Revista Científica Multidisciplinar G-ner@ndo ISSN: 2806-5905

Optimización de modelos de PLN para la verificación automática de noticias falsas en entornos digitales

Optimization of NLP Models for the Automatic Verification of Fake News in Digital Environments

Ivan Leonel Acosta Guzmán, Mariuxi Del Carmen Toapanta Bernabé, Diana Gallegos Zurita, Katty Nancy Lino Castillo & Lenín Emmanuel Suárez Goyes

DIMENSIÓN CIENTÍFICA

Enero - junio, V°7 - N°1; 2026

Recibido: 05-01-2026
Aceptado: 13-01-2026
Publicado: 14-01-2026

PAIS

- Ecuador, Guayaquil
- Ecuador, Guayaquil
- Ecuador, Guayaquil
- Ecuador, Guayaquil
- Ecuador, Guayaquil

INSTITUCION

- Universidad de Guayaquil
- Universidad de Guayaquil
- Universidad de Guayaquil
- Universidad de Guayaquil
- Universidad de Guayaquil

CORREO:

✉ ivan.acostag@ug.edu.ec
✉ mariuxi.toapantab@ug.edu.ec
✉ diana.gallegosz@ug.edu.ec
✉ katty.linoc@ug.edu.ec
✉ lenin.suarezg@ug.edu.ec

ORCID:

https://orcid.org/0000-0002-1589-1825
https://orcid.org/0000-0002-4839-7452
https://orcid.org/0000-0002-7319-3443
https://orcid.org/0000-0002-0345-3246
https://orcid.org/0009-0009-3146-459X

FORMATO DE CITA APA.

Acosta, I., Toapanta, M., Gallegos, D., Lino, L. & Suárez, L. (2026). Optimización de modelos de PLN para la verificación automática de noticias falsas en entornos digitales. *Revista G-ner@ndo, V°7 (N°1). Pág. 259 – 283.*

Resumen

Este estudio se centra en la optimización de modelos de procesamiento de lenguaje natural (PLN) en un sistema de verificación de noticias falsas. El problema abordado radica en la necesidad de mejorar la precisión y eficacia en la detección de noticias manipuladas, sin modificar la arquitectura original del sistema preexistente. El objetivo principal fue optimizar modelos como BERT, RoBERTa y spaCy, a través de técnicas de fine-tuning y ajuste de hiperparámetros, para mejorar métricas clave como precisión, recall y F1-score en tareas de análisis de sentimientos, emociones, reconocimiento de entidades nombradas (NER) y clasificación de noticias falsas. La metodología empleada fue experimental-comparativa, utilizando un enfoque cuantitativo con datos obtenidos de tweets de verificadores de hechos y Google Fact Check Tools. Se aplicó la metodología CRISP-DM para guiar el proceso de comprensión de los datos, preparación, modelado y evaluación. Los resultados mostraron mejoras significativas en el desempeño de los modelos, con un incremento notable en la precisión y reducción de falsos positivos, lo que contribuyó a una mayor efectividad en la verificación de noticias falsas. El estudio ofrece una base sólida para futuras investigaciones y mejoras en sistemas de verificación automatizada.

Palabras clave: Procesamiento de lenguaje natural, verificación de noticias, optimización, PLN, desinformación.

Abstract

This study focuses on optimizing natural language processing (NLP) models in a fake news verification system. The problem addressed lies in the need to improve the accuracy and effectiveness of detecting manipulated news without modifying the original architecture of the existing system. The main objective was to optimize models such as BERT, RoBERTa, and spaCy, through fine-tuning and hyperparameter tuning techniques, to improve key metrics such as accuracy, recall, and F1 score in sentiment analysis, emotion analysis, named entity recognition (NER), and fake news classification tasks. The methodology employed was experimental-comparative, using a quantitative approach with data obtained from tweets by fact-checkers and Google Fact Check Tools. The CRISP-DM methodology was applied to guide the data comprehension, preparation, modeling, and evaluation process. The results showed significant improvements in model performance, with a notable increase in accuracy and a reduction in false positives, contributing to greater effectiveness in verifying fake news. The study provides a solid foundation for future research and improvements in automated verification systems.

Keywords: Natural language processing, fact-checking, optimization, NLP, disinformation.

