UNIVERSIDAD PERUANA UNIÓN

ESCUELA DE POSGRADO

Unidad de Posgrado de Ingeniería y Arquitectura

Fine-Tuning de Modelos Monolingües BERT Preentrenados para el Análisis de Sentimientos en Contextos de Jerga Peruana

Tesis para obtener el Grado Académico de Maestro en Ingeniería de Sistemas con Mención en Dirección y Gestión de Tecnología de Información

Autor:

Sergio Elvis Calizaya Milla

Jair Samuel Santos Gonzales

Asesor:

Fredy Abel Huanca Torres

Lima, Septiembre de 2024

DECLARACIÓN JURADA DE ORIGINALIDAD DE TESIS

Yo Fredy Abel Huanca Torres, docente de la Unidad de Posgrado de Ingeniería y Arquitectura, Escuela de Posgrado de la Universidad Peruana Unión.

DECLARO:

Que la presente investigación titulada: **"Fine-Tuning de Modelos Monolingües BERT Preentrenados para el Análisis de Sentimientos en Contextos de Jerga Peruana"** de los autores Sergio Elvis Calizaya Milla y Jair Samuel Santos Gonzales, tiene un índice de similitud de 8% verificable en el informe del programa Turnitin, y fue realizada en la Universidad Peruana Unión bajo mi dirección.

En tal sentido asumo la responsabilidad que corresponde ante cualquier falsedad u omisión de los documentos como de la información aportada, firmo la presente declaración en la ciudad de Juliaca, a los 10 días del mes de octubre del año 2024

Fredy Abel Huanca Torres

ACTA DE SUSTENTACIÓN DE TESIS

En Lima, Ñaña, Villa unión a 30 días del mes de septiembre del año 2024, siendo las 8:40 horas, se reunieron de forma online sincrónica, bajo la dirección del presidente del jurado Mg. Lizeth Geanina Huanca López, el secretario PhD. Javier Linkolk López Gonzales y los demás miembros: Mg. Nemias Saboya Ríos, Dr. Soria Quijaite Juan Jesús y el asesor M.Sc. Fredy Abel Huanca Torres, con el propósito de administrar el acto académico de sustentación de Tesis de Maestría titulada "Fine-Tuning de Modelos Monolingües BERT Preentrenados para el Análisis de Sentimientos en Contextos de Jerga Peruana", conducente a la obtención del grado de Magíster en Ingeniería de Sistemas con mención en Dirección y Gestión de Tecnologías de Información.

El presidente inició el acto académico de sustentación invitando a los candidatos a hacer uso del tiempo determinado para su exposición. Concluida la exposición, el presidente invitó a los demás miembros del jurado a efectuar las preguntas, cuestionamientos y aclaraciones pertinentes, los cuales fueron absueltos por los candidatos. Luego, se produjo un receso para las deliberaciones y la emisión deldictamen del jurado. Posteriormente, el jurado procedió a dejar constancia escrita sobre la evaluación en la presente acta, con el dictamen siguiente:

Candidatos: Sergio Elvis Calizaya Milla y Jair Samuel Santos Gonzales

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	19	A	Con nominación de excelente	Excelencia

Finalmente, el presidente del jurado invitó a los candidatos a ponerse de pie para recibir la evaluación final. Además, el presidente concluyó el acto académico de sustentación, procediéndose a registrar las firmas respectivas.

Presidente

Secretario

Asesor

Miembro

Miembro

Candidato

Candidato

Fine-Tuning de Modelos Monolingües BERT Preentrenados para el Análisis de Sentimientos en Contextos de Jerga Peruana

Fine-Tuning Monolingual Pre-trained BERT Models for Sentiment Analysis in Peruvian Slang Contexts

Resumen

La innovación en el procesamiento del lenguaje natural (NLP) ha llevado a la creación de modelos como BERT, RoBERTa, GPT-4o, Llama 3 y Gemini. Sin embargo, la adaptación de estos modelos a dialectos específicos, especialmente en lenguas distintas del inglés, sigue siendo poco explorada, especialmente con jergas o lenguaje informal. En respuesta a esta necesidad, nuestra investigación evalúa modelos monolingües al español que mejor se adapten a las expresiones coloquiales peruanas, siendo la mejor alternativa RoBERTuito, un modelo pre-entrenado en un extenso corpus de tweets en español que destaca su eficacia en tareas de clasificación de texto. Afinamos y comparamos este modelo para reflejar las características del español peruano. Implementamos un proceso de recolección y preprocesamiento de datos de Facebook, enfocándonos en comentarios en español peruano. Este dataset especializado con más de 11,000 comentarios etiquetados fueron usados para entrenar modelos monolingües en la tarea de análisis de sentimientos y obtener una detección más precisa de la polaridad en textos que incluyen jergas peruanas. RoBERTuito obtuvo un F1-score equilibrado de 0.750, con una precisión de 0.858, un recall de 0.870 y una exactitud de 0.789. En comparación, BETO alcanzó una precisión de 0.794, recall de 0.725 y exactitud de 0.669; BERTuit, una precisión de 0.751, recall de 0.869 y exactitud de 0.722; y RoBERTa-BNE, una precisión de 0.783, recall de 0.759 y exactitud de 0.750. Este estudio no solo proporciona una solución para el análisis de sentimientos en español peruano, sino que también establece una base para adaptar modelos monolingües a contextos lingüísticos específicos.

Palabras clave: Fine-tuning, Análisis de sentimiento, Transformers, BERT, Español, Jergas

Abstract

The continuous advancements in natural language processing (NLP) have led to the development of highly effective models such as BERT, RoBERTa, GPT-4, Llama 3, and Gemini. However, adapting these models to specific dialects, especially in languages other than English, remains underexplored, particularly in the context of slang or informal language. In response to this need, our research evaluates monolingual Spanish models that best fit Peruvian colloquial expressions. Our approach involved constructing a specialized dataset of 11,276 manually annotated social media comments, preprocessed to retain the unique features of Peruvian slang. We also expanded the models' vocabulary using a dictionary of Peruvian slang, ensuring better recognition of local expressions. The dataset was used to fine-tune the models, with RoBERTuito demonstrating superior performance, achieving an F1-score of 0.750, significantly outperforming BETO (0.661), BERTuit (0.700), and RoBERTa-BNE (0.696). This research not only offers a robust solution for sentiment analysis in Peruvian Spanish but also sets a benchmark for adapting monolingual models to linguistic contexts, with applications extending to other dialects and informal language variants.

Keywords: Fine-tuning; Sentiment analysis; Transformers; BERT; Spanish; Slang

Índice

1. Introducción

El área de procesamiento del lenguaje natural (NLP) ha avanzado significativamente en los últimos años, permitiendo la creación de modelos basados en Transformers como BERT, RoBERTa, GPT-4o, Llama 3 y Gemini, entre otros, entrenados con cientos de millones de parámetros, los cuales han demostrado una capacidad para procesar el lenguaje natural y brindar información de manera muy efectiva. Los modelos grandes de lenguaje (LLMs) en su mayoría están preentrenados en un gran conjunto de datos en inglés y en otros idiomas, con el propósito de que el modelo aprenda funciones genéricas o de alto nivel que se puedan transferir y ajustar para tareas específicas. Los LLMs pueden realizar una amplia gama de tareas de NLP que incluyen traducir texto de un idioma a otro, generar, resumir y clasificar texto, responder preguntas, entre otras. Para el caso del lenguaje español, los modelos multilingües provenientes de grandes compañías privadas mostraron mejores resultados en ciertas tareas que los modelos monolingües (Agerri and Agirre 2023), por lo que, en un principio, no sería necesario dedicar recursos a construir modelos para el español. Sin embargo, no existe garantía de que las grandes compañías mantendrán sus modelos actualizados, provocando que estos se vuelvan obsoletos muy rápidamente. Además, se demostró en estudios de casos que modelos como el Multilingual BERT tienen una gran propensión por preferir oraciones parecidas al inglés (Papadimitriou, Lopez, y Jurafsky 2022). En otros idiomas, Multilingual BERT ha sido superado tras fallar en rendimiento en tareas de NLP frente a modelos monolingües, tales como FinBERT (Virtanen et al. 2019), un modelo basado en BERT y entrenado con tres diferentes corpus de texto en idioma finlandés.

Se ha demostrado que, cuando se complementa BERT con un método eficaz de preprocesamiento, se logran precisiones de clasificación superiores. Un ejemplo de esto se observa en la lengua italiana, donde el análisis de sentimientos de tweets en este idioma

ha mostrado resultados positivos (Pota et al. 2021). Esta estrategia de preprocesamiento podría ser aplicable y explorada en otros idiomas utilizando modelos preentrenados.

