187)	4,81%	42,25%	18,72%	32,09%	2,14%	100%
Datos obtenidos en agosto del 2019						

Figura 1. Localización de los medios de verificación. Información recuperada de (Rodríguez,2020)

Al enfatizar la proliferación de estos medios, dice:

Tanto los medios de comunicación como las iniciativas de periodistas independientes que se dedican a la verificación han proliferado en los cinco continentes, bien en países con sistemas democráticos estables o bien en aquellos que restringen derechos y libertades fundamentales tal y como se observa a partir de las bases de datos de medios que manejan las dos principales entidades: el Instituto Poynter a través del International Fact-Checking Network (IFCN) y el Duke Reporters' Lab. (Rodríguez, 2020, p.249)

En la actualidad, el Fact-Checking se ha llegado a considerar incluso como una oportunidad laboral para los profesionales del periodismo que buscan laborar con calidad. “Así lo revela la encuesta realizada a un total de 316 alumnos, de los que un 94,54% considera que es una nueva salida profesional y un 71,42% trabajaría exclusivamente verificando datos” (Ufarte Ruiz et. al, 2018).

En cuanto a nivel nacional, existen verificadores de hechos como Ecuador Chequea, el cual surgió para dar respuesta a una necesidad de incluir un proyecto especializado en la objetividad, en la campaña presidencial del 2016 se dedicó a comprobar la veracidad del discurso de los candidatos, en la actualidad también se dedica a corroborar la veracidad de las noticias (Dols Hernández, 2018).

Comúnmente para comprobar o desmentir algún contenido estos verificadores tienen la siguiente metodología:

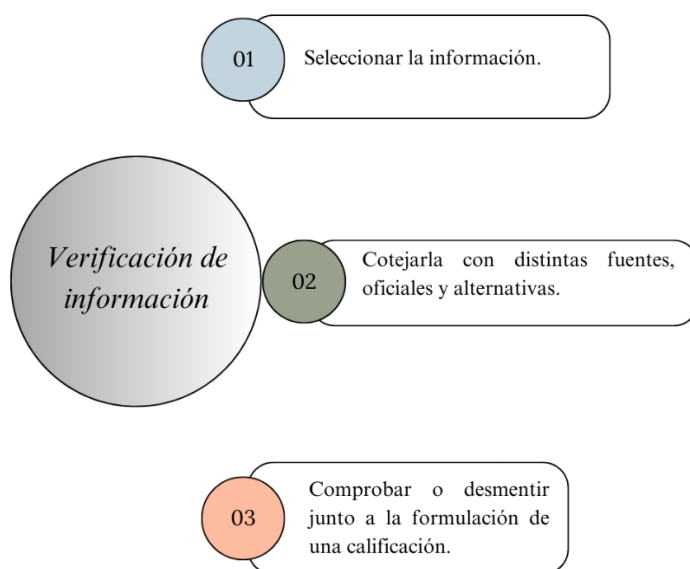


Figura 2. Proceso de verificación de información. Información adaptada de (Hernández, 2017). Elaborado por Jiménez Kimberly.

El proceso de verificación de una noticia falsa conlleva una inversión de recursos, la metodología utilizada conlleva varios pasos y su respectivo tiempo para verificar dicha información, por lo que el trabajo de investigación a nivel macro al cual pertenece el presente trabajo de titulación busca desarrollar un sistema de ayuda para los verificadores de hechos en Ecuador que contribuya al proceso de comprobación de veracidad de las noticias, brindando un soporte para poder optimizar este proceso.

Con relación a este problema general, el presente trabajo de titulación está enfocado en el desconocimiento de las emociones existentes en los comentarios de los usuarios en las noticias publicadas en Twitter y que son seleccionadas para su comprobación por los verificadores acreditados en el Ecuador por la IFCN, siendo este un problema que es necesario abordar.

Santamaría (2017), afirma que las noticias falsas son “Generadas para crear incertidumbre, miedo, desestabilización para apoyar o desacreditar. Y así obtener un capital de seguidores, perfiles y usuarios de redes sociales para avivar un movimiento, un interés económico o de marca o una persona”. En este sentido, se puede determinar que las fake-news intervienen en las emociones de los usuarios, y en algunas ocasiones provocan sentimientos negativos, lo que puede desencadenar una desestabilización en las personas.

El problema de las noticias falsas consiste en el aprovechamiento de las emociones para conseguir atención y tiempo de visualización, lo que posteriormente se convierte en ingresos de publicidad, a esto se le llama economía de la emoción (Bakir & McStay, 2018).

Para afirmar la teoría de la economía de las emociones, dice:

La teoría de la economía de la emoción propone que los titulares de noticias falsas se crean para evocar respuestas emocionales en los lectores que harán que interactúen con el artículo de una manera que permita al creador obtener una ganancia. (Horner et. al, 2021)

En resumen, las fake-news son creadas también con el objetivo de generar ingresos económicos, existen personas malintencionadas que buscan beneficiarse a nivel económico de las emociones que pueda sentir algún usuario.

1.2. Formulación del Problema

¿De qué manera, la implementación de un análisis de sentimientos podría contribuir con la verificación de la veracidad de una noticia en el software “Sistema de ayuda para los verificadores de hechos empleando la metodología utilizada por Ecuador Chequea y Ecuador Verifica”?

1.3. Objeto de Estudio

En el presente trabajo de titulación de tercer nivel se ha determinado como objeto de estudio las noticias seleccionadas para su comprobación por las verificadoras acreditadas en Ecuador y los comentarios realizados por los usuarios en Twitter.

1.4. Delimitación del objeto de investigación

1.4.1. Delimitación Geográfica.

La data de Twitter con la que se trabajará el análisis de sentimientos corresponde al verificador de hechos acreditado por la IFCN, Ecuador Chequea, el cual está ubicado al norte

de la capital del país, Quito, exactamente en José Padilla N330 e Iñaquito, Edificio Platinum, Piso 10, Oficina 1002.

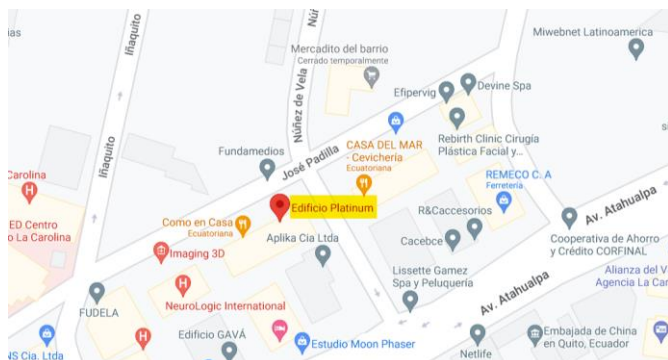


Figura 3. Ubicación del verificador de hechos Ecuador Chequea. Información recuperada de (Google maps, 2023).

Adicionalmente, en el presente trabajo investigativo también se analizarán las noticias y comentarios del medio que acoge herramientas de verificación de datos y cuenta con aval de la International Fact Checking Network (IFCN), Ecuador Verifica, el cual es una coalición que articula a medios de comunicación, organizaciones de la sociedad civil (OSC) y universidades, con el objetivo de verificar el discurso político y promover la transparencia en las instituciones públicas, que se encuentra localizado al norte de la ciudad de Quito, también conocida como Luz de América, precisamente en José Padilla, Quito 170135.



Figura 4. Ubicación del verificador de hechos Ecuador Verifica. Información recuperada de (Google maps, 2023).

1.4.2. Delimitación en Tiempo – Espacio.

El presente trabajo de titulación se ha llevado a cabo en el primer periodo de titulación del año 2023, desde mayo hasta agosto del año anteriormente mencionado. En este rango de tiempo se han realizado las tareas necesarias para alcanzar los objetivos propuestos.

1.4.3. Delimitación Semántica.

En el presente trabajo de investigación se ha establecido un marco bibliográfico comprendido en definiciones y terminologías que se detallan a continuación:

IFCN: Es un foro para verificadores a nivel mundial organizado por el Poynter Institute for Media Studies, su función es inspeccionar las tendencias y formatos en la verificación de datos.

Duke Reporters' Lab: Es un centro de investigación periodística en la Escuela de Políticas Públicas de Sanford en la Universidad de Duke, uno de sus proyectos principales es la verificación de hechos.

Fact-Checking: Conocida también como la verificación de hechos, es un producto de la era digital que consiste en validar y desmentir mitos, rumores y noticias que circulan en internet.

Fact-Checker: Es un verificador que mantiene una postura neutral respecto a la política, tiene como fin desmentir las declaraciones realizadas en público ya sea por partidos políticos o personas que no se ajusta a la realidad objetiva, de tal manera ayuda a corregir apreciaciones incorrectas de la realidad.

Inteligencia Artificial: Es una combinación de algoritmos que se realizan con el objetivo de imitar la función cognitiva humana por medio de máquinas.

Procesamiento de Lenguaje Natural (PLN): Es el campo de conocimiento de la Inteligencia Artificial que se ocupa de investigar la manera de comunicar las máquinas con las personas mediante el uso de lenguas naturales.

Preprocesamiento: Es una etapa que se realiza antes de algún proceso de minería de datos, por lo general en esta fase se realizan actividades que contienen el tratamiento de cabeceras y filas, lo que incluye eliminación de columnas sobrantes, datos duplicados, inconsistentes o inadecuados.

Identificación de tendencias: Es un procedimiento que se ocupa de comprender cuales son las señales que crean un patrón que pueda indicar un cambio colaborando con las necesidades, deseos o expectativas de las personas.

Análisis de sentimientos: Es un elemento del campo de Procesamiento del Lenguaje Natural que se dedica específicamente a entender opiniones subjetivas o sentimientos inmersos sobre algún tema.

Polaridad: En la ciencia de datos, la polaridad es una medida que indica la dirección y el grado del sentimiento, opinión o actitud inmersa en un texto, por lo general se utiliza en el análisis de sentimientos y en la minería de opinión.

Sátira: Es un género literario que se identifica por ridiculizar un personaje y sus actuaciones con el fin de evidenciar su fragilidad y expresar la indignación y la crítica.

Ironía: Manera de dar a entender algo diciendo lo contrario de lo que se piensa o se quiere decir.

Tristeza: Es una de las emociones fundamentales que tiene el ser humano y se relaciona con un bajo estado de ánimo y frustración.

Felicidad: Se interpreta como un estado emocional de ánimo positivo, alegría y plenitud que se manifiesta de maneras diferentes, dependiendo a la personalidad y el carácter.

Python: Es un lenguaje de programación que se utiliza en aplicaciones web, desarrollo de software, ciencia de datos y machine learning.

Google Colaboratory: También conocido como Google Colab, es una herramienta con arquitectura en la nube, perteneciente a Google Research que permite escribir códigos Python.

1.5. Justificación

El principal motivo que ha llevado a desarrollar el presente trabajo es la problemática que existe referente a las fake-news y el proceso que se realiza para corroborar la veracidad de una noticia. Se considera fundamental desarrollar un análisis de sentimientos de las noticias seleccionadas para la comprobación por los verificadores de hechos nacionales, Ecuador Chequea y Ecuador Verifica y lo que comentan los usuarios de Twitter en estas, dado que las noticias falsas interfieren en los sentimientos de los usuarios.

Ecuador Chequea (2018) indica que las noticias falsas incluso pueden originar conflictos, dividir a la sociedad y manipular a los ciudadanos. Por estas razones, cuando una noticia es publicada en internet se debe ratificar su veracidad en el menor tiempo posible, por consiguiente, evitar la confrontación entre los ciudadanos. Es importante recalcar que las fake-news por lo general provocan emociones negativas en los usuarios. Chavero & Intriago (2021) argumentan que las noticias falsas suelen ser de tono negativista.

Actualmente, el proceso de verificación de noticias seleccionadas por el verificador de contenido Ecuador Chequea es aproximadamente de cinco minutos y con los experimentos realizados se disminuye notablemente el tiempo de verificación (Toapanta, 2023).

El propósito de realizar un análisis de sentimientos en las noticias seleccionadas por los verificadores de hechos nacionales es conocer lo que el usuario, en la mayoría de las ocasiones ciudadanos ecuatorianos pueden sentir al visualizar dichas noticias, dado que, si la noticia genera un sentimiento negativo en el usuario, probablemente esta sea falsa, lo que

serviría de aporte al fact-checker para optimizar el tiempo en la verificación de una noticia que circule en redes sociales.

En cuanto a la información que se utilizará para realizar el análisis de sentimientos, es importante mencionar que se propone aplicarlo a una data de noticias en Twitter porque en la misma se difunde un gran porcentaje de fake-news. Chavero & Intriago (2021) menciona que “Twitter se presenta como la red en la que más se distribuyen los distintos tipos de información falsa: el gobierno ecuatoriano considera que el 43,8% de las noticias que circulan por Twitter tiene contenido engañoso al observar contenidos que pretenden causar daño” (p.28).

Adicionalmente, se dice que “El surgimiento de las redes sociales como Twitter ha sido aprovechada por la inteligencia artificial y el Big Data para el tratamiento de información con respecto a los tuis”. (Alva Segura, 2020, p.1)

Respecto a la metodología que aplican los verificadores de hechos para llevar a cabo el proceso de verificación de las fake-news, por lo general no varía totalmente y se concentra en los siguientes cinco pasos:

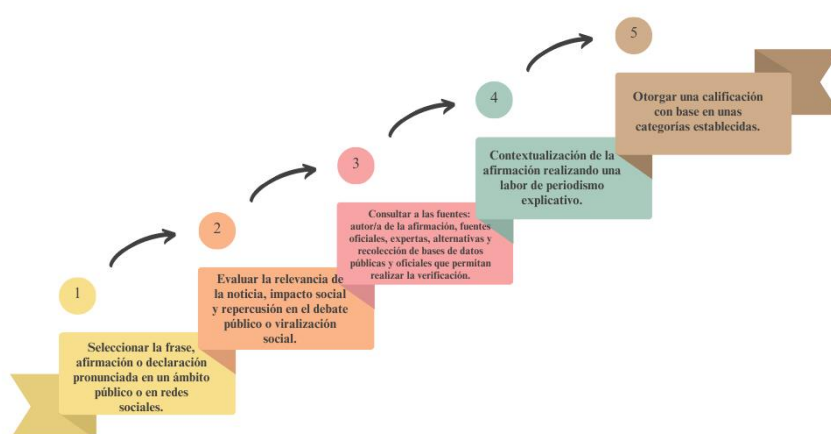


Figura 5. Pasos para verificar la veracidad de una noticia. Información adaptada de (Vélez-Bermello, 2020). Elaborado por Jiménez Kimberly.

El portal Ecuador Chequea utiliza una metodología algo sencilla para corroborar las noticias falsas, empleada en una página interna que cuenta con disponibilidad para los usuarios, adicionalmente, pretende disponer de un proceso objetivo, transparente y replicable en todo el procedimiento, el cual se inicia con una búsqueda minuciosa de la información que comprende la noticia y se procede a buscar el contenido original, en cuanto a imágenes, utilizan herramientas tecnológicas como Google Images para realizar las búsquedas insertando la imagen correspondiente para que el buscador encuentre otras

similares con sus respectivas fuentes. En caso de tratarse de un actor político local, este verificador de noticias compara información con la persona a la que haga referencia y para finalizar, ya contextualizado el contenido lo publica con la etiqueta #falseta. (Vélez-Bermello, 2020) y (Dols Hernández, 2018).

Además de eso, Ecuador Chequea cuenta con un sistema de calificación que se compone de las categorías que se detallan a continuación:

- **CIERTO:** El contenido que muestra información precisa y congruente con los datos fuentes. Las declaraciones son exactas y no omiten datos ni limitan la contextualización.
- **FALSO:** El contenido no tiene un fundamento real. Una declaración incongruente con los datos y fuentes.
- **ALTERADO:** Contenido de imagen, audio o video editado o modificado que podría engañar a las personas. Esta categoría incluye imágenes alteradas, audios con mensajes falsos, videos sacados de contexto, cadenas de WhatsApp.
- **ENGAÑOSO:** El contenido no es completamente falso, Contiene mayores elementos de falsedad que de certeza. Es engañoso cuando la afirmación es congruente con la información o coincide parcialmente con ciertos datos, pero demuestra que pudo haber sido manipulada a fin de engañar.
- **IMPRECISO:** Tiene algunas imprecisiones fácticas. Es impreciso cuando la afirmación es consistente con los datos disponibles, pero se omite u oculta información.
- **INVERIFICABLE:** Una afirmación donde no se puede identificar las fuentes o argumentos de las cuales se haya desprendido dicho enunciado.
- **SÁTIRA:** Contenido que muestra exageración, ironía, ridiculización. Usado particularmente en contexto de temas políticos, religiosos o sociales. (Ecuador Chequea, 2023)

En el caso de Ecuador Verifica, de la misma manera que Ecuador Chequea, acoge las herramientas de Fact-Checking para el desarrollo de las actividades de verificación del

discurso público o los contenidos que transitan en Internet, el proceso se realiza en las siguientes etapas:

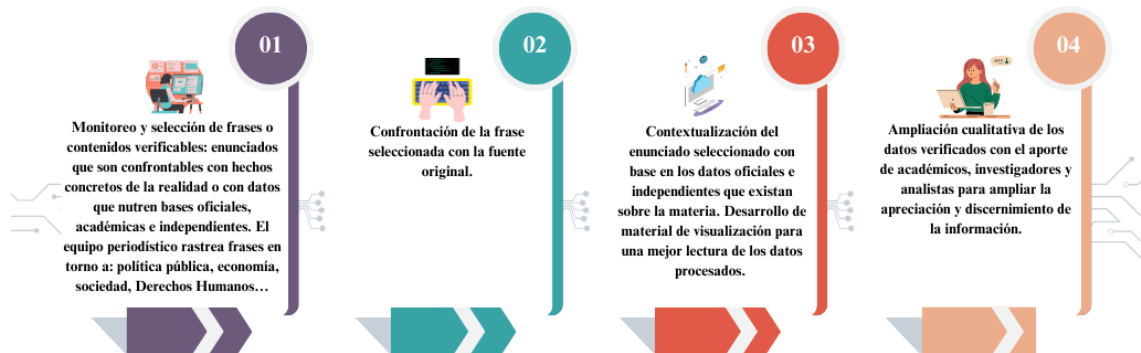


Figura 6. Etapas del proceso de verificación de Ecuador Verifica. Información adaptada de (Ecuador Verifica, 2023). Elaborado por Jiménez Kimberly.

Es importante mencionar, que Ecuador Verifica utiliza las mismas categorías que Ecuador Chequea.

Al visualizar la metodología de comprobación de veracidad de contenido que tienen estos verificadores, los pasos que se necesitan realizar, el tiempo aproximado que conlleva corroborar la veracidad de una noticia, la calificación “sátira”, la asociación que tienen los sentimientos con las fake-news y lo que puede llegar a incitar en las personas, se ha considerado de gran relevancia realizar el presente trabajo de titulación que consiste en desarrollar un componente de análisis de sentimientos utilizando herramientas de Procesamiento de Lenguaje Natural (PLN) que aporte al sistema de ayuda para los verificadores de hechos del Ecuador, buscando optimizar el tiempo al fact-checker en el proceso de verificación.

1.6. Alcance

La presente propuesta forma parte de una investigación a nivel macro que consiste en el desarrollo de un sistema de ayuda para los verificadores de hechos en Ecuador. Mediante el desarrollo del componente análisis de sentimientos se identificarán las emociones inmersas en las noticias seleccionadas y los comentarios de los usuarios en éstas. La presente propuesta tiene como alcance los siguientes puntos:

- Obtener las noticias de la base de datos.

Extraer las noticias seleccionadas y los comentarios de la base de datos.

- Preprocesar las noticias.

Aplicar el preprocesamiento de datos con énfasis en el análisis de sentimientos.

- Seleccionar algoritmos, parámetros e hiperparámetros.

Identificar los algoritmos de clasificación a utilizar, tales como: Regresión Lineal, Máquinas de Vectores de Soporte SVM, Naive Bayes, redes neuronales, BERT, Bag of words BOW, Tf-Idf, entre otros.

- Categorizar los sentimientos.

Los sentimientos que se utilizarán son: negativo y positivo. esperanza, alegría, tristeza, enfado y miedo.

- Identificar las emociones.

Se aplicará un análisis de sentimientos a las noticias preprocesadas utilizando los algoritmos parametrizados en base a la categoría de sentimiento, identificando emociones tales como: esperanza, alegría, tristeza, enfado y miedo.

- Presentar los resultados del análisis de sentimientos.

Se presentará un listado de sentimientos.

- Almacenar los resultados del análisis de sentimientos.

Los resultados del análisis de sentimientos se almacenarán en una fuente de información no estructurada, con estos, se pretende demostrar que cuando una noticia o texto de un tweet contiene un sentimiento negativo, tiene mayor probabilidad de ser falsa.

1.7. Objetivos

1.7.1. Objetivo General.

Desarrollar un componente de software para el análisis de sentimientos de las noticias en Twitter seleccionadas por las verificadoras de hechos en Ecuador, utilizando Procesamiento de Lenguaje Natural.

1.7.2. Objetivos Específicos.

1. Obtener las noticias que se encuentran almacenadas en la base de datos no estructurada para el análisis de sentimientos.
2. Implementar técnicas y herramientas de análisis de sentimientos para identificar las emociones en las noticias y en los comentarios realizados por los usuarios en la red social Twitter.
3. Presentar en un cuaderno de Google Colaboratory el resultado del análisis de los sentimientos de las noticias y comentarios de los usuarios.

4. Almacenar la información de los resultados del análisis de sentimientos en formato JSON para que sea utilizado posteriormente en el software “Sistema de ayuda para los verificadores de hechos empleando la metodología utilizada por Ecuador Chequea y Ecuador Verifica”.

1.8. Marco Teórico

1.8.1. Desinformación en las redes sociales.

Según Celaya (2008) las redes sociales son lugares donde comúnmente las personas publican y comparten cualquier información, personal y profesional indistintamente si las personas que visualizan este contenido compartido son conocidas o desconocidas. Valverde-Berrocoso et. al (2022) afirman que la libertad de expresión es un valor sustancial de las sociedades democráticas, no obstante, si se lo realiza de manera incorrecta, se convierte en desinformación, por lo tanto, pone en peligro la democracia.

Las redes sociales han tenido un gran auge en la vida de las personas, una de las razones por las que se puede dar esta gran acogida es por la facilidad que tienen los usuarios para interactuar con otras personas. Por lo tanto, al tener disponibilidad de difundir cualquier tipo de información en estas redes virtuales, los usuarios pueden ser víctimas de la desinformación.

