



**UNIVERSIDAD DE GUAYAQUIL
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA DE INGENIERÍA EN TELEINFORMÁTICA**

**TRABAJO DE TITULACIÓN
PREVIO A LA OBTENCIÓN DEL TÍTULO DE
INGENIERA EN TELEINFORMÁTICA**

**ÁREA
TECNOLOGÍA DE LOS ORDENADORES**

**TEMA
“EVALUACIÓN DE CONTENIDO DE ENTRADA Y
POSIBLE SALIDA EN CONVERSACIONES TEXTUALES
DE PERSONAS CONTAGIADAS DE COVID-19 PARA
IDENTIFICAR UN MODELO PLN EFECTIVO POR MEDIO
DEL ENTRENAMIENTO DE MACHINE LEARNING DEL
FCI 010-2021”**

**AUTORA
SOLÓRZANO MONSERRATE MIRIAN ESTEFANÍA**

**DIRECTOR DEL TRABAJO
ING. COMP. ACOSTA GUZMÁN IVÁN LEONEL, MSIG.**

GUAYAQUIL, ABRIL 2022



ANEXO XI.- FICHA DE REGISTRO DE TRABAJO DE TITULACIÓN



FACULTAD DE INGENIERÍA INDUSTRIAL CARRERA INGENIERÍA EN TELEINFORMÁTICA

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA			
FICHA DE REGISTRO DE TRABAJO DE TITULACIÓN			
TÍTULO Y SUBTÍTULO:			
Evaluación de contenido de entrada y posible salida en conversaciones textuales de personas contagiadas de covid-19 para identificar un modelo PLN efectivo por medio del entrenamiento de machine learning del FCI 010-2021.			
AUTOR(ES) (apellidos/nombres):		Solórzano Monserrate Mirian Estefanía	
REVISOR(ES)/TUTOR(ES) (apellidos/nombres):		Ing. Zurita Hurtado Harry Alfredo, Mg. / Ing. Comp. Acosta Guzmán Iván, MSIG.	
INSTITUCIÓN:		Universidad de Guayaquil	
UNIDAD/FACULTAD:		Facultad Ingeniería Industrial	
MAESTRÍA/ESPECIALIDAD:			
GRADO OBTENIDO:		Ingeniería en Teleinformática	
FECHA DE PUBLICACIÓN:		21 de abril del 2022	No. DE PÁGINAS: 108
ÁREAS TEMÁTICAS:		Tecnología de los Ordenadores	
PALABRAS CLAVES/ KEYWORDS:		Recomendaciones, Virus, Algoritmos, NLP, Modulo	
<p>RESUMEN/ABSTRACT (150-250 palabras):</p> <p>Resumen</p> <p>El text mining es el proceso de examinar grandes volúmenes de documentos para descubrir nueva información o ayudar a responder preguntas de investigación específicas. Este ha generado un gran aporte tecnológico dando lugar a la creación de portales donde era compartida la información detallada de la situación que pasaban diversos países en la pandemia, tomando información de páginas web oficiales donde publicaban los contagios por Covid-19 y fallecidos actualizados permitiendo de esta manera a la población conocer cómo se expandía a diario el virus y así este mayormente informada de nuevas variantes y posibles olas de contagio.</p> <p>En este trabajo de titulación desarrollara diversos modelos como son Long-Short Term Memory (LSTM), modelo básico Artificial Neural Network (ANN), Random Forest para multilabel y KNeighborsClassifier (kNN) en los cuales se pretende proporcionar un módulo</p>			

aplicando algoritmos supervisados con arquitecturas NLP con la finalidad de detectar patrones de recomendaciones más comunes dadas por los médicos especialistas del Ecuador a los pacientes contagiados del virus Covid-19 en la fase de acompañamiento médico o post Covid para superar este virus de esta manera descubrir mediante el modelo NLP de forma cuantitativa las recomendaciones más efectivas dadas por los médicos entre el mes de marzo del año 2020 a marzo del año 2022.

Abstract

Text mining is the process of examining large volumes of documents to discover new information or help answer specific research questions. This has generated a great technological contribution leading to the creation of portals where detailed information on the situation in different countries during the pandemic was shared, taking information from official web pages where they published updated Covid-19 infections and deaths, thus allowing the population to know how the virus was spreading daily and thus be better informed of new variants and possible waves of infection.

In this degree work we will develop several models such as Long-Short Term Memory (LSTM), Artificial Neural Network (ANN) basic model, Random Forest for multilabel and KNeighborsClassifier (kNN) in which it is intended to provide a module applying supervised algorithms with NLP architectures in order to detect patterns of most common recommendations given by medical specialists in Ecuador to patients infected with the Covid-19 virus in the phase of medical support or post Covid to overcome this virus in this way discover through the NLP model quantitatively the most effective recommendations given by doctors between the month of March 2020 to March 2022.

ADJUNTO PDF:	SI (X)	NO
CONTACTO CON AUTOR/ES:	Teléfono: 593-94510778	E-mail: mirian.solorzanom@ug.edu.ec
CONTACTO CON LA INSTITUCIÓN:	Nombre: Ing. Ramón Maquilón Nicola, Mg.	
	Teléfono: 593- 2658128	
	E-mail: direcciónTi @ug.edu.ec	



**ANEXO XII 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 INGENIERÍA EN TELEINFORMÁTICA**

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

Yo, **SOLÓRZANO MONSERRATE MIRIAN ESTEFANÍA**, con C.C. No. **094140080-6**, certifico que los contenidos desarrollados en este trabajo de titulación, cuyo título es **“EVALUACIÓN DE CONTENIDO DE ENTRADA Y POSIBLE SALIDA EN CONVERSACIONES TEXTUALES DE PERSONAS CONTAGIADAS DE COVID-19 PARA IDENTIFICAR UN MODELO PLN EFECTIVO POR MEDIO DEL ENTRENAMIENTO DE MACHINE LEARNING DEL FCI 010-2021.”** 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.

Mirian Solórzano

Solórzano Monserrate Mirian Estefanía

C.C.No. 0941400806



**ANEXO VII.- CERTIFICADO PORCENTAJE DE SIMILITUD
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA INGENIERÍA EN TELEINFORMÁTICA**



Habiendo sido nombrado ING. COMP. ACOSTA GUZMÁN IVÁN LEONEL, MSIG tutor del trabajo de titulación certifico que el presente trabajo de titulación ha sido elaborado por SOLÓRZANO MONSERRATE MIRIAN ESTEFANÍA C.C. 0941400806, con mi respectiva supervisión como requerimiento parcial para la obtención del título de Ingeniera en Teleinformática. .

Se informa que el trabajo de titulación: EVALUACIÓN DE CONTENIDO DE ENTRADA Y POSIBLE SALIDA EN CONVERSACIONES TEXTUALES DE PERSONAS CONTAGIADAS DE COVID-19 PARA IDENTIFICAR UN MODELO PLN EFECTIVO POR MEDIO DEL ENTRENAMIENTO DE MACHINE LEARNING DEL FCI 010-2021, ha sido orientado durante todo el periodo de ejecución en el programa Antiplagio URKUND quedando el 7% de coincidencia.

Documento	EXTRATO Tesis Capítulo I 2 3 Mirian Solórzano V10.docx (D130405703)
Presentado	2022-03-14 20:43 (-05:00)
Presentado por	Ivan Acosta (ivan.acostag@ug.edu.ec)
Recibido	ivan.acostag.ug@analysis.orkund.com
Mensaje	TESIS MIRIAN SOLORZANO Mostrar el mensaje completo
	7% de estas 33 páginas, se componen de texto presente en 3 fuentes.

Lista de fuentes Bloques		➔ Abrir sesión
+	Categoría	Enlace/nombre de archivo
+	>	EXTRACTO - TESIS - JOSELYN DENISSE TUMBACO BRAVO - VERSION 2.1.docx ✓
+		ICI 6541[9] Proyecto de Titulo - Jamett - Martinez.pdf ✓
+		75.583_20192_Practica_12561997.txt ✓
+	Fuentes alternativas	
+	Fuentes no usadas	

<https://secure.orkund.com/view/124561468-640655-963634>



Firmado electrónicamente por:

**IVAN LEONEL
ACOSTA GUZMAN**

ING. COMP. ACOSTA GUZMÁN IVÁN LEONEL, MSIG.
DOCENTE TUTOR
C.C. 0914940812
FECHA: LUNES 14 MARZO 2022



**ANEXO VI. - CERTIFICADO DEL DOCENTE-TUTOR DEL
TRABAJO DE TITULACIÓN
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA INGENIERÍA EN TELEINFORMÁTICA**



Guayaquil, 14 de marzo del 2022

Sr (a).

Ing. Annabelle Lizarzaburu Mora, MG.

Director (a) de Carrera Ingeniería en Teleinformática / Telemática

**FACULTAD DE INGENIERÍA INDUSTRIAL DE LA UNIVERSIDAD DE
GUAYAQUIL**

Ciudad. -

De mis consideraciones:

Envío a Ud. el Informe correspondiente a la tutoría realizada al Trabajo de Titulación **EVALUACIÓN DE CONTENIDO DE ENTRADA Y POSIBLE SALIDA EN CONVERSACIONES TEXTUALES DE PERSONAS CONTAGIADAS DE COVID-19 PARA IDENTIFICAR UN MODELO PLN EFECTIVO POR MEDIO DEL ENTRENAMIENTO DE MACHINE LEARNING DEL FCI 010-2021** de la estudiante **SOLÓRZANO MONSERRATE MIRIAN ESTEFANÍA**, 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 titulación con la respectiva calificación.

Dando por concluida esta tutoría de trabajo de titulación, **CERTIFICO**, para los fines pertinentes, que la estudiante está apta para continuar con el proceso de revisión final.



Firmado electrónicamente por:

**IVAN LEONEL
ACOSTA GUZMAN**

ING. COMP. ACOSTA GUZMÁN IVÁN LEONEL, MSIG.
DOCENTE TUTOR
C.C. 0914940812
FECHA: LUNES 14 MARZO 2022



**ANEXO VIII.- INFORME DEL DOCENTE REVISOR
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA INGENIERÍA EN TELEINFORMÁTICA**



Guayaquil, 26 de marzo de 2022.

Sr (a).

Ing. Annabelle Lizarzaburu Mora, MG.

Director (a) de Carrera Ingeniería en Teleinformática / Telemática

FACULTAD DE INGENIERÍA INDUSTRIAL DE LA UNIVERSIDAD DE GUAYAQUIL

Ciudad. -

De mis consideraciones:

Envío a Ud. el informe correspondiente a la REVISIÓN FINAL del Trabajo de Titulación **“EVALUACIÓN DE CONTENIDO DE ENTRADA Y POSIBLE SALIDA EN CONVERSACIONES TEXTUALES DE PERSONAS CONTAGIADAS DE COVID-19 PARA IDENTIFICAR UN MODELO PLN EFECTIVO POR MEDIO DEL ENTRENAMIENTO DE MACHINE LEARNING DEL FCI 010-2021”** del estudiante **SOLORZANO MONSERRATE MIRIAN ESTEFANÍA**. 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 32 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 titulación. Particular que comunicamos a usted para los fines pertinentes.

Atentamente,



Firmado electrónicamente por:
**HARRY ALFREDO
ZURITA HURTADO**

ING. ZURITA HURTADO HARRY ALFREDO, MG
C.C:0910561372

FECHA: 26/03/2022

Dedicatoria

A Dios

Por brindarme la salud y bienestar para concluir mi carrera universitaria.

A mi abuelita Rosa Veliz

Por estar conmigo siempre apoyándome en todo momento con sus palabras de fortaleza brindándome su cariño y amor.

A mi mamá Flor Monserrate

Por su amor, trabajo y sacrificio a lo largo de los años, gracias a ella he podido progresar en la vida. Es su orgullo y privilegio ser su hija, es la mejor mamá del mundo, siempre ha estado conmigo apoyándome incondicionalmente con sus consejos para ser una mejor persona, ella es uno de los pilares fundamentales en mi vida.

Agradecimiento

Agradezco a Dios por bendecir mi vida, guiarme a lo largo ella y brindarme fortaleza en tiempos de dificultad y debilidad.

Gracias a mi mama Flor Monserrate quien es la principal impulsora de mis sueños, confiando y creyendo en mis expectativas, sugerencias, valores y principios que me inculco.

A mi abuelita Rosa Veliz, Tías Paola, Juana y Yoconda quienes de una u otra manera siempre estuvieron conmigo apoyándome para cumplir mis sueños, a mis hermanos Jefferson Solórzano y Alexander Mariscal quienes estuvieron ahí cuando más lo necesite, a mi futuro esposo Néstor Santana por sus palabras fortalecedoras en momentos que ya no podía y me daba las fuerzas necesarias para continuar y no darme por vencida.

A mi amiga Alexandra Piza quien me ayudo en momentos difíciles dándome su apoyo a la distancia para que logre mi meta, gracias a los docentes quienes compartieron sus conocimientos de manera íntegra y dedicada durante el proceso de preparación profesional en especial al Ing. Rodolfo Parra y Lcdo. William Navas quienes más que docentes fueron amigos fuera de las aulas. A mi asesor de tesis Ing. Iván Acosta, que brindo su guía. Gracias por su apoyo, paciencia e integridad como docente, así como las valiosas contribuciones de los docentes integrantes del FCI. a esta investigación.

Declaración de Autoría

“La responsabilidad del contenido de este Trabajo de Titulación, me corresponde exclusivamente; y el patrimonio de este a la Facultad de Ingeniería Industrial de la Universidad de Guayaquil”

Índice general

N°	Descripción	Pág.
	Introducción	1

Capítulo I

El problema

N°	Descripción	Pág.
1.	El problema	3
1.1.	Planteamiento del problema	3
1.2.	Formulación del problema	3
1.3.	Sistematización del problema	4
1.4.	Objetivos	4
1.4.1.	Objetivo general.	4
1.4.2.	Objetivos específicos.	4
1.5.	Justificación	4
1.6.	Delimitación del problema	5
1.7.	Alcance	5
1.8.	Premisas de investigación	6
1.9.	Operacionalización de Variables	6
1.9.1.	Variable independiente	6
1.9.2.	Variable dependiente	6

Capítulo II

Marco teórico

N°	Descripción	Pág.
2.	Marco teórico	7
2.1.	Antecedentes del estudio	7
2.2.	Fundamentación teórica	8
2.2.1.	Inicios de la Inteligencia Artificial	8
2.2.2.	Inteligencia Artificial	9
2.2.3.	Aprendizaje Automático (Machine Learning)	11
2.2.3.1	Ciclo de vida del ML	11
2.2.3.2	Niveles de madurez de ML	12
2.2.3.3	Librerías usadas en ML	12
2.2.4	Aprendizaje Profundo (Deep Learning)	13

2.2.5	Tipos de Aprendizaje	14
2.2.5.1	Aprendizaje supervisado	14
2.2.5.2	Aprendizaje no supervisado	14
2.2.6	Algoritmos de Aprendizaje no supervisado	15
2.2.7	Datasets	16
2.2.8	Natural language processing (NLP)	17
2.2.8.1	Librerías utilizadas en NLP	18
2.2.8.2	Pre-procesamiento de texto	19
2.2.8.3	Técnicas en NLP	23
2.2.8.4	Modelos en NLP	23
2.2.9	Redes Neuronales Recurrentes (RNN)	29
2.2.10	Web Scraping	30
2.2.11	Lenguajes y herramientas en Inteligencia Artificial	30
2.2.11.1	Lenguajes de Programación	31
2.2.11.2	Entorno de desarrollo Integrado (IDE)	35
2.2.11.2.1	IDE en la nube	35
2.2.11.2.2	IDE de escritorio	37
2.3	Fundamentación Legal	39
2.3.1	Revisión de Normativas Nacionales	39
2.3.2	Revisión de Normativas Internacionales del uso de la IA	41

Capítulo III

Propuesta

N°	Descripción	Pág.
3.	Metodología	42
3.1.	Propuesta tecnológica	42
3.1.1.	Descripción del proceso metodológico	42
3.2.	Tipos de investigación	43
3.2.1.	Investigación exploratoria.	44
3.3.	Metodología de investigación	44
3.3.1.	Metodología para la Revisión Bibliográfica	44
3.3.2.	Metodología Cuantitativa	44
3.3.2.1	Técnicas de Investigación	45
3.3.2.1.1	Encuesta	45

3.3.2.2	Descripción del procedimiento metodológico	45
3.3.3	Metodología Cualitativa	56
3.3.3.1	Técnicas de investigación	56
3.3.3.1.1	Entrevista	56
3.4.	Construcción del modelo de Machine Learning	57
3.4.1	Importación de datos	58
3.4.2	Tratamientos de datos	60
3.4.2.1	Aplicación de Técnicas	63
3.4.2.2	Elección del modelo	65
3.4.2.3	Configuración de parámetros	66
3.4.2.4	Evaluación del modelo	66
3.4.2.5	Comparación de métricas	68
3.5	Conclusiones	69
3.6	Recomendaciones	69
	Anexos	71
	Referencia bibliográfica	

Índice de tablas

Nº	Descripción	Pág.
1	Delimitación del problema	5
2	Historia de la Inteligencia Artificial del año 1854 al 2015	8
3	Historia de la Inteligencia Artificial del año 2017 al 2021	9
4	Características del Lenguaje Python	32
5	Versiones de Python destacando sus características	32
6	Características del Lenguaje C	33
7	Características de Lenguaje Java	34
8	Características de Lenguaje C++	35
9	Características de Jupyter	37
10	Características de Spyder	38
11	Edad	47
12	Género	48
13	Residentes de la Zona 8	48
14	Importancia de conocer síntomas de Coronavirus	49
15	Efectividad de Vacunas	50
16	Información sobre el Covid-19	51
17	Conocimientos sobre Hábitos Saludables para pacientes contagiados de Covid-19	52
18	Posesión de Smartphone de los Habitantes	53
19	Conocimientos sobre Inteligencia Artificial	54
20	Importancia de herramientas tecnológicas	55

Índice de figuras

Nº	Descripción	Pág.
1	Ejemplos de dónde podría utilizarse la inteligencia artificial	10
2	Mapa mental de la Clasificación de la Inteligencia Artificial	11
3	Ciclo de vida de ML	12
4	Niveles de madurez del ML	12
5	Clasificación de los tipos de Aprendizaje	14
6	Representación de la Regresión lineal	15
7	Representación de la Regresión lineal con una recta	15
8	Pasos para creación de un modelo	16
9	Mapa mental de la Clasificación de NLP	18
10	Aplicación de un token a una oración en Google Colaboraty	20
11	Aplicación de un stemming en un conjunto de palabras en Google Colaboraty	20
12	Aplicación de Lematización a una oración en Google Colaboraty	22
13	Numero de neuronas utilizadas en el inspector visual de incrustación de palabras	24
14	Matrices de peso del Inspector Visual de palabras	25
15	Vectores resultantes del Inspector Visual de palabras	25
16	Arquitectura del modelo FastText	26
17	Arquitectura del modelo Transformer	27
18	Arquitectura del modelo ELMo	28
19	Arquitectura del modelo GPT	28
20	Arquitectura del modelo BERT	29
21	Algoritmo KNN	31
22	Índice de la comunidad de programación TIOBE	32
23	Índice de la comunidad de programación TIOBE	32
24	Entorno Google Colab	35
25	Google Colab	36
26	Microsoft Azure en Ignite	36
27	Entorno Jupyter Notebook	37

28	Entorno Spyder	38
29	Entorno Pycharm	39
30	Niveles de Investigación	43
31	Tipos de Estudio	44
32	Proceso de la Metodología Cuantitativa	45
33	Edad de los encuestados	47
34	Genero de los encuestados	48
35	Residentes de la Zona 8	49
36	Importancia de conocer síntomas de Covid	50
37	Vacunas de Covid	51
38	Información sobre el Covid-19	52
39	Conocimientos sobre Hábitos Saludables para pacientes contagiados de Covid-19	53
40	Posesión de Smartphone de los Habitantes	54
41	Conocimientos sobre la Inteligencia Artificial	55
42	Importancia de herramientas tecnológicas	56
43	Importación de datos a Google Colab	58
44	Dataset subida a Google Colab	59
45	Revisión rápida de la Dataset	59
46	Tipo de dato que contiene el Dataset	60
47	Consulta de Missing Values - NaN	60
48	Gráfico de valores perdidos	60
49	Definición de función para quitar las tildes, convertir de mayúscula a minúscula con la función .lower, eliminación de caracteres especiales y datos basura que se hayan pasado del texto	61
50	Comparación de las columnas de texto	61
51	Separación de variables a utilizar del Dataset	62
52	Datos que contiene x	62
53	Datos que contiene y	62
54	Datos sin StopWords	62
55	Frecuencia de palabras	63
56	Palabras diferentes recibidas	63

57	Aplicación de Tokenización	64
58	Arquitectura del modelo Básico ANN	65
59	Entrenamiento de la Red ANN	66
60	accuracy del modelo básico ANN	66
61	loss del modelo básico ANN	67
62	Dataframe evaluación_métricas	67
63	Métricas del modelo efectivo	68
64	df_metrics_rfl	68

Anexos

N°	Descripción	Pág.
1.	Entrevista realizada a los Profesionales de IA y afines	71
2.	Modelo de la encuesta realizada a la población de la zona 8 de la provincia del Guayas	80



**ANEXO XIII.- RESUMEN DEL TRABAJO DE
TITULACIÓN (ESPAÑOL)
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA INGENIERÍA EN TELEINFORMÁTICA**



“EVALUACIÓN DE CONTENIDO DE ENTRADA Y POSIBLE SALIDA EN CONVERSACIONES TEXTUALES DE PERSONAS CONTAGIADAS DE COVID-19 PARA IDENTIFICAR UN MODELO PLN EFECTIVO POR MEDIO DEL ENTRENAMIENTO DE MACHINE LEARNING DEL FCI 010-2021”

Autor: Mirian Estefanía Solórzano Monserrate

Tutor: Ing. Comp. Acosta Guzmán Iván Leonel. MSIG.

Resumen

El text mining es el proceso de examinar grandes volúmenes de documentos para descubrir nueva información o ayudar a responder preguntas de investigación específicas. Este ha generado un gran aporte tecnológico dando lugar a la creación de portales donde era compartida la información detallada de la situación que pasaban diversos países en la pandemia, tomando información de páginas web oficiales donde publicaban los contagios por Covid-19 y fallecidos actualizados permitiendo de esta manera a la población conocer cómo se expandía a diario el virus y así estar mayormente informada de nuevas variantes y posibles olas de contagio.

En este trabajo de titulación desarrollo diversos modelos como son Long-Short Term Memory (LSTM), modelo básico Artificial Neural Network (ANN), Random Forest para multilabel y KNeighborsClassifier (kNN) en los cuales se pretende proporcionar un módulo aplicando algoritmos supervisados con arquitecturas NLP con la finalidad de detectar patrones de recomendaciones más comunes dadas por los médicos especialistas del Ecuador a los pacientes contagiados del virus Covid-19 en la fase de acompañamiento médico o post Covid más efectivas dadas por ellos entre el mes de marzo del año 2020 a marzo del año 2022. Llegando a la conclusión que el modelo básico ANN con los parámetros ajustados llega a ser el más efectivo dando los siguientes resultados accuracy train de 88%, una precisión train de 94%, recall train 92% y auc train 98%, seguido de los datos de test accuracy test 75%, precisión test 82%, recall test 78% y auc test de 93%.

Palabras claves: recomendaciones, virus, algoritmos, NLP, modulo.



**ANEXO XIV.- RESUMEN DEL TRABAJO DE
TITULACIÓN (INGLÉS)
FACULTAD DE INGENIERÍA INDUSTRIAL
CARRERA INGENIERÍA EN TELEINFORMÁTICA**



“EVALUATION OF INPUT CONTENT AND POSSIBLE OUTPUT IN TEXTUAL CONVERSATIONS OF COVID-19 INFECTED PERSONS TO IDENTIFY AN EFFECTIVE NLP MODEL BY MEANS OF MACHINE LEARNING TRAINING OF THE FCI 010-2021”

Author: Mirian Estefanía Solórzano Monserrate

Advisor: Ing. Comp. Acosta Guzman Ivan Leonel. MSIG.

Abstract

Text mining is the process of examining large volumes of documents to discover new information or help answer specific research questions. This has generated a great technological contribution leading to the creation of portals where detailed information on the situation in different countries during the pandemic was shared, taking information from official web pages where they published updated Covid-19 infections and deaths, thus allowing the population to know how the virus was spreading daily and thus be better informed of new variants and possible waves of infection.

In this degree work I developed several models such as Long-Short Term Memory (LSTM), Artificial Neural Network (ANN) basic model, Random Forest for multilabel and KNeighborsClassifier (kNN) in which it is intended to provide a module applying supervised algorithms with NLP architectures in order to detect patterns of most common recommendations given by medical specialists in Ecuador to patients infected with the Covid-19 virus in the phase of medical monitoring or post Covid most effective given by them between March 2020 to March 2022. It was concluded that the basic ANN model with the adjusted parameters was the most effective, giving the following results: accuracy train 88%, accuracy train 94%, recall train 92% and auc train 98%, followed by test accuracy test 75%, test accuracy 82%, recall test 78% and auc test 93%.

Keywords: recommendations, virus, algorithms, NLP, module.

Introducción

El NLP es un componente del text mining que comprende el proceso de adquirir datos significativos partiendo de un texto de lenguaje natural generado por humanos, el cual está encargado de realizar un análisis lingüístico que ayuda a sistemas o maquinas a procesar el texto que llega utilizando una metodología de comprensión diferente a la de un humano. (Subramanian, 2019)

Este ha generado un gran aporte tecnológico para impartir datos sobre la pandemia del Covid-19 dando lugar a la creación de portales donde era compartida la información detallada de la situación que pasaban diversos países tomando información de páginas web oficiales donde publicaban los contagios y fallecidos actualizados permitiendo de esta manera a la población conocer cómo se expandía a diario el virus y así este mayormente informada de nuevas variantes y posibles olas de contagio.

De acuerdo con Torres A. J., (2021) Destaca la necesidad de analizar el material disponible en estas redes sociales para entender personalidades, relaciones humanas y las condiciones sociales realizando un análisis de opinión de tuits generados en Ecuador que tienen relación con el COVID-19 en el año 2020 para la gestión de los datos y para el descubrimiento de patrones ocultos en el conjunto de datos.

El presente trabajo de investigación busca proporcionar un módulo aplicando algoritmos supervisados con arquitecturas NLP con la finalidad de detectar patrones de recomendaciones más comunes dadas por los médicos especialistas del Ecuador a los pacientes contagiados del virus Covid-19 en la fase de acompañamiento médico o post Covid para superar este virus.

De esta manera descubrir mediante el modelo NLP de forma cuantitativa las recomendaciones más efectivas dadas por los médicos entre el mes de marzo del año 2020 a marzo del año 2022. Y a su vez esta herramienta sirva de aporte como consulta a la comunidad médica, de manera que estas conversaciones entre médicos y pacientes pueda realizarse la detección de esquemas de recomendaciones y así contribuya a diversos médicos cuyos pacientes presenten casos similares con síntomas de las diferentes variantes presentes y futuras del Covid-19.

Capítulo I

El problema

1.1 Planteamiento del problema

A inicios del año 2020 la insuficiente información referente al virus Covid-19 impacto al punto que muchas personas se contagiaran, las autoridades públicas tienen la necesidad de proveer a toda la población información correcta, en los actuales momentos a pesar de que existe mucha información disponible referente a actualización de casos, medidas de protección y prevención, documentos normativos proporcionados por el ministerio de salud pública del Ecuador (Ecuador, 2021), se destaca la necesidad de analizar el material disponible que se encuentra en redes sociales para entender personalidades, relaciones humanas y las condiciones sociales realizando un análisis de recomendaciones generados por médicos en Ecuador que tienen relación con el COVID-19.

El NLP es un componente del text mining que comprende el proceso de adquirir datos significativos partiendo de un texto de lenguaje natural generado por humanos, el cual está encargado de realizar un análisis lingüístico que ayuda a sistemas o maquinas a procesar el texto que llega utilizando una metodología de comprensión diferente a la de un humano. (Subramanian, 2019)

En PLN hay dos modelos de lenguaje principales utilizados: modelos lógicos basados en gramática y modelos probabilísticos basados en datos. (Moreno S. A., 2017) los cuales aún no cuentan con un procesamiento de lenguaje natural para mantener una conversación fluida, proporcionando la respuesta que necesitan los usuarios de la zona 8 de la provincia del Guayas, cantones Guayaquil, Duran y Samborondón.

1.2 Formulación del problema

¿La evaluación de contenido e identificación de un modelo PLN ayudará a los chat-bots actuales a optimizar sus respuestas de manera efectiva con las diferentes interrogantes que presenten los usuarios de la zona 8 respecto al virus Covid-19?

1.3 Sistematización del problema

A continuación se mostrará la sistematización del problema:

¿Cuán difícil será desarrollar un modelo de machine learning orientado a Procesamiento de lenguaje Natural (PLN o NLP)?

¿Cuán factible es el desarrollo e identificación de un modelo de machine learning orientado a PLN?

1.4 Objetivos

1.4.1 Objetivo General

Evaluar los contenidos de entrada y salida de las conversaciones textuales de personas contagiadas de Covid-19 para la identificación de un modelo efectivo basados en Machine Learning.

1.4.2 Objetivos Específico

- Levantar información bibliográfica de la evolución de Covid-19, algoritmos y herramientas disponibles para conocer las interacciones del procesamiento del lenguaje natural (PLN).
- Alimentar la data set creada para utilizar en el modelo de entrenamiento por medio de los grupos focales on line de un sector vulnerable de la zona 8.
- Preparar el dataset mediante carga, limpieza, depuración y etiquetado de datos previo al entrenamiento del modelo.
- Preparar opciones de arquitectura de modelo y realizar las pruebas respectivas.
- Medir los resultados del modelo efectivo por medio del entrenamiento de Machine Learning para los contenidos de entrada y posible salida de la data set construida.

