UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE QUITO

CARRERA DE COMPUTACIÓN

DESARROLLO DE UN ALGORITMO PARA EL ANÁLISIS DE SENTIMIENTOS DE TEXTOS EN KICHWA EN EL ÁMBITO ECUATORIANO

Trabajo de titulación previo a la obtención del
Título de Ingenieros en Ciencias de la Computación

AUTORES: MARÍA FERNANDA ALBÁN MORALES
BRYAN XAVIER GUALOTO FUENTES

TUTOR: DIEGO FERNANDO VALLEJO HUANGA

Quito - Ecuador
2024

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotros, María Fernanda Albán Morales con documento de identificación N°1718567413 y Bryan Xavier Gualoto Fuentes con documento de identificación N°1725417032; manifestamos que:

Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Quito, 1 de marzo del 2024

Atentamente,

.....
María Fernanda Albán Morales
1718567413

.....
Bryan Xavier Gualoto Fuentes
1725417032

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Nosotros, María Fernanda Albán Morales con documento de identificación N°1718567413 y Bryan Xavier Gualoto Fuentes con documento de identificación N°1725417032, expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Artículo Académico: "Desarrollo de un algoritmo para el análisis de sentimientos de textos en Kichwa en el ámbito ecuatoriano", el cual ha sido desarrollado para optar por el título de: Ingenieros en Ciencias de la Computación, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribimos este documento en el momento que hacemos la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Quito, 1 de marzo del 2024

Atentamente,

.....
María Fernanda Albán Morales
1718567413

.....
Bryan Xavier Gualoto Fuentes
1725417032

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Diego Fernando Vallejo Huanga con documento de identificación N°1720162708, docente de la Universidad Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: DESARROLLO DE UN ALGORITMO PARA EL ANÁLISIS DE SENTIMIENTOS DE TEXTOS EN KICHWA EN EL ÁMBITO ECUATORIANO, realizado por María Fernanda Albán Morales con documento de identificación N°17185767413 y por Bryan Xavier Gualoto Fuentes con documento de identificación N°1725417032, obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Quito, 1 de marzo del 2024

Atentamente,

.....
Ing. Diego Fernando Vallejo Huanga, MSc
1720162708

Desarrollo de un algoritmo para el análisis de sentimientos de textos en Kichwa en el ámbito ecuatoriano.

1st María Fernanda Albán Morales malbanm@est.ups.edu.ec
2nd Bryan Xavier Gualoto Fuentes bgualotof@est.ups.edu.ec
3st Diego Fernando Vallejo Huanga dvallejoh@est.ups.edu.ec

***Resumen*—En el marco de la diversidad cultural y lingüística en Ecuador, la Constitución reconoce oficialmente al Kichwa como un medio para fomentar el diálogo intercultural. Sin embargo, este idioma ha sido objeto de prejuicios y estigmatización en la cultura e identidad de los pueblos indígenas, ergo, el estado ha implementado acciones para promover el uso del idioma, incluyendo programas educativos bilingües y la creación de materiales culturales. Dada la naturaleza mayoritariamente oral de este idioma, existen pocos datos textuales disponibles, por lo que se presentan desafíos para el desarrollo de algoritmos de análisis computacional lingüístico. En esta investigación se generó un *dataset* en idioma Kichwa, etiquetado manualmente con criterios de valencia léxica, para evaluar la carga emocional de los tokens contenidos en el diccionario entre positivos, negativos o neutros. Este conjunto de datos permite realizar, a-posteriori, el análisis de sentimientos de un nuevo texto ingresado por el usuario, en un prototipo web desarrollado mediante *Flask* y Python. En la metodología de desarrollo, el pre-procesamiento de datos utiliza técnicas de Procesamiento del Lenguaje Natural (NLP) y se aplican métricas de similitud como el coeficiente de Jaccard y Coseno Vectorial para cuantificar la polaridad del texto ingresado en Kichwa. Para la validación de este sistema se llevó a cabo una fase de experimentación que permitió evaluar el rendimiento de nuestra herramienta frente a otras dos herramientas de análisis de sentimientos construidas con modelos de texto pre-entrenados. Los resultados muestran que, al analizar polaridades de textos en Kichwa, el modelo desarrollado alcanzó una exactitud máxima del 95% y una mejora del 6% y 18% en comparación con los modelos de *ChatGPT* y *Bard*, respectivamente.**

Palabras Clave*—Procesamiento del Lenguaje Natural, Kichwa, Jaccard, Coseno Vectorial, *ChatGPT*, *Bard

***Abstract*—In the context of cultural and linguistic diversity in Ecuador, the Constitution officially recognizes Kichwa as a means to promote intercultural dialogue. However, this language has been the object of prejudice and stigmatization in the culture and identity of indigenous peoples. Therefore, the state has implemented actions to promote the use of the language, including bilingual educational programs and the creation of cultural materials. Given the primarily oral nature of this language, there is little textual data available, which presents challenges for the development of computational linguistic analysis algorithms. In this research, a dataset was generated in the Kichwa language, manually labeled with lexical valence criteria, to evaluate the emotional charge of the tokens contained in the dictionary between positive, negative, or neutral. This dataset allows performing a sentiment analysis of a new text entered by the user in a web prototype developed using *Flask* and Python. In the development methodology, data pre-processing uses Natural Language Processing (NLP) techniques and similarity metrics such as the Jaccard coefficient and Vector Cosine are applied to quantify the polarity of the text entered in Kichwa. To validate this system, an experimentation phase was carried out that allowed us to evaluate the performance of our tool against two other sentiment analysis tools built with pre-trained text models. The results show that, when analyzing polarities of Kichwa texts, the developed model achieved a maximum accuracy of 95% and an improvement of 6% and 18% compared to the *ChatGPT* and *Bard* models, respectively.**

Keywords*—Natural Language Processing, Kichwa, Jaccard, Vector Cosine, *ChatGPT*, *Bard

I. Introducción

El análisis de sentimientos, también denominado como análisis de percepciones, es una técnica computarizada para la identificación del carácter emocional de un contenido textual. Puede ser clasificado como una técnica derivada de la rama del Procesamiento del Lenguaje Natural (NLP), utilizada para extraer y determinar información de distintas fuentes de texto. El NLP, de manera general, se puede emplear en una variedad de textos, incluidas publicaciones en redes sociales, reseñas de clientes y artículos de noticias [1].

