PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Análisis de Sentimiento para lenguajes de bajos recursos, Dominio: Shipibo-Konibo

Trabajo de investigación para obtener el grado académico de Maestro en Informática con mención en Ciencias de la Computación que presenta:

Jose Alejandro Florez Tapia

Asesor:

Mg Gerardo Cardoso Yllanes

Lima, 2025

Informe de Similitud

Yo, Gerardo Cardoso Yllanes, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor(a) de la tesis/el trabajo de investigación titulado Análisis de Sentimiento para lenguajes de bajos recursos, Dominio: Shipibo-Konibo, del/de la autor(a) / de los(as) autores(as) Jose Alejandro Florez Tapia, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 7%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 11/09/2025.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:
Lima, 16 de septiembre del 2025

Apellidos y nombres del asesor / de la asesora: Cardoso Yllanes, Gerardo	
DNI: 48155961	Firma: 
ORCID: 0009-0009-5200-1906	

RESUMEN

Con el objetivo de apoyar a comunidades con bajos recursos digitales en su integración a la sociedad, se desarrolló un modelo de análisis de sentimiento para lenguas indígenas, permitiendo la implementación de tecnologías como chatbots y asistentes virtuales que puedan operar en su lengua materna. Esta propuesta busca no solo facilitar un mayor acceso a servicios esenciales en áreas como educación, salud y gobierno, sino también promover la preservación cultural y lingüística de comunidades históricamente marginadas. La incorporación de herramientas de este tipo representa una estrategia para reducir la brecha digital y garantizar un acceso más equitativo a los beneficios de la transformación tecnológica.

Para el idioma Shipibo-Konibo, se utilizaron diversas técnicas de aumento de datos basadas en errores controlados, incluyendo alteraciones aleatorias, proximidad de teclado, ambigüedad fonema-grafema y similitud silábica. Estas técnicas contribuyeron significativamente a incrementar la diversidad y representatividad del corpus, permitiendo que el modelo entrenado reflejara de manera más realista la variabilidad natural del lenguaje. Asimismo, se evaluaron modelos de embeddings multilingües como XLM-Roberta, LaBSE y SIMCSE, seleccionando finalmente el más adecuado por su capacidad de generalización y desempeño en escenarios multilingües.

Los experimentos realizados lograron superar el desafío de clasificar oraciones en categorías positivas, negativas y neutras, incluso en contextos de datos limitados. Este avance constituye un paso importante hacia la inclusión tecnológica de comunidades indígenas, ofreciendo herramientas adaptadas a sus necesidades lingüísticas y fomentando un ecosistema digital más diverso e inclusivo.

Palabras clave -- Lenguas indígenas, Shipibo-Konibo, análisis de sentimiento, aumento de datos, errores controlados, inclusión tecnológica, brecha digital.

ABSTRACT

With the objective of supporting communities with low digital resources in their integration into society, a sentiment analysis model for indigenous languages was developed, allowing the implementation of technologies such as chatbots and virtual assistants that can operate in their mother tongue. This proposal seeks not only to facilitate greater access to essential services in areas such as education, health, and government, but also to promote the cultural and linguistic preservation of historically marginalized communities. The incorporation of such tools represents a strategy to reduce the digital divide and to guarantee more equitable access to the benefits of technological transformation.

For the Shipibo-Konibo language, various data augmentation techniques based on controlled errors were applied, including random alterations, keyboard proximity, phoneme-grapheme ambiguity, and syllabic similarity. These techniques significantly contributed to increasing the diversity and representativeness of the corpus, allowing the trained model to more realistically reflect the natural variability of the language. Likewise, multilingual embedding models such as XLM-Roberta, LaBSE, and SIMCSE were evaluated, ultimately selecting the most suitable one for its generalization capacity and performance in multilingual scenarios.

The experiments carried out managed to overcome the challenge of classifying sentences into positive, negative, and neutral categories, even in low-data contexts. This advancement constitutes an important step toward the technological inclusion of indigenous communities, offering tools adapted to their linguistic needs and fostering a more diverse and inclusive digital ecosystem.

Keywords -- Indigenous languages, Shipibo-Konibo, sentiment analysis, data augmentation, controlled errors, technological inclusion, digital divide.

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. INTRODUCCION

El análisis de sentimiento es una herramienta fundamental en el procesamiento de lenguaje natural (PLN), al permitir la interpretación automatizada de opiniones y emociones expresadas en texto [1]. Sin embargo, los avances en este campo se han concentrado en lenguas con abundante documentación y recursos digitales, dejando rezagadas a las lenguas indígenas que enfrentan una representación limitada en el ámbito tecnológico [2]. Este déficit perpetúa la exclusión de estas comunidades en la era digital, afectando su acceso a servicios y tecnologías modernas. Dentro de este panorama, la comunidad Shipibo-Konibo, compuesta por aproximadamente 34 152 personas hablantes, aunque solo 25 222 se identifican como parte de esta etnia [4], es un claro ejemplo de las dificultades que enfrentan las lenguas de bajos recursos.

El Shipibo-Konibo, como muchas lenguas indígenas, posee una representación digital limitada y carece de recursos tecnológicos, situación que dificulta e incluso impide su integración en herramientas modernas de procesamiento de lenguaje natural [3]. La elección de esta lengua para este estudio radica en su vulnerabilidad y en la urgencia de preservar su uso en un contexto donde las lenguas indígenas enfrentan el riesgo de desaparición [5]. Este trabajo propone un enfoque que utiliza un corpus paralelo en español y Shipibo-Konibo para desarrollar un modelo de análisis de sentimiento, orientado a mejorar herramientas tecnológicas.

Entre las posibles aplicaciones prácticas de un modelo de análisis de sentimiento se encuentran diferentes servicios, como asistentes en educación o chatbots multilingües, que podrían aprovechar la capacidad del modelo para comprender el tono emocional de los usuarios y proporcionar respuestas culturalmente adecuadas, ayudando a identificar emociones como satisfacción o preocupación [6]. En contextos comunitarios, estos sistemas servirían igualmente para recopilar retroalimentación valiosa [7] y apoyar la toma de decisiones en ámbitos como salud, educación y servicios públicos, como se observa en la Figura 1.

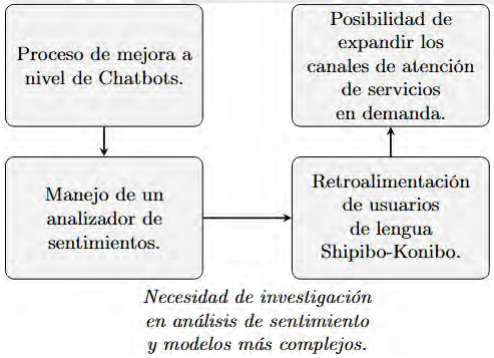

Figura 1. Relación Análisis de sentimiento Shipibo-Konibo y chatbots en otros idiomas

Más allá de las posibles aplicaciones, disponer de un analizador de sentimiento para una lengua aglutinante como el Shipibo-Konibo abre la puerta a investigaciones lingüísticas avanzadas y a la evaluación automática de contenido generado por usuarios en medios digitales [8]. Asimismo, contribuye a la expansión de los canales de atención en Shipibo-Konibo y, en general, a la inclusión digital de otras lenguas en situación de riesgo.