Bajo este contexto, se deben investigar el efecto del tamaño del corpus, la calidad y las técnicas de preentrenamiento para que los modelos monolingües en español sean significativamente mejores que los multilingües (Agerri y Agirre 2023). Tras esta problemática surgieron modelos monolingües en español, tales como BETO, un modelo basado en BERT para el español (Cañete et al. 2023), TwilBERT, una adaptación de BERT entrenado con tweets en español (González, Hurtado, y Pla 2021), MarIA (Gutiérrez-Fandiño et al. 2022), un modelo de lenguaje basado en RoBERTa y desarrollado con un corpus masivo de 570 GB de texto de la Biblioteca Nacional de España, RigoBERTa (Serrano et al. 2022), RoBERTuito (Pérez et al. 2021), basado en la arquitectura de RoBERTa y entrenado también con tweets, ALBETO y DistilBETO (Cañete et al. 2022), basados en las arquitecturas ALBERT y DistilBERT para el lenguaje español, los cuales obtienen mejores resultados en comparación con modelos multilingües base.

En Latinoamérica, los trabajos dedicados al análisis de sentimiento en su mayoría están basados en lexicones con polaridad. Aplicaciones tales como SAET (Utitiaj, Morillo, y Huanga 2020), en un contexto ecuatoriano, demuestran un rendimiento similar comparado con herramientas comerciales de análisis de sentimientos. Sin embargo, por ser modelos lexicon-based, no alcanzan gran precisión al no considerar más parámetros, como las estructuras gramaticales del lenguaje español. En el Perú, hasta la fecha de elaboración de este estudio, no se han identificado estudios comparables que apliquen NLP utilizando modelos monolingües de arquitectura Transformers para analizar el lenguaje coloquial, incluyendo jergas locales en este mismo contexto o enfoque.

El análisis de sentimientos es útil para entender opiniones en diversos temas como política, entretenimiento, deportes, productos y servicios locales, pero su precisión en el español peruano se ve limitada por las singularidades del lenguaje, la cultura y por los desafíos que existen al trabajar con jergas que a menudo perciben como ruido y puede ser excluida del corpus (Bhattacharyya, Dhuliawala, y Kanojia 2016), Además, la ambigüedad causada por abreviaciones y la falta de contexto complica la interpretación (Teodorescu y Saharia 2015), Por otro lado, las jergas no suelen ser parte de un vocabulario estándar de un lenguaje (Gómez-Adorno et al. 2016), por lo que dificulta su reconocimiento, Así mismo, el significado de una jerga puede variar entre positivo y negativo dependiendo del contexto en que se use (Mao, Liu, y Zhang 2024). Por ello, los modelos que procesan textos necesitan de diccionarios específicos para cada idioma en estudio (Gómez-Adorno et al. 2016), y adaptar un modelo de Transformers al español peruano podría mejorar la comprensión y precisión en el análisis de sentimientos. Esta investigación aborda cómo adaptar modelos monolingües Transformers para optimizar su eficacia en el contexto peruano, evaluando su rendimiento en el procesamiento del lenguaje natural.

Las secciones que componen este estudio están organizadas de la siguiente manera: en Trabajos Relacionados, revisamos el estado del arte y los modelos relevantes, contextualizando la necesidad de adaptar un modelo de lenguaje a jergas peruanas. Posteriormente, describimos detalladamente el proceso metodológico, incluyendo la construcción del conjunto de datos, las técnicas de preprocesamiento, la anotación de los datos y el ajuste fino de los modelos. En Resultados, presentamos la evaluación cuantitativa del rendimiento de los modelos y los resultados obtenidos. En Discusión, examinamos la aplicabilidad de los hallazgos y su impacto en el análisis de sentimientos

en contextos lingüísticos específicos. Finalmente, en Conclusión, sintetizamos las principales contribuciones del estudio y esbozamos futuras líneas de investigación.

2. Trabajos Relacionados

La tarea de clasificación de textos dentro del análisis de sentimientos se ha convertido en un campo de investigación activo en el procesamiento del lenguaje natural (Pota et al. 2021), con aplicaciones en diversas áreas como el marketing, la política, la atención al cliente y el turismo. De igual forma, es relevante para los profesionales e investigadores que buscan estudiar la interacción humano-computador (Hutto y Gilbert 2014).

En el procesamiento del lenguaje natural existen diferentes técnicas que se aplican para la clasificación de la polaridad de textos. Dentro de las tradicionales encontramos el enfoque lexicon-based, una técnica muy útil para el análisis de sentimientos (Hutto y Gilbert 2014). Consiste en el uso de un diccionario léxico que asigna una polaridad (positiva, negativa o neutral) a cada término o expresión. También encontramos los algoritmos de deep learning más utilizados en este campo, como Long Short Term Memory (LSTM), Recurrent Neural Network (RNN) y Convolutional Neural Network (CNN) (Wu et al. 2023).

En la actualidad, la arquitectura Transformers, y en particular los modelos basados en BERT representan una de las técnicas más avanzadas en el procesamiento del lenguaje natural. Estos modelos se entrenan con grandes cantidades de datos textuales y luego se ajustan para tareas específicas, como la clasificación de polaridad en textos. (Karfi y Fkihi) mencionan que los modelos basados en Transformers han alcanzado resultados de vanguardia en muchas de las tareas de procesamiento automático del lenguaje. Por otro lado, (Pota et al. 2021) afirman que los modelos de codificación bidireccional Transformer logran resultados sobresalientes en el reconocimiento y clasificación de texto.

El modelo TwilBERT (González, Hurtado, y Pla 2021), derivado de BERT y especializado en el idioma español dentro del dominio de Twitter, aborda 14 tareas de clasificación de texto donde utiliza un predictor de orden de respuestas para aprender la coherencia en cada tweet. De esta forma, supera a la versión multilingüe de BERT hasta en +11.07 en la métrica F1-Score. Esta investigación proporciona un framework utilizando la librería Keras para entrenar y aplicar fine-tuning al modelo TwilBERT.

Por otro lado, RoBERTuito (Pérez et al. 2021), un modelo Transformer que utiliza la arquitectura de los modelos RoBERTa y BERTweet, está entrenado con tweets en español y es considerado por los autores competitivo en el multilenguaje. Para el entrenamiento, se utilizaron 500 millones de tweets en español y conjuntos de datos específicos, como variantes regionales o temáticas. Los pesos del modelo están publicados en la plataforma HuggingFace para uso libre. Los resultados de la métrica F1-Score promediada, en comparación con otros modelos como RoBERTa y BETO, superan a estos en las tareas de análisis de sentimientos y detección de discurso de odio.

Entrenar un modelo desde cero, como BERT y su versión en español BETO, conlleva el uso de grandes recursos computacionales y tiempo. Para evitar esto, se han creado modelos Transformers ligeros con parámetros reducidos, como ALBERT (Lan et al. 2020) y DistilBERT (Jiao et al. 2020), pero entrenados en idioma inglés. (Cañete et al. 2022) propone ALBETO y DistilBETO, modelos preentrenados exclusivamente con corpus en español. Los resultados de estos modelos están a la altura del modelo BETO. Los autores han publicado libremente el modelo para investigaciones futuras.

También existen diferentes investigaciones que han creado modelos Transformers preentrenados derivados de BERT con el objetivo de entrenarlos con un solo idioma específico. CamemBERT (Martin et al. 2020) y FlauBERT (Le et al. 2020) para el idioma francés, RoBERT (Delobelle, Winters, y Berendt 2020) para el idioma holandés y

FinBERT, (Virtanen et al. 2019) para el idioma finlandés. Por otro lado, el modelo indoBERT en idioma indonesio está ajustado para el reconocimiento de jergas en ese idioma (Fernandez et al. 2022). Todos estos modelos preentrenados han demostrado ser mejores que los modelos multilingües.

A pesar de que hay numerosas investigaciones previas que han aplicado modelos basados en Transformers para clasificar la polaridad de textos, la mayoría de estos estudios se han centrado en textos en inglés y muy pocos abordan el español (Utitiaj, Morillo, y Huanga 2020). Esto ha revelado una notable falta de modelos dedicados a la clasificación de polaridad en textos en español, especialmente en un contexto de lenguaje coloquial peruano.

3. Desarrollo

Esta investigación es de carácter explicativo y se centra en un aspecto poco explorado dentro del análisis de sentimientos: la adaptación de modelos de lenguaje monolingües a un conjunto de datos que contiene jergas del español peruano. Aunque el análisis de sentimientos ha sido ampliamente estudiado en otros contextos y lenguas, hasta la fecha no se han encontrado estudios específicos que aborden las particularidades del español peruano y sus expresiones coloquiales. Por lo tanto, nuestro trabajo busca llenar este vacío para variantes dialectales, con un enfoque en la jerga peruana.