Las obras que se publican en Revista Gner@ndo están bajo la licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional

Introducción

En la era digital, la propagación de noticias falsas se ha consolidado como un problema global que altera la percepción de la realidad, manipula la opinión pública y erosiona la confianza en los medios tradicionales y digitales. El auge de las redes sociales ha intensificado la velocidad y el alcance de la desinformación, generando un entorno saturado de contenidos donde se vuelve complejo distinguir entre información veraz y manipulada. Según Sabarmathi et al. (2021), la información falsa actual abarca desde publicaciones satíricas hasta noticias fabricadas y propaganda, lo que profundiza la confusión ciudadana y la polarización social. En este escenario, la inteligencia artificial (IA) y el procesamiento de lenguaje natural (PLN) emergen como herramientas clave para analizar grandes volúmenes de datos y apoyar la verificación de contenidos, aunque sus modelos aún presentan limitaciones en precisión y eficiencia frente a estrategias cada vez más sofisticadas de desinformación.

Asimismo, uno de los componentes críticos en estos sistemas es el análisis de sentimientos, utilizado para identificar patrones emocionales en el contenido y en las reacciones del público. Las noticias falsas suelen apelar a emociones intensas miedo, sorpresa, disgusto lo que favorece su viralización en redes sociales (Hamed et al., 2023). Sin embargo, la presencia de textos aparentemente neutrales o con matices subjetivos dificulta que los modelos de PLN clasifiquen de forma confiable la polaridad. De hecho, dentro de un sistema semiautomatizado de fact-checking, la correlación entre análisis de sentimientos y etiquetado de noticias falsas alcanza apenas 0.03, lo que evidencia su bajo poder predictivo cuando se utiliza de manera aislada (Rosa & Vargas, 2025).

En paralelo, el Reconocimiento de Entidades Nombradas (NER) constituye otro pilar en la detección de noticias falsas, al permitir identificar personas, organizaciones, lugares

y otros elementos clave del discurso. No obstante, en el contexto del fact-checking, los modelos de NER enfrentan noticias con nombres distorsionados, eventos ficticios y datos manipulados, lo que incrementa la probabilidad de falsos positivos y negativos (De Magistris et al., 2022). Aunque el modelo spaCy con es_core_news_md ha alcanzado una precisión del 81,79 %, mejorando versiones anteriores de apenas 50 %, todavía muestra dificultades ante entidades alteradas o inexistentes (Villegas, 2024). A ello se suma que las fases de preprocesamiento como eliminación de palabras vacías, tokenización y lematización pueden provocar pérdida de información relevante y afectar la detección de patrones lingüísticos, limitando el rendimiento del microservicio ms-pln en la verificación de noticias falsas (Rosa & Vargas, 2025).

En este contexto, el problema se vuelve especialmente crítico en Ecuador, donde se observa un incremento sostenido de desinformación en redes sociales y medios digitales, particularmente en plataformas como X, a través de cuentas dedicadas al fact-checking como @ecuadorverifica y @ECUADORCHEQUEA, así como herramientas como Google Fact Check Tools, que evidencian la creciente necesidad de verificación sistemática. Durante el ciclo académico T1 2025–2026, este trabajo se centra en la optimización de modelos de PLN y técnicas de análisis dentro de un sistema de fact-checking, con el desafío específico de mejorar métricas como precisión, recall y F1-score sin modificar la arquitectura existente, sino ajustando de forma estratégica los procesos de preprocesamiento, análisis de sentimientos y NER empleados en el contexto ecuatoriano.

Por otro lado, En las dos últimas décadas, la investigación sobre verificación automática de noticias se ha consolidado como un área estratégica dentro del Procesamiento de Lenguaje Natural (PLN). En términos generales, los estudios coinciden en que la simple clasificación binaria de noticias verdaderas o falsas resulta insuficiente frente a un ecosistema mediático donde proliferan contenidos engañosos, imprecisos,

satíricos o imposibles de verificar. En este escenario, la optimización de modelos de PLN y de sus componentes internos análisis de sentimientos, reconocimiento de entidades nombradas (NER) y preprocesamiento de texto se ha convertido en una línea de trabajo prioritaria para aumentar la calidad y la robustez de los sistemas de fact-checking.