Para poder realizar el entendimiento de emociones y contextos que tiene el lenguaje Shipibo-Konibo, se plantea la implementación de un análisis de sentimientos. Para alcanzar dicho objetivo, se plantea completar las siguientes tareas:

La primera tarea es crear un corpus monolingüe etiquetado (positivo, negativo o neutral) del lenguaje Shipibo-Konibo; este corpus se obtendrá iniciando con la recopilación de un corpus no etiquetado en Shipibo-Konibo y su traducción en un corpus paralelo en español de alrededor de 26 000 oraciones, la cual fue realizada en otra investigación apoyada por el grupo Chana, que se preocupa por la preservación de las lenguas amazónicas [19].

Partiendo de este corpus, se realizará un etiquetado manual de aproximadamente 3000 muestras del corpus en español, creándose un conjunto de datos que sirve como base para el entrenamiento de un modelo inicial.

Estos datos serán luego pre procesados, aplicando normalización, eliminación de ruido y conversión a representaciones computacionales mediante embeddings, asegurando su adecuación para el análisis.

Para seleccionar el modelo inicial se evaluarán aquellos enfocados en análisis de sentimiento en español, como dccuchile/bert-base-spanish-wwm-uncased, un modelo basado en BERT, entrenado con whole word masking en español, el cual es ampliamente reconocido por su versatilidad y robustez en tareas de procesamiento de lenguaje en español [20], y pysentimiento/robertuito-sentiment-analysis, un modelo especializado en el análisis de sentimientos, el cual fue diseñado específicamente para capturar emociones y matices sentimentales en español con mayor precisión [21].

El modelo con mejores resultados será usado para obtener el etiquetado del resto del corpus en español pendiente, 23 000 oraciones.

Una vez completado el etiquetado del corpus en español, se integran los datos al corpus paralelo en Shipibo-Konibo, obteniendo así un corpus monolingüe etiquetado de Shipibo-Konibo.

En la segunda tarea se buscará aumentar la cantidad de datos disponibles en el corpus monolingüe etiquetado de Shipibo-Konibo, para poder entrenar modelos de sentimiento. Como estrategia, se propone generar datos mediante la introducción controlada de errores en las frases, simulando variaciones reales como errores

ortográficos o gramaticales [9]. Esto permitirá entrenar modelos más robustos, capaces de adaptarse a datos imperfectos y reflejar mejor el uso práctico del idioma.

En la tercera tarea se diseñará un modelo de análisis de sentimiento adaptado a las características del lenguaje Shipibo-Konibo, teniendo en cuenta su morfología aglutinante y las particularidades culturales de la comunidad. Utilizaremos los modelos como XLM-Roberta elegido por su capacidad de generalización multilingüe y soporte para lenguajes de bajo recurso [22], LaBSE especializado en búsqueda semántica multilingüe, optimizando la representación de frases en Shipibo-Konibo [23], y SIMCSE modelos optimizados para tareas de similitud textual [24].

Finalmente, en la cuarta tarea se implementará un prototipo del modelo de análisis de sentimiento Shipibo-Konibo accesible. La interfaz permitirá ingresar un texto en Shipibo-Konibo; al enviarlo, el sistema lo analizará y determinará si el sentimiento es positivo, negativo o neutral. El sistema devuelve la categoría de sentimiento asignada. Este prototipo inicial ayudará a la integración de herramientas de procesamiento de lenguaje natural en lenguas amazónicas.

El documento se estructura de la siguiente manera: en la Sección 2, se revisa el estado del arte, explorando investigaciones previas sobre análisis de sentimiento y lenguas de bajos recursos. La Sección 3 describe los conjuntos de datos utilizados, incluyendo el corpus paralelo y el monolingüe etiquetado. En la Sección 4, se detalla el aumento de datos, seguido de la Sección 5, donde se presenta la metodología empleada. La Sección 6 aborda la experimentación y resultados obtenidos, con la Sección 7 se propone un prototipo inicial y, finalmente, en la Sección 8 y 9 se presentan las conclusiones y futuros trabajos.

Resumen:

- Objetivo 1: Crear un corpus monolingüe etiquetado de lenguaje Shipibo-Konibo.
    - Resultado 1.1 Método de creación del corpus monolingüe etiquetado de lenguaje Shipibo-Konibo.
    - Resultado 1.2 Base de datos para almacenar el corpus monolingüe etiquetado del lenguaje Shipibo-Konibo.
- Objetivo 2: Aumentar la data para entrenar modelos de sentimiento del lenguaje Shipibo-Konibo.
    - Resultado 2.1 Algoritmo para el aumento de la data para entrenar el modelo de sentimiento del lenguaje Shipibo-Konibo.
    - Resultado 2.2 Base de datos para almacenar el corpus aumentado.
- Objetivo 3: Diseñar un modelo de análisis de sentimiento para el lenguaje Shipibo-Konibo.
    - Resultado 3.1 Diseño del modelo de análisis de sentimiento neuronal para lenguaje Shipibo-Konibo.

    - Resultado 3.2 Modelo de red neuronal para el análisis de sentimiento para lenguaje Shipibo-Konibo.
- Objetivo 4: Implementar un prototipo de análisis de sentimiento del lenguaje Shipibo-Konibo.
    - Prototipo de análisis de sentimiento del lenguaje Shipibo-Konibo.

2. REVISION DE LITERATURA

La falta de recursos digitales y estudios específicos sobre el Shipibo-Konibo plantea un desafío importante para su análisis en el ámbito del procesamiento de lenguaje natural (PLN) [10]. Para evaluar el estado actual de las investigaciones en este ámbito, se llevó a cabo una revisión de literatura siguiendo un enfoque adaptado de la propuesta de Petersen, Feid Mujtaba y Mattson [12]. Este análisis incluyó búsquedas exhaustivas en bases de datos especializadas como SCOPUS e IEEE, resultando en una recopilación inicial de 233 artículos relevantes entre 2015 y 2024, usando las cadenas de búsqueda vistas en el Figura 2.

SCOPUS:

```
( TITLE-ABS-KEY ( "minority
language") OR TITLE-ABS-KEY (
amazoni* ) OR TITLE-ABS-KEY
( indige* ) OR TITLE-ABS-KEY (
shipibo ) AND TITLE-ABS-KEY
( "natural language processing") )
```

IEEE:

```
(("All Metadata":"natural
language processing") AND (("All
Metadata":"minority language")
OR ("All Metadata":amazoni*)
OR ("All Metadata":indige*)))
```

Figura 2. Cadena de Búsqueda en bases de datos

Posteriormente, se aplicaron dos etapas de filtrado. En la primera, se seleccionaron trabajos enfocados en lenguas de bajos recursos, priorizando aquellos relacionados con el PLN y lenguas indígenas; este proceso permitió reducir la selección inicial a 129 artículos. En la segunda, se evaluó su relevancia en función de su relación con el uso de técnicas o modelos para análisis de sentimiento, enfoques basados en lenguajes secundarios (como el español) para interpretar lenguajes principales (como lenguas indígenas), y trabajos enfocados en el PLN para Shipibo-Konibo, resultando en una selección final de 55 artículos.