El objetivo del análisis de sentimientos es categorizar el sentimiento general de una parte del texto como positivo, negativo o neutro. Esto se lo realiza mediante la uso de procedimientos como el NLP y aprendizaje automático. Para ejecutar esta tarea el texto se procesa y se limpia, luego se entrena un modelo de clasificación derivado de un conjunto de datos de ejemplos etiquetados. Las técnicas comunes incluyen bolsas de palabras y léxicos emocionales, que son listas de palabras y sus puntajes de sentimiento asociados [2].

Según una investigación realizada en el año 2022 por Levey et al., se anticipa que el mercado mundial de análisis de sentimientos crezca en una tasa anual del 20% hasta 2027. Mientras que en el mercado global de NLP se espera un crecimiento de hasta un 17% anual. El crecimiento de estos mercados se ve impulsado por varios factores, como la demanda de análisis de datos, popularidad del uso de plataformas sociales y la necesidad de ampliar la comunicación entre la población [3].

Actualmente, las empresas de América Latina pueden usar el análisis de sentimientos para dar seguimiento a su marca y determinar el grado de satisfacción del cliente, mientras que gobiernos y organizaciones pueden usarlo para monitorear la opinión pública sobre temas particulares [4]. Además, el NLP

se puede utilizar en el análisis financiero para predecir las tendencias del mercado de valores mediante el análisis de noticias y publicaciones en redes sociales sobre compañías que cotizan en la bolsa [5] [6]. Una de las áreas de utilización del análisis de sentimientos es la identificación de emociones y polaridades en diferentes idiomas ancestrales o nativos [7], cuya tasa de uso no suele ser masivo o ampliamente difundido.

Por otro lado, la constitución del Ecuador ha reconocido oficialmente al Kichwa como un medio para fomentar el diálogo intercultural y construir una sociedad más equitativa e imparcial. Su valoración para la diversidad cultural ha contribuido a fortalecer la identidad cultural de las comunidades originarias de regiones andinas, amazónicas y a fomentar el respeto y lingüística en el país. Se trata de la segunda lengua más hablada en el Ecuador, después del español, y es utilizado por más de un millón de personas [8].

A pesar de los prejuicios y la estigmatización a sus hablantes, el idioma Kichwa ha sobrevivido y continúa desempeñando un papel importante en el patrimonio y singularidad de las comunidades indígenas de Ecuador. En la actualidad, existen esfuerzos para promover el uso de este idioma, incluyendo programas educativos bilingües y la creación de materiales culturales [9] [10].

Es importante comprender las perspectivas y opiniones de los hablantes nativos del idioma Kichwa sobre una variedad de temas, ergo, tener una herramienta de análisis de sentimientos para textos escritos en este idioma es menester. Sin embargo, debido a que es un idioma de tradición y transmisión oral, la falta de datos textuales disponibles dificulta la generación de algoritmos de propósito general, para el análisis computacional lingüístico [11].

El propósito de esta investigación es diseñar un algoritmo enfocado en el análisis de sentimientos basado en técnicas de NLP con la integración de un aplicativo web. La herramienta permitirá comprender la opinión pública en las comunidades que usan este lenguaje y cómo estos resultados podrían aplicarse a otros contextos.

Este proyecto de investigación tiene como objetivo primordial la conservación y fomento de la pluralidad lingüística del idioma nativo Kichwa en Ecuador. Al desarrollar un algoritmo de análisis de sentimientos en Kichwa, se busca reducir brechas digitales y fomentar la inclusión de esta comunidad en la era digital. Esto, a su vez, facilitará la comprensión de las necesidades y demandas de esta comunidad. Además, ofrecerá un recurso para investigadores y académicos que trabajan en casos específicos de procesamiento del lenguaje natural, contribuyendo así al desarrollo de la investigación en este ámbito y creando nuevas perspectivas para la aplicación de NLP en otros idiomas nativos de baja difusión.

A. Trabajos Relacionados

La exploración de estudios previos en el ámbito científico en el campo de interés de esta investigación, permitió identificar algunos trabajos vinculados con el análisis de sentimientos de textos escritos en Kichwa. Estos trabajos brindan información valiosa del tema y, además, permiten contextualizar nuestro trabajo para establecer relaciones entre diferentes lineas de exploración que se llevan en el campo del aprendizaje de máquina.

El artículo de Satyendra et al. examina datos de redes sociales con el fin de analizar los sentimientos expresados en varios textos. La metodología propuesta consta de tres fases: en la primera etapa, se realiza el preprocesamiento de los datos para limpiar la información que se encuentra en formato de texto; la segunda fase, utiliza la técnica *TF-IDF* para extraer atributos relevantes del texto en formato matricial y caracterizar a los documentos como vectores; estas características se utilizan en la tercera fase para hacer predicciones utilizando un clasificador que combina los métodos de *Random Forest* y Máquinas de Soporte Vectorial (SVM). Los resultados se informan en términos de diferentes métricas de evaluación, como *precision*, *recall*, *F1-Score* y *accuracy*. Este estudio fue realizado en Uttarakhand, India y se desarrolló en idioma inglés.

En el trabajo de Li et al. [12] se propone un modelo de red neuronal llamado *Sentiment Information based Network Model* (SINM) para el análisis de sentimiento en textos escritos en idioma chino. El enfoque metodológico empleado en la investigación se fundamenta en la combinación de redes neuronales, específicamente un codificador Transformer y LSTM. Además, se emplea un diccionario de emociones chino para ayudar al modelo a identificar los sentimientos dentro de los textos. Los resultados obtenidos muestran que el modelo SINM logra un mejor rendimiento y una mayor capacidad de generalización que la mayoría de los métodos existentes.

