



**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
“ENTRENAMIENTO CON MACHINE LEARNING: CASO
DE IDENTIFICACIÓN DE LA TÉCNICA PLN PARA LA
APLICACIÓN DE EVALUACIÓN DE CONTENIDOS EN
LAS CONVERSACIONES TEXTUALES, NUMÉRICOS DE
PERSONAS CONTAGIADAS DE COVID-19 DEL FCI 010-
2021”**

**AUTORA
PIZA GUALE ALEXANDRA ESTEFANIA**

**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

REPOSITORIONACIONAL EN CIENCIA Y TECNOLOGÍA			
FICHA DE REGISTRO DE TRABAJO DE TITULACIÓN			
TÍTULO Y SUBTÍTULO:			
ENTRENAMIENTO CON MACHINE LEARNING: CASO DE IDENTIFICACIÓN DE LA TÉCNICA PLN PARA LA APLICACIÓN DE EVALUACIÓN DE CONTENIDOS EN LAS CONVERSACIONES TEXTUALES, NUMÉRICOS DE PERSONAS CONTAGIADAS DE COVID-19 DEL FCI 010-2021			
AUTOR(ES) (apellidos/nombres):	Piza Guale Alexandra Estefania		
REVISOR(ES)/TUTOR(ES) (apellidos/nombres):	Ing. Zurita Hurtado Harry, Mgs. / 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:	100
ÁREAS TEMÁTICAS:	Tecnología de los Ordenadores		
PALABRAS CLAVES/ KEYWORDS:	Aplicación web, nutrición, alimentación, red neuronal. Web application, nutrition, food, neural network.		
RESUMEN/ABSTRACT (150-250 palabras):			
<p>Debido a que, en marzo del 2020, se declaró una pandemia a nivel mundial por el virus de Covid-19, en la cual fueron desarrollados varios Chatbot para el área de Salud los cuales estan preparados para brindar información a los usuarios, pero asi mismo los asistentes virtuales presentan cierto problema con la información almacenada por falta de entrenamiento con las respectivas respuestas.</p> <p>El siguiente proyecto a presentar incluye modelos predictivos que utilizan herramientas de Aprendizaje Automático para realizar el procesamiento del lenguaje natural para pacientes de los cantones de Guayaquil, Durán y Samborondón que conforman la zona 8 de la provincia del Guayas, que fueron diagnosticados con Covid-19, a través de algoritmos de aprendizaje. Los principales objetivos específicos implican preparar el DataSet mediante obtención, carga, limpieza y depuración de datos previo al entrenamiento del modelo, como también evaluar el entrenamiento del algoritmo de PLN basado en Machine Learning para medir la efectividad</p>			

en las conversaciones textuales etiquetadas en personas contagiadas de Covid-19. En donde el entrenamiento del algoritmo será desarrollado a nivel de Python.

Palabras claves: DataSet, Pre-procesamiento de datos entrenamiento de algoritmo, técnicas PLN.

ABSTRACT

Due to the fact that, in March 2020, a worldwide pandemic was declared by the Covid-19 virus, in which several Chatbot were developed for the Health area which are prepared to provide information to users, but also the virtual assistants have some problem with the stored information due to lack of training with the respective answers.

The following project to be presented includes predictive models that use Machine Learning tools to perform natural language processing for patients from the cantons of Guayaquil, Durán and Samborondón that make up zone 8 of Guayas province, who were diagnosed with Covid-19, through learning algorithms. The main specific objectives involve preparing the DataSet by obtaining, loading, cleaning and debugging data prior to training the model, as well as evaluating the training of the PLN algorithm based on Machine Learning to measure the effectiveness in labeled textual conversations in people infected with Covid-19. The test of the algorithm will be developed with the Python programming language.

Keywords: Dataset, data pre-processing, algorithm testing, PLN techniques.

ADJUNTO PDF:	SI X	NO
CONTACTO CON AUTOR/ES:	Teléfono: 593-967992080	E-mail: pgalexandra6@gmail.com
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, **PIZA GUALE ALEXANDRA ESTEFANIA**, con C.C. No. **0955992268**, certifico que los contenidos desarrollados en este trabajo de titulación, cuyo título es **“ENTRENAMIENTO CON MACHINE LEARNING: CASO DE IDENTIFICACIÓN DE LA TÉCNICA PLN PARA LA APLICACIÓN DE EVALUACIÓN DE CONTENIDOS EN LAS CONVERSACIONES TEXTUALES, NUMÉRICOS DE PERSONAS CONTAGIADAS DE COVID-19 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.

PIZA GUALE ALEXANDRA ESTEFANIA
C.C.NO. 0955992268



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



Habiendo sido nombrado ING. COMP. ACOSTA GUZMAN IVAN LEONEL, MSIG tutor del trabajo de titulación certifico que el presente trabajo de titulación ha sido elaborado por PIZA GUALE ALEXANDRA ESTEFANIA C.C. 0955992268, con mi respectiva supervisión como requerimiento parcial para la obtención del título de Ingeniera en Teleinformática.

Se informa que el trabajo de titulación: **ENTRENAMIENTO CON MACHINE LEARNING: CASO DE IDENTIFICACIÓN DE LA TÉCNICA PLN PARA LA APLICACIÓN DE EVALUACIÓN DE CONTENIDOS EN LAS CONVERSACIONES TEXTUALES, NUMÉRICOS DE PERSONAS CONTAGIADAS DE COVID-19 DEL FCI 010-2021**, ha sido orientado durante todo el periodo de ejecución en el programa Antiplagio URKUND quedando el 5% de coincidencia.

URKUND

Documento	Extracto Tesis ALEXANDRA PIZA Versión 10.docx (D130405763)
Presentado	2022-03-14 20:45 (-05:00)
Presentado por	Ivan Acosta (ivan.acostag@ug.edu.ec)
Recibido	Ivan.acostag.ug@analysis.urkund.com
Mensaje	TESIS ALEXANDRA PIZA Mostrar el mensaje completo

5% de estas 34 páginas, se componen de texto presente en 4 fuentes.

Lista de fuentes Bloques		Abrir sesión	
Categoría	Enlace/nombre de archivo		
	EXTRATO Tesis Capitulo 1 2 3 Mirian Solórzano V10.docx		
100%	ANEXO XIII - RESUMEN DEL TRABAJO DE TITULACIÓN (ESPAÑOL) FAC		✓
100%	ANEXO XIV - RESUMEN DEL TRABAJO DE TITULACIÓN (INGLÉS) FACUL		✓
90%	Preparar el dataset mediante carga, limpieza, depuración y etiquetado de datos previo al entrenamien		✓
76%	Preparar el dataset mediante carga, limpieza, depuración y etiquetado de datos previo al entrenamien		✓
89%	Tabla 11 Delimitación del problema		✓

<https://secure.urkund.com/old/view/124561525-643689-592588#DcY5CoAwFAXAu6R+yF/NchWxkKCSwjQpxbsbmGLE8lxQNgILOIzhCAOyVAoQQU649AVNkVYgtOOMNrd29Xq0esZCi1kZMlzJk4ubP79>



**IVAN LEONEL
ACOSTA GUZMAN**

ING. COMP. ACOSTA GUZMAN IVAN 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 **ENTRENAMIENTO CON MACHINE LEARNING: CASO DE IDENTIFICACIÓN DE LA TÉCNICA PLN PARA LA APLICACIÓN DE EVALUACIÓN DE CONTENIDOS EN LAS CONVERSACIONES TEXTUALES, NUMÉRICOS DE PERSONAS CONTAGIADAS DE COVID-19 DEL FCI 010-2021** de la estudiante PIZA GUALE ALEXANDRA ESTEFANIA, 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.

Atentamente,



**IVAN LEONEL
ACOSTA GUZMAN**

ING. COMP. ACOSTA GUZMAN IVAN LEONEL, MSIG
TUTOR DE TRABAJO DE TITULACIÓN
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 **“ENTRENAMIENTO CON MACHINE LEARNING: CASO DE IDENTIFICACIÓN DE LA TÉCNICA PLN PARA LA APLICACIÓN DE EVALUACIÓN DE CONTENIDOS EN LAS CONVERSACIONES TEXTUALES, NUMÉRICOS DE PERSONAS CONTAGIADAS DE COVID-19 DEL FCI 010-2021”** del estudiante **PIZA GUALE ALEXANDRA ESTEFANIA**. 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 31 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 digitalmente por:
**HARRY ALFREDO
ZURITA HURTADO**

ING. ZURITA HURTADO HARRY ALFREDO, MG

C.C:0910561372

FECHA: 26/03/2022

Dedicatoria

Este trabajo no sería posible sin el apoyo incondicional de nuestro Padre Jesucristo por haberme dado la vida y permitirme logrado alcanzar este período importante de mi formación profesional.

A mi papá Medardo Piza Z., mi mamá Ana Guale Q., por siempre brindarme su apoyo moral, económico y su confianza en todo lo que me he propuesto realizar por eso cada uno de mis logros son dedicados a ellos.

A mis hermanos Elvira, Fabiola y Fernando que han sido las personas más cercanas en mi proceso de formación académica y personal. También a mis sobrinos, que me inspiran seguir adelante a través de su alegría.

Agradecimiento

Agradezco a en primer lugar a nuestro Padre Jesucristo por darme fortaleza, sabiduría y darme una familia maravillosa.

A mis padres que son mis pilares fundamentales, siempre creyeron en mí, me dieron ejemplo de superación, humildad, sacrificio, me enseñaron a valorar todo lo que tengo.

Gracias a mis hermanos y mi abuela que han sido parte de mi formación académica, me han brindado su apoyo moralmente y lo más importante que siempre confiaron en mí.

A cada uno de los docentes que brindaron sus conocimientos, en especial a la Ing. Ximena Trujillo, podría decir que, no fueron fáciles las materias que implanto, pero gracias a su audaz carácter de enseñar adquirí conocimientos, el Ing. Ángel Plaza, por siempre darme la oportunidad de ser parte de las competencias de robótica y mi tutor Ing. Iván Acosta, que sin su ayuda y conocimientos no hubiese sido posible llevar a cabo este proyecto.

A mi amigo Paúl Ayovi, por siempre tenerme paciencia y ayudarme en cada momento, a mi amiga Mirian Solórzano por tenerme paciencia y ayudarme en alguna u otra forma cuando lo necesite y cada uno de mis compañeros y amigos que logre hacer en las aulas de clases, que de una u otro modo me brindaron su ayuda durante mi carrera universitaria.

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”

PIZA GUALE ALEXANDRA ESTEFANIA

C.C. 0955992268

Índice general

N°	Descripción	Pág.
	Introducción	1

Capítulo I

El problema

N°	Descripción	Pág.
1.1	Planteamiento del problema	3
1.2	Formulación del problema	4
1.3	Sistematización del problema	4
1.3.1	Objetivos	4
1.4	Justificación	4
1.5	Delimitación del problema	5
1.6	Alcance	5
1.7	Premisa de la investigación	6
1.8	Operacionalización	6
1.8.1	Variable independiente	6

Capítulo II

Marco teórico

N°	Descripción	Pág.
2.1.	Antecedentes del estudio	7
2.2	Fundamentación teórica	8
2.2.1	Orígenes de la Inteligencia Artificial	8
2.2.2	Inteligencia Artificial	10
2.2.3	Aprendizaje Automático	11
2.2.4	Fundamentación legal	38

Capítulo III

Metodología

N°	Descripción	Pág.
3.1.	Propuesta tecnológica	40
3.1.1	Descripción del proceso metodológico	40
3.3	Metodología de investigación	42
3.3.1	Metodología bibliográfica	42

3.3.2	Metodología Cualitativa	42
3.3.3	Metodología Cuantitativa	43
3.3.4	Metodología Mixta	44
3.4	Técnicas de investigación	45
3.4.1	Encuesta	45
3.4.2	Entrevista	45
3.5	Descripción del procedimiento metodológico	46
3.5.1	Población	46
3.5.2	Análisis de las encuestas	47
3.5.3	Resumen de la entrevista	56
3.6	Construcción de las técnicas de Machine Learning	57
3.6.1	Importación de datos	58
3.6.2	Tratamiento de datos	62
3.6.3	Elección de la técnica	64
3.6.4.	Elección del modelo	66
3.6.5.	Entrenamiento del algoritmo	67
3.6.6.	Evaluación del algoritmo	68
	Conclusiones	69
	Recomendaciones	70
	Anexos	71

Índice de Tablas

Nº	Descripción	Pág.
1.	Determinación del problema	5
2.	Técnicas de procesamiento de lenguaje natural	23
3.	Característica de C++.	28
4.	Característica de Python	29
5.	Característica de Java	29
6.	Característica de R	30
7.	Característica de PHP	30
8.	Característica de Se	32
9.	Característica de Jupyter	35
10.	Característica de Google Colab	35
11.	Característica de IBM Watson	36
12.	Característica de Azure Machine Learning	37
13.	Característica de Amazon Machine Learning	38
14.	Edad	47
15.	Género	48
16.	Lugar que reside de la zona 8 del Ecuador	49
17.	Conocer cómo el CORONAVIRUS (Covid-19) afecta nuestra salud	50
18.	Vacunas contra el coronavirus (Covid-19)	50
19.	Información adecuada y actualizada	51
20.	Hábitos saludables de una persona contagiada.	52
21.	Posee Smartphone	53
22.	Aplicaciones móviles que permiten interactuar y mantener información	54
23.	Aplicación móvil que mantenga actualizada la información	55

Índice de Figuras

Nº	Descripción	Pág.
1.	El comienzo de la era de la Inteligencia Artificial. ¡Error! Marcador no definido.	
2.	AI Mind Map	11
3.	Aprendizaje Automático En Acción	11
4.	Evolution Of Natural Language Processing(NLP)	15
5.	Evolution Of Natural Language Processing (NLP)	15
6.	Natural Language Processing Workflow	18
7.	Procesamiento del Lenguaje Natural (PLN) con Python.	19
8.	Procesamiento del Lenguaje Natural (PLN) con Python.	19
9.	Introducción a SpaCy	25
10.	Tecnología de procesamiento de lenguaje natural	27
11.	Índice TIOBE de enero de 2022.	28
12.	Mostrar sugerencias de bombilla.	33
13.	Introducción a IBM Watson Studio.	36
14.	Tipos de investigación y características	41
15.	Tipos de investigación.	42
16.	¿Qué es la investigación cualitativa?.	43
17.	Investigación cuantitativa	44
18.	Metodología mixta	45
19.	Pregunta 1.1	48
20.	Pregunta 1.2	49
21.	Pregunta1.3	49
22.	Pregunta2.1	50
23.	Pregunta 2.2	51
24.	Pregunta 2.3	52
25.	Pregunta2.4	53
26.	Pregunta 3.1	54
27.	Pregunta 3.2	55
28.	Pregunta 3.3	56
29.	Importación de datos	58
30.	Importación de librería nltk y desinstalación de versión 3.2.5	59
31.	Instalación de la versión 3.7 con todos sus paquetes	59
32.	Verificación de la Base de Datos Síntomas	59

Nº	Descripción	Pág.
33.	Inspección de los datos que contiene de la Base de Datos Síntomas	60
34.	Exploración de datos numéricos de Base de Datos Síntomas	60
35.	Visualización de columnas numéricas y columnas categóricas	60
36.	Inspección rápida de estadística descriptiva	61
37.	Inspección de datos de la columna 10	61
38.	Visualización de cantidad de datos	61
39.	Suma de valores perdidos de la Base de Datos Síntomas	61
40.	Visualización de datos perdidos por medio de mapa de calor	62
41.	Método personalizado y creación de una nueva columna	62
42.	Visualización de la columna 10	63
43.	Tipo de datos	63
44.	Descripción de percentiles	64
45.	Variables Y	64
46.	Exploración de datos en X e Y	64
47.	Elección de combinaciones de técnicas	65
48.	Exploración de datos en la variable x_train	66
49.	Información en variable y_train	66
50.	Elección del tipo de modelo	67
51.	Utilización de métodos y observación del entrenamiento	67
52.	Visualización de gráfico para aprendizaje de datos y de prueba	68
53.	Evaluación métrica	68
54.	Exploración de predicciones y cantidad de salidas	69
55.	Cantidad de variables de salida	69



ANEXO XIII.- RESUMEN DEL TRABAJO DE TITULACIÓN (ESPAÑOL)



FACULTAD DE INGENIERÍA INDUSTRIAL CARRERA INGENIERÍA EN TELEINFORMÁTICA

“ENTRENAMIENTO CON MACHINE LEARNING: CASO DE IDENTIFICACIÓN DE LA TÉCNICA PLN PARA LA APLICACIÓN DE EVALUACIÓN DE CONTENIDOS EN LAS CONVERSACIONES TEXTUALES, NUMÉRICOS DE PERSONAS CONTAGIADAS DE COVID-19 DEL FCI 010-2021”

Autor: Piza Guale Alexandra Estefanía

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

Resumen

En marzo del 2020, se declaró una pandemia a nivel mundial por el virus de Covid-19, en la cual fueron impulsando varios Asistentes Virtuales para el área de Salud orientados a esta temática, los cuales están capacitados para brindar información a los usuarios. Sin embargo, surgieron otras variantes, como Beta, Delta, Omicron, con síntomas diferentes, desencadenando una nueva ola de contagios y muertes. Por ello, este estudio tiene como objetivo crear un prototipo de NLP para analizar las experiencias que tuvieron las personas infectadas por el Covid-19 o una de sus variantes y poder detectar los principales síntomas. Para ello, se usó el lenguaje Python, desarrollado en Google Colab, por lo cual, se puso a prueba varias combinaciones de técnicas de procesamiento de texto usando clasificadores, en conclusión, se descubrió que la combinación de Tokenización, Stop Word y Lematización SpaCy dieron los mejores resultados usando el clasificador LSTM logrando las métricas más efectivas entre las opciones de los datos de salidas de etiquetas múltiples.

Palabras claves: Lenguaje Python, Google Colab, DataSet, Pre-procesamiento de datos, entrenamiento de algoritmo, técnicas PLN, clasificador LSTM.



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



**“TRAINING WITH MACHINE LEARNING TESTING: IDENTIFICATION
CASE OF PLN TECHNIQUE FOR THE APPLICATION OF CONTENT
ASSESSMENT IN TEXTUAL, NUMERICAL CONVERSATIONS OF COVID-19
COUNTED PERSONS FROM FCI 010-2021”**

Author: Piza Guale Alexandra Estefanía

Advisor: Acosta Guzmán Iván Leonel. MSIG, Engineer

Abstract

In March 2020, a worldwide pandemic was declared due to the Covid-19 virus, in which several Virtual Assistants were promoted for the Health area oriented to this topic, which are trained to provide information to users. However, other variants emerged, such as Beta, Delta, Omicron, with different symptoms, triggering a new wave of infections and deaths. Therefore, this study aims to create an NLP prototype to analyze the experiences of people infected by Covid-19 or one of its variants and to detect the main symptoms. For this purpose, the Python language, developed in Google Colab, was used, for which, several combinations of text processing techniques were tested using classifiers, in conclusion, it was found that the combination of Tokenization, Stop Word and SpaCy Lemmatization gave the best results using the LSTM classifier achieving the most effective metrics among the options of the multi-label output data.

Keywords: Python language, Google Colab, DataSet, Data preprocessing, algorithm training, PLN techniques, LSTM classifier.

Introducción

Debido al conocido virus de COVID-19, el mundo vive actualmente una situación sin precedentes. Desde que las autoridades sanitarias chinas notificaron a la comunidad internacional del acontecimiento de esta enfermedad en Wuhan el 31 de diciembre de 2019, las personas han reconocido la existencia de la anomalía.

Desde que la Organización Mundial de la Salud anunció que el actual contagio por Covid-19 estalló el 11 de marzo de 2020. Los países declararon un estado de excepción para prevenir la propagación de este virus. Por lo tanto, todos los países, independientemente de que existan casos confirmados, tomaron nuevas medidas además de las que se hayan propuesto desde el principio, Ecuador es uno de los países afectados por ella, asimismo, las autoridades del Ministerio de Salud informaron que la ciudad de Guayaquil es uno de los principales sitios con mayor contagio.

Según Sunniva Labarthe (2020) indico que, Guayaquil es la ciudad más grande de Ecuador con (2,3 millones de habitantes) y la ciudad más perjudicada por el virus (70% de las muertes). A diferencia con otras partes del país, la intensidad de la epidemia de Guayaquil es un caso real a nivel de la nación.

La medida elegida por las autoridades ecuatorianas fue la imposición de un confinamiento total, que tuvo un gran impacto económico, dejando únicamente la comunicación digital y el acceso a la información. Ante esta situación sanitaria, la sociedad ecuatoriana encuentra en las TIC alternativas para continuar con el proceso de trabajo, educación y capacitación en el hogar, así mismo, la insuficiencia de comunicación dio paso a que se implementen Tecnologías de Inteligencia Artificial para habilitar canales de comunicación que permitan obtener información relacionada con la Pandemia del Covid-19. En este mismo sentido, se buscó utilizar dicha tecnología para que toda la información relacionada con la pandemia se brinde a través de diferentes plataformas. Por otro lado, el objetivo de esta investigación es crear una infraestructura para PLN utilizando algoritmos de aprendizaje automático para detectar los síntomas más dominantes mencionados en las conversaciones textuales de las personas afectadas por Covid-19.

Capítulo I: Contiene el planteamiento del problema, la formación del problema y la sistematización del problema, el objetivo de la investigación (general y específica), la justificación de la investigación, delimitación, hipótesis o premisa y la operacionalización.

Capítulo II: Implica los antecedentes de la investigación, marco teórico, marco contextual, marco conceptual, marco legal, etc.

Capítulo III: Incluye la metodología a utilizar en el desarrollo de trabajos de grado y el desarrollo de la propuesta de investigación.

Capítulo I

El problema

1.1 Planteamiento del problema

En marzo de 2020, el gobierno ecuatoriano anunció un estado de excepción por la emergencia sanitaria referente al virus Covid-19, el cual afectó a la gran mayoría de la población a nivel mundial. Todo esto estaba muy desesperado por la falta de información u orientación. En ese momento, la Organización Mundial de la Salud (OMS) clasificó al Covid-19 como una pandemia global.

Comercio (2021) comunicó lo siguiente:

El Presidente de la república de Ecuador Lenin Moreno, pidió a la ciudadanía que cumpliera las normas que implicaba la emergencia sanitaria, de lo contrario anunció que habría sanciones. También hizo un llamado a que las iglesias para que realizaran las misas o cultos por televisión, Internet o por otros medios.

Por lo tanto, en Ecuador se cerraron centros comerciales, escuelas, universidades, parque de diversiones, etc., provocando un gran impacto económico, por lo que se estableció un confinamiento en casa y sin otros medios más que los digitales para comunicarse e informarse, surgió una nueva tendencia tecnológica que dio paso a varios cambios sociales, económicos, sanitarios y educativos donde las TIC'S se convirtieron en uno de los recursos más importantes para continuar con los procesos desde el hogar como el teletrabajo, educación y entrenamiento. Un ejemplo claro, el uso de las tecnologías de Inteligencia Artificial como medio de comunicación tuvieron un papel importante dentro del contexto de la pandemia Covid-19, con la implementación de Asistentes Virtuales, por medio de acceso de una conversación textual no lograron brindar la suficiente información a población en general como, sintomatología, prevenciones y uso de medicamentos, asimismo, no se compartían las cifras estadísticas de casos confirmados e información sobre los puntos de vacunación.

¿Es posible contar con un lugar o un sitio web que centralice información en el cual el ciudadano común pueda obtener respuestas de alta calidad de como poder desenvolverse durante una pandemia?