De los artículos relacionados con lenguajes con baja cantidad de datos digitalizados, de los 55 la mayoría emplea técnicas de traducción automática (56.36%), lo que indica una tendencia hacia el uso de métodos de transferencia de conocimiento desde lenguajes con mayores recursos, Tabla 1.

Modelo	Porcentaje
Machine translation	56.36 %
BERT	7.27 %
Cross language	7.27 %
POS-tagging	3.64 %
Word2vec	3.64 %
Transformers	1.82 %
SVM, Random Forest	1.82 %
N-gram	1.82 %
Embeddings from Language Models	1.82 %
Transfer learning	1.82 %
Grammatical framework	1.82 %
CRF	1.82 %
Rethinking SDG-related scientific notion	1.82 %
Not-specified	7.27 %

Tabla 1. Modelos de mayor uso en lenguajes de bajos recursos.

Solo el 7.2% (4 artículos) de los trabajos seleccionados aborda directamente el análisis de sentimiento, cada uno enfocado en contextos diversos.

El primer estudio [13], aborda a Sudáfrica, con más de 60 millones de personas y más de 2,000 lenguas, con 11 oficiales, no logrando tener avances significativos en el análisis de sentimiento a diferencia del inglés o el chino, debido a la diversidad lingüística. Para resolver este problema, se propone la creación de un corpus etiquetado basado en modelos multilingües, con el inglés como idioma de referencia, utilizando herramientas como AFFIN, NRC y VADER para etiquetar datos en las lenguas Sepedi y Setswana. Se recopilaron más de 250,000 tuits entre estos 3 idiomas, de los cuales 40,000 fueron anotados manualmente. Un 63.4 % eran monolingües y el resto presentaba code-switching entre inglés y las lenguas africanas.

El segundo estudio [14], aborda la falta de recursos para el análisis de sentimiento en bosnio, presentando el primer léxico anotado en este idioma. Para evaluar su cobertura, se usaron dos corpus de referencia. La utilidad del léxico ya fue validada en un análisis de Twitter. Se aplicaron dos enfoques: uno basado en la frecuencia de las palabras y otro considerando todas las ocurrencias sin priorizar la frecuencia. Los resultados muestran una cobertura del 27.25% y 24.34% en cada corpus, respectivamente, demostrando la viabilidad del léxico para futuros estudios.

De manera similar, un tercer artículo [15], aborda la falta de análisis de sentimiento basado en aspectos (ABSA) para idiomas indígenas como el Afaan Oromoo, a diferencia de lenguas con más recursos como el inglés o el francés. Se recopilaron 2,800 reseñas de películas en YouTube y se etiquetaron manualmente en aspectos positivos y negativos, los modelos que se usaron fueron random forest, SVM, regresión logística y naïve Bayes con BoW y TF-IDF. Los mejores resultados alcanzaron hasta un 88% de precisión tras aplicar ajustes de hiper parámetros, para los siguientes trabajos toca enfocar el impacto en corpus con mayor cantidad de datos.

Finalmente, un cuarto estudio [16], evalúa la efectividad de un pipeline estándar de extracción de información en lenguas bantúes, que presentan gramática compleja y pocos recursos. Se analizaron 20,000 tuits en inglés y otros 20,000 en una combinación de inglés y seis lenguas bantúes. Los resultados mostraron que estos idiomas requieren pasos adicionales, como análisis morfológico y etiquetado gramatical. Además, el enfoque en análisis de sentimiento debe realizarse sin la lematización para evitar la pérdida de morfemas de negación, dando así varios trabajos a futuro en lenguas bantúes.

Viendo estos resultados se revela que ninguno de ellos aborda lenguas indígenas amazónicas, ni mucho menos el Shipibo-Konibo.

Ante este panorama, nuestro trabajo propone una contribución única y necesaria: desarrollar un análisis de sentimiento específico para el Shipibo-Konibo, si bien encontramos paralelismos con el estudio del sur de África [13]], nuestro enfoque se adapta a una realidad completamente distinta, ya que los datos provienen de textos indígenas en lugar de fuentes digitales como tuits. Además, el español actúa como un puente lingüístico clave en nuestro análisis, permitiéndonos abordar las limitaciones inherentes de los recursos disponibles.

Además, la ausencia de investigaciones específicas en el análisis de sentimiento para el Shipibo-Konibo subraya la necesidad de desarrollar modelos y herramientas adaptadas a las características únicas de este lenguaje. En este contexto, modelos como BERT [17], ampliamente utilizados para abordar desafíos en contextos multilingües y en code-switching [18], podrían servir como una base sólida. Su capacidad para manejar datos donde coexisten múltiples idiomas dentro de un mismo enunciado resulta especialmente relevante para lenguas como el Shipibo-Konibo, que a menudo se relacionan con el español en su uso práctico.

De esta manera, este trabajo busca llenar una brecha importante en el PLN, aportando herramientas para lenguas indígenas amazónicas que hasta ahora han sido escasamente exploradas.

3. CONJUNTO DE DATOS

En esta sección se describe el conjunto de datos utilizado, compuesto por 26 269 oraciones en Shipibo-Konibo y sus respectivas traducciones al español, corpus obtenido de trabajos previos del grupo Chana [19]. El corpus está conformado por oraciones de tres dominios temáticos—flashcards, educación y religión—y, para realizar el análisis de sentimiento, necesitamos que cada oración esté clasificada como negativa (NEG), positiva (POS) o neutra (NEU).

Debido a que el corpus consta de 26 269 oraciones, para evitar un etiquetado manual completo, se aplicó:

- Selección y anotación manual de 3 000 oraciones en español del corpus, con etiquetas de sentimiento (positivo, negativo, neutral).
- Entrenamiento de un modelo de sentimiento basado en estas 3 000 instancias.
- Etiquetado automático de las 23 269 oraciones restantes mediante el modelo de sentimiento entrenado.
- Dado que cada oración en Shipibo-Konibo posee su traducción al español, se transfirieron las etiquetas al lado Shipibo-Konibo, generando así un corpus monolingüe anotado en la lengua objetivo.

El resultado es un recurso de 26 269 oraciones Shipibo-Konibo etiquetadas, empleado como corpus base experimental en este trabajo.

En la Tabla 2 podemos ver la suma total de oraciones para cada etiqueta (NEG, POS, NEU), sin distinción por tema, para tener una vista general del balance de los sentimientos en todo el conjunto de datos.

LABEL	TOTAL
NEG	10060
POS	5300
NEU	10909

Tabla 2. Cantidad de oraciones por etiqueta.

Dentro de la información obtenida, en la Tabla 3 podemos ver la data etiquetada por cada tema que se tomó del corpus y en la Tabla 4 a nivel de porcentajes.

LABEL	FLASHCARD	EDUCACIÓN	RELIGIOSO
NEG	3873	646	5541
POS	1571	817	2912
NEU	2770	4536	3603

Tabla 3. Cantidad de oraciones por etiqueta.

LABEL	FLASHCARD(%)	EDUCACIÓN(%)	RELIGIOSO (%)
NEG	44.8	8.4	44.9
POS	29.6	15.3	55.1
NEU	25.6	76.3	45.0

Tabla 4. Porcentaje de etiquetas dentro de cada tema.