El objetivo del estudio de Muñoz et al. [13] fue analizar el corpus obtenido de las redes sociales Facebook, YouTube y Twitter sobre el tema de Hirak 19, que fue una protesta popular en Argelia llevada a cabo en el año 2019 y que fue escrita en dialecto argelino. Los autores utilizaron algoritmos tradicionales de *Machine Learning* y *Deep Learning* como *Convolutional Neural Networks* (CNN) y *Recurrent Neural Networks* (RNN), obteniendo una exactitud del 63.28% y 60.97%, respectivamente.

Otro tópico importante sobre el cual se realizó varias investigaciones relacionadas con análisis de sentimientos fue el de COVID-19 y sus implicaciones sobre la población. Así, el artículo [14] minó las emociones expresadas en *tweets* relacionados con el COVID-19 en idioma inglés. La metodología se basa en el uso de tres algoritmos, regresión logística para el proceso de clasificación, *VADER* para el análisis de percepciones y el análisis de actitudes utiliza *BERT*. Se estableció un rango entre -1 y 1 para realizar la comparación entre los algoritmos, donde el modelo *BERT* obtuvo el mejor rendimiento con un 92% de exactitud en comparación con *VADER*.

Uno de los cambios producidos por la epidemia global del Covid-19 fue la necesidad de reevaluar y transformar los sistemas de educación. Así, en el artículo [15] se realizó un análisis de sentimientos posterior al Covid-19, para evaluar la educación en línea en India. Para recopilar información, se emplearon métodos que incluyeron cuestionarios y entrevistas

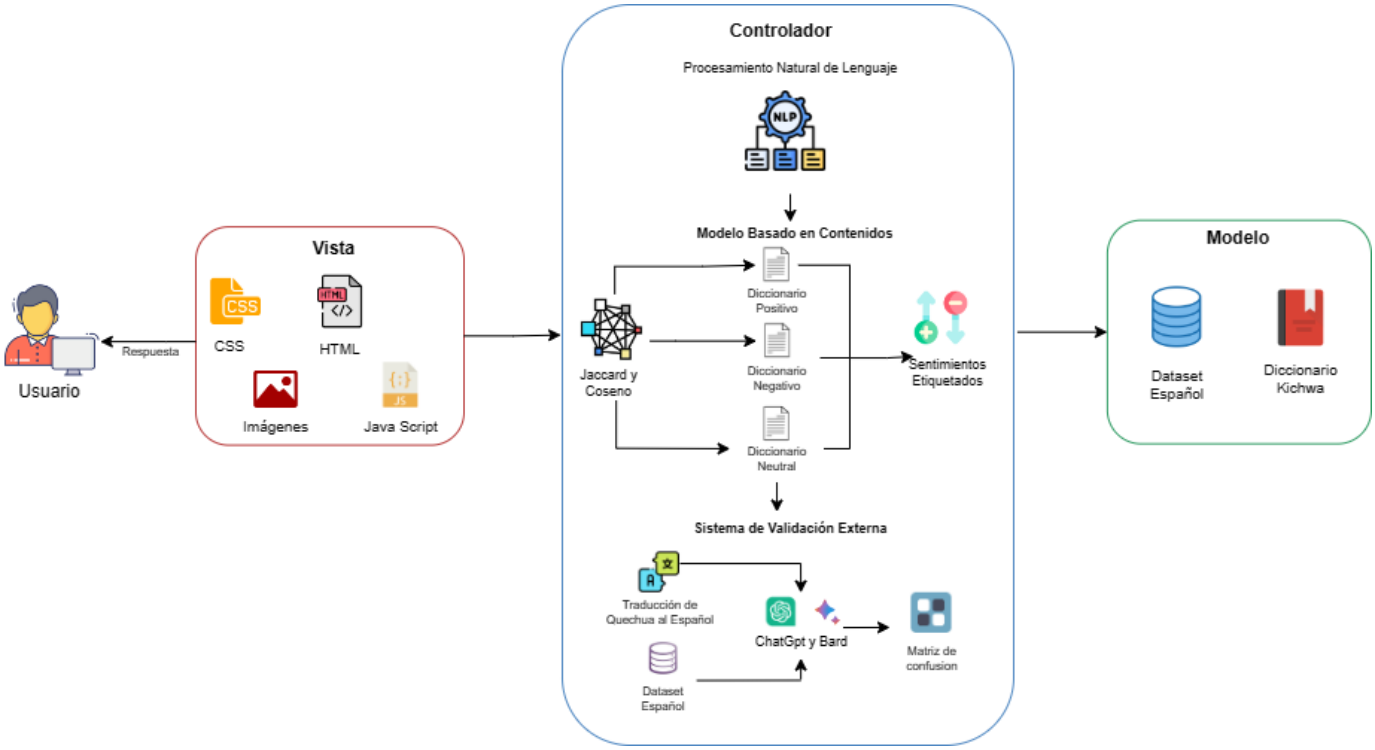

Fig. 1: Arquitectura MVC y proceso metodológico de interacciones en la herramienta web.

para obtener datos primarios. Además, se utilizó la red social Twitter para recopilar un corpus de 5000 *tweets* como fuente de datos secundaria. Sobre los datos recopilados se ejecutaron tareas de preprocesamiento, se aplicaron técnicas de NLP y análisis de bolsa de palabras para identificar los sentimientos expresados en el texto, centrándose en el estudio de la polaridad y subjetividad. Los hallazgos indican que la mayor parte de la información primaria recopilada reflejaba sentimientos negativos. En contraste, los datos secundarios indicaron una polaridad neutra y una subjetividad positiva.