Wardle & Derakhshan (2017) definen el término «desinformación» como una información falsa, que se origina y difunde intencionalmente para provocar algún tipo de daño, confundir y falsear.

Así mismo, Galdón (2001) menciona que la desinformación es la ausencia de la verdad o puede implicar un tema que genera confusión e impide a la persona llevar a cabo acciones con libertad. Adicionalmente afirma que existen dos tipos: aquella que es sin intención alguna y está causada por la falta de calidad y criterio de quien la construye y consume; y, por otro lado, la desinformación creada de forma deliberada y consumida sin conocimiento, a la que llama manipulación.

1.8.1.1. Noticias falsas como parte de la desinformación en la red social Twitter.

El fenómeno de las noticias falsas conocidas también como fake-news en inglés, y la desinformación han existido desde hace mucho tiempo y se difunden de manera acelerada. Según Marín (2019) las noticias falsas se propagan rápidamente mediante las redes sociales, incluso pueden llegar a medios de comunicación oficiales, por lo tanto, al público le cuesta distinguir los hechos de la ficción.

Hoy en día, la desinformación ocasionada por la propagación de noticias falsas es común, a menudo las personas transmiten o comparten información sin verificar previamente las fuentes de las cuales ha obtenido la noticia.

Una de las redes sociales con mayor influencia de noticias falsas, es Twitter. Vosoughi et al. (2018) llevaron a cabo un estudio en el periodo comprendido entre 2006-2017 sobre la propagación de contenidos verdaderos y falsos en Twitter, el cual tiene por conclusión que las noticias falsas se propagan a mayor velocidad y tienen más alcance que las noticias verdaderas y sus efectos son más pronunciados si versan en especial sobre política.

Las fake-news provocan efectos negativos en las personas, ya que al visualizar un contenido que no es real en las redes sociales, pueden emitir comentarios en base al contenido falso.

1.8.1.2. Factores asociados a las noticias falsas.

Allcott & Gentzkow (2017) describen cuatro factores que usualmente se asocian al término fake-news: errores involuntarios, rumores, sátira y declaraciones falsas.

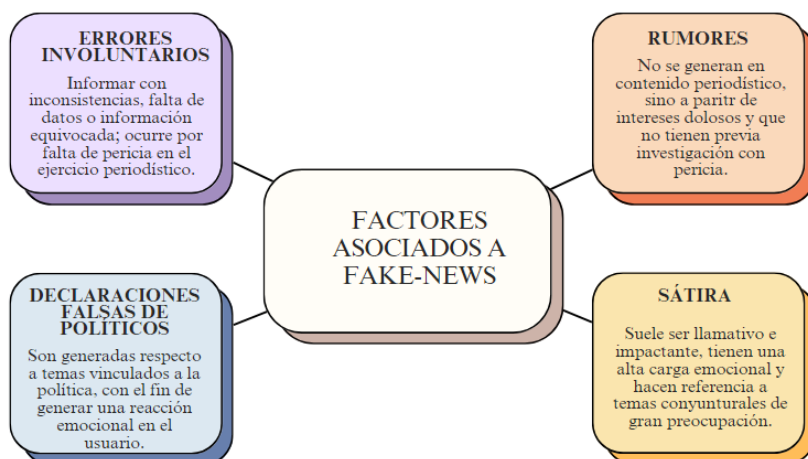


Figura 7. Factores asociados a fake news. Información adaptada de (Allcott y Gentzkow, 2017). Elaborado por Jiménez Kimberly.

Como se puede visualizar en la figura 7, son diversos los factores que se asocian a las noticias falsas, en muchas ocasiones por equivocación se difunden contenidos erróneos, por lo tanto, esto provoca desinformación. Los usuarios tienden a creer en rumores que visualizan en fuentes no oficiales o en noticias que tienen titulares impactantes o raros.

1.8.2. Fact-Checking para la corroboración de contenido.

El periodismo de verificación de hechos, también conocido en inglés como fact-checking, surgió en el 2003 en Estados Unidos con la creación del portal FactCheck.org. (Zommer, 2015).

Hace veinte años que se tomó en cuenta una necesidad de contrarrestar la desinformación, como consecuencia de la sobreinformación y la velocidad con la que se propaga información a nivel de internet en los últimos años ha provocado una proliferación de medios que se dedican a corroborar la veracidad de información.

Para enfatizar la proliferación del Fact-checking. (Zommer, 2014), dice:

A nivel mundial, se generaron múltiples plataformas de chequeo de datos (fact-checking) que han sido reconocidas por su valor periodístico. Entre ellas se destacan las experiencias en los Estados Unidos de FactCheck.org (pionero de la actividad, creado en 2003 al interior de la Universidad de Pensilvania) y Politifact.com (ganadora de un premio Pulitzer en 2009, dos años después de su surgimiento) y en Europa las de Les Décodeurs en Francia, Channel 4 en el Reino Unido y, bastante más tarde, Pagella Política en Italia. (p.6)

Ufarte-Ruiz et al. (2018), estiman que las funciones de fact-checking son necesarias para el periodismo actual que se ha centrado en un trabajo más de escritorio. Además, confirman que con el surgimiento del fact-checking se encuentran disponibles más ofertas de trabajo y, por ende, la adaptación de nuevas habilidades y competencias para poder ejercerlo.

El mundo en el que vivimos en el día de hoy se ve en la necesidad de disponer de verificadores de hechos, es indispensable que el periodismo actual se capacite y disponga de herramientas que le permitan corroborar información antes de ser compartida.

En la actualidad, es necesaria la existencia de verificadores de hechos debido a la propagación de noticias falsas que se cargan a Internet diariamente, según (Zommer, 2014) la verificación de hechos pretende brindar a la ciudadanía más elementos para que puedan analizar la información a la que tienen acceso y tomen decisiones más conscientes y sea menos manipulable.

Vélez-Bermello (2020) señala que la labor de verificar y hacer frente a la información diferencia el periodismo convencional del que se realiza en el fact-checking. En sentido de que el periodismo tradicional, se procede a evaluar los contenidos antes de ser publicados; en cambio, el fact-checking se aplica después de que se haya propagado una información.

En ese sentido, es necesaria la existencia de los verificadores de hechos debido a la información falsa en la red, de tal manera que esta información pueda ser corroborada, adicionalmente, el fact-checking se ha convertido en una oportunidad de trabajo en la actualidad.

1.8.2.1. Métodos para la verificación de hechos.

Buttry (2016) afirma que el método de comprobación varía con cada hecho e ilustra algunas pautas importantes, principalmente se debe realizar la pregunta ¿cómo lo sabes?, esto hace referencia a las fuentes de donde proviene la información, la fuente es la base del proceso de comprobación para un verificador de hechos.

Es primordial realizar esta interrogante, porque se debe tener el conocimiento de origen de la información que se está receptando, con mayor importancia aún si es una información que se pretende difundir. Las plataformas verificadoras de hechos disponen de un proceso de verificación de noticias, que comúnmente contiene las siguientes tácticas:

Tabla 1. *Tácticas del proceso de verificación de noticias.*

Táctica	Descripción
Examinar la acción	Investigar el sitio web, objetivo e información de contacto.
Profundizar la lectura	Leer la noticia completa, no dejarse llevar por un titular llamativo.
Reconocer al autor	Asegurarse de la existencia del autor, realizando una búsqueda rápida del mismo.
Fuentes adicionales	Realizar una búsqueda adicional para comprobar si hay datos que respalden la
Verificar la fecha	Comprobar que la noticia se encuentre vigente, que sea actual.
Distinguir si es una broma	Si es una noticia muy rara se procede a investigar al autor y el sitio web.
Considerar su sesgo	Es importante considerar las creencias del usuario porque podrían alterar su opinión.
Consultar con un experto	Consultar un sitio web de verificación.

Información adaptada de (Cusot & Palacios, 2019), Elaborado por Jiménez Kimberly.

La metodología de trabajo utilizada por las organizaciones que se dedican a corroborar la veracidad de una información se lleva a cabo en los siguientes pasos:

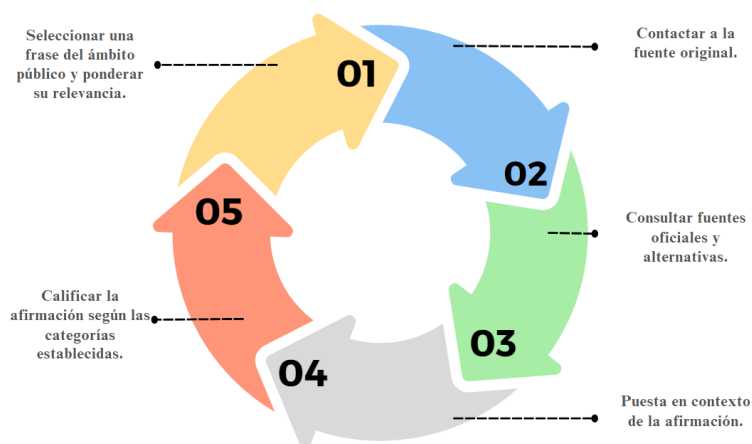


Figura 8. Metodología de trabajo general en Fact-Checking. Información adaptada de (Moreno-Gil, Ramon, & Rodríguez-Martínez, 2021). Elaborado por Jiménez Kimberly.

El Instituto Poynter mantiene el Código de Principios Deontológicos, el cual se encuentra formado por compromisos que cuidan la transparencia e imparcialidad del método de Fact-checking, los mismos que se presentan a continuación:

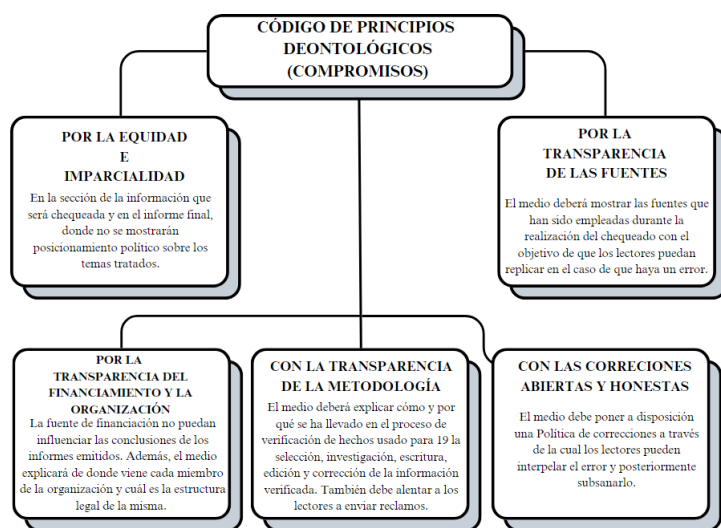


Figura 9. Códigos de Principios Deontológicos. Información adaptada de (Dols Hernández, 2018). Elaborado por Jiménez Kimberly.

Para realizar la verificación de la veracidad de una información es necesario seguir ciertas recomendaciones y estas dependen del formato en el que se encuentre y el contenido.

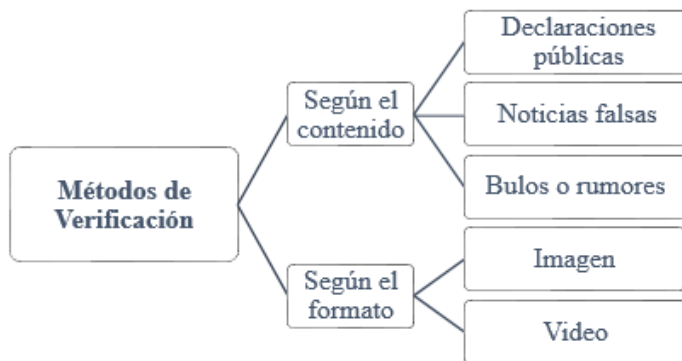


Figura 10. *Métodos de verificación. Información adaptada de (Dols Hernández, 2018). Elaborado por Jiménez Kimberly.*

Dols Hernández (2018) indica que “El método del fact-checking se basa en tres bloques generales de contenido: declaraciones públicas, noticias falsas y bulos o rumores.”

1.8.2.2. Plataformas de fact-checking.

Las organizaciones de fact-checking son conceptualizadas como intervenciones que aparecen “cuando se percibe una amenaza” (Amazeen, 2020, p. 98).

De acuerdo con Quintana Pujalte y Pannunzio (2021) existen distintas entidades que se enfrentan a los bulos informativos o noticias falsas que se difunden en las redes sociales, a continuación, se mencionan algunas.

- La Unión Europea dispone de una acción de vigilancia para luchar en contra de las noticias falsas que se propaguen dentro de su territorio.
- La ONU cuenta con una iniciativa llamada Verified que busca la colaboración de los ciudadanos con la denuncia de información falsa.
- Adicionalmente, el Instituto Poynter ha creado la Red Internacional de Verificación de Datos (IFCN por sus siglas en inglés). Hoy en día, existen setenta y cuatro organizaciones que forman parte de esta red, pertenecientes a setenta países y realizan verificaciones en cuarenta idiomas. Tiene como funcionalidad corroborar las declaraciones de figuras públicas, instituciones importantes y otras afirmaciones de interés para la sociedad de amplia circulación (Quintana Pujalte y Pannunzio, 2021).

Actualmente, según el censo elaborado por el Reporters’ Lab de la Sanford School of Public Policy de la Duke University, existen 290 plataformas de fact-checking activas en 83 países y existen dos modelos de organizaciones que se dedican a verificar hechos: aquellas que son impulsadas por medios de comunicación, conocidas también como newsroom

model en inglés y aquellas que han surgido de manera independiente, NGO model, en el idioma inglés, aunque los modelos sean distintos, la International Fact-Checking Network (IFCN) dispone de un código de principios en donde se establece que todas las entidades dedicadas al fact-checking deben caracterizarse por su transparencia e imparcialidad (Stencel y Luther, 2020; Graves y Cherubini, 2016).

Es importante mencionar que en Latinoamérica no es muy conocido el concepto de Fact-checking, sin embargo, a nivel de continente se tiene uno de los pioneros en el tema como Chequeado. Indica Zommer (2014) que Chequeado “empezó a modelarse a fines de 2009 cuando sólo existían en el mundo unas pocas organizaciones dedicadas a contrastar discurso público con hechos y datos, y que pasó a estar disponible en línea en 2010”. (p.11)

1.8.3. Verificadores de hechos acreditados en Ecuador.

A nivel nacional existen algunos medios que se dedican al fact-checking debido a la propagación de contenido falso en Internet, entre ellos se encuentra Ecuador Chequea como el verificador pionero a nivel nacional y la coalición Ecuador Verifica.

1.8.3.1. Verificador pionero a nivel nacional – Ecuador Chequea.

Ecuador Chequea nace en octubre del 2016. Actualmente, además de trabajar en el chequeo de declaraciones públicas, se dedica a comprobar las noticias falsas (Dols Hernández, 2018).

Ecuador Chequea es el primer medio ecuatoriano que se dedica por completo a la verificación del discurso público y contenidos engañosos que circulan en Internet. Es importante mencionar que este verificador desde enero de 2019 es parte de la International Fact Checking Network (IFCN). Además de esto, las fuentes que utiliza este verificador son citadas e incluyen enlaces a documentos, segmentos con códigos abiertos y a los datos originales para que el lector pueda ingresar de manera directa (Ecuador Chequea, 2023).

1.8.3.2. Verificador de hechos Ecuador Verifica.

Vélez-Bermello (2020) indica que hasta septiembre del 2020, Ecuador Chequea era el único verificador de información a nivel nacional, sin embargo en la actualidad también existe el portal denominado Ecuador Verifica, que se creó entre septiembre de 2020 y abril de 2021, es importante mencionar que Ecuador Verifica es coordinado por Ecuador Chequea y surgió con el fin de luchar contra la desinformación en tiempo de elecciones presidenciales en el 2021 y está conformado por medios de comunicación, organizaciones de la sociedad civil (OSC) y universidades del país, con el objetivo de verificar el discurso político y promover la transparencia en las instituciones públicas.

1.8.4. Metodologías para la comprobación de veracidad de noticias.

Los verificadores nacionales Ecuador Chequea y Ecuador Verifica se enfocan en la misma metodología, la cual consiste en desarrollar el proceso de verificación en cuatro etapas.

1.8.4.1. Metodología de Ecuador Chequea para la comprobación de veracidad de las noticias.

El verificador pionero a nivel nacional menciona en su página oficial que dispone de un proceso de verificación que se desarrolla en cuatro etapas:

- 1. Monitoreo y selección de frases o contenidos verificables: enunciados que son confrontables con hechos concretos de la realidad o con datos que nutren bases oficiales, académicas e independientes. El equipo periodístico rastrea frases en torno a: política pública, economía, sociedad, Derechos Humanos...**
- 2. Confrontación de la frase seleccionada con la fuente original.**
- 3. Contextualización del enunciado seleccionado con base en los datos oficiales e independientes que existan sobre la materia. Desarrollo de material de visualización para una mejor lectura de los datos procesados.**
- 4. Ampliación cualitativa de los datos verificados con el aporte de académicos, investigadores y analistas para ampliar la apreciación y discernimiento de la información.** (Ecuador Chequea, 2023)

Este verificador acreditado por la IFCN, una vez realizado el proceso de verificación, clasifica los contenidos en las siguientes categorías:



Figura 11. Calificaciones o categorías de Ecuador Chequea. Información adaptada de (Ecuador Chequea, 2023). Elaborado por Jiménez Kimberly.

1.8.4.2. Metodología de Ecuador Verifica para la detección de noticias falsas.

El portal de verificación de información Ecuador verifica, emplea las herramientas de fact checking, para confrontar la información y determinar si lo que pronuncian los candidatos es cierto, falso o merece contexto, adicionalmente verifican si los contenidos que se encuentran en redes sociales y plataformas de mensajería instantánea respecto al contexto electoral son ciertas. Cabe recalcar que para el proceso de verificación del discurso público y los contenidos que circulan en Internet, se utiliza la metodología de Ecuador Chequea (Ecuador Verifica, 2023)

Al utilizar la misma metodología que Ecuador Chequea, el procedimiento de verificación incluye siete categorías que se detallan a continuación.



Figura 12. Calificaciones o categorías de Ecuador Verifica. Información adaptada de (Ecuador Verifica, 2023). Elaborado por Jiménez Kimberly.

Ecuador Verifica utiliza las siguientes etapas para llevar a cabo el fact checking:

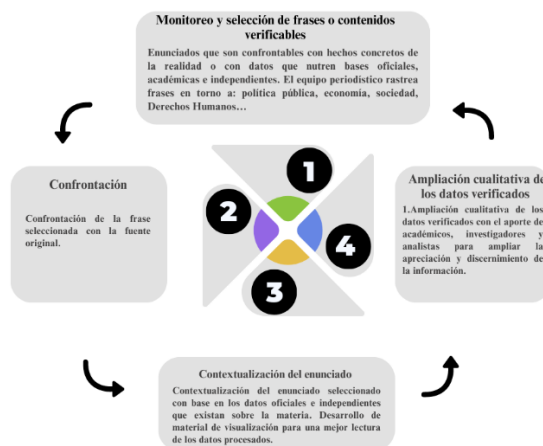


Figura 13. Etapas del proceso de verificación de Ecuador Verifica. Información adaptada de (Ecuador Verifica, 2023). Elaborado por Jiménez Kimberly.

1.8.5. PLN y aplicación de su campo de desarrollo para la detección de sentimientos y emociones.

Según Agarwal & Mital (2016) el componente análisis de sentimientos es el estudio que se ocupa de analizar la opinión y el sentimiento que tienen las personas hacia los servicios, productos y entidades que se encuentren en un texto.

Este campo de desarrollo denominado análisis de sentimientos, el cual consiste en detectar los sentimientos y emociones que puede tener una persona. Además, el análisis de sentimientos es emocionante, debido a que en la actualidad existe gran cantidad de

situaciones reales en las que se puede implementar este campo. Este campo de desarrollo se ha puesto en ejecución en proyectos y trabajos de investigación relacionados al tratamiento de información en la red social Twitter.

En el trabajo de fin de grado “Análisis de Sentimientos con Twitter: Turismo y Política Electoral” realizado por Sánchez del Hoyo (2019) en la ciudad de Sevilla, se realizó un análisis de sentimientos enfocado a las elecciones de España del mes de abril de 2019, en donde los datos se obtuvieron de Twitter, el código se realizó en lenguaje de programación R.

Por otro lado, Alarcón (2022) en su trabajo de Grado aplica un análisis de sentimientos en la red social Twitter utilizando procesamiento de lenguaje natural, realizando una construcción y transformación del conjunto de datos, utilizando algoritmos tales como, Linear SVC, XGBoost y Random Forest.

En la misma línea de investigación, en su Trabajo Fin de Máster, Alva Segura (2020) desarrolló un sistema en análisis de sentimiento político referente a las elecciones congresales del 2020 en Perú utilizando datos extraídos de la red social Twitter, enfocándose en la metodología de desarrollo de proyectos de ciencia de datos CRISP-DM y el lenguaje de programación Python, realizando el preprocesamiento de los datos, en cuanto a herramientas, empleó Textblob con el apoyo del método de Naive Bayes (NB), las máquinas Vectores de Soporte (SVM), Random Forest y las redes neuronales convolucionales (RCN).

1.9. Marco Conceptual

1.9.1. La Inteligencia Artificial.

Rouhiainen (2018) afirma que la inteligencia artificial, también conocida por sus siglas IA, es la capacidad que tienen las máquinas para utilizar algoritmos, aprender de los datos y ejercer lo aprendido en la toma de decisiones de la misma manera que lo haría un ser humano. Es decir que las máquinas utilizan algoritmos de aprendizaje para poder analizar una situación determinada y tomar decisiones en base a los resultados obtenidos en el análisis de datos.

Es importante mencionar que en la actualidad existen varias tecnologías basadas en inteligencia artificial que tienen como objetivo colaborar en las actividades que realizan las personas, por ende, los humanos tienen mayor tiempo libre para realizar otras actividades (Rouhiainen, 2018).

En otras palabras, la inteligencia artificial ha sido creada para ayudar al ser humano a realizar distintas tareas, aprendiendo mediante algoritmos de aprendizaje para poder tener una mejor toma de decisiones en las diferentes situaciones que se pueda encontrar una persona, incluso también beneficia en la toma de decisiones en las empresas.