¿Es posible identificar la técnica más adecuada de tratamiento de información escrita relacionada con el covid-19 para que a través de proceso NLP y/o ML se puedan obtener información relevante que apoye a la ciudadanía durante el tiempo de la pandemia?

1.2 Formulación del problema

¿Qué tipo de técnica PLN ayudará a la evaluación de contenidos en las conversaciones textuales, numéricos de personas contagiadas de Covid-19 de la zona 8 de la provincia del Guayas?

¿Pueden las técnicas PLN ayudar a la evaluación de contenidos en las conversaciones textuales, numéricos de personas contagiadas de Covid-19 de la zona 8 de la provincia del Guayas?

1.3 Sistematización del problema

1.3.1 Objetivos

1.3.1.1 Objetivo General

Entrenar algoritmos de Machine Learning por medio de la identificación de la técnica PLN para la evaluación de la manera más efectiva de contenidos en las conversaciones textuales etiquetadas de personas contagiadas de Covid-19.

1.3.1.2 Objetivos Específicos

- Recopilar de fuentes bibliográficas la evolución del Covid-19, técnicas PLN y algoritmos disponibles de ML para la creación del modelo de procesamiento de lenguajes.
- Preparar el DataSet mediante obtención, carga, limpieza, depuración y etiquetado de datos previo al entrenamiento del modelo.
- Crear el modelo PLN para la aplicación de las técnicas de procesamiento de lenguaje natural.
- Evaluar el entrenamiento del algoritmo de PLN basado en ML para medir la efectividad de la detección de contenidos detectados en las conversaciones textuales etiquetadas en personas contagiadas de Covid-19.

1.4 Justificación

En el presente trabajo tiene como propósito, tomar conversaciones textuales de personas que hayan superado el Covid-19 entre el año 2020 – 2022, extrayendo información relevante referente a la sintomatología que fueron apareciendo en las diversas variantes de

esta manera que este módulo puede estar disponible para la población y pueda ayudar a detectar futuras sintomatologías de futuras variantes del Covid-19.

1.5 Delimitación del problema

La Tabla 1 a continuación describe el campo, aspectos y áreas de los temas que se llevarán a cabo en el siguiente proyecto.

Tabla 1. Determinación del problema

Campo	Aplicación de tecnología de la información
Área	Tecnología de los ordenadores
Aspecto	Evaluación de contenidos en las conversaciones textuales, numéricos de personas contagiadas de Covid-19 por medio de entrenamientos de Machine Learning.
Tema	Entrenamiento con Machine Learning: Caso de identificación de la técnica PLN para la aplicación de evaluación de contenidos en las conversaciones textuales, numéricos de personas contagiadas de Covid-19 del FCI 010-2021

Elaborado por: Piza Guale Alexandra

1.6 Alcance

En el presente proyecto tiene como alcance desarrollar un modelo de entrenamiento en Machine Learning enfocado el uso de técnicas de Procesamiento del Lenguaje Natural (PLN), lo que permitirá fortalecer la arquitectura existente de las conversaciones textuales, numéricas de las personas que han sido contagiadas por Covid-19.

- Levantar información utilizando recursos bibliográficos para identificar las técnicas PLN que permita fortalecer la arquitectura de las conversaciones textuales, numéricas de las personas contagiadas por Covid-19.
- Realizar el análisis y diseño del modelo de entrenamiento con Machine Learning.
- Realizar pruebas de medición de calidad y resultados del prototipo de entrenamiento.
- Documentar el resultado de las pruebas del modelo y comparar el fortalecimiento de la arquitectura respecto a las anteriores.

En el presente alcance no se considerará las fases puestas en producción en un servidor debido que esa infraestructura no está dispuesta en la Facultad Ingeniería Industrial.

1.7 Premisa de la investigación

Entrenamiento con Machine Learning: Caso de identificación de la técnica PLN para la aplicación de evaluación de contenidos en las conversaciones textuales, numéricos de personas contagiadas de Covid-19 del FCI 010-2021

1.8 Operacionalización

Cuando los usuarios hacen preguntas basadas en Covid-19, los Asistentes Virtuales como los Chatbots no tienen suficiente información para dar suficientes respuestas, por lo que es importante saber que los programas de lenguaje natural no transfieren inteligencia a los Chatbots, solo les da procesamiento y generación. capacidad del lenguaje humano. Si desea llevar inteligencia a los Asistentes Virtuales, debe recurrir a sistemas como reglas o redes neuronales, que les permitirán obtener una mayor formación y poder brindar a los usuarios respuestas más específicas y seguras.

1.8.1 Variable independiente

Entrenamiento con Machine Learning

1.9.2 Variable dependiente

- Evaluación de técnicas PLN.

Capítulo II

Marco teórico

2.1. Antecedentes del estudio

Bonilla (2020) A causa de la aparición del virus Covid-19, Ecuador fue uno de los tantos países afectados por el Covid-19, en febrero 29 del 2020 se identificó el primer caso y la expansión comenzó a incrementarse de una manera muy rápida, donde hubo mucho personal del área de salud los cuales dieron positivo al contagio del Covid-19, lo que intensificó aún más la atención a los pacientes, por lo cual, se necesitó herramientas que puedan reducir el diagnóstico, los seguimientos y tratamientos de Covid-19. Aquella demanda fue realizada de manera esencial en ambientes donde los recursos son relativamente insuficientes.

De esta manera Coronel y Perez (2020) indicaron que, para garantizar la prestación de servicios públicos básicos, salud, seguridad, protección contra incendios, riesgo, centros de transporte, bancos, alimentos, departamentos estratégicos y otros servicios necesarios para combatir la pandemia el gobierno autorizó que las empresas de esos sectores estarían autorizadas a mantener la movilidad y sostener jornadas laborales presenciales.

Una de las tecnologías más factibles y usadas para este tipo de emergencia para poder obtener un monitoreo eficiente acerca del virus de Covid-19, es la Inteligencia Artificial (IA) que ayuda a analizar los datos disponibles, también ayuda a promover la investigación sobre el virus. La Inteligencia Artificial puede ayudar a formular planes de tratamiento adecuados, estrategias preventivas y el desarrollo de medicamentos y vacunas, según Haleem (2020).

Como indica Nicolás A. Núñez (2021), en su investigación “Uso de minerías de textos para comparar los contenidos relacionados a calidad y acreditación generados en redes sociales por universidades de Perú y Chile”, La minería de datos cubre una gama de técnicas de modelado (predicción de resultados, análisis de texto, visión artificial, reconocimiento de voz, etc.) para diferentes necesidades de las organizaciones, y propone un método de minería de texto para extraer información de artículos de investigación.

Además, describe el uso de algoritmos de clasificación para predecir usuarios en entornos educativos virtuales. Se utilizará una técnica de inteligencia artificial llamada minería de texto. Para recolectar los datos utilizados en el estudio, se consideraron dos

fuentes: Facebook y Twitter. La extracción de los datos se realiza de dos formas: en el caso de Twitter, utilizando la biblioteca `tweepy` en Python 3.7, que permite la extracción automática, lo que significa reducir esta tarea número de veces. Por el contrario, para Facebook se realiza manualmente porque no existen herramientas específicas que puedan automatizar esta tarea.

Por otro lado, después de dar uso de NLP, se aplica otra técnica de aprendizaje automático llamada clasificación binaria, que tiene como objetivo dividir el conjunto de datos en dos clases, donde cada clase debe tener la menor similitud entre sí. sus características para la clasificación se utilizará el algoritmo Random Forest, el cual, incluye una técnica para segmentar entidades de un conjunto de datos según criterios estadísticos, según el método utilizado en el estudio.

Como lo indico Torres (2021), para llevar a cabo el trabajo de analisis de opinion, referido a la metodologia KDD, se proponen las siguientes etapas: selección de datos, preprocesamiento de datos, transformacion de datos, mineria de datos e interpretacion de datos. Los lenguajes de programación elegidos fueron R y Python, para completar las etapas de minería e interpretación de datos. En la etapa de selección de datos se obtiene un DataSet acorde al objetivo de la investigación. En primer lugar, se obtuvieron 3.479 tuits utilizando la API Rest de Twitter gratuita a través del lenguaje R.

Para las etapas de procesamiento y transformación de datos se utilizó la herramienta OpenRefine, donde se aplicaron criterios de limpieza, reducción de clustering de datos. En el estándar de limpieza, se eliminaron símbolos, números y patrones (URL). Finalmente, en la etapa de interpretación de datos, utilice el IDE de Jupiter para visualizar los clústeres generados en la etapa anterior. Los gráficos resultantes incluyen frecuencia de palabras, nubes de palabras, regresión discreta y lineal, diagramas de caja y reglas de asociación.

2.2 Fundamentación teórica

2.2.1 Orígenes de la Inteligencia Artificial

Silva, Associate, & LATAM (2020) presentaron la siguiente información: El origen de la Inteligencia Artificial está íntimamente relacionado con el origen de las computadoras. Los grandes nombres en el campo de las tecnologías de la información han facilitado la aparición de la Inteligencia Artificial.

En los años 30 y 40 del siglo XX se publicaron los primeros libros que hablaban de una forma u otra sobre la Inteligencia Artificial. Es importante señalar que el término Inteligencia Artificial aún no existe oficialmente. El artículo *Arithmetic Numbers*, publicado en 1936 por Alan Turing, tuvo una fuerte influencia. Se considera que este texto establece las bases teóricas de la informática.

En este artículo, Alan presentó el concepto de máquinas de Turing. Además de formalizar la definición del algoritmo, las ideas presentadas en la publicación son una introducción a las computadoras digitales.

Una de las conclusiones importantes a las que llegó Turing con la ayuda de sus máquinas fue que había problemas que ninguna computadora podía resolver. Y para demostrarlo, se le considera el padre de la Teoría Aritmética. El extraordinario logro de Alan Turing en los orígenes de la Inteligencia Artificial fue la construcción de la primera computadora electromecánica en 1940.

En 1941, el ingeniero alemán Konrad Zuse, otro pionero de la Inteligencia Artificial, inventó el Z3, el primer ordenador electrónico digital completamente funcional. Zuse también es el creador del primer lenguaje de programación de alto nivel.

La primera Teoría Matemática del cerebro se creó combinando a un joven apasionado por la lógica y un neurocientífico brillante que creó el primer modelo formal de procesamiento de información en todo el cerebro. Considerado oficialmente el primer trabajo concreto en el campo de la Inteligencia Artificial, en 1943 se presentó el modelo de neurona artificial a los autores Warren McCulloch y Walter Bates.

Durante las últimas décadas, el Aprendizaje Automático ha evolucionado de forma drástica y se ha convertido en un tema de investigación muy importante en Inteligencia Artificial.

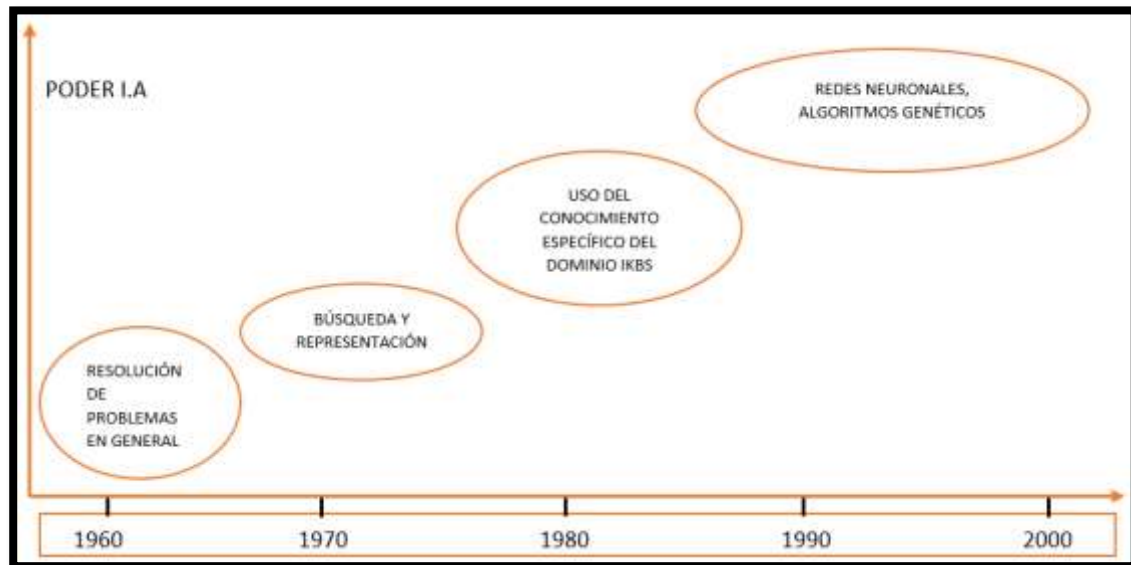


Figure 1. Información tomada de: *El comienzo de la era de la Inteligencia Artificial*. Elaborado por Piza Guale Alexandra

2.2.2 Inteligencia Artificial

Según Tomás & Varela (2020) la Inteligencia Artificial es una rama de la informática que incluye conceptos transversales relacionados con la lógica y el aprendizaje. Por tanto, se trata de diseñar herramientas informáticas que simulen los procesos de la inteligencia humana, incluyendo el aprendizaje, el razonamiento y la autocorrección.

Fernández (2019) indicó lo siguiente: la Inteligencia Artificial no es un fenómeno nuevo. De hecho, el primer trabajo sobre este tema se remonta a la década de 1950. Sin embargo, actualmente la popularidad de esta línea tecnológica ha estado en auge debido a tres factores:

La disponibilidad de grandes cantidades de datos. Digital; el aumento de la capacidad de almacenamiento y computación de costo extremadamente menor y posteriormente, el avance del desarrollo de algoritmos.

En la rama de la Inteligencia Artificial según Leyva, Vázquez & Amarandache (2018) dijo: La Inteligencia Artificial no es un campo único, sino que se divide en varias ramas, como el Aprendizaje Automático, los programas de lenguaje natural, los sistemas expertos, la visión por computadora, el reconocimiento automático de voz, la planificación y la robótica.



Figure 2. Información tomada de: AI Mind Map – Machine Learning And Artificial Intelligence Study Group – Medium. Elaborado por Piza Guale Alexandra

2.2.3 Aprendizaje Automático

Norman (2020) expresó que el Aprendizaje Automático (Machine Learning ML) es la ciencia de enseñar a las computadoras a hacer predicciones basadas en datos. El Aprendizaje Automático implica darle a una computadora un conjunto de datos y pedirle que haga predicciones. Al principio, la computadora tendrá muchas predicciones incorrectas. Sin embargo, en el proceso de miles de predicciones, la computadora actualizará los algoritmos para hacer mejores predicciones.

También se muestra que el Aprendizaje Automático está conformado por 3 tipos de aprendizajes que son los siguientes:



Figure 3. Información tomada de: APRENDIZAJE AUTOMÁTICO EN ACCIÓN. Elaborado por Piza Guale Alexandra

2.2.4 Tipos de Aprendizaje

2.4.1. Aprendizaje Supervisado

El Aprendizaje Supervisado mediante algoritmos los datos que se han clasificado u ordenado previamente para indicar cómo se clasifica la información actual. De esta manera, se requiere la intervención humana para proporcionar una retroalimentación ROUHIAINEN (2018).

2.4.2. Aprendizaje no supervisado

Según Agencia B12 (2021) anuncia que el aprendizaje no supervisado es, como era de esperar, lo opuesto al Aprendizaje Supervisado. En lugar de etiquetas, los algoritmos proporcionan grandes cantidades de datos y tienen herramientas para comprender las propiedades de los datos.

A partir de ahí, puede aprender a recopilar datos de una manera que otra inteligencia (humana o artificial) pueda piratear y comprender. Es la capacidad de transformar millones de datos sin procesar que contienen en información potencialmente útil es esencial para cualquier empresa u organización.

2.4.3. Aprendizaje reforzado

Como la Agencia B12 (2021) enuncia que el Aprendizaje por Refuerzo, es muy diferente de los dos estilos de aprendizaje anteriores. Se podría decir que el Aprendizaje por Refuerzo es lo que permite que un algoritmo aprenda de los errores. Al principio, los errores serán numerosos, pero si se presenta una serie de señales positivas y negativas relacionadas con el éxito y el fracaso respectivamente, con el tiempo el algoritmo aprenderá por sí solo, hasta que se vuelva más eficiente.

2.4.4. Aprendizaje profundo

Como Díaz, (2020) expresa que: El aprendizaje profundo es un subcampo del Aprendizaje Automático que intenta clasificar datos utilizando algoritmos relacionados. Se basa en determinadas arquitecturas de Redes Neuronales, lo que le permite priorizar la información (visual, auditiva y escrita) mediante la segmentación de patrones clasificados por nivel. Bajo

este estándar, el aprendizaje se lleva a cabo por etapas, lo que equivale a lo que les sucede a los humanos. En otras palabras, comienza con datos básicos y aprenderá a medida que se expandan los niveles más complejos.

2.2.5 Procesamiento del Lenguaje Natural (NLP)

León Esmeralda (2020) detallo lo siguiente el Procesamiento del Lenguaje Natural (NLP) es una rama de la Inteligencia Artificial que se ocupa de las interacciones entre las computadoras y los humanos utilizando el lenguaje natural. El objetivo principal es comprender el lenguaje humano a través de una computadora de una manera que sea fácil de entender. La mayoría de las técnicas de NLP se basan en el Aprendizaje Automático para poder comprender y entender los diferentes lenguajes que existen. A continuación, se muestra un ejemplo de interacción hombre-máquina mediante el procesamiento del lenguaje natural:

- La persona está hablando con la máquina.
- El dispositivo graba audio.
- El audio se convierte en texto.
- Procesamiento de datos de texto.
- El proceso de convertir datos en sonido.
- El dispositivo responde a la persona reproduciendo un archivo de audio.

Kaur (2021) expreso lo siguiente, en pocas palabras, un lenguaje puede entenderse como un conjunto de reglas o símbolos. Estos símbolos se integran y luego se utilizan para transmitir y difundir información. Aquí se aplican las reglas para suprimir símbolos. El campo del procesamiento del lenguaje natural se divide en subcampos, a saber, generación y comprensión del lenguaje natural, que, como sugiere el nombre, están relacionados con la generación y comprensión de textos.

A continuación, se muestran manera muy breve:

Como lo indico ITELLIGENT (2017) la arquitectura de un sistema de PLN se sustenta en una definición de Lenguaje Natural compuesta por 5 niveles descritas a continuación:

Nivel fonológico: Se basa en cómo las palabras tienen un nexo con los sonidos que representan.

Nivel morfológico: Cómo las palabras se componen por medio de unidades de significado más pequeñas que reciben el nombre de “morfemas”.

Nivel sintáctico: Explica cómo las palabras pueden conectarse para formular oraciones, fijando el papel estructural que cada palabra significa en la oración y qué sintagmas son parte de otros.

Nivel semántico: Significado de las palabras y de cómo se entrelazan para dar significado a una oración, también se refiere al significado independiente del contexto, es decir, de la oración aislada.

Nivel pragmático: Explica cómo las oraciones se utilizan en diferentes contextos y de cómo el uso perjudica al significado de las oraciones.

Todas las capas descritas aquí son inseparables y se refuerzan mutuamente. El objetivo de un sistema de PNL es introducir estas definiciones en una computadora y luego usarlas para crear oraciones estructuradas con un significado específico.

2.2.6 Programación neurolingüística

Delgado Paulette (2021) manifiesto lo consiguiente la Programación Neurolingüística es un método para cambiar los pensamientos y hábitos de una persona para tener éxito a través de habilidades cognitivas, conductuales y de comunicación. Este es un enfoque pseudocientífico que se basa en conexiones neuronales, específicamente en cómo se procesa el lenguaje. Se ha vuelto popular con enfoques alternativos para el desarrollo personal o la autoayuda.

Según la página de NPL Empowering Partners, NLP es aprender el lenguaje de tu cerebro o un manual de usuario. Se fundamenta de tres maneras:

- Neuro, que es la Técnica Neurológica.
- Lingüística: que es el mensaje, de forma verbal como no verbal que se exporta al cerebro.
- Programación: que es la forma en que el instinto procesa estos mensajes.

2.2.7 Evolución del NLP

Es importante destacar que Deva (2021) indico lo siguiente, el proceso de lenguaje natural (PNL) existe desde hace mucho tiempo. De hecho, en la década de 1950 se introdujo un modelo de bolsa de palabras muy simple.

Desde 2013, debido al desarrollo y progreso de los algoritmos de aprendizaje automático y la reducción de los costos de computación y memoria, se ha logrado un gran progreso en este campo.



Figure 4. Información tomada de: Evolution Of Natural Language Processing(NLP). Elaborado por Piza Guale Alexandra

Word2Vec: convierte texto en vectores, también conocido como incrustación. Cada uno de estos vectores consta de 300 valores, por lo que simplemente se denominan vectores de 300 dimensiones. Porque representa un espacio vectorial de 300 dimensiones. Luego, puede usar estas representaciones vectoriales como entrada para el aprendizaje automático.

La arquitectura se basa en una red neuronal de superficie de dos capas:

- Palabras inversas continuas (CBOW)
- Salta gramos continuos.

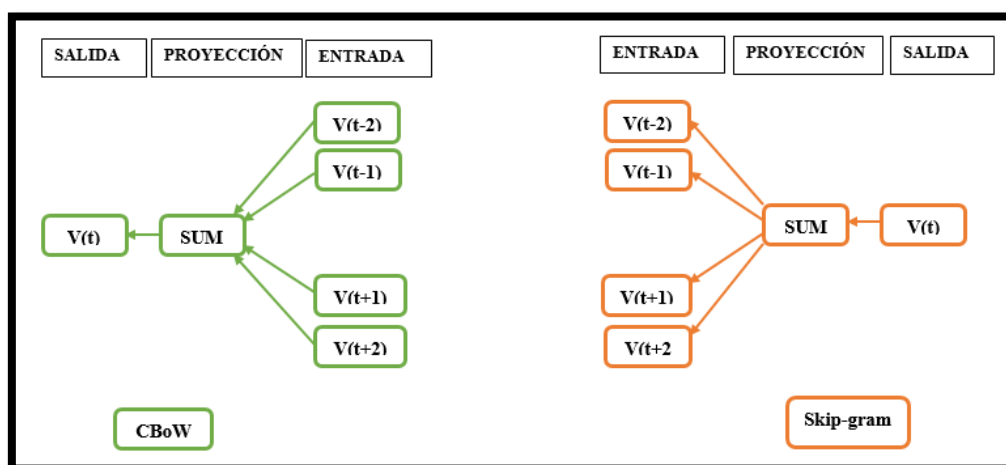


Figure 5. Información tomada de: Evolution Of Natural Language Processing (NLP). Elaborado por Piza Guale Alexandra

GloVe: Utiliza un modelo de regresión para aprender expresiones de palabras a través del aprendizaje no supervisado. La principal intuición que subyace al modelo es una simple observación de que la razón de probabilidad de coexistencia entre palabras puede codificar alguna forma de significado.

FastText: En 2016, el Laboratorio de Inteligencia Artificial (FAIR) de Facebook publicó una investigación sobre FastText. FastText se basa en Word2Vec, pero trata cada

palabra como una serie de palabras secundarias llamadas n-gramas de letras. Y ayuda a perder vocabulario como Word2Vec y Glove. Por ejemplo, la palabra "agricultura" se desglosó en n-gramas de la siguiente manera:

"Granja" => "g", "gr", "gran", "granj", "granja", "granjero"

Incluso si la palabra "agricultura" no está incluida en el vocabulario, puede ser la palabra "granja".