4. AUMENTO DE DATOS

En esta sección se describe que el aumento de datos es esencial debido a la limitada disponibilidad de información en Shipibo-Konibo, lo que dificulta la construcción de corpus representativos y afecta la capacidad de los modelos para generalizar. Sin suficientes datos, los modelos no pueden captar las complejidades del lenguaje, limitando su efectividad en tareas como el análisis de sentimiento.

Como estrategia, se propone generar datos mediante la introducción controlada de errores en las frases, simulando variaciones reales como errores ortográficos o gramaticales [9]. Esto permite entrenar modelos más robustos, capaces de adaptarse a datos imperfectos y reflejar mejor el uso práctico del idioma.

Siendo las técnicas utilizadas:

Error Aleatorio:

Se aplican modificaciones aleatorias a las palabras, como la inserción, eliminación, sustitución o intercambio de caracteres [9].

Ejemplo: Error Aleatorio (inserción)
Entrada: *rabikaayamai*
Salida: *rabikaa**n**yamai*

Proximidad de Teclado:

Este enfoque de error se basa en la observación de que, al escribir en un teclado, es común que una persona presione accidentalmente una tecla adyacente a la que pretendía pulsar. Para simular este tipo de errores, se utiliza una estructura de datos tipo diccionario, donde cada tecla se asocia con una lista de teclas vecinas según su disposición física en el teclado, permitiendo generar variaciones realistas en los textos [9].

Ejemplo: Proximidad de Teclado
Entrada: *rabikaaayamai* **Salida:** *rabikasayamai*

Ambigüedad Fonema-Grafema:

Este enfoque de error parte de la observación de que ciertos grafemas tienen una pronunciación similar, lo que puede llevar a confusiones al momento de deletrear. Con la colaboración del grupo Chana y de lingüistas especializados [9], se construye una estructura de datos tipo diccionario, donde cada grafema se asocia con una lista de grafemas que podrían ser confundidos, permitiendo generar variaciones que reflejan este tipo de errores comunes.

Ejemplo: Ambigüedad Fonema-Grafema
Entrada: *rabikaayamai* **Salida:** *rabicayamai*

Similitud de Sílabas

Este enfoque de error se basa en la observación de que, dentro de una lengua, ciertas sílabas tienen similitudes que pueden llevar a confusiones al deletrear. A partir del método de silabificación desarrollado por el grupo Chana apoyado con lingüistas para Shipibo-Konibo [11], se identifican las sílabas y se construye una estructura de datos tipo diccionario, donde cada sílaba se asocia con una lista de sílabas similares, permitiendo simular errores de este tipo de manera realista.

Ejemplo: Similitud de Sílabas
Entrada: *rabikaayamai* **Salida:** *jabikaayamai*

Los resultados obtenidos para cada tipo de error se resumen en la sección de resultados.

5. MÉTODOS Y PROCEDIMIENTOS

En esta sección veremos los métodos y procedimientos propuestos para este trabajo que se divide en tres etapas principales: creación del corpus, aumento de datos y diseño del modelo de análisis de sentimiento. Este enfoque permite abordar los desafíos específicos para el lenguaje Shipibo-Konibo y aprovechar técnicas avanzadas de PLN para construir un modelo funcional.

5.1. Creación de un corpus monolingüe etiquetado de lenguaje Shipibo-Konibo.

El desarrollo del modelo de análisis de sentimiento se inicia con la recopilación de un corpus no etiquetado en Shipibo-Konibo, la cual fue realizada en otra investigación apoyada por el grupo Chana, que se preocupa por la preservación de las lenguas amazónicas [19]. Este se complementa con un corpus paralelo en español, aprovechando la relación lingüística entre ambos idiomas para simplificar y enriquecer el proceso de etiquetado.

A través de un etiquetado manual de aproximadamente 3000 muestras, recolectando 1000 oraciones de cada dominio temático—flashcards, educación y religión—se crea un conjunto de datos en español que sirve como base para el entrenamiento del modelo, en la Tabla 5. Estos datos son luego preprocesados, aplicando normalización, eliminación de ruido y conversión a representaciones computacionales mediante embeddings, asegurando su adecuación para el análisis.

LABEL	FLASHCARD	EDUCACIÓN	RELIGIOSO
NEG	194	29	228
POS	178	172	188
NEU	628	799	584

Tabla 5. Cantidad de oraciones etiquetadas manualmente.

Para determinar el modelo más adecuado para las primeras pruebas en español, se evaluaron dos opciones principales:

- **dccuchile/bert-base-spanish-wwm-uncased**: Un modelo basado en BERT, entrenado con whole word masking en español. Es ampliamente reconocido por su versatilidad y robustez en tareas de procesamiento de lenguaje en español [20].
- **pysentimiento/robertuito-sentiment-analysis:** Un modelo especializado en el análisis de sentimientos, diseñado específicamente para capturar emociones y matices sentimentales en español con mayor precisión [21].

Ambos modelos fueron evaluados utilizando el corpus etiquetado en español. Los resultados mostraron un desempeño similar en términos de precisión, con

dccuchile/bert-base-spanish-wwm-uncased [20] alcanzando un 77% y **pysentimiento/ robertuito-sentiment-analysis** [21] un 76%.

Sin embargo, se optó por **pysentimiento** debido a su especialización en tareas de análisis de sentimiento, lo que lo posiciona como una opción más adecuada para captar emociones y polaridades en este contexto [21].

Este modelo permitió establecer un punto de partida sólido para transferir las etiquetas generadas al resto del corpus en español. Una vez completado el etiquetado, se integraron los datos al corpus paralelo en Shipibo-Konibo, sentando las bases para el desarrollo de un modelo base inicial adaptado a este idioma. Los parámetros utilizados en el modelo base en español se detallan en el Apéndice A.

5.2. Aumento de data para el corpus Shipibo-Konibo.

Con el corpus monolingüe etiquetado de Shipibo-Konibo, se entrena un modelo base para evaluar el desempeño en la tarea de análisis de sentimiento, alcanzando una precisión del 58.8%; se usó el modelo LSTM y se separó 80-20% para los datos de entrenamiento y prueba, en el Apéndice B podemos ver los parámetros usados para el modelo. Este resultado destaca la necesidad de mejorar la cobertura y robustez del corpus.

Corpus Español & Shipibo-Konibo, Negativo:
Español: *decimos "no me gusta" cuando nos hacen daño con palabras o golpes* **Label ESP:** *0 (Negativo)* **Shipibo:** *nonra yoiyai " enra costanyamake" jawetianki noa jakomaakanai join iamax timakan* **Label SHI:** *0 (Negativo)*

Corpus Español & Shipibo-Konibo, Positivo:
Español: *feliz el que tome parte en el* banquete del reino de dios **Label ESP:** *1 (Positivo)* **Shipibo:** *ja diossen ikinaton jatinkoxon icha pitiakana jato betan tsinkíxon jawékiaibora kikinbires raroshamanbo ikanti iki aki* **Label SHI:** *1 (Positivo)*

Corpus Español & Shipibo-Konibo, Neutro:
Español*: escribe la lista de tareas para los grupos* **Label ESP:** *2 (Neutro)* **Shipibo:** *wishawe ja tee akantibo* yoiya ja tsamabaon **Label SHI:** *2 (Neutro)*

Para lograrlo, se implementan técnicas de aumento de datos, detalladas en la Sección 4, que incluyen generación de datos sintéticos y ampliación del vocabulario, con el objetivo de abordar la falta de recursos en Shipibo-Konibo. Este proceso incrementa significativamente la diversidad y cantidad de ejemplos disponibles, mejorando la capacidad del modelo para generalizar.