Aunque existen varios trabajos previos en el área de análisis de sentimientos en diferentes idiomas y dialectos nativos, la revisión de las publicaciones académicas muestran que no existen trabajos previos del uso en el idioma Kichwa. Por otro lado, existen algunos artículos científicos que abordan el estudio computacional del idioma Kichwa, pero sus análisis no se centran en la minería de opiniones. Así, Yahuarcani et al. [16] presentan la experiencia de diseño, desarrollo y validación de una aplicación móvil, llamada Wawa, como herramienta educativa para el aprendizaje del idioma Kichwa en las comunidades del río Napo en Loreto, Perú. Su metodología se basó en la revisión y análisis de trabajos relacionados, comparando la efectividad de la enseñanza tradicional con la metodología experimental que utiliza la aplicación Wawa. Los resultados indican que Wawa es una herramienta efectiva para el aprendizaje del idioma Kichwa, siendo especialmente eficaz en la enseñanza de vocales y el alfabeto, superando a la enseñanza tradicional en la instrucción de expresiones, números y colores.

El artículo [17] presenta un sistema de traducción automática diseñado para traducir frases desde el idioma español al idioma Kichwa ecuatoriano y viceversa. La metodología empleada implica la comparación de dos enfoques de redes neuronales: la Red Neuronal de Memoria a Corto y Largo Plazo (LSTM-NN) y la Red Neuronal Transformadora (Transformer NN) dentro del proceso de modelado y aprendizaje. Los autores realizan un análisis de la situación tecnológica actual y seleccionan estos dos enfoques como opciones potenciales para abordar la tarea de traducción. Aunque los resultados iniciales no alcanzan el nivel de precisión deseado, debido a la limitada disponibilidad de textos bilingües, la primera aproximación muestra la viabilidad de implementar LSTM-NN en investigaciones y desarrollos en esta área. A pesar de los desafíos relacionados con la falta de textos bilingües, los autores concluyen que existe la posibilidad de continuar investigando y desarrollando la herramienta de traducción automática utilizando la red neuronal LSTM.

II. Materiales y Métodos

En esta investigación, se aborda la creación y análisis de un *dataset* en idioma Kichwa etiquetado. La recopilación de 2705 términos es respaldado por un enfoque de etiquetación manual, aplicando criterios de valencia léxica para evaluar la carga emocional de cada palabra. Este conjunto de datos es el recurso fundamental para poder continuar con el desarrollo metodológico. La estructura de este estudio se segmenta en tres etapas, destacando el pre-procesamiento de datos mediante técnicas de NLP y la aplicación de métricas de similitud

como el coeficiente de Jaccard y el Coseno Vectorial. Para validar la eficacia de la herramienta, se usa validación externa, comparando los resultados obtenidos con *ChatGPT* y *Bard*. Finalmente, integramos estos hallazgos en un sistema web, facilitando la interacción directa y sencilla con el usuario final. La arquitectura y metodología plateada para este proyecto de investigación se resumen en el esquema de bloques de la Figura 1

A. Conjunto de Datos en Idioma Kichwa Etiquetado

Para crear el conjunto de datos, se recolectaron alrededor de 2705 términos del diccionario más reciente de la Academia de la Lengua Kichwa del Ecuador (ALKI) [18]. La tarea de recolección asegura que el conjunto de palabras en Kichwa contenga una amplia gama de términos frecuentemente usados. Posteriormente, se procedió a la etiquetación manual de todos los términos recopilados en este diccionario, siguiendo criterios de valencia léxica, los cuales permiten evaluar la carga emocional inherente de cada palabra.

La valencia léxica se convierte en un elemento crucial para la caracterización de las palabras del conjunto de datos, ya que las clasifica según su capacidad para transmitir emociones. Este proceso de etiquetación permite determinar la valencia de cada palabra, i.e., si tiene una connotación positiva, negativa o neutral. Este enfoque garantiza que el conjunto de datos sea exhaustivo y representativo de la diversidad de matices emocionales presentes en el lenguaje Kichwa. Como resultado de este proceso, se clasificaron 641 términos como positivos $D_i(+)$, 461 como negativos $D_i(-)$, 1603 como neutros $D_i(\pm)$ y se unificaron en un vector D_i .

B. Proceso Metodológico para Tratamiento del Texto

El desarrollo de este estudio está segmentado en tres etapas. La primera etapa incluye el pre-procesamiento de datos mediante NLP, tal como se puede visualizar en la Figura 2.

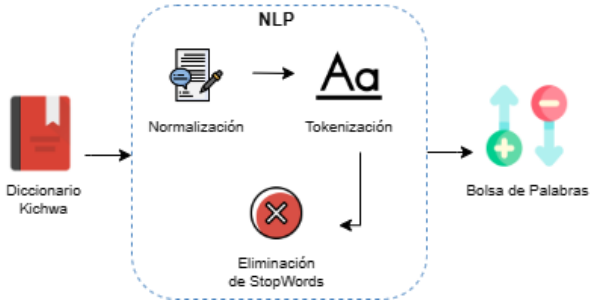

Fig. 2: Etapas de pre-procesamiento de datos mediante técnicas de NLP.

Esta primera etapa incluye tres fases para poder procesar la información no estructurada en formato textual: normalización, tokenización y eliminación de *stopwords*. La normalización implica convertir el texto a minúsculas y eliminar caracteres especiales. Es importante destacar que, en este contexto lingüístico, el idioma Kichwa, no posee palabras acentuadas. Sin embargo, existen ciertas palabras que usan el signo ortográfico virgulilla ($\sim$) que para el tratamiento del texto no serán eliminadas, ya que poseen un significado semántico en este idioma. Esto último asegura que el análisis sea coherente y libre de ambigüedades. La tokenización, por su parte, implica la subdivisión o segmentación del texto en unidades más pequeñas, denominadas tokens. Estos tokens son unidades mínimas de información textual, representadas como palabras individuales que facilitan la comprensión y posterior procesamiento del corpus. Finalmente, en el caso específico del idioma Kichwa, se construyó una lista de 24 tokens identificados como *stopwords*. Para la creación de esta lista se realizó una comparación entre las *stopwords* del idioma español y sus respectivos significados almacenados en nuestra base de datos de idioma Kichwa. El sistema se encarga de buscar coincidencias; en caso de hallarlas, estas palabras se incorporan a un vector específico de *stopwords* en Kichwa. Posteriormente, este vector se somete al proceso de normalización aplicando los mismos criterios especificados anteriormente para eliminar las palabras que carecen de significado. Los procesos de normalización y eliminación de palabras vacías, depuran el contenido del corpus y permiten que las métricas de similitud se centren en palabras clave y conceptos relevantes, mejorando la calidad de los resultados en el análisis de sentimientos. Además, es menester señalar que dentro del pre-procesamiento del texto no se lleva a cabo el proceso de *stemming* debido a la complejidad inherente del proceso de reducción hacia la raíz de las palabras en Kichwa. Este procedimiento requiere una comprensión lingüística profunda y extensa del idioma, lo cual puede suponer un desafío considerable en términos temporales.