Las incrustaciones que FastText aprende para palabras son la suma de las incrustaciones para cada n-gram, FastText usa los mismos modelos CBOW y Skipgram, y FastText aumenta el impacto del vocabulario Word2Vec a más de 3 millones de palabras.

Transformer: Se implantó un nuevo tipo de arquitectura de red neuronal basada en el mecanismo de auto-atención (las sumas ponderadas ocultas se transfieren como vectores de contexto a pasos de tiempo futuros, generalmente en secuencia al decodificador RNN).

El mecanismo de atención automática de esta nueva arquitectura de Transformer se enfoca en capturar las relaciones entre todas las palabras en la secuencia de entrada, mejorando en gran medida la precisión de las tareas de comprensión del lenguaje natural, como la traducción automática. La arquitectura Transformer fue un hito muy importante para la PNL, pero otros equipos de investigación continuaron desarrollándola y la mantuvieron como la raíz de las arquitecturas alternativas.

BlazingText: Utiliza varias CPU o GPU para el escalamiento y aceleramiento del entrenamiento de Word2Vec. De manera semejante, la ejecución de BlazingText del algoritmo de ajuste de texto amplifica FastText para emplear la rapidez de GPU en un kernel CUDA personalizado. CUDA o arquitectura de dispositivo unificado de cómputo es una plataforma informática paralela y un modelo de programación establecido por Nvidia. BlazingText utiliza una colección continua de palabras para crear gramática e incrustación de caracteres, omitiendo la arquitectura de entrenamiento. También puede usar BlazingText para dejar de preparar el modelo de entrenamiento basado en el evento.

ELMO: Aquellos vectores de palabras son ocupaciones aprendidas del estado Interno de un Modelo de Lenguaje Bidireccional Profundo (BiLM) previamente entrenado en un gran corpus de texto. En ELMO, los vectores de palabras se entrenan utilizando un modelo de lenguaje bidireccional profundo. También combina modelos de lenguaje directo e inverso para comprender mejor la sintaxis y la semántica en distintos argumentos lingüísticos.

GPT: Se basa en la arquitectura de Transformer, pero realiza dos pasos de entrenamiento.

1. GPT aprende un modelo de lenguaje a partir de un gran corpus de texto sin etiquetas.
2. GPT usa datos etiquetados para realizar pasos de aprendizaje supervisado para aprender tareas específicas de NLP, como la clasificación de texto.

GPT es unidireccional.

BERT: El modelo de lenguaje de última generación de la PNL. Se deriva de representaciones contextuales posteriormente al entrenamiento, adjuntando el aprendizaje secuencial semisupervisado, BERT considera el contexto de cada ocurrencia de una palabra en particular. por ejemplo:

Considere dos oraciones "Él tiene un negocio" y "Él corre un maratón"

BERT es realmente bidireccional. En el paso de entrenamiento no supervisado, BERT aprende a mostrar texto sin etiquetar de izquierda a derecha y de derecha a izquierda. El BERT inglés original tiene dos modelos que son:

1. BERT_BASE: con 12 codificadores con 12 cabezales de autocuidado bidireccionales.
 2. BERT_LARGE: 24 codificadores con 16 cabezales de autocuidado bidireccionales.
- Los dos modelos están entrenados previamente a continuación de los datos sin etiquetar de BooksCorpus (800 millones de palabras) y Wikipedia en inglés (2.500 millones de palabras).

2.2.8 Flujo de trabajo de procesamiento de lenguaje natural (NLP Workflow)

Según Smirnov, et al. (2021) revelo que la implementación del flujo de trabajo está orientada a los servicios en la nube. Este modelo implementado está entrenado y alojado en la nube proporcionada por Google (Google AutoML) o Amazon (Amazon Comprehend) son totalmente compatibles

Flujo de trabajo de procesamiento de lenguaje natural, con la excepción del proceso de anotación. Los modelos en esta opción solo se pueden usar a través de la API REST, sin la posibilidad de descargarlos y utilizarlos de forma autónoma. El proceso de anotación El proceso de anotación debe ser organizado por una herramienta de terceros como BRAT o la herramienta Doccano.

Se utiliza una interfaz basada en la web para ingresar texto y mostrar la clasificación del texto y los resultados de las anotaciones. Además, guía a los usuarios a través de las herramientas subyacentes proporcionadas por su interfaz web para la anotación de texto manual, la capacitación y la evaluación del modelo.

El núcleo de la PNL proporciona modelos y lógica empresarial, cuyos elementos deben ser reemplazables, así que prueba varias combinaciones de herramientas para encontrar la mejor. La solución completa se empaqueta como un contenedor de Docker y se ejecuta en cualquier host físico o virtual que admita Docker/Kubernetes.

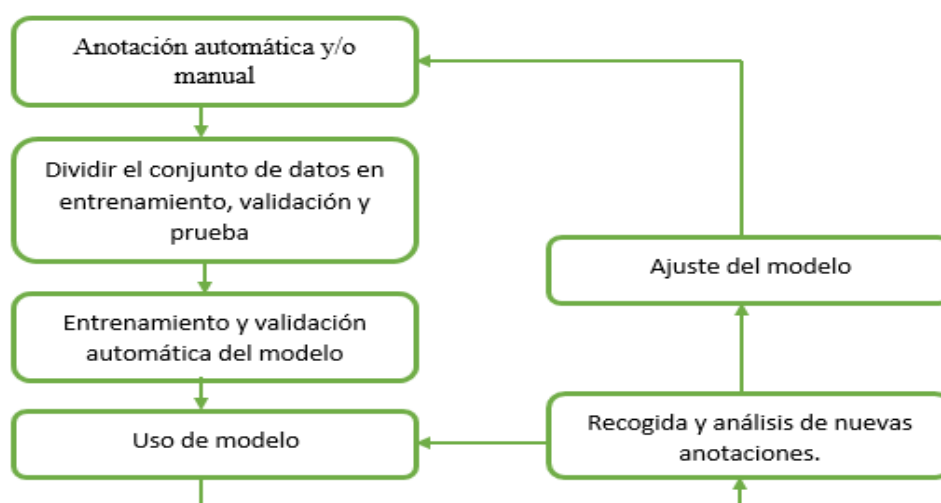


Figure 6. Información tomada de: *Natural Language Processing Workflow for Customer Request Analysis in a Company*. Elaborado por Piza Gualé Alexandra

2.2.9 Técnicas comunes usadas en NLP

Las técnicas más comunes que ha sido utilizadas en el ámbito de NLP son las siguientes:

2.2.9.1. Tagging Parts of Speech (PoS)

Se puede definir como el proceso de asignar una de las partes del discurso a una palabra dada. Según Pham (2020) comúnmente conocido como etiquetado POS. En términos simples, se entiende que el etiquetado de parte del discurso es la tarea de etiquetar cada palabra en una oración con la parte del discurso correcta. Las partes del discurso incluyen sustantivos, verbos, adverbios, adjetivos, pronombres, conjunciones y su subcategoría.

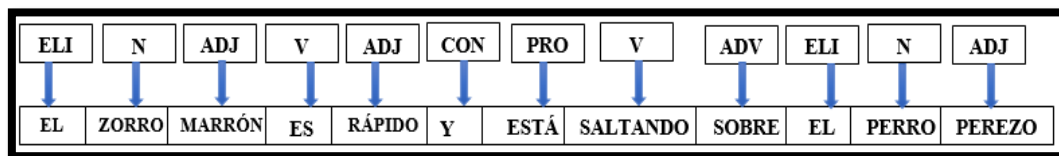


Figure 7. Información tomada de: Procesamiento del lenguaje natural (PLN) con Python. Elaborado por Piza Guale Alexandra

Hay dos tipos principales de etiquetas POS como los grupos únicos: basados en reglas y estocástico. Muchas aplicaciones de procesamiento de lenguaje natural (NLP) utilizan las técnicas estocásticas para establecer la posición del discurso. El interés de las técnicas estocásticas acerca de las técnicas tradicionales establecidas en reglas proviene de la facilidad de obtener estadísticas automatizadas.

2.2.9.2. Shallow parsing/ Chunks

Se utiliza para comprender la gramática en una oración. Los tokens se analizan y se construye un árbol de estructura a partir de su PoS.

Significado: Semántico.

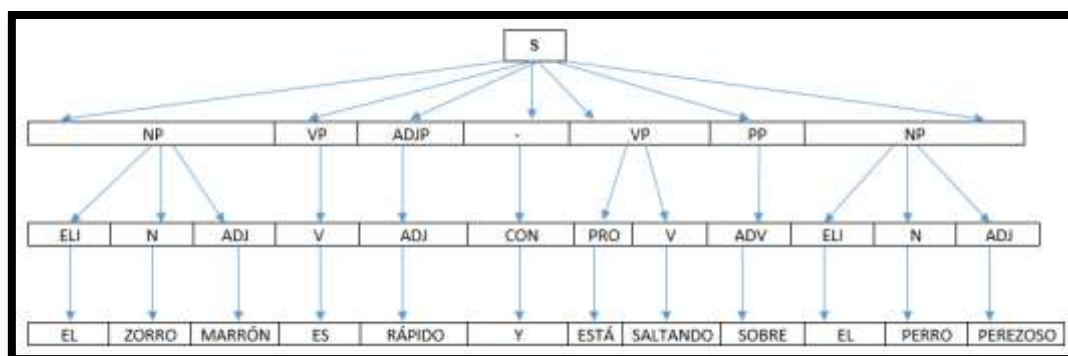


Figure 8. Información tomada de: Procesamiento del lenguaje natural (PLN) con Python. Elaborado por Piza Guale Alexandra

2.2.9.3. Tokenización

La tokenización es una tarea común en el procesamiento del lenguaje natural (NLP) así lo indicó LIMA (2021). Es un paso fundamental en métodos tradicionales de NLP como Count Vectorizer y en arquitecturas basadas en Deep Learning avanzado como Transformers.

Los tokens son los componentes básicos del lenguaje natural.

La tokenización es un método para dividir un fragmento de texto en unidades más pequeñas llamadas tokens. Aquí, los tokens pueden ser palabras, caracteres o subpalabras. Por lo tanto, la tokenización se puede dividir aproximadamente en 3 tipos: tokenización de palabras, caracteres y subpalabras (caracteres n-gram).

Por ejemplo, considere la siguiente oración: "**Nunca te rindas**".

El método más común para formar mosaicos se basa en el espacio. Suponiendo que el espacio en blanco sea el delimitador, la tokenización de una oración produce 3 tokenizaciones: **nunca te rindas**.

Dado que cada token es una palabra, se convierte en un ejemplo de tokenización de Word.

2.2.9.4. **Pragmatic Analysis**

El análisis pragmático como lo indico Johnson (2022) permite analizar el significado básico de un texto dado. El objetivo es sacar inferencias del texto dado. El análisis de sentimiento es una de las áreas de investigación del análisis pragmático, cuyo objetivo es revelar el sentimiento en un texto dado. El análisis de sentimientos es el campo de estudio que analiza las opiniones, sentimientos, evaluaciones y actitudes de los usuarios hacia entidades tales como productos, servicios, organizaciones, individuos, temas, eventos y sus características.

A. Paradigma significa abstraer o derivar el uso significativo del lenguaje en una situación. En este análisis, el foco principal siempre está en lo que se dice al reinterpretar su significado.

El análisis pragmático ayuda a los usuarios a descubrir este efecto deseado mediante la aplicación de un conjunto de reglas que caracterizan el diálogo cooperativo.

Por ejemplo, "¿Cerrar la ventana?"

debe interpretarse como una solicitud, no como un comando.

2.2.9.5. **Bag of words**

Es una de las ideas más simples y utilizadas en el procesamiento del lenguaje natural según Brownlee (2017) dado un set de documentos o corpus, determine con qué frecuencia

aparece cada palabra del conjunto en cada documento. Esta información puede ser más o menos refinada dependiendo de los filtros previos aplicados al texto.

Es una representación textual que describe la aparición de una palabra en un documento.

Esto significa dos cosas:

- Un vocabulario de palabras conocidas.
- Medir la presencia de palabras conocidas.

Modelo Bag of words pasos:

- 1 Recopilar datos
- 2 Vocabulario de diseño
- 3 Crear un vector de documento
 - Gestión de vocabulario
 - Puntuación
 - palabra hash
 - TF-FDI

Además, en un nivel más detallado, los modelos de aprendizaje automático utilizan datos numéricos en lugar de datos textuales. Entonces, para que sea más concreto, usando la técnica (BoW), convertimos el texto a su vector numérico equivalente.

2.2.9.6. **Word2vec**

Según data (2018) la creación de un equipo de investigadores de Google dirigido por Tomas Mikolov, es uno de los modelos más populares para crear incrustaciones de palabras. Word2vec tiene dos métodos principales de contextualización de palabras: el modelo Continuous Bag-of-Words (CBOW) y el modelo Skip-Gram.

2.2.9.6.1. **CBOW**

El menos popular de los dos modelos, utiliza palabras de origen para predecir palabras de destino.

Ejemplo, toma una oración, Quiero aprender Python. En este caso, la palabra destino es Python y la palabra origen es "Quiero aprender".

2.2.9.6.2. **Skip-Gram**

Funciona de manera opuesta al modelo CBOW, utilizando la palabra objetivo para predecir el origen o el contexto de las palabras circundantes. Como lo indica León (2020) considere la frase el veloz zorro marrón saltó sobre el perro perezoso.

Skip-Gram descompone oraciones en (contexto, objetivo), lo que predice el contexto o la salida mediante el uso de palabras objetivo (entradas) de la siguiente manera: (rápido, el), (rápido, marrón), (marrón, rápido), (marrón, zorro)

Esta es una técnica para aprender leyendo grandes cantidades de texto y recordando palabras que parecen ser similares en diferentes contextos. Después de entrenar suficientes datos, se genera un vector de 300 dimensiones para cada palabra para formar un nuevo vocabulario en el que las palabras "similares" están cerca unas de otras. Utilizando vectores previamente entrenados, logramos obtener una gran cantidad de información para comprender la semántica del texto.

2.2.9.7. **Stop Word**

Las palabras reservadas son tokens así lo expresa Chen, (2019) que aparecen con frecuencia en un corpus que introducen más ruido que señal. Por otro lado, las palabras vacías tienen la misma probabilidad de aparecer en documentos relevantes para una determinada tarea de NLP, que en documentos que no son relevantes para esa tarea.

Ejemplos de palabras vacías son: "un", "una" y "la".

Las palabras negativas menos intuitivas y de uso común son las siguientes: "no" y "not" también se consideran palabras vacías. No se requiere que estas palabras ocupen espacio en nuestra base de datos o consuman tiempo de procesamiento. Por lo tanto, se podrían eliminarlas fácilmente almacenando una lista de palabras que se puedan considerar palabras vacías.

El análisis del agente de diálogo según Celi-Parraga, Varela-Tapia, & Montaña-Pulzara (2021) se puede hacer simplemente mediante un proceso de comparación con una lista de frases explicativas. Por lo tanto, esta no es una solución completa y debe interactuar con el uso de herramientas informáticas que ayuden a mejorar la búsqueda de información, planteando grandes desafíos a la interpretación y procesamiento del lenguaje natural y la optimización del tiempo de procesamiento.

Tabla 2. Técnicas de procesamiento de lenguaje natural

Técnicas	Función	Objetivo
Preprocesamiento de datos	Se basa en el contenido de los ficheros de entradas clasificado utilizando bibliotecas desarrolladas para este propósito.	Toma la decisión de cuáles serán los patrones a utilizarse para la extracción de concepto.
Análisis Superficiales	Está dirigido inicialmente a la identificación de concepto.	Representan las estructuras gramaticales.
Análisis de Dependencia	Establecen las relaciones de dependencia concurrentes entre las estructuras gramaticales.	Diferencia estructuras gramaticales e identifica o construye.
Recuperación de información	Análisis de información textual ingresada por los usuarios basados en sintaxis y semánticas.	Obtención de información de buena calidad para tratar mejor en conjunto a los usuarios.
Búsqueda semántica de información	La búsqueda semántica predice lo que los usuarios expresan públicamente y lo ajusta según sea necesario.	Mejora de investigación y se ajusta a la necesidad de usuario.
Técnicas basadas en el análisis de deletreo y distancia	Se encarga de identificar la diferencia que existe entre las cadenas de caracteres.	Mediciones basadas en la diferencia de caracteres en la cadena.

Información tomada de Técnicas de procesamiento de lenguaje natural en la inteligencia artificial conversacional textual. Elaborado por Piza Gualé Alexandra

2.2.10 Herramientas usadas en Python para NLP

Según Aprende (2018) explica que, en artículos futuros, veremos ejemplos de NLP más detallados usando Python, pero aquí hay una breve revisión de las herramientas utilizadas en Python:

2.2.10.1. NLTK

Natural Language Toolkit, es la plataforma líder que permite crear programas de Python para procesar datos de lenguaje humano. Es una biblioteca que todo el mundo empieza a usar. Es muy útil para pre-procesar, crear etiquetas, lematiza, etc. Gestiona una interfaz más factible de utilizar para más de 50 corpus y procesos léxicos (como Word Net), así como un conjunto de bibliotecas de procesamiento de textos para la clasificación.

2.2.10.2. Text Blob

Es una librería de Python simple para realizar tareas de NLP como tokenización, extracción de frases nominales, análisis de sentimientos, etiquetado de POS y derivación de palabras, N-gramas. Es como NLTK (Natural Language Toolkit), asimismo, cuenta con más funciones como: revisión ortográfica, análisis de opiniones, creación de resúmenes breves de texto, traducción y detección de idioma.

Text Blob es similar a una cadena de Python que se puede manejar de una manera más sencilla. Es una herramienta que contiene todos los materiales importantes para el procesamiento del lenguaje natural (NLP), los Text Blobs pueden interactuar con NLTK. Por lo tanto, permite cambiar sencillamente a una implementación pre-entrenada de la biblioteca NLTK para el análisis de opiniones.

2.2.10.3. Gensim

Es una biblioteca de Python gratuita y de código abierto escrita por Radim Rehurek, para el modelado de temas sin supervisión y el procesamiento del lenguaje natural para representar documentos como vectores semánticos del modo más eficaz. Construido específicamente para el modelado de temas, incluidas múltiples tecnologías Latent Dirichlet Assignment y Large Scale Integrated (LDA y LSI).

Puede manejar grandes cantidades de texto. Como tal, se diferencia de otros paquetes de aprendizaje automático que se centran en el procesamiento en memoria. Gensim también facilita implementaciones multinúcleo eficientes de varios algoritmos para amplificar la velocidad de procesamiento. Facilita herramientas más convenientes para el procesamiento de texto que otros paquetes como Scikit-learn, R, etc.

2.2.10.4. SpaCy

Es una biblioteca de Python que le permite crear aplicaciones de procesamiento de lenguaje natural (NLP) como lo manifestó Emilio (2020). SpaCy otorga modelos pre-entrenados en diferentes idiomas, junto con una sintaxis clara, ideal para principiantes en el campo de NLP. Asimismo, le permite crear nuevos modelos o volver a entrenar los modelos que proporciona con sus propios datos para entrenar modelos en dominios específicos. Diseñado para uso en producción, SpaCy

ofrece un framework para crear aplicaciones completas que requieren procesamiento de lenguaje natural, a diferencia de distintas bibliotecas como TensorFlow, que le permiten experimentar con diversas arquitecturas de redes neuronales o implementar modelos desarrollados recientemente.

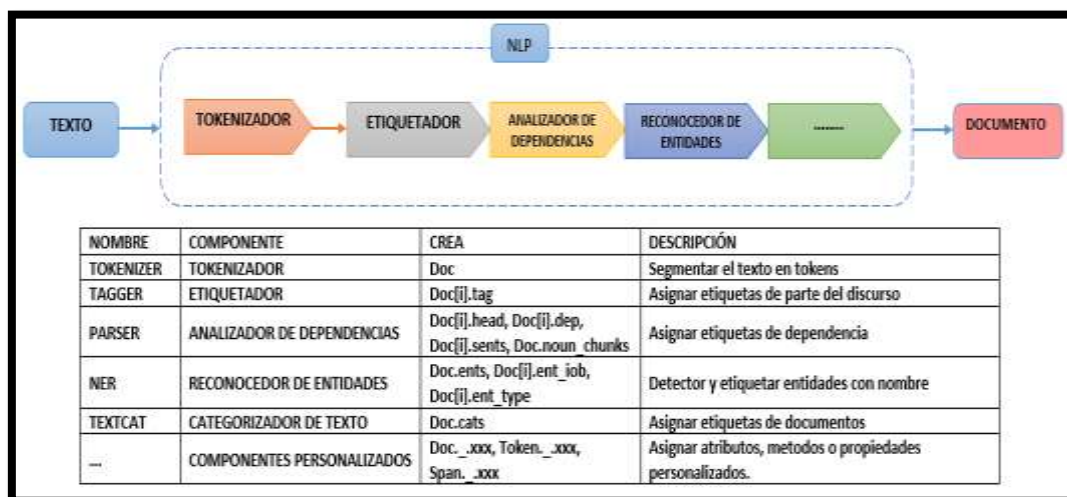


Figure 9. Información tomada de: *Introducción a SpaCy (Industrial Strength Natural Language Processing en Python)*.
Elaborado por Piza Guale Alexandra

2.2.10.5. Web Scrapping

Web Scrapping recopila datos e información web de forma automatizada según desarrollo (2020) básicamente, es minería de datos web. Web Scrapping manipula la recuperación de información, la recopilación de noticias, el monitoreo web, el marketing competitivo y más.

El Web Scrapping puede tener aplicaciones muy diversas. Además de la indexación del motor de búsqueda, el raspado web se puede utilizar para los siguientes propósitos, que incluyen:

- Crear una base de datos de contactos.
- Ver y comparar ofertas en línea.
- Recopile datos de varias fuentes en línea.
- Observa la evolución de tu presencia y reputación online
- Recopile datos financieros, meteorológicos u otros.
- Esté atento a los cambios en el contenido web.
- Recoger datos con fines de investigación.
- realizar exploración de datos o minería de datos

2.2.11 Algoritmos de procesamiento de NLP

Por lo tanto, el Equipo Expert.ai (2017) anuncio que, el algoritmo de PNL basado en tecnologías cognitivas como la tecnología semántica coloca al diccionario en el centro de su capacidad para comprender el lenguaje. Para eliminar la ambigüedad de las palabras, este método reconoce todos los aspectos estructurales del lenguaje, consulta la base de datos del diccionario o la red semántica para revelar todos los significados posibles de las palabras, y utiliza toda la otra información para eliminar la ambigüedad en el lenguaje.

Así mismo como lo revelo Mike Attal (2021) lo siguiente, en el escenario principal, se encuentran:

- **Limpieza:** Varía según la fuente de datos. Esta etapa incluye tareas como eliminar URL y emojis.

- **Normalización de datos:**

1. Tokenización, o división del texto en varias partes llamadas tokens.

Ejemplo:

"En el adjunto, encontrará el archivo correspondiente", "Encontrará",
"Adjunto", "Archivo", "Problema".

2. Tallo: Según género (masculino o femenino), número (singular o plural), persona (yo, tú, ellos ...), etc., la misma palabra se puede encontrar de diferentes formas. La derivación generalmente especifica un proceso heurístico simple de cortar las terminaciones para mantener solo la raíz.