1.9.1.1. Capacidades de la Inteligencia Artificial.

García (2012) indica que en 1950 Turing publicó un artículo denominado Computing machinery and intelligence donde mencionaba que se puede llamar a una máquina inteligente si es que esta tiene la capacidad de actuar como un humano. A pesar del tiempo que ha transcurrido desde la propuesta de Turing, sigue siendo importante porque establece capacidades que necesita una máquina para ser inteligente, lo que hace referencia a lo que se conoce hoy en día como Inteligencia Artificial, a continuación, se muestran las capacidades que según el Test de Turing debe tener una máquina para considerarse inteligente.

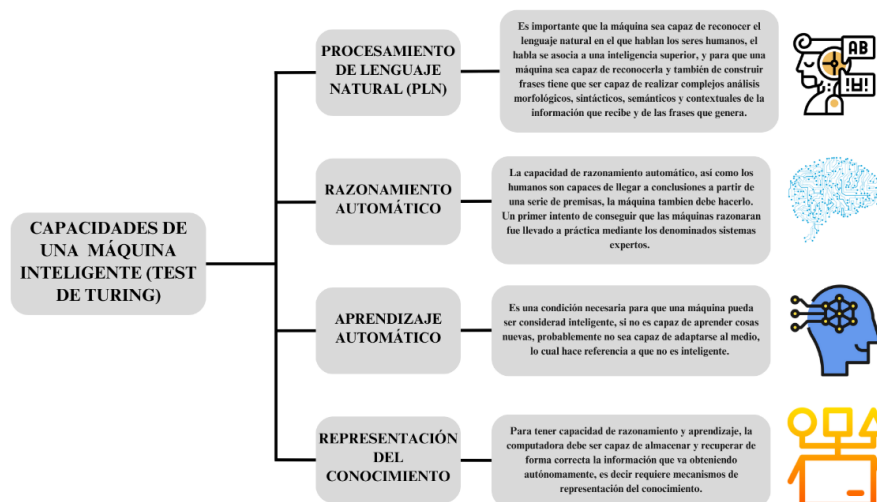


Figura 14. Capacidad de una máquina inteligente según el test de Turing. Información adaptada de (García, 2012). Elaborado por Jiménez Kimberly.

1.9.2. Machine learning.

El aprendizaje automático, conocido como machine learning en el idioma inglés, es cuando a una computadora se le da una cantidad masiva de datos y un algoritmo para que los pueda analizar y la computadora puede aprender a realizar una tarea de una manera similar a la capacidad de aprendizaje de un ser humano (Komuro et. al, 2023).

1.9.3. Procesamiento del lenguaje natural (PLN).

García (2012) menciona que el procesamiento del lenguaje natural o NLP es una rama de la Inteligencia Artificial que se responsabiliza de las capacidades de comunicación de las máquinas con las personas haciendo uso de su propio lenguaje.

El procesamiento del lenguaje natural consiste en la habilidad de una máquina para procesar información comunicada mediante el uso del lenguaje natural. Crean modelos computacionales del lenguaje suficientemente detallados que permitan escribir programas informativos que realicen distintas órdenes o peticiones donde interviene el lenguaje natural. Se podría decir que el NLP consiste en usar una expresión natural que pueda tener comunicación con la computadora directamente, por medio escrito o comando de voz, facilitando las órdenes o peticiones con el lenguaje, o seguir desarrollando modelos que ayuden a la comprensión humana y sus mecanismos que se relacionan al lenguaje (Gelbukh, 2010).

1.9.3.1. Campos del procesamiento de lenguaje natural.

Los distintos campos de proceso del PLN (Procesamiento de Lenguaje Natural), también conocido como NLP (Natural Language Processing) por sus siglas en inglés, son los siguientes: el restablecimiento y extracción de información; el análisis de sentimientos; el modo de investigación y consultas; la reproducción de síntesis automáticas; y la minería de datos (Hernández & Gómez, 2013).

Por otro lado, Hernández & Gómez (2013) argumentan que los campos de desarrollo de PLN son: la recuperación y extracción de información, traducción automática, sistemas de búsquedas de respuestas, generación de resúmenes automáticos, minería de datos, análisis de sentimientos, entre otras.

1.9.4. Análisis de sentimientos.

El análisis de sentimientos conocido también como minería de opinión es un área de la minería de textos que se sostiene de otros campos como la inteligencia artificial y el aprendizaje automático. En este campo del procesamiento del lenguaje natural, la opinión es evaluada según su polaridad en positiva, negativa, neutral y conflictiva (Sánchez del Hoyo, 2019).

El análisis de sentimientos es la disciplina que tiene como objetivo principal analizar todas las emociones que un humano puede expresar, tales como, sentimientos, opiniones,

actitudes y valoraciones hacia determinadas entidades y atributos que se encuentran inmersos en un texto.

1.9.4.1. Niveles del análisis de sentimiento.

Según Liu (2015) el análisis de sentimiento se ha realizado principalmente en tres niveles: nivel de documento, nivel de oración y nivel de aspecto.

- Nivel de documento: Consiste en clasificar que en un documento exprese la opinión completa si el sentimiento que se encuentra en el mismo es de forma positiva o negativa.
- Nivel de oración: Se trata de determinar si cada oración expresa una opinión positiva, negativa o neutral, considerando que la “opinión neutral” comúnmente significa “sin opinión”, se relaciona con la clasificación de subjetividad.
- Nivel de aspecto: Para tener un nivel de resultados detallados, ir más allá de la clasificación positiva o negativa, el análisis de nivel de aspecto se concentra directamente en la opinión y su objetivo.

1.9.4.2. Tareas del análisis de sentimiento.

Es importante considerar que este campo perteneciente al procesamiento del lenguaje natural está compuesto por el desarrollo de las siguientes tareas (Pozzi et al., 2017).



Figura 15. Tareas del análisis del sentimiento. Información adaptada de (Pozzi et al., 2017). Elaborado por Jiménez Kimberly.

1.9.4.3. Técnicas para la clasificación de los sentimientos.

Los enfoques existentes dentro de un análisis de sentimientos se pueden juntar en cuatro categorías que son:

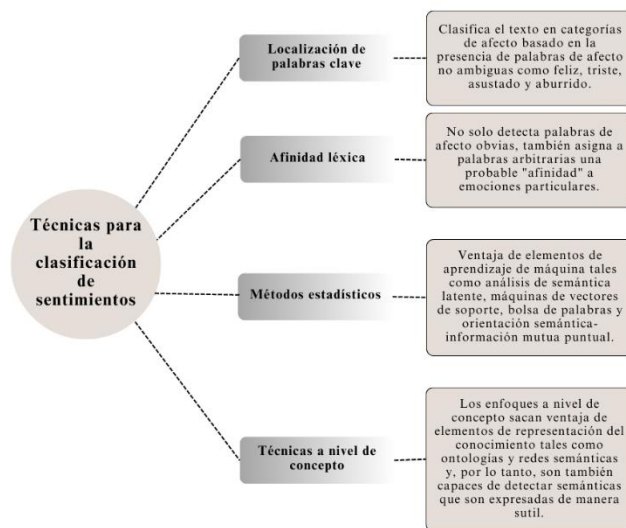


Figura 16. Técnicas para la clasificación de sentimientos. Información adaptada de a (Cambria et al., 2013). Elaborado por Jiménez Kimberly.

Primordialmente lo que se trata de hacer con el análisis de sentimientos es encontrar opiniones, identificar los sentimientos que expresan las personas, y luego clasificar su polaridad, por lo tanto, se considera un proceso de clasificación.

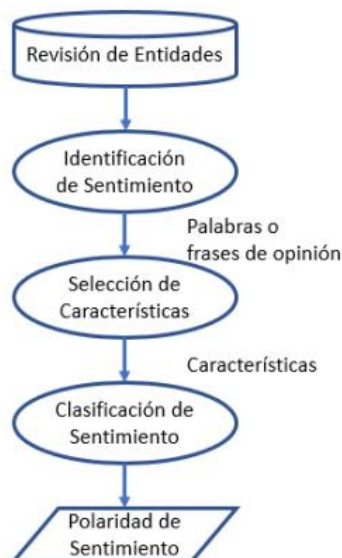


Figura 17. Proceso de análisis de sentimiento. Información recuperada de (Romero Moreno et al., 2020)

1.9.5. Preprocesamiento de los datos.

Es necesario preprocesar los datos con los que realizará el análisis de sentimientos, según Alva Segura (2020) es fundamental contar con un proceso para normalizar el texto y de igual manera para el corpus de entrenamiento y pruebas, es importante que el texto preprocesado

no pierda la polaridad inicial del mensaje, a continuación, se muestran algunas de las tareas de preprocesamiento que se deben realizar.

1. Eliminación de tildes: Existen usuarios que no utilizan adecuadamente el uso de las tildes, por eso al utilizarlas de manera incorrecta, el algoritmo puede considerarla de una manera distinta. Se reemplazan las vocales con tilde por vocales sin tilde.
2. Normalización de mayúsculas y minúsculas: Se normalizan las palabras que estén en mayúsculas o que contengan una mayúscula, se las convierte en minúsculas.
3. Tratamiento de duplicidad de caracteres: En ocasiones los usuarios de las redes sociales repiten los caracteres como por ejemplo “Gaaaanamooos”, por lo que una palabra escrita de esta manera no podrá ser entendido por el algoritmo.
4. Normalización de risas jergas: En las redes sociales las personas se expresan de manera rápida, por ende, existen personas que se expresan en sus comentarios con jergas como “TQM”, “q”, en algunos casos cuando se normalizan pueden aportar sentido de polaridad para el posterior análisis de sentimientos.
5. Tratamiento de emoticones: Los usuarios en Twitter suelen expresar sentimientos en manera de emoticones, por lo tanto, es necesario sustituir el emoticón por su significado textual.

tweetId	Texto Normal.	Texto Preprocesado
1218980593343639552	RT @CesarBejarano21: @CayetanaAljovin Por eso no podemos volver a votar por #FuerzaPopular #Apra #PodemosPerú #VamosPerú #PerúPatriaSegura	Por eso no podemos volver a votar por fuerza popular apra podemosperu Vamos Perú Perupatriasegura
1217866192389623808	#AHORA @DanielUrresti1 candidato al Congreso por #PodemosPerú ofrece conferencia de prensa luego de que el @JNE_Peru lo reincorporara a las elecciones #Elecciones2020 #EleccionesCongresales https://t.co/JGopHhmZTi	ahora candidato al congreso por podemosperu ofrece conferencia de prensa luego de que el lo reincorporara a las elecciones Elecciones2020 elecciones Congresales

Figura 18. Ejemplo de texto preprocesado. Información recuperada de (Alva Segura, 2020).

1.9.5.1. Lematización.

Alva Segura (2020) afirma que la lematización es un proceso que cambia el lema de una palabra haciendo uso del diccionario, esto se realiza con el fin de obtener una oración gramatical no sujeta a variabilidad, por ejemplo, la palabra perdedores se podría remplazar por la palabra perdedor.

Por ende, el proceso de lematizar es obviar las diferencias y unir todas las variantes en un solo término.

1.9.5.2. Tokenización.

Según (Mayo, 2018) la tokenización consiste en que se fragmentan las cadenas de texto más largas en piezas más pequeñas o tokens. Los trozos de texto más grandes pueden ser convertidos en oraciones, las oraciones pueden ser tokenizadas en palabras.

La tokenización también se conoce como segmentación de texto o análisis léxico. En la etapa de la tokenización las frases se dividen en pequeñas unidades denominadas tokens.

1.9.6. Metodologías en el desarrollo de proyectos de ciencia de datos.

Existen varias metodologías para realizar proyectos en Data Science, entre ellas se muestran las siguientes:

- KDD (Knowledge Discovery in Databases)
- CRISP-DM (Cross-Industry Standard Process for Data Mining)
- SEMMA (Sample, Explore, Modify, Model and Access)

1.9.6.1. Metodología KDD (*Knowledge Discovery in Databases*).

En la actualidad se pueden encontrar varias definiciones acerca de la metodología KDD (Knowledge discovery in databases). Entre los mejores conceptos se encuentra el siguiente:

El descubrimiento de conocimiento en bases de datos, también conocido como KDD, es un campo de la inteligencia artificial de rápido crecimiento, que combina técnicas del aprendizaje de máquina, reconocimiento de patrones, estadística, bases de datos, y visualización para automáticamente extraer conocimiento o información, de una base de datos. (Nigro et. al, 2004)

La metodología KDD (Knowledge Discovery in Databases) está compuesta por cinco pasos que se muestran a continuación:

1. Selección: Se empieza con la selección de un data set principal, del mismo se selecciona un subconjunto de variables que sea útil para el tema que se está estudiando.
2. Preprocesamiento: Se realiza la limpieza y normalización de datos.
3. Transformación: El método sugiere que se reduzcan las dimensiones con técnicas estadísticas para manipular la menor cantidad de variables.

4. Minería de datos: En minería se buscan patrones de interés o representativos en relación con el objetivo de la minería de datos, posteriormente, para colar al conocimiento se procede proceso de interpretación y evaluación de modelo.
5. Evaluación: Para finalizar, se otorga una calificación al modelo y si no se cumplen de manera satisfactoria los objetivos se repite una y otra vez hasta que sean logrados.

1.9.6.2. Metodología CRISP-DM.

DATLAS (2020) indica que la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), es el método más utilizado en la industria de la ciencia de datos.

- Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.
- Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos. (IBM Documentation, 2021)

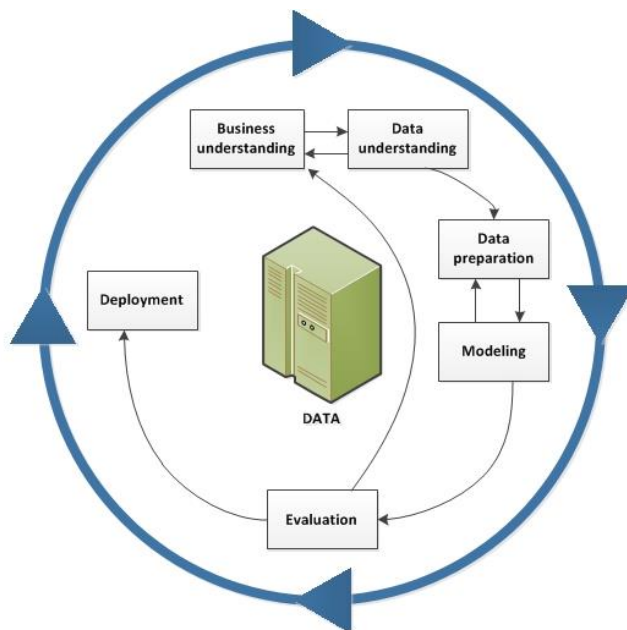


Figura 19. Ciclo de vida de minería de datos. Información recuperada de (IBM Documentation, 2017)

CRISP-DM (Cross-Industry Standard Process for Data Mining) se divide en las siguientes etapas:

1. Entendimiento de negocio: En esta etapa se determinan los objetivos de negocio, objetivos estratégicos.

2. Entendimiento de los datos: Se recolectan los datos iniciales, se detallan cada uno de estos, se exploran y para finalizar se verifica la calidad de la información.
3. Preparación de los datos: Se selecciona la información más razonable, se limpia, etc. Para poder obtener un entregable que sería un conjunto de datos.
4. Modelamiento: Se experimenta con distintas técnicas, se consideran supuestos, se realizan pruebas, se definen parámetros y se revisa la funcionalidad de los modelos.
5. Evaluación: Se considera si la evaluación es positiva o negativa, se determinan los siguientes pasos y se toma las decisiones necesarias.
6. Despliegue: Sólo se activa si el proyecto tuvo una evaluación positiva.

1.9.6.3. Metodología SEMMA (*Sample, Explore, Modify, Model and Access*)

DATLAS (2020) agumenta que esta metodología está conformada por las siguientes fases:

1. Muestra: En esta fase se realiza un muestreo de la gran base de datos principal para poder manipular un pequeño conjunto de datos de una manera ágil.
2. Explorar: Se exploran los datos para entender y generar ideas respecto a la información, además de buscar cualquier anomalía, patrones y tendencias.
3. Modificar: Esta fase se enfoca en crear, seleccionar y transformar variables para enfocarse en un proceso de selección, aquí también se busca reducir el número de variables.
4. Modelo: Consiste en aplicar los distintos métodos estadísticos evaluando sus fortalezas y cumplimiento de objetivos.
5. Acceso: Se evalúa la utilidad y fiabilidad de los hallazgos.

A continuación, se muestra una tabla de comparación de las metodologías utilizadas en proyectos de ciencia de datos.

Tabla 2. *Comparación de las metodologías en ciencia de datos.*

KDD	SEMMA	CRISP-DM
		Entendimiento de negocio
Selección	Muestra	Entendimiento de los datos
Preprocesamiento	Explorar	
Transformación	Modificar	Preparación de los datos
Minería de datos	Modelo	Modelamiento

Evaluación	Acceso	Evaluación
		Despliegue

Información adaptada de (DATLAS, 2020), Elaborado por el Jiménez Kimberly

Es importante mencionar que, al comparar las metodologías, se considera que CRISP-DM es más completa como flexible y se puede adaptar fácilmente.

1.9.7. Algoritmos para análisis de sentimientos.

Existen distintos algoritmos para realizar el análisis de sentimientos, a continuación, se detallan algunos.

Tabla 3. *Algoritmos de análisis de sentimientos.*

ALGORITMO	DESCRIPCIÓN
BERT	Es un modelo de incrustación de palabras que, en comparación con otros modelos de incrustación de palabras (como Word2Vec), es capaz de distinguir el significado de palabras iguales o similares que se usan en diferentes contextos. Al igual que con otros modelos de incrustación, debe usarse junto con un clasificador. En la literatura, el uso del término BERT solo se asocia con un modelo hecho de BERT seguido de una red neuronal densa.
Frecuencia de término - Frecuencia de documento invertida (TF-IDF)	Es el producto de Term Frequency (TF) e Inverse Document Frequency (IDF); donde IDF es una puntuación que mide la importancia de un término. Las puntuaciones que ocurren con poca frecuencia tienen una puntuación IDF alta. Por tanto, TF-IDF aumenta proporcionalmente al número de veces que aparece una palabra en el documento y se compensa con el número de documentos del corpus que contienen la palabra.
Bolsa de palabras (BOW)	Es una representación de texto que describe la ocurrencia de palabras dentro de un documento. Implica dos cosas: 1) un vocabulario de palabras conocidas; 2) una medida de la presencia de palabras conocidas. Se descarta cualquier información sobre el orden o la estructura de las palabras en el documento.

Información adaptada de (Capuano et. al, 2023), Elaborado por el Jiménez Kimberly.

1.9.8. Cuadernos colaborativos para ciencia de datos.

Los cuadernos colaborativos son entornos de programación que por lo general permiten escribir código en distintos lenguajes de programación y suelen ejecutarse en un navegador web, de esta manera el usuario interactúa con el código y esto le permite analizar con mayor facilidad los datos.

1.9.8.1. Deepnote.

Priya (2023) afirma que Deepnote es un entorno de notebook Jupyter que se encuentra basado en la nube y diseñado para la ciencia de datos, es gratuito y dispone de características útiles como:

- Aprovisionamiento para realizar consultas de los datos utilizando SQL de BigQuery, PostgreSQL y Snowflake.
- Permite utilizar SQL y Python en la misma interfaz de notebook sin necesidad de cambiar de aplicación.
- Dispone de soporte para distintos lenguajes de programación como Julia, Python y R.
- Es compatible con marcos de aprendizaje profundo como PyTorch y TensorFlow.

1.9.8.2. Jupyter.

Priya (2023) indica que Jupyter es un cuaderno interactivo basado en la web, comúnmente se utiliza en proyecto de ciencia de datos con lenguajes de programación como Python, Scala y R. A continuación, se detallan algunas características:

- Permite la recopilación, limpieza, análisis y visualización de datos.
- Puede construir e interpretar modelos de aprendizaje automático.
- Permite adicionar ecuaciones matemáticas y texto enriquecido.

1.9.8.3. Google Colaboratory.

Según Bisong (2019) indica que Google Colaboratory, más conocido como "Google Colab" o "Colab", es un proyecto de investigación para desarrollar prototipos de modelos de aprendizaje automático en potentes opciones de hardware, como GPU y TPU y es gratuito.

Una de las mayores ventajas de Colab es que permite a cualquier usuario escribir y ejecutar código en lenguaje Python en el navegador de Google, además, es adecuado para tareas de aprendizaje automático, análisis de datos y educación (Google Colab, 2023).

1.9.9. IDEs de programación.

Un entorno de desarrollo integrado (IDE) es una aplicación de software que permite a los programadores desarrollar de manera eficiente, además, los desarrolladores pueden combinar capacidades tales como, editar, crear, probar y empaquetar software en una aplicación sencilla de utilizar (Amazon Web Services, 2023).

1.9.9.1. Visual Studio Code.

Visual Studio Code es un editor de código fuente desarrollado por Microsoft para Windows, Linux y macOS, en el que se puede escribir código en distintos lenguajes de programación, tales como, Python, C++, JavaScript, entre otros. Además, es rápido y liviano lo que permite ver, editar, ejecutar y depurar código fuente para aplicaciones (IT, 2021).

1.9.9.2. RStudio.

RStudio es un IDE de desarrollo completo que permite desarrollar web con R y otros lenguajes de programación que se enfocan al tratamiento de grandes cantidades de datos y dispone de una serie de herramientas que se centran en la gestión de proyectos (Zúñiga, 2020).

1.9.9.3. Spyder.

Spyder es un entorno de desarrollo integrado (IDE), conocido también como el IDE científico de Python para la ciencia de datos, es de código abierto basado en Python, tiene como enfoque principal el análisis de datos, la investigación y la creación de paquetes científicos junto con el desarrollo de Python (Redessy, 2021).

1.9.10. Lenguajes de programación.

1.9.10.1. Python.

Python es un lenguaje interpretado que tiene como objetivo enfatizar la legibilidad del código usando una sintaxis simple y evitar casos especiales y excepciones. Utiliza un sistema de tipo dinámico y admite múltiples paradigmas de programación (Guzzi, 2019).

1.9.10.2. R.

R es un lenguaje de programación de código abierto orientado al trabajo con datos y su análisis estadístico, por lo general se utiliza en el contexto de investigación matemática, minería de datos y aprendizaje automático (Zúñiga, 2020).

1.9.11. APIs y librerías.

Para llevar a cabo un proyecto de ciencia de datos enfocado al análisis de sentimientos, es importante utilizar librerías que permitan el desarrollo del proyecto.

Una librería es uno o varios archivos escritos en algún lenguaje de programación delimitado y que proporciona varias funcionalidades (Think Technology Consulting, 2018).