El presente estudio comenzó con la extracción de textos provenientes de comentarios en la plataforma Facebook, utilizando la Graph API de Meta, a través del software libre Facepager[1]. Esta etapa inicial permitió conformar un corpus en español. Posteriormente, el corpus fue sometido a un riguroso proceso de limpieza y filtrado,

[1] https://github.com/strohne/Facepager

donde se aplicó un diccionario especializado en jergas peruanas para identificar y excluir comentarios que contenían jerga y expresiones coloquiales. Una vez refinado el corpus, procedimos a la construcción del dataset mediante un proceso de anotación manual. Este dataset fue utilizado para llevar a cabo un ajuste fino (full finetuning) de 4 modelos preentrenados al español: RoBERTuito, BETO, BERTuit y RoBERTa-BNE. Por último, pusimos a prueba los modelos ajustados en la tarea de clasificación de texto para el análisis de sentimiento, enfocándonos en determinar la polaridad de los comentarios como positiva, negativa o neutral. Para la evaluación del desempeño de los modelos, se emplearon las métricas estándar de clasificación: precisión, recall, exactitud y F1-score. Estos pasos se detallan en la figura 1.

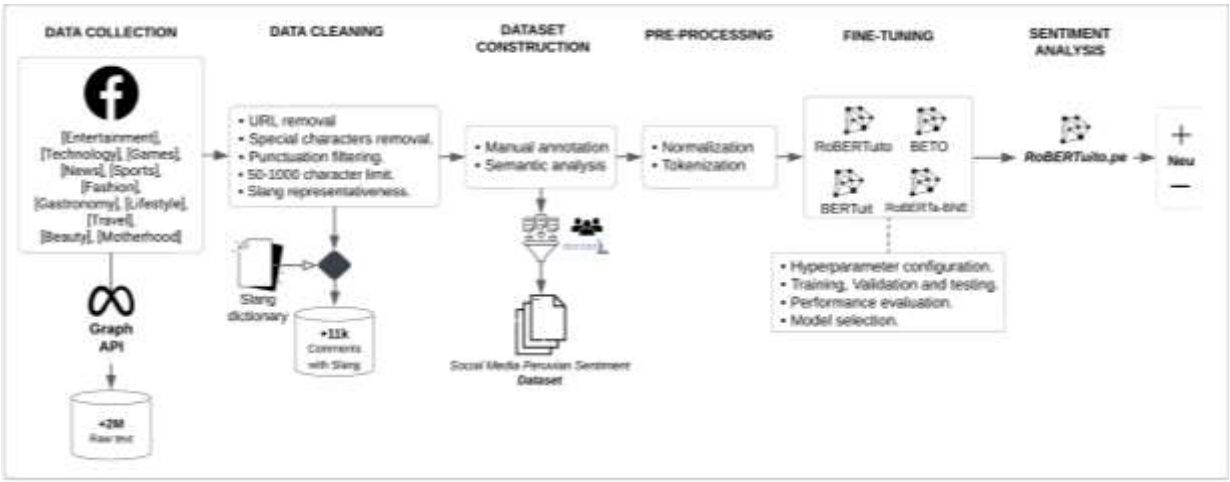

Figura 1. Diagrama del método propuesto de Fine-Tuning de los modelos BERT

3.1. Construcción del conjunto de datos

La población de estudio consistió en comentarios publicados por usuarios de Facebook en Perú. La recolección de datos se llevó a cabo utilizando la Graph API de Meta y Facepager, basándonos en la metodología y criterios de un estudio sobre influencers peruanos (Forbes Perú 2023), que evaluó factores como número de seguidores,

interacción, impacto e influencia en redes sociales. Se seleccionaron comentarios de los creadores de contenido más influyentes en áreas como entretenimiento, tecnología, juegos, noticias, deportes, moda, gastronomía, estilo de vida, farándula, viajes, belleza y maternidad. El periodo de recolección abarcó cinco años (2019-2023), lo que resultó en un total de 2.2 millones de comentarios recopilados.

Tabla 1. Distribución de dominios temáticos del conjunto de datos

Dominios temáticos	**%**
Deportes	13%
Entretenimiento/Gamers	20%
Gastronomía/Foodies	13%
Maternidad	13%
Moda/Belleza/Estilo de vida	13%
Noticias	6%
Populares/farándula	12%
Tecnología	3%
Viajes	7%
Total	**100%**

Para conformar el conjunto de datos definitivo, se llevó a cabo un proceso de limpieza y filtrado sobre los 2.2 millones de comentarios iniciales, y tomando como referencia el estudio de (Cuzcano et al. 2020), consideramos incluir comentarios con emojis para reforzar las emociones en los mensajes. Este procedimiento nos proporcionó un conjunto con más de 90,000 comentarios, cada uno caracterizado por la inclusión de jergas identificadas mediante nuestro diccionario de jergas. Con el objetivo de obtener una visión preliminar, decidimos seleccionar una muestra representativa de estos comentarios. Para esto, aplicamos una técnica de muestreo basándonos en la determinación de muestra en poblaciones finitas (Montesinos et al. 2010).

$$n = \frac{NZ^2\, p(1-p)}{(N-1)E^2 \;+\; Z^2 p(1-p)} \tag{1}$$

Donde:

- *N*: 90000, el tamaño de la población
- *Z*: 2.58, el nivel de confianza del 99%
- *p*: 0.5, proporción estimada de la característica de interés
- *E*: 0.01, margen de error

Como resultado obtuvimos alrededor de 14,044 comentarios. Tras el proceso de limpieza y depuración de comentarios, la muestra final se redujo a 11,276 comentarios válidos, los cuales conformaron nuestro dataset final al cual se ha denominado SocialMediaPeruvianSentiment (SMPS). Este procedimiento nos permitió obtener un conjunto de datos diverso y, al mismo tiempo, manejable para análisis posteriores.

Para el filtrado de comentarios, construimos un diccionario con jergas que recopilamos de foros, redes sociales, chats, páginas y del diccionario de americanismos de la Asociación de Academias de la Lengua Española (ASALE 2024). Este diccionario de jergas peruanas[2] puede ser descargado libremente.

3.2. Preprocesamiento

La limpieza y normalización del texto incluyó la eliminación de ruido, como caracteres especiales y errores tipográficos. Se evitó el uso de técnicas como stemming o lematización para no alterar la estructura semántica de las jergas y comprometer su correcta interpretación. Para asegurar la calidad y coherencia del conjunto de datos, se implementaron los siguientes procedimientos:

[2] https://huggingface.co/datasets/pyupeu/peruvian-slangs-dictionary

- **Estandarización del uso de mayúsculas y minúsculas:** Todo el texto fue convertido a minúsculas para garantizar la uniformidad y minimizar discrepancias, lo que reduce la variabilidad del corpus. Esta estandarización simplifica el vocabulario y disminuye el número de palabras únicas, mejorando así el rendimiento de los modelos de clasificación.
- **Eliminación de acentos:** Los acentos y caracteres no-ASCII, a excepción de los emojis, fueron eliminados debido a su uso inconsistente en el lenguaje informal. Algunos modelos de clasificación experimentan ligeras mejoras de rendimiento cuando el texto es procesado sin acentos.
- **Reducción de caracteres repetitivos.** Se estableció un límite de tres caracteres consecutivos para mitigar la repetición excesiva en palabras como “Paltaaaaaaaaaaa”, que se transformó en "Paltaaa", y “Causaaaaaaa”, que se normalizó a “Causaaa”.
- **Estandarización de expresiones de risa.** Las expresiones de risa como "jajajajajajja" o "hahaha" fueron convertidas a "jajaja", estandarizando su formato para mantener la consistencia en el dataset.
- **Estandarización de emojis.** En el caso del modelo RoBERTuito los emojis se convirtieron a una representación textual, por ejemplo “🤮” a “emoji cara vomitando” o “😊” a “emoji cara sonriendo” debido a que el modelo base esta entrenado de esta forma para esto usamos la librería *emoji*[3], para el resto de los modelos se entrenaron con los emojis en formato hexadecimal, se evaluó eliminar

[3] https://github.com/carpedm20/emoji/

los emojis para todos los modelos, sin embargo, no hubo una mejora significativa en rendimiento al momento de eliminarlos.

- **Eliminación de menciones, hashtags y links**: Las menciones de usuarios fueron reemplazadas por el texto "@username", los hashtags se normalizaron manteniendo solo el contenido de texto (por ejemplo, "#BreakingNews" se convirtió en "breaking news"), y las URLs fueron eliminadas por completo. Este proceso contribuyó a la consistencia del dataset, preservó el anonimato de los usuarios mencionados y eliminó datos irrelevantes contenidos en las URLs. Sin embargo, en algunos casos. las menciones hechas en Facebook incluían nombres parciales o completos, lo que permitió identificar esos comentarios como inválidos durante el proceso de anotación. Como en el ejemplo: "***Juan Andres*** *es tu jato jajaja*". Debido a la naturaleza del lenguaje informal, estas menciones no se identificaron como etiquetas o enlaces específicos. En su lugar, se evaluó la relevancia del comentario en función de su contenido y longitud.