Ejemplo:

"Encontrarás" -> "Lo encontré"

- **Lematización:** Incluye realizar la misma tarea, pero utilizando vocabulario y un análisis exhaustivo de la estructura de las palabras. La derivación le permite eliminar solo las terminaciones inflexibles, aislando así la forma canónica de la palabra, llamada lema.

Ejemplo:

"Encontrarás" -> "Buscar"

- Otras operaciones: eliminar números, puntuación, símbolos y palabras vacías y cambiarlas a minúsculas.

Para aplicar métodos de aprendizaje automático a problemas relacionados con el lenguaje natural, los datos de texto deben convertirse en datos digitales.

2.2.12 Arquitectura

Microsoft (2021) manifestó que es importante destacar que, el procesamiento del lenguaje natural (NLP) se utiliza para tareas como el análisis de sentimientos, el reconocimiento de temas, el reconocimiento de voz, la extracción de palabras clave y la clasificación de documentos.

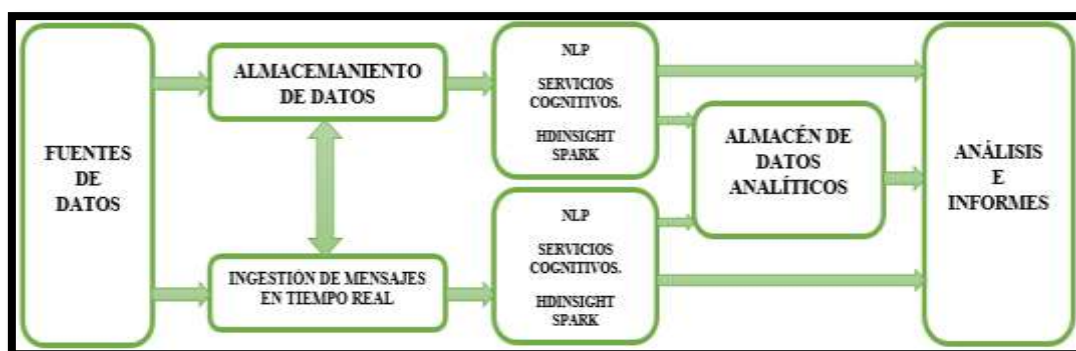


Figure 10. Información tomada de: Tecnología de procesamiento de lenguaje natural. Elaborado por Piza Guale Alexandra

Puede utilizar el procesamiento del lenguaje natural para clasificar documentos, por ejemplo, marcándolos como confidenciales o spam. La salida del procesamiento del lenguaje natural se puede utilizar para un procesamiento o una búsqueda posteriores.

2.2.13 Herramientas y lenguaje empleados en la Inteligencia Artificial

2.2.13.1. Lenguaje de Programación

Según menciona Akker (2021) que los lenguajes de programación se utilizan a menudo para evaluar las últimas tendencias, lo que orienta a la comunidad de desarrolladores sobre qué perfiles son los más necesarios en la tecnología actual.

Según el ranking TIOBE, Python también lidera el lenguaje de programación más utilizado con un sistema de puntuación del 11,27%, seguido de C con un 11,16% y finalmente Java con un 10,46%.



Figure 11. Índice TIOBE de enero de 2022. Información tomada de: <https://www.tiobe.com/tiobe-index/>

Según ARABA 4.0 (2020) denomina que cada vez hay más aplicaciones que te permiten utilizar Big Data o Inteligencia Artificial sin conocimientos de codificación, lo cierto es que siempre es mejor utilizar estas tecnologías en todo su potencial y adaptarlas a tus proyectos. Escribe tu propio código y aplicaciones.

Estos son algunos de los más utilizados:

2.2.13.1.1. Lenguaje C

ARABA 4.0 (2020) expreso que el lenguaje de programación más popular del mundo según el índice TIOBE. Sin duda está influenciado por el lenguaje en el que está escrito el sistema operativo Linux. Debe ser muy minucioso cuando se trata de escribir, pero es un estándar simple y limpio.

Tabla 3. *Característica de C++.*



- Simple
- Sintaxis flexible
- Flujo de control estructurado
- Posibilidad de abstracción de datos
- Variedad de operadores

Información tomada de Los lenguajes de programación más usados en Big Data e Inteligencia Artificial. Elaborado por Piza Gualé Alexandra

2.2.13.1.2. Lenguaje Python

Visus (2020) indico lo siguiente, Python es un lenguaje de programación

interpretado cuya idea principal es que cualquier persona con conocimientos básicos de programación pueda leerlo. Además, tiene una serie de propiedades que la hacen muy especial, lo que sin duda le otorga muchas ventajas y es la raíz de su uso generalizado:

Tabla 4. Característica de Python



- Lenguaje de programación multiparadigma
- Código abierto
- Es gratuito
- Soporte e integración con distintos lenguajes
- Posee integradas muchas bibliotecas estándar

Información tomada de ¿Para qué sirve Python? Razones para utilizar este lenguaje de programación. Elaborado por Piza Guale Alexandra

2.2.13.1.3. Lenguaje Java

Según ARABA 4.0 (2020) el lenguaje de programación más popular en el podio después de C y Python. Un lenguaje multiplataforma, polivalente, polivalente y muy estable. Se utiliza para crear sistemas operativos como Android, aplicaciones e incluso páginas web. Su programación orientada a objetos lo hace muy útil en el campo de la Inteligencia Artificial.

Tabla 5. Característica de Java



- Sintaxis similar a C y C++
- Orientado a Objeto
- Multiplataforma
- Portable
- Multihilo
- Solido
- Seguro


Información tomada de Los lenguajes de programación más usados en Big Data e Inteligencia Artificial. Elaborado por Piza Guale Alexandra

2.2.13.1.4. Lenguaje R

Por lo tanto Tecnología (2019) manifiesto que, R es un entorno de software libre (licencia GNU GLP) y un lenguaje de programación interpretado, es decir, ejecuta

instrucciones directamente sin compilar primero el programa en instrucciones de lenguaje máquina. En R, el término entorno se refiere a un sistema completamente planificado y consistente, en lugar de la acumulación de herramientas ad hoc e inflexibles que a menudo se encuentran en otro software de análisis de datos.

Tabla 6. Característica de R

	<ul style="list-style-type: none"> • Lenguaje de programación para análisis estadístico • Manejo y almacenamiento efectivo de datos • Contiene lenguaje propio (R) • Se distribuye gratuitamente con la licencia de GNU • El código está escrito en C para sistemas Unix y Linux • Contiene archivos binarios seleccionados para Windows
--	--

Información tomada de Lenguaje R, ¿qué es y por qué es tan usado en Big Data? Elaborado por Piza Guale Alexandra

2.2.13.1.5. Lenguaje PHP

Según Deyimar (2020) estableció que PHP es un lenguaje de secuencias de comandos creado para la comunicación del lado del servidor. Por lo tanto, puede manejar varias funciones del lado del servidor, como recopilar datos de formularios, administrar archivos en el servidor, modificar bases de datos y más.

El lenguaje PHP fue creado originalmente por Rasmus Lerdorf para rastrear a los visitantes de su página de inicio personal. Lerdorf finalmente lo lanzó como un proyecto de código abierto.

Tabla 7. Característica de PHP



- Lenguaje muy fácil
- Integración con 8 servidores HTTP
- Acceso a 20 tipos de Bases de Datos
- Ampliación más eficiente en diseño modular.
- Licencia abierta

Información tomada de ¿Qué es PHP? Una guía para principiantes. Elaborado por Piza Guale Alexandra

2.2.13.2. Entorno de Desarrollo Integrado (IDE)

2.2.13.2.1. Entorno de desarrollo integrado (IDE) de escritorio

Acharya (2021) expreso lo siguiente, los IDE son software en sí mismos, y consisten en herramientas de desarrollo utilizadas para desarrollar y probar software. Proporciona un entorno de desarrollo en el que todas las herramientas están disponibles en una interfaz gráfica de usuario (GUI) fácil de usar.

Un IDE incluye principalmente:

- Editor de código para escribir código de software
- Automatización de la construcción local
- Depurador de programas.

2.2.13.2.1.1. SlickEdit

Zamora (2021) manifesto que, SlickEdit es un editor de código fuente comercial multiplataforma, editor de texto, editor de código Revisión: Slickedit comenzó en 1988 como un editor de modo de caracteres para DOS y OS/2. Proporciona un editor de código potente y altamente personalizable y un IDE que puede editar rápidamente hasta 2 TB de datos. Es muy apreciado por sus amplias herramientas de codificación y sus potentes funciones de programación que ahorran tiempo.

Si necesita flexibilidad para codificar en múltiples lenguajes de programación en múltiples plataformas, SlickEdit es una buena opción. Las características incluyen:

Tabla 8. Característica de Se

- Ampliación de sintaxis.
- Plantillas de código.
- Finalización automática.
- Atajos de entrada personalizados con alias.
- Ampliación funcional utilizando el lenguaje de macros Slick-C.
- Barras de herramientas personalizables, acciones del ratón, menús y combinaciones de teclas.
- Soporta Perl, Python, XML, Ruby, COBOL, Groovy, etc.

Información tomada de SlickEdit, Historia, Características más relevantes. Elaborado por Piza Guale Alexandra

2.2.13.2.1.2. KDevelop

Por lo tanto, Acharya (2021) indico que, basado en tecnologías modernas de código abierto, proporciona un entorno de desarrollo perfecto para los desarrolladores que trabajan en proyectos de cualquier tamaño. En esencia, es una combinación de un editor sofisticado y un análisis de código semántico que brinda una buena experiencia de programación.

Además, KDevelop proporciona diferentes flujos de trabajo necesarios para ayudar a los desarrolladores, mejorar la calidad de su código, permite verificar la funcionalidad y facilita la implementación que se necesita construir. KDevelop es un IDE de Python extensible y admite otros lenguajes de programación como C, C++, PHP, entre otros.

2.2.13.2.1.3. PyCharm

Acharya (2021) anuncio que, tiene un editor de código inteligente que brinda soporte de primera clase no solo para Python, sino también para JavaScript, TypeScript, CoffeeScript, CSS, lenguajes de plantillas conocidos, Node.js, AngularJS y otros más. Si desea saltar a una clase, uso, implementación, prueba, etc. específicos, obtiene una búsqueda inteligente. PyCharm viene

con una serie de herramientas de desarrollo, incluido un ejecutor de pruebas y un depurador, terminal y generador de perfiles de Python.

2.2.13.2.1.4. Visual Studio Code

Como Acharya (2021) informó que, proporciona un rendimiento mejorado a través de programas IntelliSense para C++. Le permite escribir variables con precisión y rapidez utilizando sugerencias de código. Mantiene la velocidad y supera la complejidad al navegar a archivos, componentes, tipos o símbolos. También puede mejorar su código a través de posibles sugerencias que realiza el editor, sugiriendo acciones como agregar parámetros, renombrar funciones y más. Con CodeLens, encuentra información básica, como los cambios realizados en el código y su impacto, y verifica que el método se haya probado por unidad.

Como señalo Microsoft (2021) puede usar sugerencias para presentar las acciones sugeridas. Por ejemplo, puede proporcionar acciones para mover llaves abiertas a una nueva línea o moverlas al final de la línea anterior. Las siguientes instrucciones explican cómo mostrar una lámpara en la palabra actual y realizar dos acciones: convertir a mayúsculas y convertir a minúsculas.



Figure 12. Información tomada de Tutorial: Mostrar sugerencias de bombilla. Elaborado por Piza Guale Alexandra

2.2.13.2.1.5. LiClipse

Es perfecto para usted porque le brinda una experiencia completamente nueva con una funcionalidad lista para usar. Además de Python, su editor rápido admite plantillas Java, JavaScript, CSS, PHP, PERL, C, C++, HTML,

Go, Ruby, Django y unos 30 lenguajes más. Cuenta con múltiples cursores, barras de desplazamiento temáticas, guías con sangría vertical, capacidad de búsqueda mejorada, filtrado adicional, editores abiertos y compatibilidad con carpetas externas.

La última versión de LiClipse es 7.0.1, que incluye PyDev 8.0.0, Python 3.9 actualizado, mejora en el depurador, una solución rápida para convertir cadenas en cadenas f y EGit actualizado.

2.2.13.2.1.6. Spider

Es un potente entorno científico de Python diseñado para desarrolladores, científicos de datos e ingenieros. Combina perfectamente la edición, el análisis y la depuración avanzados con la exploración de datos, la inspección profunda, las mejores visualizaciones y la ejecución interactiva.

Permite trabajar de manera eficiente con un editor multilingüe con navegador de clase/función, análisis de código, finalización automática de código, definiciones de acceso y división vertical/horizontal.

2.2.13.2.1.7. Wing

Entorno de desarrollo Smart Python: Python IDE está diseñado para brindarle más productividad. Wing se encarga de escribir el código de Python brindándole comentarios instantáneos e interactivos en tiempo de ejecución. Puede escribir fácilmente documentación y código de navegación. Con un análisis de código en profundidad, puede evitar errores comunes y detectar problemas temprano.


Wing también puede emular vi, Eclipse, emacs, Visual Studio, MATLAB y XCode. Su depurador permite corregir código multiproceso desde IDE alojados en marcos web. Además, Wing proporciona una matriz y un visor de marcos de datos para realizar análisis de datos y tareas científicas.

2.2.13.2.1.8. Jupyter

Según Desarrollo web (2019) indico que Jupiter es una aplicación cliente-servidor lanzada en 2015 que le permite crear y compartir documentos web con formato JSON que siguen un esquema versionado y una lista ordenada

de unidades de entrada y salida. Estas celdas contienen código, texto, fórmulas y ecuaciones matemáticas, contenido multimedia y más. El programa se ejecuta desde una aplicación web del lado del cliente que se ejecuta en cualquier navegador estándar. El requisito previo es que el servidor Jupyter esté instalado y ejecutándose en el sistema. Los documentos creados en Jupyter se pueden exportar a formatos como HTML, PDF, Markdown o Python, o se pueden compartir con otros usuarios por correo electrónico, Dropbox, GitHub o el visor de Jupyter integrado. Los dos componentes principales de un Jupyter Notebook son un conjunto de núcleos (intérpretes) y un tablero. Cada núcleo o kernel es el motor de ejecución de un lenguaje, responsable de procesar las solicitudes y devolver las respuestas adecuadas.

Tabla 9. *Característica de Jupyter*

	<ul style="list-style-type: none"> • Código abierto • Gratuito • Trabaja en la nube • Control de versiones • Comprende más de 50 lenguajes de programación
--	---

Información tomada de Digital Guide IONOS. Elaborado por Piza Guale Alexandra

2.2.13.2.2. Entorno de desarrollo integrado (IDE) en la nube

2.2.13.2.2.1. Google Colab

Baume (2021) indico que Google Colab es una herramienta para escribir y ejecutar código Python 2.7 y 3.6 en la nube, R y Scala aún no están disponibles.

Aunque tiene algunas limitaciones y puede consultarse en su página de preguntas frecuentes, es una herramienta ideal no solo para practicar y mejorar los conocimientos de técnicas y herramientas de ciencia de datos, sino también para el desarrollo de aplicaciones de aprendizaje automático (piloto). Aprendizaje profundo sin invertir en hardware o recursos en la nube.

Con Colab, puede crear o importar cuadernos que haya creado, y puede compartirlos y exportarlos en cualquier momento.

Tabla 10. *Característica de Google Colab*



- No requiere configuración adicional
- Obtiene las bibliotecas de Python para desarrollar, analizar o visualizar proyectos
- Acceso al hardware de Google de forma gratuita
- Puede guardar y compartir desde la nube y se enlaza al Drive

Información tomada de Breve introducción a Goolge Colab. Elaborado por Piza Guale Alexandra

2.2.13.2.2. IBM Watson Studio

De modo que Hagarty & Karlsen (2019) anunciaron que contiene desde un enfoque semiautomático que emplea la herramienta AutoAI Experiment que comprende un enfoque esquemático utilizando flujos de SPSS Modeler y un modelo completamente programado utilizado por Jupyter Notebooks para Python.

Para la mayoría de las actividades que realizamos utilizamos IBM Watson Studio. Ofreciéndonos el medio y las herramientas necesarias para solucionar problemas de negocio trabajando en colaboración con datos. Existe la posibilidad de escoger herramientas para examinar y observar datos; depurar y dar forma a los datos; ingresar datos de transmisión; o generar, instruir y extender modelos de aprendizaje automático.



Figure 13. Información tomada de Introducción a IBM Watson Studio. Elaborado por Piza Guale Alexandra

Tabla 11. Característica de IBM Watson



- Implementa proyectos para administrar los recursos.
- Obtiene acceso a datos a partir de las conexiones a la nube
- Establece y conserva listados de datos para descubrir, equilibrar y distribuir datos
- Optimiza datos depurados y modela los datos

Información tomada de Introducción a IBM Watson Studio. Elaborado por Piza Guale Alexandra

2.2.13.2.2.3. Azure Machine Learning

Microsoft (2021) señalo que Azure Machine Learning es un servicio en la nube, su función es agilizar y dirigir el ciclo de vida del proyecto de enseñanza instantáneo. Los expertos de este proceso de aprendizaje, científicos y los ingenieros son autorizados en utilizarlo en sus flujos de trabajo diarios: preparar, aplicar modelos y administrar MLOps.

Es factible instaurar un modelo en Azure Machine Learning o utilizar un modelo ya establecido tomando en consideración una plataforma de código abierto, como Pytorch, TensorFlow o scikit-learn. Las herramientas MLOps le ayudan a inspeccionar, volver a preparar y a implementar prototipos.

Tabla 12. Característica de Azure Machine Learning



- Ver ejecuciones, métricas, registros, resultados y más.
- Crear y editar cuadernos y archivos.
- Administrar bienes públicos tales como:
 - Recibo de datos
 - Calcular
 - Alrededores
- Ver métricas, resultados e informes de rendimiento.
- Ver canalizaciones creadas a través de la interfaz de desarrollador.
- Crear trabajos de AutoML.

Información tomada de ¿Qué es el aprendizaje automático de Azure? Elaborado por Piza Guale Alexandra

2.2.13.2.2.4. Amazon Machine Learning

Perrier (2017) expreso que, Amazon Machine Learning es un servicio en línea de Amazon Web Services(AWS) que desarrolla un aprendizaje controlado para el análisis predictivo. Este servicio en línea tuvo origen en abril de 2015 en la cima de AWS, Amazon ML cuenta con una lista cada vez más amplia de servicios de aprendizaje automático fundamentados en la nube, como Microsoft Azure, Google Prediction, IBM Watson, Prediction 10, BigML, etc. Tienen una oferta comúnmente conocida como aprendizaje automático como servicio o MLaaS siguiendo un patrón de denominación de varios servicios como SaaS, PaaS y IaaS, ya sea para Software, plataformas o infraestructura como servicio.

Amazon ML sintetiza significativamente el proceso de análisis predictivo y su implementación que conlleva cuatro decisiones de diseño integradas en la plataforma:

Tabla 13. *Característica de Amazon Machine Learning*



- Límite de tareas: clasificación binaria, múltiple y regresión.
- Cuenta con un solo algoritmo lineal.
- Selección limitada de métricas para estimar la calidad de la predicción.
- Conjunto simple de parámetros de ajuste para el algoritmo predictivo subyacente.

Información tomada de Effective Amazon Machine Learning. Elaborado por Piza Guale Alexandra

2.2.4 Fundamentación legal

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."

Artículo 386.- El Sistema Nacional, de Ciencia, Tecnología, Innovación, y; Saberes Ancestrales "Comprenderá programas, políticas, recursos, acciones, e incorporará 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."

Ley orgánica de la contraloría general del estado, Arts. 5, 8, 31

Art. 298.- Se establecen preasignaciones presupuestarias destinadas a los gobiernos autónomos descentralizados, al sector salud, al sector educación, a la educación superior; y a la investigación, ciencia, tecnología e innovación en los términos previstos en la ley. Las transferencias correspondientes a preasignaciones serán predecibles y automáticas. Se prohíbe crear otras preasignaciones presupuestarias.

Sección octava

Ciencia, tecnología, innovación y saberes ancestrales

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.
2. Recuperar, fortalecer y potenciar los saberes ancestrales.
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.

Ministerio de Telecomunicaciones y de la Sociedad de la Información

Rectoría del sector

Art. 140.- El Ministerio encargado del sector de las Telecomunicaciones y de la Sociedad de la Información es el órgano rector de las telecomunicaciones y de la sociedad de la información, informática, tecnologías de la información y las comunicaciones y de la seguridad de la información. A dicho órgano le corresponde el establecimiento de políticas, directrices y planes aplicables en tales áreas para el desarrollo de la sociedad de la información, de conformidad con lo dispuesto en la presente Ley, su Reglamento General y los planes de desarrollo que se establezcan a nivel nacional.

Capítulo III

Metodología

3.1. Propuesta tecnológica

En el presente capítulo se expone la metodología que se empleó en la investigación y a la vez se procede a revisar las técnicas NLP disponibles identificando las adecuadas para el desarrollo de un modelo NLP. Asimismo, se efectuó una encuesta para el nivel de aceptación del producto final a construir realizada a la población de la zona 8 de la provincia de Guayas, la cual está distribuida por los cantones de Guayaquil, Samborondón y Durán, se realizó la recopilación de datos necesarios para la investigación que conformarían la DataSet con variables de entrada y variables de salida, para este efecto se estableció la población y la muestra para este trabajo. Los estudios utilizados son mixtos porque se revisará información bibliográfica de técnicas NLP, se recogerán datos cuantitativos y cualitativos.

3.1.1 Descripción del proceso metodológico

El proceso metodológico como lo explicó Antonio (2017) es la transformación de la realidad en datos comprensibles y cognoscibles destinados a hacer comprensible para el objetivo de investigación. Los procesos metodológicos a menudo se confunden con técnicas o herramientas para registrar o recopilar datos.

Una metodología es una guía para resolver problemas y encontrar alternativas de solución. Un proceso metodológico se apoya en: indicar la forma de entender el objeto de investigación, orientar la generación de conceptos teóricos brindar criterios para determinar los procedimientos y procesos más adecuados para comprender el objeto de investigación ser capaz de desarrollar soluciones conceptuales para explicar el objeto de investigación o la investigación propuesta.

3.2. Tipos de investigación

Según Molina (2021) indicó que, la clasificación puede variar según el propósito, es decir, su clasificación depende de las metas propuestas, la profundidad del proyecto que se llevará a cabo, el tipo de datos que se debe analizar (ya sean cuantitativos o cualitativos), y otros factores que determinarán el tipo de investigación realizada.