1.9.11.1. Natural language toolkit (NLTK).

El Kit de herramientas de lenguaje natural es una plataforma utilizada para crear programas de Python para trabajar con datos de lenguaje humano, la cual ofrece interfaces sencillas y dispone de un conjunto de bibliotecas de procesamiento de texto para

clasificación, tokenización, lematización, etiquetado, análisis y razonamiento semántico (NLTK, 2023).

1.9.11.2.Textblob.

TextBlob es un modelo de biblioteca basado en léxico de análisis de sentimientos disponible en Python que proporciona un procesamiento de texto simplificado.

TextBlob es una popular biblioteca de código abierto enfocada en el procesamiento de datos de texto, proporcionando una API simple para entrar en procesamiento de lenguaje natural (NLP), una de las tareas que incluye es el análisis de sentimientos (kexugit, 2018).

1.9.11.3.Scikit-Learn.

Scikit-Learn es un módulo de Python que integra una amplia gama de algoritmos de aprendizaje automático de última generación para problemas supervisados y no supervisados de mediana escala. Este paquete se centra en llevar el aprendizaje automático a los no especialistas mediante un lenguaje de alto nivel de propósito general (Pedregosa, et. al, 2011).

1.9.12. Arquitectura de la propuesta.

La arquitectura de la presente propuesta se encuentra planificada de la siguiente manera:

En primer lugar, en el conjunto de datos tratado se encontrarán incluidas variables referentes al texto de la noticia, a los comentarios y a la identificación de la noticia.

Considerando estas variables, se procede a realizar el proceso de preprocesamiento de datos, en donde se ejecutará la limpieza de stop words, eliminación de caracteres duplicados, tratamiento de emoticones, lematización y traducción de idioma. Luego, se tokenizará el texto de las noticias y los comentarios para proceder a aplicar algoritmos de PLN que servirán para categorizar los sentimientos en sátira, humor, felicidad y tristeza.

Posteriormente, se tendrán como resultados, una lista de sentimientos en el cuaderno Google Colaboratory que muestre las noticias y comentarios junto con la categoría de sentimiento. Y para finalizar, los resultados del análisis de sentimientos se almacenarán en una base de datos no relacional.

1.10. Marco Legal

En el presente trabajo investigativo se fundamenta en la ley, políticas y normas que se especifican a continuación.

1.10.1. Ley orgánica de comunicación (LOC).

Es fundamental tener en consideración la ley orgánica de comunicación, porque los ciudadanos ecuatorianos tienen derecho a libertad de expresión, siempre que se realice con responsabilidad, además tienen derecho a recibir información veraz.

LEY ORGANICA DE COMUNICACION

TITULO I

Disposiciones preliminares y definiciones

Art. 1.- Esta ley tiene por objeto desarrollar, proteger, promover, garantizar, regular y fomentar, el ejercicio de los derechos a la comunicación establecidos en los instrumentos de derechos humanos y en la Constitución de la República del Ecuador.

Además, el objeto de esta Ley comprenderá la protección del derecho a ejercer la libertad de expresión, y a buscar, recibir y difundir información e ideas de toda índole a través de medios de comunicación. (Asamblea Nacional, 2019)

1.10.2. Código de principios de la IFCN.

La International Fact-Checking en Poynter, dispone de un Código de principios que consiste en una serie de compromisos que las organizaciones cumplen para promover la excelencia en la verificación de datos. A continuación, se mencionan los principios de la IFCN.

1. Compromiso con la independencia e imparcialidad.
2. Transparencia de las fuentes.
3. Transparencia en la financiación.
4. Transparencia con la metodología.
5. Compromiso con una corrección abierta y honesta. (IFCN Code of Principles, 2023)

Es importante considerar que estos principios son utilizados por los verificadores de hechos nacionales Ecuador Chequea y Ecuador verifica tal como lo señala la IFCN. Además, en las páginas de los verificadores los usuarios pueden reportar la violación de estos mandatos comunicándose con la IFCN (Ecuador Chequea, 2023; Ecuador Verifica, 2023).

1.10.2.1. Política de rectificaciones de Ecuador Chequea y Ecuador Verifica.

“Ecuador Chequea y Ecuador Verifica no acogen promesas, propuestas proselitistas ni deseos, sino enunciados que son confrontables con hechos concretos de la realidad o con datos que nutren bases oficiales, académicas e independientes” (Ecuador Chequea, 2023).

- 1. Enviar un correo o mensaje con la solicitud correspondiente. En el asunto es necesario incluir: “Solicitud de rectificación”. En el cuerpo del mensaje se señalará el contenido que se considera errado, anexando el enlace a la nota publicada. *Es importante adjuntar las fuentes a través de las cuales se concluye que la categoría consignada a la frase verificada no es la correcta o que los datos expuestos están errados.**
- 2. El equipo de Ecuador Chequea revisará y someterá a análisis la información, en un plazo máximo de dos días después de recibida la solicitud.**
- 3. Tras el plazo mencionado, en caso de que se concluya que la verificación merece una rectificación, inmediatamente se incluirá en la publicación los contenidos nuevos, agrupados bajo el subtítulo “Rectificación”. Una vez actualizado el artículo, inmediatamente se difundirá a través de las redes de Ecuador Chequea (Facebook, Twitter e Instagram), con el sello de “Actualización” y “Rectificación”.**
- 4. Tras el plazo de máximo dos días mencionado, en caso de que se concluya que la verificación no merece rectificación, se responderá al remitente del correo electrónico con los argumentos periodísticos por los cuales el equipo de Ecuador Chequea mantiene la categoría con que se ha catalogado la frase o los datos que se han expuesto.**

1.10.3. Política de Twitter para desarrolladores.

Twitter tiene como objetivo brindar acceso abierto a los datos para los desarrolladores, para que las personas interesadas puedan hacer uso de estos datos cumpliendo con la siguiente política de uso.

Uso de esta política

Hemos estructurado esta política para que sea lo más fácil de seguir posible.

Tenga presente la información de las siguientes secciones de la política cuando utilice la API de Twitter y el contenido de Twitter:

1. Prepárese para el éxito: usted es responsable de cumplir con todas las políticas de Twitter. Es importante que revise y comprenda esta Política, así como las políticas a las que nos vinculamos en este documento, antes de acceder a la API de Twitter y al Contenido de Twitter. El tiempo dedicado a revisar nuestras políticas puede ahorrarle horas de reelaboración en el futuro.

2. La privacidad y el control son esenciales: proteger y defender la privacidad de las personas en Twitter está integrado en el ADN central de nuestra empresa. Como tal, prohibimos el uso de datos de Twitter de cualquier manera que sea inconsistente con las expectativas razonables de privacidad de las personas. Al desarrollar la API de Twitter o acceder al contenido de Twitter, usted tiene un papel especial que desempeñar para salvaguardar este compromiso, lo que es más importante, respetando la privacidad de las personas y brindándoles transparencia y control sobre cómo se utilizan sus datos.

3. Siga las pautas de uso de la plataforma: obtener la aprobación para acceder a la API de Twitter y al contenido de Twitter es solo el primer paso. Nuestras Pautas de uso de la plataforma deben ser su primera parada cada vez que tenga preguntas sobre cómo garantizar el cumplimiento de la política para su uso planificado de la plataforma de Twitter. (Twitter Developers, 2023)

Es necesario tomar en consideración la política que tiene Twitter para los desarrolladores, porque se va a trabajar el análisis de sentimientos con datos extraídos de la red social Twitter.

Capítulo II

Metodología

2.1. Tipo de investigación

Considerando que el presente trabajo de titulación se enfoca en realizar el desarrollo de un análisis de sentimientos, se ha determinado llevar a cabo una investigación descriptiva y exploratoria con un enfoque cualitativo.

2.1.1. Investigación descriptiva.

La investigación descriptiva tiene como objetivo describir las características fundamentales de conjuntos homogéneos de fenómenos, utilizando criterios sistemáticos que permiten establecer el comportamiento de los fenómenos en estudio, puede emplearse con un enfoque cuantitativo o cualitativo (Guevara Alban et. al, 2020; Valle et. al, 2022).

Por estas razones, se ha considerado pertinente aplicar el tipo de investigación descriptiva para analizar y recopilar información útil con relación a las noticias falsas y el campo de procesamiento de lenguaje natural para la detección de emociones, que sirva de ayuda en el desarrollo del componente de análisis de sentimientos en las noticias de Twitter que son seleccionadas por las verificadoras acreditadas en Ecuador y los comentarios de los usuarios en estas noticias.

2.1.2. Investigación exploratoria.

La investigación exploratoria consiste en investigar una problemática que no está claramente definida, se emplea para poder comprender de mejor manera el problema existente.

Desde esta perspectiva, se ha considerado emplear este tipo de investigación para poder obtener mayor conocimiento referente a la intervención que tienen las noticias falsas en los sentimientos de las personas.

2.2. Pregunta de investigación

¿Las noticias asociadas con emociones negativas pueden ser un indicativo para determinar que probablemente estas noticias sean falsas?

2.2.1. Enfoque de la investigación.

En la presente investigación se ha determinado utilizar el enfoque cualitativo.

2.2.1.1. Enfoque cualitativo.

Koh y Owen (2000) argumentan que el enfoque cualitativo se centra en cómo suceden las situaciones o hechos, es decir que se centra en las actitudes, las creencias o las formas en

las que las personas dan sentido e interpretan las experiencias que atraviesan y el mundo que las rodea.

En cuanto a esto, la investigación cualitativa se emplea para conocer las opiniones sobre el tema de investigación analizado, comúnmente se utilizan las entrevistas como herramienta para recolectar información importante respecto a un determinado tema.

2.3. Técnicas de recolección de datos

Para poder obtener información en el presente trabajo de investigación, en el contexto de las noticias falsas, el tiempo que se necesita para llevar a cabo la verificación de una noticia y los sentimientos que se podrían asociar a una noticia falsa, se han utilizado las técnicas de entrevista y análisis documental como herramientas de recolección de datos cualitativos.

2.3.1. Entrevista.

La entrevista es una técnica de investigación que sirve para recolectar datos cualitativos, utilizando esta herramienta en el presente proyecto se puede profundizar en los sentimientos, motivaciones, y sobre todo en las opiniones de distintos perfiles para la obtención de información.

2.3.1.1. Aplicación de entrevista.

La entrevista se realizará al perfil de periodista, el cual se considera de gran importancia para el proceso de verificación del medio de comunicación especializado en la verificación de hechos Ecuador Chequea y de la coalición Ecuador Verifica, ambos certificados por la International Fact Checking Network (IFCN).

De esta manera, se conseguirá la información conveniente que servirá de apoyo para realizar el componente de análisis de sentimientos en las noticias de Twitter que son seleccionadas por las verificadoras acreditadas en Ecuador y los comentarios de los usuarios en estas noticias, tal componente que forma parte de la investigación a nivel macro que tiene con fin desarrollar el software “Sistema de ayuda para los verificadores de hechos empleando la metodología utilizada por Ecuador Chequea y Ecuador Verifica”.

Tabla 4. *Entrevista dirigida a la periodista de Ecuador Chequea.*

Perfil de Entrevista	
Cargo	Periodista
Profesión	Licenciada en Periodismo multimedia
Nombre	Lcda. Paola Simbaña Ramos
Lugar de Trabajo	Ecuador Chequea
Aporte Cualitativo	

Proporcionará información respecto al Fact-checking y la desinformación que actualmente se presentan a nivel de redes sociales, las noticias falsas y la asociación y repercusión que tienen con las emociones de las personas.

Información adaptada por la Investigación de Campo. Elaborado por Jiménez Kimberly.

2.2.1.1 Resumen de la entrevista realizada a la periodista de Ecuador Chequea.

Por medio del siguiente resumen, se expone la entrevista realizada a la periodista y Fact-checker Paola Simbaña, quien participa en el proceso de verificación de hechos en el verificador pionero a nivel nacional, Ecuador Chequea. Esta entrevista permitirá ampliar el conocimiento referente al proceso de Fact-checking que se lleva a cabo en el portal Ecuador Chequea y la coalición coordinada por el mismo, Ecuador Verifica.

Tabla 5. *Entrevista dirigida a la periodista de Ecuador Chequea.*



Universidad de Guayaquil
Sistemas de Información

Fecha de Elaboración
21/06/2023

Entrevista para el desarrollo del componente de análisis de sentimientos de las noticias comprobadas en Twitter por las verificadoras acreditadas en Ecuador utilizando procesamiento de lenguaje natural.

Lugar o Canal de Entrevista: Plataforma Zoom.

Entrevistadora: Kimberly Jiménez

Entrevistada: Lcda. Paola Simbaña Ramos

Establecimiento: Ecuador Chequea

Cargo: Periodista

Resumen

En la entrevista realizada a la Lcda. Simbaña se argumenta que la desinformación es un fenómeno muy complejo y quienes fabrican este tipo de contenido tienen como objetivo engañar a la audiencia que consume contenido en internet a través de las redes sociales, por lo que, es de gran importancia contrarrestar este inconveniente, en consecuencia de la proliferación existente de la desinformación se necesita todo un ejercicio periodístico relacionado con la investigación, para poder realizar un proceso de verificación bien sustentado, el cual cuenta con cuatro etapas, de las cuales se considera que el acceso a las fuentes para poder realizar la confrontación del texto seleccionado a verificar es la más complicada.

Además, se menciona que Ecuador Chequea cumple con los principios de la IFCN y todos los criterios que ellos exponen, sin embargo, son mucho más rigurosos con la información que se verifica, lo cual se relaciona con el tiempo que se utiliza para realizar el proceso de

Fact-checking, el cual dependerá de varias circunstancias al momento de verificar, tales como: la temática del contenido a corroborar, la respuesta de las fuentes consultadas y documentadas para poder informarse, entre otras, por estas razones, no se tiene un tiempo determinado.

En cuanto a la asociación de las emociones con la desinformación se argumenta que, las noticias falsas siempre apelan a las emociones de la ciudadanía que visualiza este tipo de contenido para lograr el engaño, ya que al acudir a las emociones las personas se dejan llevar por lo que sienten en ese momento y reaccionan al sentimiento anulando la parte lógica y analítica que se pueda aplicar al contenido antes de creer y replicar esa desinformación, sobre todo si es una noticia que ha causado algún sentimiento de temor, enojo, miedo, entre otros, ocasionando incluso afectaciones a nivel de sociedad, aumentando el nivel de polarización lo que puede provocar una sociedad fragmentada, más que todo en la política, ya que en este contexto nace la desinformación.

Adicionalmente, se menciona que actualmente se utilizan distintas herramientas que sirven de apoyo para el proceso de verificación de hechos en algunas ocasiones, en el contexto de verificar si una imagen ha sido manipulada, para identificar el origen de una fotografía o video, entre otras. Para terminar la entrevista, la periodista Simbaña brindó como sugerencia dudar de toda información que circule en internet y no difundirla sin antes verificarla en fuentes confiables.

Información adaptada por la Investigación de Campo. Elaborado por Jiménez Kimberly.

2.3.2. Análisis documental.

La recopilación documental se realiza por medio de la consulta de documentos tales como: libros, revistas, periódicos, registros, constituciones, conferencias escritas, encuestas, documentos fílmicos, documentas grabados, entre otros.

En el presente trabajo de titulación se recolectará información seleccionada de distintas documentaciones consultadas en varias fuentes basadas en el Fact Checking y en la asociación que pueden llegar a tener las noticias falsas con los sentimientos.

2.2.1.2 Aplicación del análisis documental.

El análisis documental se empleará para recopilar información respecto al Fact Checking, la desinformación, noticias falsas y los sentimientos o emociones, por lo tanto, se ha considerado importante revisar documentaciones que se enfoquen en estos temas, tales como el proyecto de investigación “PANDEMIA DE DESINFORMACIÓN Y

HERRAMIENTAS DE VERIFICACIÓN DE NOTICIAS” realizado por la Srta. Martín, en donde entrevista a varios actores del ámbito periodístico respecto al tema de verificación de hechos.

En esta documentación, se expone el criterio del editor general de Ecuador Chequea y Ecuador Verifica, Serrano Carmona (2021) indica que el fact-checking se trata de ir más allá de los flancos periodísticos.

De la misma manera, se realizó una entrevista al periodista español Mark Amorós, un referente importante que ha realizado estudios del fenómeno de las fake news y cómo influye en la vida de las personas, escribió un libro denominado Fake news, la verdad de las noticias falsas y Por qué las fake news nos joden la vida, en donde conecta las emociones, la razón y la neuropsicología para comprender la magnitud de este fenómeno. Amorós (2021) argumenta que el fact-checking es imprescindible en la era digital y debido a la infoxicación provocada por el internet tiene como consecuencia que el proceso de verificación ya no sea solo una exigencia de la noticia sino una costumbre colectiva. Respecto a la asociación de las noticias falsas con los sentimientos, Amorós (2021) indica que, las noticias falsas buscan hackear la razón a través del corazón.

2.4. Recopilación del conjunto de datos

Es importante mencionar que, en el presente trabajo de titulación se trabajarán con datos que se obtuvieron por medio de un componente partícipe del proyecto macro “Sistema de ayuda para los verificadores de hechos empleando la metodología utilizada por Ecuador Chequea y Ecuador Verifica”, estos datos se encuentran almacenados en la base de datos no relacional MongoDB, los cuales fueron previamente extraídos de la red social Twitter utilizando la librería Tweepy de Python.

2.5. Materialización de variables

En el presente proyecto se utilizarán los objetos tweet y user con sus respectivos atributos o variables extraídas de la base de datos no relacional MongoDB Atlas del trabajo de titulación RECOPIACIÓN Y EXTRACCIÓN DE TWEETS REALIZADOS POR LAS VERIFICADORAS ACREDITADAS EN EL ECUADOR POR LA IFCN PARA LA GENERACIÓN DEL CONJUNTO DE DATOS USANDO TWEETPY realizado por la Srta. Jenny Cercado. A continuación, se detallan los atributos pertenecientes a cada uno de estos objetos.

Tabla 6. *Materialización de variables.*

Objeto	Descripción	Atributo
Tweet	Los son el bloque de construcción atómico básico de todas las cosas de Twitter, conocidos también como “actualizaciones de estado”. Los objetos Tweet son el objeto principal de diversos objetos secundarios.	<ul style="list-style-type: none"> • ID • created_at • tweet_text • favorite_count • retweet_count
User	Contiene metadatos de la cuenta de Usuario de Twitter que describen al Usuario de Twitter al que se hace referencia. Los usuarios pueden crear tweets, retwittear, citar tweets de otros usuarios, responder a tweets, seguir a usuarios, ser @mencionados en tweets y pueden agruparse en listas.	<ul style="list-style-type: none"> • user_id • user_name • user_screen_name

Información adaptada por la Investigación de Campo. Elaborado por Jiménez Kimberly.

2.6. Descripción del conjunto de datos

Del conjunto de datos utilizado se procederá a utilizar las variables que se describen a continuación:

Tabla 6. *Descripción del conjunto de datos de Tweets.*

Variable	Tipo	Descripción
tweet_id	Int64	Identificador único del tweet.
created_at	String	Fecha y hora UTC cuando se creó el Tweet.

tweet_text	String	El texto del tweet (actualización de estado).
favorite_count	Integer	Cantidad aproximada de las veces en que los usuarios de Twitter han dado me gusta a determinado Tweet.
retweet_count	Integer	Cantidad aproximada de las veces en que los usuarios de Twitter han retuiteado determinado Tweet.
user_id	Int64	Identificador único del usuario.
user_name	String	El nombre del usuario, tal como se ha definido. No necesariamente el nombre de una persona. Por lo general, tiene un límite de 50 caracteres, pero está sujeto a cambios.
user_screen_name	String	El nombre de pantalla, identificador o alias con el que el usuario se identifica. Los screen_names son únicos, pero están sujetos a cambios.
user_location	String	La ubicación definida por el usuario para el perfil de esta cuenta.
user_verified	Boolean	Corroborar si el usuario es verificado.

Información adaptada por la Investigación de Campo. Elaborado por Jiménez Kimberly.

Tabla 7. Descripción del conjunto de datos Sentimientos.

Variable	Tipo	Descripción
tweet_id	Int64	Identificador único del tweet.

tweet_comment	String	Comentario de usuario en algún determinado tweet.
category	String	Categoría de sentimiento.
sentiment	String	Sentimiento identificado en el tweet.

Información adaptada por la Investigación de Campo. Elaborado por Jiménez Kimberly.

2.7. Análisis exploratorio

El análisis exploratorio también conocido como EDA, consiste en la manipulación de los datos pertenecientes a un dataset o conjunto de datos, es de gran importancia realizarlo previo a cualquier tipo de análisis que se requiera ejecutar, de esta manera se puede comprender de qué se trata el conjunto de datos.