El uso de un diccionario especializado en jergas peruanas fue esencial para identificar casos en los que las palabras contenían jergas incrustadas en su estructura léxica. Por ejemplo, en el comentario *"Yo mirando todo el vídeo y diciendo ahorita le da una patada a la pared"*, aquí, la palabra "*patada*" encapsula la jerga "*pata*", que en el argot peruano se interpreta como "Amigo". De manera similar, se detectaron términos polisémicos en numerosos comentarios. Por ejemplo, "*mi perro también se lastimó la pata..."*, el término "*pata*" se entiende en su sentido literal. Sin embargo, en un contexto diferente, como en "*Yo concuerdo en muchas cosas con el pata...*", el mismo término "*pata*" adquiere su significado coloquial, refiriéndose a un amigo. Esta observación resalta la complejidad y la riqueza semántica propia del lenguaje coloquial y la jerga que puede referirse a más de un significado a la vez (Teodorescu y Saharia 2015).

Para abordar esta dificultad, implementamos una validación adicional basada en la similitud de coseno (Kwartler 2017), la fórmula general (2) da como resultado el coseno del ángulo θ entre dos vectores, que varían entre 0 y 1, una similitud de coseno cercana a 1 indica que la palabra y la jerga están altamente alineadas semánticamente, mientras que un valor cercano a 0 indicaría poca o ninguna similitud.

$$Similitud\ Coseno = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2)$$

Durante esta validación, observamos que el 91% de los comentarios del conjunto de datos presentan una similitud coseno mayor a 0.7, lo que indica una alta precisión en la identificación de las jergas dentro de los comentarios. La tabla 2 muestra algunos ejemplos del cálculo en cada jerga identificada.

Tabla 2. Ejemplos de similitud Coseno entre la jergas y palabras reconocidas

Comentarios	Jerga	Palabra	Similitud coseno
Joel Apolaya broder😅😅	broder	Joel	0.292
Literal fría yara🤣🤣🤣	yara	fría	0.325
Cholo.. En Premios la Zona estabas Bien Charlie... 😎😎😂	charli	Charlie...	0.444
Te quedo de lujo!! Me has dado varias ideas de como darle mas sabor a mi chaufita, Gracias!!😄👌	chau	chaufita,	0.478
…la teletón es una estafa no entiendo chaparrón	chapa	chaparrón	0.565
	chapar	chaparrón	0.738
… CONERAS los alaracos de mi barrio 🤣	habla	hablar	0.734
BUENO ESO ENTENDÍ XD 1ra vez ENTENDÍ bien hablar los chilenos. SALUDOS	conera	CONERAS	0.877
Melanie Arce Rodriguez mira amor que capo!! 👏	capo	capo!!	0.858
	la firme	firme	0.790
O chino pagales con un turrón y a la firme que feo mostró..... 🤣🤣	chino	chino	1.000
	firme	firme	1.000
	turrón	turrón	1.000

En la figura 2, podemos observar los términos predominantes, mostrando los desafíos de NLP ante la riqueza del idioma.

Figura 2. N-Grama de las Jergas del conjunto de datos.

3.3. Proceso de Anotación

Para la anotación de los comentarios, se utilizó Doccano[4], una herramienta de software libre, con funciones para el etiquetado de texto de forma colaborativa. La anotación consistió en una revisión por parte de 8 personas, peruanos nativos con edades entre 18 y 40 años, a quienes se le dio acceso a la herramienta Doccano para comenzar a anotar los comentarios de forma aleatoria, donde 1=positivo, 2=neutral y 3=negativo. (ver tabla 3). Una vez terminada la anotación de comentarios, obtuvimos la distribución de polaridades, (ver tabla 4), de igual forma la figura 4 muestra las jergas con mayor frecuencia en cada clase positiva, negativa y neutral. El conjunto de datos SMPS ha sido compartido públicamente en Hugging Face[5].

[4] https://github.com/doccano/doccano

[5] https://huggingface.co/datasets/pyupeu/social-media-peruvian-sentiment

Cada anotador revisó los comentarios de manera aleatoria, y estos fueron etiquetados en tres categorías: positivo (1), neutral (2) y negativo (3). A continuación, definimos de manera clara cada una de estas polaridades:

- **Positivo:** Se clasificaron como positivos aquellos comentarios que expresaban opiniones favorables, satisfacción o elogios hacia una persona, evento o idea. En el contexto de las jergas peruanas, se incluyeron términos coloquiales que comúnmente denotan aprecio o admiración, tales como "capo", "buenazo" y "trome". Estos términos suelen utilizarse para resaltar las cualidades positivas de una persona, como en el caso de "Buena Carlitos, eres un capo" (expresando que alguien es muy talentoso o sobresaliente). Asimismo, se incluyeron como positivos los comentarios que, aunque en tono humorístico o de broma, no contenían insultos o mala intención (chascarrillo).
- **Negativo:** Los comentarios negativos reflejaban opiniones de desaprobación, críticas o descontento. En muchos casos, se identificaron insultos característicos de la cultura urbana, burlas o un uso elevado de sarcasmo. Términos como "palta" (en referencia a una situación embarazosa o problemática) fueron comúnmente etiquetados como negativos, al igual que los comentarios que expresaban insatisfacción o rechazo, ya sea de manera explícita o mediante el uso de ironía. Estos comentarios podían incluir lenguaje directo o implicar descontento a través de connotaciones culturales específicas.
- **Neutral:** Los comentarios fueron etiquetados como neutrales cuando no mostraban una opinión clara ni positiva ni negativa. En la mayoría de los casos, estos comentarios consistían en descripciones o información sin una carga emocional significativa. En cuanto a las jergas, se consideraron neutros aquellos términos que no transmitían una emoción específica o que eran utilizados de

forma descriptiva, como en "Su chapa es el gordo" (su apodo es el gordo) o "Con solo ver este video de comida ya me dio la bajada" (me dio hambre al ver el video), donde las palabras "gordo" y "bajada" no contenían juicio ni intención ofensiva. Algunas jergas, como "pulpín" (joven), fueron particularmente difíciles de clasificar debido a su ambigüedad, ya que su interpretación depende del contexto. Por ejemplo, en "Ese pulpín, ¿a dónde se fue?" (¿Ese joven a dónde se fue?), "pulpín" puede ser neutral si no se expresa ningún juicio sobre la edad de la persona.

Se identificó la necesidad de gestionar de manera rigurosa las muestras que contenían únicamente emojis, comentarios demasiado cortos o largos que carecían de sentido, mensajes de spam o publicidad, abreviaciones intencionalmente distorsionadas, o menciones de usuarios sin contenido relevante. En estos casos, los anotadores fueron instruidos para evaluar la intención general del comentario y, si se consideraba que carecía de sentido o relevancia, clasificarlo como "inválido". Estos comentarios inválidos fueron eliminados del conjunto de datos SMPS utilizado para el entrenamiento de los modelos, asegurando así la calidad y coherencia del corpus.

Tabla 3. Ejemplo de comentarios etiquetados del conjunto de datos SMPS

ID	Comentarios	Jerga	Polaridad
2037	Buena carlitos!!! Eres un capo  .	capo	Positivo
7041	Que éxito!!!!!unos tromes todos !!no hay como nuestro chifa !!! Es delicioso  .	trome	Positivo
3810	Mira Camilo Ahi tienes que empujarte tu cevillano.	cevillano	Neutral
19964	Con solo ver este video de comida ya me dio la bajada.	bajada	Neutral
18615	Jajajajajajaja  ahora ya quiere pasar piola para que no la sigan atacando  .	piola	Negativo
244536	 causa! Tú con todas las publicaciones q realizas y más palta los gobiernos que hemos elegido hasta antes de Vizcarra  .	palta, causa	Negativo

Tabla 4. Distribución de polaridades en el dataset etiquetado

Polaridades	Cantidad de comentarios	%
Positive	4,736	42%
Negative	3,280	29%
Neutral	3,260	29%
Total	**11,276**	**100%**

La jerga más frecuente en el análisis es "chino", un término coloquial utilizado comúnmente para referirse a un amigo o conocido, o como apodo. En algunos casos, también puede aludir a rasgos faciales asociados con ascendencia asiática. Este término aparece en las tres categorías de polaridad, lo que indica su ambigüedad semántica y su capacidad para adquirir diferentes interpretaciones según el contexto. De manera similar, la palabra "carajo" se utiliza predominantemente en las categorías positiva y negativa, y su connotación varía según el uso. En ciertos contextos, "carajo" puede expresar frustración o enojo, como en "¡No puede ser que hayan perdido el partido otra vez, carajo!". En otros casos, puede emplearse para expresar entusiasmo o afecto, como en "¡Viva el Perú, carajo!". Por otro lado, se observa que la palabra "buenazo" es la más frecuente dentro de la categoría positiva, reflejando su uso en situaciones de aprobación o elogio.