La segunda etapa del estudio aplica métricas de similitud para identificar la divergencia entre dos textos que previamente han pasado por la etapa de pre-procesamiento con NLP. Las dos métricas utilizadas en esta investigación son el coeficiente de Jaccard (J) y el Coseno Vectorial (CV). Estos dos coeficientes de similitud son empleados, ampliamente, en procesos de búsqueda y recomendación de textos [19].

El modelo de similitudes, utiliza un diccionario con todos los tokens únicos contenidos en el *dataset* de idioma Kichwa y los representa mediante un vector unificado de términos D_i y los indexa juntos con el texto proporcionado por el usuario, denotado como T_j. Este vector resultante se identifica como $V_k : \{1, 2, 3, ..., n\}; k = i + j$. Este enfoque permite la integración y análisis conjunto de elementos provenientes de los diccionarios, D_i, y del texto ingresado en Kichwa, T_j, para el análisis de sentimientos, dentro de un rango acotado, para evaluar la similitud entre los elementos de ambas fuentes de datos.

Sobre el vector unificado V_k se aplican las métricas de similitud como se puede visualizar en la Figura 3.

Primero, el coeficiente de Jaccard es una métrica ampliamente empleada para cuantificar la similitud entre dos conjuntos de tokens en un texto. Se calcula como la tasa de la intersección de términos comunes entre dos textos sobre la unión total de los términos. En el contexto de nuestra herramienta de análisis de sentimientos para el idioma

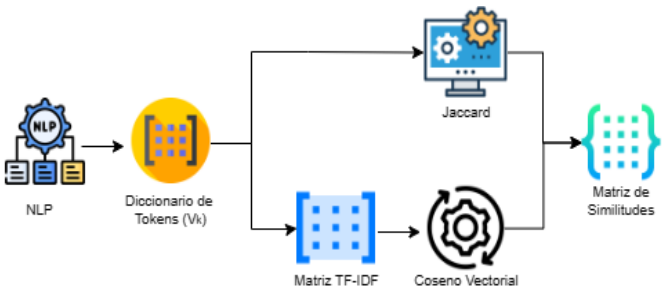

Fig. 3: Generación de la Matriz de Similitud mediante el uso del coeficiente de Jaccard y Coseno Vectorial

Kichwa, se utiliza para evaluar la similitud entre $D_i(+)$, $D_i(-)$, $D_i(\pm)$ y T_j, respectivamente. Un valor de J cercano a 1 se interpreta como un mayor nivel de similitud, lo que permite clasificar a un texto T_j dentro de las tres posibles polaridades que tiene D_i. Un texto T_j, es clasificado como positivo $T_j(+)$ cuando su nivel de similitud $J(D_i(+), T_j)$ supera al $J(D_i(-), T_j) \vee J(D_i(\pm), T_j)$. Para un sentimiento negativo el nivel de similitud $J(D_i(-), T_j)$ del $T_j(-)$ supera al $J(D_i(+), T_j) \vee J(D_i(\pm), T_j)$. Finalmente, para el caso de que los niveles de similitud sean iguales $J(D_i(-), T_j) \vee J(D_i(+), T_j)$ se considera como un sentimiento neutro $T_j(\pm)$. Esta métrica se representa mediante la Ecuación 1.

$$J(D_i, T_j) = \begin{cases} \frac{|D_i(+) \cap T_j|}{|D_i(+) \cup T_j|} \\ \\ \frac{|D_i(-) \cap T_j|}{|D_i(-) \cup T_j|} \\ \\ \frac{|D_i(\pm) \cap T_j|}{|D_i(\pm) \cup T_j|} \end{cases} \quad (1)$$

La segunda métrica de similitud utilizada en esta investigación es el Coseno Vectorial, que se implementa mediante el enfoque de Bolsa de Palabras, también conocido como *Bag of Words*, con ponderación TF-IDF. Esta medida evalúa la semejanza entre vectores representativos de documentos en función de la frecuencia de términos específicos en dichos documentos. En el análisis de sentimientos, se emplea una matriz TF-IDF con los V_k tokens, asignando un peso a cada término según su relevancia relativa, para los vectores D_i y T_j. De esta manera, se compara la lista fija V_k con los términos del vector T_j y se asigna la polaridad en función de la puntuación obtenida, seleccionando aquella que mejor refleje el significado de la frase.

El cálculo de esta métrica se encuentra representada por la Ecuación 2.

$$CV = \frac{\sum_{k=1}^{n} D_k T_k}{\sqrt{\sum_{k=1}^{n} D_k^2} \sqrt{\sum_{k=1}^{n} T_k^2}} \quad (2)$$

Una vez aplicadas estas dos técnicas se obtienen dos matrices de similitud M_{s_J} y $M_{s_{CV}}$ que permiten identificar el nivel de divergencia que existe entre los textos ingresados y los diccionarios de polaridades. El Algoritmo 1 muestra el pseudocódigo para el análisis de sentimientos de textos en Kichwa en el ámbito ecuatoriano.