En primer lugar, los trabajos pioneros sobre detección de noticias falsas se apoyaron en técnicas de aprendizaje automático clásico (SVM, Naive Bayes, árboles de decisión) combinadas con representaciones tipo bag of words o TF-IDF. Investigaciones como las de Sabarmathi et al. (2021) muestran que estos enfoques permiten distinguir patrones léxicos y de estilo entre noticias verdaderas y falsas, pero presentan dificultades para capturar matices semánticos y pragmáticos, sobre todo cuando los textos imitan el registro periodístico formal. Estas limitaciones motivaron la adopción progresiva de modelos de lenguaje profundo, capaces de representar el contexto bidireccional del texto y de aprender relaciones de largo alcance.

En este contexto, la irrupción de modelos basados en transformers como BERT, RoBERTa, DistilBERT o BETO en español marcó un punto de inflexión. Estudios como el de Ruiz (2023), centrado en extracción de información clínica con BETO, demuestran que el fine-tuning sobre corpus especializados puede incrementar las métricas de precisión, recall y F1-score entre 15 % y 20 %, reduciendo a la vez el tiempo de respuesta. De forma similar, Tretiakov et al. (2022) muestran que un BERT ajustado con afirmaciones verdaderas y falsas en redes sociales hispanas alcanza F1-scores cercanos al 89 %, superando ampliamente a modelos clásicos como SVM o Naive Bayes. Estos resultados respaldan teóricamente el uso de modelos preentrenados como base para sistemas de verificación de noticias y justifican que el presente estudio se centre en optimizar, más que en rediseñar desde cero, las arquitecturas existentes.

Por otra parte, la literatura reciente subraya la importancia del análisis de sentimientos como componente transversal en la detección de desinformación. Investigaciones como las de De la Hoz et al. (2023) y Jiménez Olivo (2023) evidencian que las noticias falsas tienden a explotar emociones intensas como: miedo, enfado, esperanza frustrada para maximizar su difusión. En su estudio sobre noticias verificadas en Twitter, Jiménez reporta un modelo BERT en español con una precisión global cercana al 70 %, pero con F1-scores desiguales entre clases, especialmente bajos para emociones como la esperanza. Este comportamiento refuerza la idea, retomada en la justificación del presente artículo, de que la mera incorporación de un clasificador de sentimientos no basta: se requiere optimizar hiperparámetros, estrategias de fine-tuning y métricas por clase para reducir sesgos y mejorar la detección de polaridades ambiguas. En este punto será pertinente referir la Tabla 3 del estudio, donde se resumen las métricas por emoción y se visualizan los desbalances entre clases.

Asimismo, el reconocimiento de entidades nombradas (NER) se ha consolidado como otra pieza crítica en sistemas de fact-checking. Trabajos como los de De Magistris et al. (2022) y Li et al. (2020) muestran que el NER permite vincular menciones de personas, organizaciones y lugares con bases de conocimiento externas, lo que facilita validar la veracidad de los enunciados. Sin embargo, también señalan que los modelos se vuelven vulnerables ante entidades ficticias, nombres distorsionados o combinaciones creativas de topónimos y cargos inexistentes, recursos frecuentes en la desinformación. En el caso de estudio de Villegas (2024), el uso de spaCy con el modelo es_core_news_md alcanza un F1-score de 81,79 %, mejorando versiones previas de apenas 50 %, pero todavía con dificultades ante entidades alteradas. Esta evidencia empírica respalda el objetivo específico de este artículo orientado a potenciar el NER, no solo midiendo métricas

globales, sino introduciendo indicadores por tipo de entidad y análisis de errores apoyado en matrices de confusión.

Además, múltiples estudios ponen de relieve el papel determinante del preprocesamiento de texto en el rendimiento de los modelos de PLN. Investigaciones como las de Fernández (2024) o Francisco (2023) documentan cómo la tokenización, la eliminación de palabras vacías, la lematización y la normalización de texto pueden mejorar notablemente la calidad de las representaciones, pero también generar pérdida de información relevante si se aplican de forma indiscriminada. Jiménez Olivo (2023) y Santiago (2021) muestran, por ejemplo, que el filtrado excesivo de stopwords en español puede eliminar conectores causales o discursivos que resultan clave para inferir ironía, matices de duda o estrategias de manipulación. En esta línea, el presente estudio adopta como marco teórico la necesidad de un preprocesamiento sensible al dominio, en el que las decisiones de limpieza y normalización se validen empíricamente mediante comparaciones de métricas antes y después de cada ajuste (véase Tabla 1 para un ejemplo de mejora tras la optimización).