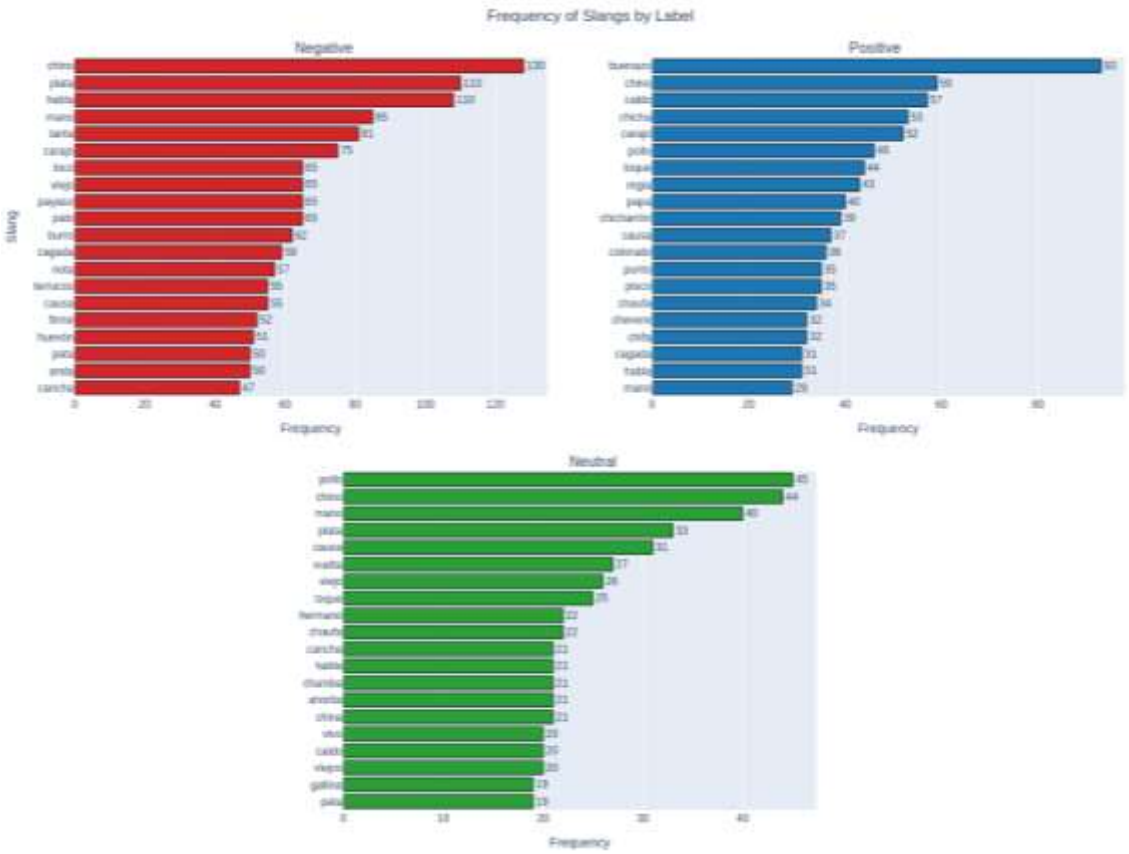

Figura 3. Frecuencia de las jergas en cada clase positiva, negativa y neutral.

3.4. Proceso de Tokenización

Se implementó la técnica de tokenización de palabras (Word Tokenization) para descomponer los textos en unidades mínimas o palabras individuales. Durante este proceso, se observó que los tokenizadores predefinidos de los modelos de lenguaje no reconocían adecuadamente las jergas peruanas presentes en el dataset. Un ejemplo claro de esta deficiencia se observa en el caso de la frase "no es pituca causa", donde el tokenizador de los modelos no tenían indexada la palabra "pituca" u otros peruanismos que no están en el tokenizador.

Para superar estas limitaciones, se realizó una expansión manual del vocabulario base de los modelos, originalmente compuesto por 30,000 tokens. Este proceso implicó la incorporación de nuevos tokens extraídos del diccionario de jergas peruanas, resultando en un vocabulario ampliado de 31,886 tokens. Esta expansión permitió una

descomposición más precisa de los textos que incluían jergas o palabras nuevas, mejorando significativamente la calidad del preprocesamiento de datos previo a la fase de entrenamiento. La figura 4 ilustra esta mejora.

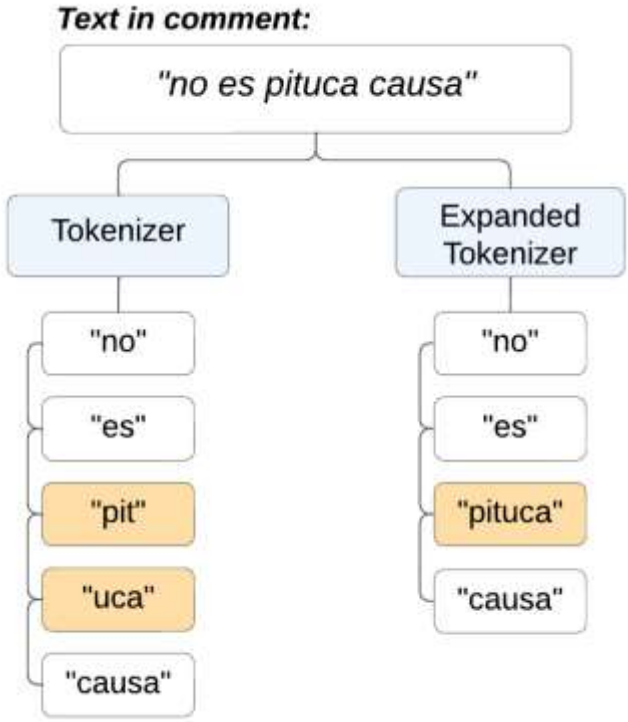

Figura 4. Tokenización de palabras.

La arquitectura de los modelos RoBERTuito, BETO, BERTuit y RoBERTa-BNE pertenece a la familia de modelos Transformer, introducida originalmente por (Vaswani et al. 2017). Esta arquitectura, representada en la Figura 5, permite a los modelos procesar secuencias extensas de texto. Gracias a su mecanismo de self-attention y multi-headed attention, los modelos son capaces de capturar relaciones complejas dentro del texto. En particular el modelo RoBERTuito basado en la arquitectura RoBERTa (Liu et al. 2019) y BERTweet (Nguyen, Vu, y Nguyen 2020) ha sido entrenado con 12 “self-attention layers”, cada una con 12 “attention heads”, y un “hidden size” de 768 dimensiones. Para

RoBERTuito usamos la versión *deacc* donde el modelo fue entrenado en minúsculas y elimina acentos.

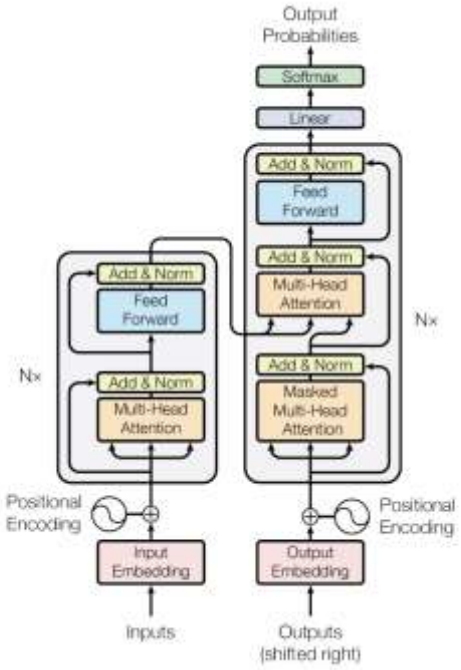

Figura 5. Arquitectura Transformer

3.5. Entrenamiento de los Modelos

Con el objetivo de adaptar los modelos preentrenados en español RoBERTuito, BETO, BERTuit y RoBERTa-BNE a nuestra tarea específica de clasificación de texto, realizamos un ajuste de los parámetros de los modelos. Este proceso incluyó la configuración de una tasa de aprendizaje de (5e-5), siguiendo las directrices de (Devlin et al. 2019), que sugieren una tasa pequeña para cambios graduales y efectivos en los pesos preentrenados del modelo, train batch size de 32 y evaluation batch size de 16. El proceso de ajuste se llevó a cabo durante un total de 5 épocas, lo cual se consideró suficiente para permitir que los modelos capturen las particularidades de la tarea específica sin perder la riqueza de las representaciones lingüísticas adquiridas durante la fase de preentrenamiento. Este enfoque también ayuda a mitigar el riesgo de sobreajuste.

Además, se implementó un calentamiento (warm up) del 10% de los pasos de entrenamiento, lo que permitió un inicio suave del proceso de optimización.