5.3. Diseño de un modelo de análisis de sentimiento para el lenguaje Shipibo-Konibo.

Con los resultados del aumento de datos, el corpus enriquecido se utilizó para diseñar y mejorar el modelo de análisis de sentimiento. Para las representaciones vectoriales, se emplearán embeddings generados con modelos pre entrenados, seleccionados por su capacidad de trabajar con múltiples idiomas y tareas específicas de similitud semántica, las opciones consideradas incluyen:

- **XLM-Roberta**: Elegido por su capacidad de generalización multilingüe y soporte para lenguajes de bajo recurso [22].
- **LaBSE:** Especializado en búsqueda semántica multilingüe, optimizando la representación de frases en Shipibo-Konibo [23].
- **SIMCSE Roberta large y SIMCSE Bert:** Modelos optimizados para tareas de similitud textual [24].

El diseño del modelo principal se basó en una arquitectura de redes LSTM, configurada para clasificar oraciones en tres categorías: positivo, negativo y neutral. Para evaluar su desempeño, el corpus fue dividido en un 80% para entrenamiento y un 20% para pruebas, garantizando una evaluación balanceada y confiable.

Durante las pruebas, XLM-Roberta mostró ser el modelo con menor tendencia a sobre entrenarse, incluso en escenarios con datos limitados. Esto permitió obtener un equilibrio entre precisión en los datos de entrenamiento y prueba, siendo el modelo final para embeddings.

6. EXPERIMENTACION Y RESULTADOS

En esta sección se presentan las experimentaciones realizadas para evaluar y optimizar el modelo de análisis de sentimiento en Shipibo-Konibo. A partir de un corpus inicial derivado del español, el cual se obtuvo en la Sección 5.1, este modelo base alcanzó una precisión del 58.8%, se puede ver más detalles del modelo en el apéndice B. Con el objetivo de mejorar esta métrica, se llevaron a cabo experimentos en tres etapas principales: aumento de datos, evaluación de modelos de embeddings y optimización con el embedding seleccionado.

Considerar que todos los modelos fueron ejecutados en una proporción de 80-20% respectivamente entre entrenamiento y validación, además de que esta división fue

hecha de manera aleatoria sin considerar ningún peso para algún dominio del corpus.

6.1. Pruebas de Aumento de Datos, Etapa I

El propósito de esta etapa es el aumento de datos mencionados en la Sección 4 para el modelo base.

EXPERIMENTO I:

Se implementaron estrategias de aumento de datos introduciendo errores sintéticos como:

- Error aleatorio
- Proximidad de teclado
- Ambigüedad fonema-grafema
- Similitud de sílabas

RESULTADOS:

Los resultados muestran que ciertos errores, como Fonema-Grafema, mejoran consistentemente la precisión en entrenamiento y validación en comparación con los demás tipos de errores; pero aun así la mejora no es significativa en comparación con la precisión del modelo base de Shipibo-Konibo, en la Tabla 6 se observan los resultados obtenidos.

Errores	**Corpus**	**Etiquetas**			**Accuracy (%)**	
		NEG	**NEU**	**POS**	**Train**	**Val**
Shipibo-Konibo Modelo Base	26K	10060	10909	5300	76.5 %	58.8 %
Aleatorio	31K	12442	12164	6663	82.6 %	58.3 %
Proximidad	31K	12442	12164	6663	82.5 %	57.9 %
Fonema-Grafema	31K	12442	12164	6663	82.9 %	59.2 %
Similitud de Sílabas	31K	12442	12164	6663	82.8 %	57.7 %
Aleatorio, Proximidad	36K	14824	13419	8026	81.2 %	56.6 %
Aleatorio, Fonema-Grafema	36K	14824	13419	8026	81.1 %	57.5 %
Aleatorio, Similitud de Sílabas	36K	14824	13419	8026	81.3 %	56.4 %
Proximidad, Fonema-Grafema	36K	14824	13419	8026	81.5 %	57.2 %
Proximidad, Similitud de Sílabas	36K	14824	13419	8026	81.8 %	55.8 %
Fonema-Grafema, Similitud de Sílabas	36K	14824	13419	8026	82.3 %	57.0 %
Todos los errores	46K	19588	15929	10752	78.9 %	52.2 %

Tabla 6. Aumento de datos basado en errores. Experimento I.

6.2. Pruebas de Modelos de Embeddings, Etapa II

El propósito de esta etapa es desarrollar un modelo de embeddings robusto, capaz de adaptarse eficazmente al idioma Shipibo-Konibo, minimizando el riesgo de sobre entrenamiento y garantizando un rendimiento equilibrado.

EXPERIMENTO II:

Se evaluaron diferentes modelos de embeddings pre entrenados para determinar cuál era el más adecuado para el análisis de sentimiento en Shipibo-Konibo. Los modelos considerados incluyeron a XLM-Roberta, LaBSE, SIMCSE Roberta Large y SIMCSE Bert. Cada modelo fue probado utilizando el corpus base en Shipibo-Konibo. Además, se incorporaron 5000 muestras generadas mediante técnicas de aumento de datos, introduciendo errores del tipo aleatorio, en el Apéndice podemos ver los parámetros usados para los modelos.

RESULTADOS:

De los resultados mostrados en la Tabla 7, el modelo XLM-Roberta destacó, esto gracias a su capacidad para equilibrar la precisión en las fases de entrenamiento y validación, mostrando un menor nivel de sobre entrenamiento en comparación con los demás modelos. Este resultado subraya la eficacia de los embeddings multilingües para capturar relaciones semánticas complejas en lenguas indígenas, haciéndolo especialmente adecuado para Shipibo-Konibo.

Errores	**Corpus**	**Accuracy (%)**	
		Train	**Val**
XLM-Roberta	**31K**	**60.0 %**	**55.2 %**
LaBSE	31K	83.0 %	56.0 %
SIMCSE Roberta large	31K	75.8 %	55.6 %
SIMCSE Bert	31K	74.6 %	55.5 %

Tabla 7. Evaluación de modelos de embeddings. Experimento II.

6.3. Pruebas con Modelo Embedding Seleccionado, Etapa III

El propósito de esta etapa es mejorar el modelo de entrenamiento ya habiendo seleccionado a XLM-Roberta como el modelo de embeddings. Para lograr este propósito, se diseñaron los experimentos III y IV para optimizar aún más la precisión del modelo:

EXPERIMENTO III:

El propósito de este experimento fue evaluar el impacto del aumento masivo de datos en el rendimiento del modelo de análisis de sentimiento en Shipibo-Konibo. Para ello, se generaron 20,000 muestras únicas por cada tipo de error previamente

definido: error aleatorio, proximidad de teclado, ambigüedad fonema-grafema y similitud de sílabas. Cada conjunto de datos fue utilizado para entrenar un modelo independiente, permitiendo analizar de manera específica el efecto de cada tipo de error. Los parámetros empleados en cada modelo se detallan en el Apéndice E.