2.7.1. Visualización del conjunto de datos

En el presente trabajo de titulación, se ha utilizado la herramienta Google Colab y el lenguaje de programación Python para tratar el conjunto de datos, el cual se ha denominado “data_tweets”, es importante visualizar de manera general el conjunto de datos, en este caso se visualizarán las 5 primeras filas del dataset utilizando la línea de código data_tweets.head(10)

	tweet_id	tweet_id_str	tweet_text	user_id	user_screen_name	user_name	user_location	user_created_at	user_verified	date	source	retweets	likes	source_url	truncated	is_reply	is_retweet	coordinates	place	is_quote_status	entities
0	1578027178915847	1578027178915847	@EQUADORQUEJEA El pueblo no estaba...	1585402011232592	Maria_Romero	Maria Romero	Azuay Ecuador	False	False	2023-05-11 19:23:37-05:00	Tweet for Android	0	0	http://twitter.com/download/android	False	ECUADORQUEJEA	NaN	NaN	False	['hashtag', 'user', 'mention', 'url']	
1	1578024404880502	1578024404880502	@EQUADORQUEJEA El pueblo no estaba...	1585402011232592	Maria_Romero	Maria Romero	Azuay Ecuador	False	False	2023-05-11 19:23:37-05:00	Tweet for Android	0	0	http://twitter.com/download/android	False	ECUADORQUEJEA	NaN	NaN	False	['hashtag', 'user', 'mention', 'url']	
2	1578020323884402	1578020323884402	@EQUADORQUEJEA El pueblo no estaba...	1585402011232592	Maria_Romero	Maria Romero	Azuay Ecuador	False	False	2023-05-11 19:23:37-05:00	Tweet for Android	0	0	http://twitter.com/download/android	False	ECUADORQUEJEA	NaN	NaN	False	['hashtag', 'user', 'mention', 'url']	
3	1578020323884402	1578020323884402	@EQUADORQUEJEA El pueblo no estaba...	1585402011232592	Maria_Romero	Maria Romero	Azuay Ecuador	False	False	2023-05-11 19:23:37-05:00	Tweet for Android	0	0	http://twitter.com/download/android	False	ECUADORQUEJEA	NaN	NaN	False	['hashtag', 'user', 'mention', 'url']	
4	1578020323884402	1578020323884402	@EQUADORQUEJEA El pueblo no estaba...	1585402011232592	Maria_Romero	Maria Romero	Azuay Ecuador	False	False	2023-05-11 19:23:37-05:00	Tweet for Android	0	0	http://twitter.com/download/android	False	ECUADORQUEJEA	NaN	NaN	False	['hashtag', 'user', 'mention', 'url']	
5	1578020323884402	1578020323884402	@EQUADORQUEJEA El pueblo no estaba...	1585402011232592	Maria_Romero	Maria Romero	Azuay Ecuador	False	False	2023-05-11 19:23:37-05:00	Tweet for Android	0	0	http://twitter.com/download/android	False	ECUADORQUEJEA	NaN	NaN	False	['hashtag', 'user', 'mention', 'url']	
6	1578020323884402	1578020323884402	@EQUADORQUEJEA El pueblo no estaba...	1585402011232592	Maria_Romero	Maria Romero	Azuay Ecuador	False	False	2023-05-11 19:23:37-05:00	Tweet for Android	0	0	http://twitter.com/download/android	False	ECUADORQUEJEA	NaN	NaN	False	['hashtag', 'user', 'mention', 'url']	
7	1578020323884402	1578020323884402	@EQUADORQUEJEA El pueblo no estaba...	1585402011232592	Maria_Romero	Maria Romero	Azuay Ecuador	False	False	2023-05-11 19:23:37-05:00	Tweet for Android	0	0	http://twitter.com/download/android	False	ECUADORQUEJEA	NaN	NaN	False	['hashtag', 'user', 'mention', 'url']	
8	1578020323884402	1578020323884402	@EQUADORQUEJEA El pueblo no estaba...	1585402011232592	Maria_Romero	Maria Romero	Azuay Ecuador	False	False	2023-05-11 19:23:37-05:00	Tweet for Android	0	0	http://twitter.com/download/android	False	ECUADORQUEJEA	NaN	NaN	False	['hashtag', 'user', 'mention', 'url']	
9	1578020323884402	1578020323884402	@EQUADORQUEJEA El pueblo no estaba...	1585402011232592	Maria_Romero	Maria Romero	Azuay Ecuador	False	False	2023-05-11 19:23:37-05:00	Tweet for Android	0	0	http://twitter.com/download/android	False	ECUADORQUEJEA	NaN	NaN	False	['hashtag', 'user', 'mention', 'url']	

Figura 20. EDA - Visualización general del conjunto de datos. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

2.7.2. Preprocesamiento del conjunto de datos

Es fundamental realizar un preprocesamiento del conjunto de datos, en primer lugar, se verifica el número de filas y columnas existentes en el dataset, esto se puede realizar con la sentencia data_tweets.shape, en donde se identifica que existen 813 filas y 21 columnas.

data_tweets.shape
(813, 21)

Figura 21. EDA – Número de columnas y filas en el conjunto de datos sin preprocesar. Información adaptada de a (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Además, es importante visualizar los nombres de las columnas o atributos que se pertenecen al conjunto de datos, utilizando la sentencia `data_tweets.columns`, como se muestra a continuación.

```
data_tweets.columns
Index(['tweet_id', 'tweet_id_str', 'tweet_text', 'user_id', 'user_sreen_name',
      'user_name', 'user_location', 'user_contributors_enabled',
      'user_verified', 'date', 'lang', 'source', 'retweets', 'likes',
      'source_url', 'truncated', 'in_reply_to_screen_name', 'coordinates',
      'place', 'is_quote_status', 'entities'],
      dtype='object')
```

Figura 22. EDA – Nombres de columnas en el conjunto de datos sin preprocesar. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Al visualizar los nombres de las columnas, identificamos las columnas o atributos que se necesitarán para realizar el análisis exploratorio y posteriormente el análisis de sentimientos, las que no se utilizarán se eliminarán con la sentencia de código:

```
data_tweets = data_tweets.drop(['user_contributors_enabled', 'source', 'source_url',
                                'truncated', 'in_reply_to_screen_name', 'coordinates',
                                'place', 'is_quote_status'], axis=1 )
```

Figura 23. EDA – Eliminación de columnas. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Luego, se procede a renombrar algunas columnas o variables para su mejor comprensión, con la línea de código:

```
data_tweets = data_tweets.rename(columns={"date": "created_at", "retweets": "retweet_count",
                                           "likes": "favorite_count"})
```

Figura 24. EDA – Renombramiento de columnas. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Una de las acciones importantes que se debe ejecutar, es la verificación del tipo de datos que tienen las variables, eso se realiza con la sentencia `data_tweets.dtypes`, también es necesario corroborar la existencia de valores faltantes o nulos en el conjunto de datos, por lo tanto, se utiliza la línea de código `data_tweets.isnull().sum()`.

```
data_tweets.dtypes
```

tweet_id	int64
tweet_id_str	int64
tweet_text	object
user_id	int64
user_sreen_name	object
user_name	object
user_location	object
user_verified	bool
date	object
lang	object
retweets	int64
likes	int64
entities	object
dtype:	object

Figura 25. EDA – Visualización del tipo de variables. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

```
data_tweets.isnull().sum()
```

tweet_id	0
tweet_id_str	0
tweet_text	0
user_id	0
user_sreen_name	0
user_name	0
user_location	324
user_verified	0
created_at	0
lang	0
retweet_count	0
favorite_count	0
entities	0
dtype:	int64

Figura 26. EDA – Corroboración de valores faltantes o perdidos. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Se ha verificado que existen valores faltantes en la columna `user_location`, por lo tanto, es necesario tratarlos, debido a la gran cantidad de información que se tiene en el dataset, se ha procedido con la eliminación de estos valores con la sentencia `data_tweets.isnull().sum()` y posteriormente comprobar que ya no existan estos valores faltantes con la línea de código `data_tweets.isnull().sum()`.

```
data_tweets.dropna(inplace=True)
data_tweets.isnull().sum()
```

tweet_id	0
tweet_id_str	0
tweet_text	0
user_id	0
user_sreen_name	0
user_name	0
user_location	0
user_verified	0
created_at	0
lang	0
retweet_count	0
favorite_count	0
entities	0
dtype: int64	

Figura 27. EDA – Corroboración de valores faltantes o perdidos. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Posteriormente, es necesario verificar después del preprocesamiento, el número de columnas y filas resultantes, se lo verifica con la sentencia `data_tweets.shape`.

2.7.3. Desarrollo del análisis exploratorio del conjunto de datos.

El análisis exploratorio en el presente trabajo se encuentra enfocado en un conjunto de datos que fueron previamente extraídos de Twitter utilizando Tweepy de Python, posteriormente obtenido de la base de datos no relacional MongoDB.

En primer lugar, se corrobora la cantidad de tweets que tiene el conjunto de datos, posteriormente, por medio de un gráfico de barras se muestran los diez primeros nombres de pantalla de usuario que se encuentran en el dataset tratado, con la sentencia de código `data_tweets['user_sreen_name'].value_counts().head(10).plot.bar()`.

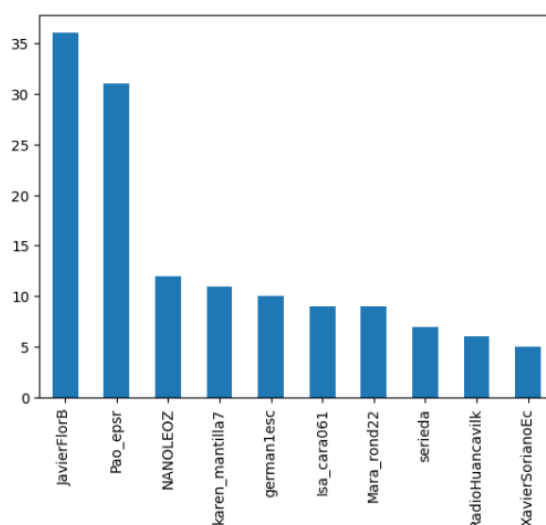


Figura 28. EDA – Gráfico de barras de Screen name de usuarios. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

A continuación, se verifica si entre los usuarios que tiene el conjunto de datos se encuentra algunos de los integrantes de Ecuador Chequea y Ecuador Verifica, en este caso, se tomará como ejemplo el usuario del editor general “Alexis Serrano Carmona”, se lo filtrará de la siguiente manera.

```
Integrante_Ecuador_Chequea = data_tweets [data_tweets["user_name"]=="Alexis Serrano Carmona"]
print (Integrante_Ecuador_Chequea)
```

	tweet_id	tweet_id_str \	tweet_text	user_id \
4	1678590115206119424	1678590115206119424	RT @ECUADORCHEQUEA: El impacto de El Niño se d...	233340840
430	1676363307265974272	1676363307265974272	RT @ECUADORCHEQUEA: #ATENCIÓN 1 Ante la imine...	233340840
454	1676345544103305216	1676345544103305216	RT @ECUADORCHEQUEA: #AlGrano Una comisión f...	233340840
455	1676345519910658049	1676345519910658049	RT @ECUADORCHEQUEA: #AlGrano No hay un cron...	233340840
525	1676262265807306752	1676262265807306752	RT @ECUADORCHEQUEA: #AHORA El presidente, @L...	233340840

	user_sreen_name	user_name	user_location	user_verified \
4	alexsserranocar	Alexis Serrano Carmona	Quito	False
430	alexsserranocar	Alexis Serrano Carmona	Quito	False
454	alexsserranocar	Alexis Serrano Carmona	Quito	False
455	alexsserranocar	Alexis Serrano Carmona	Quito	False
525	alexsserranocar	Alexis Serrano Carmona	Quito	False

	created_at	lang	retweet_count	favorite_count \
4	2023-07-11 02:20:44+00:00	es	4	0
430	2023-07-04 22:52:12+00:00	es	5	0
454	2023-07-04 21:41:37+00:00	es	3	0
455	2023-07-04 21:41:31+00:00	es	4	0
525	2023-07-04 16:10:42+00:00	es	39	0

Figura 29. EDA – Verificación de user_name. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

En el siguiente gráfico de barras se observan los diez primeros usuarios que están presentes en el conjunto de datos analizado, en el cual también se identifican usuarios de fact-checkers de Ecuador Chequea y Ecuador Verifica, ejecutando la sentencia `data_tweets['user_name'].value_counts().head(10).plot.bar()`

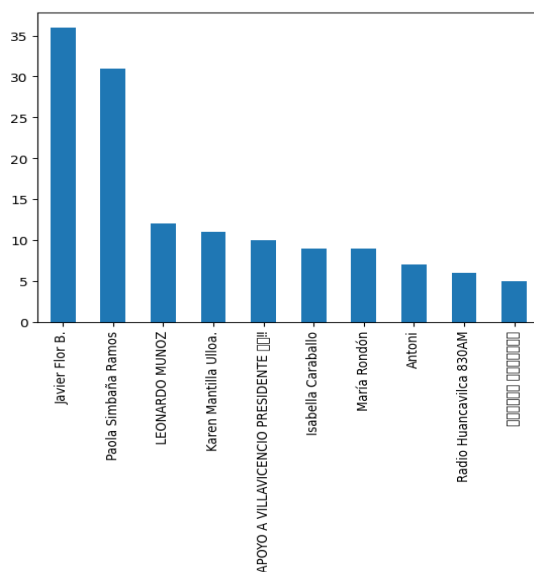


Figura 30. EDA – Gráfico de barras de nombres de usuarios. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

A continuación, mediante un gráfico boxplot se realiza una comparación entre un usuario que forma parte de Ecuador Chequea y Ecuador Verifica (fact-checker) y un usuario común. En este gráfico se puede visualizar que una cuenta de un verificador tiene más retweets que una de un usuario en común, se podría indicar que esto se debe a la credibilidad que tiene la cuenta al ser de un usuario integrante de la Verificadora de hechos Ecuador Chequea y Ecuador Verifica.

```
plt.figure(figsize=(10,8))
plt.subplot (121)
plt.boxplot (Integrante2_Ecuador_Chequea["retweet_count"])
plt.ylabel ("Cantidad de retweets")
plt.title ("Usuario fact-checker de Ecuador Chequea y Ecuador Verifica")
plt.grid (True)

plt.subplot(122)
plt.boxplot (Usuario_comun["retweet_count"])
plt.title ("Usuario común")
plt.grid (True)
```

Figura 31. EDA – Sentencia de boxplot para la comparativa. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

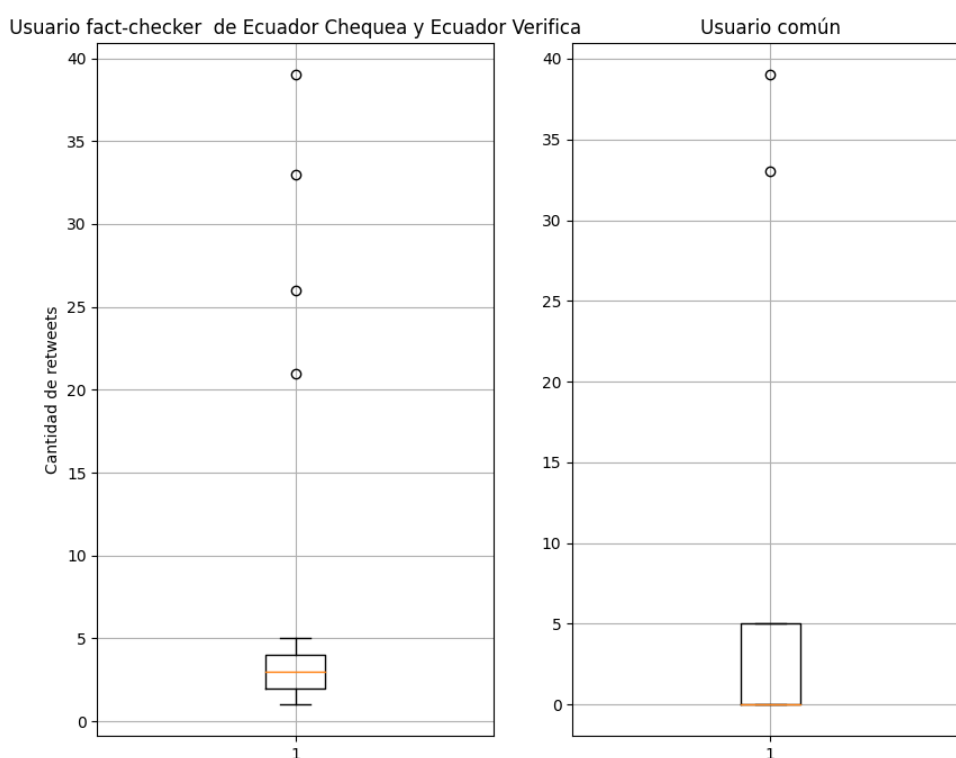


Figura 32. EDA – Boxplot comparativo de la cantidad de retweets entre un usuario fact-checker y un usuario común. Información adaptada de a (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Se procede a ejecutar la siguiente sentencia de código:

```
data_tweets.groupby(['tweet_id'])['favorite_count'].sum().sort_values(ascending
```

`=False).head(5)`, con el fin de visualizar el id del tweet que tiene la mayor cantidad de likes o favoritos, siendo el tweet_id “1676984953697968129”, con una cantidad de 109 likes.

```
data_tweets.groupby(['tweet_id'])['favorite_count'].sum().sort_values(ascending=False).head(5)
```

tweet_id	favorite_count
1676984953697968129	109
1676983580013019139	44
1676044892735254528	28
1676003412796661760	26
1676074764744679425	23

Name: favorite_count, dtype: int64

Figura 33. EDA – Tweet_id con mayor número de likes. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Por medio de un gráfico de pastel, también se puede visualizar el id del tweet con mayor cantidad de likes o marcado como favorito.

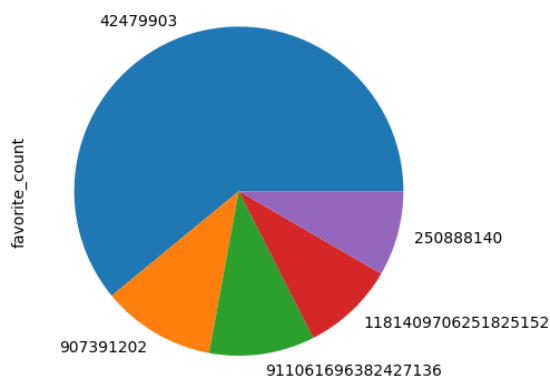


Figura 34. EDA – Gráfico Pastel Tweet_id con mayor número de likes. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Posteriormente, se visualiza el nombre de usuario y toda la información referente al tweet_id con mayor número de likes, en el que se puede identificar que el usuario que publicó este tweet se llama Ecuadorinmediato.

```
tweet_like = data_tweets [data_tweets["tweet_id"]== 1676984953697968129 ]
print (tweet_like)
```

	tweet_id	tweet_id_str \	tweet_text	user_id \	user_screen_name	user_name	user_location	user_verified \	created_at	lang	retweet_count	favorite_count \
275	1676984953697968129	1676984953697968129	!!#URGENTE\nJosé Villavicencio, presidente del...	42479903	ecuainm_oficial	Ecuadorinmediato	Quito - Ecuador	False	2023-07-06 16:02:24+00:00	es	69	109

Figura 35. EDA – Información Tweet_id con mayor número de likes. Información adaptada de a (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Para finalizar, se procede a visualizar los cinco primeros tweets que están el conjunto de datos.

```
data_tweets['tweet_text'].value_counts().head(5)
```

RT @ecuainm_oficial: !!#URGENTE\nJosé Villavicencio, presidente del FUT, anuncia movilizaciones si el Gobierno decide dar paso a las propues...

RT @ECUADORCHEQUEA: #AHORA | El presidente, @LassoGuillermo, firmará un Decreto Ley en beneficio de ciudadanos en coactiva por créditos edu...

RT @ECUADORCHEQUEA: ●#FALSO \n\nEl candidato presidencial Yaku Pérez no dijo que se detendrá inmediatamente toda explotación de recursos natu...

RT @ECUADORCHEQUEA: #ATENCIÓN I Tras la captura en Colombia de Luis Arboleda, alias 'Gordo Lucho', cabecilla la banda de delincuencia orga...

RT @ECUADORCHEQUEA: #URGENTE | La Policía Nacional informó que recapturó a alias el 'Gordo Luis', cabecilla de la banda "Los lobos". https...

Figura 36. EDA – Información Tweet_id con mayor número de likes. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Capítulo III

Desarrollo de la Propuesta

3.1. Selección de Variables

La etapa de selección de variables es necesaria en el desarrollo de cualquier análisis en ciencia de datos, en el presente trabajo de titulación se realiza esta fase con el objetivo de determinar las variables que tienen relevancia para el análisis de sentimientos de las noticias seleccionadas en Twitter por las verificadoras acreditadas en Ecuador utilizando Procesamiento de Lenguaje Natural. Es importante mencionar que, para una mejor comprensión al momento de analizar el conjunto de datos, se procedió a renombrar los atributos o variables, los mismos que se obtuvieron de la base de datos no relacional MongoDB Atlas, denominada “Twitter_principal”, en donde se encuentran almacenados los documentos que contienen la información de los objetos.

Para obtener los datos necesarios, se implementaron tres módulos en Python utilizando el cuaderno colaborativo Google Colab, el primero “conectar_mongodb_cloud”, para establecer la conexión a la base de datos, el segundo “obtener colección”, para obtener las colecciones tweets_f y media_f de la base de datos Twitter_principal, ya obtenidas las colecciones se procede a consultar los documentos que se encuentran en éstas, utilizando el módulo “consultar_documentos”. Es importante mencionar que, cada una de las colecciones consultadas se convierten y se muestran en un DataFrame, luego se concatenan en un solo DataFrame denominado df_unida, utilizando la sentencia de código `df_unida = pd.concat([df_tweets, df_media], axis=1)` y posteriormente se muestra con `df_unida.head()`.

ID	Fecha de Creación	ID del Autor	Texto	Retweets	Respuestas	Likes	Citas	Nombre de Usuario del Autor	Nombre del Autor	URL de la Imagen de Perfil del Autor	Fuente	_id	ID del Tweet
1685999109184901121	2023-07-31T13:01:26.000Z	1305529399215902723	 Carta abierta a los candidatos para las Elecciones	1	0	0	1	ecuadorverifica	Ecuador Verifica	https://pbs.twimg.com/profile_images/135136469...	None	64c2eb942d62b43e6228d74	1684671988961316965
1685990756796178432	2023-07-31T13:00:02.000Z	1305529399215902723	¿Es cierto que los presos de la Penitenciaría...	1	0	1	1	ecuadorverifica	Ecuador Verifica	https://pbs.twimg.com/profile_images/135136469...	None	64c2eb942d62b43e6228d75	1684665916271756227
1685822685412966880	2023-07-31T01:20:23.000Z	1305529399215902723	 La desinformación disminuye la confianza en...	0	0	1	0	ecuadorverifica	Ecuador Verifica	https://pbs.twimg.com/profile_images/135136469...	None	64c2eb942d62b43e6228d76	1684659386696765920
1685742463208570880	2023-07-30T20:01:37.000Z	1305529399215902723	#LoMásVisto La imagen fue editada para...	0	0	2	0	ecuadorverifica	Ecuador Verifica	https://pbs.twimg.com/profile_images/135136469...	None	64c2eb942d62b43e6228d77	1684628729400553472
1685621692771524608	2023-07-30T12:01:43.000Z	1305529399215902723	#LoMásVisto La creación de créditos es una p...	0	0	0	0	ecuadorverifica	Ecuador Verifica	https://pbs.twimg.com/profile_images/135136469...	None	64c2eb942d62b43e6228d78	1684620857648336896

Figura 37. Concatenación de DataFrames. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Al visualizar los atributos en el DataFrame df_unida se identifica que existen algunos que no son relevantes para el presente caso de estudio, por lo tanto, se procede a seleccionar las variables o atributos necesarios para el análisis de sentimientos.