A continuación, se detallan los tipos de investigación:

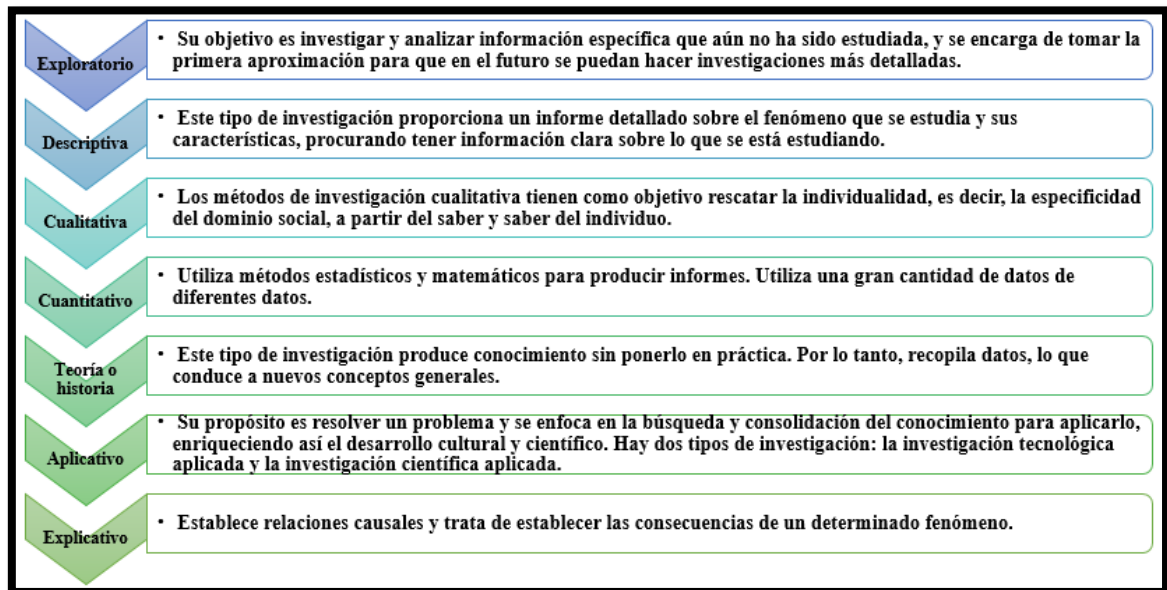


Figure 14. Información tomada de: Tipos de investigación y características. Elaborado por Piza Guale Alexandra

Como indico Arias (2020) que, dependiendo del propósito de la investigación, podemos distinguir:

- Investigación metodológica
- Investigación análisis jerárquico de la información
- Investigación fuente de información
- Investigación de campo de estudio
- Investigación profunda
- Investigación tipo de razonamiento
- Investigación de tiempo

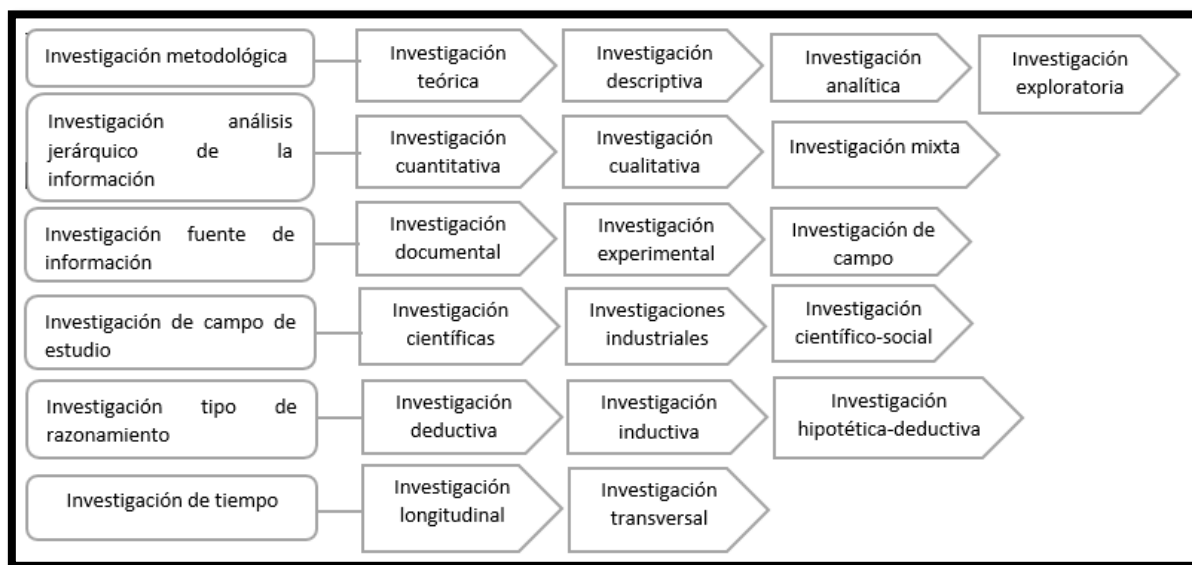


Figure 15. Información tomada de: Tipos de investigación. Elaborado por Piza Guale Alexandra

3.3 Metodología de investigación

3.3.1 Metodología bibliográfica

Como lo expreso Ocampo Campos (2017) que, el uso del método bibliográfico o documental es aquel que hace uso de textos como fuentes primarias para la obtención de datos, sin embargo, la misma no solo se trata del uso de libros, sino que también se usa cualquier tipo de fuente documental como películas, música, pinturas, microfilmes, sitios en la Internet.

3.3.2 Metodología Cualitativa

La investigación cualitativa como lo indico Latinoamérica (2019), es un conjunto de técnicas de investigación utilizadas para obtener una visión general del comportamiento y las percepciones de las personas sobre un tema en particular. Las ideas y suposiciones que genera pueden ayudar a comprender cómo la población objetivo percibe un problema y ayudar a definir o identificar opciones relevantes para ese problema. También le permite analizar los datos utilizados en las ciencias sociales y obtener un conocimiento profundo a través del análisis de texto.

A continuación, se presentan algunas características de la investigación cualitativa:

- La investigación cualitativa se enfoca en comprender o explicar el comportamiento de grupos, eventos o temas. Estas son algunas de las características de la investigación cualitativa.

- La investigación cualitativa tiene como objetivo describir y analizar la cultura y el comportamiento de los seres humanos y sus grupos desde la perspectiva del investigador.
- La investigación cualitativa se basa en estrategias de investigación flexibles e interactivas.
- Es un método de investigación más descriptivo que se centra en la interpretación, la experiencia y su significado.
- Los datos derivados de tales estudios no son medibles estadísticamente y deben interpretarse subjetivamente.
- Este tipo de investigación utiliza métodos como observaciones, entrevistas y grupos focales o ideas sobre debates en las comunidades.

¿Cómo analizar los resultados de una investigación cualitativa?

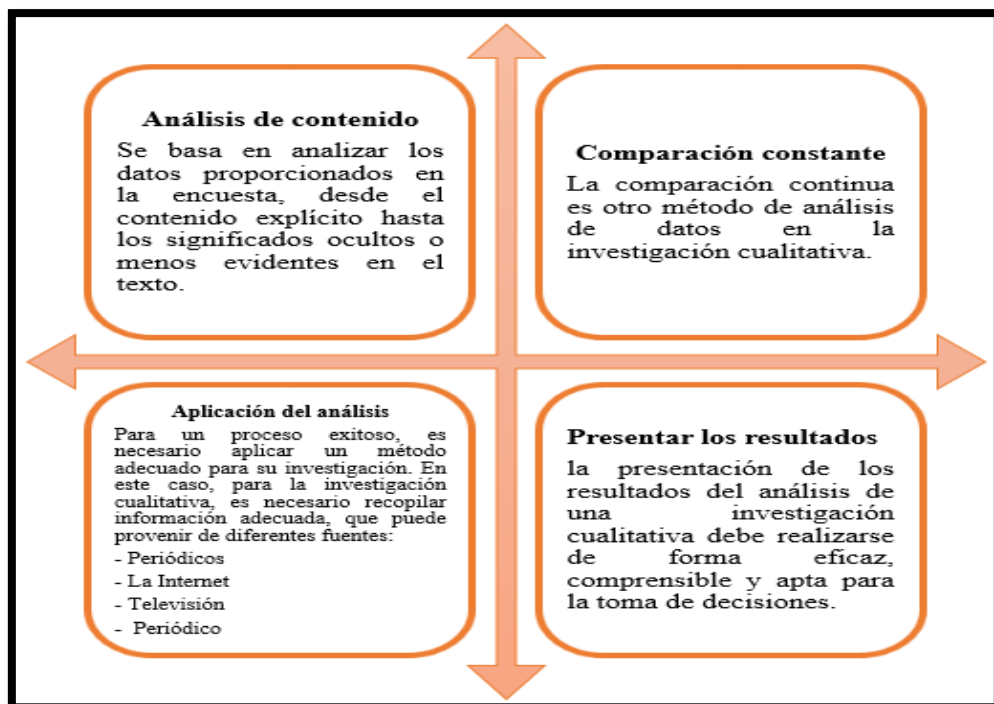


Figure 16. Información tomada de: ¿Qué es la investigación cualitativa? Elaborado por Piza Guale Alexandra

3.3.3 Metodología Cuantitativa

La metodología cuantitativa como lo manifestó Campos (2020) es el que permite examinar datos numéricamente, necesario en el campo de la estadística. Para que esto se lleve a cabo, la metodología cuantitativa requiere una dependencia lineal entre los fundamentos de la propuesta de exploración. Es decir, debe haber claridad entre los

elementos de la pregunta de estudio que constituyen el problema, poder definirlo, limitarlo y saber exactamente dónde comienza el inconveniente, en qué dirección va, y qué tipo de asociaciones existen entre sus elementos. Los términos que componen un problema de investigación lineal se denominan: variables, relaciones entre variables y unidades de observación.

Los métodos más comunes son los siguientes:

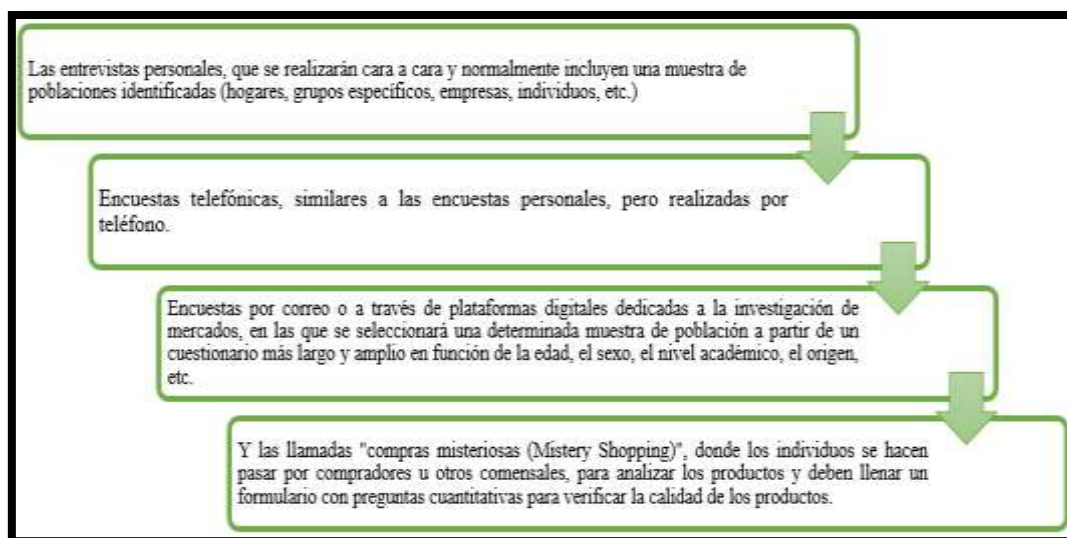


Figure 17. Información tomada de: *Investigación cuantitativa: qué es y características*. Elaborado por Piza Guale Alexandra

3.3.4 Metodología Mixta

La metodología mixta según Jiménez, Pacheco, & García (2019), la combinación de estrategias de investigación cualitativas y cuantitativas se convierte en una de las formas más apropiadas de investigación científica sobre la contravención y su impacto en la sociedad. Los métodos mixtos son un enfoque pragmático en el campo de la investigación empírica que asume que la recopilación de datos desde diferentes perspectivas (cualitativa y cuantitativa) conduce a una mejor comprensión del fenómeno en estudio.

Un estudio puede comenzar con una encuesta exploratoria para averiguar la prevalencia.

Las características del método híbrido son las siguientes:

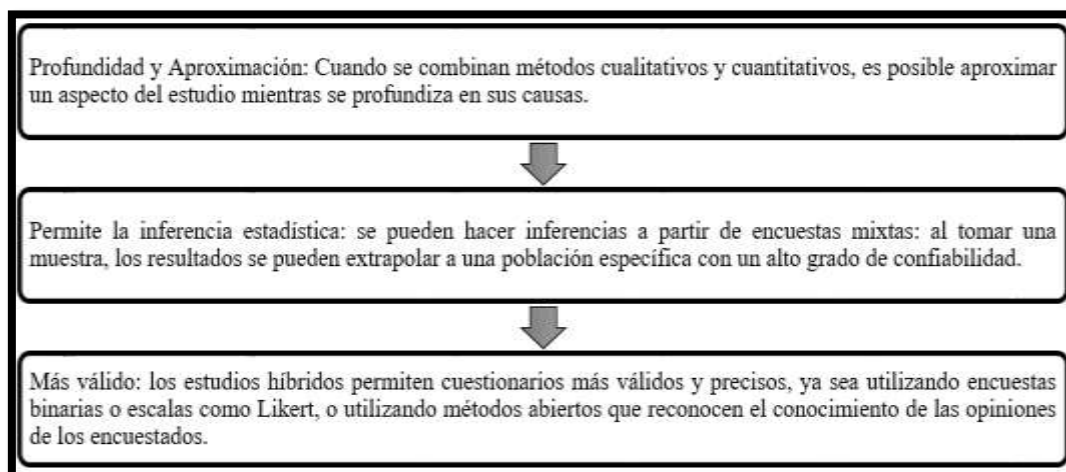


Figure 18. Información tomada de: Metodología mixta; estudios de caso. Elaborado por Piza Gualé Alexandra

3.4 Técnicas de investigación

3.4.1 Encuesta

Una encuesta como lo declaró Infobae (2021), es una técnica realizada mediante la realización de un informe de cuestionario de una muestra de una población. Las encuestas proporcionan información sobre las opiniones, actitudes y comportamientos de los ciudadanos.

La encuesta es aplicable a la necesidad de probar hipótesis o descubrir soluciones a problemas, y de identificar e interpretar un conjunto de testimonios que sirvan a un propósito declarado de la manera más organizada posible.

Para poder realizar la recaudación de información necesaria de datos solicitados para el entrenamiento de las técnicas NLP, se efectuó 5000 encuestas la cual se utilizó un formulario virtual, donde se empleó en la herramienta de Google Forms, ya que es una de las herramientas gratuita y muy factible que brinda un link del cual puede ser distribuido por medio de redes sociales o correos electrónicos, también, porque es fácil de manipular las tabulaciones en cuestiones de respuestas del formulario establecido.

3.4.2 Entrevista

La entrevista como lo expreso Solís (2020) es una técnica muy útil para recolectar datos para la investigación cualitativa, se define como una conversación con un propósito específico, más que el simple acto de hablar. Es una herramienta técnica en forma de diálogo hablado. Canales la define como “la comunicación interpersonal que se establece entre el investigador y el sujeto de investigación con el fin de obtener una respuesta verbal a la pregunta planteada por la pregunta.

Para la recolección de datos, se llevó a cabo una técnica muy útil como son las entrevistas, en las que se hacía una variedad de preguntas para conocer los criterios de especialistas de tecnología con conocimientos en IA, médicos generales o médicos nutricionistas y para las personas común que hayan interactuado con un asistente virtual. Entre ellos, se seleccionó una muestra aleatoria de expertos correspondientes de la zona 8 de la provincia del Guayas.

3.5 Descripción del procedimiento metodológico

3.5.1 Población

La población en donde se realizará la investigación está dirigida para los moradores de zona 8 de la provincia del Guayas, la cual está distribuida por los cantones de Guayaquil, Samborondón y Durán, también es destinada para especialistas de tecnología con conocimientos en IA, médicos generales o médicos nutricionistas, en donde, se tiene un estimado de 387 personas encuestadas. Como lo indico el Instituto Nacional de Estadísticas y Censos (INEC), (2010), Como institución responsable de las estadísticas oficiales, aparte de brindar información estadística relevante, oportuna, confiable y de alta calidad, es la entidad encargada de planificar, regular y certificar la producción de los sistemas estadísticos nacionales, además de diseñar, implementar y evaluar de manera innovadora las estadísticas necesarias para la planificación nacional métodos, métricas y análisis de la información. Por medio de la base de datos que contiene INEC se visualizó que, en el cantón del Guayas poseo 2.350.915 habitantes, en el cantón de Samborondón obtuvo 67.590 habitantes y en el cantón Durán tuvo 235.769 habitantes, donde se alcanzó un total de 2.654.274 habitantes en la zona 8 de la provincia del Guayas.

En las encuestas donde la variable principal es el tipo de estudio cualitativo y proporcional fenómeno estudiado en población de referencia. (Aguilar, 2005)

la muestra se calcula mediante la siguiente fórmula:

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

Donde:

Z = nivel de confianza

p = proporción aproximada del fenómeno en estudio población de referencia.

q = proporción de la población de referencia que no está presente. El fenómeno objeto de estudio $(1-p)$.

N = tamaño del universo

e = margen de error de estimación

n = tamaño de muestra

3.5.1. Muestra

El muestreo se divide en dos categorías probabilísticos y no probabilísticos así lo indica Ávila (2019). Algunos son probabilísticos, basados en una base de igual probabilidad. El método que utilizan es buscar que todos los objetos de la población tengan la misma probabilidad de ser seleccionados para representarla y formar parte de la muestra, y suelen ser los más utilizados porque buscan una mayor representatividad.

En un enfoque no probabilístico, los sujetos se seleccionan cuidadosamente de la población utilizando criterios específicos, buscando ser lo más representativos posible. Aun así, no pueden utilizarse para inferir resultados generales.

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

$$n = \frac{2.654.274 * 1.96_{\alpha}^2 0.50 * 0.50}{0.05^2 (2.654.274 - 1) + 1.96_{\alpha}^2 * 0.50 * 0.50}$$

$$n = 384$$

Por otra parte, se realizó la recopilación de datos por medio de una DataSet, la cual se adquirió alrededor de 4924 datos, los cuales fueron compartidos para que se llevara a cabo el entrenamiento de la técnica NLP establecida, por la gran recopilación de información que se recibió a través de entrenamiento se logró adquirir resultados con efectividad.

En donde se aplicó la técnica del 80-20. En el cual fue distribuido el 80% para el entrenamiento y el 20% para pruebas.

3.5.2 Análisis de las encuestas

Pregunta 1.1 Seleccione la Edad

Tabla 14. Edad

Respuestas	Encuestados	Porcentaje
18 - 25 años	263	68%
26 - 40 años	90	23%
41 - 50 años	23	6%

Restante	11	3%
Total	387	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

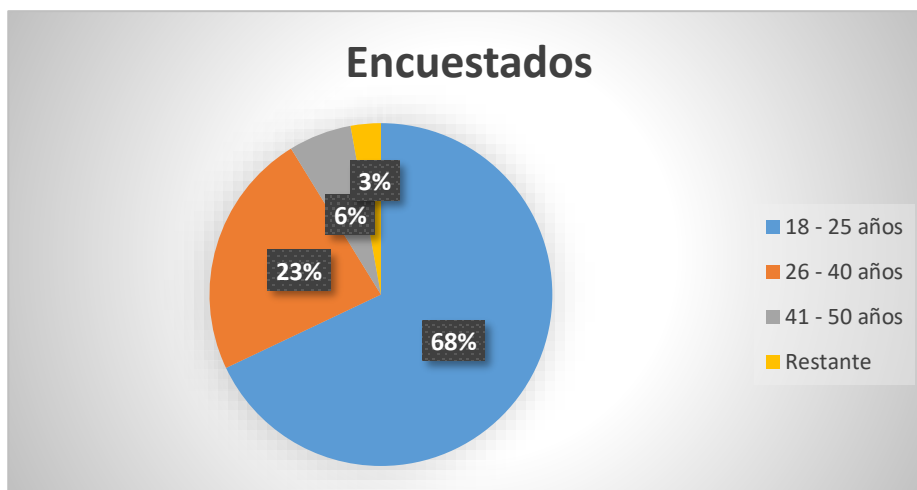


Figure 19. Edad de los encuestados. Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

Respecto al resultado del Gráfico Estadístico (N°19) correspondiente relacionada a la selección de la edad, se observa que de la muestra de 387 encuestados la cantidad es predominada por adolescentes y jóvenes correspondiente a la edad de 18 a 25 años y el rango más bajo corresponde a las personas mayores de 50 años.

Pregunta 1.2 género

Tabla 15. Género

Respuestas	Encuestados	Porcentaje
Femenino	220	57%
Masculino	167	42%
Total	387	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

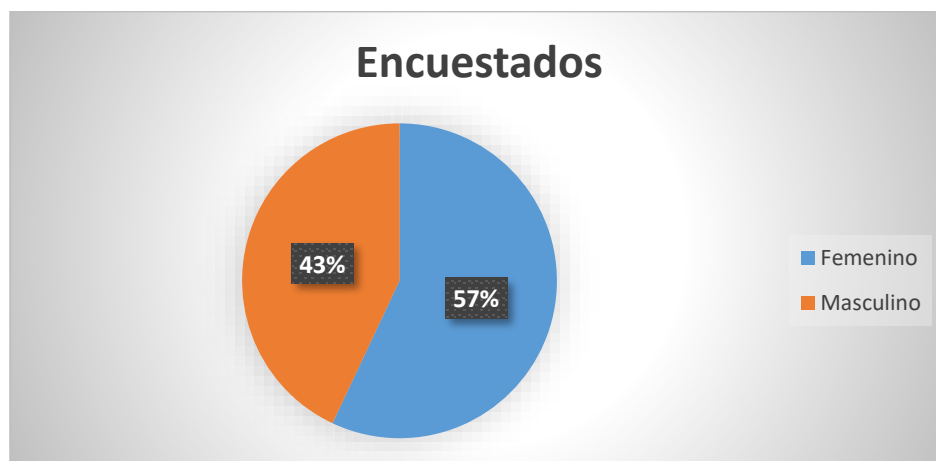


Figure 20. Género de los encuestados. Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

En relación con los resultados del gráfico estadístico (N°20) acerca del género de las personas encuestadas, se determina que, el género femenino es el que predomina las encuestas con un 57%, mientras que el género masculino tiene un rango inferior que corresponde al 43%, información que puede ser un indicativo muy importante.

Pregunta 1.3 Lugar que reside de la zona 8 del Ecuador.

Tabla 16. Lugar que reside de la zona 8 del Ecuador

Respuestas	Encuestados	Porcentaje
Durán	26	7%
Samborondón	8	2%
Guayaquil	330	85%
Otra ciudad del Ecuador	20	5%
Otro país	3	1%
Total	387	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

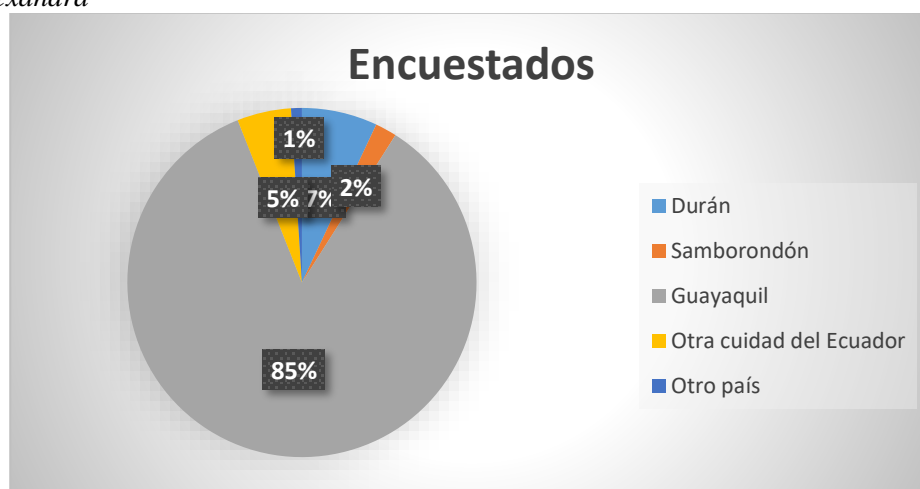


Figure 21. Lugar donde reside el encuestado. Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

Respecto al gráfico estadístico (N°21) en relación con lugar donde residen los encuestados, se determina que la mayor cantidad de personas son residentes de Guayaquil que alcanza un 85%, es la ciudad en la que prácticamente se ha dedicado esta Tesis de Titulación, mientras que el porcentaje más bajo corresponde las personas que no pertenecía a los lugares que estaba dirigida la encuesta.