1.5 Justificación

La problemática ha dado origen a la necesidad de suministrar soluciones informáticas para ello se propone evaluar el contenido en conversaciones textuales de personas contagiadas, mediante la creación de varios modelos de procesamiento de lenguaje natural a través de la utilización de algoritmo de Machine Learning, se identificará el modelo más idóneo, este algoritmo ayudara a la población de la zona 8 a poder contar con una alternativa para realizar consultas relacionadas con el Covid-19 permitiendo obtener respuestas confiables para el manejo de situaciones de contagio o posible contagio.

1.6 Delimitación del problema

Tabla 1 Delimitación del problema

Campo	Aplicación de tecnología de la información
Área	Tecnología de los ordenadores
Aspecto	Modelo PLN efectivo por medio del entrenamiento de machine learning.
Tema	Evaluación de contenido de entrada y posible salida en conversaciones textuales de personas contagiadas de covid-19 para identificar un modelo PLN efectivo por medio del entrenamiento de machine learning del FCI 010-2021.

Elaborado por: Solórzano Monserrate Mirian

La realización del presente trabajo evaluara las entradas Xs que den un impacto significativo a las salidas Ys mediante modelos de PLN utilizando el algoritmo de Machine Learning de esta manera se analizará el modelo que presente las predicciones más acertadas, dando un beneficio a la población de la zona 8 de la provincia del Guayas aplicándolo a un asistente virtual que presente un mayor porcentaje de predicción.

1.7 Alcance

La presente investigación se desarrollará mediante investigación, encuesta, entrevista a la población de la provincia del Guayas, Zona 8 los cantones Guayaquil, Duran y Samborondón para posteriormente alimentar la base de datos.

La cual va dirigida a personas adultas de una zona contagiadas de covid-19, y realizar las pruebas de entrenamiento, se va a generar un modelo de machine learning mediante la configuración de un modelo el cual va a modificarse para tener diversas propuestas de arquitectura de modelo y se van a evaluar cual es la que da mejores resultados en cuanto a calidad de predicción usando el lenguaje de programación Python.

No se considera en el presente alcance la puesta en un servidor de producción por cuanto no se cuenta con esa infraestructura disponible a nivel de la carrera de Ing. Telemática.

1.8 Premisa de la investigación

La evaluación de contenido e identificación de un modelo efectivo utilizando entrenamiento de machine learning aportara a los asistentes virtuales actualmente utilizados una mayor precisión en cuanto a respuesta a los usuarios de la zona 8 sobre el Covid-19.

1.9 Operacionalización de Variables

En PLN existen modelos principales utilizados los cuales aún no cuentan con un procesamiento de lenguaje natural para mantener una conversación fluida, proporcionando la respuesta optima que necesita el usuario de la zona 8 y de esta manera contribuir a que todos estén informados sobre el virus Covid-19 y sus futuras variantes.

1.9.1 Variable Independiente:

En la presente investigación se halló una variable independiente, la actual se detalla a continuación:

- Frases de las personas

1.9.2 Variable dependiente:

- Predicción de Recomendaciones Médicas

Capítulo II

Marco teórico

2.1. Antecedentes del estudio

La expansión de las soluciones de IAC, como asistentes virtuales y asistentes de voz, es fundamental tanto para empresas como para personas naturales. (Marc, 2020)

A raíz de la pandemia de Covid-19 que llegó al país el 29 de febrero de 2020, el distanciamiento social que a fuerza se tenía que llevar a cabo y más que todo el confinamiento de las personas en cada uno de sus hogares, llevaron a diversos desarrolladores a generar soluciones informáticas las cuales den una herramienta de apoyo a las personas contagiadas de este virus y a la población en general; dando lugar a los chat-bots el cual está creado como un agente de servicio automatizado o un agente de recomendación de productos que puede ayudar a los usuarios a aprender las características del producto, la usabilidad del producto, la aplicabilidad, la resolución de problemas y otras tareas relacionadas. (Mutiwokuziva et al., 2017)

La Inteligencia Artificial Conversacional (I.A.C) toma cada vez mayor importancia ya que proporciona una interfaz simple para la interacción persona y computadora. Por su enorme potencial y atractivo valor comercial como asistente virtual, asistente virtual y / o chatbot social. (Yan, 2018)

De acuerdo con González, Estrada, & Febles, 2018 al emplear la tecnología de I.A en el análisis de padecimientos, se utiliza en investigaciones complejas para lograr un grado de certeza reconocido. Entre los resultados obtenidos en términos de identificación de resultados de enfermedades específicas, la aplicación ha logrado resultados favorables debido a la simplicidad de la investigación sistemática que puede aprender y mejorar diferentes conjuntos de datos en el proceso de clasificación y predicción de enfermedades.

Las redes neuronales se basan en datos, pueden encontrar relaciones (patrones) de manera inductiva a través de algoritmos de aprendizaje basados en datos existentes, sin la ayuda de modeladores para especificar formas funcionales y sus interacciones. (Pacela, Semeraro, & Anglania, 2004)

El uso de redes neuronales es diverso aplicado en el ámbito de la medicina como en otras ramas para analizar, dar diagnósticos, en imágenes y en el área de farmacología las cuales están inspiradas en el comportamiento de la razón de una persona, las RNA son una rama

significativa en la inteligencia artificial. El objetivo es intentar encontrar modelos que resuelvan problemas de difícil resolución mediante técnicas de algoritmos tradicionales.

2.2. Fundamentación teórica

2.2.1 Inicios de la IA.

Es de gran importancia conocer el origen de la IA durante la historia, dado a que esta rama dio el desarrollo de diversas técnicas de manejo del conocimiento y a través de eso se realizaron diferentes aportes tecnológicos y llevaron a un gran progreso a la I. A., además surgieron nuevas áreas como son la percepción y el lenguaje natural.

Según datos publicados de National Geographic, (2020) indica una breve historia de la IA a través de los años la cual se muestra en la **Tabla 1**.

Tabla 2. Historia de la Inteligencia Artificial

Año	Detalle
1854	Una lógica matemática , el matemático George Boole en este año por primera vez argumento que el razonamiento lógico sería posible estructurarse de forma que resolviera sistema de ecuaciones.
1921	Idea de un robot , en la obra de teatro R.U.R de Karek Apek usa el término “robot” el cual en muchas lenguas esclavas de acuerdo con su etimología significa “trabajo duro”
1936	Concepto de algoritmo , Alan Turing el cual fue considerado el padre de la computación moderna publicó un artículo introduciendo el significado de algoritmo el cual constaba de números computables.
1941	Z3 , creada por Konrad Zuse siendo la primera computadora automática y programable. Ley de la robótica , las cuales son 3 y nacieron de un relato denominado “circulo vicioso” escrito por Isaac Asimov.
1950	Diferenciación hombre-maquina , Alan Turing plantea un ensayo que llevo por título Computing Machinery an intelligence el que fue llamado el Test de Turing.
1956	Nacimiento del término Inteligencia Artificial , durante una conferencia del informático John McCarthy indica el termino de Inteligencia artificial la cual fue estimada como el germen de la disciplina.
1957	Se imito a una mente , diseñada por primera vez una red neuronal artificial por Frank Rosenblatt.
1966	Se da voz a las computadoras mediante ELIZA , siendo así el chatbot principal del mundo creado por Joseph Weizenbaum, incorporando el nlp.
1969	Perceptrones , redactado por el cofundador del MIT Marvin Minsky.
1996	Deep Blue , la cual fue creada por IBM una supercomputadora.

- 1979 Cart de Stanford**, siendo el primer vehículo que recorrió con éxito en una pista de obstáculos de manera automática
- 2005 Maquinas más inteligentes que el hombre**, Raymond Kurzweil empleando la ley de moore logro predecir que las maquinas llegarían a un nivel de inteligencia mayor en 2029 y para el 2045 superarían la inteligencia de las personas.
-
- 2012 Deep learning el verdadero poder**, se crea un superordenador el cual está capacitado de instruirse mediante YouTube a reconocer caras, cuerpos humanos y gatos.
- 2014 Test de Turing superado por una IA**, un Bot capacitado engañando a 30 de los 150 jueces a los cuales fueron sometido durante el Test de Turing haciendo que creyeran que hablaban con un niño de 13 años.
- 2015 AlphaGo**, primera máquina que le gano a un competidor profesional de Go.

Elaborado por: Solórzano Monserrate Mirian. Información tomada de National Geographic.

Según Berzal, (2016) indica lo que se ha desarrollado en los últimos años hasta la actualidad

Tabla 3. Historia de la Inteligencia Artificial

Año	Detalle
2017	Libratus , desarrollado por la Universidad Carnegie Mellon el algoritmo que venció a 4 de los mejores jugadores profesionales de póquer.
2018	En sectores importantes del tejido productivo se asienta la inteligencia artificial.
2019	La utilización de soluciones informáticas basadas en inteligencia artificial se vio con mayor intensidad.
2020	Dado a la situación de la pandemia el desarrollo de la IA son significativos en el ámbito de la salud y se da el crecimiento del uso del “Low Code”.
2021	

Elaborado por: Solórzano Monserrate Mirian. Información tomada de un trabajo de investigación de la web

2.2.2 Inteligencia Artificial

La I. A. es el talento que tienen las máquinas para aprender de los datos y luego realizar una toma de decisiones a través de lo que ha aprendido como lo haría una persona, proporcionando una diferencia entre humano - máquina que estas no necesitan tomar descanso y pueden llegar a analizar grandes paquetes de datos a la vez. (Rouhiainen, 2018)

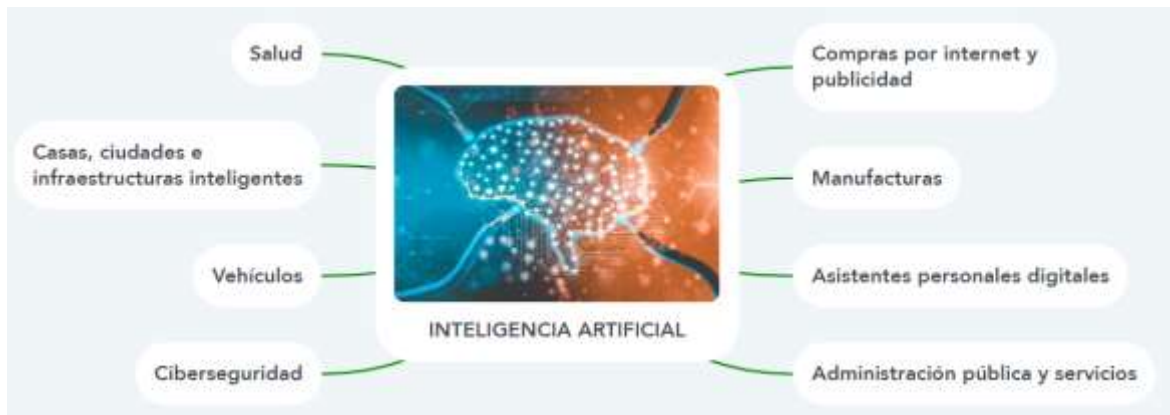


Figura 1. Ejemplos de dónde podría utilizarse la inteligencia artificial. Elaborado por Solórzano Monserrate Mirian

Según Boden, (2016) indica que la IA tiene como principal objetivo realizar lo mismo que haría una mente humana. Además la IA ha hecho posible que sea aplicada a diversas ramas dando lugar al desarrollo de diferentes teorías.

Es un tipo de inteligencia que puede mostrar varios artefactos creados por humanos y generalmente se menciona en los sistemas informáticos. Pero también se refiere al campo de la investigación científica destinada a crear un entorno que utilice dicha inteligencia. (Redacción APD, 2021)

La IA pretende llegar a resolver diversos problemas informáticos realizando un aporte significativo a diferentes ramas de la ciencia aplicadas desde una inteligencia artificial débil hasta una inteligencia artificial fuerte.

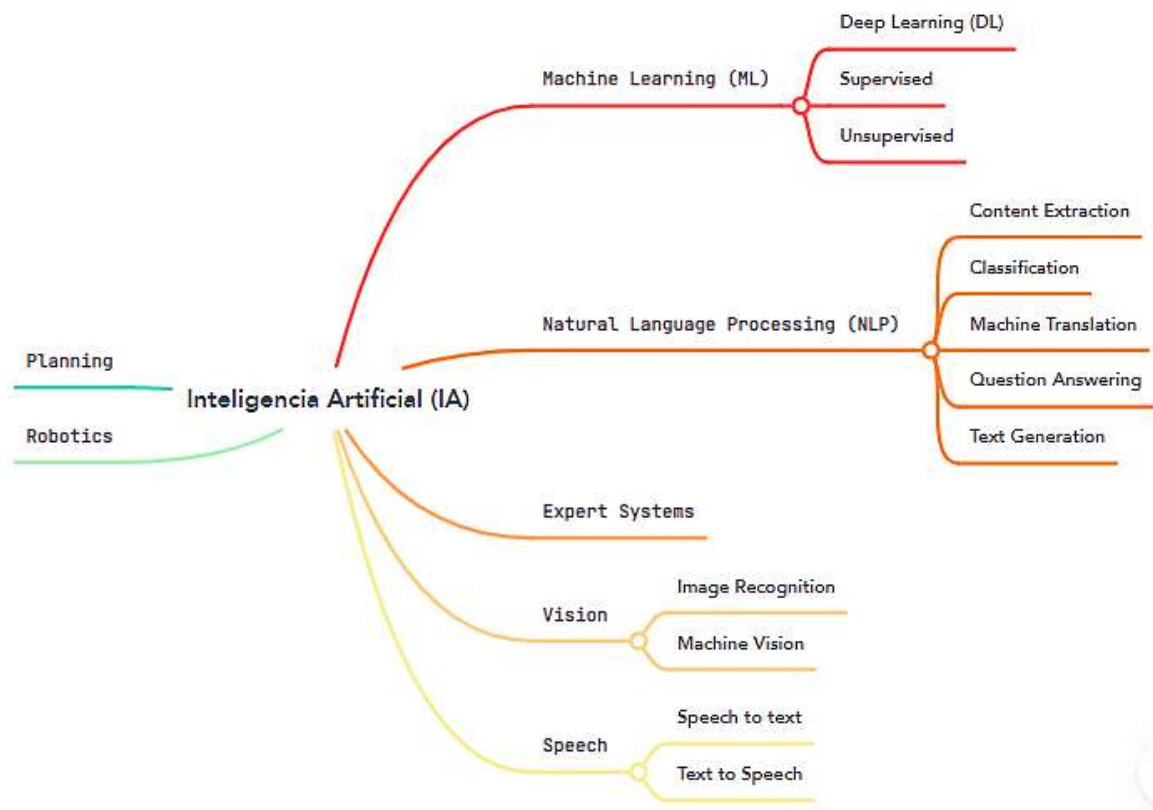


Figura 2. Mapa mental de la Clasificación de la Inteligencia Artificial.
Elaborado por Solórzano Monserrate Mirian

2.2.3 Aprendizaje automático (Machine Learning)

Según Lugo, Maldonado, & Murata, (2014) indican que el Aprendizaje Automático con sus siglas en inglés ML Machine Learning, la información procesada la cual ha aprendido dando así un reconocimiento de patrones explorando un sin número de datos y descubrir previamente relaciones de estos sin que sea necesario una hipótesis.

2.2.3.1 Ciclo de vida del ML

Son ciclos iterativos entre la mejora de datos, el modelado y la evaluación que en realidad nunca terminan. Este ciclo es crítico para desarrollar modelos de ML porque se enfoca en usar los resultados y la evaluación del modelo para optimizar su conjunto de datos. Estos datos de alta calidad es la manera más segura de entrenar un modelo de máxima calidad. La velocidad de iterar de este ciclo determina el costo, favorablemente existen herramientas que pueden ayudar a aligerar el ciclo sin sacrificar la calidad. (Hofesmann, 2021)

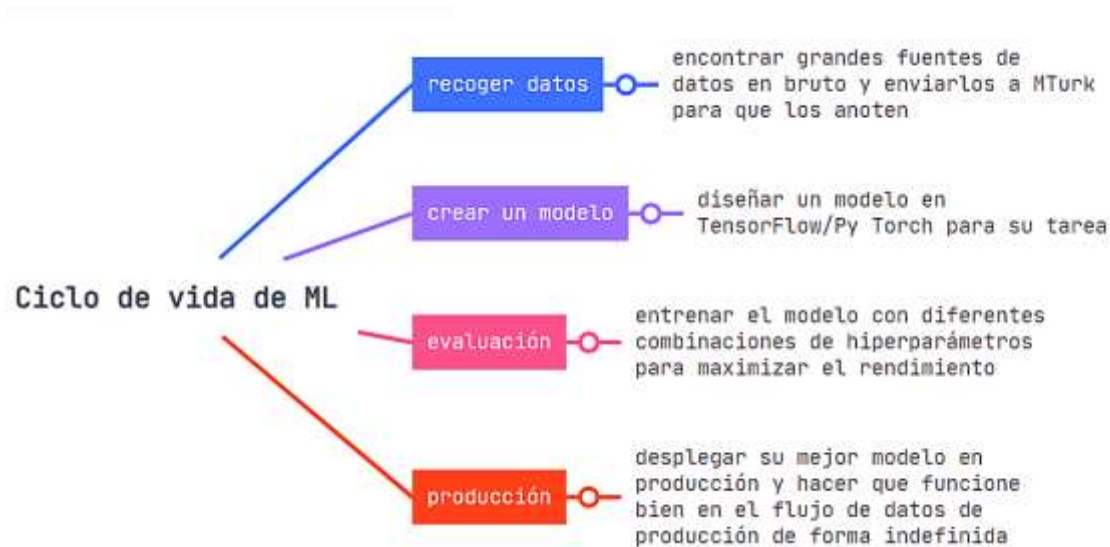


Figura 3. Ciclo de vida de ML. Elaborado por Solórzano Monserrate Mirian

2.2.3.2 Niveles de Madurez de ML

Entre los niveles se tiene desde la evaluación de los casos de uso hasta los modelos más sofisticados puestos en producción durante 5 años y más, los cuales se detallan a continuación en la figura 4.

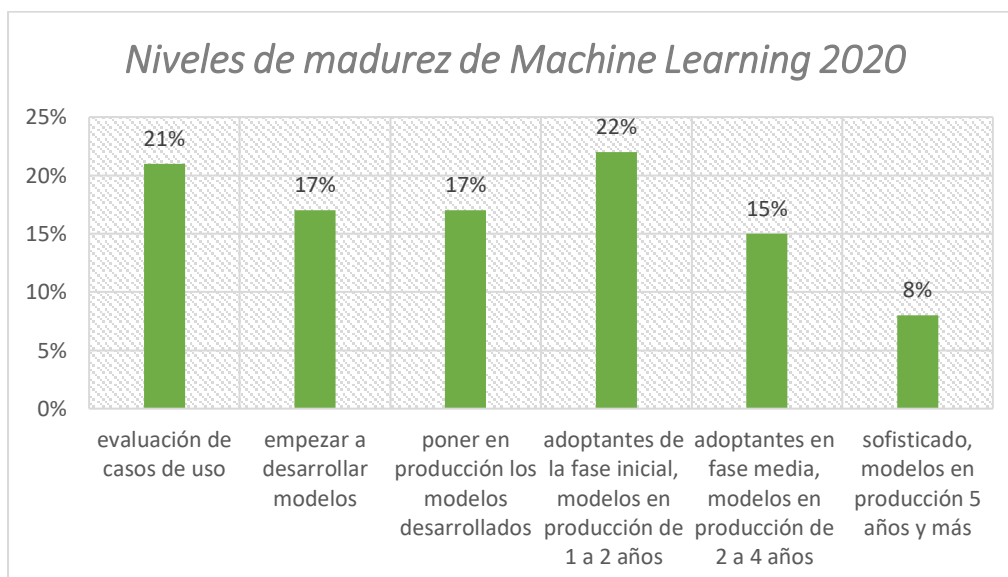


Figura 4. Información tomada de un artículo Niveles de madurez del ML. Elaborado por Solórzano Monserrate Mirian

2.2.3.3 Librerías usadas en ML

2.2.3.3.1 Pandas

Pandas es un paquete de Python que proporciona estructuras de datos rápidas, alto rendimiento, estructuras de datos flexibles y confiables y herramientas de análisis de datos. El diseño expresivo facilita la labor con información que pueden ser "relacionales" o "etiquetados" y su manejo es intuitivo. Está diseñado para ser un bloque de construcción avanzado para

efectuar análisis de estos datos. Mundo práctico y real en Python. (Martínez, Cruz, y López, 2021)

2.2.3.3.2 Numpy

Una biblioteca de funciones matemáticas usando operaciones matriciales. Numpy significa "Numerical Python" y es la biblioteca informática científica líder que proporciona grandes estructuras de datos. (Moreno F. S., 2020)

2.2.3.3.3 SciPy

De acuerdo con Moreno F. S., (2020) indica que esta librería es un paquete científico, incluye bibliotecas de matemáticas, ciencias e ingeniería basadas en la ciencia. Es una recopilación de cajas de herramientas específicas de dominio y algoritmos numéricos, que contienen proceso de señales, estadísticas, optimización y más.

2.2.3.3.4 Matplotlib

Esta es una librería apta para generar vistas estáticas e interactivas en Python. Lo que permite crear gráficos con calidad de publicación con solo unas pocas líneas de código, se tiene control total sobre el estilo del gráfico, colores, las propiedades del eje. (Moreno F. S., 2020)

2.2.3.3.5 Tensor Flow

TensorFlow es una librería creada por Google para entrenar y construir RN. Influenciado por Theano, fue el predecesor de DistBelief, esta es otra biblioteca establecida por Google para la producción e investigación de los propios productos de Google. La arquitectura que maneja es flexible, es decir que puede ejecutarse en muchas plataformas (GPU, CPU o TPU), así como en computadoras, clústeres de servidores e incluso dispositivos móviles. (Conde, 2018)

2.2.4 Aprendizaje Profundo (Deep learning)

El Aprendizaje Profundo con sus siglas en inglés DL Deep Learning, consiste en aprender representaciones de datos con diversos niveles de abstracción y permite que estos modelos computacionales compuestos de varias capas de procesamiento mejoren de manera drástica el estado de la técnica de reconocimiento de voz, detección de objetos, entre otros dominios. (LeCun, Bengio, & Hinton, 2015)

2.2.5 Tipos de Aprendizaje

A continuación se detalla la clasificación de los tipos de aprendizaje existentes en la *Figura 3*.

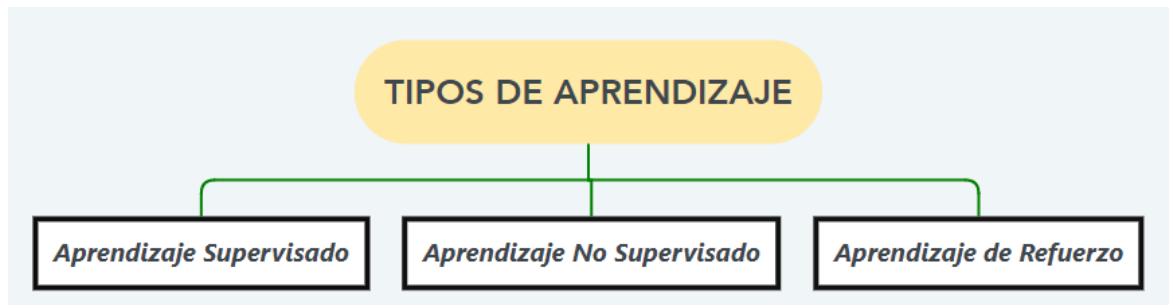


Figura 5. Clasificación de los tipos de Aprendizaje. Elaborado por Solórzano Monserrate Mirian

2.2.5.1 Aprendizaje Supervisado.

Los algoritmos utilizan datos previamente etiquetados u organizados para indicar cómo se debe clasificar la información nueva. El uso de este método requiere la intervención humana para proporcionar retroalimentación. (Rouhiainen, 2018)

2.2.5.2 Aprendizaje No Supervisado.

El algoritmo no utiliza ningún dato previamente etiquetado u organizado para indicar cómo clasificar nueva información, pero debe averiguar cómo clasificarla por sí mismo. Por tanto, este método no requiere intervención manual. (Rouhiainen, 2018)

2.2.5.2.1 Clustering (K-means)

En particular, el algoritmo de K-means es un método de partición, donde cada grupo está representado por un centro (centroide), que es la media de los puntos de datos en el grupo. La idea es clasificar dónde los objetos dentro del mismo grupo son lo más similares posible (alta cohesión dentro de la clase) y los objetos dentro de diferentes grupos son lo más diferentes posible (baja correlación entre clases). (González, Varela, y Sandra, 2017)

Según Chavez, (2020) la agrupación en clústeres de K-means es un método de agrupación no supervisado, que se usa más comúnmente para dividir un conjunto de datos determinado en k grupos, donde k representa la cantidad de grupos especificados previamente por el analista. En el agrupamiento de K-means, cada conglomerado está representado por su centro o baricentro, que corresponde a la media de los puntos asignados al conglomerado.

2.2.5.3 Aprendizaje Por refuerzo.

Los algoritmos aprenden de la experiencia. En otras palabras, debemos darles un "refuerzo positivo" cada vez que hagan lo correcto. El estilo de aprendizaje de estos algoritmos se puede comparar con el estilo de aprendizaje de un perro que aprende a sentarse. (Rouhiainen, 2018)

2.2.6 Algoritmos de aprendizaje supervisado

2.2.6.1 Regresión Lineal

Esta basada en analizar los cambios en una variable la cual no es aleatoria y se asigna una variable de estas variables en caso de que exista una relacion funcional entre estas variables se representa mediante una funcion lineal como se muestra en la **Figura 4**.

La estructura del modelo de Regresion Lineal Simple es el siguiente:

$$y = wx + b$$

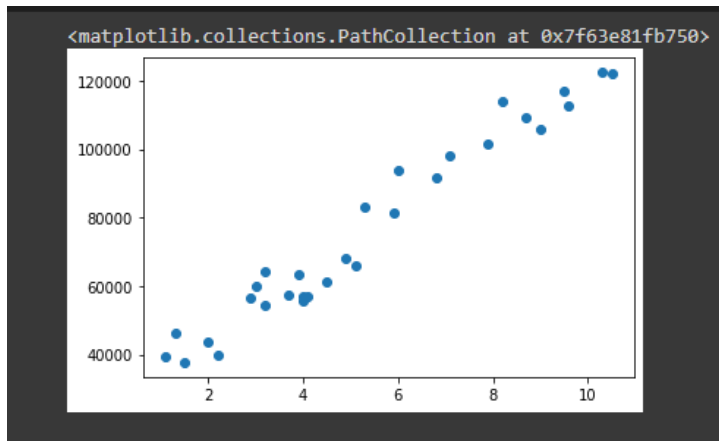


Figura 6. Representación de la Regresión lineal.
Elaborado por Solórzano Monserrate Mirian

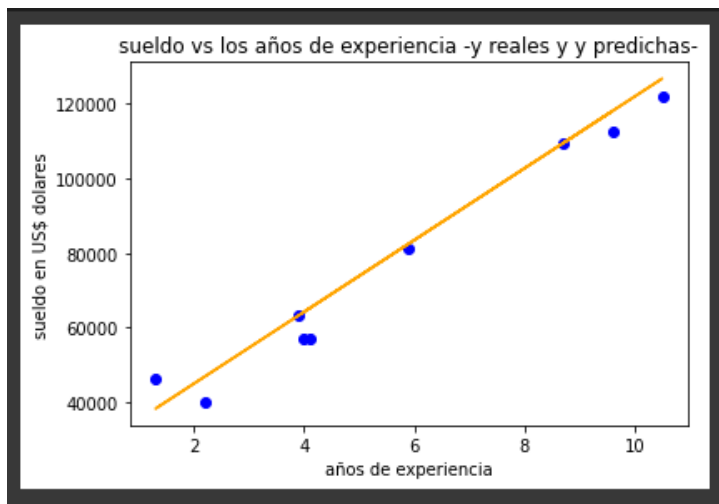


Figura 7. Representación de la Regresión lineal con una recta.
Elaborado por Solórzano Monserrate Mirian

2.2.6.2 Regresión Polinomial

La Regresion Polinomial es una regresion lineal a diferencia que en esta se calcula atributos polinomicos con lo que ya se conoce de la RL.

La estructura del modelo de Regresion Polinomial es el siguiente:

$$x = [a + bx + cx^2 + \dots + Nx^n]$$

2.2.6.3 Regresión de soporte vectorial (SVR)

La máquina de Soporte Vectorial se usa para resolver problemas multidimensionales. Con la misma estrategia de clasificación, puede estar cerca de la regresión. La idea general es usar un hiperplano regresor que se ajuste mejor al conjunto de datos de entrenamiento, de modo que no haya No hay clases para separar. A su vez, todos los conjuntos de vectores de soporte tienen solo un lado del hiperplano. (Delgado & Ochoa, 2018)

Pasos para la creación de un modelo en Machine Learning

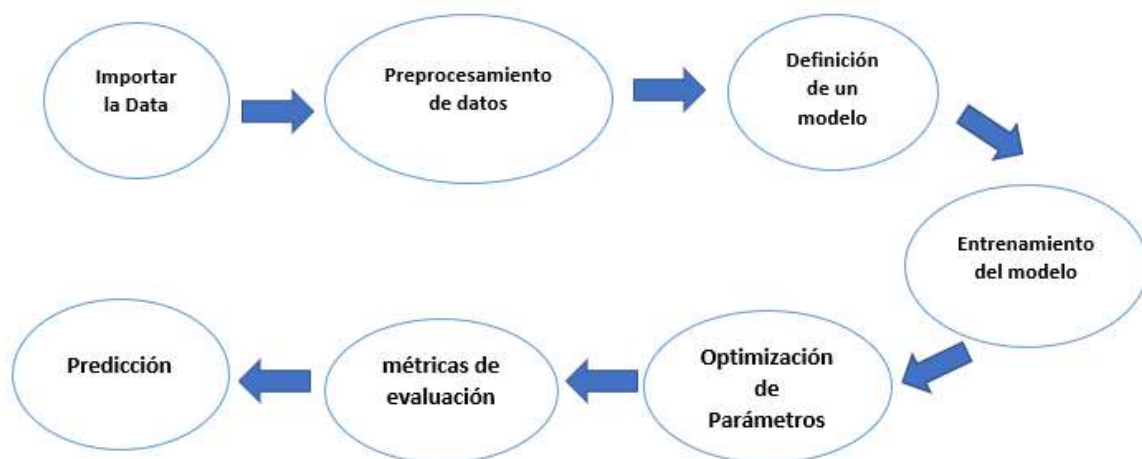


Figura 8. Pasos para creación de un modelo. Elaborado por Solórzano Monserrate Mirian

2.2.7 Datasets

Existen Datasets disponibles en la web entre los cuales están Kaggle Dataset, Dataset Search, SQuAD en ellos tienen datos almacenados para realizar evaluaciones y además contienen códigos para proporcionar un mayor entendimiento del procesamiento que conlleva cada análisis efectuado en este conjunto de datos.