Algoritmo 1 Algoritmo para el análisis de sentimientos de textos en Kichwa

Entrada: D_i, T_j
Salida: $M_{s_{J_k}}$ y $M_{s_{CV_k}}$
1: Paso 1: Inicializar Variables
2: $D_i \leftarrow [1, ..., i]$;
3 : $T_j \leftarrow [1, ..., j]$;
4: Paso 2: Procesamiento NLP
for *i ← 1, n* **do**
  6: $D_i[i] \leftarrow Normalizacion(D_i[i])$;
  7 : $D_i[i] \leftarrow Tokenizacion(D_i[i])$;
  8 : $D_i[i] \leftarrow EliminacionStopwords(D_i[i])$;
  9 : $T_j[i] \leftarrow Normalizacion(T_j[i])$;
  10 : $T_j[i] \leftarrow Tokenizacion(T_j[i])$;
  11 : $T_j[i] \leftarrow EliminacionStopwords(T_j[i])$;
end
13: Paso 3: Calcular Matriz de similitud J
14: $M_{s_{J_k}} \leftarrow Jaccard(Tj)$;
15: Paso 4: Matriz de similitud CV
16: Paso 5: Calcular TF y DF
17: $df \leftarrow [cont_1, cont_2, ..., cont_n]$;
18: Paso 6: Calcular IDF
19: idf $\leftarrow IDF(len(Tj), df)$;
20: Paso 7: Calcular TF - IDF
for *i ← 1,* **do**
  22: $tfidf[i] = TFIDF(wtf[i], idf)$
end
24: Paso 8: Vectorizar textos de salida
for *i ← 1,* **do**
  26: $V_k[i] = VectorUnitario(tfidf[i])$
end
28: Paso 9: Calcular Matriz de similitud CV
29: $M_{s_{CV_k}} \leftarrow Coseno(Vk)$;

C. Desarrollo Full-Stack del Prototipo

El proyecto se desarrolló utilizando Python 3.8 como lenguaje de programación, por su sintaxis clara, legible y su amplia compatibilidad con bibliotecas y *frameworks* que agilizan el desarrollo de aplicaciones y prototipos. El sitio web está desarrollado bajo la arquitectura Modelo, Vista, Controlador (MVC) e integra tres módulos, que son: *front-end, back-end* y un Interfaz de Programación de Aplicaciones *API REST*. Los experimentos se llevaron a cabo en una computadora con un procesador Core i7, 16 GB de RAM y 2TB de almacenamiento.

En la creación del *front-end* se emplearon lenguajes como HTML5, CSS3 y JavaScript, para la organización y funcionalidad del sistema. Además, se usó la librería *Bootstrap* para garantizar un diseño web adaptable. El sitio web analizará el sentimiento del texto en idioma Kichwa ingresado por el usuario, mostrándole el porcentaje de polaridad. La polaridad más alta determina el sentimiento general del texto ingresado.

El núcleo del sistema, también conocido como *back-end*, actúa como el controlador en el patrón de diseño MVC, gestionando las solicitudes del usuario a través de métodos *POST* y generando respuestas mediante métodos *GET*. Su tarea principal consiste en manejar las solicitudes originadas cuando el usuario introduce texto en Kichwa a través de la interfaz del *front-end*. Mediante el *framework Flask* se facilita la interacción entre la interfaz y el controlador, estableciendo conexiones eficientes mediante solicitudes sincrónicas al servidor. Este enfoque proporciona una experiencia global para el usuario, donde las funciones del modelo responden de manera efectiva a las solicitudes del *front-end*.

La API ofrece un servicio que integra los procesos de NLP y las métricas de similitud, al comparar el texto ingresado por el usuario con un diccionario de términos en Kichwa previamente etiquetados. Este modelo ha priorizado la accesibilidad, asegurando una utilización eficiente del servicio por parte de los usuarios y permitiéndoles aprovechar sus funcionalidades.

D. Descripción del Funcionamiento de la Herramienta Web

En la página principal de la herramienta web, se puede visualizar una transición que muestra el título de la herramienta y una caja de texto que permite al usuario ingresar una frase en Kichwa. Una vez que el texto ha sido ingresado por el usuario, puede hacer click en el botón *Análisis* para visualizar los resultados del algoritmo que clasifica las polaridades de los sentimientos. La ejecución de este proceso permite que la API realice, internamente, las tareas de NLP sobre el texto ingresado por el usuario. Además, aplica la métrica de similitud que en el proceso experimental obtuvo el mayor rendimiento. Los resultados finales son presentados en una interfaz dedicada, la cual presenta la polaridad mayoritaria de la frase ingresada e incluye un gráfico que compara los porcentajes asociados a cada polaridad identificada, mediante el *front-end*. Esta herramienta web se encuentra disponible en el siguiente enlace: https://ukumay-fec773715828.herokuapp.com/

III. Experimentos y Resultados

El objetivo de la experimentación es examinar el rendimiento del prototipo, que comprende tanto el modelo de clasificación como la herramienta web asociada. Para este fin, se ha recopilado un nuevo conjunto de datos, S_t, que consta de 83 frases. De estas, 24 fueron catalogadas como positivas, 28 como negativas y 21 como neutras. Estas expresiones, provenientes del libro Vocabulario Elemental y Expresiones Frecuentes [20] y del libro Kichwa [21], fueron analizadas para determinar su polaridad mediante la aplicación de criterios de valencia léxica. Este conjunto de datos se utiliza para realizar pruebas exhaustivas y cuenta con un promedio de 25 *tokens* por frase. Además, cada frase que ha sido agregada en S_t cuenta con su respectiva traducción al castellano, permitiendo así evaluar la capacidad del sistema para abordar una variedad de tonalidades lingüísticas.

A. Primer Experimento: Evaluación de la Mejor Métrica de Similitud

Dado que en la metodología se consideraron dos métricas de similitud, Jaccard y Coseno Vectorial, para el análisis de sentimientos, es menester determinar su rendimiento frente a nuevas instancias textuales contenidas en el *dataset* S_t. Las 83 instancias de S_t son evaluadas, de forma individual, con las dos medidas de similitud y el resultado del análisis de sentimientos arrojará la polaridad positiva, negativa o neutra de cada texto. Se considera que las etiquetas de polaridad manuales en S_t son la clase real y se usará los resultados de la clasificación, con las dos medidas de similitud, J y CV, como clases predichas. La evaluación del rendimiento se mide con cuatro métricas *accuracy*, *precision*, *recall* y *F1-Score*; y los resultados se sintetizan en la Tabla I.