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

Tabla 17. Conocer cómo el CORONAVIRUS (Covid-19) afecta nuestra salud

Respuestas	Encuestados	Porcentaje
De acuerdo	52	13%
Parcialmente de acuerdo	28	7%
Totalmente de acuerdo	305	79%
Totalmente en desacuerdo	2	1%
Total	387	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

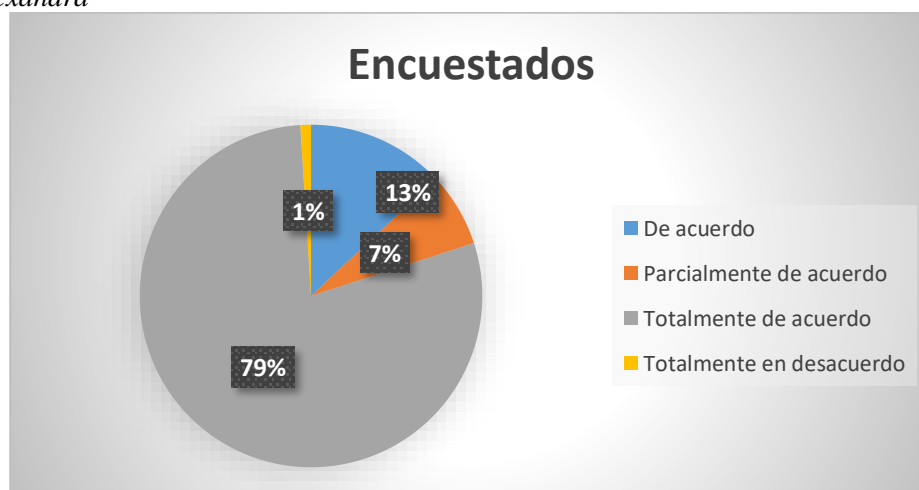


Figure 22. Considera importante CONOCER cómo el CORONAVIRUS (Covid-19) afecta nuestra salud. Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

Respecto al gráfico estadístico (N°22) correspondiente a la importancia de conocer como CORONAVIRUS (Covid-19) afecta la salud, la población se encuentra en un total acuerdo con la pregunta formulada alcanzando un 79%, lo que significa que la mayoría de la población considera importante que se dé a conocer esta información, mientras que solo el 7% se mantienen parcialmente de acuerdo y un 1% menciona su inconformidad.

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

Tabla 18. Vacunas contra el coronavirus (Covid-19)

Respuestas	Encuestados	Porcentaje
De acuerdo	91	24%
Parcialmente de acuerdo	134	35%
Parcialmente en desacuerdo	27	7%
Totalmente de acuerdo	109	28%
Totalmente en desacuerdo	26	7%
Total	387	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

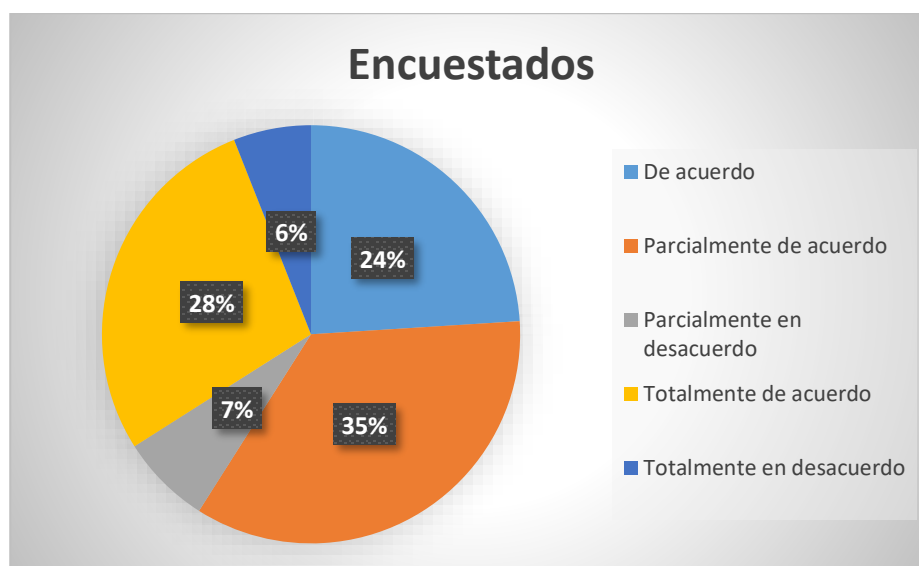


Figure 23. De acuerdo que las vacunas contra el coronavirus (Covid-19) son efectiva eliminando el virus. Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra.

Respecto al gráfico (N°23) con relación a efectividad de las vacunas contra el coronavirus (Covid-19), se determina que un 35% de la población encuestada está parcialmente de acuerdo que, con la efectividad de la vacuna, un 28% está totalmente de acuerdo, le sigue un 24% de acuerdo con la pregunta, sin embargo, una pequeña cantidad de la población no está de acuerdo con lo que manifiesta la pregunta.

Pregunta 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, por ejemplo: Los hábitos saludables, ¿evolución del virus etc.?

Tabla 19. Información adecuada y actualizada

Respuestas	Encuestados	Porcentaje
De acuerdo	95	25%
Parcialmente de acuerdo	128	33%
Parcialmente en desacuerdo	31	8%

Totalmente de acuerdo	122	32%
Totalmente en desacuerdo	11	2%
Total	387	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

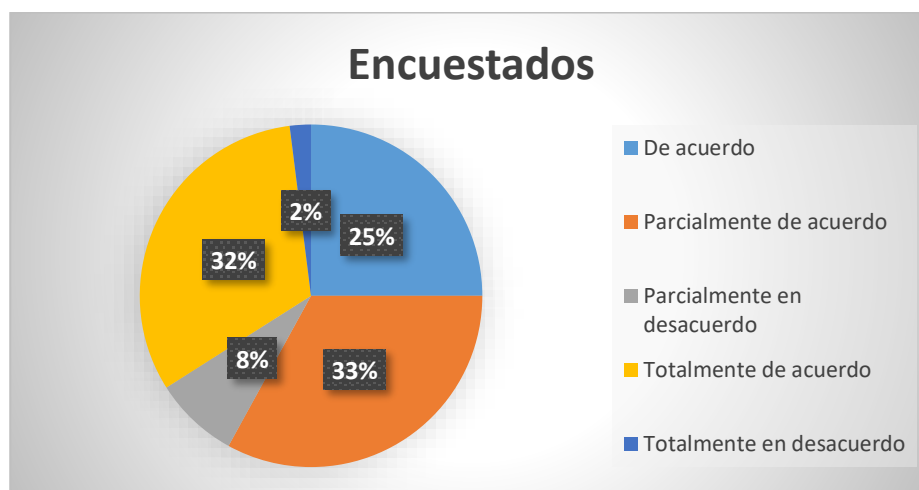


Figure 24. . De acuerdo que la información del coronavirus (Covid-19) que recibe del ministerio de salud o subcentro de salud por cualquier medio de comunicación es la adecuada y actualizada. Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra.

Respecto al gráfico (N°24) la cantidad de personas que se encuentran parcialmente de acuerdo con la pregunta es del 33% mientras le siguen con 25% las personas de acuerdo. La cantidad más baja corresponden a los encuestados que se encuentran parcialmente en desacuerdo.

Pregunta 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?

Tabla 20. Hábitos saludables de una persona contagiada.

Respuestas	Encuestados	Porcentaje
No tenía conocimiento	30	8%
Posee alto conocimiento del tema	191	49%
Posee bajo conocimiento del tema	166	43%
Total	387	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

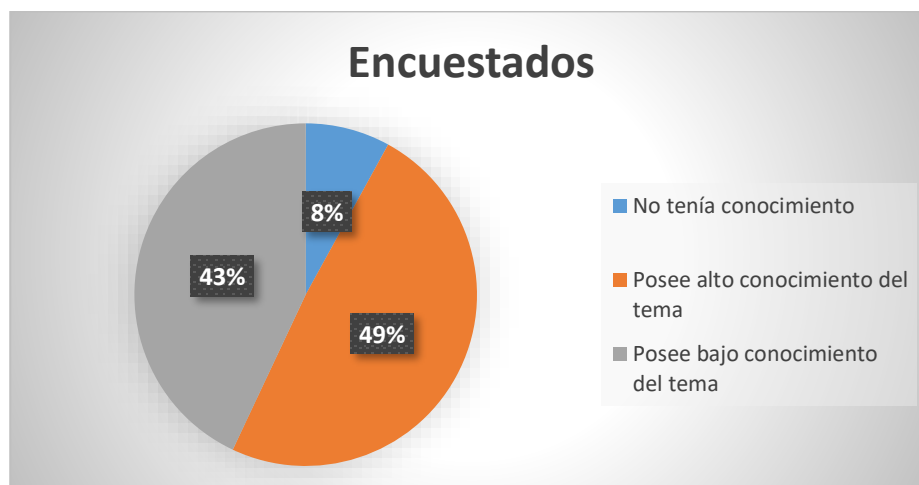


Figure 25. Sabía usted que aplicar **HÁBITOS SALUDABLES** cuando una persona esta contagiada de coronavirus (covid-19) disminuye el riesgo de afecciones graves incluso descartando hasta la muerte. Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

Respecto al gráfico (N°25) estadístico de los encuestados nos refleja que con un 43% poseen bajo conocimiento del tema junto con un 8 % que desconocen en su totalidad el cual mantiene una preocupación importante en el desarrollo de la tesis.

Pregunta 3.1. ¿Usted posee un Smartphone básico (Teléfono móvil con acceso a internet)?

Tabla 21. Posee Smartphone

Respuestas	Encuestados	Porcentaje
No poseo con internet, pero estoy en proceso de adquirir uno	9	2%
No poseo con internet, pero estoy en proceso de adquirir uno	16	4%
Si poseo uno de Gama Alta (costo > \$501)	64	17%
Si poseo uno de Gama Baja con internet (costo < \$200)	129	33%
Si poseo uno de Gama Media (costo \$201 a \$500)	169	44%
Total	387	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

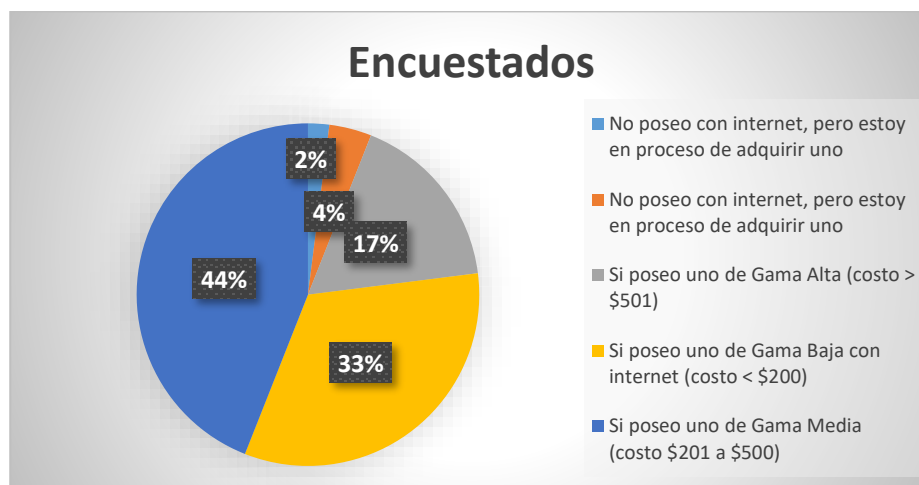


Figure 26. ¿Usted posee un Smartphone básico (Teléfono móvil con acceso a internet)??. Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

Respecto al gráfico estadístico (N°26) muestra un importante valor del 44% que si posee una gama media mientras le sigue con un 33% quienes poseen una gama baja con internet de menos de 200 dólares. A continuación, un 17% quienes si poseen una gama alta mayor a \$501 y el ultimo 6% No posee alguna de las 2 opciones.

Pregunta 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?

Tabla 22. *Sabe de aplicaciones móviles que permiten interactuar y mantener información*

Respuestas	Encuestados	Porcentaje
No tenía conocimiento	50	13%
Posee alto conocimiento del tema	154	40%
Posee bajo conocimiento del tema	183	47%
Total	387	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

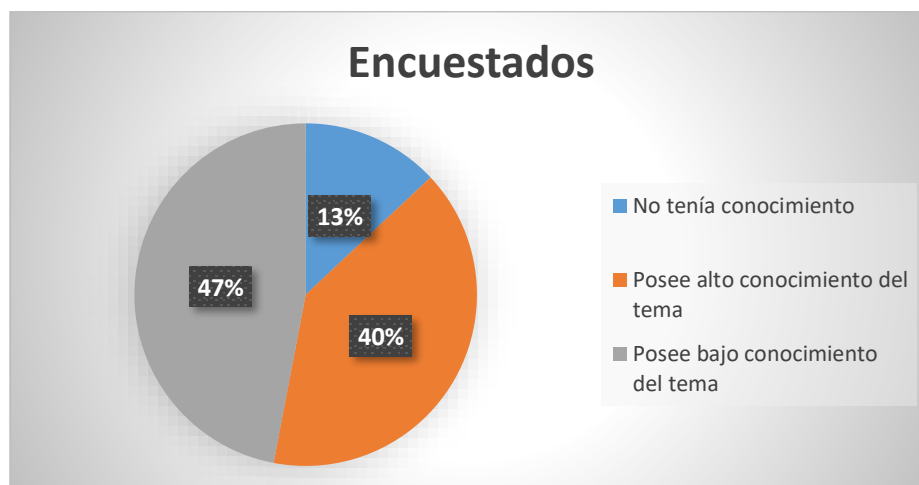


Figure 27. ¿Sabía usted que la tecnología de INTELIGENCIA ARTIFICIAL es capaz de desarrollar aplicaciones móviles que permita interactuar y mantener información de forma actualizada para combatir el coronavirus (Covid-19) en cualquier lugar del mundo, a cualquier hora, incluyendo fechas de feriado? Información obtenida de formulario de Google y de la investigación directa. *Elaborado por Piza Guale Alexandra*

Respecto al gráfico estadístico (N°27) el 47% de las personas encuestadas posee bajo conocimiento del tema mientras que el 40% posee alto conocimiento del tema.

Pregunta 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 SALUDABLE utilizando la tecnología de inteligencia artificial de forma GRATUITA?

Tabla 23. *Le gustaría contar con una aplicación móvil que mantenga actualizada la información*

Respuestas	Encuestados	Porcentaje
Ni de acuerdo ni en desacuerdo	33	18%
Parcialmente de acuerdo	68	37%
Parcialmente en desacuerdo	3	2%
Totalmente de acuerdo	77	42%
Totalmente en desacuerdo	2	1%
Total	183	100%

Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

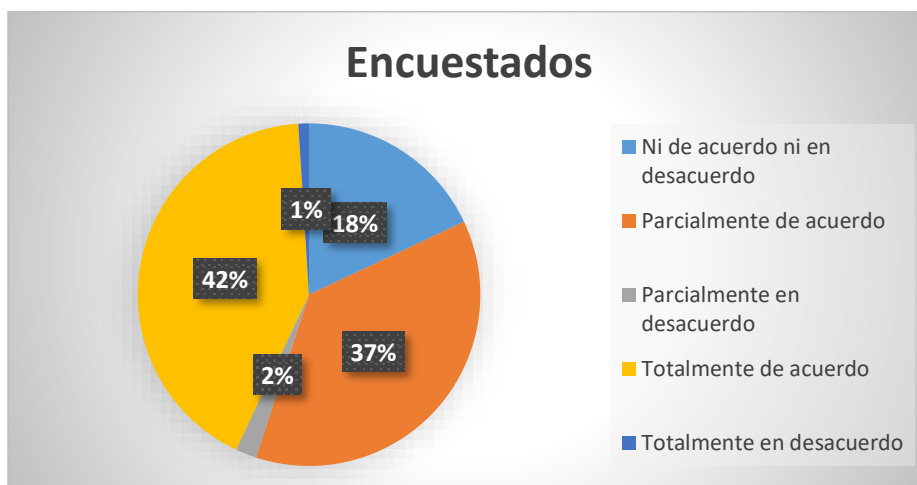


Figure 28. ¿Le gustaría contar con una aplicación móvil que le permita interactuar, mantener informado de forma actualizada para combatir al coronavirus (COVID-19) sobre los Hábitos Saludable utilizando la tecnología de inteligencia artificial de forma GRATUITA? Información obtenida de formulario de Google y de la investigación directa. Elaborado por Piza Guale Alexandra

Respecto al gráfico estadístico (N°28) la cantidad reflejada predomina con un 42% de las personas totalmente de acuerdo con la pregunta mientras que le sigue con un 37% parcialmente de acuerdo y la cantidad más baja que se encuentra totalmente en desacuerdo es del 1%.

3.5.3 Resumen de la entrevista

Se realizó una entrevista a tres especialistas de tecnología con conocimientos en IA, la ingeniera Liseth Jiménez Valencia, ellas indicaron que, desde su experiencia laboral el mundo de la tecnología no solo se debe aprender si no también practicarla y cada día aprender más de aquella sin limitar la diferencia de género, dando referencia que para el procesamiento de lenguaje natural desde su perspectiva los algoritmos más adecuados son, árbol de decisiones ya que en la actualidad hay librerías que contienen modelos matemáticos que hace mucho más fácil la utilización del mismo, aprendizaje semi-supervisado y redes neuronales las cuales abarcan gran información y están dirigidas a conversaciones textuales que se usan para el procesamiento de lenguaje natural. Asimismo, determino que en el ámbito de las técnicas NLP desde su punto de vista la segmentación por palabras y Tokenización son las técnicas que mejor resultados han brindado. Por lo tanto, ella opina que es importante realizar proyectos de análisis de técnicas de procesamiento de lenguaje natural NLP para clasificación de texto de conversaciones textuales para personas contagiadas con Covid-19, aclarando que en otros países ya existe tal proyecto, pero en Ecuador aún no es implementado por el motivo que aun los registros hospitalarios siguen siendo manuales la

cual es difícil recaudar información y poder realizar la clasificación de información, por otro lado, para implementar dicho proyecto las instalaciones hospitalarias deberían manejarse con información abierta que sería más útil para el algoritmo poderlos procesar.

La Srta. Karla Avilés Mendoza manifiesto que, es importante el uso de las tecnologías como la Inteligencia Artificial referente al Covid-19 la cual será de gran ayuda al departamento de salud, además indico que, desde su experiencia, las técnicas de procesamiento de lenguaje natural en manejo de información textual clasificada con mejores funcionamientos son, Stop Word, Reconocimiento de Entidades Nombradas (NER) y Tokenización. De la misma manera ella indica que es importante plasmar trabajos de investigación futuros ya que ayudara en el uso de los análisis de procesamiento de lenguaje natural en el procedimiento de clasificación conversacional textuales a personas infectadas con Covid-19.

La Srta. Nayelhi de Anda, determino que desde sus conocimientos el algoritmo más adecuado para el uso de arquitectura en NLP son las Redes Neuronales las cuales reflejan el comportamiento del cerebro humano que permiten que los programas informáticos exploren patrones y den soluciones a los problemas comunes en Inteligencia Artificial, machine Learning y Deep Learning. También indico que, desde sus conocimientos las técnicas de procesamiento de lenguaje natural que cuentan con mayor funcionalidad son, los Stop Word, Reconocimiento de Entidades Nombradas (NER) y Tokenización. Asimismo, expreso que está de acuerdo a que se empleen nuevos estudios para los análisis de técnicas NLP para la clasificación de textos conversacionales a personas contagiadas con Covid-19.

3.6 Construcción de las técnicas de Machine Learning

Para realizar el siguiente estudio de investigación se necesita de una Base de Datos, la cual cuenta con información asertiva, la cual otorgará el entrenamiento a las técnicas NLP que se haya escogido, para que de una u otra manera pueda mostrar o reflejar la esperada previsión. Se efectuó una encuesta la cual la base de datos, por lo tanto, adquirió mucha información la cual fue muy satisfactorio para poder realizar el entrenamiento de las técnicas NLP.

La encuesta fue empleada el pasado diciembre del 2021 que fue dirigida para los moradores de la zona 8 de la provincia del Guayas, la cual está distribuida por los cantones

de Guayaquil, Samborondón y Durán. Los datos se guardan en un archivo de Excel con la extensión .csv y se almacenan en el ordenador. Dado que funciona en Google Colab, cada vez que hay un período prolongado de inactividad, el conjunto de datos debe cargarse y la línea de código debe volver a ejecutarse.

3.6.1 Importación de datos

Se realiza la importación de la base de datos desde el ordenador hacia el portal de Google Colab, ya que será el repositorio por el cual se trabajará.

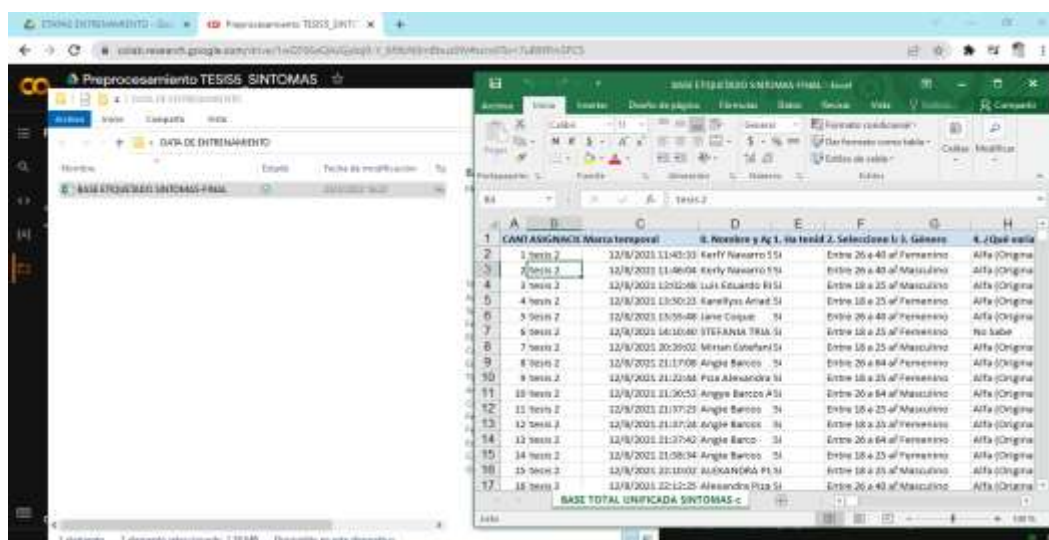


Figure 29. Información tomada de: Computador personal, Importación de datos. Elaborado por Piza Guale Alexandra

A continuación, se muestra la importación de librerías que son utilizadas para el tratamiento de datos, también se importa librerías las cuales se usan para el preprocesamiento de datos y técnicas NLP, asimismo, se importaron las librerías para crear el modelo LSTM (Long Short-Term Memory). Que son las librerías a utilizar en mi proyecto para el Procesamiento de Lenguaje Natural (PLN) de la Base de datos de Síntomas.