Kaggle Datasets Explora, comparte y analiza datos de calidad y se obtiene más información sobre los tipos de datos, la creación y la colaboración.

Dataset Search Este conjunto de datos contiene información sobre brotes de enfermedades causados por un nuevo virus de la familia Coronaviridae llamado SARS-CoV-2. La enfermedad provocada por este nuevo virus ha sido denominada por consenso internacional como COVID-19.

SQuAD es un conjunto de datos de comprensión de lectura que consta de preguntas planteadas por un colaborador a un conjunto de artículos, contiene más de 100000 pares de preguntas y respuestas en más de 500 artículos.

2.2.8 Natural language processing (NLP)

El Procesamiento de Lenguaje Natural con sus siglas en ingles NLP comenzó en la década de 1950 y es la intersección de la inteligencia artificial y la lingüística, era originalmente diferente de la recuperación de información textual, que utilizaba técnicas basadas en estadísticas altamente escalables para indexar y buscar de manera eficiente grandes cantidades de texto. (Prakash y Ohno-Machado, 2011)

Alias & Cassanelli, (2019) indica que el NLP es una rama de la IA, el propósito de esta es procesar y entender datos del lenguaje, pretende la interacción entre las computadoras y el lenguaje humano, comúnmente designado como lenguaje natural. Característicamente, se llega a crear programas o aplicaciones capaces de analizar grandes cantidades de datos en dicho lenguaje.

De acuerdo con Zhao, et al (2020) el PLN o NLP es un conjunto de técnicas computacionales basadas en la teoría para analizar y representar texto natural en uno o más niveles de análisis lingüístico con el fin de lograr un procesamiento del lenguaje similar al humano para una variedad de tareas o aplicaciones. En esta definición, el concepto "niveles de análisis del lenguaje" se refiere al análisis fonológico, morfológico, léxico, sintáctico, semántico, discursivo y pragmático del lenguaje.

Sobre la base de que los humanos suelen utilizar todos estos niveles para generar o comprender el lenguaje, los sistemas de NLP pueden admitir diferentes combinaciones de niveles o niveles de análisis lingüístico. (Zhao, et al, 2020)

Según Lloret, (2019) indica que el NLP es una manera en que las computadoras comprendan, analicen y obtengan el significado del lenguaje humano de una forma útil e inteligente, con NLP, los programadores logran desarrollar y organizar instrucciones para realizar labores como la síntesis automática, el reconocimiento de voz, la traducción, la extracción de relaciones, el reconocimiento de entidades nombradas, el análisis de opiniones y la segmentación de temas.

El NLP es un campo de investigación interdisciplinario que se basa en el conocimiento de la lingüística, la informática y muchas otras disciplinas relacionadas que han desarrollado métodos para analizar datos del lenguaje humano. (McKenzie & Adams, 2021)

Recientemente, el campo se ha desplazado al uso del ML, incluido el DL para el entrenamiento de algoritmos aplicando métodos y técnicas modelando su proceso para lograr un rendimiento efectivo.

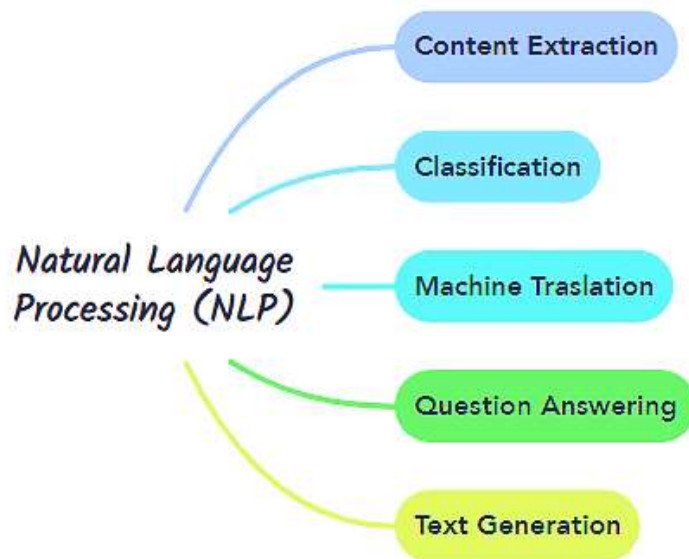


Figura 9. Mapa mental de la Clasificación de NLP. Elaborado por Solórzano Monserrate Mirian

2.2.8.1 Librerías utilizadas en NLP

2.2.8.1.1 Natural Language Toolkit (NLTK)

Es la plataforma líder para crear programas Python con datos de lenguaje humano, dando una interfaz fácil de usar con más de cincuenta cuerpos y medios léxicos como WordNet y un serie de bibliotecas de procesamiento de texto para la clasificación, tokenización, análisis sintáctico y razonamiento semántico. Este fue creado con el propósito de efectuar los siguientes objetivos: simplicidad, consistencia, extensibilidad, modularidad. (Camacho y Navarro, 2020)

2.2.8.1.2 Scikit Learn (SKLEARN)

Es una biblioteca gratuita de software de aprendizaje automático en lenguaje Python que contiene varias utilidades para trabajar con conjuntos de datos, algoritmos de clasificación, agrupamiento, regresión y más. Contiene un clasificador que utiliza técnicas de poda de conjunto es decir como un árbol de soluciones para optimizar el tiempo de clasificación y la precisión de los metaestimadores, como los bosques aleatorios o el embolsado. (Messina, 2018)

Esta librería contiene diversos módulos útiles que son: conjuntos de datos, preprocesamiento, métricas, selección de modelos, selección de características, agrupación, multiclase entre otros.

2.2.8.2 Pre-Procesamiento de texto

2.2.8.2.1 Lista de pasos para el pre-procesamiento de texto

Para el pre-procesamiento manual del texto se quitan etiquetas HTML, elimina espacios en blanco adicionales, convertir caracteres acentuados en caracteres ASCII, expande las contracciones, quita caracteres especiales, minúscula todo el texto, convierte números a palabras numéricas, quita números, quita palabras vacías. (Weng, 2019)

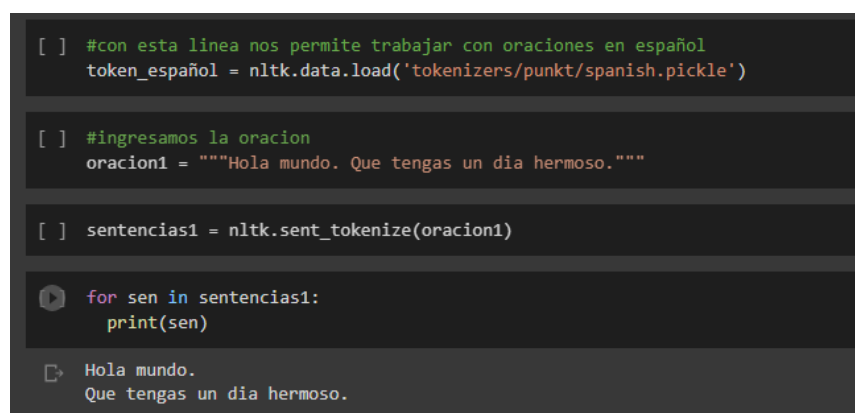
De acuerdo con Weng, (2019) muestra que el esquema general del preprocesamiento de texto existe tres principales componentes que son: Tokenización, Lematización, Eliminación de Ruido.

2.2.8.2.1.1 Tokenización

La primera etapa del NLP incluye la identificación de tokens, que son unidades básicas que no necesitan descomponerse en el procesamiento posterior. La palabra entidad es una muestra de NLP, la más básica. Sin embargo, lo que nos importa es el uso de computadoras para identificar aquellas etiquetas que no tienen separadores diferentes y expresiones fijas. (Webster & Kit, 1992)

Se cree que al tomar modismos y expresiones fijas como una unidad básica al mismo nivel que las palabras, la tokenización debería ganar un significado más universal y realista, haciendo que los sistemas de PNL y TM sean más robustos y prácticos. (Webster & Kit, 1992)

La Tokenización además de ser el primer paso para la clasificación de las palabras o frases es el nivel más importante ya que aquí inicia el moldeamiento de lo que será las entradas del algoritmo a entrenar.



```
[ ] #con esta línea nos permite trabajar con oraciones en español
    token_español = nltk.data.load('tokenizers/punkt/spanish.pickle')

[ ] #ingresamos la oracion
    oracion1 = """Hola mundo. Que tengas un día hermoso."""

[ ] sentencias1 = nltk.sent_tokenize(oracion1)

for sen in sentencias1:
    print(sen)

Hola mundo.
Que tengas un día hermoso.
```

Figura 10. Aplicación de un token a una oración en Google Colaboraty. Elaborado por Solórzano Monserrate Mirian

2.2.8.2.1.2 STEMMING o Derivación

Es el proceso de reducir las inflexiones o cambio de una palabra a su forma de raíz, es como asignar un grupo de palabras a la misma raíz, incluso si la raíz no es una palabra válida en el idioma. (Jabeen, 2018)

Esta se puede realizar mediante la librería NLTK la cual mientras procesa tareas de PLN va encontrando varios escenarios en los que estarán varias palabras de igual raíz y aquí es donde se aplica el stemming.



```
[8] import nltk

[11] from nltk.stem.porter import *

[12] stemmer = PorterStemmer()

[15] tokens = ['compute', 'computer', 'computed', 'computing']

for token in tokens:
    print(token + ' --> ' + stemmer.stem(token))

compute --> comput
computer --> comput
computed --> comput
computing --> comput
```

Figura 11. Aplicación de un stemming. Elaborado por Solórzano Monserrate Mirian

Algoritmos de STEMMING

De acuerdo con Molina, (2018) indica que entre los principales algoritmos de Stemming son los siguientes:

Truncado

Es el menos complejo de los algoritmos de Stemming, el cual toma los primeros n caracteres como raíz y el resto se elimina. Si el valor de n es mayor o igual que la longitud de la palabra, la palabra en cuestión no cambia. Para valores más pequeños de n, se producen más errores de lematización, mientras que valores de n superiores a 7 aumentan el número de errores de lematización.

Lovins

Fue el más efectivo y popular de los algoritmos de Stemming, Este elimina el sufijo más largo de la palabra, una vez que se eliminan las terminaciones de las palabras, se realizan varios ajustes a la palabra usando diferentes tablas para garantizar que la raíz sea una palabra válida. Por su naturaleza de recorrer el corpus una sola vez, siempre se eliminan un máximo de 15 sufijos de cada palabra.

Posters

Basado en la idea de los sufijos ingleses, hay alrededor de 1200, en su mayoría compuestos por sufijos más pequeños y simples. Contiene cinco pasos y en cada uno se aplica reglas hasta que una de ellas este cumplida. Si la regla es admitida el sufijo es anulado de acuerdo con la misma y se realiza el siguiente paso. Al final del último paso se genera el stem resultante.

Paice/Husk

Este algoritmo es un núcleo iterativo cuya tabla contiene 120 reglas indexadas, por cada última letra del sufijo en cada iteración, intenta encontrar la regla adecuada en función de la última letra de la palabra. Cada regla especifica el final de una eliminación o cambio, Si ninguna regla es válida, el algoritmo termina.

Dawson

Este Stemming es una extensión del algoritmo de Lovins, excepto que incluye una lista de alrededor de 1200 sufijos. Al igual que Lovins, solo viaja a través del corpus una vez, por lo que es muy rápido. Los sufijos se almacenan de manera invertida a su longitud y última letra. En realidad, están organizados como un conjunto de ramas de árboles de caracteres para un acceso rápido.

N-gram

Son series de elementos tomados de un texto, documento, publicación. En este caso se da que la letra N representa el número de elementos que se deben tomar a consideración, en otras palabras es el tamaño de la secuencia del grama. Este modelo se puede usar en varias aplicaciones, como secuenciación de ADN, reconocimiento de voz, detección en errores de ortografía, entre otros. Dentro de este existen los uni-gramas que está conformado por un elemento que pueden ser lemas o palabras, bi-gramas que está formado por 2 gramas, tri-gramas compuesto por 3 gramas y así sucesivamente. (Sidorov, 2013)

2.2.8.2.1.3 Lematización

Se puede realizar mediante la librería Spacy, para realizarlo se debe usar el atributo lemma_ que viene incluido en el paquete de la librería antes mencionada.



```
[1] import spacy

[2] sp = spacy.load('en_core_web_sm')

[3] sentence6 = sp(u'compute computer computed computing')

#Podemos encontrar las raíces de todas las palabras usando la lematización espacial de la siguiente manera
for word in sentence6:
    print(word.text, word.lemma_)

compute compute
computer computer
computed compute
computing computing
```

Figura 12. Aplicación de Lematización. Elaborado por Solórzano Monserrate Mirian

2.2.8.2.1.4 Normalización de palabras

Un enfoque común para lidiar con la normalización de palabras es usar un diccionario para designar las palabras que se van a convertir. Este proceso incluye operaciones como separación de sílabas, combinación de raíces, corrección ortográfica y la conversión de ortografías raras o alteradas en vocabulario de uso común. (Hamdy, 2021)

2.2.8.2.1.5 Eliminación de signos de puntuación y caracteres especiales

Los caracteres no alfanuméricos a menudo no agregan ningún valor a la comprensión del texto y generan ruido que afecta negativamente el rendimiento del modelo de aprendizaje automático. Es decir limpiar los datos de texto de estos caracteres suele resultar de provecho para la precisión y la velocidad del modelo y presentar mayor precisión y velocidad de este. (Hamdy, 2021)

2.2.8.2.1.6 Transformación de símbolos especiales

Esto implica reemplazar emojis con representaciones de texto que pueden mejorar mucho el contexto y cambiar el significado de una oración o confirmar la intención del autor. También es importante investigar la relación entre la emoción de una oración y los símbolos especiales, como los emojis. (Hamdy, 2021)

2.2.8.2.1.7 Representación en bolsa de palabras (BOW, Bag Of Words)

Considerado un método eficiente y popular para representar texto en actividades de recuperación de información es la representación en bolsa de palabras o (BoW). Cuando se usa esta representación, se descarta la estructura del texto de entrada, como capítulos, párrafos, oraciones y formato, se cuenta las veces que aparece cada palabra en cada texto del corpus.

Descartar la estructura del texto y solo contar las ocurrencias de palabras produce una imagen mental que representa el texto como una "bolsa". (Díaz, 2019)

2.2.8.3 Técnicas en NLP

Una técnica de NLP es un método práctico, un enfoque, un proceso o un procedimiento para realizar una tarea de NLP concreta, como el etiquetado POS, el análisis sintáctico o la tokenización. Una biblioteca de software que soporta una o más técnicas de NLP, como Stanford CoreNLP, NLTK u OpenNLP. (Zhao, et al, 2020)

Las técnicas de NLP se refieren a diferentes áreas de investigación y desarrollo de aplicaciones de inteligencia artificial. (Redacción APD, 2021)

2.2.8.4 Modelos en NLP

En PLN hay dos modelos de lenguaje principales utilizados: modelos lógicos basados en gramática y modelos probabilísticos basados en datos. (Moreno S. A., 2017)

Teniendo en cuenta la creciente popularidad del deep learning (DL) para manejar problemas de clasificación, han aparecido diversos modelos de DL con arquitecturas complejas. (Khasanah, 2021)

2.2.8.4.1 Meta Embeddings

El objetivo es combinar la información obtenida por incrustaciones de palabras generadas por diferentes métodos existentes y diferentes fuentes de información durante el entrenamiento. Esto trata de conseguir Mejoras en la calidad de incrustación de palabras. (García, 2018)

2.2.8.4.2 Word Embeddings

La incrustación de palabras es un enfoque semántico distribuido para representar palabras como vectores reales. Esta representación tiene propiedades de agrupación útiles, ya que agrupa palabras similares semántica y sintácticamente. Los modelos de espacio vectorial se han utilizado para la semántica distribuida desde la década de 1990. Desde entonces, se han desarrollado diferentes modelos para estimar representaciones continuas de palabras, como el Análisis Semántico Latente (LSA). (García, 2018)

Estas son replantaciones vectoriales de palabras las cuales son de gran utilidad para variedades de tareas en relación con el NLP y contienen una arquitectura de redes neuronales las cuales están basadas en enfoques actuales.

2.2.8.4.3 Word2Vec

Word2vec no es ni el primero ni el último ni el mejor cuando se trata de espacios vectoriales, incrustaciones, escalado, métricas de similitud, etc. Pero word2vec es muy simple y accesible, a menudo juega un papel de apoyo relativamente menor en estas tareas, en gran medida cerrando la brecha entre la entrada ascii y el formato de entrada más adecuado para las redes neuronales. (Iglesia, 2016)

Se ha demostrado que la representación vectorial de las palabras adquiridas utilizando el modelo word2vec es lingüísticamente útil en varias tareas de NLP. A medida que más y más investigadores quieren experimentar con word2vec o técnicas similares, he notado una falta de literatura que detalle el proceso de aprendizaje de los parámetros del modelo de incrustación de palabras, lo que imposibilita que los no expertos en redes neuronales puedan entender cómo estos modelos funcionan. (Rong, 2016)

Según García, (2018) Word2Vec es un modelo predictivo para generar incrustaciones de palabras e implementa dos modelos neuronales: CBOW y Skip-gram. En CBOW dado el contexto de la palabra objetivo, intenta predecirla y en Skip-gram se intenta predecir el contexto dado a una palabra.

INSPECTOR VISUAL DE INCRUSTACIÓN DE PALABRAS

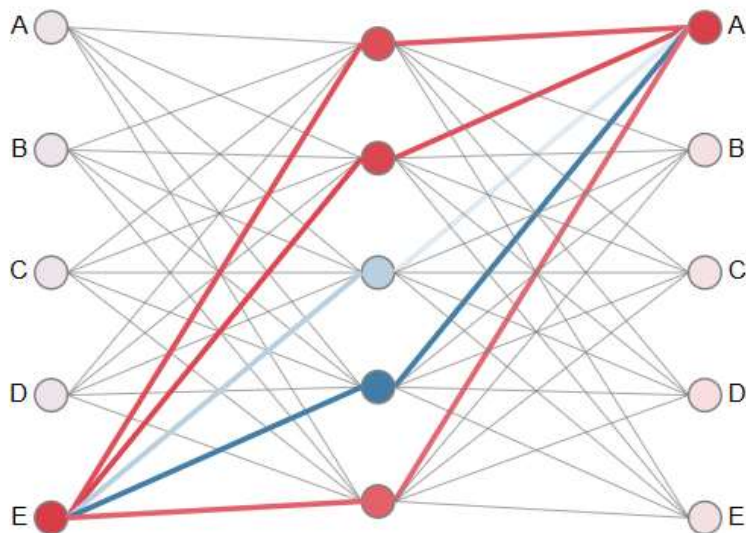


Figura 13. Numero de neuronas utilizadas. Información tomada del artículo *Explicación de los parámetros de word2vec*

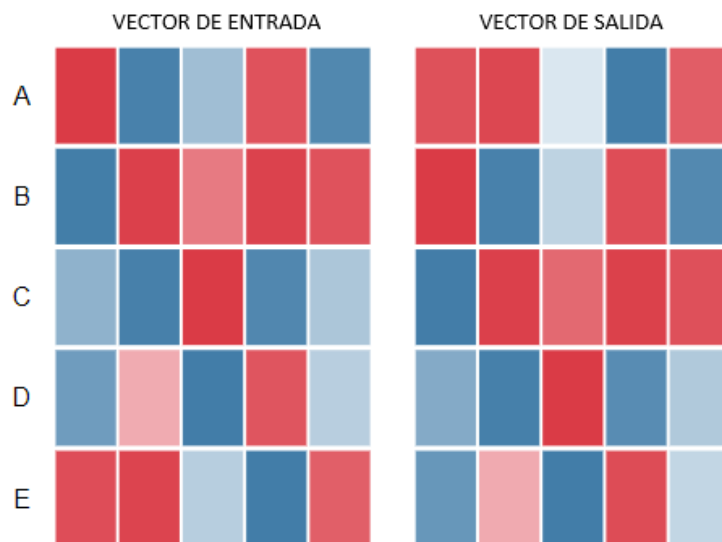


Figura 14. Matrices de peso del Inspector Visual de palabras
 Información tomada del artículo Explicación de los parámetros de word2vec

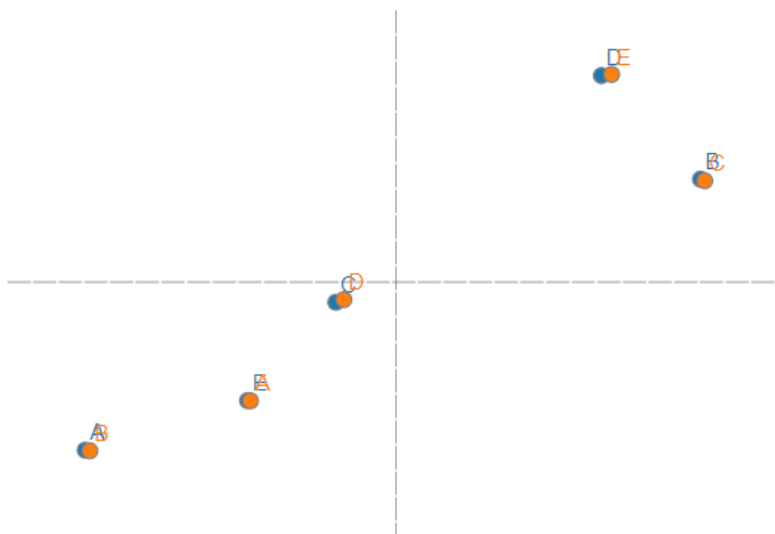


Figura 15. Vectores resultantes del Inspector Visual de palabras
 Información tomada del artículo Explicación de los parámetros de word2vec

2.2.8.4.4 Global Vectors for Word Representation (GloVe)

Este a diferencia de Word2Vec, es un modelo basado en conteo. GloVe genera una gran matriz donde la información concurrencia entre palabras y contextos. Por lo tanto, para cada palabra, contamos el número de veces que aparece esa palabra en algún contexto. El objetivo de entrenamiento de la matriz es aprender vectores tales que el producto escalar entre palabras sea igual al logaritmo de la probabilidad de coocurrencia entre palabras. (García, 2018)

En otras palabras la complejidad computacional del modelo depende del número de elementos de la matriz X , como este número es menor que el número total de entradas de la matriz.

2.2.8.4.5 FastText

FastText es una extensión del modelo Word2Vec. Cada palabra se trata como la suma de sus caracteres, lo que se denomina ngrama. El vector de una palabra está dado por la suma de sus ngramas. (García, 2018)

Es ligeramente inferior en cuanto a la precisión pero supera al estado de la técnica con un margen en términos de compromiso entre uso de memoria y precisión.

ARQUITECTURA DEL MODELO FastText

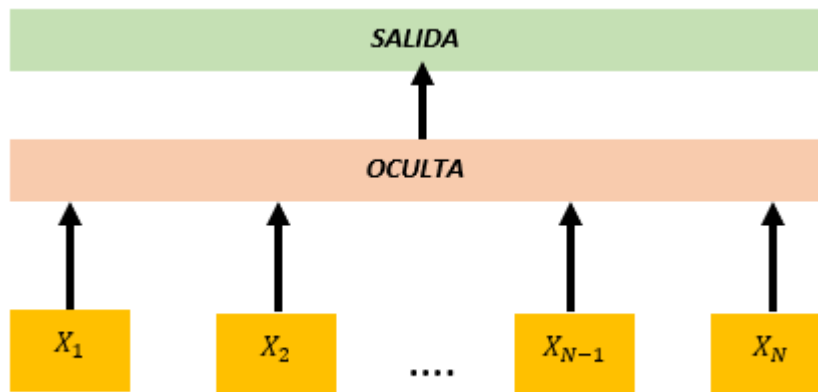


Figura 16. Arquitectura del modelo FastText. Información tomada de Google. Elaborado por Solórzano Monserrate Mirian

2.2.8.4.6 Transformer

Wang, et al, (2019) mencionan que Transformer es el modelo de última generación en las últimas evaluaciones de traducción automática. Hay varios estudios que prometen mejorar las pruebas de dichos modelos: el primero utiliza una amplia gama de redes y ha sido el estándar o norma para el desarrollo de los sistemas Transformer, el otro utiliza representaciones de lenguaje más profundas, pero enfrenta los desafíos de aprender redes profundas. Este modelo es una opción a las RNN la cual está basada en la atención que ha logrado resultados de vanguardia en una serie de tareas de NLP.

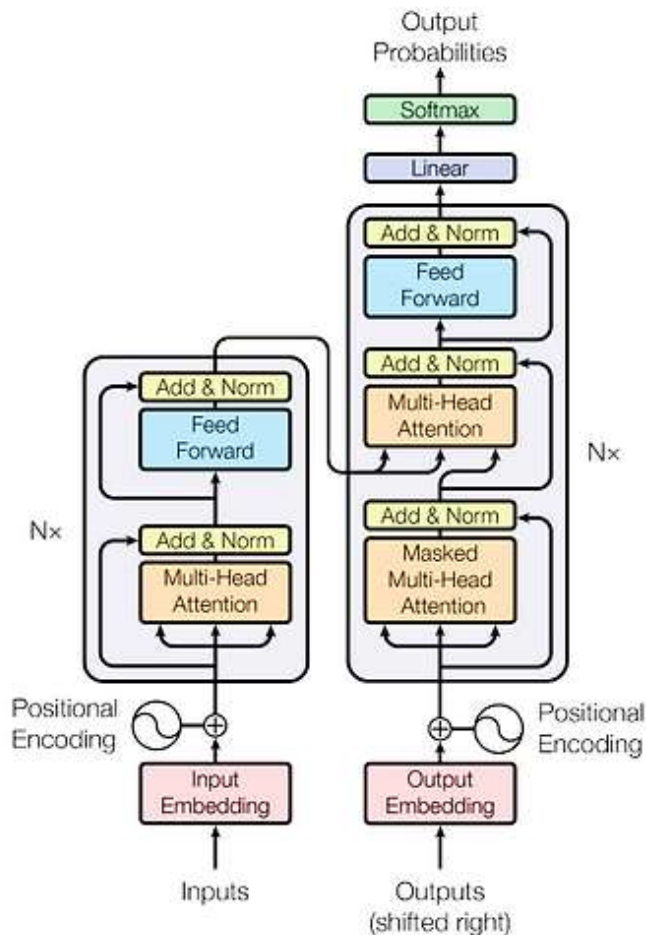


Figura 17. Arquitectura del modelo Transformer. Información tomada de Google. Elaborado por el autor.