	Accuracy	Precision	Recall	F1 Score	Clase
Clase Real		0.96	0.96	0.96	Positiva
vs	0.90	0.81	0.93	0.87	Negativa
Jaccard		0.96	0.84	0.90	Neutra
Clase Real		0.92	0.96	0.94	Positiva
vs	0.95	1.00	0.93	0.96	Negativa
Coseno		0.94	0.97	0.95	Neutra

TABLE I: Evaluación de rendimiento de las métricas de similitud entre la clase real vs. Jaccard y Coseno Vectorial

Los resultados indican que ambas medidas de similitud muestran un desempeño similar en términos de *precision* y *recall* para cada clase. Ya que los resultados no muestran un sesgo hacia alguna clase, el *accuracy* será la métrica que permitirá definir a la medida de similitud con mejor rendimiento. Esto último porque la exactitud se fundamenta en la tasa de aciertos en comparación con el total de predicciones realizadas y un valor más elevado sugiere una mayor probabilidad de que el modelo clasifique correctamente a los documentos. En este contexto, el modelo basado en la métrica coseno vectorial tiene un *accuracy* de 0.95, mientras que el modelo basado en el índice de Jaccard presenta un *accuracy* de 0.90.

Las matrices de confusión, de las medidas de similitud J y CV, en la Figura 4 evidencia que las dos aproximaciones tienden a clasificar, de forma similar, las instancias en las clases positiva y negativa. Mientras que, al abordar la clasificación de la clase neutra, el modelo Coseno Vectorial demuestra una mayor exactitud en su predicción, ergo, CV es la medida de similitud que será implementada en el *front-end* del sistema web.

B. Segundo Experimento: Validación Externa empleando ChatGPT y Bard

Para evaluar la eficiencia de la herramienta propuesta en esta investigación, se lleva a cabo una validación externa utilizando dos analizadores de sentimientos implementados con el el uso de herramientas de inteligencia artificial, *ChatGPT* y *Bard*. Estas herramientas se utilizan como servicios dentro del entorno de *Google Colab*.

ChatGPT y *Bard* son dos modelos, de uso libre, para generación de texto con algoritmos pre-entrenados. Por lo

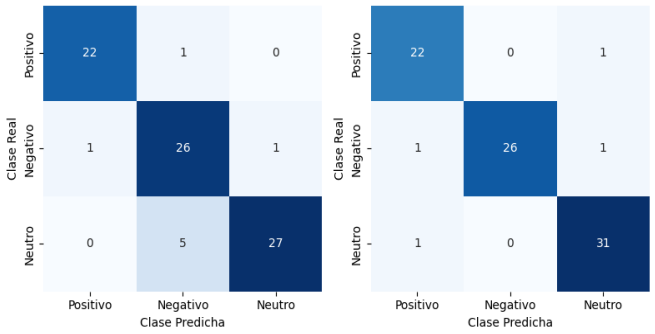

Fig. 4: Matrices de Confusión entre la clase real Vs. Jaccard y Coseno Vectorial

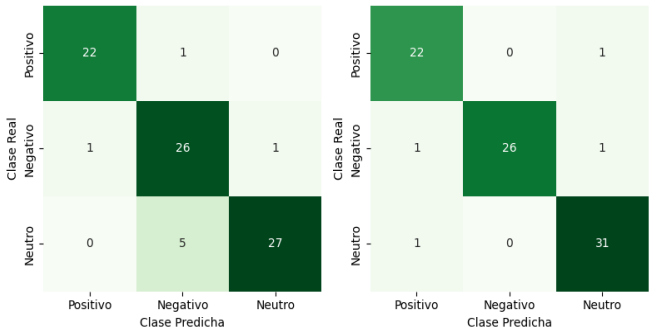

Fig. 5: Matrices de Confusión entre la clase real Vs. *Bard* y *ChatGPT*

tanto, en primera instancia, se establece un *prompt* de manera que estos modelos actúen como analizadores de sentimientos en castellano y clasifiquen las instancias de acuerdo a las clases positiva, negativa y neutra. Posteriormente, se extrae de S_t la columna que contiene la traducción para cada frase y se las procesa con las instrucciones del *prompt*. En este contexto, las herramientas asumen la responsabilidad de determinar la polaridad general de cada frase de S_t. Finalmente, luego del análisis, se compara la polaridad asignada a las frases de S_t con la clase real.

Los resultados presentados en la Tabla II permiten observar que el *accuracy* del modelo *Bard* alcanza el 77%, mientras que el modelo *ChatGPT* tiene un rendimiento superior con un 89% de *accuracy*. En términos generales, estos resultados muestran un mejor rendimiento del modelo generador de texto *ChatGPT*, para la tarea de clasificación de instancias en el contexto de un análisis de sentimientos en castellano.

	Accuracy	Precision	Recall	F1 Score	Clase
Clase Real		0.55	1.00	0.71	Positiva
vs	0.77	1.00	0.96	0.98	Negativa
Bard		1.00	0.44	0.61	Neutra
Clase Real		0.74	1.00	0.85	Positiva
vs	0.89	0.97	1.00	0.98	Negativa
ChatGPT		1.00	0.72	0.84	Neutra

TABLE II: Evaluación de rendimiento de las métricas de similitud entre la clase real vs. *Bard* y *ChatGPT*

La Figura 5 muestra que tanto el modelo *Bard* como el modelo *ChatGPT* tienen una alta tasa de *recall* para las clases positiva y negativa. Sin embargo, para la clase neutra, estos valores descienden al 44% y 72%, respectivamente.

Al analizar los resultados obtenidos a través del *accuracy* de los dos modelos pre-entrenados y de la herramienta propuesta, se observa que nuestra herramienta tiene un rendimiento superior del 6% y 18% en comparación con *ChatGPT* y *Bard*, respectivamente.