Como segundo punto, se importa la librería NLTK que será usada para el respectivo entrenamiento del Procesamiento del Lenguaje Natural, también se verifica que la versión que por defecto brinda el repositorio de Google Colab.

Como la versión 3.2.5 que proporciona Google Colab no contiene todos los paquetes necesarios, se realizó la respectiva desinstalación.

```
Found existing installation: nltk 3.2.5
Uninstalling nltk-3.2.5:
  Would remove:
    /usr/local/lib/python3.7/dist-packages/nltk-3.2.5.dist-info/*
    /usr/local/lib/python3.7/dist-packages/nltk/*
Proceed (y/n)? y
Successfully uninstalled nltk-3.2.5
```

Con la función **.info**, se realiza una inspección rápida de los tipos de datos que contiene la Base de Datos Síntomas.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4924 entries, 0 to 4923
Data columns (total 91 columns):
#   Column                                                                                               Non-Null Count  Dtype
---  -
0   CANTIDAD                                                       4924 non-null   int64
1   ASIGNACIÓN DE PREPROCESAMIENTO                                4140 non-null   object
2   Marca temporal                                                  4293 non-null   object
3   0. Nombre y Apellido del encuestador (Persona que te ha pedido que llene la encuesta)  4140 non-null   object
4   1. Ha tenido coronavirus?                                       4323 non-null   object
5   2. Seleccione la Edad                                           4282 non-null   object
6   3. Género                                                       4722 non-null   object
7   4. ¿Qué variante del Virus lo contagió?                        4139 non-null   object
8   5. ¿En qué fecha se contagió? MM-DD-YYYY                      4077 non-null   object
9   6. ¿Nivel de intensidad que tuvo los síntomas?                4137 non-null   object
10  7. ¿En qué lugar o evento considera que se contagió?          3958 non-null   object
11  8. ¿En caso de haber estado vacunado al momento de contagiarse cuantas dosis tenía aplicadas al contagiarse?  3961 non-null   object
12  9. ¿En caso de haber estado vacunado al momento de contagiarse qué vacuna recibió?      3955 non-null   object
13  10. Describe lo más detallado ¿Qué síntomas ha tenido?-SIN DEPURAR  4886 non-null   object
14  10. Describe lo más detallado ¿Qué síntomas ha tenido?-DEPURADO  4884 non-null   object
15  Afectación psicológica                                          4924 non-null   int64
16  Alergia                                                         4924 non-null   int64
17  Alucinación                                                     4924 non-null   int64
18  Amigdalitis                                                     4924 non-null   int64
19  Ansiedad                                                        4924 non-null   int64
20  Apatía                                                          4924 non-null   int64
21  Arritmia                                                        4924 non-null   int64
22  Asintomatismo                                                   4924 non-null   int64
23  Bronquitis                                                      4924 non-null   int64
24  Cambio coloración de piel                                       4924 non-null   int64
25  Cansancio                                                       4924 non-null   int64
26  Colitis                                                         4924 non-null   int64
```

Figure 33. Información tomada de: entorno de desarrollo integrada Google Colab, Inspección de los datos que contiene de la Base de Datos Síntomas. Elaborado por Piza Guale Alexandra

A continuación, se realizó una exploración rápida de datos numéricos que tiene la Base de Datos Síntomas.

centro	afectación psíquica	alergia	alucinación	amigdalitis	ansiedad	apatía	arritmia	asintomatismo	bronquitis	...	evento	mareado	retención de líquidos	resaca	seguimiento	toque al respirar	hidratación excesiva	temperatura	tos	vómito	hormona	alimento
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	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	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
...	
4819	4820	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4820	4821	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4821	4822	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4822	4823	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4823	4824	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figure 34. Información tomada de: entorno de desarrollo integrada Google Colab, Exploración de datos numéricos de Base de Datos Síntomas. Elaborado por Piza Guale Alexandra

Como se visualiza en la siguiente imagen, refleja que existen 77 columnas numéricas y 14 columnas categóricas. Que al sumarmas muestran las 91 columnas que hay en la Base de Datos Síntomas.

77 14

Figure 35. Información tomada de: entorno de desarrollo integrada Google Colab, Visualización de columnas numéricas y columnas categóricas. Elaborado por Piza Guale Alexandra

Se utiliza el comando `describe()` para realizar una inspección rápida de estadística descriptiva, medias de tendencia central como lo son: media (mean), desviación estándar (std), mínima (min), máxima.

Donde indica que, el máximo es 1 y el mínimo es 0 siendo así la representación de una base de datos integral.

	CANTIDAD	Afectación psicológica	Alergia	Alucinación	Amigdalitis	Anisidul	Apetía	Arritmia	Asintomatismo	Bronquitis	...	Prurito	Resfriado	Retención de líquidos
count	4524.000000	4524.000000	4524.000000	4524.000000	4524.000000	4524.000000	4524.000000	4524.000000	4524.000000	4524.000000	...	4524.000000	4524.000000	4524.000000
mean	2462.500000	0.005135	0.009139	0.000406	0.004468	0.015028	0.000609	0.000406	0.05788	0.006782	...	0.001628	0.006499	0.004671
std	1421.500093	0.005160	0.005169	0.020152	0.006700	0.121678	0.024678	0.020152	0.23354	0.001506	...	0.042716	0.000361	0.008150
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
25%	1231.750000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
50%	2462.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
75%	3693.250000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
max	4524.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000

0 rows × 17 columns

Figure 36. Información tomada de: entorno de desarrollo integrada Google Colab, Inspección rápida de estadística descriptiva. Elaborado por Piza Guale Alexandra

Se inspecciona los datos que contiene la columna 10. Describa lo más detallado ¿Qué síntomas ha tenido? -SIN DEPURAR, de la Base de Datos Síntomas.

Una vez, que se haya realizado la verificación se realiza una copia de la Base de Datos Síntomas para no dañar la base de datos central.

```
0
1
2
3
4
...
4919
4920
4921
4922
4923
Name: 10. Describa lo más detallado ¿Qué síntomas ha tenido?-SIN DEPURAR, Length: 4924, dtype: object
```

Figure 37. Información tomada de: entorno de desarrollo integrada Google Colab, Inspección de datos de la columna 10. Describa lo más detallado ¿Qué síntomas ha tenido?- SIN DEPURAR y realización de una copia de la Base de Datos Síntoma. Elaborado por Piza Guale Alexandra

Se visualiza la cantidad de datos que almacena la columna 10. Describa lo más detallado ¿Qué síntomas ha tenido?- SIN DEPURAR, hay 37 filas con 91 columnas.

(38, 91)

Figure 38. Información tomada de: entorno de desarrollo integrada Google Colab, Visualización de la cantidad de datos. Elaborado por Piza Guale Alexandra

Se realiza una búsqueda de celdas con valores NAN o NULOS, al momento de imprimir los valores perdidos con lo que cuenta la Base de Datos Síntomas, los muestra como valores booleanos, es decir, con verdadero y falso. Se realiza la suma de todos los valores verdaderos, que indican las columnas que contienen datos perdidos en la Base de Datos Síntomas. En donde, 14 es el total de valores perdidos.

14

Figure 39. Información tomada de: entorno de desarrollo integrada Google Colab, Suma de valores perdidos de la Base de Datos Síntomas. Elaborado por Piza Guale Alexandra

Se verifica las columnas de datos perdidos que contiene la base de datos Síntomas a través de un representativo gráfico de un mapa de calor.

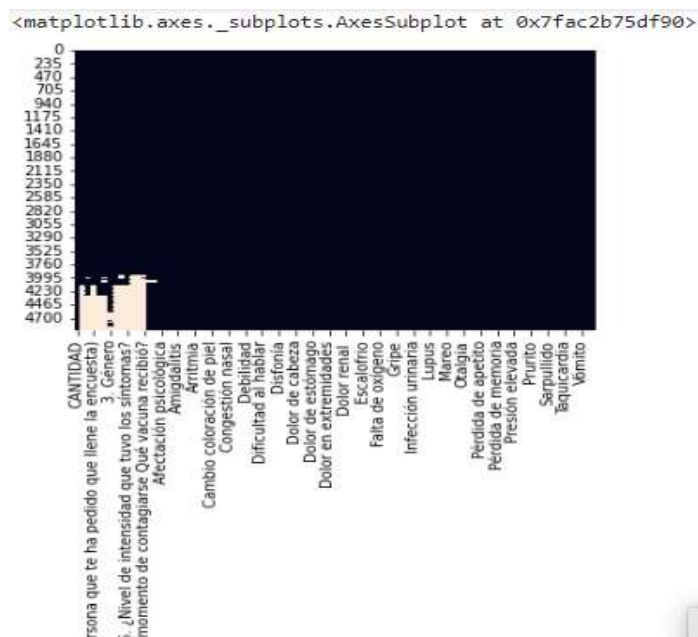


Figure 40. Información tomada de: entorno de desarrollo integrada Google Colab, Visualización de datos perdidos por medio de gráfico. Elaborado por Piza Guale Alexandra

3.6.2 Tratamiento de datos

Posteriormente se aplica el método **def**, para realizar un método personalizado el cual permitirá el reemplazo de las palabras con tilde a palabras sin tilde, también, excluirá cualquier tipo de caracteres que estén en la columna 10. Describa lo más detallado ¿Qué síntomas ha tenido? -SIN DEPURAR.

Una vez que se realice la ejecución se crea una columna llamada Sintomas—PROCESADA, donde reflejará el cambio que se aplicó a través del método personalizado.

[illegible]

Figure 41. Información tomada de: entorno de desarrollo integrada Google Colab, Método personalizado y creación de una nueva columna. Elaborado por Piza Guale Alexandra

Después de realizar la ejecución de la limpieza de datos en la columna, se procede a hacer una exploración rápida de palabras vacías que hayan quedado durante el

proceso. Luego, se refleja la comparación de los datos entre la columna 10. Describa lo más detallado ¿Qué síntomas ha tenido? –SIN DEPURAR y la columna Sintomas-PROCESADA, en la cual, se puede apreciar el cambio de las palabras mayúsculas a minúsculas, el reemplazo de las palabras con tildes a palabras sin tildes, la extracción de caracteres y exclusión de números.

	10. Describa lo más detallado ¿Qué síntomas ha tenido?-SIN DEPURAR	Sintomas_PROCESADA
0	Perdida de gusto y olfato	perdida de gusto y olfato
1	Dolor leve de espalda y fiebre	dolor leve de espalda y fiebre
2	Dolor al momento de respirar por algunos días	dolor al momento de respirar por algunos dias
3	Dolor de cabeza tos seca y mucha fiebre no ten...	dolor de cabeza tos seca y mucha fiebre no ten...
4	Sin sabor	sin sabor
...
4919	La coagulación de la sangre, la inflamación o ...	la coagulation de la sangre la inflamacion o l...
4920	El COVID-19 ataca también a la piel: alertan d...	el covid ataca tambien a la piel alertan de nu...
4921	Los síntomas de esta enfermedad son como los d...	los sintomas de esta enfermedad son como los d...
4922	quedan nuevos síntomas del Covid-19: pérdida d...	quedan nuevos sintomas del covid perdida del o...
4923	rueda de prensa en el Hospital Ramón de Lara s...	rueda de prensa en el hospital ramon de lara s...

4924 rows x 2 columns

Figure 42. Información tomada de: entorno de desarrollo integrada Google Colab, Visualización de la columna 10. Describa lo más detallado ¿Qué síntomas ha tenido? -SIN DEPURAR y columna Sintomas –PROCESADA. Elaborado por Piza Guale Alexandra

Se realizó nuevamente una copia de la base de datos anterior, para realizar la separación de variables como lo es X y Y.

```

<class 'pandas.core.frame.DataFrame'>
Int64Index: 4919 entries, 0 to 4923
Data columns (total 93 columns):
 #   Column                                                                                               Non-Null Count  Dtype
---  -
0   CANTIDAD                                                       4919 non-null   int64
1   ASIGNACION DE PREPROCESAMIENTO                               4139 non-null   object
2   Marca temporal                                                 4291 non-null   object
3   0. Nombre y Apellido del encuestador (Persona que te ha pedido que llene la encuesta)  4139 non-null   object
4   1. Ha tenido coronavirus?                                       4321 non-null   object
5   2. Seleccione la edad                                           4260 non-null   object
6   3. Género                                                       4717 non-null   object
7   4. ¿Qué variante del Virus lo contagio?                       4158 non-null   object
8   5. ¿En qué fecha se contagio? MM-DD-YYYY                     4076 non-null   object
9   6. ¿Nivel de intensidad que tuvo los síntomas?               4136 non-null   object
10  7. ¿En qué lugar o evento considera que se contagio?         3957 non-null   object
11  8. ¿En caso de haber estado vacunado al momento de contagiarse cuentas dosis tenía aplicadas al contagiarse?  3960 non-null   object
12  9. ¿En caso de haber estado vacunado al momento de contagiarse Qué vacuna recibió?       3954 non-null   object
13  10. Describa lo más detallado ¿Qué síntomas ha tenido?-SIN DEPURAR  4881 non-null   object
14  10. Describa lo más detallado ¿Qué síntomas ha tenido?-DEPURADO  4879 non-null   object
15  Afectación psicológica                                         4919 non-null   int64
16  Alergia                                                         4919 non-null   int64
17  Alucinación                                                    4919 non-null   int64
18  Amigdalitis                                                    4919 non-null   int64
19  Ansiedad                                                       4919 non-null   int64

```

Figure 43. Información tomada de: entorno de desarrollo integrada Google Colab, tipo de datos que almacena la Nueva copia de base de datos. Elaborado por Piza Guale Alexandra

Al realizar la separación de variables X e Y. Se aplicó el método **iloc**, el cual indica, las posiciones de filas y columnas que se va a establecer.

En la variable Y, se guardó los síntomas más relevantes que indicaron las personas por medio de una encuesta. Después, se establece la suma todos los valores que se

entrenamiento. Sin embargo, el código utilizado para definir la técnica que se manejó es el siguiente, Técnica Tokenización con SpaCy, Stop Word y Lematización con Spacy, como se puede observar en la siguiente imagen, se importa las librerías **nltk**. Las cuales ayudan a realizar el procedimiento de Tokenización por medio de código, también se importa la función stopwords, por otro lado, se importa la función de Lematización SpaCy, donde, en cada técnica utilizada se crea una variable que guardara la información de la clase utilizada por el lenguaje español.

Así mismo, se efectúa una exploración rápida de palabras frecuentes en el idioma español y poderlas excluir de la información almacenada en la Base de Datos Síntomas.

Se importó la librería DataFrame, luego se realiza una copia de la variable que almacena la nueva columna procesada. Asimismo, se efectúa una exploración rápida de la información que contiene la nueva variable.

	Síntomas_PROCESADA	Tokenizado	SinStopwords	Lematizacion
0	perdida de gusto y olfato	[perdida, de, gusto, y, olfato]	perdida gusto olfato	perdido gustar olfato
1	dolor leve de espalda y fiebre	[dolor, leve, de, espalda, y, fiebre]	dolor leve espalda fiebre	dolor levar espalda fiebre
2	dolor al momento de respirar por algunos días	[dolor, al, momento, de, respirar, por, alguno...	dolor momento respirar dias	dolor momento respirar dias
3	dolor de cabeza tos seca y mucha fiebre no ten	[dolor, de, cabeza, tos, seca, y, mucha, febr...	dolor cabeza tos seca mucha fiebre tenia gusto	dolor cabeza tos seco muchu fiebre tenia gusta
4	sin sabor	[sin, sabor]	sabor	saber
...				
4918	yo diciendo que soy asintomatica y tengo un chi...	[yo, diciendo, que, soy, asintomatica, y, teng...	diciendo asintomatica chingo sintomas atipicos...	decir asintomatica chingar sintomas atipicos c...
4919	la coagulación de la sangre la inflamación o l...	[la, coagulación, de, la, sangre, la, inflamac...	coagulation sangre inflamacion cambios colorac...	coagulation sangre inflamacion cambio colorac...
4920	el covid ataca también a la piel alertan de nu...	[el, covid, ataca, también, a, la, piel, alert...	covid ataca tambien piel alertan nuevos simbro...	covid atacar tambien piel alertar nuevos sintem...
4921	los síntomas de esta enfermedad son como los d...	[los, sintomas, de, esta, enfermedad, son, com...	sintomas enfermedad infeccion decir asociados...	sintomas enfermedad infeccion decir asociar fi...
4922	quedan nuevos síntomas del covid perdida del o...	[quedan, nuevos, sintomas, del, covid, perdida...	quedan nuevos sintomas covid perdida olfato gu...	quedar nuevo sintomas covid perdido olfato gas...

4903 rows = 4 columns

Figure 47. Información tomada de: entorno de desarrollo integrada Google Colab, elección de combinaciones de técnicas. Elaborado por Piza Guale Alexandra

Una vez que se haya realizado el procedimiento de la primera combinación de técnicas, se procede a realizar la separación de variables **Train – test**, el cual, ayuda a medir la precisión que se realiza en el modelo LSTM. En donde, el conjunto de datos se divide en X y Y, donde el 80% calcula el entrenamiento y el 20% es para la medición de pruebas. También se realiza una exploración de los datos que guarda la variable x_train.

	Síntomas_PROCESADO	Tokenizado	Síntopwords	Lesatización
279	fiebre y dolor de cabeza	[fiebre, y, dolor, de, cabeza]	fiebre dolor cabeza	fiebre dolor cabeza
3925	malestar general	[malestar, general]	malestar general	malestar general
4086	estufa con fiebrícula fatiga y dolor de cabeza	[estufa, con, fiebrícula, fatiga, y, dolor, de, ...]	fiebrícula fatiga dolor cabeza	fiebrícula fatiga dolor cabeza
4031	me agito rapido y me duelen los huesos del cor...	[me, agito, rapido, y, me, duelen, los, huesos, ...]	agito rapido duelen huesos corazon recien vino...	agito rapido dolor hueso corazon recien venir...
1467	decaimiento y falta de aire	[decaimiento, y, falta, de, aire]	decaimiento falta aire	decaimiento falta aire
...
2172	dolor muscular fiebre perdida del olfato y si...	[dolor, muscular, fiebre, perdida, del, olfato, ...]	dolor muscular fiebre perdida olfato gusto	dolor muscular fiebre perdido olfato gustar
1430	dolores musculares y de cabeza fiebre leve	[dolores, musculares, y, de, cabeza, fiebre, l...]	dolores musculares cabeza fiebre leve	dolor muscular cabeza fiebre llevar
3701	sin sintomas	[sin, sintomas]	sintomas	sintomas
3174	perdida de gusto dolores de cabeza y cuerpo	[perdida, de, gusto, dolores, de, cabeza, y, c...]	perdida gusto dolores cabeza cuerpo	perdido gustar dolor cabeza cuerpo
2275	dolor de cabeza fiebre y malestar	[dolor, de, cabeza, fiebre, y, malestar]	dolor cabeza fiebre malestar	dolor cabeza fiebre malestar

3682 rows x 4 columns

Figure 48. Información tomada de: entorno de desarrollo integrada Google Colab. Exploración de datos en la variable x_train. Elaborado por Piza Guale Alexandra

Además, se elabora una indagación de la información recopilada en la variable y_train.

	Asintomatismo	Conmoción	Conmoción	Conmoción	Conmoción	Congestión nasal	Debilidad	Diarrea	Dificultad para Respirar	Dolor articular	...	Fiebre	Gripe	Malestar general	Pérdida de apetito	Pérdida de gusto	Pérdida de olfato
279	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0
3925	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0
4086	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0
4031	0	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0
1467	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0
...
2172	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	1
1430	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0
3701	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
3174	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	1
2275	0	0	0	0	0	0	0	0	0	0	...	1	0	1	0	0	0

3682 rows x 19 columns

Figure 49. Información tomada de: entorno de desarrollo integrada Google Colab, Indagación de información en variable y_train. Elaborado por Piza Guale Alexandra

3.6.4. Elección del modelo

Se escogió el modelo LSTM (Long Short-Term Memory), en donde, se realizó la comprobación de las técnicas antes mencionadas y poder descubrir que técnica es

más eficiente para este tipo de entrenamiento. Asimismo, se especifica el código que se trató para definir la técnica a utilizar.

Cuando se trabaja en Red Neuronal se escoge el modelo, en este caso, se escogió el modelo secuencial, indicando cada uno de sus indicadores con límites, se realiza la activación Softmax de la función logística matemática.

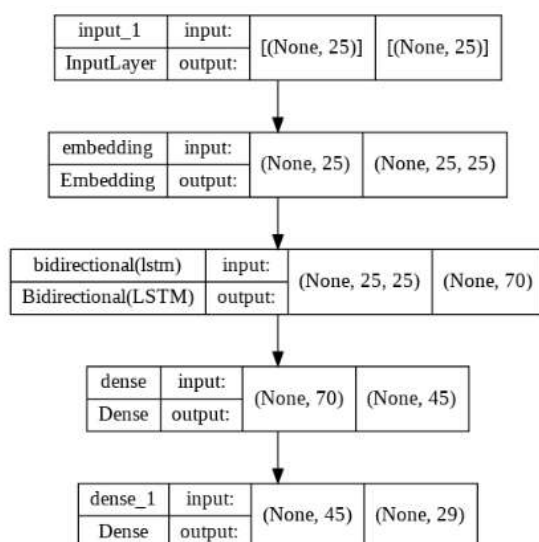


Figure 50. Información tomada de: entorno de desarrollo integrada Google Colab, Elección del tipo de modelo y activación de función. Elaborado por Piza Guale Alexandra

3.6.5. Entrenamiento del algoritmo

Se utiliza el método unique que permite encontrar elementos únicos, por otro lado, se estableció las entradas, salidas y cantidades de iteraciones de aprendizaje y prueba, donde, se puede manipular la cantidad de salidas que mejor resultados nos brinden.