Los embeddings de cada modelo fueron redimensionados a un tamaño de 31,886, de acuerdo con el vocabulario específico de su propio tokenizador, asegurando una representación precisa de los peruanismos. Para la división de los datos, se empleó el 80% del conjunto total para el entrenamiento y el 20% restante se destinó a la prueba y validación.

Los modelos fueron entrenados utilizando una computadora equipada con una tarjeta NVIDIA RTX 3060 con 12 GB de RAM, así como una instancia con GPU Nvidia T4 en un entorno de Google Colab. El tiempo de entrenamiento por modelo osciló entre 15 y 20 minutos, lo que demuestra que el ajuste fino de modelos monolingües preentrenados, enfocados en dialectos y jergas específicas, puede realizarse de manera eficiente sin requerir recursos computacionales excesivamente altos.

4. Evaluación y Resultados

Para la evaluación de los modelos de clasificación de análisis de sentimientos, se aplicaron las métricas estándar de precisión, recall, accuracy, F1 Score y Macro F1 Score (Murphy 2012). Estas métricas proporcionan una visión integral de la efectividad de los modelos en la clasificación correcta de los comentarios analizados. La exactitud (Accuracy) indica la proporción de predicciones correctas en general, lo que significa que es la cantidad total de casos que el modelo clasifica correctamente, considerando tanto los positivos como los negativos. La precisión, por su parte, mide cuán exactas son las predicciones positivas, indicando la proporción de casos que fueron correctamente identificados como positivos en comparación con el total de predicciones realizadas. La sensibilidad (Recall), por su parte, evalúa la capacidad del modelo para identificar todos los casos positivos, representando la proporción de casos positivos que fueron

pronosticados correctamente en relación con el número total de casos positivos reales. La puntuación F1 proporciona una medida equilibrada que considera tanto la precisión como el recall, calculada como la media armónica de ambas métricas (Hidayat et al. 2021).

$$Accuracy = \frac{tp \ + \ tn}{tp \ + \ tn \ + \ fp \ + \ fn} \tag{3}$$

$$Precision = \frac{tp}{tp \ + \ fp} \tag{4}$$

$$Recall = \frac{tp}{tp \ + \ fn} \tag{5}$$

$$F1 \ Score = 2 \cdot \frac{precision \ \times \ recall}{precision \ + \ recall} \tag{6}$$

$$Macro \ F1 = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{7}$$

Donde *tp* representa el número de verdaderos positivos, es decir, los casos donde el modelo ha identificado correctamente un comentario como positivo; *tn* denota el número de verdaderos negativos, o comentarios negativos correctamente clasificados; *fp* corresponde a los falsos positivos, donde un comentario negativo es incorrectamente identificado como positivo; y *fn* refleja los falsos negativos, situaciones en las que comentarios positivos son erróneamente clasificados como negativos. El Macro F1 Score, por su parte, calcula la media aritmética de los F1 Scores obtenidos para cada una de las clases del modelo, ofreciendo así un panorama equilibrado que no está influenciado por el desbalance entre clases. En la tabla 4 podemos observar que RoBERTuito muestra el desempeño más destacado en términos de precisión y recall para comentarios positivos y negativos. En particular, RoBERTuito logra una precisión de 0.858 para comentarios positivos y 0.828 para negativos, con un recall de 0.870 y 0.798, respectivamente. Estos resultados indican que RoBERTuito es altamente efectivo en identificar correctamente

tanto comentarios positivos como negativos, y su F1-Score de 0.864 para positivos y 0.813 para negativos.

Tabla 5. Resultados de evaluación de rendimiento en los modelos ajustados

Métricas	RoBERTuito			BETO			BERTuit			RoBERTa-BNE		
	P	Neu	N	P	Neu	N	P	Neu	N	P	Neu	N
Precision	0.858	0.565	0.828	0.794	0.502	0.701	0.751	0.629	0.746	0.783	0.561	0.746
Recall	0.870	0.578	0.798	0.725	0.588	0.669	0.869	0.492	0.736	0.759	0.590	0.740
F1-Score	0.864	0.572	0.813	0.758	0.542	0.685	0.806	0.552	0.741	0.771	0.575	0.743
Accuracy	0.789			0.669			0.722			0.750		
MacroF1	0.750			0.661			0.700			0.696		

BETO y BERTuit también muestran resultados prometedores. BETO logra un F1-Score de 0.758 para comentarios positivos y 0.685 para negativos, con una precisión de 0.794 para positivos y 0.701 para negativos. BERTuit obtiene un F1-Score de 0.806 para positivos y 0.741 para negativos, con una precisión de 0.751 y 0.746, respectivamente. Estos modelos demuestran su capacidad para clasificar correctamente los comentarios, aunque con un rendimiento ligeramente inferior al de RoBERTuito.

El modelo RoBERTa-BNE presenta un rendimiento equilibrado, con una precisión de 0.783 para comentarios positivos y 0.746 para negativos, y un recall de 0.759 y 0.740, respectivamente. Su F1-Score de 0.771 para positivos y 0.743 para negativos indica un buen equilibrio entre precisión y recall en estas categorías.

Se ha evidenciado que todos los modelos presentan dificultades en la clasificación de comentarios neutrales, con puntuaciones de precisión y recall significativamente más bajas en esta categoría. RoBERTa-BNE tiene el mejor rendimiento en comentarios neutrales entre los modelos evaluados, con un F1-Score de 0.575. RoBERTuito y BERTuit le siguen con F1-Scores de 0.572 y 0.552, respectivamente. BETO también muestra un rendimiento moderado en esta categoría, con un F1-Score de 0.542. Estas

dificultades sugieren la necesidad de un mayor refinamiento de los modelos para mejorar la clasificación de comentarios neutrales.

Las matrices de confusión presentadas en la figura 6 muestran el rendimiento de los modelos RoBERTuito, BERTuit, RoBERTa-BNE y BETO en la clasificación de comentarios en tres categorías: negativos, neutrales y positivos. El modelo RoBERTuito demuestra una capacidad sólida para clasificar comentarios negativos (31.10%) y positivos (21.17%), aunque, al igual que los otros modelos, presenta desafíos al identificar comentarios neutrales, con un 13.31% de aciertos y una tendencia a confundir comentarios negativos con neutrales (14.04%).