El objetivo principal fue determinar cómo la inclusión de datos masivos con diferentes tipos de errores influye en la capacidad del modelo para generalizar y reconocer patrones lingüísticos propios del idioma Shipibo-Konibo.

RESULTADOS:

Los resultados indicaron una mejora significativa en la precisión del modelo tanto en entrenamiento como en validación. En particular:

Los errores de Ambigüedad Fonema-Grafema y Proximidad de Teclado proporcionaron los mayores beneficios en términos de precisión, ya que capturan patrones frecuentes y consistentes de confusión lingüística en Shipibo-Konibo. El aumento masivo de datos permitió al modelo manejar mejor la variabilidad y los errores introducidos, lo que mejoró su capacidad de generalización. El análisis detallado de los resultados se encuentra en la Tabla 8, donde se observan que no hay mucha diferencia en los incrementos de precisión en comparación con el corpus base.

Errores	Corpus	Etiquetas			Accuracy (%)	
		NEG	NEU	POS	train	val
Shipibo-Konibo Modelo Base	26K	10060	10909	5300	76.5 %	58.8 %
Aleatorio	46K	19127	17577	9565	62.7 %	54.5 %
Proximidad	46K	19127	17577	9565	70.8 %	55.5 %
Fonema-Grafema	46K	19127	17577	9565	75.9 %	56.2 %
Similitud de Sílabas	46K	19127	17577	9565	74.5 %	53.6 %

Tabla 8. Aumento de datos de manera masiva basado en XLM-Roberta, exceptuando el Modelo Base. Experimento III.

EXPERIMENTO IV:

En este experimento se diseñó un enfoque progresivo y sistemático para aumentar los datos del corpus, utilizando combinaciones escalonadas de errores. El procedimiento fue el siguiente:

Incrementos iniciales de 5000 muestras por error, abarcando los cuatro tipos de errores, error aleatorio, proximidad de teclado, ambigüedad fonema-grafema y similitud de Sílabas, todos usan las mismas muestras modificadas para cada error. Posteriormente, las muestras de cada error fueron combinadas de manera balanceada, aumentando gradualmente la cantidad total de datos hasta alcanzar 20

000 muestras combinadas, en el Apéndice F podemos ver los parámetros usados para cada modelo.

El objetivo de este método era analizar cómo el equilibrio entre los diferentes tipos de errores impacta el rendimiento del modelo, comparándolo con el enfoque masivo del Experimento III. Este enfoque buscaba reducir posibles sesgos en el modelo derivados de un solo tipo de error, explorar si una combinación más balanceada y escalonada de datos generaba un efecto acumulativo positivo en el aprendizaje del modelo.

El análisis detallado de los resultados se encuentra en la Tabla 9.

RESULTADOS:

Los resultados demostraron que este método fue más efectivo por las siguientes razones:

- Reducción de la varianza: El enfoque balanceado logró disminuir las discrepancias entre las métricas de entrenamiento y validación, lo que indica una mejora en la capacidad del modelo para generalizar a datos no vistos.
- Mayor estabilidad del modelo: La incorporación escalonada y balanceada de errores permitió que el modelo aprendiera de manera más uniforme y progresiva, evitando que un solo tipo de error dominara el proceso de entrenamiento.
- Efecto acumulativo positivo: La mezcla gradual de datos permitió al modelo adaptarse mejor a la diversidad de errores, en lugar de recibir un aumento masivo de datos de un solo tipo de error de forma abrupta, como en el Experimento III.

Errores	**Corpus**	**Etiquetas**			**Accuracy (%)**	
		NEG	**NEU**	**POS**	**train**	**val**
Shipibo-Konibo Modelo Base	26K	10060	10909	5300	76.5 %	58.8 %
Aleatorio	31K	12442	12164	6663	64.3 %	55.6 %
Proximidad	31K	12442	12164	6663	75.8 %	55.5 %
Fonema-Grafema	31K	12442	12164	6663	81.4 %	55.3 %
Similitud de Sílabas	31K	12442	12164	6663	84.0 %	54.2 %
Aleatorio, Proximidad	36K	14824	13419	8026	82.3 %	65.4 %
Aleatorio, Fonema-Grafema	36K	14824	13419	8026	86.3 %	62.3 %
Aleatorio, Similitud de Sílabas	36K	14824	13419	8026	87.3 %	60.9 %
Proximidad, Fonema-Grafema	36K	14824	13419	8026	88.4 %	61.3 %
Proximidad, Similitud de Sílabas	36K	14824	13419	8026	89.6 %	60.2 %
Fonema-Grafema, Similitud de Sílabas	36K	14824	13419	8026	90.6 %	61.9 %
Aleatorio, Proximidad, Fonema-Grafema, Similitud de Sílabas	**46K**	**19588**	**15929**	**10752**	**87.6 %**	**74.4 %**

Tabla 9. Aumento de datos de manera escalonada basado en XLM-Roberta, exceptuando el Modelo Base. Experimento IV.

7. PROTOTIPO DE ANÁLISIS DE SENTIMIENTO PARA SHIPIBO-KONIBO

Con el objetivo de abordar la carencia de herramientas tecnológicas orientadas al análisis de sentimiento en lenguas amazónicas peruanas, se desarrolló un prototipo de análisis de sentimiento enfocado en el idioma Shipibo-Konibo. Este prototipo utiliza el modelo del Experimento IV, el cual obtuvo el mejor resultado obtenido y procesado durante esta investigación, Figura 3.

Figura 3. Prototipo de análisis de sentimiento para Shipibo-Konibo

El prototipo fue implementado como una aplicación web utilizando el framework Flask [25]. La interfaz permite al usuario ingresar un texto en Shipibo-Konibo, una vez ingresado, el texto es enviado al backend de la aplicación, donde se procesa para determinar la polaridad del sentimiento (positivo, negativo o neutral). El sistema devuelve como resultado el texto analizado junto con la categoría de sentimiento asignada. Este prototipo inicial representa un paso importante hacia la integración de herramientas de procesamiento de lenguaje natural en lenguas amazónicas, contribuyendo no solo a su revitalización sino también a su aplicación práctica en contextos educativos, culturales y sociales.

8. CONCLUSIONES

En los experimentos realizados, el modelo base alcanzó una precisión inicial del 58.8% en validación. Sin embargo, tras la incorporación de XLM-Roberta como modelo de embeddings y la aplicación de un aumento de datos escalonado y balanceado, se logró una mejora significativa, alcanzando una precisión final del 74.4%. Este avance refleja la efectividad de los embeddings multilingües y de las estrategias de aumento progresivo para abordar los desafíos del análisis de sentimiento en lenguas indígenas como el Shipibo-Konibo.

El enfoque balanceado permitió reducir la variación entre las métricas de entrenamiento y validación, favoreciendo una mayor capacidad de generalización del modelo. Este resultado subraya la importancia de combinar estrategias modernas en PLN con un diseño cuidadoso de los datos, adaptado a las particularidades lingüísticas del idioma.