2.2.8.4.7 BlazingText

BlazingText facilita implementaciones mayormente optimizadas del modelo Word2vec simple pero muy rápido y eficiente presenta algoritmos de clasificación de texto basado en fastText. (Hamdy, 2021)

2.2.8.4.8 Embeddings from Language Models (ELMo)

En ELMo, las representaciones de palabras de la misma oración se vuelven más similares a medida que aumenta la especificidad del contexto superior. Es decir este genera representaciones contextuales de cada símbolo al vincular los estados internos de un biLSTM entrenado en dos capas en una tarea de modelado lingüístico bidireccional. (Matthew, et al, 2018)

ARQUITECTURA DEL MODELO ELMo

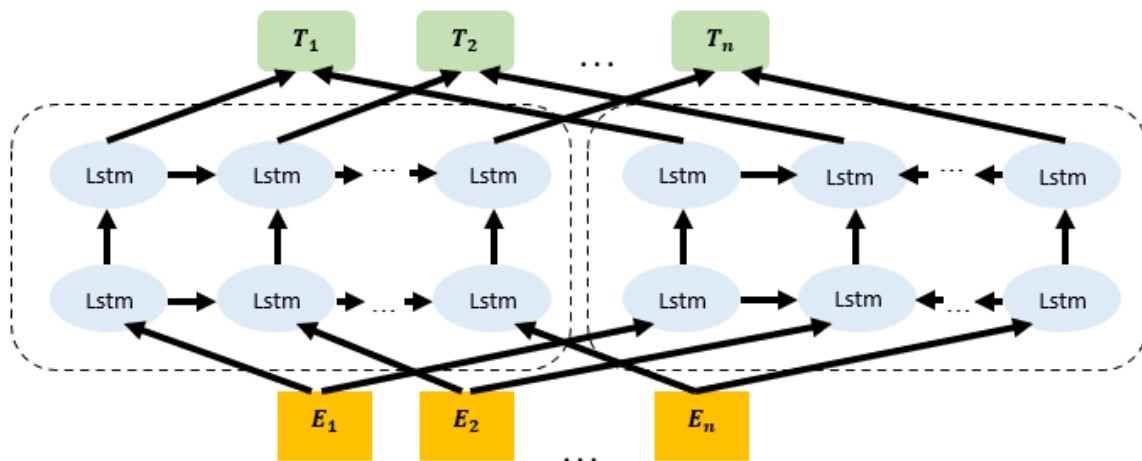


Figura 18. Arquitectura del modelo ELMo. Información tomada de Google.
Elaborado por Solórzano Monserrate Mirian

2.2.8.4.9 GPT-n

GPT-2 es un decodificador de transformador apilado que introduce una secuencia de marcadores y aplica incrustaciones de posición y marcador, seguido de varias capas de decodificación. Cada capa combina una red de avance, normalización de capas y conexiones residuales para hacer cumplir el autocuidado de varios cabezales. El modelo GPT-2 tiene 12 capas y 12 cabezales. (Vig & Belinkov, 2019)

ARQUITECTURA DEL MODELO GPT

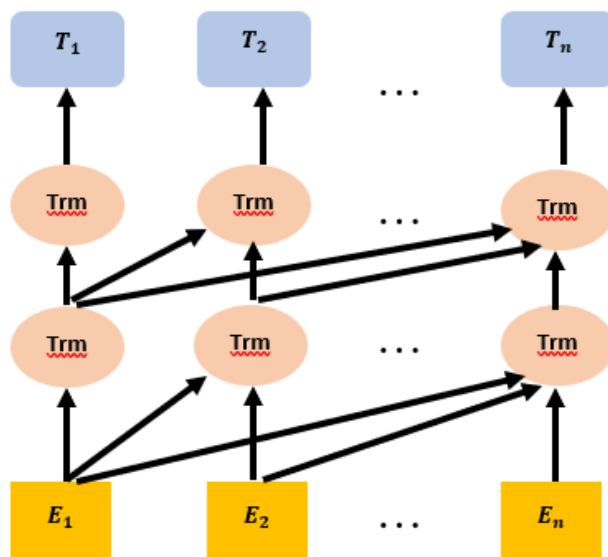


Figura 19. Arquitectura del modelo GPT Información tomada de Google.
Elaborado por Solórzano Monserrate Mirian

2.2.8.4.10 BERT

BERT está diseñado para entrenar previamente representaciones de dos profundidades a partir de texto sin marcar, con contexto izquierdo y derecho establecido en todas las capas. Por lo tanto, un modelo BERT pre-entrenado se puede adaptar con solo una capa de salida adicional para crear modelos modernos para una alta escala de labores, como responder preguntas y analizar datos. Pregunta, respuesta y razonamiento lingüístico, sin realizar cambios significativos en la estructura de tareas específicas. (Devlin, Chang, & Lee, 2019)

En otras palabras, este modelo es una tecnología de la IA el cual permite que el sistema realice una mejor comprensión del lenguaje humano.

ARQUITECTURA DEL MODELO BERT

Devlin, Chang, & Lee, (2019) mencionan que la arquitectura utilizada de este modelo es un codificador bidireccional capas múltiples Transformer las cuales están basadas en la implementación original.

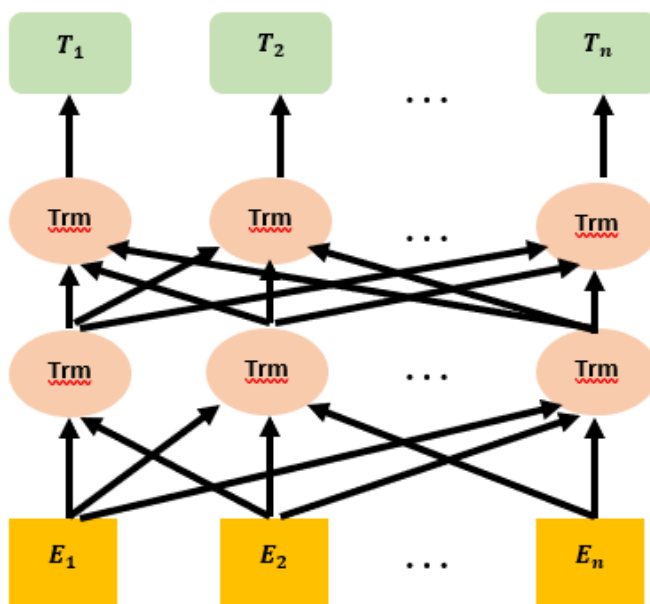


Figura 20. Arquitectura del modelo BERT Información tomada de la Información tomada de Google.
Elaborado por Solórzano Monserrate Mirian

2.2.9 Redes Neuronales Recurrentes – RNN

Las RNN son redes que permiten el procesamiento de datos. Es necesaria la gestión del contexto y se almacena el estado temporal, por lo que RNN se puede utilizar para resolver tareas relacionadas con los datos, donde el orden es importante. (Ordóñez & López, 2021)

Son denominadas de esta manera ya que cada neurona procesa un conjunto de datos de forma secuencial y luego utiliza los datos de salida de su predecesor como entrada. Ya que las

neuronas almacenan información a través de estados internos, este tipo de modelos tienen memoria al pasar del tiempo. Por esto, se dice que las RNN administran estados de bucle porque básicamente expanden la memoria para aprender de experiencias o información importante que sucedió hace mucho tiempo. (Ordóñez & López, 2021)

2.2.10 Web Scraping

Según Martínez, et al. (2019) indican que prácticamente, si los datos de una página web se copian y almacenan en una base de datos, se considera un proceso de extracción de datos. El Web Scraping está estrechamente relacionado con la indexación web, que utiliza robots para indexar información web y es una tecnología común utilizada por la mayoría de los buscadores. No obstante, este se enfoca más en la conversión de datos que no constan de estructura en la web.

2.2.10.1 Éticas en Web Scraping

La extracción de datos de sitios web es una motivación para aprender a programar, Pero una vez que adquiere la capacidad de recopilar grandes cantidades de datos en un corto período de tiempo, siempre surgen problemas éticos:

- ¿Puedo tomar estos datos?
- ¿Puedo volver a publicar estos datos?
- ¿Estoy sobrecargando los servidores del sitio web?
- ¿Para qué puedo usar estos datos?

Densmore, (2017) menciona que para un raspado ético se tiene que cumplir una serie de puntos:

- tener una API pública que proporcione los datos que se está buscando.
- Se proporciona una cadena de agente de usuario que explique las intenciones y proporcione una forma de que se contacte si tiene inquietudes.
- Guardar los datos que estrictamente se necesitan de la página.
- Nunca hacer pasar el contenido como propio.

2.2.11 Random Forest para multilabel o multi etiqueta

Random Forest es un modelo de clasificación de conjunto que consta de diferentes árboles de decisión. Se desarrolla embolsando y seleccionando variables aleatorias. El principio de construcción del árbol en RF es el mismo que el del árbol de decisión, que se basa en la división recursiva. En la partición recursiva, la ubicación exacta del punto de corte y la elección de las variables de partición dependen en gran medida de la distribución de las observaciones en la muestra de entrenamiento. (Wu, Gao, & Jiao., 2019)

2.2.12 Support Virtual Machine para multilabel o multietiqueta

Support Virtual Machine (SVM) originalmente inicio como un clasificador de etiqueta única de dos clases basado en la geometría de margen máximo. Desde entonces, las SVM binarias han sido ampliamente estudiadas y aplicadas con éxito en muchos campos diferentes. (Shajari, Hoda & Rangarajan, 2020)

La SVM realiza un aprendizaje con una superficie de decisión para dos puntos de entrada diferentes, como clase de clasificadores, la descripción dada por los datos del vector de soporte puede establecer un límite de decisión en torno al dominio de dato aprendido, con poco o ningún conocimiento de los datos fuera de este límite. Los datos asignados mediante de un núcleo gaussiano u otro tipo de núcleo a un espacio de características en un espacio dimensional superior donde se busca la máxima separación entre clases. (Betancourt, Gustavo, 2005).

2.2.13 K-Nearest Neighbors(KNN)

En este modelo k vecino cercano se realiza una clasificación midiendo la distancia que existe entre características, si la mayor parte de k muestras que muestren mayor similitud pertenecerán a una clase. La decisión que toma este algoritmo es que la muestra que toma de k la clasifica basándose en los elementos mas cercanos a esta característica o clase de muestra. (Zhu,. 2020)

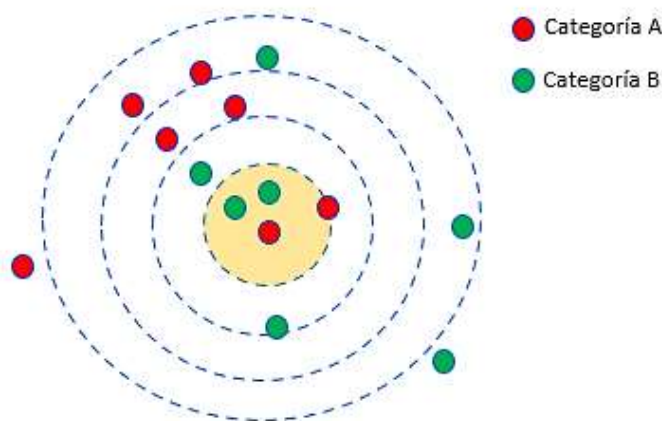


Figura 21. Algoritmo KNN. Información tomada de un trabajo de investigación de Data Science. Elaborado por Solórzano Monserrate Mirian

2.2.14 Long Short Term Memory (LSTM)

es un tipo de red neuronal recurrente que tiene como significado memoria a largo y corto plazo, pero es superior que la red neuronal recurrente habitual en el ámbito de memoria, estas tienen un control fino sobre el recuerdo de ciertos patrones y con mejor calidad. Como cualquier otro NN, un LSTM puede

tener varias capas ocultas a medida que pasa por cada capa, la información relevante se conserva y todos los demás datos irrelevante se descartan en cada unidad. (Shekhar, 2021)

2.2.15 Lenguajes y Herramientas en Inteligencia Artificial

En el ámbito de la IA existen varias herramientas las cuales ayudan al desarrollo de proyectos tanto a nivel universitario como empresarial, para de esta manera organizar algoritmos con diversas estructuras y combinatorias llegando a conseguir la capacidad de un ser humano.

2.2.15.1 Lenguajes de Programación

La tabla de la entidad de programación TIOBE es un cuadro del renombre de los lenguajes de programación esta se actualiza mensualmente. La evaluación está basada en la cantidad de ingenieros calificados, cursos y proveedores externos en todo el mundo. Los buscadores más populares como Wikipedia, Google, Amazon, Bing, YouTube entre otros son utilizados para computarizar las evaluaciones.

Dec 2021	Dec 2020	Change	Programming Language	Ratings	Change
1	3	▲	 Python	12.90%	+0.69%
2	1	▼	 C	11.80%	-4.69%
3	2	▼	 Java	10.12%	-2.41%
4	4		 C++	7.73%	+0.82%

Figura 22. Índice de la comunidad de programación TIOBE. Información tomada de. <https://www.tiobe.com/tiobe-index/>. Elaborado por TIOBE.



Figura 23. Índice de la comunidad de programación TIOBE. Información tomada de. <https://www.tiobe.com/tiobe-index/>. Elaborado por TIOBE.

Cabe indicar que esta tabla no se trata del mejor lenguaje de programación, ni es el lenguaje para editar la mayoría de las líneas de código.

2.2.15.1.1 Python



Python representa una herramienta de máximo nivel, multipropósito e interpretado. En los últimos tiempos la utilización ha incrementado de manera constante y actualmente es uno de los lenguajes de programación más utilizados para desarrollar sistemas de software. (Fernández, 2012)

La razón por la que es tan popular es su versatilidad y ser un lenguaje de código abierto, lo que lo hace fácil de entender e implementar. Además, Python cuenta con una gran cantidad de recursos, por lo que actualmente es la estrella del mundo de la programación. (Ortiz, 2021)

Según Challenger, Díaz, & Becerra, (2014) mencionan que Python tiene las funciones de programación orientada a objetos, imperativa y funcional, por lo que se considera un lenguaje de múltiples paradigmas y se basa en el lenguaje ABC.

La página oficial de Python se puede encontrar en www.python.org

Tabla 4. Características del Lenguaje Python

Características	Python
Lenguaje de código abierto	 
Multipropósito	
Versatilidad	
Extensa elección de biblioteca	
Sintaxis simple, fácil de entender e implementar	
Librerías para la Inteligencia artificial	
Multiplataforma	

Información tomada de la investigación. Elaborado por Solórzano Monserrate Mirian

Tabla 5. Versiones de Python destacando sus características

Versiones	Fecha de Lanzamiento	características
Versión 0. 9. 0	febrero de 1991	incluía clases con herencia, funciones y los tipos modulares
Versión 1. 0	enero de 1994	herramientas de la programación funcional
Versión 1. 6	septiembre de 2000	Presento inconvenientes con la licencia
Versión 2. 0	octubre de 2000	Incluyo generación de listas

Versión 1. 6. 1		Se trata de la misma versión que la 1.6, con la diferencia de una nueva licencia compatible con GPL
Versión 3.1	junio de 2009	Realizado dando importancia en eliminar constructores dobles y módulos
Versión 2.7	julio de 2010	Está centrada en mejoramiento del rendimiento del lenguaje y dar soporte a las versiones anteriores
Versión 3.7	junio de 2010	incluye un modo desarrollo en el intérprete para la mejora de la limpieza del código
Versión 3.8	octubre de 2019	Contiene nuevo operador de asignación con este puede devolver y establecer variables en línea
Versión 3.9	octubre de 2020	Se pueden fusionar diccionarios, contiene métodos para la eliminación de prefijos y sufijos
Versión 3.10.1	diciembre de 2021	La más actual corrigiendo errores de versiones anteriores

Información tomada de la página oficial de Python. Elaborado por Solórzano Monserrate Mirian

2.2.15.1.2 Python para Ciencia de Datos (DC)

En los últimos años, el análisis de datos ha cobrado una gran importancia como herramienta para tomar decisiones dentro de una organización. A través de este programa integral, los participantes aprenderán a usar el lenguaje Python para el análisis y la visualización de datos, utilizando las bibliotecas scikit-learn y matplotlib. Python es el lenguaje preeminente más utilizado por los analistas de datos de todo el mundo. (Nizama, 2020)

Usando Python, se puede analizar datos y hacer predicciones para hacer recomendaciones y así mejorar los procesos comerciales de diversos ámbitos.

2.2.15.1.3 C

Este lenguaje es uno de los más antiguos el cual se encuentra en uso en la actualidad, es estimado por la eficiencia del código que crea y es el más popular utilizado para crear software de sistema, sin embargo también se utiliza para la creación de aplicaciones en la actualidad; tiene una versatilidad que le ha permitido seguir siendo uno de los lenguajes más utilizados en todo tipo de desarrollo actualmente. A medida que se ha transformado de un lenguaje para el desarrollo específico de UNIX a un lenguaje de propósito general, y esta la razón por la cual tiene 50 años siendo utilizado. (Yañez, 2021)

Tabla 6. Características de Lenguaje C

Características	C
-----------------	---

Es un lenguaje estructurado

Programación de nivel medio

Flexible

No depende del hardware

Control absoluto de todo lo que pasa en la Pc

Acceso a memoria es de bajo nivel



Información tomada de la investigación. ¿Qué es lenguaje C? Elaborado por Solórzano Monserrate Mirian

2.2.15.1.4 Java

El desarrollo de este lenguaje de programación denominado Java ha sido muy rápida, esta plataforma de desarrollo ha ido ampliando y cada vez integra a mayores programadores en el mundo. Por otra parte este no solo es un lenguaje para programar, además un entorno de ejecución, plataforma de desarrollo y cuenta con librerías para desarrollo de programas sofisticados. (Martínez L. J., 2013)

Tabla 7. Características de Lenguaje Java

Características	Java
Portable	
Uso sencillo	
Multiplataforma	
Orientado a Objetos	
librerías para desarrollo de programas sofisticados	

Información tomada de la investigación. Elaborado por Solórzano Monserrate Mirian


2.2.15.1.5 C++

Según Jiménez, (2017) menciona que es un lenguaje de POO el cual tiene como base lenguaje C. A mediados de los años 80 fue diseñado por Bjarne Stroustrup.

C++ es un lenguaje de programación de nivel medio cabe recalcar que no es por ser con menor potencial que otros, sino que esta junta la programación estructurada de los lenguajes de alto nivel con la manera flexible del ensamblador, además es un lenguaje orientado a objetos

y lenguaje estructurado cuenta con las funciones y sentencias de control if, while, etc. (Osorio, 2006)

Tabla 8. Características de Lenguaje C++

Características	C++
Orientado a Objetos	
Sintaxis del lenguaje C	
Cuenta con las funciones y sentencias de control if, while, etc	
Flexible, Compila en C	

Información tomada de la investigación. Elaborado por Solórzano Monserrate Mirian

En cuanto a el análisis realizado en cuanto a los importantes lenguajes de programación expuestos por TIOBE se llega a la conclusión que el lenguaje Python es el más idóneo para efectuar los modelos de NLP dado a que cuenta con las librerías necesarias y algoritmos aplicados en la IA.

2.2.15.2 Entorno de Desarrollo Integrado (IDE)

Un IDE por sus siglas en inglés, es un programa informático formado por un conjunto de herramientas de programación. Este Puede dedicarse exclusivamente a un solo lenguaje de programación o puede utilizarse para varios. (Delgado C. , 2021)

2.2.15.2.1 IDE en la nube

2.2.15.2.1.1 Google Colaboraty

Es un entorno en la nube gratis el cual no solicita configuración alguna y es ejecutado o compilado netamente en línea. (De la Fuente, 2019)

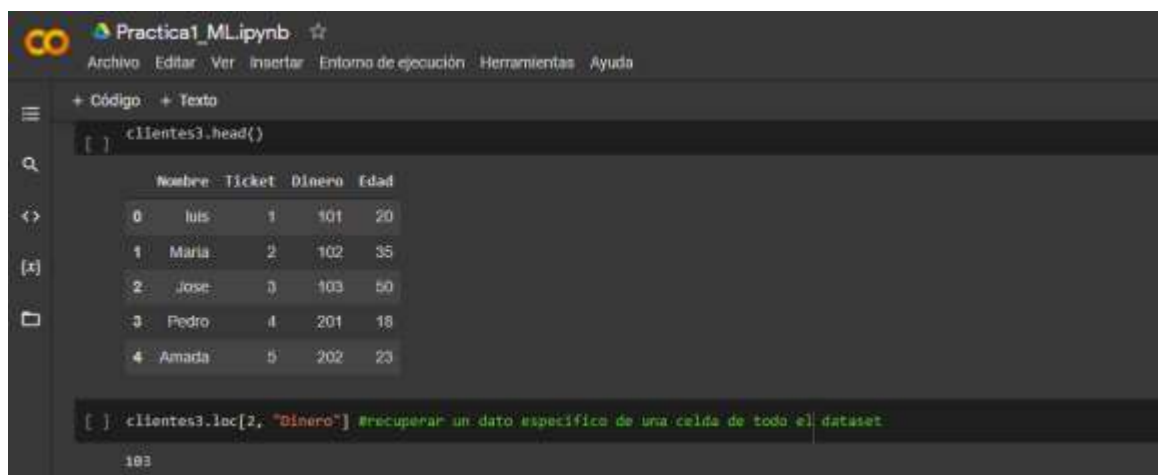


Figura 24. Entorno Google Colab. *Elaborado por Solórzano Monserrate Mirian*

Google Colab Ofrece 12 horas de ejecución continua, después de esto se elimina todo lo que se haya cargado y compilado en la nube, en el es posible trabajar con diversas instancias como CPU, GP Y TPU, en la Figura 11 se puede visualizar las especificaciones de cada una.

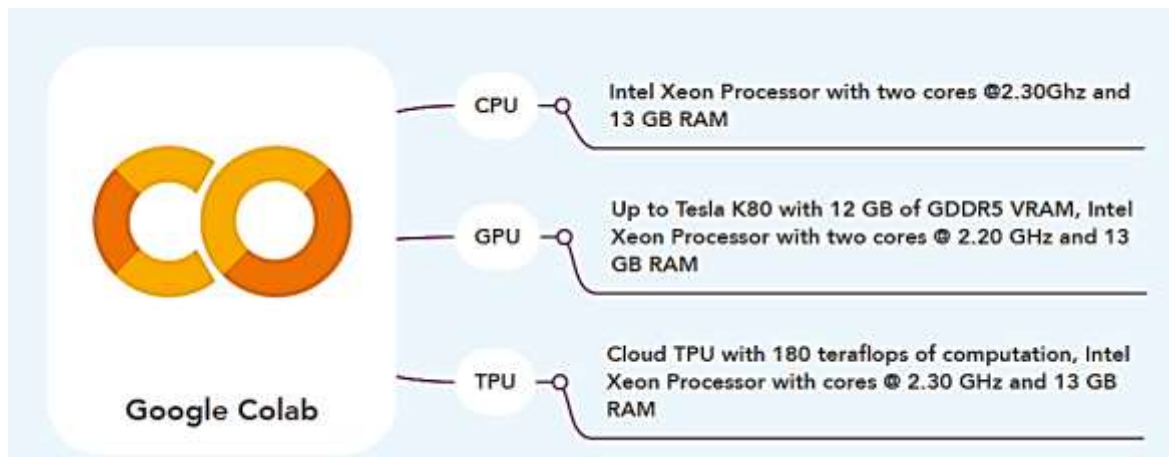


Figura 25. Google Colab. Elaborado por Solórzano Monserrate Mirian

2.2.15.2.1.2 Azure Machine Learning

Según se menciona en la página de Microsoft (2021) es de paga esta plataforma no obstante existe la posibilidad de probar el servicio de manera gratuita. Está basado en utilización de forma interactiva, cuenta con capacidades que ejecuta la creación y expansión de modelos a escala, automatizándolos. Contiene componentes para crear, probar e implementar resultado de análisis que predicen datos.

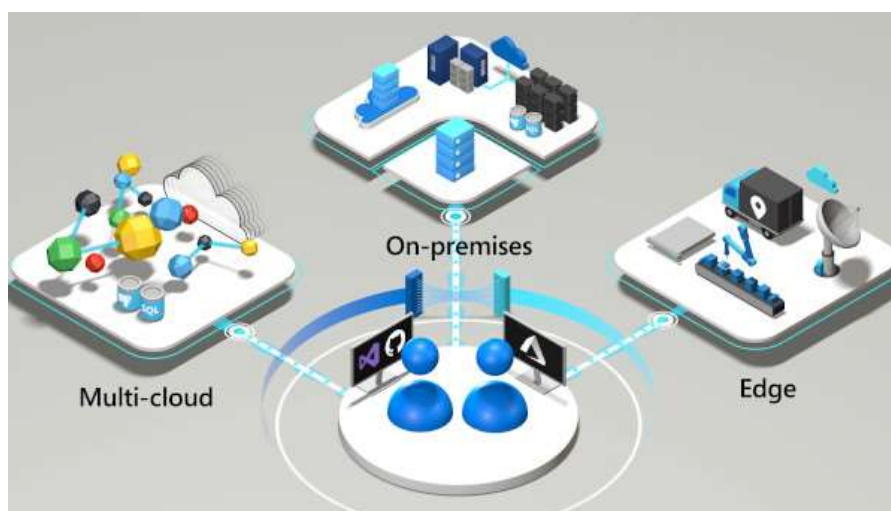


Figura 26. Microsoft Azure en Ignite. Tomada de Microsoft Azure en Ignite: nube híbrida. Elaborado por el autor

2.2.15.2.1.3 SageMaker (SM)

Es un servicio de ML completamente administrado, con Amazon SageMaker, los científicos de datos y los desarrolladores pueden crear y entrenar modelos de ML de manera sencilla y ágil, luego implementarlos directo en entornos de alojamiento listos para producción. Proporciona una instancia integrada de cuadernos de creación de Jupyter, lo que proporciona

un fácil acceso a sus fuentes de datos para exploración y análisis, por lo que no necesita administrar servidores. También proporciona algoritmos comunes de aprendizaje automático que están optimizados para operar de manera eficiente con datos muy grandes en un entorno distribuido. Beneficiarse del soporte nativo para marcos y algoritmos patentados. (Amazon, 2019)

Ofrece opciones flexibles de capacitación distribuida para adaptarse a su flujo de trabajo específico. Implemente modelos en un entorno seguro y escalable lanzando un modelo con un solo clic desde la consola de Amazon SageMaker. La capacitación y el hospedaje se facturan por minuto de uso, sin tarifas mínimas ni compromisos iniciales. (Amazon, 2019)

2.2.15.2.2 IDE de Escritorio

2.2.15.2.2.1 Jupyter Notebook

Jupyter es un ambiente interactivo y de código abierto que admite desarrollar código Python de manera dinámica. el cual se ejecuta en local como una aplicación cliente-servidor y posibilita tanto la ejecución de código como la escritura de texto, favoreciendo así la interacción del entorno y que se pueda comprender el código realizado como la lectura de un documento. (De la Fuente, 2019)

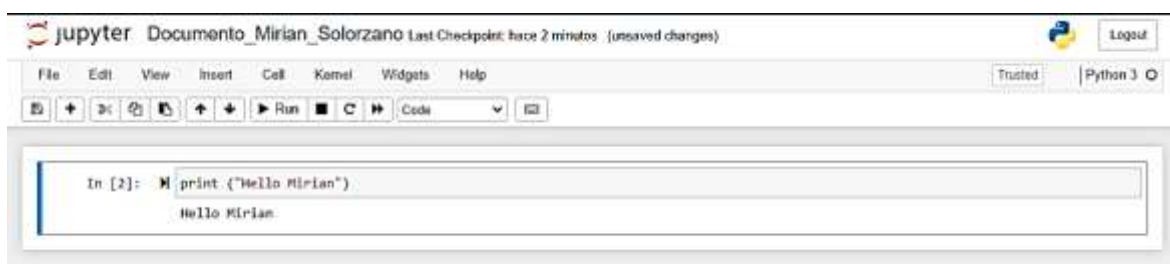


Figura 27. Entorno Jupyter Notebook. Elaborado por Solórzano Monserrate Mirian

Tabla 9. Características de Jupyter

Características	C++
Contiene librerías científicas como son Numpy, Matplotlib, SciPy, etc	
Ambiente interactivo	
Código abierto	
Utiliza lenguaje para ML	

Información tomada de la investigación. Elaborado por Solórzano Monserrate Mirian

2.2.15.2.2.2 Spyder

El nombre proviene del acrónimo en inglés “Scientific PYthon Develoment EnviRonment”. Es utilizado para la ejecución de Python enfocado en el análisis de datos, investigación y elaboración de paquetes científicos, es de acceso libre y está escrito en Python, y contiene una interfaz bien planificada con secciones intercambiables, opciones interactivas y diseños personalizables. (Gonzalez, 2018)

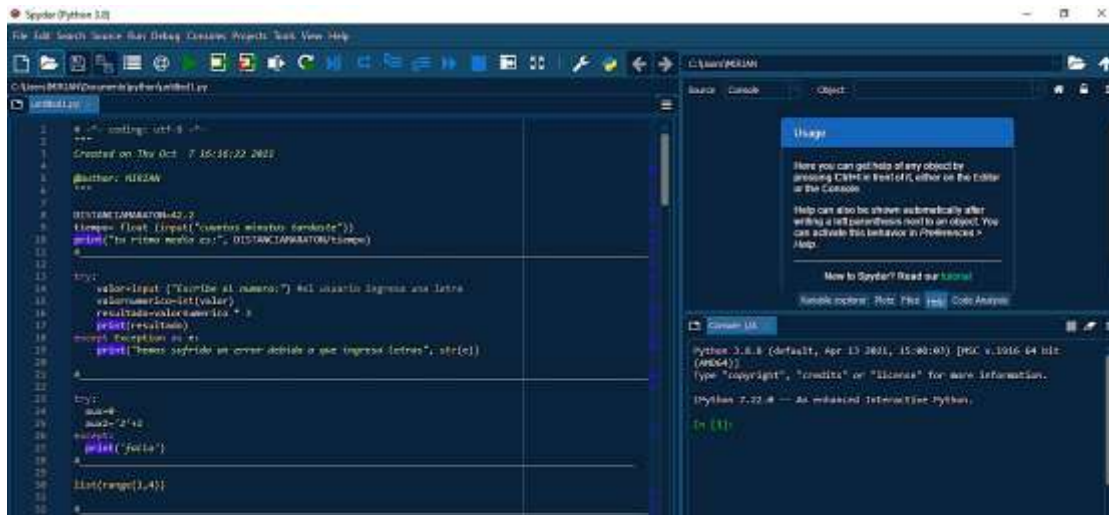



Figura 28. Entorno Spyder. Elaborado por Solórzano Monserrate Mirian

Tabla 10. Características de Spyder

Características	C++
Aplicable a varios idiomas	 SPYDER The Scientific Python Development Environment
Código gratuito	
Contiene visualización de gráficos, documentos y datos	
Compilador interactivo	

Información tomada de aprendeIA. Elaborado por Solórzano Monserrate Mirian

2.2.15.2.2.3 PyCharm

Es un IDE de Python dedicado el cual brinda una extensa gama de herramientas principales para los desarrolladores de Python, estrechamente integradas para generar un entorno adecuado para el desarrollo productivo de Python, ciencia de datos y web. Se encuentra disponible en tres versiones para la comunidad que es código de manera gratuita, profesional que es de pago y finalmente Edu que es para aprender lenguajes y tecnologías en relación con herramientas educativas. (Pycharm, 2021)

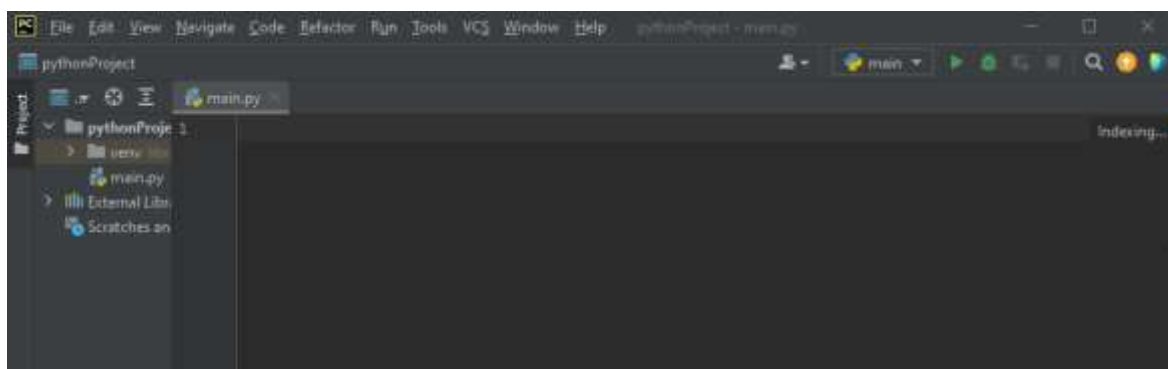


Figura 29. Entorno Pycharm. Elaborado por Solórzano Monserrate Mirian

En vista a lo investigado se llega a la conclusión de trabajar en Google Colab debido a que es la mejor opción en cuanto a los recursos a usar del proyecto además que cuenta con librerías certificadas y aplicables a ML, por otra parte todo el material queda almacenado en la nube con la posibilidad de manipularlo en cualquier bloque del algoritmo sin la necesidad de la instalación de un programa adicional.