A continuación, se determinan las variables seleccionadas:

- ID: En el conjunto de datos tratado, este campo hace referencia al tweet_id o el identificador único del texto del tweet.
- Fecha de Creación: Hace referencia al atributo created_at, es la fecha en la que fue creado el tweet o noticia.
- ID del Autor: Es el user_id o identificador único del usuario que ha publicado el tweet o noticia.
- Texto: Se refiere al texto del tweet o “noticia”, hace referencia al tweet_text.
- Nombre del Autor: Es el nombre del usuario, tal como se ha definido, no necesariamente el nombre de una persona hace referencia al user_name.
- Nombre de Usuario del Autor: El nombre de pantalla, identificador o alias con el que el usuario se identifica, hace referencia a user_screen_name.
- Tipo de media: Es el tipo de archivo media que contiene el tweet publicado.
- URL de la Media: El enlace para acceder al archivo media que contiene el tweet.

Con los campos seleccionados se crea un nuevo DataFrame con el nombre df_nuevo, utilizando la siguiente línea de código.

```
# Crear un nuevo DataFrame con las columnas seleccionadas
df_nuevo = df_unida[["ID", "ID del Tweet", "Fecha de Creación", "ID del Autor", "Texto",
                    "Nombre de Usuario del Autor", "Nombre del Autor",
                    "Tipo de Media", "URL de la Media" ]]
```

Figura 38. Creación del DataFrame df_nuevo con variables seleccionadas. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.1.1. Preprocesamiento del texto de las noticias.

En el presente trabajo, el análisis de sentimientos se ejecuta en el texto del tweet, es decir en la noticia, por lo tanto, es necesario realizar un preprocesamiento de los textos que se encuentran en el conjunto de datos previamente a la aplicación de los modelos.

El preprocesamiento del texto incluye una limpieza en los textos o tweets, para esto se ha aplicado una función denominada “limpiar_texto”, utilizando la librería “re” incorporada en Python para trabajar con las expresiones regulares que tiene el texto, esta limpieza consiste en convertir a cadena de texto, eliminar enlaces URL, eliminar menciones a usuarios, eliminar hashtags y convertir el texto a minúsculas, teniendo como resultado un texto limpio como se muestra en la imagen a continuación.

```
print(df_nuevo['Texto_Limpio'].head())
```

```
0  📄 carta abierta a los candidatos para las ele...
1  ¿es cierto que los presos de la penitenciaría ...
2  ❌ la disminuye la confianza en las institucio...
3  | la imagen fue editada para insinuar nexos e...
4  | la creación de créditos es una propuesta de...
Name: Texto_Limpio, dtype: object
```

Figura 39. Texto limpio de expresiones regulares. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Luego se eliminan las stopwords con el fin de disminuir el ruido en el texto utilizando Natural Language Toolkit (NLTK), para lo cual es necesario descargar las stopwords en idioma español utilizando la siguiente sentencia de código.

```
# Descargar las stopwords
nltk.download('stopwords')
stop_words = set(stopwords.words('spanish'))
```

Figura 40. Descarga de stopwords. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Para eliminar las stopwords se ha creado una función denominada “eliminar_stopwords”, que se aplica en la columna “Texto_Limpio” que se creó cuando se realizó la limpieza del texto y posteriormente se muestran los textos sin stopwords.

```
# Función para eliminar stopwords
def eliminar_stopwords(texto):
    palabras = texto.split()
    palabras_filtradas = [palabra for palabra in palabras if palabra.lower() not in stop_words]
    return " ".join(palabras_filtradas)

# Aplicar la función a la columna "Texto_Limpio" del DataFrame
df_nuevo["Texto_Limpio"] = df_nuevo["Texto_Limpio"].apply(eliminar_stopwords)
```

Figura 41. Eliminación de stopwords. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

```
# Mostrar algunos ejemplos después de eliminar stopwords
print("\nEjemplos después de eliminar stopwords:")
for i in range(5):
    print(df_nuevo["Texto_Limpio"].iloc[i])
```

```
Ejemplos después de eliminar stopwords:
📄 carta abierta candidatos elecciones adelantadas 2023 lee texto completo:
¿es cierto presos penitenciaría litoral piden vuelta rafael correa? información encuestras resumen semanal .
❌ la disminuye confianza instituciones menoscaba democracia países. objetivo combatirla. ¿quieren conocer cómo hacemos? contamos📄
| imagen editada insinuar nexos políticos. presidente guillermo lasso negado apoyar algún candidato varias ocasiones. imagen corresponde acto territorio protagonizado yaku Pérez. 📄
| creación créditos propuesta sonnenholzner. propuesta campaña, plantea creación crédito jóvenes. embargo, logos bancos forman parte publicidad oficial candidato. 📄
```

Figura 42. Textos sin stopwords. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Algo importante de tratar cuando se está realizando un análisis de sentimientos son los emoticones, ya que pueden contener información emocional e incidir en la interpretación

del texto, para tratarlos se instala la biblioteca emoji. En primer lugar, se verifica si los textos o noticias contienen emoticones, para esto se crea la función “contiene_emoticones”, se aplica a la columna “Texto_Limpio” y se crea una nueva columna booleana en el DataFrame llamada “Contiene Emoticones”, la que indica si el texto contiene o no emoticones.

```
def contiene_emoticones(texto):
    # Contar el número de emoticones en el texto
    count = emoji.emoji_count(texto)

    # Verificar si el texto tiene emoticones
    if count > 0:
        return True
    else:
        return False

# Aplicar la función contiene_emoticones a la columna 'Texto_Limpio'
df_nuevo['Contiene Emoticones'] = df_nuevo['Texto_Limpio'].apply(contiene_emoticones)
```

Figura 43. Función para verificar existencia de emoticones. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.



	Texto_Limpio	Contiene Emoticones
0	 carta abierta candidatos elecciones adelant...	True
1	¿es cierto presos penitenciaría litoral piden ...	False
2	 la disminuye confianza instituciones menosca...	True
3	imagen editada insinuar nexos políticos. pre...	True
4	creación créditos propuesta sonnenholzner. p...	True

Figura 44. Verificación de existencia de emoticones. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Posteriormente, se corrobora si existen emoticones que hagan referencia a las emociones que se pretenden identificar en el análisis de sentimientos, que son: alegría, esperanza, miedo, enfado, tristeza, para esto se ha creado una función denominada “detectar_emoticones_sentimientos” que se aplica en la columna “Texto_Limpio” y se crea una nueva columna en el DataFrame llamada “Emoticones_Detectados” en donde se verifica la existencia de emoticones referentes a los sentimientos anteriormente mencionados.

```
# Mostrar algunos ejemplos de la columna "Emoticones_Detectados"
(df_nuevo[["Texto_Limpio", "Emoticones_Detectados"]].head(5))
```



	Texto_Limpio	Emoticones_Detectados
0	 carta abierta candidatos elecciones adelant...	[]
1	¿es cierto presos penitenciaria litoral piden ...	[]
2	 la disminuye confianza instituciones menosca...	[]
3	imagen editada insinuar nexos políticos. pre...	[]
4	creación créditos propuesta sonnenholzner. p...	[]

Figura 45. Verificación de existencia de emoticones referentes a sentimientos tratados. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Al verificar algunos ejemplos de texto con emoticones detectados, haciendo referencia a los emoticones de esperanza, alegría, tristeza, enfado, miedo, entre otros, se puede identificar que no hay emoticones de esta clase en el conjunto de datos utilizados. De todas maneras, para corroborar que la función está ejecutándose de manera correcta, se procede a aplicarla en un texto de ejemplo.

```
texto_ejemplo = "Estoy 😊 y 😞 al mismo tiempo, 😊 y algo 😞."
emoticones_detectados = detectar_emoticones_sentimientos(texto_ejemplo)
print(emoticones_detectados)

['Felicidad', 'Tristeza', 'Enojo', 'Satira']
```

Figura 46. Texto ejemplo aplicando la función `detectar_emoticones_sentimientos`. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Es importante considerar los emojis de pulgar arriba y pulgar abajo, porque podrían interpretarse como algo negativo o positivo en el texto, por lo tanto, se aplica una función denominada “eliminar_emoticones_no_deseados” en la que se indica los emojis o emoticones que se desean mantener en el texto, y se eliminan los emojis que no son relevantes para el análisis de sentimientos, luego esta función se aplica a la columna “Texto_Limpio” y se verifican los cambios.

```
# Aplicar la función a la columna "Texto_Limpio" y guardar los resultados en la columna "Texto_Limpio"
df_nuevo["Texto_Limpio"] = df_nuevo["Texto_Limpio"].apply(eliminar_emoticones_no_deseados)
df_nuevo["Texto_Limpio"]

0      carta abierta candidatos elecciones adelantada...
1      es cierto presos penitenciaria litoral piden v...
2      la disminuye confianza instituciones menoscaba...
3      imagen editada insinuar nexos politicos presi...
4      creación créditos propuesta sonnenholzner pro...
...
123     ecuador luto, afirma presidente república, , t...
124     gobierno decreta excepción cantón durán guaya...
125     fin semana registraron varios incidentes penit...
126     twitter cambió icónico logo pajarito azul x ca...
127     entidad asegura guías penitenciarios retenidos...
Name: Texto_Limpio, Length: 128, dtype: object
```

Figura 47. Texto sin emoticones. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Después de haber tratado los emoticones, se procede a limpiar los caracteres especiales, tales como puntos, coma, etc. Se lo ejecuta después de todo lo anteriormente realizado porque es importante recalcar que entre los patrones de los emoticones se tienen patrones tales como :) :(, entre otros que podrían contener información relevante, para proceder con la limpieza de estos caracteres, se aplica la siguiente función.

```
def limpiar_caracteres_especiales(texto):
    # Eliminar caracteres especiales y puntuación
    texto_limpio = re.sub(r'^\w\s', '', texto)
    return texto_limpio
df_nuevo["Texto_Limpio"] = df_nuevo["Texto_Limpio"].apply(limpiar_caracteres_especiales)
```

Figura 48. Función para limpiar caracteres especiales en los textos. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Se corrobora que el texto esté limpio en su totalidad tomando un texto aleatorio para proceder a efectuar la aplicación de los modelos.

```
texto_limpiof = df_nuevo["Texto_Limpio"].iloc[i]
print(texto_limpiof)

creación créditos propuesta sonnenholzner propuesta campaña plantea creación crédito jóvenes embargo logos bancos forman parte publicidad oficial candidato
```

Figura 49. Texto sin caracteres especiales. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2. Modelos Aplicados

En el presente trabajo se utilizaron modelos preentrenados BERT de la librería Transformers de Hugging Face para corroborar la efectividad que tienen al realizar el análisis de sentimientos enfocado a las noticias seleccionadas por las verificadoras de hecho en el Ecuador.

Para seleccionar posteriormente el modelo base, se hizo una comparación entre los siguientes modelos preentrenados.

- Modelo Preentrenado BERT en español “uncased”.

- Modelo Preentrenado BERT multilingual o multilingües “cased”.

Al ejecutar el análisis de sentimientos en los textos de las noticias con el modelo Preentrenado BERT multilingual o multilingües “cased”, utilizando Bert Tokenizer para tokenizar los textos, determinando un padding con una longitud máxima de 55, verificando previamente que el texto más largo tiene 52 tokens y codificando los textos, se pudo visualizar que los resultados al realizar el análisis de sentimientos en los tweets o noticias no eran tan efectivos, como se puede observar en la siguiente imagen, en donde 1 es un sentimiento positivo y 0 un sentimiento negativo.

```
df_nuevo[["Texto", "Binaria_categoria_multilingual"]].head(5)
```



	Texto	Binaria_categoria_multilingual
0	 Carta abierta a los candidatos para las Ele...	1
1	¿Es cierto que los presos de la Penitenciaría ...	0
2	 La #desinformación disminuye la confianza en...	1
3	#LoMásVisto La imagen fue editada para insin...	0
4	#LoMásVisto La creación de créditos es una p...	1

Figura 50. Análisis de sentimientos con Bert en español – categoría binaria multilingual. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Para una mejor comprensión, se crea un nuevo campo en el DataFrame denominado “Categoria_sentimiento_multilingual”, en el que indica si el sentimiento identificado en el texto o noticia es negativo o positivo, como se muestra en la figura a continuación.

```
df_nuevo[["Texto", "Categoria_sentimiento_multilingual"]].head(5)
```



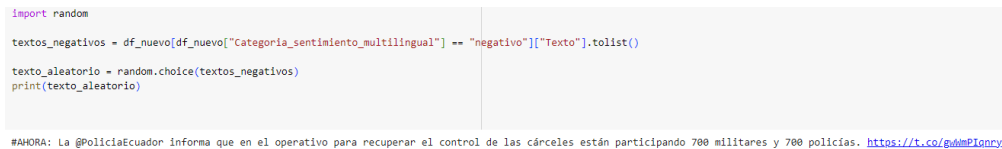
	Texto	Categoria_sentimiento_multilingual
0	 Carta abierta a los candidatos para las Ele...	positivo
1	¿Es cierto que los presos de la Penitenciaría ...	negativo
2	 La #desinformación disminuye la confianza en...	positivo
3	#LoMásVisto La imagen fue editada para insin...	negativo
4	#LoMásVisto La creación de créditos es una p...	positivo

Figura 51. Análisis de sentimientos con Bert en español – categoría sentimiento multilingual. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

El modelo Preentrenado de Bert no brinda resultados óptimos para el análisis de sentimientos en este caso de estudio, por ejemplo, en la fila número 2 de la columna Texto que se visualiza en la figura 51, trata de desinformación y la disminución de confianza, lo que se podría interpretar como algo negativo, sin embargo, el modelo lo considera positivo. Para corroborar esto, se ha seleccionado una noticia de manera aleatoria en la que el modelo

ha identificado un sentimiento negativo, cuando el contexto de la noticia podría indicar que tiene un sentimiento positivo.



```
import random

textos_negativos = df_nuevo[df_nuevo["Categoria_sentimiento_multilingual"] == "negativo"]["Texto"].tolist()

texto_aleatorio = random.choice(textos_negativos)
print(texto_aleatorio)
```

#AHORA: La @PoliciaEcuador informa que en el operativo para recuperar el control de las cárceles están participando 700 militares y 700 policías. <https://t.co/gwMPlagry>

Figura 52. Texto identificado con sentimiento negativo con Bert multilingual. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

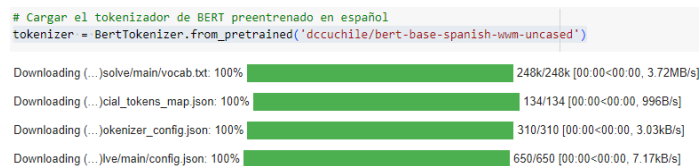
3.2.1. Modelo Base.

Al realizar pruebas con el modelo preentrenado BERT en español “uncased”, a diferencia del modelo preentrenado BERT multilingual o multilingües “cased”, se obtuvieron resultados favorables respecto a los sentimientos identificados en los textos de las noticias en Twitter de los verificadores de hecho Ecuador Chequea y Ecuador Verifica, por lo tanto, se ha seleccionado este algoritmo como modelo base para el presente proyecto.

Previo a la aplicación del modelo preentrenado BERT en español “uncased”, es necesario realizar una tokenización en los textos que se encuentran limpios en el DataFrame `df_nuevo`, posteriormente, se debe padear y codificar estos textos con el objetivo de que se encuentren listos para aplicar el modelo.

3.2.1.1. Tokenización en el modelo base.

Para realizar la tokenización en los textos se utiliza Bert Tokenizer con el modelo preentrenado en español, utilizando la sentencia de código que se visualiza en la figura 53.



```
# Cargar el tokenizador de BERT preentrenado en español
tokenizer = BertTokenizer.from_pretrained('dccuchile/bert-base-spanish-wwm-uncased')
```

Downloading (...)solve/main/vocab.txt: 100% 248k/248k [00:00<00:00, 3.72MB/s]
 Downloading (...)cial_tokens_map.json: 100% 134/134 [00:00<00:00, 996B/s]
 Downloading (...)tokenizer_config.json: 100% 310/310 [00:00<00:00, 3.03kB/s]
 Downloading (...)lve/main/config.json: 100% 650/650 [00:00<00:00, 7.17kB/s]

Figura 53. Carga de BertTokenizer en español – Modelo Base. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Se ha creado una función para tokenizar los textos, la que se ha aplicado en el campo “Texto_Limpio” del DataFrame “`df_nuevo`”, creando también un campo nuevo denominado “`Texto_Tokenizado_bert_spanish`”, en donde se guardan los textos divididos en tokens.

```
# Función para tokenizar el texto sin convertir en cadena
def tokenizar_texto(texto):
    return tokenizer.tokenize(texto)

# Aplicar el tokenizador a la columna "Texto_Limpio"
df_nuevo["Texto_Tokenizado_bert_spanish"] = df_nuevo["Texto_Limpio"].apply(tokenizar_texto)
```

Figura 54. Función para tokenizar texto con Bert Tokenizer – Modelo Base. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

De manera aleatoria se selecciona un texto para corroborar que el texto se haya dividido en tokens, como se observa en la figura 55, el tokenizador está fragmentando algunas palabras en varios tokens, esto lo realiza con el fin de permitir una mejor representación de palabras desconocidas o complejas que pueden aparecer en el corpus del modelo BERT. Es importante mencionar que los prefijos "##" se utilizan para indicar que el token es parte de una palabra más larga y que necesita combinarse con otros tokens para obtener la palabra completa.

```
# Acceder al texto tokenizado de cualquier registro en la columna "Texto_Tokenizado_bert_spanish"
texto_tokenizado = df_nuevo["Texto_Tokenizado_bert_spanish"].iloc[i]

# Imprimir los tokens resultantes
print(texto_tokenizado)

['creación', 'créditos', 'propuesta', 'son', '##nen', '##hol', '##z', '##ner', 'propuesta', 'campaña', 'plantea', 'creación', 'crédito', 'jóvenes']
```

Figura 55. Texto tokenizado con Bert Tokenizer – Modelo Base. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.1.2. Padding en el modelo base.

El padding se realiza con el fin de asegurar que todas las secuencias de entrada tengan la misma longitud y sean compatibles con el modelo, esto permite un preprocesamiento en lotes, de esta manera la manipulación de secuencias más largas es un poco más sencilla. Antes de padear y codificar, es recomendable verificar cuántos tokens tiene el texto más largo que se está analizando para determinar la longitud máxima en el padding.

```
Texto con más tokens:
['sna', '##i', 'informó', 'logrado', 'liberación', '10', '##6', 'guías', 'penitenciario', '##s', 'reten'
Número de tokens: 38
```

Figura 56. Texto con más tokens – Modelo Base. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Se identifica que el número de tokens del texto más largo del conjunto de datos tratado es de 38, por lo que se determinará una longitud máxima del padding de 40, posteriormente, se aplica el padding a la columna "Texto_Tokenizado_bert_spanish", creando una nueva columna "Texto_Pad_spanish" en donde se guardarán los textos padeados.


```
# Definir la longitud máxima para el padding
max_length = 40

# Obtener los textos padding de la columna "Texto_Pad_spanish" y convertirlos en una lista
textos_padding = df_nuevo["Texto_Pad_spanish"].tolist()

# Aplicar la función de codificación para obtener los textos codificados
textos_codificados = pad_and_encode_texto_bert_spanish(textos_padding, max_length)

# Crear una nueva columna "Texto_Codificado" en el DataFrame con los textos codificados
df_nuevo["Texto_Codificado_spanish"] = textos_codificados
```

Figura 60. Codificación de los textos – Modelo Base. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

A continuación, se muestran los textos codificados de la columna "Texto_Codificado_spanish" del DataFrame df_nuevo.

```
df_nuevo["Texto_Codificado_spanish"]

0      [4, 3350, 6051, 7962, 4552, 5703, 3166, 22268,...
1      [4, 1028, 2551, 12706, 27381, 2706, 14296, 161...
2      [4, 1032, 18753, 5264, 3673, 29467, 1525, 6537...
3      [4, 4620, 29184, 19568, 1020, 1504, 23727, 456...
4      [4, 3553, 7701, 3498, 1318, 20587, 19056, 3098...
...
123     [4, 8225, 26531, 6631, 2078, 2542, 1894, 6196...
124     [4, 2022, 5955, 1047, 7294, 11896, 3926, 1197...
125     [4, 1346, 2859, 21682, 2909, 14107, 27381, 270...
126     [4, 1027, 1004, 16805, 8143, 8014, 8772, 16265...
127     [4, 8156, 12207, 18482, 24618, 30958, 13643, 1...
Name: Texto_Codificado_spanish, Length: 128, dtype: object
```

Figura 61. Visualización de los textos codificados – Modelo Base. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.1.4. Análisis de sentimientos con el Modelo Preentrenado de BERT en español.

Teniendo los textos codificados, se procede a realizar el análisis de sentimientos en las noticias seleccionadas por los verificadores de hecho a nivel nacional, Ecuador Chequea y Ecuador Verifica, utilizando como modelo base el modelo preentrenado BERT en español “uncased”.

Es importante mencionar que este modelo ya está preentrenado por defecto pero no se lo ha entrenado ajustando hiperparámetros y evaluado con las métricas correspondientes, es un modelo base para posteriormente proceder con el entrenamiento que satisfaga los requerimientos para el análisis de sentimientos del presente caso de estudio, para esto se crea una función denominada “análisis_sentimientos_binario_bert” a la que se le envía como parámetro el texto que se encuentra codificado, es importante mencionar que previo a la creación de la función, es necesario importar la librería TensorFlow, ya que se utilizará en la función.

```
def analisis_sentimientos_binario_bert(texto_codificado):
    # Cargar el tokenizer y el modelo preentrenado de BERT en español
    tokenizer = BertTokenizer.from_pretrained('dccuchile/bert-base-spanish-wwm-uncased')
    model = BertForSequenceClassification.from_pretrained('dccuchile/bert-base-spanish-wwm-uncased')

    # Convertir los enteros en tokens
    tokens = tokenizer.convert_ids_to_tokens(texto_codificado)

    # Convertir los tokens en texto
    texto = tokenizer.convert_tokens_to_string(tokens)

    # Preparar los datos (codificar con padding)
    inputs = tokenizer(texto, return_tensors='pt', padding=True, truncation=True)

    # Realizar predicciones
    outputs = model(**inputs)
    logits = outputs.logits
    predicted_labels = torch.argmax(logits, dim=1)

    # Obtener el resultado de la predicción (por ejemplo, 1 para positivo, 0 para negativo)
    predicted_sentiment = predicted_labels.item()

    return predicted_sentiment
```

Figura 62. Función de análisis de sentimientos con Bert – Modelo Base. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Esta función se aplica a la columna en donde se encuentran los textos codificados, en este caso se llama “Texto_Codificado_spanish” y se crea una nueva columna denominada “Binaria_categoria_spanish”.

```
df_nuevo[["Texto", "Binaria_categoria_spanish"]].head(5)
```




	Texto	Binaria_categoria_spanish
0	 Carta abierta a los candidatos para las Ele...	0
1	¿Es cierto que los presos de la Penitenciaría ...	0
2	  La #desinformación disminuye la confianza en...	0
3	#LoMásVisto La imagen fue editada para insin...	0
4	#LoMásVisto La creación de créditos es una p...	1

Figura 63. Análisis de sentimientos con Bert en español categoría binaria – Modelo Base. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Para una mejor comprensión, se crea un nuevo campo en el DataFrame denominado “Categoria_spanish”, en el que indica si el sentimiento identificado en el texto o noticia es negativo o positivo, como se muestra en la figura a continuación.

```
df_nuevo[["Texto", "Categoria_spanish"]].head(5)
```




	Texto	Categoria_spanish
0	 Carta abierta a los candidatos para las Ele...	negativo
1	¿Es cierto que los presos de la Penitenciaría ...	negativo
2	 La #desinformación disminuye la confianza en...	negativo
3	#LoMásVisto La imagen fue editada para insin...	negativo
4	#LoMásVisto La creación de créditos es una p...	positivo

Figura 64. Análisis de sentimientos con Bert en español categoría spanish. – Modelo Base Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Para seguir corroborando que este modelo puede determinar de una manera más efectiva los sentimientos que se encuentran inmersos en las noticias de los verificadores de hecho a nivel nacional, se visualiza un texto en el que se ha identificado un sentimiento negativo.

```
texto_sentimiento_negativo = df_nuevo[df_nuevo["Categoria_spanish"] == "negativo"]["Texto"].iloc[0]

# Imprimir ltexto_sentimiento_negativo
print(texto_sentimiento_negativo)
```

 La #desinformación disminuye la confianza en las instituciones y menoscaba la democracia de los países.