Más allá de la mejora cuantitativa del 58.8% al 74.4% de precisión, este trabajo demuestra que los métodos de PLN de última generación pueden adaptarse con éxito a lenguas escasamente representadas. Al validar que los embeddings multilingües (XLM-RoBERTa) y el aumento de datos escalonado reducen el sesgo y la varianza, ofrecemos una ruta replicable para otras comunidades lingüísticas que enfrentan carencia de datos.

Esto acorta la brecha tecnológica entre los idiomas dominantes y las lenguas indígenas, fomentando la creación de aplicaciones —como chatbots, asistentes de voz o sistemas de alerta temprana— que funcionen en el idioma de la comunidad y no requieran traducción intermedia.

Finalmente, los experimentos resaltaron la necesidad de integrar datos adicionales provenientes de hablantes nativos y contextos reales para seguir fortaleciendo el modelo, ampliando su aplicabilidad en escenarios prácticos y su contribución al procesamiento de lenguas indígenas.

9. TRABAJOS FUTUROS

Como futuros trabajos, se debe buscar ampliar el corpus con datos de hablantes nativos y contextos reales, incorporando variación dialectal y metadatos. También se recomienda implementar etiquetado variado como emociones de miedo, alegría, tristeza, etc.; y basarse en datos que contengan información suficiente para identificar una emoción. Se sugiere evaluar robustez y equidad por dominio, dialecto, longitud y ruido, con análisis de errores accionables. Además, optimizar modelos de aplicaciones prácticas (chatbots, asistentes de voz, alarmas) para entornos con recursos limitados, así ampliando la investigación en lenguajes de bajos recursos.

REFERENCIAS BIBLIOGRÁFICAS

[1] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pages 79–86. Association for Computational Linguistics.

[2] Shah, S.R., and Kaushik, A. (2019). Sentiment Analysis On Indian Indigenous Languages: A Review On Multilingual Opinion Mining. ArXiv, abs/1911.12848.

[3] Diccionario Shipibo-Castellano, Serie Lingüística Peruana N31, Editorial Mary Ruth Wise, Recopiladores: J. Loriot, E. Lauriault, D. Day, Ministerio de Educación Instituto Lingüístico de Verano Peru 1993, 2da Edición 2008.

[4] INSTITUTO NACIONAL DE ESTADÍSTICA E INFORMÁTICA (INEI) (2017) Censos Nacionales 2017: XII de Población, VII de Vivienda y III de Comunidades nativas y comunidades campesinas. Lima: Instituto Nacional de Estadística e Informática (INEI).

[5] Organización de las Naciones Unidas, Reporte Inteligencia Artificial centrada en los Pueblos Indígenas: Perspectivas desde América Latina y el Caribe, 2023.

[6] Anna Grøndahl Larsen, Asbjørn Følstad, The impact of chatbots on public service provision: A qualitative interview study with citizens and public service providers, Government Information Quarterly, Volumen 41, Issue 2, 2024, 101927, ISSN 0740-624X, https://doi.pucp.elogim.com 0.1016/j.giq.2024.101927.

[7] Jhan, J., C. Liu, S. Jeng, et al. Cheerbots: Chatbots toward empathy and emotionusing reinforcement learning. CoRR, abs/2110.03949, 2021.

[8] Yu, S., Kulkarni, N., Lee, H., & Kim, J. (2017). Syllable-level Neural Language Model for Agglutinative Language. SWCN@EMNLP. https://doi.org/10.48550/arXiv.1708.05515

[9] Arturo Oncevay, Gerardo Cardoso, Carlo Alva, César Lara Ávila, Jovita Vásquez Balarezo, Saúl Escobar Rodríguez, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Juan López Bautista, Nimia Acho Rios, Remigio Zapata Cesareo, Héctor Erasmo Gómez Montoya, and Roberto Zariquiey. 2022. SchAman: Spell-Checking Resources and Benchmark for Endangered Languages from Amazonia. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 411–417, Online only. Association for Computational Linguistics.

[10] Challenges of language technologies for the indigenous languages of the Americas, August 2018, Conference: 27th International Conference on Computational Linguistics (COLING 2018) Santa Fe, New Mexico, USA.

[11] Carlo Alva and Arturo Oncevay. 2017. Spell-Checking based on Syllabification and Character-level Graphs for a Peruvian Agglutinative Language. In Proceedings of the First Workshop on Subword and Character Level Models in NLP, pages 109–116, Copenhagen, Denmark. Association for Computational Linguistics.

[12] K. Petersen, R. Feldt, Mujtaba, S. and M. Mattsson, “Systematic Mapping Studies in Software Engineering”, EASE'08 Proceedings of the 12th international conference on Evaluation and Assessment in Software Engineering, Bari, Italy, June 2008.

[13] Mabokela, Koena, Schlippe, Tim. (2022). A Sentiment Corpus for South African Under-Resourced Languages in a Multilingual Context.

[14] Sead Jahi, Jernej Vičič, Annotated Lexicon for Sentiment Analysis in Bosnian Language, Faculty of Mathematics, Natural Science and Information Technologies, University of Primorska, Koper, Slovenia, The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing, https://ceur-ws.org/Vol-3315/

[15] Horsa, Obsa Gelchu, Tune, Kula Kekeba, Aspect-Based Sentiment Analysis for Afaan Oromoo Movie Reviews Using Machine Learning Techniques, Applied Computational Intelligence and Soft Computing, 2023, 3462691, 12 pages, 2023.https://doi.org/10.1155/2023/3462691

[16] Nchabeleng M, Byamugisha J. Evaluating the Effectiveness of the Standard Insights Extraction Pipeline for Bantu Languages. Advances in Information Retrieval. 2020 Mar 17; 12035:159–72. doi: 10.1007/978-3-030-45439-5_11. PMCID: PMC7148238.

[17] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 10.48550/arXiv.1810.04805.

[18] Robert Pugh, Francis Tyers, The ITML Submission to the IberLEF2023 Shared Task on Guarani-Spanish Code Switching Analysis, Department of Linguistics, Indiana University, Bloomington, IN, U.S.A, https://ceur-ws.org/Vol-3496/

[19] Shipibo-Konibo Data, Grupo Chana, La Pontificia Universidad Católica del Perú (PUCP), la Universidad de Zurich y el Instituto Max Planck para la Antropología Evolutiva.

[20] Cañete, José and Chaperon, Gabriel and Fuentes, Rodrigo and Ho, Jou-Hui and Kang, Hojin and Pérez, Jorge, inproceedings Canete CFP 2020, Spanish Pre-Trained BERT Model and Evaluation Data, PML4DC at ICLR 2020

[21] Pérez, Juan & Giudici, Juan & Luque, Franco. (2021). pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. 10.48550/arXiv.2106.09462.

[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.747 (Conneau et al., ACL 2020)

[23] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

[24] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[25] Welcome to Flask — Flask Documentation (2.0.x).