2.3 Fundamentación legal

2.3.1 Revisión de Normas Nacionales

Para realizar el desarrollo de la investigación y posteriormente aplicarlos se debe tener en cuenta varios aspectos legales que rige en leyes o reglamentos conocidos dentro de la constitución del Ecuador.

2.3.1.1 Constitución de la República del Ecuador

Artículo 350.- "El sistema de educación superior tiene como finalidad la formación académica y profesional con visión científica y humanista; la investigación científica y tecnológica; la innovación, promoción, desarrollo y difusión de los saberes y las culturas; la construcción de soluciones para los problemas del país, en relación con los objetivos del régimen de desarrollo."

Art. 262.- Los gobiernos regionales autónomos tendrán las siguientes competencias exclusivas, sin perjuicio de las otras que determine la ley que regule el sistema nacional de competencias:

6. Determinar las políticas de investigación e innovación del conocimiento, desarrollo y transferencia de tecnologías, necesarias para el desarrollo regional, en el marco de la planificación nacional.

Art. 277.- Para la consecución del buen vivir, serán deberes generales del Estado:

6. Promover e impulsar la ciencia, la tecnología, las artes, los saberes ancestrales y en general las actividades de la iniciativa creativa comunitaria, asociativa, cooperativa y privada

Art. 385.- El sistema nacional de ciencia, tecnología, innovación y saberes ancestrales, en el marco del respeto al ambiente, la naturaleza, la vida, las culturas y la soberanía, tendrá como finalidad:

1. Generar, adaptar y difundir conocimientos científicos y tecnológicos. [...]
3. Desarrollar tecnologías e innovaciones que impulsen la producción nacional, eleven la eficiencia y productividad, mejoren la calidad de vida y contribuyan a la realización del buen vivir.

Artículo 386.- El Sistema Nacional, de Ciencia, Tecnología, Innovación, y; Saberes Ancestrales "Comprenderá programas, políticas, recursos, acciones, e incorporaré a instituciones del Estado, universidades y escuelas politécnicas, institutos de investigación públicos y particulares, empresas públicas y privadas, organismos no gubernamentales y personas naturales o jurídicas, en tanto realizan actividades de investigación, desarrollo tecnológico, innovación y aquellas ligadas a los saberes ancestrales."

2.3.1.2 Reglamento de Régimen Académico

Art. 78.- "Se entenderá como pertinencia de carreras y programas académicos a la articulación de la oferta formativa, de investigación y de vinculación con la sociedad, con el régimen constitucional del Buen Vivir, el Plan Nacional de Desarrollo, los planes regionales y locales los requerimientos sociales en cada nivel territorial y las corrientes internacionales científicas y humanísticas de pensamiento.

2.3.1.3 Ley Orgánica de Comunicación

Art. 35.- Derecho al acceso universal a las tecnologías de la información y comunicación. - Todas las personas tienen derecho a acceder, capacitarse y usar las tecnologías de información y comunicación para potenciar el disfrute de sus derechos y oportunidades de desarrollo.

2.3.1.4 Ley Orgánica de Telecomunicaciones

Art. 88.- Promoción de la Sociedad de la Información y del Conocimiento.

5. Promover el desarrollo y masificación del uso de las tecnologías de información y comunicación en todo el territorio nacional.
6. Apoyar la educación de la población en materia de informática y tecnologías de la información, a fin de facilitar el uso adecuado de los servicios o equipos.

2.3.2 Revisión de Normas Internacionales del uso de la IA

Debido a que la normativa referente al manejo, entrenamiento y uso de la IA a nivel de legislación ecuatoriana aún no se encuentra desarrollada se verifican las leyes que están en vigencia en otros continentes.

2.3.2.1 Resolución del Parlamento Europeo sobre uso policial de la Inteligencia Artificial

El pasado 6 de octubre de 2021 fue aprobada con 377 votos a favor, 248 en contra y 62 abstenciones, la Resolución del Parlamento Europeo sobre la IA en el Derecho Penal y su utilización por las autoridades policiales y judiciales.

El presente Reglamento también debe aplicarse a las instituciones, oficinas y organismos de la Unión cuando actúen como proveedores o usuarios de un sistema de IA. Los sistemas de IA desarrollados o utilizados exclusivamente con fines militares deben quedar excluidos del ámbito de aplicación del presente Reglamento cuando dicho uso sea competencia exclusiva de la Política Exterior y de Seguridad Común regulada en el título V del Tratado de la Unión Europea (TUE).

Art. 6 *Normas de clasificación de los sistemas de IA de alto riesgo*

Independientemente de si un sistema de IA se introduce en el mercado o se pone en servicio independientemente de los productos mencionados en las letras a) y b), dicho sistema de IA se considerará de alto riesgo cuando se cumplan las dos condiciones siguientes:

(a) el sistema de IA está destinado a ser utilizado como componente de seguridad de un producto, o es en sí mismo un producto, cubierto por la legislación de armonización de la Unión enumerada en el anexo II;

(B) el producto cuyo componente de seguridad es el sistema de IA, o el propio sistema de IA como producto, debe someterse a una evaluación de la conformidad por un tercero con vistas a la comercialización o puesta en servicio de dicho producto de conformidad con la armonización de la Unión legislación enumerada en el Anexo II.

Art. 10 *Datos y gobernanza de datos*

Los sistemas de IA de alto riesgo que utilicen técnicas que impliquen el entrenamiento de modelos con datos se desarrollarán sobre la base de conjuntos de datos de entrenamiento, validación y ensayo que cumplan los criterios de calidad.

2.3.2.2 Reglamento general de protección de datos

Art. 22 En función de los datos que se utilizan para entrenar los sistemas de IA, sus resultados pueden estar sesgados. La utilización de la IA para crear obras puede tener repercusiones sobre los derechos de propiedad intelectual y plantear cuestiones en relación, por ejemplo, con la patentabilidad, los derechos de autor y los derechos de propiedad.

Capítulo III

Metodología

3.1 Propuesta tecnológica

En esta sección se detallará de acuerdo con las metodologías empleadas a lo largo de la investigación, la arquitectura y modelos usados mostrando cada una de las predicciones, a su vez comprende el tipo de investigación, población y muestra utilizada que comprende la zona 8 de la provincia del Guayas que son Duran, Guayaquil y Samborondón. La investigación utilizada fue exploratoria aplicando el método de revisión bibliográfica y desarrollo de software.

3.1.1. Descripción del proceso metodológico

El método de investigación científica es un conjunto de métodos, leyes y procedimientos que guían los esfuerzos de investigación para resolver problemas científicos con la máxima eficiencia. El método se apoya en diferentes etapas de desarrollo para extraer importantes conclusiones a partir de la confirmación de las hipótesis y de la investigación realizada. (Otzen, Manterola, Rodriguez, & Garcia, 2017)

3.2 Tipos de investigación



Figura 30. Niveles de Investigación. Información tomada de Diseños de Investigación. Elaborado por Solórzano Monserrate Mirian

Según Monjarás, et al., (2019) detallan que los tipos de estudio de la investigación se clasifican en base a:

- Características de Datos
- Dimensión Temporal
- Intervención
- Obtención de Datos

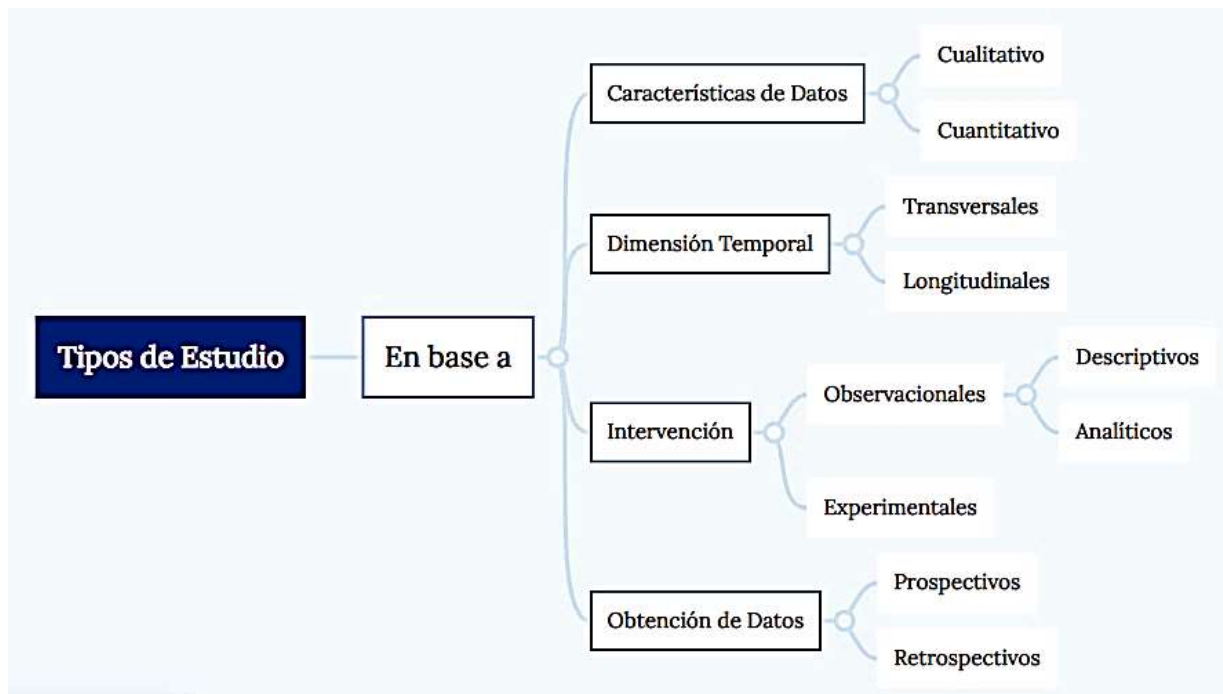


Figura 31. Tipos de Estudio. Información tomada de Diseños de Investigación. Elaborado por Solórzano Monserrate Mirian

3.2.1. Investigación exploratoria.

De acuerdo con Nieto, (2018) menciona que el propósito de la investigación exploratoria es familiarizarnos con fenómenos relativamente desconocidos, obtener información sobre la posibilidad de examinar más a fondo un contexto particular, descubrir nuevas preguntas, identificar conceptos o variables prometedores, priorizar futuras investigaciones o sugerir afirmaciones y consejos u suposición.

3.3 Metodología de investigación

3.3.1 Metodología para la Revisión Bibliográfica

Es un paso esencial en cualquier proyecto de investigación y corresponde a una descripción detallada del tema o la tecnología, es esencial para garantizar que se obtenga la información más relevante la cual puede provenir de diferentes escenarios de desarrollo de tecnología y permite que se tome decisiones estratégicas sobre la investigación. (Gómez, et al., 2014)

3.3.2 Metodología Cuantitativa

Según Hernández y Vásquez, (2018) Indican que la metodología cuantitativa engloba un conjunto de métodos y técnicas que intentan acercarse al conocimiento de la realidad social mediante el análisis de la extensión, alcance y trascendencia de los hechos. Los métodos cuantitativos son adecuados cuando queremos estimar la magnitud o la ocurrencia de un fenómeno y probar hipótesis.

El método cuantitativo se basa en el campo de la estadística analizando y mirando la realidad objetiva, a través de mediciones y evaluación numérica para obtener datos de seguridad, encontrar evidencia y generalizar la causa. (Escudero y Cortez, 2018).

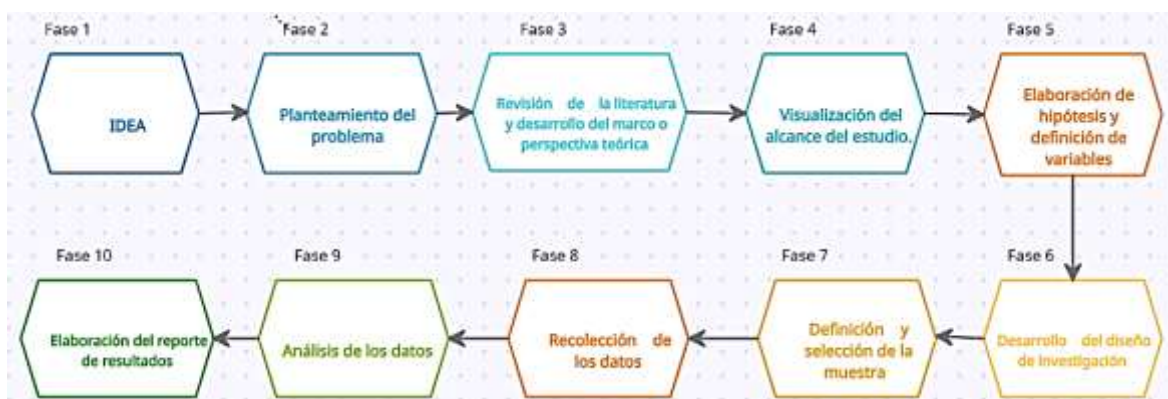


Figura 32. Proceso de la Metodología Cuantitativa. Información tomada de METODOLOGÍA DE LA INVESTIGACIÓN. Elaborado por Solórzano Monserrate Mirian

3.3.2.1 Técnicas de Investigación

3.3.2.1.1 Encuesta

La encuesta representa uno de los métodos de recopilación de datos de la investigación cuantitativa que forman un término medio entre "observación y experimentación" por ello el investigador se basa en datos y no en suposiciones. (Torres, Paz, y Salazar, 2019)

En este caso se utilizó la encuesta impulsando 4141 encuestas para la obtención de datos requeridos para el entrenamiento del modelo NLP, la cual fue realizada en un formulario virtual utilizando la herramienta denominada Google Forms, esta herramienta cuenta con el apoyo de proporcionar las tabulaciones de la encuesta y mostrando las estadísticas de cada pregunta y cuenta con la facilidad de poder compartirlo mediante el link tanto a correos electrónicos como redes sociales.

3.3.2.2 Descripción del Procedimiento Metodológico

3.3.2.2.1 Población

La población tomada para esta investigación está constituida por los ciudadanos que residen en la zona 8 de la provincia del Guayas, Según el Instituto nacional de estadística y censos (INEC), (2010), en la base de datos (BD) en la cual están almacenados los datos del recuento de población ecuatoriana en el año 2010, el cantón Durán consto de 235.769 habitantes, el cantón Guayaquil consto de 2.350.915 y finalmente el cantón Samborondón consto de 67.590, en total suman 2.654.274 habitantes en la zona 8 de la provincia del Guayas.

La variable principal es el tipo de investigación cualitativa, expresada como la proporción del fenómeno de investigación que es relevante para la población. (Aguilar, 2005)
Para ello se empleó técnicas estadísticas de poblaciones infinitas, que se encargan de obtener datos de una población que se desconoce el total que la integran o supera de 10000 habitantes.

$$n = \frac{N * Z_{\alpha}^2 p * q}{d^2 (N - 1) + Z_{\alpha}^2 * p * q}$$

Donde:

n= Tamaño de la Muestra

NC = nivel de confianza deseado

Z= se define según el NC

S= varianza de la población en estudio (que es el cuadrado de la desviación estándar y puede obtenerse de estudios similares o pruebas piloto)

d = intervalo de confianza deseado en la determinación del valor promedio de la variable en estudio.

N= tamaño de la población

3.3.2.2.2 Muestra

Según Otzen y Manterola, (2017) muestran que la representación de la muestra permite la extrapolación de los resultados observados en ella y, por tanto, la generalización a poblaciones accesibles; de aquí, a los blancos. Por lo tanto, esta es representativa; solo si se selecciona aleatoriamente, es decir, todos los sujetos de la población objetivo tienen la misma probabilidad de ser seleccionados en esta muestra e incluidos en el estudio.

Es la porción de unidades que representan una población o universo que se seleccionan al azar y se observan para censos válidos. (López y Fachelli, 2015)

$$n = \frac{N * Z_{\alpha}^2 p * q}{d^2 (N - 1) + Z_{\alpha}^2 * p * q}$$

$$n = \frac{2.654.274 * 1.96_{\alpha}^2 * 0.5 * 0.5}{0.05^2 (2.654.274 - 1) + 1.96_{\alpha}^2 * 0.5 * 0.5}$$

$$n = 384,10$$

$$n = 384$$

Según especialistas de IA consultados al trabajar con un Dataset en Data Science, para elevar la calidad de los resultados que arroja el modelo posterior al entrenamiento, entre más datos recopilados se tenga, mayor calidad se obtendrá en la salida del modelo, tanto con datos conocidos (train) y datos desconocidos (test) por lo cual para este estudio se cuenta con 4924 datos recopilados que fueron empleados en el aprendizaje del modelo, de los cuales se aplicó la división de 80-20, 80% proporcionado para entrenamiento (training) y el 20% para las pruebas (testing).

3.3.2.2.3 Análisis de la Encuesta

Pregunta N° 1.- Seleccione la Edad

Tabla 11. Edad

Opciones de respuestas	Encuestados	Porcentaje
De 18 a 30 años	290	76%
De 31 a 40 años	48	13%
De 41 a 50 años	28	7%
De 51 a 78 años	18	5%
Total	384	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Solórzano Monserrate Mirian

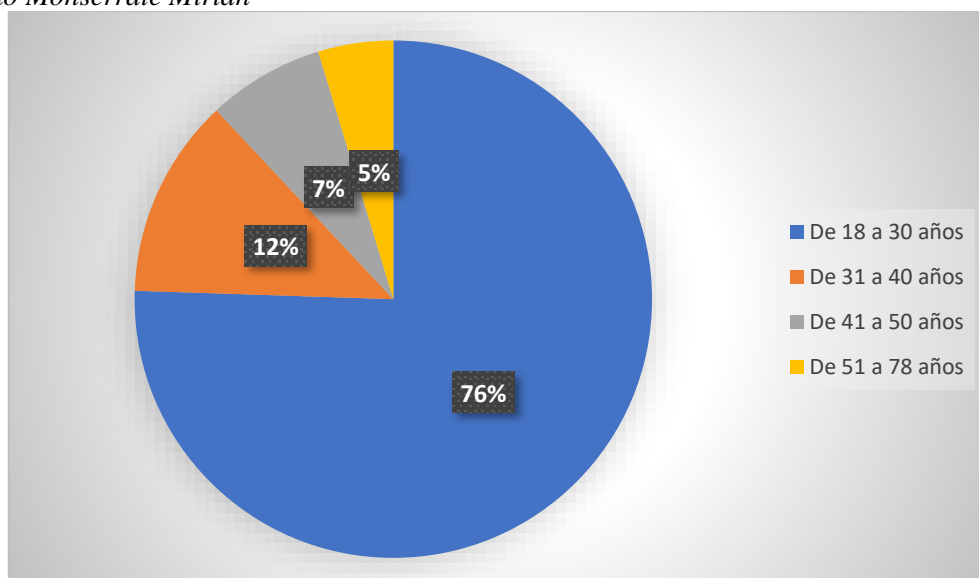


Figura 33. Edad de los encuestados. Información tomada de formulario de Google Elaborado por Solórzano Monserrate Mirian

De acuerdo con el gráfico estadístico No. 1 se puede observar la muestra de 384 de los habitantes de la zona 8, otras provincias y país, se identifica que el 95% de los encuestados son de 18 a 50 años.

Pregunta N° 2.- Género

Tabla 12. Género

Opciones de respuestas	Encuestados	Porcentaje
Masculino	159	41%
Femenino	225	59%
Total	384	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Solórzano Monserrate Mirian

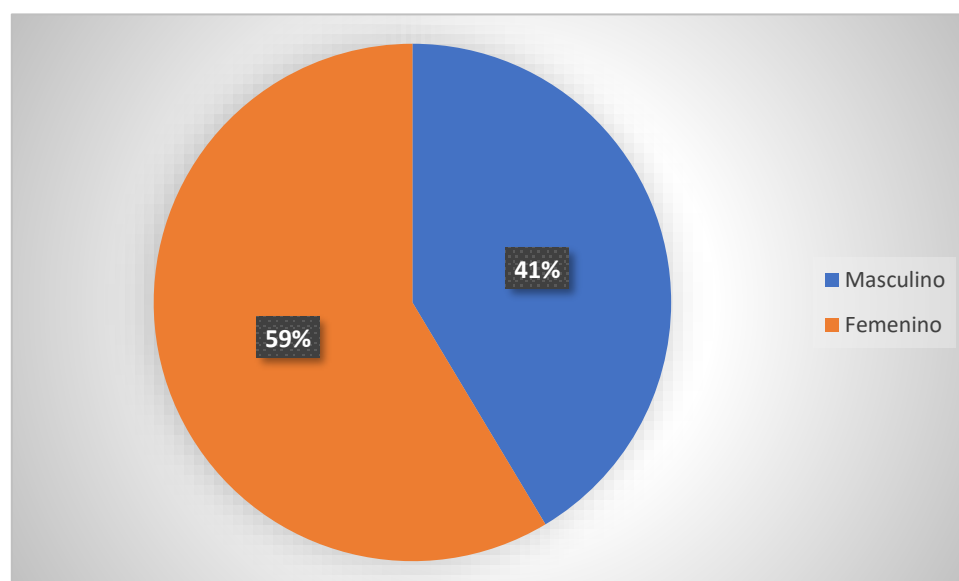


Figura 34. Género de los encuestados. Información tomada de formulario de Google Elaborado por Solórzano Monserrate Mirian

Conforme al gráfico estadístico No. 2 de la muestra de 384 encuestados de la zona 8 de la provincia del Guayas correspondiente a los cantones Duran, Guayaquil y Samborondón entre otras se identifica que el 59% corresponde al género femenino y a continuación el 41% del género masculino.

Pregunta N° 3.- Lugar que reside de la zona 8 del Ecuador

Tabla 13. Residentes de la Zona 8

Opciones de respuestas	Encuestados	Porcentaje
Guayaquil	299	78%
Duran	23	6%
Samborondón	7	2%

Otra ciudad del Ecuador	45	12%
Otro País	10	3%
Total	384	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Solórzano Monserrate Mirian

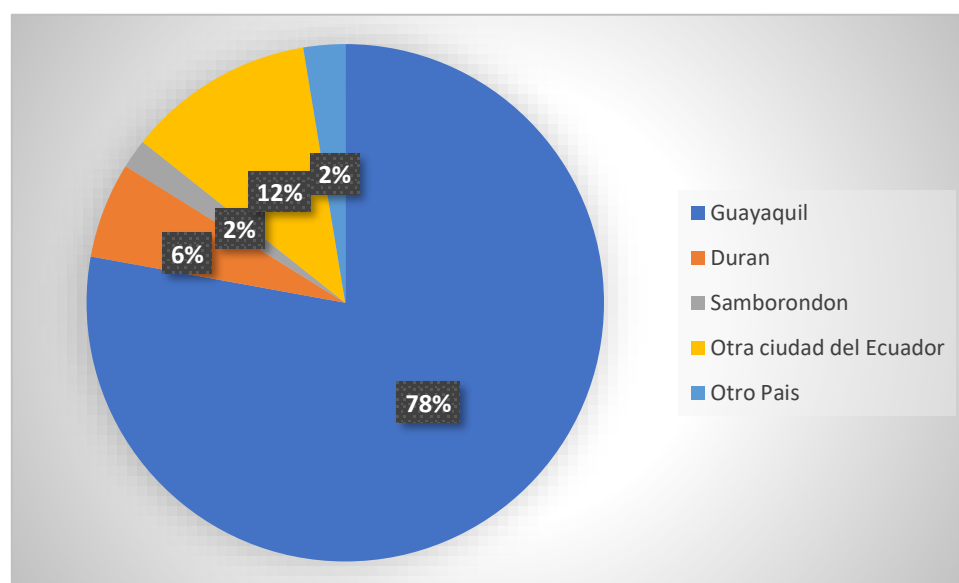


Figura 35 Residentes de la Zona 8. Información tomada de formulario de Google Elaborado por Solórzano Monserrate Mirian

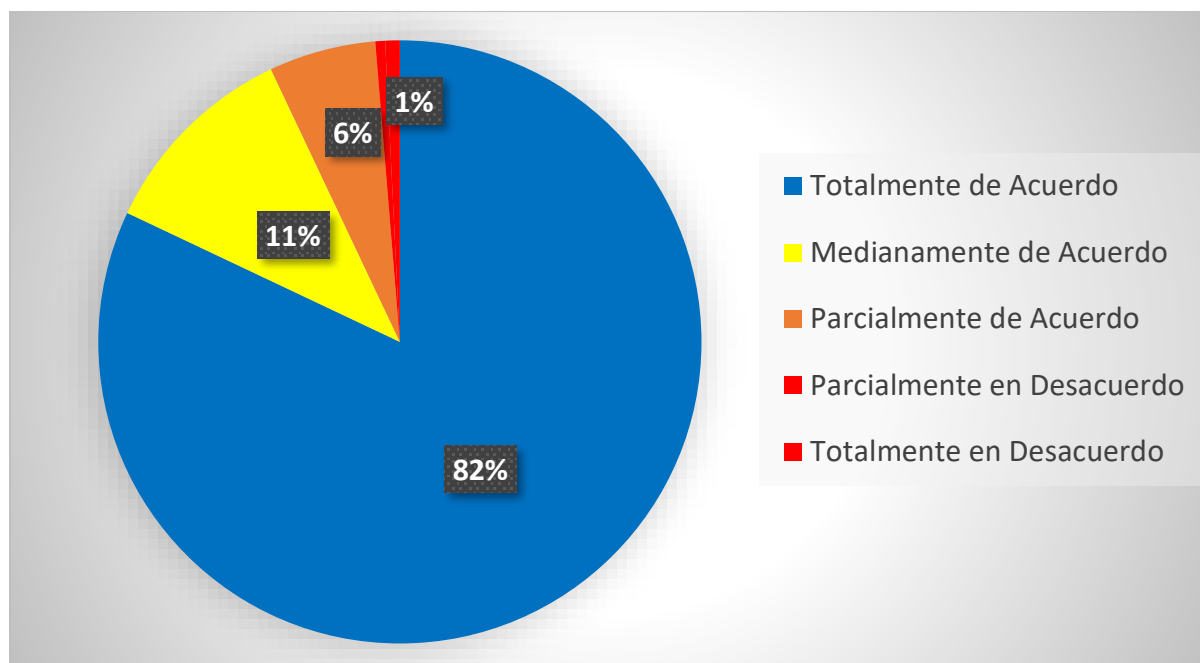
Respecto al gráfico estadístico No. 3 de la muestra de 384 encuestados se puede identificar que el 86% corresponden a la Zona 8 de la provincia del Guayas, seguido del 12% y 2% correspondientes a otras ciudades del Ecuador y otros países.

Pregunta N° 4.- ¿Usted considera importante CONOCER cómo el CORONAVIRUS (Covid-19) afecta nuestra salud?

Tabla 14. Importancia de conocer síntomas de Coronavirus

Opciones de respuestas	Encuestados	Porcentaje
Totalmente de Acuerdo	315	82%
Medianamente de Acuerdo	42	11%
Parcialmente de Acuerdo	22	6%
Parcialmente en Desacuerdo	2	1%
Totalmente en Desacuerdo	3	1%
Total	384	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Solórzano Monserrate Mirian



*Figura 36. Importancia de conocer síntomas de Covid. Información tomada de formulario de Google
Elaborado por Solórzano Monserrate Mirian*

De acuerdo con el gráfico estadístico No. 4 de la muestra de encuestados $n=384$ el 99% está totalmente de acuerdo de la importancia de tener conocimientos de cómo el Coronavirus (COVID-19) afecta a la salud.

Pregunta N° 5.- ¿Usted está de Acuerdo que las vacunas contra el coronavirus (Covid-19) son efectiva eliminando el virus ?