En #EcuadorVerifica nuestro objetivo es combatirla. ¿Quieres conocer cómo lo hacemos?


Te contamos  <https://t.co/aAtqyIHge>

Figura 65. Texto con sentimiento negativo con Bert en español – Modelo Base. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.2. Modelo Propuesto

El modelo propuesto se basa en el modelo preentrenado de BERT en español de la librería Transformers de Hugging Face y PyTorch, en este modelo se propone identificar y etiquetar sentimientos en negativos y positivos, y etiquetar emociones como alegría, esperanza, tristeza, enfado y miedo.

Para proceder con el entrenamiento del modelo propuesto, se ha creado un subconjunto del conjunto de datos “df_nuevo”, el cual se denomina “subconjunto” que contiene una muestra aleatoria de 100 datos, esto se realiza con el objetivo de etiquetarlos con los sentimientos y emociones correspondientes para luego entrenar el modelo con estos datos. Este pequeño DataFrame contiene la variable “Texto_Limpio” del conjunto de datos más grande “df_nuevo”, en el mismo que ya se encuentran los textos de las noticias preprocesados.

subconjunto.head(2)

	ID	ID del Tweet	Fecha de Creación	ID del Autor	Texto	Nombre de Usuario del Autor	Nombre del Autor	Tipo de Medios	URL de la Medios	Texto_Limpio	Contiene Emoticones	Emoticones_Detectados	Sentimiento	Emoción
115	1683550436538589192	1685093191182172160	2023-07-24T18:51:17.000Z	777891400297897985	El fin de semana se registraron varios incidentes.	ECUADORCHEQUEA	Ecuador Chequea	photo	https://pbs.twimg.com/media/F2X0uRgWIAAAQJN.jpg	En semana registraron varios incidentes pent.	True	☹️	negativo	entado
120	168391285279551295	1683824789089054211	2023-07-25T18:51:24.000Z	777891400297897985	La @PoliciaEcuador ha confirmado que tras rec...	ECUADORCHEQUEA	Ecuador Chequea	photo	https://pbs.twimg.com/media/F1485QVIAAARCAF.jpg	confirmado que tras recibir alerta disturbios...	False	☹️	negativo	meado

Figura 66. Subconjunto de datos etiquetados. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.3. Fase de Entrenamiento.

Para el entrenamiento del modelo, es necesario preparar el conjunto de datos, por esta razón, previamente se creó un conjunto de datos etiquetados para poder entrenar el modelo propuesto.

3.2.3.1. Tokenización.

La tokenización es fundamental en el Procesamiento de Lenguaje Natural (NLP), que tiene como fin convertir los textos en secuencias numéricas, conocidas como tokens, para que puedan ser entendidas en el modelo de aprendizaje.

En primer lugar, es necesario cargar el tokenizador preentrenado de BERT en español, utilizando la sentencia de código `tokenizer = BertTokenizer.from_pretrained("dccuchile/bert-base-spanish-wwm-uncased")`, para proceder a tokenizar los textos como se visualiza en la figura 67.

```
# Tokenizar los textos
input_texts = subconjunto["Texto_Limpio"].tolist()
sentimiento_labels = subconjunto["Sentimiento"].tolist()
emocion_labels = subconjunto["Emocion"].tolist()
```

Figura 67. Tokenización de textos en Modelo Propuesto. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.3.2. Padding y Codificación.

El proceso de codificación de los textos se realiza porque permite convertir los tokens en índices numéricos, para que puedan ser comprendidos por el modelo. Esto conlleva a crear un diccionario llamado “encodings” que contiene los tokens de los textos, con opción a truncado en caso de que los textos superen la longitud máxima permitida que es de 40 tokens, respecto a esta longitud, es recomendable que se corrobore previamente el número de tokens del texto más largo del conjunto de datos.

Además, se utiliza el parámetro padding, lo que indica que, si existen textos que sean más cortos que la longitud máxima definida, se deban agregar tokens de relleno a estos textos.

```
# Codificar los textos
encodings = tokenizer(input_texts, truncation=True, padding=True, max_length=40)
```

Figura 68. Codificación de textos en Modelo Propuesto. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Del mismo modo, las etiquetas referentes a los sentimientos y emociones se deben convertir en índices numéricos, los que se guardan en dos listas “sentimiento_label_ids” y “emocion_label_ids”, se debe realizar este proceso porque estos índices se utilizarán después para entrenar y evaluar el modelo de aprendizaje automático.

```
# Convertir las etiquetas a números
sentimiento_label2id = {label: idx for idx, label in enumerate(set(sentimiento_labels))}
emocion_label2id = {label: idx for idx, label in enumerate(set(emocion_labels))}
sentimiento_label_ids = [sentimiento_label2id[label] for label in sentimiento_labels]
emocion_label_ids = [emocion_label2id[label] for label in emocion_labels]
```

Figura 69. Conversión de etiquetas a números. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.3.3. Creación de tensores de PyTorch.

Por otra parte, se deben convertir las variables resultantes del paso anterior en tensores de PyTorch para sustentar al modelo durante la etapa de entrenamiento y la evaluación. Las entradas input_ids y attention_mask hacen referencia a las secuencias tokenizadas y sus máscaras de atención, a diferencia de las etiquetas de sentimiento y emoción que se utilizan como objetivos para calcular la pérdida y medir el rendimiento del modelo.

```
# Crear tensores de PyTorch a partir de los encodings
input_ids = torch.tensor(encodings["input_ids"])
attention_mask = torch.tensor(encodings["attention_mask"])
sentimiento_labels = torch.tensor(sentimiento_label_ids)
emocion_labels = torch.tensor(emocion_label_ids)
```

Figura 70. Creación de tensores de PyTorch. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.3.4. División de datos en conjuntos de entrenamiento y prueba.

Para dividir los datos en conjuntos de entrenamiento y prueba se ha utilizado la función “train_test_split” de Scikit-learn, en este sentido la división se realiza en las secuencias tokenizadas y sus máscaras de atención en las etiquetas de sentimiento y emoción.

Para determinar la proporción del conjunto de datos para prueba, se ha utilizado el parámetro “test_size” con un valor del 20% de los datos, con una aleatoriedad de la división de 42, siendo un valor fijo y garantizando que la división sea reproducible, de esta manera se obtendrán los mismos resultados en cada ejecución.

Los datos de entrenamiento se usarán para ajustar el modelo, y los datos de prueba se utilizarán para evaluar su rendimiento. Esta división es fundamental para evaluar la capacidad de generalización del modelo a datos no vistos durante el entrenamiento.

```
# Dividir los datos en conjuntos de entrenamiento y prueba
train_input_ids, test_input_ids, train_attention_mask, test_attention_mask, train_sentimiento_labels, test_sentimiento_labels, train_emocion_labels, test_emocion_labels = train_test_split(
    input_ids, attention_mask, sentimiento_labels, emocion_labels, test_size=0.2, random_state=42
)
```

Figura 71. División en conjunto de entrenamiento y prueba. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.3.5. Creación de conjuntos de datos de entrenamiento y prueba.

Estos conjuntos de datos se crean en formato PyTorch haciendo uso de la clase `TensorDataset` de la librería “`torch.utils.data`”, se crea un conjunto de datos de entrenamiento denominado “`train_dataset`” con cuatro tipo de datos, “`train_input_ids`” que se refiere a las secuencias tokenizadas, “`train_attention_mask`” que son los máscaras de atención que corresponden a las secuencias, “`train_sentimiento_labels`” las etiquetas de sentimiento para las entradas y “`train_emocion_labels`” que respecta a las etiquetas de emoción para las entradas.

Por otro lado, se crea el conjunto de datos de prueba llamado “`test_dataset`”, en el que de manera similar al conjunto de datos de entrenamiento se tienen los mismos tipos de datos, pero para prueba.

```
# Crear conjuntos de datos
train_dataset = TensorDataset(train_input_ids, train_attention_mask, train_sentimiento_labels, train_emocion_labels)
test_dataset = TensorDataset(test_input_ids, test_attention_mask, test_sentimiento_labels, test_emocion_labels)
```

Figura 72. Creación de conjuntos de datos de entrenamiento y prueba. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.3.6. Definición del Modelo Preentrenado de BERT.

Después de la creación de los conjuntos de datos de entrenamiento y prueba, es indispensable cargar el modelo preentrenado de BERT para el análisis de sentimientos y emociones utilizando la arquitectura “`bert-base-spanish-wwm-uncased`”.

Por consiguiente, se crea una instancia del modelo preentrenado “`BertForSequenceClassification`”, utilizando la función “`from_pretrained`” con el fin de cargar los pesos preentrenados del modelo utilizado.

```
# Definir el modelo preentrenado de BERT para análisis de sentimientos y emociones
model = BertForSequenceClassification.from_pretrained("dccuchile/bert-base-spanish-wwm-uncased", num_labels=len(sentimiento_label2id))
```


Downloading pytorch_model.bin: 100%  440M/440M [00:20<00:00, 22.6MB/s]

Figura 73. Definición del modelo preentrenado de BERT. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.4. Selección de parámetros e hiperparámetros.

La selección de parámetros e hiperparámetros en un modelo de aprendizaje automático con Procesamiento de Lenguaje Natural (PLN), es relevante porque controlan algunos aspectos del proceso de entrenamiento y la inferencia dentro del modelo, es por esta razón, que es necesario realizar una elección adecuada para obtener resultados óptimos en la tarea del análisis de sentimientos.

3.2.4.1. Hiperparámetros del modelo.

Los hiperparámetros se establecen antes de comenzar el proceso de entrenamiento del modelo, es decir que son valores que no son aprendidos directamente del conjunto de datos como los parámetros, en este caso se han implementado dos hiperparámetros importantes, el “batch_size” que representa el tamaño del lote, esto se refiere al número de ejemplos de entrenamiento que se utilizarán en cada interacción durante el entrenamiento, en este caso se utilizará un tamaño de lote de 20.

En el mismo tema de hiperparámetros, se tiene el “num_epochs”, es decir el número de veces que el modelo recorrerá todo el conjunto de datos de entrenamiento mientras se esté entrenando, en este caso se trabajará con 5 épocas, es importante definir este hiperparámetro porque controla la cantidad de veces que el modelo ajusta sus pesos en función de los datos de entrenamiento.

```
# Definir hiperparámetros de entrenamiento
batch_size = 20 # Tamaño del lote
num_epochs = 5 # Número de epochs o épocas que se va a repetir el proceso de entrenamiento
```

Figura 74. Hiperparámetros de entrenamiento. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.4.2. Creación de cargadores de datos.

Es importante crear los cargadores de datos para el conjunto de entrenamiento y el conjunto de prueba, ya que forman parte del flujo de trabajo en PyTorch.

Los cargadores de datos van a permitir iterar de una forma eficiente sobre los datos de entrenamiento y prueba en lotes y esto es fundamental para entrenar el modelo y por consiguiente las evaluaciones.

```
# Crear cargadores de datos
train_loader = DataLoader(train_dataset, batch_size=batch_size, shuffle=True)
test_loader = DataLoader(test_dataset, batch_size=batch_size, shuffle=False)
```

Figura 75. Creación de cargadores de datos. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.4.3. Definición de la función de pérdida y optimizador.

Para definir la función de pérdida o criterio, se ha utilizado la función “CrossEntropyLoss”, que se utiliza regularmente en los problemas de clasificación, en el presente caso, se utiliza para medir la discrepancia entre las predicciones del modelo y las etiquetas verdaderas conocidas como “ground truth”, con esto se puede guiar el proceso de ajuste de los parámetros del modelo.

En el presente entrenamiento, se ha utilizado el optimizador AdamW que pretende evitar el sobreajuste, además se utiliza el “model.parameters()” que indica cuáles son los parámetros que deben ser actualizados durante el entrenamiento, determinando la tasa de aprendizaje para poder controlar cuánto se ajusta los parámetros en función de la magnitud del gradiente, de esta manera se afina el rendimiento del modelo.

```
# Definir función de pérdida y optimizador
criterion = torch.nn.CrossEntropyLoss()
optimizer = torch.optim.AdamW(model.parameters(), lr=2e-5)
```

Figura 76. Función de pérdida y optimizador. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.2.4.4. Entrenamiento del modelo.

El proceso de entrenamiento del modelo comienza con la iteración a través de las épocas o epochs, es decir que se repite el proceso de entrenamiento por el número determinado de épocas, luego se debe poner el modelo en modo de entrenamiento, con el fin de que se habiliten capas, tales como: dropout y batch normalization para que actúen acorde al entrenamiento.

En este proceso se itera por medio de los lotes de datos de entrenamiento, posteriormente se procede con la optimización, se restablecen los gradientes del optimizador a cero para evitar acumulación en iteraciones previas, para las salidas se pasa un lote de datos de entrenamiento al modelo y se obtienen las predicciones, luego se calcula la pérdida entre las predicciones del modelo y las etiquetas verdaderas para ese lote. Además, se calculan los gradientes de la pérdida en contexto de parámetros del modelo y se actualizan los parámetros del modelo.

```

# Entrenar el modelo
for epoch in range(num_epochs):
    model.train()
    total_loss = 0
    for batch in train_loader:
        input_ids, attention_mask, sentimiento_labels, emocion_labels = batch
        optimizer.zero_grad()
        outputs = model(input_ids, attention_mask=attention_mask, labels=sentimiento_labels)
        loss = outputs.loss
        loss.backward()
        optimizer.step()
        total_loss += loss.item()
    print(f"Epoch {epoch+1}, Loss: {total_loss / len(train_loader)}")

```

Figura 77. Entrenamiento del modelo. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Para finalizar el proceso de entrenamiento, se acumula la pérdida total en la variable “total_loss” y al final de cada época se imprime el promedio de la pérdida a lo largo de todos los lotes de entrenamiento. Como se puede observar en la figura 78, al completar la quinta época de entrenamiento se obtiene como resultado una pérdida promedio de $2.3856729285398615 \times 10^{-6}$, una pérdida diminuta que se encuentra establecida en notación científica, donde “e-06” significa que el número es multiplicado por 10 elevado a la potencia de -6, lo que lo hace muy cercano a cero.

```

Epoch 1, Loss: 1.1531936934261466e-05
Epoch 2, Loss: 7.069084063004993e-06
Epoch 3, Loss: 4.798161739927309e-06
Epoch 4, Loss: 3.36765646125059e-06
Epoch 5, Loss: 2.3856729285398615e-06

```

Figura 78. Pérdida promedio del modelo. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly

Por esta razón, se puede corroborar que el modelo se está ajustando de forma precisa a los datos de entrenamiento y está logrando relacionarse de manera correcta con las etiquetas verdaderas, cabe recalcar que para asegurarse que el modelo tiene un buen rendimiento es importante evaluarlo y revisar sus métricas, tales como precisión, entre otras.

3.3. Evaluación

La evaluación es una fase que se debe realizar después de entrenar un modelo, esto se realiza con el objetivo de verificar el rendimiento real del modelo y corroborar si el modelo generaliza bien a nuevos ejemplos.

Para evaluar el modelo entrenado, en primer lugar, se debe colocar el modelo en modo evaluación para que no realice cambios en los pesos durante la evaluación, luego se utilizan

las listas “sentimiento_predicted_labels” y “emocion_predicted_labels” para almacenar las etiquetas predichas por el modelo, y para almacenar las etiquetas verdaderas se utilizan las listas “true_sentimiento_labels” y “true_emocion_labels”. Es importante mantener un seguimiento del número total de predicciones correctas para los sentimientos y emociones y llevar un registro del número total de ejemplos en el conjunto prueba.

Al finalizar la evaluación, se tiene el número total de predicciones correctas para sentimientos y emociones, de la misma forma se tiene el número total de ejemplos en el conjunto de prueba.

```
# Evaluar el modelo
model.eval()
sentimiento_predicted_labels = []
emocion_predicted_labels = []
true_sentimiento_labels = []
true_emocion_labels = []
correct_sentimiento = 0
correct_emocion = 0
total = 0
with torch.no_grad():
    for batch in test_loader:
        input_ids, attention_mask, sentimiento_labels, emocion_labels = batch

        # Predicciones para sentimientos
        sentimiento_outputs = model(input_ids, attention_mask=attention_mask)
        sentimiento_predicted_batch_labels = torch.argmax(sentimiento_outputs.logits, dim=1)
        sentimiento_predicted_labels.extend(sentimiento_predicted_batch_labels.cpu().numpy())
        true_sentimiento_labels.extend(sentimiento_labels.cpu().numpy())
        correct_sentimiento += (sentimiento_predicted_batch_labels == sentimiento_labels).sum().item()

        # Predicciones para emociones
        emocion_outputs = model(input_ids, attention_mask=attention_mask)
        emocion_predicted_batch_labels = torch.argmax(emocion_outputs.logits, dim=1)
        emocion_predicted_labels.extend(emocion_predicted_batch_labels.cpu().numpy())
        true_emocion_labels.extend(emocion_labels.cpu().numpy())
        correct_emocion += (emocion_predicted_batch_labels == emocion_labels).sum().item()
    total += sentimiento_labels.size(0)
```

Figura 79. Evaluación del modelo. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Posteriormente, para facilitar la comparación y generación de métricas de evaluación, se convierten las listas de etiquetas de predicción y etiquetas verdaderas en matrices NumPy.

3.3.1. Métricas.

Las métricas son necesarias para poder evaluar el modelo entrenado, entre las métricas utilizadas se tienen las siguientes:

3.3.1.1. Precisión (Accuracy).

La precisión es una métrica utilizada para evaluar el rendimiento de un modelo de clasificación y se calcula dividiendo la cantidad de predicciones correctas entre el total de ejemplos, con la sentencia de código $\text{accuracy} = \text{correct} / \text{total}$.

```
# Calcular el accuracy
accuracy = correct / total
print(f"Accuracy o precisión general en el conjunto de prueba: {accuracy:.2f}")

Accuracy o precisión general en el conjunto de prueba: 0.70
```

Figura 80. Precisión general del modelo. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Teniendo como resultado una precisión general de 0.70 en el conjunto de prueba, lo que significa que el modelo ha realizado predicciones correctas para aproximadamente el 70% de los ejemplos.

Por lo tanto, este valor de precisión indica que el modelo está acertando de manera correcta aproximadamente el 70% de los casos en el conjunto de prueba.

3.3.1.2. Precisión, recall y F1-score para sentimientos.

Aunque se tiene un modelo con un 70% de precisión general, se ha realizado un reporte de clasificación para la etiqueta sentimiento con la sentencia de código:

```
# Calcular el classification report para sentimiento
sentimiento_report = classification_report(true_sentimiento_labels, sentimiento_predicted_labels,
                                          target_names=sentimiento_label2id.keys())
```

Figura 81. Calcular métricas en sentimientos. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

En donde se evalúan las métricas de precisión, recall y F1-score, respecto a la precisión indica la proporción de predicciones positivas que son realmente positivas, para la clase "positivo", la precisión es 0.60, lo que significa que del total de predicciones que el modelo hizo como "positivo", aproximadamente el 60% de ellas eran correctas.

El recall o sensibilidad, hace referencia a la proporción de ejemplos reales de la clase que el modelo ha identificado correctamente, para la clase "positivo", el recall es 1.00, lo que significa que el modelo ha identificado todos los ejemplos reales de la clase "positivo".

En cuanto al F1-score, combina la precisión y recall, brindando una medida equilibrada del rendimiento del modelo, para la clase "positivo", el F1-score es 0.75, lo que indica que el modelo tiene un buen equilibrio entre precisión y recall para esta clase.

Classification Report para Sentimiento:				
	precision	recall	f1-score	support
positivo	0.60	1.00	0.75	6
negativo	1.00	0.71	0.83	14
accuracy			0.80	20
macro avg	0.80	0.86	0.79	20
weighted avg	0.88	0.80	0.81	20

Figura 82. Reporte de clasificación para sentimientos . Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.3.1.3. Precisión, recall y F1-score para emociones

De la misma manera que para los sentimientos, en las emociones también se hace un reporte de clasificación para la etiqueta emoción con la siguiente sentencia de código:

```
# Calcular el classification report para emocio
emocion_report = classification_report(true_emocion_labels, emocion_predicted_labels,
                                       labels=list(emocion_label2id.values()),
                                       target_names=emocion_label2id.keys())
```

Figura 83. Calcular métricas en emociones. Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

En este caso, la precisión varía para los diferentes tipos de emociones, se puede observar que para la emoción “tristeza”, la precisión es 0.40, lo que significa que el 40% de las predicciones de este tipo son correctas, es importante mencionar que en el presente proyecto de titulación se asocia más a los sentimientos negativos, por lo que es un modelo factible para lo que se requiere.