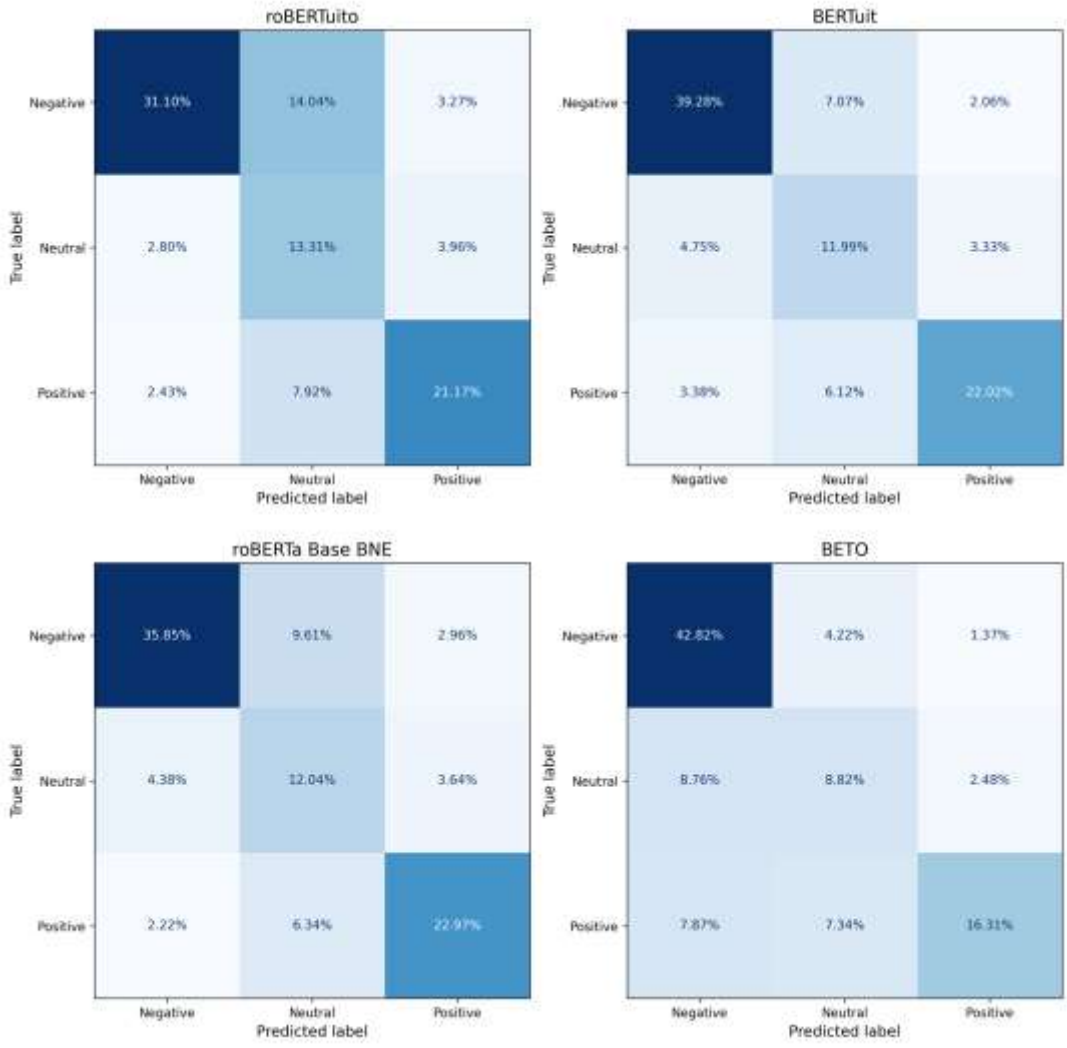

Figura 6. Matrices de confusión de los Modelos Ajustados

Por otro lado, BERTuit muestra el mejor rendimiento en la clasificación de comentarios negativos, con un 39.28% de precisión, pero también una mayor confusión al clasificar comentarios neutrales como negativos (7.07%) y positivos (22.02%). Esto indica que BERTuit tiene una inclinación a clasificar de manera más extrema, pero mantiene una buena precisión en los extremos de la polaridad.

RoBERTa-BNE, aunque similar en su desempeño a BERTuit en la clasificación de comentarios positivos (22.97%), muestra una menor capacidad para identificar comentarios neutrales, con un 12.04% de aciertos y una confusión significativa al clasificar comentarios negativos como neutrales (9.61%). Este modelo parece encontrar un equilibrio en la clasificación de comentarios negativos y positivos, pero sufre en la identificación precisa de neutralidad.

Finalmente, BETO, con un desempeño notable en la clasificación de comentarios negativos (42.82%), muestra la mayor confusión en la categoría de neutrales, con un 8.76% de aciertos y un considerable número de falsos positivos y negativos. BETO también tiene la menor precisión en la identificación de comentarios positivos (16.31%), lo que demuestra que este modelo tiene dificultades para captar correctamente el contexto positivo.

Tabla 6. Evaluación de Modelos BERT con Fine-Tuning en la Clasificación de Textos, los números en negrita indican puntajes altos.

Modelos	Macro F1	Macro F1 (Fine-Tuned)
RoBERTuito	0.707	**0.750**
BETO	0.665	0.661
BERTuit	0.632	**0.700**
RoBERTa-BNE	0.669	0.696

Los resultados globales mostrados en la Tabla 6, comparados con los resultados reportados en las tareas de clasificación de texto de cada modelo respectivo (Pérez et al.

2021) y (Huertas-Tato, Martin, y Camacho 2023), indican que el modelo con mejor desempeño en la clasificación de polaridad es RoBERTuito, alcanzando un Macro F1 de 0.750, seguido por BERTuit, que exhibe una puntuación de 0.700. Estos dos modelos demuestran una gran capacidad para la clasificación de textos dentro del conjunto de datos. Por otro lado, BETO muestra un rendimiento respetable con una puntuación de 0,661. El modelo RoBERTa-BNE también presenta un sólido rendimiento con un Macro F1 de 0,696, lo que indica que, aunque es ligeramente inferior a BERTuit y RoBERTuito, sigue siendo eficaz en la clasificación de la polaridad.

5. Discusión

Esta investigación destaca la naturaleza desafiante en la clasificación de sentimientos en textos coloquiales, y la relevancia de considerar las particularidades lingüísticas y culturales en los modelos de NLP. A pesar de los avances alcanzados por modelos como RoBERTuito, BERTuit, BETO y RoBERTa-bne en el procesamiento de textos en español estándar, persisten desafíos al adaptarlos a variantes dialectales, tal como lo evidencia nuestro análisis de jergas peruanas y sus resultados. En nuestro estudio, hemos empleado modelos de arquitectura Transformers, obteniendo resultados prometedores. En línea con nuestras observaciones, la investigación de (Merayo et al. 2024) ha demostrado mejoras significativas en la clasificación de sentimientos y emociones al evaluar a RoBERTuito en comparación con algoritmos tradicionales dentro de un contexto de lenguaje informal como es el gaming donde se destaca que la falta de contexto y el uso del lenguaje informal como lo son las jergas en los videojuegos son las limitantes debido a que prima las abreviaturas y repeticiones de letras. Por consiguiente, se sugiere abordar no solo a las jergas peruanas, sino que también incluir las jergas propias de redes sociales con enfoque más detallado y contextualizado en futuras investigaciones.

Al contrastar con investigaciones previas, nuestro estudio remarca la importancia de un conjunto de datos bien definido, específico y no necesariamente extenso para mejorar la clasificación de sentimientos. A diferencia de enfoques generales, que pueden no captar las idiosincrasias del lenguaje coloquial y su contexto, nuestro enfoque orientado a un dialecto particular muestra que la precisión en la clasificación de polaridades puede mejorar con mejores técnicas de preprocesamiento y adaptaciones de BERT. Un claro ejemplo es la investigación de (Kannan y Kothamasu 2022), quienes alcanzaron en promedio hasta un 25% más en el rendimiento de su modelo modificado comparado con un modelo base BERT tras usar un procedimiento de preprocesamiento que incluía convertir la jerga de Twitter, incluidos los emojis y emoticones, en texto sin formato y usar una versión de BERT que fue entrenada con texto sin formato. Por otro lado, (Fernandez et al. 2022) lograron una Accuracy de 60.35% en un contexto similar de lenguaje coloquial, aplicando fine-tuning a IndoBERT, una versión de BERT para el idioma indonesio con un corpus relativamente pequeño. En contraste, nosotros obtuvimos una accuracy de 78.9%. Aunque no se compara el mismo lenguaje, es cierto que ambas investigaciones abordan un contexto informal y obtienen resultados prometedores. Esto refuerza la idea de que el fine-tuning de modelos monolingües no requiere necesariamente grandes volúmenes de texto para obtener resultados destacados. Es fundamental considerar la complejidad de las jergas y la ironía, ya que estos elementos dependen en gran medida del contexto y no tanto del significado literal de las palabras, como mencionan (González, Hurtado, y Pla 2020). A partir de esto, se abren nuevas posibilidades para investigar cómo los modelos monolingües, junto con técnicas de anotación más precisas, pueden abordar mejor los desafíos que presentan los comentarios ambiguos o emocionalmente confusos.

6. Conclusiones y Trabajo Futuro

Esta investigación aborda la problemática de la clasificación de textos con jergas propias de la cultura peruana, aplicando un conjunto de datos de textos etiquetado manualmente con polaridades y entrenado con modelos monolingües al idioma español. El modelo con mejor rendimiento fue RoBERTuito, alcanzando un Macro F1 Score de 0.750 después del ajuste fino. A este modelo ajustado lo denominamos RoBERTuito.pe[6]. Este hallazgo demuestra que el ajuste de modelos pre-entrenados al español con léxicos específicos propios del español en Perú puede mejorar significativamente la clasificación de textos. Sin embargo, se identificaron áreas de mejora, particularmente en la clasificación de comentarios neutrales, lo que indica la necesidad de un mayor refinamiento del modelo explorando otras técnicas avanzadas de NLP que puedan manejar mejor las ambigüedades y el contexto para este tipo de datos.

A lo largo de este estudio, se ha confirmado que la variedad y la integridad del conjunto de datos ejercen una influencia en los resultados. Por consiguiente, se planea ampliar la recolección de datos para expresiones coloquiales más actuales y llevar a cabo más experimentos para perfeccionar los hallazgos. El conjunto de datos[7] desarrollado puede utilizarse y descargarse libremente desde la plataforma Hugging Face y hemos puesto a disposición tanto el corpus de texto[8] como los scripts utilizados en todo el proceso en un repositorio[9] de GitHub. Esto facilita su uso en investigaciones futuras para otras tareas que aborden el español peruano, tales como la clasificación de emociones o

---

[6] https://huggingface.co/pyupeu/robertuito-peruvian-sentiment

[7] https://huggingface.co/datasets/pyupeu/social-media-peruvian-sentiment

[8] https://huggingface.co/datasets/pyupeu/social-media-peruvian-corpus

[9] https://github.com/jairleo95/peruvian-slang-sentiment

desafíos que persisten en NLP, especialmente en dialectos específicos como la ironía, sarcasmo o discurso de odio.

Por otro lado, la extensa variedad de comentarios en nuestro estudio presentó desafíos para establecer una concordancia entre anotadores. Dado que los anotadores trabajaron aleatoriamente en más de 11,000 comentarios, no se calculó directamente el índice de acuerdo. Por tanto, proyectamos como línea de investigación futura el análisis detallado de la concordancia entre tres o más anotadores. Este esfuerzo permitirá no solo validar la consistencia de las etiquetas asignadas, sino también afinar la precisión de nuestro conjunto de datos anotados. Por último, planteamos comparar el rendimiento de los modelos monolingües frente a los LLMs más recientes en la tarea de clasificación de texto con el dataset propuesto.