Tabla 15. Efectividad de Vacunas

Opciones de respuestas	Encuestados	Porcentaje
Totalmente de Acuerdo	96	25%
Parcialmente de Acuerdo	127	33%
De Acuerdo	105	27%
Parcialmente en Desacuerdo	34	9%
Totalmente en Desacuerdo	22	6%
Total	384	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Solórzano Monserrate Mirian

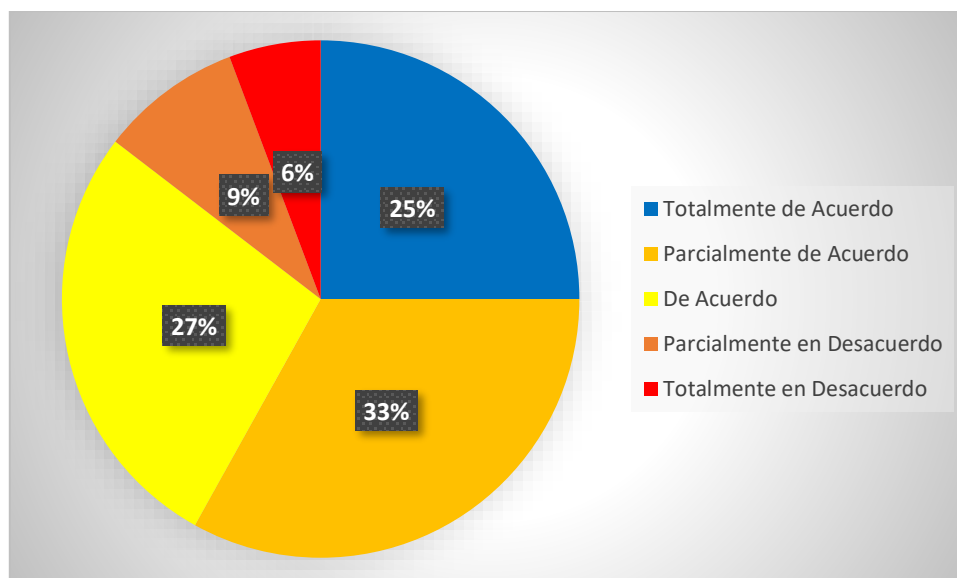


Figura 37. Vacunas de Covid. Información tomada de formulario de Google Elaborado por Solórzano Monserrate Mirian

Conforme al grafico estadístico No. 5 el 85% del total de encuestados $n=384$ expresa que esta de acuerdo en que las vacunas contra en Covid-19 sean efectivas eliminando este virus mientras que el 6% está totalmente en desacuerdo.

Pregunta N° 6.- ¿Está de acuerdo que la información del coronavirus (Covid-19) que recibe del ministerio de salud o sub-centro de salud por cualquier medio de comunicación es la adecuada y actualizada como por ejemplo: Los hábitos saludables, evolución del virus etc.?

Tabla 16. Información sobre el Covid-19

Opciones de respuestas	Encuestados	Porcentaje
Totalmente de Acuerdo	134	35%
Parcialmente de Acuerdo	89	23%
De Acuerdo	113	29%
Parcialmente en Desacuerdo	32	8%
Totalmente en Desacuerdo	16	4%
Total	384	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Solórzano Monserrate Mirian

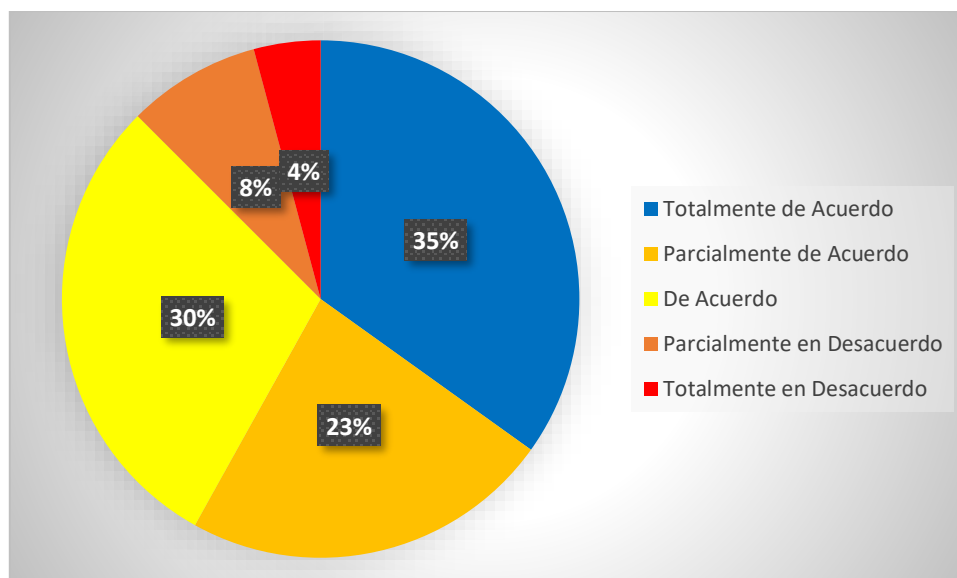


Figura 38. Información sobre el Covid-19. Información tomada de formulario de Google Elaborado por Solórzano Monserrate Mirian

Respecto al grafico estadístico No. 6 de un total de 384 habitantes encuestados el 88% se encuentra acuerdo que la información que recibe sea apropiada y actualizada sobre el Covid-19 por parte del ministerio de salud o sub-centro de salud y otro medio de comunicación.

Pregunta N° 7.- ¿Sabía usted que aplicar HÁBITOS SALUDABLES cuando una persona esta contagiada de coronavirus (covid-19) disminuye el riesgo de afecciones graves incluso descartando hasta la muerte?

Tabla 17. Conocimientos sobre Hábitos Saludables para pacientes contagiados de Covid-19

Opciones de respuestas	Encuestados	Porcentaje
Poseo alto conocimiento del tema	212	55%
Poseo bajo conocimiento del tema	136	35%
No tenía conocimiento	36	9%
Total	384	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Solórzano Monserrate Mirian

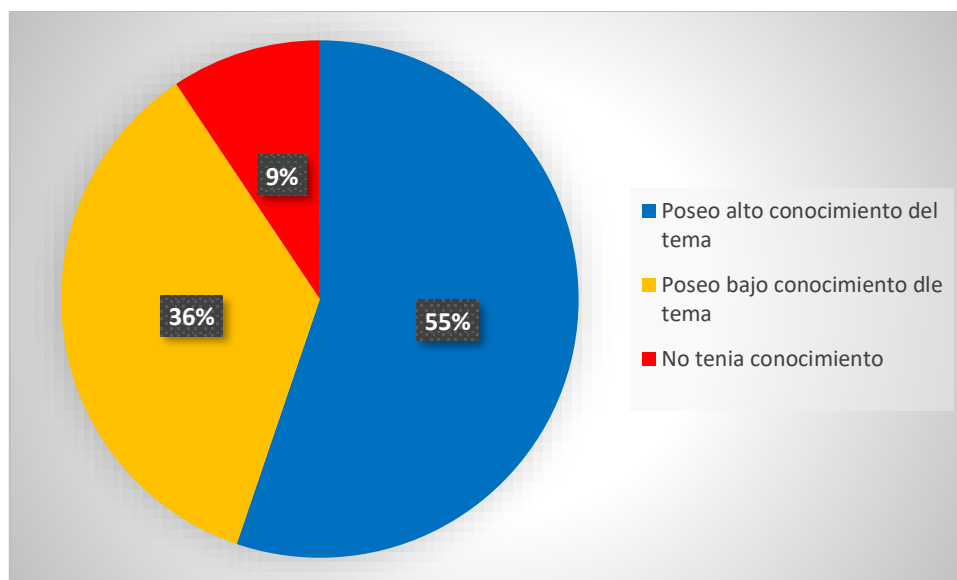


Figura 39. Conocimientos sobre Hábitos Saludables para pacientes contagiados de Covid-19. Información tomada de formulario de Google Elaborado por Solórzano Monserrate Mirian

De acuerdo con el gráfico estadístico No. 7 se visualiza que de la muestra de 384 encuestados el 91% posee alto y bajo conocimiento sobre hábitos saludables para pacientes contagiados de Covid-19 mientras que el 9% desconoce del tema.

Pregunta N° 8.- ¿Usted posee un smartphone básico (Teléfono móvil con acceso a internet)?

Tabla 18. Posesión de Smartphone de los Habitantes

Opciones de respuestas	Encuestados	Porcentaje
Si poseo uno de Gama Alta (costo > \$501)	62	16%
Si poseo uno de Gama Media (costo \$201 a \$500)	174	45%
Si poseo uno de Gama Baja con internet (costo < \$200)	129	34%
No poseo con internet pero estoy en proceso de adquirir uno	14	4%
No poseo con internet y no planeo adquirirlo	5	1%
Total	384	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Solórzano Monserrate Mirian

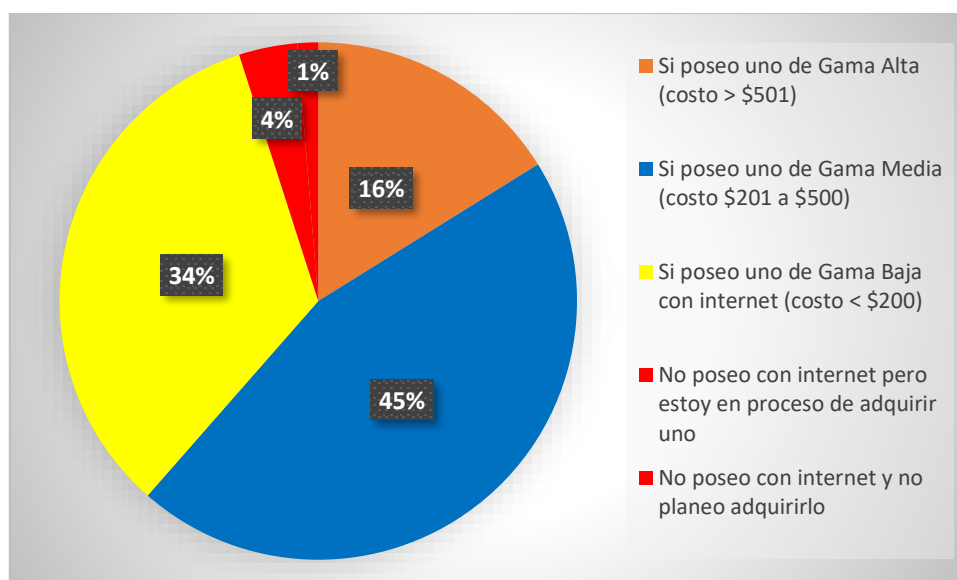


Figura 40. Posesión de Smartphone de los Habitantes. Información tomada de formulario de Google Elaborado por Solórzano Monserrate Mirian

Respecto al gráfico estadístico No. 8 del total de habitantes encuestados $n=384$, el 95% indica que posee un dispositivo smartphone ya sea de alta, media y baja gama con acceso a internet.

Pregunta N° 9.- ¿ Sabía usted que la tecnología de INTELIGENCIA ARTIFICIAL es capaz de desarrollar aplicaciones móviles que permita interactuar y mantener información de forma actualizada para combatir el coronavirus (Covid-19) en cualquier lugar del mundo, a cualquier hora, incluyendo fechas de feriado?

Tabla 19. Conocimientos sobre Inteligencia Artificial

Opciones de respuestas	Encuestados	Porcentaje
Poseo alto conocimiento del tema	174	45%
Poseo bajo conocimiento del tema	170	44%
No tenía conocimiento	40	10%
Total	384	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Solórzano Monserrate Mirian

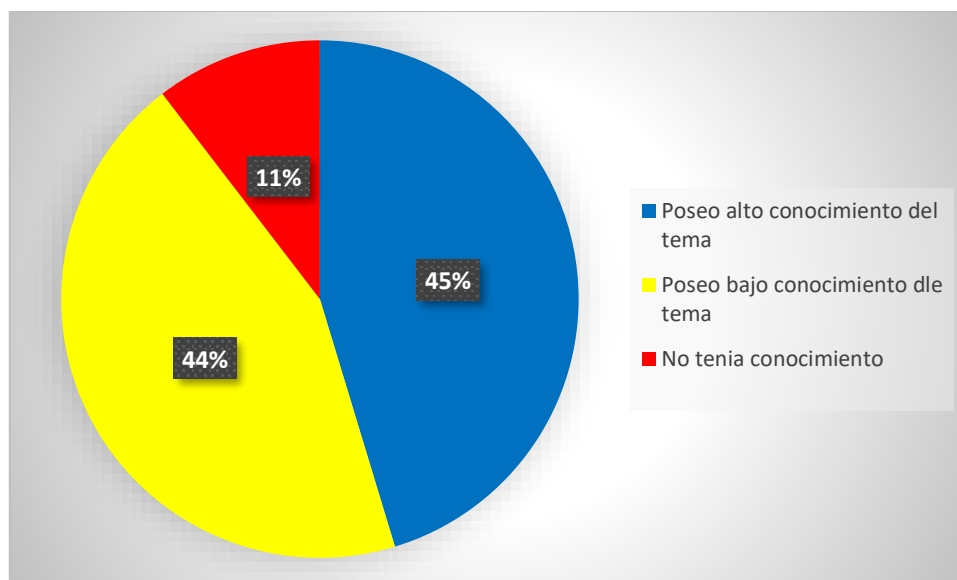


Figura 41. Conocimientos sobre la Inteligencia Artificial. Información tomada de formulario de Google Elaborado por Solórzano Monserrate Mirian

Acorde al grafico estadístico No. 9 se puede identificar que de los 384 encuestados el 89% tiene alto y bajo conocimiento que por medio de la tecnología de IA existe la capacidad de poder desarrollar aplicaciones móviles para la interacción y poder estar mayormente informado sobre cómo combatir el covid-19 en cualquier parte del mundo, a toda hora inclusive fechas festivas de manera actualizada, seguido del 11% que no conoce del tema.

Pregunta N° 10.- ¿Le gustaría contar con una aplicación móvil que le permita interactuar, mantener informado de forma actualizada para combatir al coronavirus (COVID-19) sobre los HÁBITOS SALUDABLE utilizando la tecnología de inteligencia artificial de forma GRATUITA?

Tabla 20. Importancia de herramientas tecnológicas

Opciones de respuestas	Encuestados	Porcentaje
Totalmente de Acuerdo	203	53%
Parcialmente de Acuerdo	110	29%
De Acuerdo	60	16%
Parcialmente en Desacuerdo	5	1%
Totalmente en Desacuerdo	6	2%
Total	384	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Solórzano Monserrate Mirian

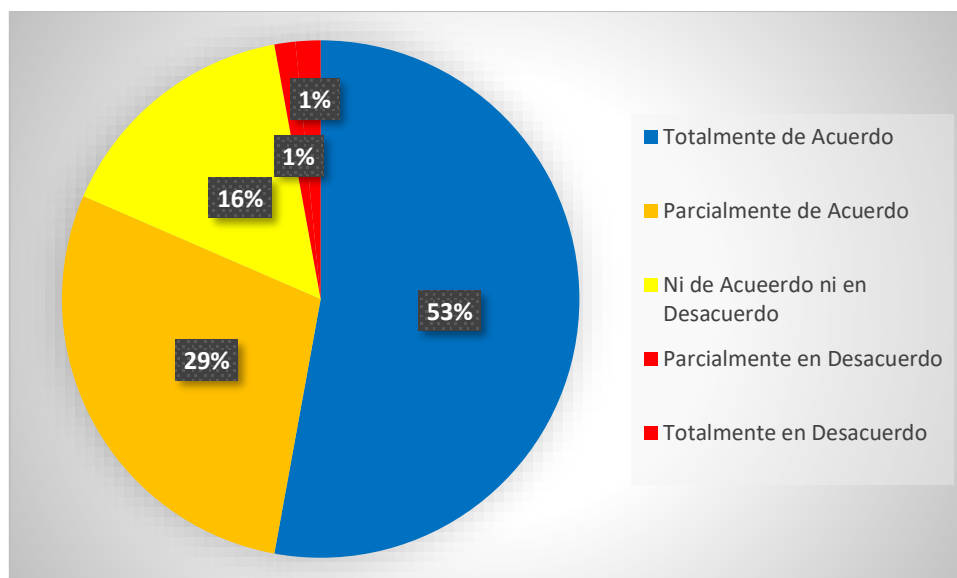


Figura 42. Importancia de herramientas tecnológicas. Información tomada de formulario de Google Elaborado por Solórzano Monserrate Mirian

Referente al gráfico estadístico No. 10 se puede identificar que de la muestra tomada de 384 encuestados de la Zona 8 y otras ciudades/países el 82% expresan que están Total y parcialmente de acuerdo con una aplicación móvil que les permita mantenerse informado de una manera actualizada de todo lo referente al Covid-19 de forma gratuita, utilizando la tecnología de IA orientada a NLP.

3.3.3 Metodología Cualitativa

Según Lugo, (2021) Busca lograr una descripción holística, es decir, intenta analizar en detalle un problema o actividad específica, en ella se estudia la eficacia de actividades, problemas, medios, materiales o herramientas en una determinada situación.

3.3.3.1 Técnica de Investigación

3.3.3.1.1 Entrevista

Suele realizarse entre dos personas, el entrevistador y el entrevistado. Cualquier pregunta planteada a los encuestados debe registrarse en el cuestionario o se puede efectuar mediante un video para ayudar a registrar la información y que esta pueda obtenerse mediante la grabación. (Torres et al, 2019)

Para la recolección de datos en este punto se utilizó una técnica de gran utilidad como es la entrevista la cual se realizó un formulario de preguntas con la finalidad de conocer la opinión de los especialistas de Tecnología con conocimientos en IA y afines quienes hayan interactuado con este tipo de tecnología los cuales corresponden a la zona 8 de la provincia del Guayas.

3.3.3.1.2 Análisis de la Entrevista

Se realizó la entrevista a 3 profesionales de la especialidad de Ingeniero en Electrónica Ing. Jostin Marcelo Maldonado Flores, Magister en Inteligencia Artificial Msg. Ricardo Manuel Prieto Galarza y Msg. Aristo Cabrera Torres, llevan entre 3 a 4 años trabajando en proyectos y temas de Inteligencia Artificial y poseen conocimiento de Machine Learning y de la rama de la IA denominada Procesamiento de Lenguaje Natural (NLP), en dicha entrevista indican que estos temas son extremadamente importantes para la superación de la actual pandemia ya que han surgido diversos proyectos aplicando la IA ya sea en el ámbito de la salud, laboral y comercio empresarial.

En su opinión consideran que las redes neuronales son el algoritmo más adecuado para utilizarlo en una arquitectura NLP para clasificación de conversaciones textuales ya que estas solo se deben tener en cuenta que la calidad y cantidad de los datos proporcionados al entrenamiento sea lo suficientemente amplia para un mejor aprendizaje.

Los modelos que consideran 2 de los profesionales con mayor beneficio es Tranformer y GPT3 ya que facilita la manera de aprendizaje y al utilizarlas en información textual clasificada sea más sencillo de usar, seguido de los modelos Long-Short Term Memory (LSTM), Recurrent Neural Networks (RNN), ELMo, Global Vectors for Words Representations (GloVe).

Además comentan que sería de gran importancia que haya más investigación y nuevas propuestas de construcción de NLP para de esta manera crear modelos efectivos de conversaciones de texto con respecto a los existentes relacionados a combatir el Covid-19.

3.4 Construcción del modelo de Machine Learning

Para el posterior entrenamiento del modelo se requiere de una dataset, la cual proporcionara la información para entrenar los diferentes modelos obteniendo un aprendizaje y medir la asertividad. Se utilizó una base de datos proporcionada por una encuesta realizada a los habitantes de la zona 8 de la provincia del Guayas correspondiente a los cantones de Duran, Guayaquil y Samborondón, denominada BASE ETIQUETADO SINTOMAS-FINAL.csv que esta almacenada en Google Drive y posteriormente descargada y alojada en el computador.

La base de datos usada fue realizada en el año 2021, contiene datos estructurados, el cual es un modelo de data con formato tabular, en documento de Excel, extensión .csv, al trabajar en Google Colab cada vez que se deja de realizar alguna actividad o acción por un largo tiempo, se borran las líneas ejecutadas, por lo cual se debe subir la dataset y proceder a ejecutar cada una nuevamente.

Se efectuará una comparación de modelos, los cuales son Long short-term memory (LSTM), modelo básico NCC, , Modelo Bosque Aleatorio o Random Forest y finalmente pero no menos importante Support Vector Machine (SVM). Se va a realizar un análisis respecto al mayor porcentaje de asertividad que presentan cada uno de ellos sin sobreentrenar el módulo.

Para ejecutar los mencionados módulos se efectuará en lenguaje Python, ya que es un lenguaje versátil, posee una sintaxis simple, es de código abierto, fácil implementación y a su vez proporciona librerías de IA utilizadas en Data Science.

3.4.1 Importación de datos

Se importan los datos de Recomendaciones desde el computador hacia el entorno de trabajo, el cual se determinó usar Google Colab.

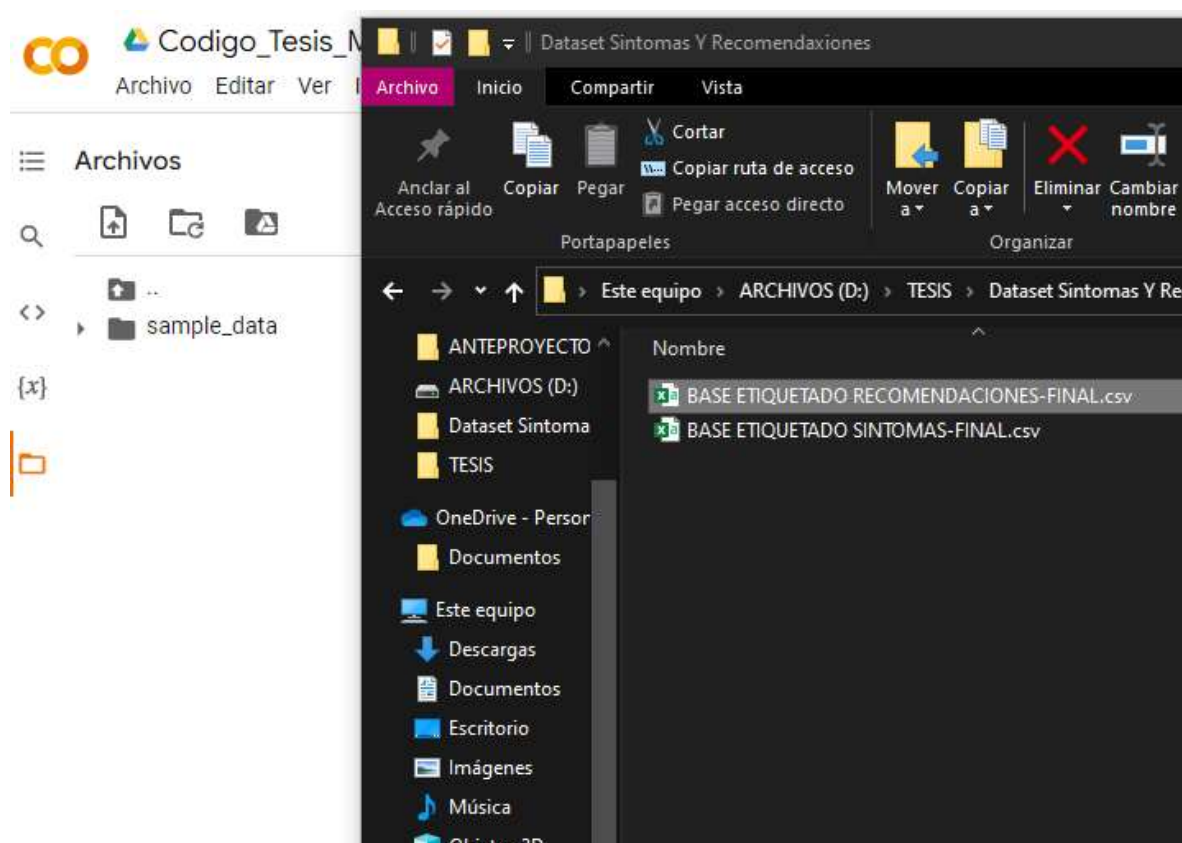


Figura 43. Importación de datos a Google Colab. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

A continuación se muestra la dataset cargada en la bandeja de archivo de Google Colab.



Figura 44. Dataset subida a Google Colab. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

La data se encuentra cargada en un archivo llamado BASE ETIQUETADO SINTOMAS-FINAL, la cual tiene de extensión .csv, en otras palabras es un documento de Excel separados por comas.

Se procede a importar la librería Panda con la sintaxis de import pandas as pd, que quiere decir esto que para llamar la librería no será necesario escribir el nombre completo sino la abreviatura pd, a continuación se nombra dataframe_Recom_Medic donde se van a almacenar los datos, para realizar una revisión rápida de la data se utiliza con .info la cual muestra la estructura interna que contiene la data, con .head se muestra la cabecera de la data que corresponde las primeras cinco filas del dataframe y con .tail se muestra las últimas cinco filas que contiene.

En la siguiente imagen se puede visualizar las columnas que contiene:



Figura 45. Revisión rápida de la Dataset. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

```
[8] muestra el tipo de dato que contiene la dataset
dataframe_recom_medico.types
```

CANTIDAD	tipo de dato
ASIGNACION DE PREPROCESAMIENTO	int64
Marca temporal	object
8. Nombre y Apellido del encuestador (persona que te ha pedido que lleres la encuesta)	object
1. Ha tenido coronavirus?	object
2. Selecciona la edad	object
3. Género	object
4. ¿Qué variante del virus lo contagió?	object
5. ¿En qué fecha se contagió? MM-DD-YYYY	object
6. ¿Nivel de intensidad que tuvo los síntomas?	object
7. ¿En qué lugar o evento considero que se contagió?	object
8. ¿En caso de haber estado vacunado al momento de contagiarse cuántas dosis tenía aplicadas al contagiarse?	object
9. ¿En caso de haber estado vacunado al momento de contagiarse qué vacuna recibió?	object
10. Describe ¿Qué recomendaciones saludables le dio a conocer su médico? ejemplo: MI doctor me recomendó que caminará 4 veces por semana por 40 minutos - SIN DEPURAR	object
11. Describe ¿Qué recomendaciones saludables le dio a conocer su médico? ejemplo: MI doctor me recomendó que caminará 4 veces por semana por 40 minutos - DEPURADA	object
ejercicios	float64
Reposo	float64
vitaminas	float64
terapia respiratoria	float64
hidratado	float64
alimentación	float64
mascarilla	float64
ejercicios respiratorios	float64
no esfuerzo físico	float64
Desestresarme	float64

Figura 46. Tipo de dato que contiene el Dataset. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

3.4.2 Tratamiento de Datos

Se procede a consultar los valores perdidos de la dataset con el siguiente código:

```
#verificacion de variables con valores NaN
val_NaN = dataframe_Recom_Medic.isnull().any().sum()
val_NaN
```

13

Figura 47. Consulta de Missing Values - NaN. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian



Figura 48. Gráfico de valores perdidos. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

El valor total de datos perdidos es de 13, los cuales pertenecen a las primeras columnas del Dataset.

Se añadió una función en la cual realiza el preprocesamiento del texto como son eliminación de caracteres especiales, cambio de palabras con tilde a sin tilde, cambio de palabras mayúsculas a minúsculas.

```
✓ 0s #definicion de una funcion para cambiar datos que contengan tilde por una sin tilde,
#convertir de mayuscula a minuscula
#eliminacion de caracteres especiales
def Clean_txt(text):
```

Figura 49. Definición de función para limpiar el texto. Elaborado por Solórzano Monserrate Mirian

Se realizó una comparación de las dos columnas la Data cruda sin depurar y la columna procesada la cual se le aplicó las funciones definidas y se usó el .head para mostrar las 20 primeras filas para una mejor visualización.

```
#comparacion de datos crudos con datos depurados
dataframe_Recom_Medic.iloc[:, [13, 41]].head(20)
```

	15. Describa ¿qué recomendaciones saludables le dio a conocer su médico? ejemplo: mi doctor me recomendó que caminará 4 veces por semana por 40 minutos - Sin depurar	15. Describa ¿qué recomendaciones saludables le dio a conocer su médico? ejemplo: mi doctor me recomendó que caminará 4 veces por semana por 40 minutos - depurada
0	Dormir boca abajo	dormir boca abajo
1	Me dijo que haga terapias de respiración usual	me dijo que haga terapias de respiracion usual
2	Ninguna	ninguna
3	Me dijo que aglutina la respiración y soltera	me dijo que aglutina la respiracion y soltera
4	Si	si
5	Que tomara cosas calientes	que tomara cosas calientes
6	Me recomendó que me cuidara que me podría dar	me recomendo que me cuidara que me podia dar
7	Evitar lugares con aglomeraciones, distanciamiento	evitar lugares con aglomeraciones distanciamiento
8	No me atendió con un médico	no me atendio con un medico
9	Hacer ejercicio caminar tomar el sol de 10 a	hacer ejercicio caminar tomar el sol de a de
10	Que haga actividad física constante	que haga actividad fisica constante
11	Que respiren	que respiren

Figura 50. Comparación de las columnas de texto. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

Se separó la información del Dataframe denominado dataframe_Recom_Medic, las x y y con las que se trabajó como x se ubicó la columna 41 que corresponde a '15. Describa ¿Qué Recomendaciones saludables le dio a conocer su médico? ejemplo: Mi doctor me recomendó que caminará 4 veces por semana por 40 minutos - Depurada' y las columnas de la 15 hasta la 41 se eligió las más significativas que corresponden a las columnas de "ejercicios", "Reposo", "Vitaminas", "terapia respiratoria", "Hidratado", "alimentación", "mascarilla", "ejercicios respiratorios", "vaporizaciones", "aislarme", "distanciamiento", "aseo", "No visite al médico", "Ninguna".


```

[76] # entrada columna depurada de la dataset
x_inicial = dataframe_Recom_Medic.iloc[:, [41]]

[79] # columna 15 afectacion psicologica hasta columna y = df_new_sintomas.iloc[:, 15:91] se eligen valores mayores a 200 datos
y_inicial = dataframe_Recom_Medic.loc[:, ["ejercicios", "Reposo", "Vitaminas", "terapia respiratoria", "Hidratado", "alimentacion", "mascarilla",

```

Figura 51. Separación de variables a utilizar del Dataset. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

```

[78] x_inicial.head(10)

```

	15. Describa ¿Qué Recomendaciones saludables le dio a conocer su médico? ejemplo: Mi doctor me recomendó que caminará 4 veces por semana por 40 minutos - Depurada
0	duermir boca abajo
1	me dijo que haga terapias de respiracion cuand...
2	ninguna
3	me dijo que aguantara la respiracion y soflara...
4	si
5	que tomara cosas calientes
6	me recomendo que me cuidara que me podria dar...
7	evitar lugares con aglomeraciones distanciamas...
8	no me atendí con un medico
9	hacer ejercicios caminar tomar el sol de a de...