APÉNDICE

Apéndice A

Modelo base para corpus en español, para la implementación del sistema, utilizamos para los embeddings el modelo pysentimiento/robertuito-sentiment-analysis y para el entrenamiento utilizamos un modelo de red neuronal profunda con 50 épocas, con los siguientes hiper-parámetros:

CAPA	TIPO	NEURONAS
Entrada	Densa	768
Capa Densa 1	Densa	256
Dropout 1	Dropout	-
Capa Densa 2	Densa	64
Dropout 2	Dropout	-
Salida	Densa	8
CAPA	ACTIVACION	REGULARIZACION
Entrada	-	-
Capa Densa 1	ReLU	L2 (λ=0.01)
Dropout 1	-	Dropout (20 %)
Capa Densa 2	ReLU	L2 (λ=0.01)
Dropout 2	-	Dropout (20 %)
Salida	Softmax	-
Tamaño del batch: 16 Pérdida: Entropía cruzada categórica dispersa Optimizador: Adam		

Apéndice B

Modelo base para corpus en shipibo-konibo, para la implementación del sistema, utilizamos para los embeddings se usó TextVectorization() y para el entrenamiento utilizamos el modelo neuronal LSTM con 50 épocas, con los siguientes hiper-parámetros:

CAPA	TIPO	NEURONAS
Entrada	Densa	768
LSTM	LSTM	256
Capa Densa 1	Densa	128
Dropout 1	Dropout	-
Capa Densa 2	Densa	64
Dropout 2	Dropout	-
Salida	Densa	3
CAPA	ACTIVACION	REGULARIZACION
Entrada	-	-
Capa Densa 1	ReLU	L2 (λ=0.01)
Dropout 1	-	Dropout (20 %)
Capa Densa 2	ReLU	L2 (λ=0.01)
Dropout 2	-	Dropout (20 %)
Salida	Softmax	-
Tamaño del batch: 32 Pérdida: Entropía cruzada categórica Optimizador: Adam		

Apéndice C

Modelo base shipibo-konibo con aumento de datos, para la implementación del sistema, utilizamos para los embeddings se usó TextVectorization() y para el entrenamiento utilizamos el modelo neuronal LSTM con 50 épocas, con los siguientes hiper-parámetros:

CAPA	TIPO	NEURONAS
Entrada	Densa	768
LSTM	LSTM	256
Capa Densa 1	Densa	128
Dropout 1	Dropout	-
Capa Densa 2	Densa	64
Dropout 2	Dropout	-
Salida	Densa	3

CAPA	ACTIVACION	REGULARIZACION
Entrada	-	-
Capa Densa 1	ReLU	L2 (λ=0.01)
Dropout 1	-	Dropout (20 %)
Capa Densa 2	ReLU	L2 (λ=0.01)
Dropout 2	-	Dropout (20 %)
Salida	Softmax	-

Tamaño del batch: 32
Pérdida: Entropía cruzada categórica
Optimizador: Adam

Apéndice D

Modelo base shipibo-konibo evaluando embeddings, para la implementación del sistema, se probó con los embeddings XLM-Roberta, LaBSE, SIMCSE Roberta Large y SIMCSE Bert y para el entrenamiento utilizamos el modelo neuronal LSTM con 50 épocas, con los siguientes hiper-parámetros:

XLM-Roberta y LaBSE

CAPA	TIPO	NEURONAS
Entrada	Densa	768
LSTM	LSTM	256
Capa Densa 1	Densa	128
Dropout 1	Dropout	-
Capa Densa 2	Densa	64
Dropout 2	Dropout	-
Salida	Densa	3

CAPA	ACTIVACION	REGULARIZACION
Entrada	-	-
Capa Densa 1	ReLU	L2 (λ=0.01)
Dropout 1	-	Dropout (20 %)
Capa Densa 2	ReLU	L2 (λ=0.01)
Dropout 2	-	Dropout (20 %)
Salida	Softmax	-

Tamaño del batch: 32
Pérdida: Entropía cruzada categórica dispersa
Optimizador: Adam

SIMCSE Roberta Large y SIMCSE Bert

CAPA	TIPO	NEURONAS
Entrada	Densa	1024
LSTM	LSTM	512
Capa Densa 1	Densa	256
Dropout 1	Dropout	-
Capa Densa 2	Densa	128
Dropout 2	Dropout	-
Salida	Densa	3
CAPA	ACTIVACION	REGULARIZACION
Entrada	-	-
Capa Densa 1	ReLU	L2 (λ=0.01)
Dropout 1	-	Dropout (20 %)
Capa Densa 2	ReLU	L2 (λ=0.01)
Dropout 2	-	Dropout (20 %)
Salida	Softmax	-
Tamaño del batch: 32 Pérdida: Entropía cruzada categórica dispersa Optimizador: Adam		

Apéndice E

Modelo base shipibo-konibo con XLM-Roberta evaluando aumento de datos masivo, para la implementación del sistema, se usó para embeddings el modelo XLM-Roberta y para el entrenamiento utilizamos el modelo neuronal LSTM con 75 épocas, para el aumento de data se usaron 20 000 muestras de cada error (Error Aleatorio, Proximidad de Teclado, Ambigüedad Fonema-Grafema y Similitud de Sílabas) y fueron evaluados independientemente cada uno con los siguientes hiper-parámetros:

CAPA	TIPO	NEURONAS
Entrada	Densa	768
LSTM	LSTM	256
Capa Densa 1	Densa	128
Dropout 1	Dropout	-
Capa Densa 2	Densa	64
Dropout 2	Dropout	-
Salida	Densa	3
CAPA	ACTIVACION	REGULARIZACION
Entrada	-	-
Capa Densa 1	ReLU	L2 (λ=0.01)
Dropout 1	-	Dropout (20 %)
Capa Densa 2	ReLU	L2 (λ=0.01)
Dropout 2	-	Dropout (20 %)
Salida	Softmax	-
Tamaño del batch: 32 Pérdida: Entropía cruzada categórica dispersa Optimizador: Adam		

Apéndice F

Modelo base shipibo-konibo con XLM-Roberta evaluando aumento de datos progresivo,} Para la implementación del sistema, se usó para embeddings el modelo XLM-Roberta y para el entrenamiento utilizamos el modelo neuronal LSTM con 75 épocas, para el aumento de data se usaron muestras de cada error (Error Aleatorio, Proximidad de Teclado, Ambigüedad Fonema-Grafema y Similitud de Sílabas) desde 5000 muestras hasta 20 000 aumentando progresivamente cada una de las muestras y fueron evaluados con los siguientes hiper-parámetros:

CAPA	TIPO	NEURONAS
Entrada	Densa	768
LSTM	LSTM	256
Capa Densa 1	Densa	128
Dropout 1	Dropout	-
Capa Densa 2	Densa	64
Dropout 2	Dropout	-
Salida	Densa	3
CAPA	**ACTIVACION**	**REGULARIZACION**
Entrada	-	-
Capa Densa 1	ReLU	L2 (λ=0.01)
Dropout 1	-	Dropout (20 %)
Capa Densa 2	ReLU	L2 (λ=0.01)
Dropout 2	-	Dropout (20 %)
Salida	Softmax	-
Tamaño del batch: 32		
Pérdida: Entropía cruzada categórica dispersa		
Optimizador: Adam		