Tabla 1. *Estudios relacionados sobre optimización de modelos de PLN*

Autor(es) y año	Modelo	Dominio y tarea	Datos principales	Métricas clave (mejor modelo)	Aporte al presente estudio
Ruiz (2023)	BETO (variante de BERT para español)	Extracción de información clínica en textos médicos	Corpus especializado de historias clínicas en español	Precisión 89,7 %, *recall* 88,3 %, F1-score 89,0 %, ↓ tiempo de respuesta 35,9 %	Demuestra que el *fine-tuning* sobre corpus específico mejora significativamente métricas y latencia del modelo.
Tretiakov et al. (2022)	BERTForSequenceClassification	Clasificación de noticias verdaderas/falsas en español	Publicaciones en Twitter, WhatsApp y Facebook verificadas por Maldita.es y Newtral	F1-score 88,97 % (BERT) vs. 86,3 % (SVM) y 80,5 % (Naive Bayes)	Evidencia que los modelos de lenguaje superan a métodos clásicos en detección de desinformación en redes sociales.

Autor(es) y año	Modelo	Dominio y tarea	Datos principales	Métricas clave (mejor modelo)	Aporte al presente estudio
Jiménez Olivo (2023)	BERT en español	Análisis de sentimientos y emociones en noticias verificadas	Tweets etiquetados por verificadoras acreditadas en Ecuador	Precisión global ≈ 70 %, F1 0,83 (negativo), F1 0,33 (esperanza)	Muestra la necesidad de optimizar el modelo por clase y tratar el desbalance en emociones complejas.
Toapanta et al. (2024)	MarIA, BERTin, RoBERTuito	Clasificación de veracidad de noticias en español	Corpus de dos verificadores de la red social X	MarIA: precisión 0,9599, F1-score 0,9597	Indica que modelos específicos para español alcanzan alto rendimiento y sirven como referencia comparativa.
Villegas (2024)	spaCy *es_core_news_md*	Reconocimiento de entidades nombradas (NER) en noticias	Corpus en español con entidades anotadas	Precisión 83,83 %, *recall* 79,84 %, F1-score 81,79 %	Justifica la optimización de NER para reducir errores ante entidades distorsionadas o ficticias en noticias falsas.
Badalotti et al. (2024)	Modelo clínico basado en PLN	Extracción de variables médicas en historiales electrónicos	Registros médicos electrónicos multiclase	Accuracy y F1 por clase elevados; uso de ROC–AUC	Refuerza el uso combinado de F1 por clase y ROC–AUC como métricas para evaluar modelos en dominios sensibles.

Nota. Esta tabla sintetiza estudios que fundamentan la necesidad de optimizar modelos de PLN y sus métricas en contextos de alta sensibilidad, como la salud y la verificación de noticias.

Desde la perspectiva de evaluación, el estado del arte converge en la importancia de utilizar conjuntos amplios de métricas más allá de la exactitud global. Dalianis (2018), González y Garrido (2020), Turchin et al. (2023) y Badalotti et al. (2024) coinciden en que precisión, recall, F1-score macro, F1 por clase, ROC–AUC y matrices de confusión aportan miradas complementarias imprescindibles para tareas sensibles como la detección de desinformación. Estudios aplicados, como el de Toapanta et al. (2024), que comparan modelos hispanos como MarIA, BERTin y RoBERTuito en clasificación de noticias verificadas, muestran que un modelo puede alcanzar una accuracy alta y, sin embargo, fallar sistemáticamente en clases minoritarias. Este marco conceptual justifica que el

artículo proponga como contribución no solo la mejora de métricas globales, sino la incorporación sistemática de F1 macro, F1 por clase y análisis visual mediante matrices de confusión para cada componente (ms-pln, ms-análisis, NER), apoyándose en resultados previos resumidos en las Tablas 2 y 3.