Classification Report para Emoción:				
	precision	recall	f1-score	support
esperanza	0.20	1.00	0.33	2
tristeza	0.40	0.80	0.53	5
miedo	0.00	0.00	0.00	5
enfado	0.00	0.00	0.00	4
alegría	0.00	0.00	0.00	4
accuracy			0.30	20
macro avg	0.12	0.36	0.17	20
weighted avg	0.12	0.30	0.17	20

Figura 84. Reporte de clasificación para emociones . Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Respecto al recall o sensibilidad, para el tipo “esperanza”, el recall es 1.00, lo que significa que el modelo ha identificado correctamente todos los ejemplos reales de la clase "esperanza" y para “tristeza” el recall es 0.80, indicando que el modelo ha identificado el 80% de los ejemplos reales de la clase "tristeza".

En cuanto al F1-score, para “esperanza” es 0.33, lo que indica un equilibrio razonable entre precisión y recall para esta clase y para “tristeza” es 0.53, indicando un mejor equilibrio que la clase "esperanza".

3.3.1.4. Matriz de Confusión.

Para realizar la matriz de confusión se ha utilizado “confusion_matrix” de Sklearn para visualizar el rendimiento del modelo al comparar las etiquetas reales de las muestras con las predichas por el modelo.

```
Matriz de Confusión para Sentimiento:
[[ 6  0]
 [ 4 10]]
```

Figura 85. Matriz de confusión para sentimiento . Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

En la matriz de confusión respecto a las predicciones de los sentimientos, se tienen 6 verdaderos positivos, lo que indica que el modelo identificó 6 casos de sentimiento “positivo” de manera correcta. En cuanto a los falsos positivos, el valor es 0, por lo que se interpreta que no hubo casos donde el modelo haya errado al predecir “negativo”. Respecto a los verdaderos negativos, se observa un valor de 10, es decir que el modelo identificó 10 casos de sentimiento “negativo” correctamente.

En cuanto a la matriz de confusión para las emociones, contiene cinco filas que son representadas por las emociones de la siguiente manera.

- Fila 1: Representa la clase "esperanza".
- Fila 2: Representa la clase "alegría".
- Fila 3: Representa la clase "tristeza".
- Fila 4: Representa la clase "enfado".
- Fila 5: Representa la clase "miedo".

Entre las representaciones más relevantes, se tienen:

- Fila 1: Tiene 2 en la columna 1, que hace referencia a la esperanza y ceros en las demás columnas, es decir que hay 2 muestras de la clase esperanza que fueron clasificadas correctamente como esperanza (verdaderos positivos) y ninguna fue clasificada como ninguna de los otros tipos de emociones.
- Fila 2: Representa la emoción "alegría": Tiene 1 en la columna 1 (esperanza) y 4 en la columna 2 (alegría), lo que significa que hay 1 muestra de "esperanza" clasificada como "esperanza" y 4 muestras de "alegría" clasificadas como "alegría"

(verdaderos positivos). No hubo muestras clasificadas como ninguna de las otras clases.

```
Matriz de Confusión para Emoción:
[[2 0 0 0 0]
 [1 4 0 0 0]
 [2 3 0 0 0]
 [1 3 0 0 0]
 [4 0 0 0 0]]
```

Figura 86. Matriz de confusión para emoción . Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.4. Resultados

Para poder corroborar los resultados del presente análisis de sentimientos, es necesario guardar el modelo entrenado, por consiguiente, se ha utilizado Google Drive para almacenar el modelo dentro de un objeto denominado “ruta_modelos_entrenados” y el modelo tiene por nombre “modelo_propuesto_sentimientos_emociones”.

```
# Guardar el modelo
ruta_modelo_guardado = f"{ruta_modelos_entrenados}modelo_propuesto_sentimientos_emociones"
model.save_pretrained(ruta_modelo_guardado)
```

Figura 87. Almacenamiento del modelo entrenado . Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Posteriormente, para realizar las predicciones o inferencias de los sentimientos y emociones, se debe cargar el modelo entrenado definiendo la ruta en la que se almacenó el modelo, utilizando la clase “BertForSequenceClassification” del framework Transformers de Hugging Face.

```
# Definir la ruta del modelo guardado
ruta_modelo_guardado = "/content/drive/My Drive/modelos_entrenados_colab/modelo_propuesto_sentimientos_emociones"

# Cargar el modelo previamente entrenado
modelo_cargado = BertForSequenceClassification.from_pretrained(ruta_modelo_guardado)
```

Figura 88. Carga del modelo entrenado . Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Antes de utilizar el modelo previamente entrenado, es necesario preparar los datos de entrada, en este caso los textos de las noticias seleccionadas en Twitter por las verificadoras de hecho en Ecuador deben encontrarse preprocesados, posteriormente se debe cargar el tokenizador de Bert con el modelo preentrenado de Bert en español para tokenizar y codificar los textos, los cuales se han almacenado en el objeto “textos_entrada”.

```
# Tokenizar y codificar los textos de entrada
encodings_textos_entrada = tokenizer(textos_entrada, truncation=True, padding=True, max_length=40, return_tensors="pt")
```

Figura 89. Tokenización y codificación de textos de entrada para las predicciones . Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

3.4.1. Predicciones de sentimientos y emociones.

Para realizar predicciones de los sentimientos y emociones de los datos de entrada se utiliza el modelo previamente cargado, luego se determina “encodings_textos_entrada.input_ids” como las entradas de ID de tokens y “encodings_textos_entrada.attention_mask” como la máscara de atención correspondiente para los encodings de los textos de entrada y los resultados se almacenan en “emocion_outputs”, posteriormente se utiliza la variable “emocion_id2label” para convertir los índices de etiquetas predichos en etiquetas legibles y comprensibles para las emociones y se utiliza “torch.argmax” para conseguir el índice de la etiqueta predicha con la puntuación más alta para cada ejemplo.

```
# Realizar predicciones para emociones
with torch.no_grad():
    emocion_outputs = modelo_cargado(encodings_textos_entrada.input_ids, attention_mask=encodings_textos_entrada.attention_mask)
    emocion_predicho_indices = torch.argmax(emocion_outputs.logits, dim=-1)

# Convertir los índices de etiquetas en etiquetas legibles para emociones
emocion_predicha_etiquetas = [emocion_id2label[idx] for idx in emocion_predicho_indices.tolist()]

# Imprimir las predicciones para sentimientos y emociones
for texto, sentimiento, emocion in zip(textos_entrada, sentimiento_predicho_etiquetas, emocion_predicha_etiquetas):
    print(f"Texto: {texto}")
    print(f"Sentimiento predicho: {sentimiento}")
    print(f"Emoción predicha: {emocion}")
    print()
```

Figura 90. Tokenización y codificación de textos de entrada para las predicciones . Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Al finalizar, se obtiene como salida: el texto, el sentimiento y la emoción predichos, como se muestra unos ejemplos en la figura 90.

```
Texto: país vuelve contar muertos 31 personas asesinadas penitenciaría litoral el origen reciente ruptura lobos tiguerones reveló nueva medida presión secuestro guías penitenciarios
Sentimiento predicho: negativo
Emoción predicha: tristeza

Texto: según momento hallado 18 presos muertos penitenciaría litoral prisión registran enfrentamientos sábado
Sentimiento predicho: negativo
Emoción predicha: tristeza

Texto: mañana reportado múltiples disturbios esmeraldas registraron atentados explosivos agencia cnel balaceras cerca escuelas debido ataques cnel empresa confirmado cierre 12 oficinas servicio cliente
Sentimiento predicho: negativo
Emoción predicha: miedo

Texto: además diferentes pabellones decomisado droga municiones armas cortopunzantes armas fuego celulares electrodomésticos dinero efectivo
Sentimiento predicho: negativo
Emoción predicha: enfado
```

Figura 91. Predicciones de sentimientos y emociones . Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Posteriormente, se crea un DataFrame con las predicciones, en este caso representadas en las columnas “Sentimiento predicho” y “Emoción predicha”, también las columnas: ID del tweet, ID del Autor, “Nombre del Autor” y “Texto”, denominado “df_predicciones” y se guardan en un archivo JSON utilizando la siguiente sentencia de código:

```
import json

# Se define la ruta del archivo json
ruta_archivo_json = "ruta/del/archivo.json"

# Convertir el DataFrame a formato JSON
json_data = df_predicciones.to_json(orient="records", lines=True)

# Guardar el JSON en un archivo
with open(ruta_archivo_json, "w") as archivo:
    archivo.write(json_data)
```

Figura 92. Almacenamiento de predicciones en archivo JSON . Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

A continuación, se muestra un ejemplo del cómo se visualiza un documento guardado en formato JSON.

```
{
  "ID": "1683968410609233921",
  "ID del Autor": "777891400297897985",
  "Nombre del Autor": "Ecuador Chequea",
  "Texto": "\ud83d\udea8#URGENTE | La @FiscaliaEcuador inform\u00f3 que hasta ahora hay 31 muertos y 14 heridos, tras los enfrentamientos registrados desde el",
  "Sentimiento predicho": "negativo",
  "Emoción Predicha": "tristeza"
}
```

Figura 93. Documento almacenado en formato JSON . Información adaptada de (Google Colab, 2023). Elaborado por Jiménez Kimberly.

Para finalizar, estos documentos se almacenan en una nueva colección denominada “predicciones” en la base de datos “Twitter_principal” de MongoDB Atlas.

3.5. Análisis de Resultados

Al visualizar los resultados de las predicciones de los sentimientos y las emociones identificadas en las noticias de las verificadoras de hecho en Ecuador aplicando técnicas de Procesamiento de Lenguaje Natural (PLN), se podría indicar que se tienen resultados alentadores, brindando una visión de cómo pueden llegar a percibirse estas noticias.

Utilizando como modelo base el modelo preentrenado de Bert en español junto con datos previamente etiquetados, se ha podido entrenar un modelo que ha logrado identificar y categorizar un alto nivel de precisión los sentimientos y emociones asociadas al contenido de las noticias seleccionadas por Ecuador Chequea y Ecuador Verifica.

Es importante mencionar que, al verificar los resultados de las predicciones de los sentimientos, se pudo corroborar que existen más noticias categorizadas por el sentimiento negativo, esto se puede interpretar que es porque las noticias contienen información de los últimos sucesos que se han presentado en el país, los cuales generan sentimientos negativos y de tristeza a las personas en el Ecuador.

3.6. Conclusiones

En el transcurso de este trabajo de titulación, se ha podido verificar a profundidad el funcionamiento de un análisis de sentimientos en el Procesamiento de Lenguaje Natural (PLN), se han analizado varias técnicas de procesamiento de texto y varios modelos de clasificación, seleccionando el modelo preentrenado de BERT en español de la librería Transformers de Hugging Face, se realizaron pruebas con este modelo sin previo entrenamiento para determinar su funcionamiento en el análisis de sentimientos, se verificó que puede identificar y categorizar sentimientos en positivos y negativos, pero en el presente caso de estudio, se pretendía identificar las emociones que se perciben en las noticias seleccionadas por las verificadoras de hechos del Ecuador.

Por consiguiente, se entrenó el modelo con datos etiquetados con sentimientos en negativos y positivos y las emociones esperanza, alegría, tristeza, enfado y miedo, procediendo posteriormente con una evaluación del modelo en donde se obtuvo una precisión del 70% respecto a las predicciones del conjunto de prueba, verificando así un comportamiento adecuado para el análisis de sentimientos en las noticias.

Es importante mencionar, que el uso del modelo preentrenado de BERT, ha demostrado una precisión favorable en la clasificación o categorización de los sentimientos en las noticias, adicionalmente, un punto relevante en este trabajo de titulación es que al tener la categorización de los sentimientos en un formato JSON, esto podría ser utilizado posteriormente para verificar si una noticia cuando se etiqueta con el sentimiento negativo tiende a ser falsa.

3.7. Recomendaciones

Sobre las conclusiones indicadas, en el futuro se sugiere explorar con mayor medida el refinamiento del modelo, es decir que se puede entrenar aún más para conseguir aún mejores resultados, esto va a depender del contexto al cual se requiere aplicar.

Además, este modelo entrenado para el análisis de sentimientos de las noticias seleccionadas en Twitter -actualmente conocido con el nombre de X- por las verificadoras de hechos Ecuador Chequea y Ecuador Verifica, se podría entrenar con más datos y con diferentes emociones, cabe recalcar que para esto se deben aplicar las métricas correspondientes para poder evaluar el rendimiento del modelo.

Adicionalmente, este análisis de sentimientos puede servir para que se considere los sentimientos y emociones que puede percibir el público respecto a las noticias que se

comparten en la plataforma de Twitter -actualmente conocido con el nombre de X- por las verificadoras de hecho anteriormente mencionadas.

ANEXOS

Anexo N° 1

Preguntas de la entrevista realizada a la periodista Paola Simbaña de Ecuador Chequea.

Entrevista para el desarrollo del componente de análisis de sentimientos de las noticias comprobadas en Twitter por las verificadoras acreditadas en Ecuador utilizando procesamiento de lenguaje natural.

1. ¿Cuál es la primera señal que lleva a considerar que posiblemente una noticia sea falsa?
2. ¿Cuál es la importancia de la transparencia y la imparcialidad en el trabajo de verificación de noticias?
3. En su trayectoria como periodista del portal que se encuentra certificado por la IFCN (Ecuador Chequea) ¿cuál cree usted que es el principio de la IFCN más complejo de cumplir y por qué?
4. ¿Considera que las personas pueden experimentar emociones negativas al ser víctimas de la desinformación y por qué?
5. ¿Al momento de verificar una noticia ha podido observar que existen emociones tales como: alegría, enfado, tristeza, ¿en dicha noticia?
6. ¿Ha notado situaciones en las que las noticias falsas han causado una reacción emocional masiva en la sociedad? ¿Podría indicar algún ejemplo?
7. Sabiendo que las noticias falsas pueden tener consecuencias negativas, tales como: originar conflictos, dividir a la sociedad y manipular a los ciudadanos ¿considera usted que un indicador de que una noticia posiblemente sea falsa es el sentimiento negativo que pueda provocar en las personas?
8. En su trayectoria como periodista del portal pionero de verificación a nivel nacional (Ecuador Chequea) ¿cuál cree usted que es la red social en la que se presenta mayor cantidad de noticias falsas?
9. En la escala del 1 al 10 ¿cuán importante es el uso de herramientas tecnológicas en la verificación de contenido?
10. ¿En base a su experiencia en el Fact-checking, considera importante que se desarrolle un sistema de ayuda al Fact-checker permitiéndole optimizar el tiempo de verificación de una noticia y por qué?

Bibliografía

- Amazeen, M. A. (2020). Journalistic interventions: The structural factors affecting the global emergence of fact-checking. *Journalism*, 21(1), 95–111. <https://doi.org/10.1177/1464884917730217>
- Alcott, H., y Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspective*, 31(2), 211-236
- Alva Segura, D. A. (2020). Análisis del Sentimiento Político en Twitter durante las Elecciones Congresales 2020 en el Perú.
- Bakir, V., & McStay, A. (2018). Fake News and the Economy of Emotions. *Digital Journalism*, 6(2), 154-175. doi:<https://doi.org/10.1080/21670811.2017.1345645>
- Capuano, N., Fenza, G., Loia, V., & Nota, F. D. (2023). Content-Based Fake News Detection With Machine and Deep Learning: a Systematic Review. *Neurocomputing*, 530(ISSN 0925-2312), 91-103. doi:<https://doi.org/10.1016/j.neucom.2023.02.005>
- Celaya, J. (2008). *La empresa en la web 2.0*. España: Editorial Grupo Planeta.
- Chavero, P., & Intriago, D. (mayo de 2021). Las fake news como herramienta. 19-35.
- Cusot, G., & Palacios, I. (2019). Las FAKE NEWS y las estrategias de verificación del discurso público: Caso Ecuador Chequea. 3, págs. 88-107. doi:<https://doi.org/10.18272/pd.v3i1.1558>
- DATLAS. (16 de febrero de 2020). *4 Metodologías para proyectos de Data Science – INVESTIGACIÓN DATLAS*. Obtenido de Blog Datlas: <https://blogdatlas.wordpress.com/2020/02/16/4-metodologias-para-proyectos-de-data-science-datlas-research/>
- Dols Hernández, A. (2018). *Análisis de la metodología del fact-checking: Caso de Chequeado y Ecuador Chequea*.
- Ecuador Chequea. (17 de octubre de 2018). *Ecuador Chequea by Fundamedios*. Obtenido de Ecuador Chequea frente a los hechos ocurridos en Posorja: <https://ecuadorchequea.com/editorial-ecuadorchequea-posorja-linchamiento-fakenews/>

- Galdón, G. (2001). *Información, desinformación y manipulación*. Obtenido de https://repositorioinstitucional.ceu.es/bitstream/10637/1494/1/Informacion_Galdon_2001.pdf
- Galeano, S. (26 de enero de 2023). *El número de usuarios de internet en el mundo crece un 1,9% y alcanza los 5.160 millones (2023)*. Obtenido de Marketing 4 Ecommerce: <https://marketing4ecommerce.net/usuarios-de-internet-mundo/>
- Gleick, J. (2011). *La información: historia y realidad*. (koothrapali, Ed.)
- Graves, L., & Cherubini, F. (2016). *The Rise of Fact-Checking sites in Europe*. Reuters Institute for the Study of Journalism. Obtenido de <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/The%2520Rise%2520of%2520Fact-Checking%2520Sites%2520in%2520Europe.pdf>
- Guzzi, P. (2019). Computing Languages for Bioinformatics: Python. (M. G. Shoba Ranganathan, Ed.) *Encyclopedia of Bioinformatics and Computational Biology*(ISBN 978012811432), 195-198. doi:<https://doi.org/10.1016/B978-0-12-809633-8.20366-X>
- Horner, C. G., Galletta, D., Crawford, J., & Shirsat, A. (2021). Emotions: The Unexplored Fuel of Fake News on Social Media. *Journal of Management Information Systems*, 38(4), 1039-1066. doi:<https://doi.org/10.1080/07421222.2021.1990610>
- Hütt Herrera, H. (2012). LAS REDES SOCIALES: UNA NUEVA HERRAMIENTA DE DIFUSIÓN. *Reflexiones*, 91(2), 121-128.
- Kemp, S. (26 de enero de 2023). *DataReportal - Global Digital Insights*. Obtenido de DataReportal - Global Digital Insights: <https://datareportal.com/reports/digital-2023-global-overview-report>
- Komuro, J., Kusumoto, D., Hashimoto, H., & Yuasa, S. (2023). Machine learning in cardiology: Clinical application and basic research. *Journal of Cardiology*(ISSN 0914-5087). doi:<https://doi.org/10.1016/j.jjcc.2023.04.020>
- Liu, B. (2015). *Sentiment Analysis*. New York: Cambridge University Press.
- Mayo, M. (3 de mayo de 2018). *Preprocesamiento de datos de texto: un tutorial en Python*. Obtenido de <https://medium.com/datos-yciencia/preprocesamiento-de-datos-de-texto-un-tutorial-en-python-5db5620f1767>
- Moreno-Gil, V., Ramon, X., & Rodriguez-Martinez, R. (2021). Fact-Checking Interventions as Counteroffensives to Disinformation Growth: Standards, Values, and Practices in Latin America and Spain. *Media and Communication*, 9(1), 251-263. doi:<https://doi.org/10.17645/mac.v9i1.3443>

- Naso, F., Balbi, M. L., Di Grazia, N., & Peri, J. A. (2012). La importancia de las Redes sociales en el ámbito educativo. *Universidad Nacional del Noroeste de la Provincia de Buenos Aires*, 1-8.
- Nigro, H. O., Xodo, D., Corti, G., & Terren, D. (2004). KDD (Knowledge Discovery in Databases): Un proceso centrado en el usuario. *In VI Workshop de Investigadores en Ciencias de la Computación*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- Piñero, M. S. (2021). MARKETING EN REDES SOCIALES. *Revista de Estudios Empresariales*(2), 317-319. doi:<https://doi.org/10.17561/ree.n2.2022.7150>
- Priya, B. (31 de marzo de 2023). *Los 12 mejores cuadernos colaborativos de ciencia de datos [alternativas de Jupyter]*. Recuperado el 8 de junio de 2023, de Geekflare: <https://geekflare.com/es/best-data-science-notebooks/>
- Quintana Pujalte, L., & Pannunzio, M. F. (2021). Fact-checking en latinoamérica. Tipología de contenidos virales desmentidos durante la pandemia del coronavirus. *Revista de Ciencias de la Comunicación e Información*, 26, 27-46. doi:<http://doi.org/10.35742/rcci.2021.26.e178>
- Romero Moreno, F. Y., Sanchez Martelo, C. A., Breed Yeet, A. C., Sanchez Cifuentes, J. F., & Ospina López, J. P. (2020). Técnicas para la Clasificación de Sentimientos en Redes Sociales como Apoyo en el Marketing Digital. *RISTI*(35), 167-186.
- Rouhiainen, L. (2018). *Inteligencia artificial 101 cosas que debes saber hoy sobre nuestro futuro*. Madrid: © Editorial Planeta, S.A.
- Santamaria, P. (16 de enero de 2017). Antropología social: La diferencia y relación entre posverdad y las noticias falsas. *Merca2.0*. Obtenido de Merca2.0: <https://goo.gl/agXAMR>
- Stencel, M., & Luther, J. (22 de junio de 2020). Annual census finds nearly 300 fact-checking projects around the world. Obtenido de <https://reporterslab.org/annual-census-finds-nearly-300-fact-checking-projects-around-the-world/>
- Toapanta, M. (2023). nvestigaciones de tesis doctoral como estudiante de la Universidad de Jaén y cursante del doctorado en Tecnologías de Información y Comunicación.
- Ufarte Ruiz, M. J., Peralta García, L., & Murcia Verdú, F. J. (2018). Fact checking: un nuevo desafío para el periodismo. *El profesional de la información*, 1-9. doi:<https://doi.org/10.3145/epi.2018.jul.02>

- Valverde-Berrocoso, J., González-Fernández, A., & Acevedo-Borrega, J. (2022). Desinformación y multialfabetización: Una revisión sistemática de la literatura. *Comunicar Revista Científica de Educomunicación*.
- Vélez-Bermello, G. L. (Julio - diciembre de 2020). Inmediatez y fact-checking: análisis del portal Ecuador Chequea. *Revista ABRA*, 40(61), 63-87. doi:<https://doi.org/10.15359/abra.40-61.3>
- Wardle, C., & Derakhshan, H. (2017). Information Disorder. Toward an interdisciplinary framework for research and policymaking. *Council of Europe*. Obtenido de <https://bit.ly/2V9xsdy>
- Zommer, L. (2014). El boom del fact checking en América Latina Aprendizajes y desafíos del caso de Chequeado. *Konrad Adenauer Stiftung - Chequeando*, 1-58.