Figura 52. Datos que contiene x. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

```

[94] x_inicial

```

	ejercicios	Reposo	Vitaminas	terapia respiratoria	Hidratado	alimentacion	mascarilla	ejercicios respiratorios	aglomeraciones	aislarme	distanciamiento	asno	no visita al médico	Ninguna
0	0	1.0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0.0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	0.0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0.0	0	1	0	0	0	1	0	0	0	0	0	0
4	0	0.0	0	0	0	0	0	0	0	0	0	0	0	0
...
#126	0	0.0	0	0	0	0	0	0	0	0	0	0	0	0
#126	0	0.0	0	0	0	0	0	0	0	0	0	0	0	0
#127	0	0.0	0	0	0	0	0	0	0	0	0	0	0	0
#128	0	0.0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 53. Datos que contiene y. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

3.4.2.1 Aplicación de Técnicas

Se aplico la técnica de StopWords a la columna '15. Describa ¿Qué Recomendaciones saludables le dio a conocer su médico? ejemplo: Mi doctor me recomendó que caminará 4 veces por semana por 40 minutos - Depurada', se importó el método de la librería nltk y se descargó el tokenizador de oraciones "Punkt" este divide el texto de manera automatizada mediante un algoritmo no supervisado y el método "stopwords" a continuación se creó una lista de palabras que tiene guardada internamente el método stopwords las cuales están guardadas en la variable txt_stop_words, luego se aplica esto al dataframe_Recom_Medic con la función lambda y se crea la columna de Datos sin Stopwords.

18. describe qué recomendaciones saludables le dio a conocer su médico? ejemplo: mi doctor me recomendó que caminara 4 veces por semana por las afueras - Depende		Datos sin stopwords
0	debería hacer algo	debería hacer algo
1	me dijo que haga terapia de respiración nasal	algo haga terapia respiratoria como fumar...
2	me dijo que apretara la respiración y solara	algo apretara respiración solara totalmente
3	me recomendó que me cuidara que me podía dar	recomendó cuidara podía dar nuevamente
4	evitar lugares con aglomeraciones distantes	evitar lugares aglomeraciones distantes...
5045	que tenga mucho cuidado	cuidado
5046	apretado hasta notar	apretado hasta notar
5052	me recomendaron como terapia asada	recomendaron como terapia asada
5061	muchas aperturas respiratorias y medidas de	aperturas respiratorias medidas
6043	hacer paciencia comer bien hacer ejercicio bal	hacer paciencia comer bien hacer ejercicio bal
3644 rows x 3 columns		

Figura 54. Datos sin StopWords. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

Se definió una variable lista_serie_palabras en la cual se guardo todas las palabras recibidas y realizar un conteo del total de palabras que se recibe, e total se reciben 12.190 palabras, luego se definió un laso for para verificar la frecuencia de cada palabra.

```
resfrio:1
duele:1
tocar:1
estornudar:1
debil:1
activar:1
cerebro:1
ejercicios:1
piscina:1
desconocia:1
realiza:1
gradualmente:1
ahislado:1
amejor:1
graso:1
_
```

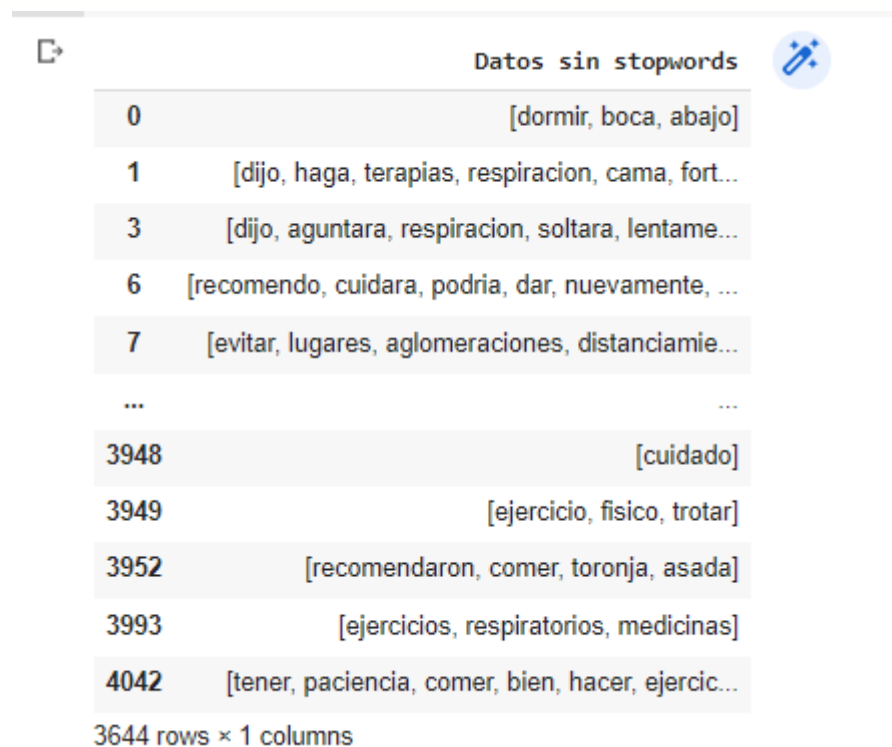
Figura 55. Frecuencia de palabras. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

Se consulto de acuerdo con los percentiles cuantas palabras diferentes se reciben en la x_st la cual es la entrada del modelo a entrenar.

```
count    3644.000000
mean      4.453622
std       3.403231
min       1.000000
25%      2.000000
50%      4.000000
75%      6.000000
85%      7.000000
90%      9.000000
95%     11.000000
97%     12.000000
99%     16.000000
max      49.000000
Name: Datos sin stopwords, dtype: float64
```

Figura 56. Palabras diferentes recibidas. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

Se uso otra técnica denominada Tokenización llamando el método `word_tokenize` y esto se aplica a la columna de Datos sin stopwords realizando una separación de palabras como se muestra a continuación:



	Datos sin stopwords
0	[dormir, boca, abajo]
1	[dijo, haga, terapias, respiracion, cama, fort...]
3	[dijo, aguntara, respiracion, soltara, lentame...]
6	[recomendo, cuidara, podria, dar, nuevamente, ...]
7	[evitar, lugares, aglomeraciones, distanciamie...]
...	...
3948	[cuidado]
3949	[ejercicio, fisico, trotar]
3952	[recomendaron, comer, toronja, asada]
3993	[ejercicios, respiratorios, medicinas]
4042	[tener, paciencia, comer, bien, hacer, ejercic...]

3644 rows × 1 columns

Figura 57. Aplicación de Tokenización. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

Para entrar a la fase de modelamiento se llamo a las respectivas librerías y se separo las x de entrenamiento / prueba y las y de entrenamiento / prueba.

Se identifico la cantidad de datos distintos que se va a manejar en la data de entrada la cual será guardada en la variable `tokenizer` y se convierten los datos a numéricos para que el modelo pueda procesarlo y aprender de estos.

3.4.2.2 Elección del Modelo

Se desarrollo diversos algoritmos Long Short-Term Memory (LSTM), Modelo básico ANN, Modelo NLP con clasificador Random Forest para multilabel o multi etiqueta y KNeighborsClassifier (kNN) k Vecino cercano. Para realizar la comparación entre ellos y exponer cual es el que presenta mayor asertividad en cuanto a recomendaciones dadas por el especialista medico a sus pacientes contagiados de covid-19.

A continuación se detalla el algoritmo utilizado para el modelo que presento el mayor accuracy en entrenamiento y prueba. El modelo Long Short-Term Memory (LSTM), en la cual se utilizó 60 epoch con 15 neuronas.

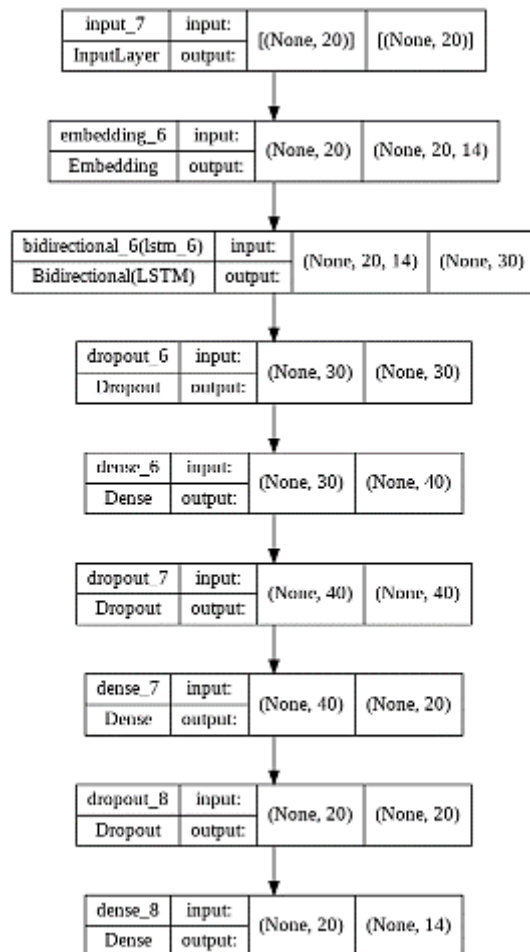


Figura 58. Arquitectura del modelo Básico ANN. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

3.4.2.3 Configuración de Parámetros

Long Short-Term Memory (LSTM)

Al definir el modelo LSTM lo siguiente ejecutado es la compilación con la librería keras ya que esta trabaja con valores numéricos, se utiliza la función de pérdida “binary_crossentropy” la cual calcula la pérdida de entropía cruzada binaria, con “adam” se usa como un optimizador de sustitución para el descenso de gradiente estocástico y una función de activación “sigmoid”.

Utilizando diversos parámetros variando el tamaño de batch_size, epochs, validación_split para detectar el modelo optimizado.

Artificial Neural Network (ANN) basic model

En la definición del modelo ANN se utilizó una función de activación “sigmoid” y 'softmax' lo cual devolvera una salida que será por la neurona dada una entrada, realizando variantes en los valores de batch_size, epochs, validación_split para revelar el modelo optimizado.

Random Forest for multilabel

En la configuración de parámetros para este modelo se utilizó número de estimadores esto refiere a la cantidad de combinaciones o arboles a usar y max_features 'sqrt' y 'log2'.

KNeighborsClassifier (kNN)

Definiendo este modelo se usó el numero de vecinos y la métrica la distancia “minkowski” la cual se utiliza por defecto en el modelo.

3.4.2.4 Evaluación del Modelo

Se ejecutaron los diversos modelos con las entradas x que corresponden a los Datos tokenizados, con la función .fit, se proporciona los ajustes del modelo para realizar el respectivo entrenamiento mediante epoch ajustando el número de iteraciones a realizar, history es una variable de devolución, validation_split corresponde a la división de los datos de validación y entrenamiento, verbose muestra la generación de cada epoch, finalmente batch_size integra la cantidad de ejemplos de capacitación utilizados en una iteración.

Entrenamiento del Algoritmo

Se ajusto los pesos que ingresan a las neuronas de la red, realizando la compilación en el modelo Básico ANN.

```
Epoch 1/200
292/292 [#####] - 14s 30ms/step - loss: 0.5903 - accuracy: 0.2889 - precision_0: 0.2382 - recall_0: 0.0995 - auc_0: 0.0483 - val_loss: 0.2898 - val_accuracy: 0.3911 - val_precision_0: 0.2067
Epoch 2/200
292/292 [#####] - 6s 21ms/step - loss: 0.2841 - accuracy: 0.3718 - precision_0: 0.4048 - recall_0: 0.0522 - auc_0: 0.7563 - val_loss: 0.2727 - val_accuracy: 0.3911 - val_precision_0: 0.000000
Epoch 3/200
292/292 [#####] - 4s 12ms/step - loss: 0.2226 - accuracy: 0.4738 - precision_0: 0.7786 - recall_0: 0.2287 - auc_0: 0.8528 - val_loss: 0.2147 - val_accuracy: 0.5077 - val_precision_0: 0.8112
Epoch 4/200
292/292 [#####] - 4s 12ms/step - loss: 0.2392 - accuracy: 0.5625 - precision_0: 0.8248 - recall_0: 0.3218 - auc_0: 0.8908 - val_loss: 0.2040 - val_accuracy: 0.5015 - val_precision_0: 0.8409
Epoch 5/200
292/292 [#####] - 4s 12ms/step - loss: 0.2818 - accuracy: 0.6222 - precision_0: 0.8222 - recall_0: 0.3738 - auc_0: 0.9375 - val_loss: 0.1943 - val_accuracy: 0.5081 - val_precision_0: 0.8040
Epoch 6/200
292/292 [#####] - 4s 13ms/step - loss: 0.1668 - accuracy: 0.6478 - precision_0: 0.8348 - recall_0: 0.4475 - auc_0: 0.9334 - val_loss: 0.1868 - val_accuracy: 0.6186 - val_precision_0: 0.7788
Epoch 7/200
292/292 [#####] - 4s 12ms/step - loss: 0.2394 - accuracy: 0.6038 - precision_0: 0.8828 - recall_0: 0.4244 - auc_0: 0.9434 - val_loss: 0.1815 - val_accuracy: 0.5989 - val_precision_0: 0.7789
Epoch 8/200
292/292 [#####] - 4s 12ms/step - loss: 0.2414 - accuracy: 0.6913 - precision_0: 0.8173 - recall_0: 0.5484 - auc_0: 0.9539 - val_loss: 0.1836 - val_accuracy: 0.6072 - val_precision_0: 0.7990
Epoch 9/200
292/292 [#####] - 4s 12ms/step - loss: 0.2234 - accuracy: 0.7054 - precision_0: 0.8168 - recall_0: 0.5812 - auc_0: 0.9688 - val_loss: 0.1752 - val_accuracy: 0.6141 - val_precision_0: 0.7624
Epoch 10/200
292/292 [#####] - 4s 11ms/step - loss: 0.2263 - accuracy: 0.7253 - precision_0: 0.8261 - recall_0: 0.6283 - auc_0: 0.9642 - val_loss: 0.1739 - val_accuracy: 0.6072 - val_precision_0: 0.7692
Epoch 11/200
292/292 [#####] - 4s 12ms/step - loss: 0.2258 - accuracy: 0.7987 - precision_0: 0.8412 - recall_0: 0.6521 - auc_0: 0.9708 - val_loss: 0.1706 - val_accuracy: 0.6581 - val_precision_0: 0.7688
Epoch 12/200
292/292 [#####] - 4s 12ms/step - loss: 0.2119 - accuracy: 0.7663 - precision_0: 0.8388 - recall_0: 0.6772 - auc_0: 0.9728 - val_loss: 0.1701 - val_accuracy: 0.6123 - val_precision_0: 0.7872
Epoch 13/200
292/292 [#####] - 4s 11ms/step - loss: 0.1676 - accuracy: 0.7822 - precision_0: 0.8586 - recall_0: 0.6998 - auc_0: 0.9745 - val_loss: 0.1663 - val_accuracy: 0.6484 - val_precision_0: 0.7948
Epoch 14/200
292/292 [#####] - 4s 12ms/step - loss: 0.2007 - accuracy: 0.7972 - precision_0: 0.8636 - recall_0: 0.7205 - auc_0: 0.9782 - val_loss: 0.1724 - val_accuracy: 0.6381 - val_precision_0: 0.7712
Epoch 15/200
292/292 [#####] - 4s 13ms/step - loss: 0.2025 - accuracy: 0.7946 - precision_0: 0.8672 - recall_0: 0.7207 - auc_0: 0.9775 - val_loss: 0.1761 - val_accuracy: 0.6638 - val_precision_0: 0.7734
Epoch 16/200
292/292 [#####] - 4s 11ms/step - loss: 0.1938 - accuracy: 0.8258 - precision_0: 0.8739 - recall_0: 0.7588 - auc_0: 0.9888 - val_loss: 0.1636 - val_accuracy: 0.6988 - val_precision_0: 0.8018
Epoch 17/200
```

Figura 59. Entrenamiento de la Red Básica ANN. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

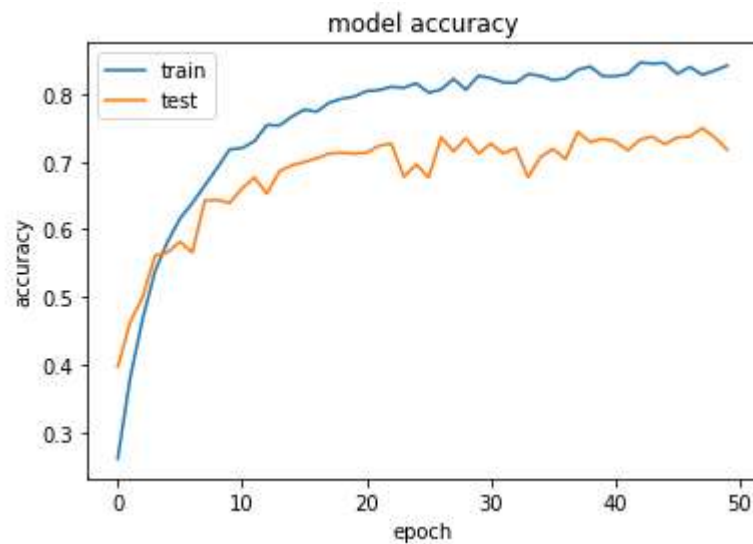


Figura 60. accuracy del modelo Básico ANN. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

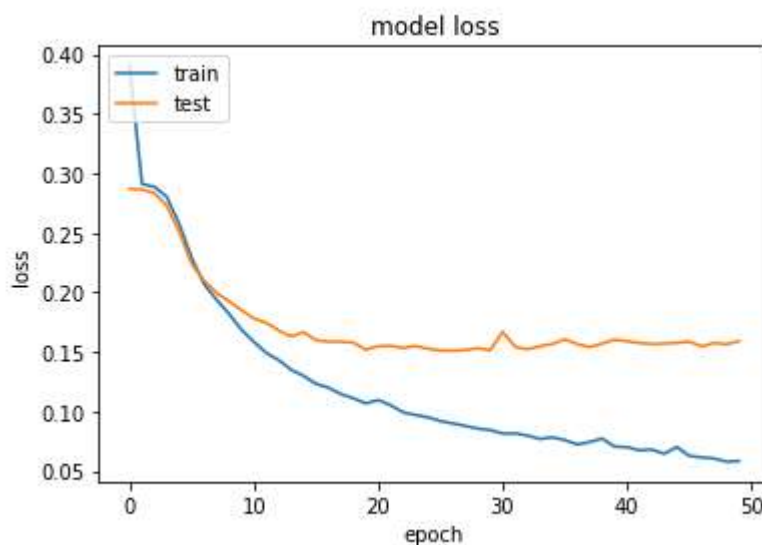


Figura 61. loss del modelo Básico ANN. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

3.4.2.5 Comparación de métricas

Para la ejecución del accuracy de los diversos algoritmos utilizados se definió un dataframe en el cual se integran los resultados de la asertividad generada en cada modelo las primeras 5 variantes son del modelo LSTM, seguido del modelo básico ANN, finalmente Random forest para multilabel y KNN vecino cercano detallando el valor de accuracy, precisión recall, auc de train y test.

Version	Clasificador	batch_size	epochs	validation_split	Accuracy_train	Precision_train	Recall_train	AUC_train	Accuracy test	Precision test	Recall test	AUC test
0	0	LSTM (15)	6.0	100.0	0.2	0.836364	0.956221	0.923590	0.983456	0.714678	0.826327	0.747 0.920944
1	1	LSTM (15)	12.0	50.0	0.2	0.839451	0.936543	0.846686	0.979191	0.721536	0.816514	0.712 0.925786
2	2	LSTM (15)	8.0	70.0	0.2	0.837050	0.941008	0.891444	0.984502	0.728395	0.822545	0.737 0.928760
3	3	LSTM (15)	8.0	50.0	0.3	0.802058	0.918050	0.875371	0.976678	0.695473	0.809989	0.746 0.934662
4	4	LSTM (15)	12.0	60.0	0.3	0.848027	0.907239	0.836795	0.977465	0.727023	0.785874	0.701 0.931502
5	5	LSTM (15)	12.0	50.0	0.3	0.796913	0.923925	0.829882	0.974849	0.705075	0.802286	0.702 0.930411
6	6	ANN	8.0	100.0	0.2	0.883019	0.947930	0.922049	0.986486	0.757202	0.822665	0.784 0.930666
7	7	ANN	8.0	50.0	0.2	0.892058	0.946223	0.913639	0.98389	0.760969	0.812371	0.788 0.937411
8	8	ANN	12.0	50.0	0.2	0.841166	0.964093	0.663947	0.971729	0.721536	0.849711	0.588 0.907352
9	9	ANN	6.0	50.0	0.3	0.798628	0.944601	0.661968	0.949225	0.696845	0.830056	0.591 0.884644

Figura 62. Dataframe evaluación_métricas. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

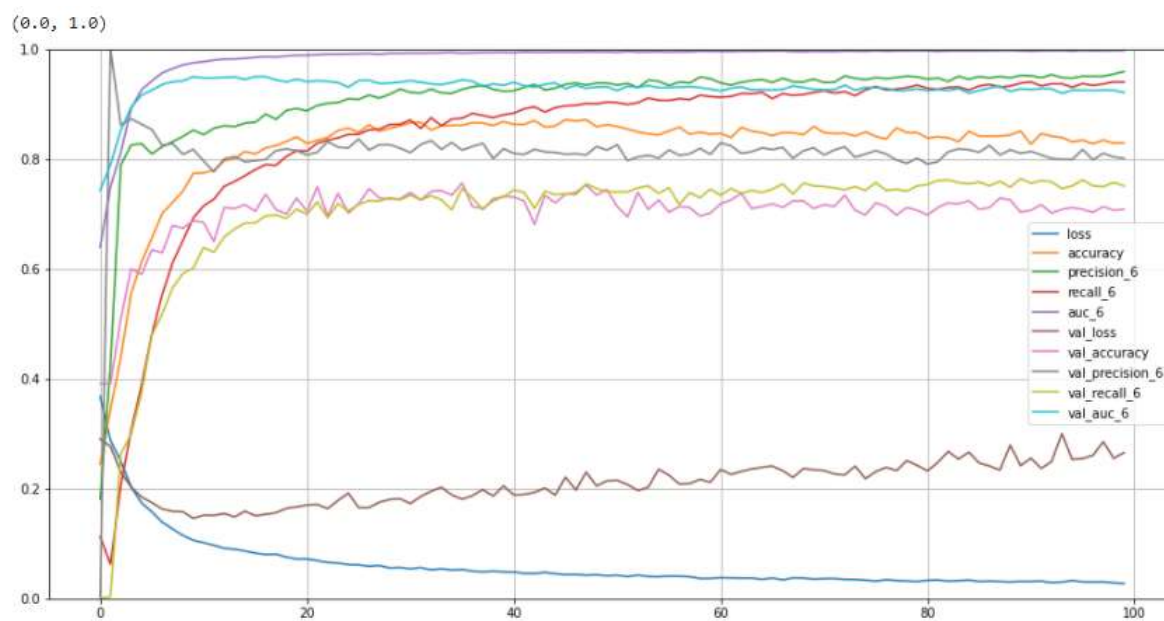


Figura 63. Métricas del modelo efectivo. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

clasificador	Version	numero estimadores	caracteristica	Accuracy score	F1 score	Precision score	Recall score	Auc score	metrica	numero vecinos
0	Random Forest	0	200.0	sqrt	0.447188	0.573099	0.818182	0.441 0.715177	NaN	NaN
1	Random Forest	1	300.0	sqrt	0.447188	0.577373	0.825279	0.444 0.716895	NaN	NaN
2	Random Forest	2	400.0	sqrt	0.444444	0.576998	0.823748	0.444 0.716840	NaN	NaN
3	Random Forest	3	100.0	log2	0.448560	0.579016	0.821691	0.447 0.718232	NaN	NaN
4	Random Forest	4	300.0	log2	0.447188	0.577373	0.825279	0.444 0.716895	NaN	NaN
5	kNN	0	NaN	NaN	0.325103	0.577373	0.825279	0.444 0.716895	minkowski	5.0
6	kNN	1	NaN	NaN	0.270233	0.577373	0.825279	0.444 0.716895	minkowski	15.0

Figura 64. df_metrics_rf1. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

Llegando a la conclusión que el modelo ANN presenta un accuracy train de 88%, una precisión train de 94%, recall de 92% y auc train de 98%, seguido de los datos de test accuracy test de 75%, precisión test de 82%, recall test 78% y auc test de 93%.

3.4.2.6 Evaluación del Modelo

Se definió una variable para evaluar el modelo efectivo con la siguiente frase `produccion_test`. Se tokenizo la frase y finalmente se aplica el modelo identificado con mayor asertividad.

```
[117] tokenizer.sequences_to_texts([x_test[produccion_test]])
```

```
['caminar hacer ejercicios comida saludable']
```

ejercicios	1.0
Reposo	0.0
Vitaminas	0.0
terapia respiratoria	0.0
Hidratado	0.0
alimentacion	1.0
mascarilla	0.0
ejercicios respiratorios	0.0
vaporizaciones	0.0
aislar me	0.0
distanciamiento	0.0
aseo	0.0
No visite al médico	0.0
Ninguna	0.0

Figura 64. Pruebas del modelo efectivo. Tomada de investigación directa. Elaborado por Solórzano Monserrate Mirian

3.5 Conclusiones

- Se tiene una revisión bibliográfica de 53 fuentes entre las cuales destacan papers, tesis doctorales, tesis universitarias, documentales etc.,.
- Se conto con una dataset de 4141 datos con información de recomendaciones empleadas por médicos especialistas a pacientes con covid-19 durante el acompañamiento y post covid
- Para la depuración de la data set obtenida se aplicó una normalización del texto la cual está compuesta por la eliminación de tildes, caracteres especiales y conversión de palabras mayúsculas a minúsculas
- Luego de realizar la comparación de varios modelos el más efectivo es básico Artificial neural network (ANN) configurando los parámetros presenta una mejora considerable arrojando las siguientes métricas precisión de 94% y un accuracy de 81% respecto al train y en el test una precisión de 81% y accuracy de 69%.

3.6 Recomendaciones

Se recomienda lo siguiente:

- Se sugiere impulsar una futura investigación que incluya implementar este modelo efectivo a una aplicación web o portal que este disponible a la comunidad medica.
- Se continúe impulsando nuevos estudios NLP aplicando los nuevos modelos basados en Transformer, Elmo, Bert para comparar si estos mejoran los resultados obtenidos con los modelos ya probados.

Anexos

Anexo 1

Entrevista realizada a los Profesionales de IA y afines



- 1. Fecha de la Entrevista:** 3/05/2022
- 2. Nombre:** Jostin Marcelo Maldonado Flores
- 3. Grado de educación:** Tercer nivel de educación culminado
- 4. Título de educación:** Ingeniero en Electrónica
- 5. Experiencia laboral:**

Se ha desempeñado como desarrollador de Software en C#, C++ o Java, también como arquitecto desarrollador referente a bases de datos y servidores en la nube.

6. Edad

29 años

7. Género

Masculino

8. Lugar de Residencia

Guayaquil

9. ¿Tiene conocimientos de Inteligencia Artificial?

Si

10. Años de experiencia trabajando en temas o proyectos de Inteligencia Artificial

4 años

11. ¿Tiene conocimientos de la rama de Inteligencia Artificial llamada Machine Learning (Aprendizaje Automático)?

Si

12. ¿Posee conocimientos de la rama de Inteligencia Artificial denominada Procesamiento de Lenguaje Natural (NLP)?

Si

13. ¿Qué tan importante considera usted el uso de tecnologías como la Inteligencia Artificial y soluciones de NLP para la superación de la actual de la pandemia?

Es extremadamente importante debido a que la Inteligencia Artificial juega un rol importante en la actual pandemia, dado que han salido diversos proyectos referentes a si las personas cumplen con los 2 metros de distancia, si existe acumulación de personas, si las personas usan correctamente la mascarilla y cuál es la temperatura corporal de las personas, todo esto a través de los algoritmos desarrollados que envían una alerta si alguna de estas condiciones no se cumple.

14. ¿Qué Algoritmo considera usted más adecuado para usarlo en una arquitectura NLP a ser creada para Clasificación de conversaciones de textos de personas contagiadas de covid-19?

Arboles de decisiones, Naive Bayes, Redes neuronales.

En el algoritmo de "Árboles de decisiones" la precisión de la predicción será mayor de acuerdo con la cantidad de almacenamiento que se esté manejando. En el algoritmo de "Redes Neuronales" conservan los pesos, pero a medida que se utilizan mejores resultados se obtendrán.

15. De la lista de modelos NLP cuál considera usted que seria los más adecuados usarlos con información textual clasificada relacionada con el Covid, para el descubrimiento de tendencias en la población que está enfrentando el Covid-19.?

Transformer, GPT3 (Generative Pre-trained Transformer 3), RNN (Recurrent Neural Networks)

Estos modelos de NLP facilitan las formas de aprendizaje y hacen que al utilizarlas en información textual clasificada sea mucho más sencillo usarlas.

16. ¿Considera usted que aún se necesita más investigación y nuevas propuestas de construcción de NLP para crear modelos más efectivos de conversaciones de textos con respecto a los existentes relacionados para combatir el Covid-19?

Totalmente de acuerdo debido a que los modelos actuales aún no cumplen con todas las expectativas.

17. ¿ Sabía usted que uno de los beneficios de aplicar Técnicas de NLP es la simplificación de interacción entre la máquina y el ser humano?