```

Epoch 122/150
369/369 [=====] - 7s 20ms/step - loss: 0.0059 - accuracy: 0.4652 - val_loss: 0.2181 - val_accuracy: 0.4220
Epoch 123/150
369/369 [=====] - 7s 20ms/step - loss: 0.0051 - accuracy: 0.4718 - val_loss: 0.2189 - val_accuracy: 0.4050
Epoch 124/150
369/369 [=====] - 7s 20ms/step - loss: 0.0060 - accuracy: 0.4676 - val_loss: 0.2196 - val_accuracy: 0.4476
Epoch 125/150
369/369 [=====] - 7s 20ms/step - loss: 0.0068 - accuracy: 0.4518 - val_loss: 0.2253 - val_accuracy: 0.4410
Epoch 126/150
369/369 [=====] - 7s 20ms/step - loss: 0.0074 - accuracy: 0.4882 - val_loss: 0.2234 - val_accuracy: 0.4643
Epoch 127/150
369/369 [=====] - 7s 20ms/step - loss: 0.0060 - accuracy: 0.4744 - val_loss: 0.2102 - val_accuracy: 0.4208
Epoch 128/150
369/369 [=====] - 7s 20ms/step - loss: 0.0064 - accuracy: 0.4889 - val_loss: 0.2270 - val_accuracy: 0.4084
Epoch 129/150
369/369 [=====] - 7s 20ms/step - loss: 0.0060 - accuracy: 0.4672 - val_loss: 0.2219 - val_accuracy: 0.4315
Epoch 130/150
369/369 [=====] - 8s 21ms/step - loss: 0.0058 - accuracy: 0.4584 - val_loss: 0.2215 - val_accuracy: 0.4437
Epoch 131/150
369/369 [=====] - 7s 20ms/step - loss: 0.0058 - accuracy: 0.4662 - val_loss: 0.2258 - val_accuracy: 0.4206
Epoch 132/150
369/369 [=====] - 7s 20ms/step - loss: 0.0057 - accuracy: 0.4689 - val_loss: 0.2251 - val_accuracy: 0.4491
Epoch 133/150
369/369 [=====] - 7s 20ms/step - loss: 0.0058 - accuracy: 0.4723 - val_loss: 0.2249 - val_accuracy: 0.4491
Epoch 134/150
369/369 [=====] - 7s 20ms/step - loss: 0.0057 - accuracy: 0.4708 - val_loss: 0.2312 - val_accuracy: 0.4303
Epoch 135/150
369/369 [=====] - 7s 20ms/step - loss: 0.0058 - accuracy: 0.4808 - val_loss: 0.2282 - val_accuracy: 0.4261
Epoch 136/150
369/369 [=====] - 7s 20ms/step - loss: 0.0057 - accuracy: 0.4754 - val_loss: 0.2302 - val_accuracy: 0.4228
Epoch 137/150
369/369 [=====] - 7s 20ms/step - loss: 0.0057 - accuracy: 0.4679 - val_loss: 0.2250 - val_accuracy: 0.4138
Epoch 138/150

```

Figure 51. Información tomada de: entorno de desarrollo integrada Google Colab, Utilización de métodos y observación del entrenamiento. Elaborado por Piza Guale Alexandra

Para verificar el conjunto de datos de aprendizaje y el conjunto de datos de pruebas. Donde, se realizó una prueba de 100 epochs, en donde, visualizamos que el 60% es de aprendizaje y un 15% de prueba.

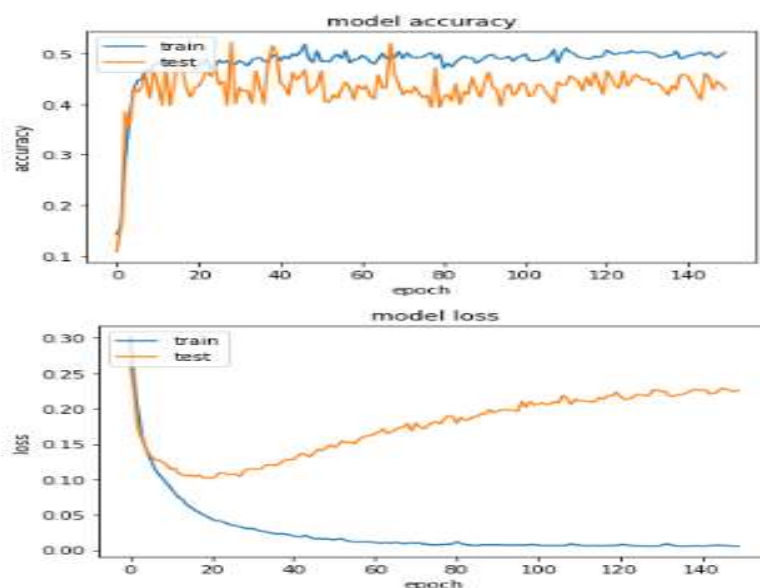


Figure 52. Información tomada de: entorno de desarrollo integrada Google Colab, Visualización de gráfico para aprendizaje de datos y de prueba. Elaborado por Piza Guale Alexandra

3.6.6. Evaluación del algoritmo

Se realiza la respectiva evaluación de datos de x_{train} , y_{train} y se realiza la respectiva exploración métrica de la evaluación establecida. Se observa que el índice de test es de 3, en donde, se procede a explorar la variable $X_{test}[\text{índice_test}]$ y se observa los resultados que refleja el arreglo.

```
array([[8.8409515e-04, 2.6516199e-02, 2.8269034e-02, 2.7707843e-02,
        2.6102390e-02, 8.6616800e-04, 3.7569203e-07, 1.0236036e-09,
        8.3715822e-06, 1.1653713e-03, 3.1741329e-02, 1.6066937e-06,
        1.2093561e-05, 3.9601600e-05, 6.6068619e-02, 3.4703952e-01,
        4.1068117e-07, 6.7630375e-04, 2.0997017e-05, 8.2760985e-04,
        5.7423435e-04, 4.3884838e-01, 1.3297038e-06, 6.9101688e-08,
        1.1113081e-04, 1.9011287e-03, 9.5369320e-12, 6.0053024e-04,
        1.5154943e-05]], dtype=float32)
```

Figure 53. Información tomada de: entorno de desarrollo integrada Google Colab, Evaluación métrica. Elaborado por Piza Guale Alexandra

También, se utiliza las predicciones de la variable `X_test[índice_test]`.

```
Asintomatismo      0
Cansancio          0
Cansancio          0
Cansancio          0
Cansancio          0
Congestión nasal   0
Debilidad          0
Diarrea           0
Dificultad para Respirar 0
Dolor articular    0
Dolor de cabeza    0
Dolor de espalda   0
Dolor de estómago  0
Dolor de garganta  0
Dolor de huesos    0
Dolor muscular     1
Escalofrío        0
Falta de oxígeno  0
Fatiga            0
Fiebre            0
Gripe             0
Malestar general  0
Nariz            0
Pérdida de apetito 0
Pérdida de Gusto   0
Pérdida de Olfato  0
Presión en el pecho 0
Tos              0
Vómito           0
Name: 1645, dtype: int64
```

Figure 54. Información tomada de: entorno de desarrollo integrada Google Colab, Exploración de predicciones y cantidad de salidas. Elaborado por Piza Guale Alexandra

Y para finalizar, consultamos cuantas redes neuronales de salidas tendremos.

29

Figure 55. Información tomada de: entorno de desarrollo integrada Google Colab, Cantidad de variables de salida. Elaborado por Piza Guale Alexandra

Conclusiones y Recomendaciones

Conclusiones

- Se realizó la revisión bibliográfica de información de documento recientes de los últimos 5 años relacionados con las Técnicas básicas de Procesamiento del Lenguaje Natural, terminología de Covid-19, algoritmo de ML en un total de 59 referencias bibliográficas.
- Para obtener una información más relevante de la base de datos depurada que se asignó para este caso de estudio, se realizó un preprocesamiento, el cual, consistió en eliminar caracteres, números, signos de puntuación, entre otros.

- Se ejecutó el modelo de red neuronal LSTM, con la finalidad de entrenar varias combinaciones de técnicas y elegir las más relevantes para este caso de estudio
- En el presente estudio se cumplió con el objetivo general de construir un módulo de técnicas, por lo que, se realizó varias pruebas y se identificó que la combinatoria de técnicas más adecuada fue Tokenización NLTK, Stop Word y Lematización SpaCy.

Recomendaciones

- Para futuras investigaciones se analice el uso de nuevas técnicas NPL, para el desarrollo de entrenamientos en otros campos que requieran una aplicación de Machine Learning. Se recomiendan en las ferias tecnológicas de la Facultad Ingeniería Industrial impulsar sesión de workshop relacionadas con NLP. Para estimular el interés de los estudiantes en el campo de la inteligencia artificial.
- Se sugiere impulsar nuevas investigaciones en el campo de NLP, el desarrollo de modelos de procesamiento del lenguaje natural.
- Que se desarrolle un modelo de Procesamiento de Lenguaje Natural que se encuentren en las líneas o sub-líneas de asistente virtual, líneas de traducción de Lenguajes u otros frentes de aplicación de NLP.

ANEXOS

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 I: 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, I 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

Llámesse 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

Bibliografía

- Acharya, D. P. (05 de Noviembre de 2021). Los 11 mejores IDE de Python para potenciar el desarrollo y la depuración. *Geekflare*. Obtenido de <https://geekflare.com/es/best-python-ide/>
- Agencia B12. (07 de Mayo de 2021). *agenciab 12*. Obtenido de agenciab 12: <https://agenciab12.com/noticia/tipos-aprendizaje-automatico-existen>
- Agencia B12. (07 de Mayo de 2021). *agenciab 12*. Obtenido de agenciab 12: <https://agenciab12.com/noticia/tipos-aprendizaje-automatico-existen>
- Akker, A. V. (21 de Diciembre de 2021). Top 3 de los lenguajes de programación en 2021. *Exploradata*. Obtenido de <https://www.exploradata.com/top-3-de-los-lenguajes-de-programacion-en-2021/>
- Antonio, G. B. (2017). *Library*. Obtenido de Library: <https://1library.co/document/q76rm7dy-debido-proceso-recursividad-multas-impuestas-director-regional-trabajo.html>
- Aprende Machine Learning. (27 de Diciembre de 2018). *Aprende Machine Learning*. Obtenido de Aprende Machine Learning: <https://www.aprendemachinelearning.com/procesamiento-del-lenguaje-natural-nlp/>
- ARABA 4.0. (18 de Diciembre de 2020). Los lenguajes de programación más usados en Big Data e Inteligencia Artificial. *ARABA 4.0*. Obtenido de <https://araba40.eus/lenguajes-de-programacion-big-data-inteligencia-artificial/>
- Arias, E. R. (05 de Diciembre de 2020). Tipos de investigación. *Economipedia haciendo fácil la economía*. Obtenido de <https://economipedia.com/definiciones/tipos-de-investigacion.html>
- Ávila, C. E. (24 de Abril de 2019). Introducción a los tipos de muestreo. *ALERTA Revista científica del Instituto Nacional de Salud*. Obtenido de <https://alerta.salud.gob.sv/introduccion-a-los-tipos-de-muestreo/>
- Baume, G. L. (2021). *Breve introducción a Google Colab*. Obtenido de <http://fcaglp.unlp.edu.ar/~gbaume/grupo/Publicaciones/Apuntes/GoogleColab.pdf>
- Bonilla, G. J. (25 de Mayo de 2020). *Las dos caras de la educación en el Covid-19*. Obtenido de Las dos caras de la educación en el Covid-19: <http://cienciamerica.uti.edu.ec/openjournal/index.php/uti/article/view/294/462>
- Brownlee, J. (9 de 10 de 2017). *Machine Learning Mastery*. Obtenido de Machine Learning Mastery: <https://machinelearningmastery.com/gentle-introduction-bag-words-model/>
- Cajal Alberto. (16 de Mayo de 2016). *lifeder*. Obtenido de lifeder: <https://www.lifeder.com/tecnicas-de-pnl/>
- Campos, C. (Febrero de 2020). Fases del proceso de investigación científica y elementos de la investigación cuantitativa y cualitativa. *SCRIBD*, 11. Obtenido de <https://www.scribd.com/document/447304281/Actividad-N-02-Fases-del-proyecto-de-Investigacion-Cientifica-inv-cualitativa-y-cuantitativa>

- Celi-Parraga, R., Varela-Tapia, E., & Montaña-Pulzara, I. A.-G. (05 de Noviembre de 2021). Técnicas de procesamiento de lenguaje natural en la inteligencia artificial conversacional textual. *Alpha publicaciones*, 9. Obtenido de <https://alfapublicaciones.com/index.php/alfapublicaciones/article/view/123/373>
- Chen, P. -H. (23 de 09 de 2019). Elementos esenciales del procesamiento del Lenguaje natural. *sci hub*, 2. Obtenido de <https://sci-hub.st/https://www.sciencedirect.com/science/article/abs/pii/S1076633219304179>
- Comercio, E. (12 de marzo de 2021). ¿Qué pasaba en Ecuador el 12 de marzo del 2020? ¿Qué pasaba en Ecuador el 12 de marzo del 2020?, pág. 1. Obtenido de <https://www.elcomercio.com/actualidad/ecuador/ecuador-pandemia-covid-emergencia-sanitaria.html>
- Coronel y Perez. (23 de Abril de 2020). *Coronel y Perez*. Obtenido de Coronel y Perez: <https://www.coronelyperez.com/2020/04/23/la-crisis-ocasionada-por-el-covid-19-y-sus-implicaciones-legales-en-el-ecuador/>
- data, S. b. (24 de 08 de 2018). *Sitio big data*. Obtenido de Sitio big data: <https://sitiobigdata.com/2018/08/24/mapeo-de-incrustaciones-de-word-con-word2vec/#>
- Delgado Paulette. (12 de Febrero de 2021). *Tecnológico de Monterrey*. Obtenido de Tecnológico de Monterrey: <https://observatorio.tec.mx/edu-news/programacion-neurolinguistica-aprendizaje>
- Desarrollo web. (28 de 02 de 2019). *Digital Guide IONOS*. Obtenido de Digital Guide IONOS: <https://www.ionos.es/digitalguide/paginas-web/desarrollo-web/jupyter-notebook/>
- Deva. (11 de Julio de 2021). *Analytics Vidhya*. Obtenido de Analytics Vidhya: <https://medium.com/analytics-vidhya/evolution-of-natural-language-processing-nlp-ac941b6523e9>
- Deyimar, A. (26 de Junio de 2020). *HOSTINGER TUTORIALES*. Obtenido de HOSTINGER TUTORIALES: https://www.hostinger.es/tutoriales/que-es-php?ppc_campaign=google_search_generic_hosting_all&bidkw=defaultkeyword&lo=9077222&gclid=Cj0KCQiAip-PBhDVARIsAPP2xc0kt6pMNGs1KvdB0vm7b41f5cwHvcCNU23ODgZRugPsUjc6xqlflaAkFSEALw_wcB
- Díaz, M. J. (2020). Inteligencia artificial y Big Data como soluciones frente a la COVID-19. *SCIELO*. Obtenido de https://scielo.isciii.es/scielo.php?pid=S1886-58872020000300019&script=sci_arttext&tlng=en
- Emilio. (24 de 06 de 2020). Intriduccion a spaCY. *Todo BI*. Obtenido de <https://todobi.com/introduccion-a-spacy/>
- Equipo Expert.ai. (22 de Agosto de 2017). *Expert.ai*. Obtenido de Expert.ai: <https://www.expert.ai/blog/natural-language-processing-algorithms/>
- Fernández, A. (29 de Marzo de 2019). Inteligencia artificial en los servicios financieros. *BOLETÍN ECONÓMICO*, 3. Obtenido de <https://core.ac.uk/download/pdf/322617455.pdf>

- Hagarty, R., & Karlsen, E. (03 de Septiembre de 2019). Intruducción a IBM Watson Studio. *IBM Developer*. Obtenido de <https://developer.ibm.com/es/articles/introduction-watson-studio/>
- Haleem, R. V. (2020, August). Artificial Intelligence (AI) applications for COVID-19 pandemic. *ScienceDirect*, 339. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1871402120300771>
- Infobae. (24 de Noviembre de 2021). *Infobae*. Obtenido de Infoabe: <https://www.infobae.com/america/peru/2021/11/24/que-es-una-encuesta-para-que-sirve-y-como-se-elabora-brainly-preguntas-y-respuestas-aprendo-en-casa-tareas-resueltas/>
- ITELLIGENT. (26 de Julio de 2017). *ITELLIGENT*. Obtenido de ITELLIGENT: <https://itelligent.es/es/procesamiento-del-leguaje-natural-aplicaciones/>
- Jiménez, M. B., Pacheco, G. F., & García, A. C. (Abril de 2019). *RESEARCHGATE*. Obtenido de RESEARCHGATE: https://www.researchgate.net/publication/340818034_Metodologia_mixta_estudios_de_caso
- Johnson, D. (01 de 01 de 2022). *Guru99*. Obtenido de Guru99: <https://www.guru99.com/nlp-tutorial.html>
- Kaur, J. (07 de Octubre de 2021). *XENONSTACK*. Obtenido de XENONSTACK: <https://www.xenonstack.com/blog/evolution-of-nlp>
- Latinoamérica, E. d. (21 de Noviembre de 2019). *QuestionPro*. Obtenido de QuestionPro: <https://www.questionpro.com/es/investigacion-cualitativa.html>
- León Esmeralda. (2020). Procesamiento del lenguaje natural (PLN) con Python. *Baoss*. Obtenido de <https://www.baoss.es/procesamiento-del-lenguaje-natural-pln-con-python/>
- León, E. (16 de Diciembre de 2020). Procesamiento del lenguaje natural (PLN) con Python. *Baoss Analytics Everywhere*. Obtenido de <https://www.baoss.es/procesamiento-del-lenguaje-natural-pln-con-python/>
- Leyva-Vázquez, D. M., & Amarandache, D. F. (Noviembre de 2018). Inteligencia Artificial: retos, perspectivas y papel de la Neutrosfía. *Educación, Política y Valores*, 5. Obtenido de <https://books.google.es/books?hl=es&lr=&id=FqqcDwAAQBAJ&oi=fnd&pg=PA5&dq=origen+de+la+inteligencia+artificial&ots=-Q1SyDeRly&sig=7qCKAD87Y6uXcjwekxrnRYsQgj4#v=onepage&q=origen%20de%20la%20inteligencia%20artificial&f=false>
- LIMA, A. (2021). PNL | CÓMO FUNCIONA LA TOKENIZACIÓN DE TEXTO, ORACIONES Y PALABRAS. *ACERVO LIMA*. Obtenido de ACERVO LIMA: <https://es.acervolima.com/pnl-como-funciona-la-tokenizacion-de-texto-oraciones-y-palabras/>
- Microsoft. (15 de Diciembre de 2021). ¿Qué es el aprendizaje automático de Azure? *Microsoft*. Obtenido de [https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-](https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-machine-)
is-azure-machine-

learning#:~:text=Azure%20Machine%20Learning%20studio%20is%20a%20web%20portal, Azure%20Machine%20Learning%20studio%20at%20ml.azure.com.%20See%20More.

Microsoft. (2021). Elección de una tecnología de procesamiento de lenguaje natural en Azure. *Microsoft*. Obtenido de <https://docs.microsoft.com/es-es/azure/architecture/data-guide/technology-choices/natural-language-processing>

Microsoft. (14 de Septiembre de 2021). *Tutorial: Mostrar sugerencias de bombilla*. Obtenido de Tutorial: Mostrar sugerencias de bombilla: <https://docs.microsoft.com/es-es/visualstudio/extensibility/walkthrough-displaying-light-bulb-suggestions?view=vs-2022>

Mike Attal. (3 de Diciembre de 2021). *DataScientest*. Obtenido de DataScientest: <https://datascientest.com/es/nlp-natural-language-processing-introduccion>

Molina, J. (23 de 09 de 2021). *Tesis y Másters COLOMBIA*. Obtenido de Tesis y Másters COLOMBIA: <https://tesisymasters.com.co/tipos-de-investigacion/>

Nicolás A. Núñez, R. A. (Febrero de 2021). Uso de minería de textos para comparar los contenidos relacionados a calidad y acreditación generados en redes sociales por universidades de Perú y Chile. *Scielo*. Obtenido de https://www.scielo.cl/scielo.php?pid=S0718-50062021000100111&script=sci_arttext

Norman, A. T. (2020). *APRENDIZAJE AUTOMÁTICO EN ACCIÓN*. Obtenido de <https://books.google.es/books?hl=es&lr=&id=iTIREAAQBAJ&oi=fnd&pg=PT14&dq=aprendizaje+autom%C3%A1tico&ots=hVTevkd7J0&sig=2dYJe65er-Wu4Ryj-hnXgwIDG-A#v=onepage&q&f=false>

Ocampo Campos , M. (2017). *Metodos de Investigacion Academica*. Universidad de Costa Rica.

Perrier, A. (2017). Effective Amazon Machine Learning. Retrieved from https://books.google.com.ec/books?hl=es&lr=&id=I0lwDwAAQBAJ&oi=fnd&pg=PP1&dq=Amazon+Machine+Learning&ots=vM7tOdImks&sig=A4rWriGscMv-U1ule4Bmn9l5sVU&redir_esc=y#v=onepage&q=Amazon%20Machine%20Learning&f=false

Pham, B. (19 de 2 de 2020). Parts of Speech Tagging: Rule-Based. *Computer and Information Sciences Undergraduate (CISC)*, 2. Obtenido de https://digitalcommons.harrisburgu.edu/cgi/viewcontent.cgi?article=1001&context=cisc_student-coursework

ROUHIAINEN, L. (2018). *INTELIGENCIA ARTIFICIAL 101 cosas que debes saber hoy sobre nuestro futuro*. Barcelona: Planeta, S.A., 2018. Obtenido de https://static0planetadelibroscom.cdnstatics.com/libros_contenido_extra/40/39308_Inteligencia_artificial.pdf

Silva, D. d., Associate, W. C., & LATAM. (10 de Noviembre de 2020). *zendesk*. Obtenido de zendesk: <https://www.zendesk.com.mx/blog/historia-inteligencia-artificial/>

Smirnov, A., Teslya, N., Shilov, N., Frank, D., Weidig, D., Minina, E., & Evers, K. (2021). Natural Language Processing Workflow for Customer Request Analysis in a Company. *ScienceDirect*, 5. Retrieved from

<https://reader.elsevier.com/reader/sd/pii/S240589632100906X?token=5F52544675EE2D73BF91BBE9A43092A3FE8F269FBA0B68123BED68833C66E2A8820A0A2BF201996DCACC5C389BD49C1E&originRegion=us-east-1&originCreation=20220130234902>

- Solís, L. D. (04 de Febrero de 2020). *investigalia*. Obtenido de investigalia:
<https://investigaliacr.com/investigacion/la-entrevista-en-la-investigacion-cualitativa/#:~:text=La%20entrevista%20en%20la%20investigaci%C3%B3n%20cualitativa%20es%20una%20t%C3%A9cnica%20para,a%20prop%C3%B3sitos%20concretos%20del%20estudio.&text=La%20entrev>
- Sunniva Labarthe. (2020). ¿Qué pasa en Ecuador? *Nueva sociedad*, 1. Obtenido de Nueva sociedad: <https://www.nuso.org/articulo/que-pasa-en-ecuador/>
- Tecnología, I. y. (29 de Noviembre de 2019). Lenguaje R, ¿qué es y por qué es tan usado en Big Data? *UNIR*. Obtenido de <https://www.unir.net/ingenieria/revista/lenguaje-r-big-data/>
- Tomás, J. A., & Varela, M. M. (11 de Julio de 2020). La inteligencia artificial y sus aplicaciones en medicina. *ELSEVIER*, 2. Obtenido de
<https://reader.elsevier.com/reader/sd/pii/S0212656720301451?token=4CDD15BCDB5906FF451F000DFA4EFDDA963ABA5DF9F64C5B8CAD960B90242C63E73B7F7E0FB2A143B37DF23970D15081&originRegion=us-east-1&originCreation=20211210211314>
- Torres, J. A. (15 de 07 de 2021). Análisis de opinión sobre tuits del COVID-19 generados por usuarios ecuatorianos. *CEDAMAZ*, 2. Obtenido de
[file:///C:/Users/Usuario/Downloads/oscar_cumbicus,+9__An_lisis_de_opini_n_sobre_tuits_del_COVID_19_generados_por_usuarios_ecuatorianos%20\(1\).pdf](file:///C:/Users/Usuario/Downloads/oscar_cumbicus,+9__An_lisis_de_opini_n_sobre_tuits_del_COVID_19_generados_por_usuarios_ecuatorianos%20(1).pdf)
- Visus, A. (Octubre de 2020). ¿Para qué sirve Python? Razones para utilizar este lenguaje de programación. *ESIC*. Obtenido de <https://www.esic.edu/rethink/tecnologia/para-que-sirve-python>
- Zamora, K. (12 de Diciembre de 2021). SlickEdit, Historia, Características más relevantes. *Slideshare*. Obtenido de <https://es.slideshare.net/KevinZamora32/slickedit-historia-caractersticas-ms-relevantes#:~:text=SlickEdit%20Es%20un%20editor%20de,editar%20r%C3%A1pidamente%20hasta%202%20TB.>