IV. Conclusiones y Trabajos Futuros

Para este proyecto de investigación, se ha desarrollado un algoritmo que permite realizar el análisis de sentimientos de textos en Kichwa en el ámbito ecuatoriano mediante una herramienta web. El enfoque de esta investigación se centra en la generación y análisis de un conjunto de datos en Kichwa, los cuales han sido etiquetados manualmente para evaluar su carga emocional. La metodología se divide en tres etapas y emplea técnicas de Procesamiento del Lenguaje Natural, así como métricas de similitud como el coeficiente de Jaccard y el Coseno Vectorial. Además, se lleva a cabo una validación externa mediante la comparación de resultados con modelos preexistentes como *ChatGPT* y *Bard*, con el fin de evaluar la eficacia de la herramienta propuesta.

Los resultados derivados de la primera fase experimental muestran que, en términos de medidas de similitud, el Coseno Vectorial logra clasificar las polaridades con una precisión del 95%, superando de esta manera al modelo basado en Jaccard. En la segunda fase experimental, al comparar los resultados obtenidos por los modelos pre-entrenados de generación de texto, *ChatGPT* y *Bard*, se observa que ambos no logran superar la exactitud obtenida por nuestra herramienta propuesta.

Esta característica se debe a que estos modelos han sido entrenados con textos diversos, lo que implica que no están especializados en realizar análisis de sentimientos específicamente en el contexto ecuatoriano. Por lo tanto, la herramienta propuesta ofrece una mejora significativa, destacando su capacidad para adaptarse y clasificar de manera más exacta las polaridades en el idioma Kichwa, ergo, la herramienta propuesta es más confiable cuando se aborda la tarea de clasificación de sentimientos de un texto.

En trabajos futuros, para mejorar la capacidad de clasificar las polaridades y hacer más preciso el modelo, se sugiere enriquecer el diccionario mediante la colaboración de hablantes nativos del idioma y expertos en lingüística. Además, se sugiere incorporar la fase de *stemming*, la cual implica un proceso complejo destinado a simplificar las palabras hacia su raíz léxica. De este modo, al optimizar la recuperación de información, se lograría una mayor precisión en el proceso de NLP.

References

[1] S. Wang and Y. Lu, "A hybrid deep learning approach for sentiment analysis of covid-19 related tweets," *Journal of biomedical informatics*, p. 125, 2022.

[2] G. H. Yılmaz, E. and C. Akkaya, "Emotion detection in turkish texts using convolutional neural networks," *Journal of Information Science*, pp. 802–821, 2021.

[3] S. Levey and L.-R. L. Cheng, "Artificial intelligence and sentiment analysis: A review in competitive research," *Computers*, vol. 2, no. 12, p. 37, 2023.

[4] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[5] X. Man, T. Luo, and J. Lin, "Financial sentiment analysis (fsa): A survey," in *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*. IEEE, 2019, pp. 617–622.

[6] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, vol. 60, pp. 617–663, 2019.

[7] R. J. Sutar and K. R. Desai, "A study on various sentiment analysis for mixed transliterated indigenous language using machine learning algorithms," 2023.

[8] P. Muñoz and L. Jacome, "El kichwa como lengua de resistencia en la literatura ecuatoriana," *Revista electrónica de estudios literarios*, pp. 1–14, 2020.

[9] S. Levey and L.-R. L. Cheng, "The impact of bias and discrimination," *Communication Disorders Quarterly*, vol. 43, no. 4, pp. 215–223, 2022.

[10] Y. P. A. Bustamante, "Intercultural bilingual education (ibe) and the image of indigenous people and quechua speakers in the peruvian andes," Ph.D. dissertation, University of Toronto (Canada), 2022.

[11] F. J. P. Siller, C. E. P. López, A. V. López, L. C. Romero, and E. R. Quesada, *Lenguaje, textos y cultura: Perspectivas de análisis y transmisión*. Ediciones Octaedro, 2021.

[12] G. Li, Q. Zheng, L. Zhang, S. Guo, and L. Niu, "Sentiment infomation based model for chinese text sentiment analysis," in *2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*, 2020, pp. 366–371.

[13] P. Muñoz and L. Jacome, "Deep learning-based sentiment analysis of algerian dialect during hirak 2019," 2021.

[14] V. G. Nair, A. and A. Vinakay, "Comparation study of twitter sentiment on covid - 19 tweets," *IEEE Explore*, 2021.

[15] D. Sonal and D. Apurva, "Post covid-19 sentiment analysis of succes of online learning: A case study of india," *IEEE Explore*, 2022.

[16] I. O. Yahuarcani, L. A. S. Llaja, A. M. N. Satalaya, J. A. G. Cruzado, A. J. A. Acuña, J. E. G. Díaz, E. G. Gómez, J. A. A. Rengifo, and R. Del Aguila, "Wawa: Mobile educational tool for learning the kichwa language for the communities of the napo river in loreto, peru," in *2022 IEEE World Engineering Education Conference (EDUNINE)*. IEEE, 2022, pp. 1–6.

[17] C. A. G. P. Garay, L. and V. Robles, "Ecuadorian kichwa to spanish automatic translation: an initial approximation based on lstm neural networks for educational applications," *IEEE Explore*, pp. 1–3, 2021.

[18] J. J. C. Aguinda, M. A. U. Velasco, and E. E. S. Andi, *Diccionario Castellano (Español) – Kichwa*. CCE Núcleo Sucumbios, 2007.

[19] L. Zahrotun, "Comparison jaccard similarity, cosine similarity and combined both of the data clustering with shared nearest neighbor method," *Computer Engineering and Applications Journal*, vol. 5, no. 1, p. 11, 2016.

[20] L. E. Cachiguango, *Vocabulario Elemental y Expresiones Frecuentes*. Cuadernos de formación y capacitación de los(as) voluntarios(as) de Cielo Azul, 2008, reedición: 2011.

[21] F. Potosí and R. V. Corral, *Kichwa*. Ministerio de Educación Ecuador, 2009.