Medianamente Informado

18. ¿Considera usted que en futuros proyectos o trabajos investigativos será de utilidad el análisis de técnicas de procesamiento de lenguaje natural NLP para clasificación de texto de conversaciones textuales de personas contagiadas con Covid-19?

Totalmente de acuerdo debido a que si seguimos utilizando el mismo modelo en cualquier momento podría haber una nueva pandemia y para ese momento ya tendremos nuestro modelo mucho más desarrollado y evitaremos cometer los mismo errores que cometimos al principio de la pandemia.



1. Fecha de la Entrevista: 3/05/2022

2. Nombre: Ricardo Manuel Prieto Galarza

3. Grado de educación: Doctorado

4. Título de educación:

Ingeniero Electrónico con una especialidad en Telecomunicaciones y el título de Máster (actualmente cursando) es en Inteligencia Artificial.

5. Experiencia laboral

He trabajado en Claro Ecuador como capacitador en el área de ciencias de datos. He trabajado en el Banco Central y en otras entidades financieras como Cooperativa Biblián, Cooperativa Jardín Azuayo en temas de seguridad informática. A nivel de instituciones ha trabajado en la Universidad Politécnica Salesiana y en la Universidad Estatal en la ciudad de Cuenca. También se ha desarrollado como profesor en el Instituto Superior Juan Montalvo en la ciudad de Loja.

6. Edad

33 años

7. Género

Masculino

8. Lugar de Residencia

Cuenca

9. ¿Tiene conocimientos de Inteligencia Artificial?

Si

10. Años de experiencia trabajando en temas o proyectos de Inteligencia Artificial

3 años

11. ¿Tiene conocimientos de la rama de Inteligencia Artificial llamada Machine Learning (Aprendizaje Automático)?

Si

12. ¿Posee conocimientos de la rama de Inteligencia Artificial denominada Procesamiento de Lenguaje Natural (NLP)?

Si

13. ¿Qué tan importante considera usted el uso de tecnologías como la Inteligencia Artificial y soluciones de NLP para la superación de la actual de la pandemia?

Es extremadamente importante porque facilita mucho el saber cuáles son las opiniones de las personas con respecto a un tema. Y en temas de salud la pandemia nos ha afectado en varios ámbitos de nuestra vida, por lo que la Inteligencia Artificial y el uso de las técnicas de procesamiento natural son sumamente importantes.

14. ¿Qué Algoritmo considera usted más adecuado para usarlo en una arquitectura NLP a ser creada para Clasificación de conversaciones de textos de personas contagiadas de covid-19?

Naive Bayes, Redes neuronales, Otra(Especifique)

El algoritmo más destacable son Redes Neuronales solo que debe tener en cuenta que la cantidad y la calidad de los datos deben ser bastante amplia, es decir se deben tener suficientes datos.

Naive Bayes también trabaja bastante bien, su desventaja es que no son tan precisas pero usan menor cantidad de datos.

LSTM de la extensión de redes neuronales recurrentes.

15. De la lista de modelos NLP cuál considera usted que seria los más adecuados usarlos con información textual clasificada relacionada con el Covid, para el descubrimiento de tendencias en la población que está enfrentando el Covid-19.?

Transformer, ELMo (Embeddings from Language Models), GPT3 (Generative Pre-trained Transformer 3)

ELMO se puede utilizar para textos en español y es muy importante. Transformers y GPT3 por su parte se encuentran relacionadas por lo que también son bastantes destacables.

16. ¿Considera usted que aún se necesita más investigación y nuevas propuestas de construcción de NLP para crear modelos más efectivos de conversaciones de textos con respecto a los existentes relacionados para combatir el Covid-19?

Totalmente de acuerdo porque siempre se necesita continuar con el desarrollo de la tecnología más aún en temas de pandemia.

17. ¿ Sabía usted que uno de los beneficios de aplicar Técnicas de NLP es la simplificación de interacción entre la máquina y el ser humano?

Totalmente Informado

18. ¿Considera usted que en futuros proyectos o trabajos investigativos será de utilidad el análisis de técnicas de procesamiento de lenguaje natural NLP para clasificación de texto de conversaciones textuales de personas contagiadas con Covid-19?

Totalmente de acuerdo porque es muy importante.



1. Fecha de la Entrevista: 3/07/2022

2. Nombre: Aristo Cabrera Torres

3. Grado de educación: Doctorado

4. Título de educación:

Maestría en administración de empresas, maestría en business analíticos.

5. Experiencia laboral

En realidad inteligencia artificial cubre algunos tópicos que uno comúnmente, piensa que inteligencia artificial es algo así como maquinas autónomas o que se desenvuelven solas. inteligencia artificial incluye otros temas como aprendizaje automático, entre otras cosas. por supuesto la parte analítica o implementación de datos, o una intersección entre la parte matemática y la parte estadística y propiamente los negocios. como experiencia yo he trabajado en Indurama que es una empresa que se dedica a la fabricación de electrodomésticos donde he trabajado temas de inteligencia, lo que se trata es básicamente es monitorear grupos de empresas que pertenecían a la corporación e ir entendiendo el desempeño, de crear los indicadores para permitir a la empresa generar conocimientos de los datos que permitan tomar mejores decisiones. he trabajo en varios proyectos de este tipo en temas de inteligencia artificial en lo que tiene que ver análisis relacionados con campañas políticas donde uno puede conocer que es lo que la gente habla en las redes sociales, allí si de lenguaje natural donde uno analiza documentos de lo que la gente expresa en redes sociales para ir tomando cambios y estrategias en las campañas electorales, también he trabajado en proyectos midiendo el desempeño de los gastos municipales en un proyecto a través de la municipalidad del Ecuador, donde debía obtener una calificación de desempeño de los gastos municipales. algunos proyectos de aprendizaje propios y de interés en empresas donde realizan estudios de mercados, donde la idea es identificar patrones para brindar retroalimentaciones sobre el comportamiento de personas o productos.

6. Edad

53 años

7. Género

Masculino

8. Lugar de Residencia

Guayaquil

9. ¿Tiene conocimientos de Inteligencia Artificial?

Si

10. Años de experiencia trabajando en temas o proyectos de Inteligencia Artificial

4 años

11. ¿Tiene conocimientos de la rama de Inteligencia Artificial llamada Machine Learning (Aprendizaje Automático)?

Si

12. ¿Posee conocimientos de la rama de Inteligencia Artificial denominada Procesamiento de Lenguaje Natural (NLP)?

Si

13 ¿Qué tan importante considera usted el uso de tecnologías como la Inteligencia Artificial y soluciones de NLP para la superación de la actual de la pandemia?

En realidad, creo que se ha visto un crecimiento enorme en el uso de la inteligencia artificial sacando provecho, de aquí en adelante seguramente las empresas usaran la inteligencia artificial para entender de mejor manera y ver desde un punto de vista remoto como lo hemos visto en los últimos 2 años, funcionando de manera remota entonces deben replantear y evitar la interacción física, un claro ejemplo es Amazon, donde usan la inteligencia artificial dándole la facilidad de adquisición a sus clientes usando campañas orientadas a cada persona, donde una persona ve cosas de su interés, esto lo podríamos aplicar en varios sectores.

14. ¿Qué Algoritmo considera usted más adecuado para usarlo en una arquitectura NLP a ser creada para Clasificación de conversaciones de textos de personas contagiadas de covid-19?

Modelos Lineales, Redes neuronales, serían los que mejor se adapten a las necesidades.

15. De la lista de modelos NLP cuál considera usted que seria los más adecuados usarlos con información textual clasificada relacionada con el Covid, para el descubrimiento de tendencias en la población que está enfrentando el Covid-19.?

GloVe (Global Vectors for Words Representations), LSTM (Long-Short Term Memory), RNN (Recurrent Neural Networks)

LSTM- una red neuronal que permite hacer reducciones, que nos permiten hacer una reducción de los textos RNN - procesamiento de lenguaje natural que están en crecimiento.

16. ¿Considera usted que aún se necesita más investigación y nuevas propuestas de construcción de NLP para crear modelos más efectivos de conversaciones de textos con respecto a los existentes relacionados para combatir el Covid-19?

La tecnología no se detiene siempre es posible hacer algo más con nuevos elementos, en caso concreto de Covid podemos analizar cosas más allá de temas estadísticos. en varios impactos psicológicos, educativos.

17. ¿Sabía usted que uno de los beneficios de aplicar Técnicas de NLP es la simplificación de interacción entre la máquina y el ser humano?

Totalmente Informado

18. ¿Considera usted que en futuros proyectos o trabajos investigativos será de utilidad el análisis de técnicas de procesamiento de lenguaje natural NLP para clasificación de texto de conversaciones textuales de personas contagiadas con Covid-19?

Yo creo que es importante, cuando estábamos en crisis de Covid surgieron muchos comentarios, algunos de ellos mencionaban que era un virus para eliminar a la humanidad lo importante es realizar análisis y ver criterios unificados ese es el beneficio de estas herramientas que nos permiten dar un paso gigante en la inundación del mundo de las redes sociales para manejar este volumen de información.

Anexo 2

Modelo de la encuesta realizada a los habitantes de la Zona 8 de la Provincia del Guayas

Encuesta dirigida público en General

La Universidad de Guayaquil a través de sus investigadores impulsa la creación de soluciones tecnológicas que buscan ayudar a la comunidad en el corto o mediano plazo, ofreciendo herramientas tales como, por ejemplo: Asistentes Virtuales que proporcione de manera gratuita información relacionada a hábitos saludables que las personas contagiadas de Covid-19 deben manejar. Por este motivo se solicita su apoyo, con la siguiente encuesta que busca recopilar información necesaria para la construcción de este tipo de soluciones tecnológicas de IA.

PARTE 1: DATOS INFORMATIVO

1.1 Indique su Edad

1.2 Seleccione su Género

- ☐ Femenino
- ☐ Masculino

1.3 Lugar que reside de la zona 8 del Ecuador *

- ☐ Guayaquil
- ☐ Durán
- ☐ Samborondón
- ☐ Otra ciudad del Ecuador
- ☐ Otro País

1.4 De haber indicado que reside en otro lugar distinto a la zona 8 del Ecuador, indique el país, la provincia y su ciudad de residencia (o cantón)

PARTE 2: CORONA VIRUS (covid-19)

2.1 ¿Usted considera importante CONOCER cómo el CORONAVIRUS (Covid-19) afecta nuestra salud?

- ☐ Totalmente de Acuerdo
- ☐ Parcialmente de Acuerdo
- ☐ De Acuerdo
- ☐ Parcialmente en Desacuerdo
- ☐ Totalmente en Desacuerdo

2.2 ¿Usted está de Acuerdo que las vacunas contra el coronavirus (Covid-19) son efectiva eliminando el virus?

- ☐ Totalmente de Acuerdo
- ☐ Parcialmente de Acuerdo
- ☐ De Acuerdo
- ☐ Parcialmente en Desacuerdo
- ☐ Totalmente en Desacuerdo

2.3 ¿Está de acuerdo que la información del coronavirus (Covid-19) que recibe del ministerio de salud o subcentro de salud por cualquier medio de comunicación es la adecuada y actualizada como hábitos saludables, evolución del virus etc.?

- ☐ Totalmente de Acuerdo
- ☐ Parcialmente de Acuerdo
- ☐ De Acuerdo
- ☐ Parcialmente en Desacuerdo
- ☐ Totalmente en Desacuerdo

2.4 ¿Sabía usted que aplicar HÁBITOS SALUDABLES cuando una persona esta contagiada de coronavirus (covid-19) disminuye el riesgo de afecciones graves incluso descartando hasta la muerte?

- ☐ Poseo alto conocimiento del tema
- ☐ Poseo bajo conocimiento del tema
- ☐ No tenía conocimiento

PARTE 3: ASISTENTES VIRTUALES

Llámesese asistente virtual, la interacción de comunicación entre un dispositivo móvil o página web y el ser humano

3.1 ¿Usted posee un smartphone básico (Teléfono móvil con acceso a internet)? *

- ☐ Si poseo uno de Gama Alta (costo > \$501)
- ☐ Si poseo uno de Gama Media (costo \$201 a \$500)
- ☐ Si poseo uno de Gama Baja con internet (costo < \$200)
- ☐ No poseo con internet, pero estoy en proceso de adquirir uno
- ☐ No poseo con internet y no planeo adquirirlo

3.2 ¿Sabía usted que la tecnología de INTELIGENCIA ARTIFICIAL es capaz de desarrollar aplicaciones móviles que permita interactuar y mantener información de forma actualizada para combatir el Coronavirus (Covid-19) en cualquier lugar del mundo, a cualquier hora, incluyendo fechas de feriado?

- ☐ Poseo alto conocimiento del tema
- ☐ Poseo bajo conocimiento del tema
- ☐ No tenía conocimiento

3.3 ¿Le gustaría contar con una aplicación móvil que le permita interactuar, mantener informado de forma actualizada para combatir al Coronavirus (COVID-19) sobre los Hábitos Saludables utilizando la tecnología de Inteligencia Artificial de forma GRATUITA?

- ☐ Totalmente de Acuerdo
- ☐ Parcialmente de Acuerdo
- ☐ Ni de Acuerdo ni en Desacuerdo
- ☐ Parcialmente en desacuerdo
- ☐ Totalmente en desacuerdo

Bibliográfica

- Berzal, F. (agosto de 2016). Breve historia de la inteligencia artificial: el camino hacia la empresa. *Cesce*(479), 46-73. Obtenido de <https://www.cesce.es/es/w/asesores-de-pymes/breve-historia-la-inteligencia-artificial-camino-hacia-la-empresa>
- Torres A. John A. (2021). Análisis de opinión sobre tuits del COVID-19 generados por usuarios ecuatorianos. *CEDAMAZ Revista del Centro de Estudio y Desarrollo de la Amazonia*. Carrera de Ingeniería en Sistemas/Computación, Universidad Nacional de Loja, Loja, Ecuador. Vol. 11, No. 1, pp. 70–77. Obtenido de <https://revistas.unl.edu.ec/index.php/cedamaz/article/download/1039/791/3185>
- Lugo Reyes, S., Maldonado, C. G., & Murata, C. (2014). Inteligencia artificial para asistir el diagnóstico clínico en medicina. *Alergia México*, 110-120. Obtenido de <https://revistaalergia.mx/ojs/index.php/ram/article/view/33/46>
- Boden, M. A. (2016). *Inteligencia artificial*. Madrid: Turner Publicaciones S.L. Obtenido de <https://books.google.com.ec/books?hl=en&lr=&id=LCnYDwAAQBAJ&oi=fnd&pg=PT3&dq=inteligencia+artificial+&ots=drSuwWeKp5&sig=dIVCAsKiMPJLYImRt670uVzD-84#v=onepage&q=inteligencia%20artificial&f=false>
- Iglesia, K. (2017). Word2Vec. Natural Language Engineering. *Cambridge University*, 155 – 162. *Ingeniería del Lenguaje Natural*, 23 (1), 155-162. doi:10.1017/S1351324916000334
- Mutiwokuziva, M.T., Chanda, M.W., Kadebu, P., Mukwazvure, A., & Gatora, T.T. (2017). A neural-network based chat bot. *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, 212-217. Obtenido de doi: 10.1109/CESYS.2017.8321268
- González, C. M., Varela, S., y Sandra, M. (2017). Aplicación de algoritmos no supervisados para la detección de tópicos de investigación. *Presentado en V Jornadas de Intercambio y Reflexión acerca de la Investigación en Bibliotecología* (págs. 1853-5631). La Plata: Instituto de Investigaciones en Humanidades y Cs Sociales- IdIHCS- (CONICET/UNLP). Obtenido de http://sedici.unlp.edu.ar/bitstream/handle/10915/73956/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y

- Otzen, Tamara y Manterola, Carlos. (2017). Técnicas de Muestreo sobre una Población a Estudio. *Revista Internacional de Morfología*, 35 (1), 227-232. Obtenido de <https://dx.doi.org/10.4067/S0717-95022017000100037>
- Otzen, Tamara, Manterola, Carlos, Rodríguez-Núñez, Iván, y García-Domínguez, Maricela. (2017). La Necesidad de Aplicar el Método Científico en Investigación Clínica: Problemas, Beneficios y Factibilidad del Desarrollo de Protocolos de Investigación. *International Journal of Morphology*, 35(3), 1031-1036. Obtenido de <https://dx.doi.org/10.4067/S0717-95022017000300035>
- Moreno, S. A. (02 de noviembre de 2017). ¿Qué es el PLN o Procesamiento de Lenguaje Natural?. Obtenido de <https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/>
- Conde, O. D. (2018). Inteligencia artificial con TensorFlow para predicción de comportamientos. *Trabajo Fin de Grado Inédito, Universidad de Sevilla, Sevilla*, 29-65. Obtenido de <https://idus.us.es/handle/11441/80122>
- García, F. I. (2018). Estudio de Word Embeddings y métodos de generación de Meta Embeddings. *Universidad de Pais Vasco*, 1-226. Obtenido de <https://core.ac.uk/download/pdf/168407311.pdf>
- Delgado, T. F., y Ochoa, C. W. (2018). Estudio de la Autonomia del vehiculo electronico kia soul aplicando máquinas de soporte vectorial en la ciudad de cuenca. Cuenca, Ecuador. Obtenido de <https://dspace.ups.edu.ec/bitstream/123456789/16162/1/UPS-CT007827.pdf>
- González, B. N., Estrada, S. V., y Febles, E. A. (12 de abril de 2018). *Scielo.sld.cu*. Obtenido de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1561-31942018000300014
- Gonzalez, L. (14 de septiembre de 2018). *AprendeIA*. Obtenido de Introducción al IDE Spyder: <https://aprendeia.com/ide-spyder-para-python/>
- Jabeen, H. (2018). Stemming y lematización en Python. *Datacamp*, 1-82.
- Gupta, R., & Jivani, A. (2018). Analyzing the Stemming Paradigm. *Information and Communication Technology for Intelligent Systems* (págs. 333-342). ICTIS. Obtenido de

https://www.researchgate.net/publication/319161033_Analyzing_the_Stemming_Paradigm

Hernández, M., y Vasquez, P. (2018). *METODOLOGÍA DE LA INVESTIGACIÓN: LAS RUTAS CUANTITATIVA, CUALITATIVA Y MIXTA*. México: McGRAW-HILL INTERAMERICANA EDITORES. Obtenido de http://www.biblioteca.cij.gob.mx/Archivos/Materiales_de_consulta/Drogas_de_Abuso/Articulos/SampieriLasRutas.pdf

Messina, V. C. (2018). Librería de métodos de poda en conjuntos de clasificadores para Scikit-Learn. *UAM_Biblioteca*, 21-71. Obtenido de <https://repositorio.uam.es/handle/10486/688291>

Molina, G. N. (2018). Infraestructura para la evaluación intrínseca de algoritmos de Stemming. *Repositorio de la Universidad de Matanzas Cuba*, 25-70. Obtenido de [http://cict.umcc.cu/repositorio/tesis/Trabajos%20de%20Diploma/Ingenier%C3%A1Da%20Inform%C3%A1tica/2018/Infraestructura%20para%20la%20evaluaci%C3%B3n%20intr%C3%ADnseca%20de%20algoritmos%20de%20stemming%20\(Noe1%20Molina%20Gonz%C3%A1lez\).pdf](http://cict.umcc.cu/repositorio/tesis/Trabajos%20de%20Diploma/Ingenier%C3%A1Da%20Inform%C3%A1tica/2018/Infraestructura%20para%20la%20evaluaci%C3%B3n%20intr%C3%ADnseca%20de%20algoritmos%20de%20stemming%20(Noe1%20Molina%20Gonz%C3%A1lez).pdf)

Matthew, P., Mark, N., Mohit, I., Matt, G., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, págs. 2227–2237. North American. Obtenido de <https://arxiv.org/abs/1802.05365>

Rouhiainen, L. P. (2018). *Inteligencia Artificial*. Barcelona: Artes Gráficas Huertas, S.A. Obtenido de https://www.planetadelibros.com/libros_contenido_extra/40/39307_Inteligencia_artificial.pdf

Yan, R. (2018). "Chitty-Chitty-Chat Bot": Deep Learning for Conversational AI. *International Joint Conference on Artificial Intelligence (IJCAI-18)*, 18, 5520-5526. Obtenido de <https://www.ijcai.org/proceedings/2018/0778.pdf>

Alias, G., y Cassanelli, R. (12 de agosto de 2019). *NLP aplicado a análisis de texto*. Obtenido de Universidad Nacional de Mar del Plata:

<http://rinfi.fi.mdp.edu.ar/bitstream/handle/123456789/354/GAlias-RCassanelli-TFG-II-2019.pdf?sequence=1&isAllowed=y>

De la Fuente, Sans. Ó. M. (4 de junio de 2019). *Google Colab: Python y Machine Learning en la nube*. Obtenido de <https://www.adictosaltrabajo.com/2019/06/04/google-colab-python-y-machine-learning-en-la-nube/>

Devlin, J., Chang, M.-W., & Lee, Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arxiv*, 2-16. Obtenido de <https://arxiv.org/pdf/1810.04805.pdf>

Lloret, Ega. J. A. (25 de agosto de 2019). *Procesamiento del Lenguaje Natural (PLN) / Procesamiento del lenguaje natural (NLP)*. Obtenido de https://www.researchgate.net/publication/337172055_Procesamiento_del_Lenguaje_Natural_PLN_Natural_Language_Processing_NLP

Díaz, Avalos, A. (2019). Herramienta automática para diferenciar zonas dialectales de México en Twitter. *Mexico, CDMX*, 16-53. Obtenido de Universidad Autonoma Metropolitana, Division de Ciencias de la Comunicaci ´ on y Dise ´ no. Obtenido de https://gabyrr.github.io/assets/documents/PT_AlejandroDiaz_2019.pdf

Martínez, R., Rodríguez, R. A., Vera, P., y Parkinson, C. (2019). Análisis de técnicas de raspado de datos en la web aplicado al Portal del Estado Nacional Argentino. *XXV Congreso Argentino de Ciencias de la Computación (CACIC)* (págs. 457-466). Córdoba-Argentina: Creative Commons Attribution-NonCommercial-ShareAlike. Obtenido de <http://sedici.unlp.edu.ar/handle/10915/91026>

Monjarás Ávila, Á. J., Bazán Suarez, A. K., Pacheco-Martínez, Z. K., Rivera Gonzaga. J. A., Zamarripa Calderón, J. E., & Cuevas Suárez, C. E. (2019). Diseños de Investigación. *Esaludyeducacion*, 119-122. Obtenido de <https://doi.org/10.29057/icsa.v8i15.4908>

Weng, J. (30 de agosto de 2019). NLP Text Preprocessing: A Practical Guide and Template. *Towards Data Science*. Obtenido de <https://towardsdatascience.com/nlp-text-preprocessing-a-practical-guide-and-template-d80874676e79>

Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. *Universidad de Cornell*, 1-14. Obtenido de <https://arxiv.org/abs/1906.04284>

- Wu, Gao, & Jiao, (2019). Multi-Label Classification Based on Random Forest Algorithm for Non-Intrusive Load Monitoring System. *Processes*. 7. 337. 10.3390/pr7060337.
- Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., & Chao, L. S. (2019). Learning Deep Transformer Models for Machine Translation. *Association for Computational Linguistics*, 1810–1822.
- Geographic National. (02 de diciembre de 2020). *Breve historia visual de la inteligencia artificial*. Obtenido de https://www.nationalgeographic.com.es/ciencia/breve-historia-visual-inteligencia-artificial_14419
- Alicante. (27 de mayo de 2020). *Revista La Vanguardia*. Obtenido de Revista La Vanguardia: <https://www.lavanguardia.com/vida/20200527/481425245148/la-onu-encarga-a-1millionbot-un-chatbot-sobre-la-covid-19-en-ecuador.html>
- Camacho, Á. M., & Navarro, Á. E. (2020). Procesamiento del lenguaje natural con Python. *Revista de Cómputo Aplicado*, 33-44. Obtenido de https://www.researchgate.net/publication/346152605_Procesamiento_del_lenguaje_natural_con_Python
- Chavez, V. L. (2020). *repositorio universitario lamolina Lima - Peru* . Obtenido de CARACTERIZACIÓN DEL PERFIL DEL INGRESANTE DE UNA UNIVERSIDAD PÚBLICA APLICANDO ALGORITMOS CLUSTERING K-PROTOTYPES Y K-MEDOIDS: <http://repositorio.lamolina.edu.pe/bitstream/handle/20.500.12996/4633/chavez-valderrama-ledvir-ayrton-walter.pdf?sequence=1&isAllowed=y>
- Nizama, J. (19 de octubre de 2020). *TECSUP*. Obtenido de Ciencia de datos con Python: <https://es.scribd.com/document/480639988/Ciencia-de-datos-con-Python>
- Diaz, C. F., & Toro, M. A. (2020). SARS-CoV-2/COVID-19: el virus, la enfermedad y la pandemia. *Salud*, 24(3), 1-205. Obtenido de <https://docs.bvsalud.org/biblioref/2020/05/1096519/covid-19.pdf>
- Zhao, L., Alhoshan, W. F., Letsholo, K., Ajagbe, M., Chioasca, E., & Riza, B. (07 de abril de 2020). *Natural Language Processing (NLP) for Requirements Engineering* . Obtenido de arxiv: <https://arxiv.org/ftp/arxiv/papers/2004/2004.01099.pdf>

- Zhao, L., Alhoshan, W., Ferrari, A., Letsholo, K., Ajagbe, M., Chioasca, E., & Riza, B. (2020). Natural Language Processing (NLP) for Requirements Engineering: A Systematic Mapping Study. *arxiv*, 3-75.
- Zheng, X., Zhang, C., & Woodland, P. C. (2021). ADAPTING GPT, GPT-2 AND BERT LANGUAGE MODELS FOR SPEECH RECOGNITION. *Cambridge University Engineering*, 2-7. Obtenido de <https://arxiv.org/pdf/2108.07789.pdf>
- Hamdy, E. (2021). Neural Models for Offensive Language Detection. *Lehrstuhl für Data Science*, 17-66. Obtenido de <https://arxiv.org/pdf/2106.14609.pdf>
- Marc. (9 de julio de 2020). *Uso de chatbots para la gestión de crisis y más*. Obtenido de onlim: <https://onlim.com/en/using-chatbots-for-crisis-management-and-beyond-part-1/>
- Moreno, F. S. (2020). Herramienta de Reconocimiento de Imágenes en Python. *Trabajo Fin de Grado Inédito, Universidad de Sevilla, Sevilla.*, 24-88. Obtenido de <https://idus.us.es/bitstream/handle/11441/102090/TFG-2877-MORENO%20FERNANDEZ.pdf?sequence=1&isAllowed=y>
- Hofesmann, E. (21 de enero de 2021). *The Machine Learning Lifecycle in 2021*. Obtenido de Towards Data Science: <https://towardsdatascience.com/the-machine-learning-lifecycle-in-2021-473717c633bc>
- Delgado, C. (29 de marzo de 2021). *Descubre que es un entorno de desarrollo integrado*. Obtenido de ourcodeworld: <https://ourcodeworld.co/articulos/leer/1469/que-es-un-ide-entorno-de-desarrollo-integrado>
- Ecuador, G. d. (2021). *Ministerio de Salud Pública del Ecuador*. Obtenido de Ministerio de Salud Pública del Ecuador: <https://www.salud.gob.ec/coronavirus-covid-19/>
- Khasanah, I. N. (2021). Sentiment Classification Using fastText Embedding and Deep Learning Model. *Procedia Computer Science*, 343-350. Obtenido de <https://www.sciencedirect.com/science/article/pii/S187705092101228X>
- Lugo, N. (2021). Métodos de investigación cualitativa. *Programa analítico de métodos cualitativos enfocados en etnografía*. (págs. 1-10). Monterrey: Departamento de medios y cultura digital.

- Martínez, S. J., Cruz, G. S., & López, C. J. (2021). Optimización de un portafolio con Python. *Pädi*, 132-135. Obtenido de <http://portal.amelica.org/ameli/jatsRepo/595/5952727024/5952727024.pdf>
- McKenzie, G. D., & Adams, B. (2021). Natural Language Processing in GIScience Applications. *The Geographic Information Science & Technology Body of Knowledge*, (pág. 13). Canadá. Obtenido de https://pages.dataiku.com/nlp-basics?utm_campaign=CONTENT%20NLP%20Basics&utm_medium=paid-search&utm_source=nam-adwords&camp=14310740159&creative=544388576341&utm_campaign=ZZZWP_G_SRCH_NOB_EXACT_AI%20BML_NAM_EN&utm_term=natural%20language%20processing&utm_source=adwords&utm_medium=ppc&hsa_ver=3&hsa_acc=2896441958&hsa_mt=p&hsa_src=g&hsa_cam=14310740159&hsa_grp=128908357631&hsa_tgt=aud-1587671435992:kwd-35561540&hsa_kw=natural%20language%20processing&hsa_ad=544388576341&hsa_net=adwords&gclid=Cj0KCQjwjN-SBhCkARIsACsrBz6oH4Yd0RgZs-RI_FXxTQbMBo7mc1Q9moT5wIfSX0sVCoqgCU6tRCYaAuy-EALw_wcB
- OPS. (5 de mayo de 2021) Hospitalizaciones y muertes por COVID-19 de adultos jóvenes se disparan en las Américas. Obtenido de <https://www.paho.org/es/noticias/5-5-2021-hospitalizaciones-muertes-por-covid-19-adultos-jovenes-se-disparan-americas>
- Ordóñez, B. J., & López, S. Y. (2021). *Repositorio Universidad ICESI*. Obtenido de PROYECTO DE GRADO: https://repository.icesi.edu.co/biblioteca_digital/bitstream/10906/89008/1/T02228.pdf
- Redacción APD. (07 de enero de 2021). Métodos y técnicas de inteligencia artificial: ¿cuáles son y para qué se usan? *APD*. Obtenido de <https://www.apd.es/tecnicas-de-la-inteligencia-artificial-cuales-son-y-para-que-se-utilizan>