Declaración de divulgación

Los autores manifiestan que no tienen conflictos de interés en relación con la publicación de este artículo.

Declaración de disponibilidad de datos

Los datos que apoyan los resultados de este estudio se encuentran disponibles públicamente en Figshare http://doi.org/10.6084/m9.figshare.26198315, número de referencia 26198315.

7. Referencias

Agerri, Rodrigo, and Eneko Agirre. 2023. Lessons Learned from the Evaluation of Spanish Language Models, December. http://arxiv.org/abs/2212.08390.

ASALE. 2024. Asociación de Academias de La Lengua Española. *Https://Www.Asale.Org/*.

Bhattacharyya, Pushpak, Shehzaad Dhuliawala, and Diptesh Kanojia. 2016. *SlangNet: A WordNet like Resource for English Slang*. http://www.reddit.com.

Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish Pre-Trained BERT Model and Evaluation Data, August. http://arxiv.org/abs/2308.02976.

Cañete, José, Sebastián Donoso, Felipe Bravo-Marquez, Andrés Carvallo, and Vladimir Araujo. 2022. ALBETO and DistilBETO: Lightweight Spanish Language Models, April. http://arxiv.org/abs/2204.09145.

Cuzcano, Marianne, Asesor Victor Hugo Ayma Quirita, Ximena M Cuzcano, and Victor H Ayma. 2020. A Comparison of Classification Models to Detect Cyberbullying in the Peruvian Spanish Language on Twitter. *IJACSA International Journal of Advanced Computer Science and Applications* 11 (10):132–138. doi:10.14569/IJACSA.2020.0111018.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt. 2020. RobBERT: A Dutch RoBERTa-Based Language Model, January. http://arxiv.org/abs/2001.06286.

Fernandez, Enrico, Anderies, Michael Gilbert Winata, Fadly Haikal Fasya, and Alexander Agung Santoso Gunawan. 2022. Improving IndoBERT for Sentiment Analysis on Indonesian Stock Trader Slang Language. In *2022 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, 240–244. doi:10.1109/IoTaIS56727.2022.9975975.

Forbes Perú. 2023. Estudio | Este Es El Top 10 de Influencers Peruanos 2023 - Forbes Perú. https://forbes.pe/lista-forbes/2023-10-26/estudio-este-es-el-top-10-de-influencers-peruanos-2023.

Gómez-Adorno, Helena, Ilia Markov, Grigori Sidorov, Juan Pablo Posadas-Durán, Miguel A. Sanchez-Perez, and Liliana Chanona-Hernandez. 2016. Improving

Feature Representation Based on a Neural Network for Author Profiling in Social Media Texts. *Computational Intelligence and Neuroscience* 2016. Hindawi Limited. doi:10.1155/2016/1638936.

González, José Ángel, Lluís F. Hurtado, and Ferran Pla. 2020. Transformer Based Contextualization of Pre-Trained Word Embeddings for Irony Detection in Twitter. *Information Processing and Management* 57 (4). Elsevier Ltd. doi:10.1016/j.ipm.2020.102262.

González, José Ángel, Lluís F. Hurtado, and Ferran Pla. 2021. TWilBert: Pre-Trained Deep Bidirectional Transformers for Spanish Twitter. *Neurocomputing* 426 (February). Elsevier B.V.:58–69. doi:10.1016/j.neucom.2020.09.078.

Gutiérrez-Fandiño, Asier, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2022. MarIA: Spanish Language Models, July. doi:10.26342/2022-68-3.

Hidayat, Tirta Hema Jaya, Yova Ruldeviyani, Achmad Rizki Aditama, Gusti Raditia Madya, Ade Wija Nugraha, and Muhammad Wijaya Adisaputra. 2021. Sentiment Analysis of Twitter Data Related to Rinca Island Development Using Doc2Vec and SVM and Logistic Regression as Classifier. In *Procedia Computer Science*, 197:660–667. Elsevier B.V. doi:10.1016/j.procs.2021.12.187.

Huertas-Tato, Javier, Alejandro Martin, and David Camacho. 2023. BERTuit: Understanding Spanish Language in Twitter through a Native Transformer. *Expert Systems* 40 (9). doi:10.1111/exsy.13404.

Hutto, C J, and Eric Gilbert. 2014. *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text*. http://sentic.net/.

Jiao, Xiaoqi, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for Natural Language Understanding, September. http://arxiv.org/abs/1909.10351.

Kannan, Eswariah, and Lakshmi Anusha Kothamasu. 2022. Fine-Tuning BERT Based Approach for Multi-Class Sentiment Analysis on Twitter Emotion Data. *Ingenierie Des Systemes d'Information* 27 (1). International Information and Engineering Technology Association:93–100. doi:10.18280/isi.270111.

Karfi, Ikram El, and Sanaa El Fkihi. *An Ensemble of Arabic Transformer-Based Models for Arabic Sentiment Analysis*. *IJACSA) International Journal of Advanced Computer Science and Applications*. Vol. 13. www.ijacsa.thesai.org.

Kwartler, Ted. 2017. *Text Mining in Practice with R*. J. Wiley & Sons Ltd.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations, September. http://arxiv.org/abs/1909.11942.

Le, Hang, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-Training for French, December. http://arxiv.org/abs/1912.05372.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://github.com/pytorch/fairseq.

Mao, Yanying, Qun Liu, and Yu Zhang. 2024. Sentiment Analysis Methods, Applications, and Challenges: A Systematic Literature Review. *Journal of King*

Saud University - Computer and Information Sciences. King Saud bin Abdulaziz University. doi:10.1016/j.jksuci.2024.102048.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A Tasty French Language Model, November. doi:10.18653/v1/2020.acl-main.645.

Merayo, Noemí, Rosalía Cotelo, Rocío Carratalá-Sáez, and Francisco J. Andújar. 2024. Applying Machine Learning to Assess Emotional Reactions to Video Game Content Streamed on Spanish Twitch Channels. *Computer Speech and Language* 88 (November). Academic Press. doi:10.1016/j.csl.2024.101651.

Montesinos, Osval A., Ignacio L. Espinoza, Carlos M. Hernandez, and Miguel A. Tinoco. 2010. *Muestreo Estadístico Tamaño de Muestra y Estimación de Parámetros*. Universidad de Colima.

Murphy, Kevin. 2012. *Machine Learning : A Probabilistic Perspective*. MIT Press.

Nguyen, Dat Quoc, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A Pre-Trained Language Model for English Tweets, May. http://arxiv.org/abs/2005.10200.

Papadimitriou, Isabel, Kezia Lopez, and Dan Jurafsky. 2022. Multilingual BERT Has an Accent: Evaluating English Influences on Fluency in Multilingual Models, October. http://arxiv.org/abs/2210.05619.

Pérez, Juan Manuel, Damián A. Furman, Laura Alonso Alemany, and Franco Luque. 2021. RoBERTuito: A Pre-Trained Language Model for Social Media Text in Spanish, November. http://arxiv.org/abs/2111.09453.

Pota, Marco, Mirko Ventura, Rosario Catelli, and Massimo Esposito. 2021. An Effective Bert-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian. *Sensors (Switzerland)* 21 (1). MDPI AG:1–21. doi:10.3390/s21010133.

Serrano, Alejandro Vaca, Guillem Garcia Subies, Helena Montoro Zamorano, Nuria Aldama Garcia, Doaa Samy, David Betancur Sanchez, Antonio Moreno Sandoval, Marta Guerrero Nieto, and Alvaro Barbero Jimenez. 2022. RigoBERTa: A State-of-the-Art Language Model For Spanish, April. http://arxiv.org/abs/2205.10233.

Teodorescu, Horia-Nicolai, and Navanath Saharia. 2015. *An Internet Slang Annotated Dictionary and Its Use in Assessing Message Attitude and Sentiments*. www.slangguide.com.

Utitiaj, Ismael, Paulina Morillo, and Diego Vallejo Huanga. 2020. Sentiment Analysis Tool for Spanish Tweets in the Ecuadorian Context. In *ACM International Conference Proceeding Series*. Association for Computing Machinery. doi:10.1145/3446132.3446424.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need, June. http://arxiv.org/abs/1706.03762.

Virtanen, Antti, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual Is Not Enough: BERT for Finnish, December. http://arxiv.org/abs/1912.07076.

Wu, Jun, Tianliang Zhu, Jiahui Zhu, Tianyi Li, and Chunzhi Wang. 2023. A Optimized BERT for Multimodal Sentiment Analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications* 19 (2s). Association for Computing Machinery (ACM):1–12. doi:10.1145/3566126.