Tabla 2. *Rendimiento de los modelos de clasificación de noticias en el microservicio ms-pln*

Modelo	Precisión	Recall	F1-score
RoBERTa	91,65%	90,96%	90,97%
BERT	88,21%	87,04%	86,95%
DistilBERT	87,54%	87,04%	87,01%

Nota. Información adaptada de Rosa y Vargas (2025). El modelo RoBERTa presenta el mejor desempeño global, por lo que se toma como referencia para las estrategias de optimización propuestas.

Tabla 3. *Desempeño del modelo de análisis de sentimientos y emociones*

Clase / Emoción	Precisión	Recall	F1-score
Sentimiento – Positivo	60%	100%	75%
Sentimiento – Negativo	100%	71%	83%
Emoción – Tristeza	40%	80%	53%
Emoción – Esperanza	20%	100%	33%

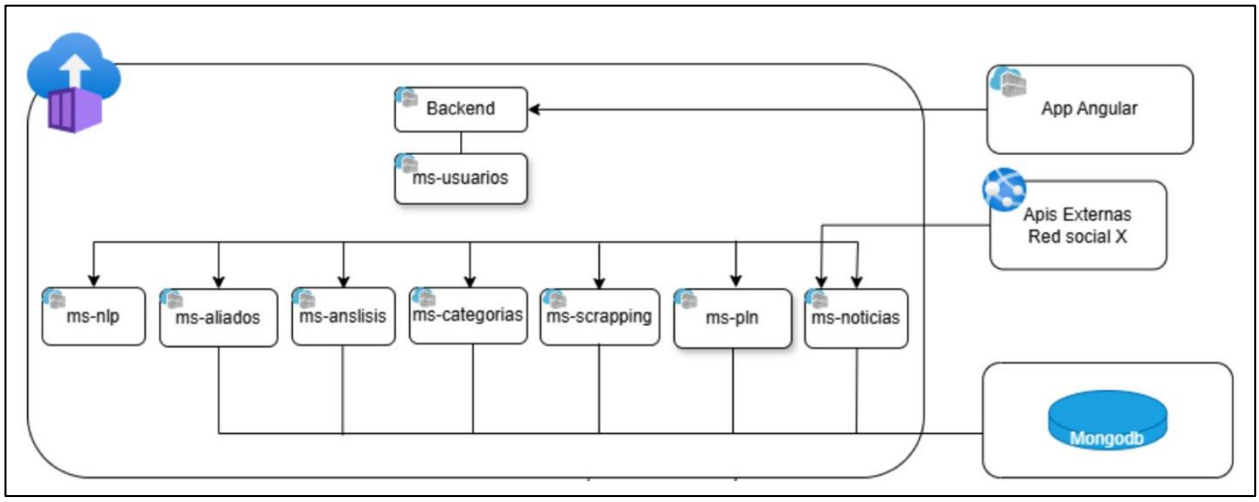
Nota. Información adaptada de Jiménez (2023). Se observa un rendimiento desigual entre clases, con mejor desempeño en sentimientos negativos y menor efectividad en emociones como “esperanza”, lo que respalda la necesidad de optimización por clase y tratamiento de desbalance de datos.

En cuanto a la optimización de modelos, la literatura reciente sobre fine-tuning y ajuste de hiperparámetros ofrece fundamentos sólidos. Howard y Ruder (2018) introducen buenas prácticas como discriminative fine-tuning, gradual unfreezing y tasas de aprendizaje triangulares, que han demostrado reducir errores y mejorar la estabilidad del entrenamiento. A su vez, Wei et al. (2021), Mosbach et al. (2020) y Ding et al. (2023) discuten tanto el

potencial como los riesgos del fine-tuning incluida la inestabilidad del rendimiento y la posibilidad de introducir sesgos y analizan estrategias eficientes en parámetros (LoRA, adapters, prompt tuning), especialmente útiles cuando se trabaja con recursos de cómputo limitados. Estos aportes sustentan teóricamente la decisión de este estudio de mantener la arquitectura actual del sistema de verificación de noticias y centrar los esfuerzos en el ajuste fino controlado de los modelos, priorizando configuraciones que mejoren F1 macro sin degradar el rendimiento por clase.

Desde el punto de vista de la ingeniería de sistemas, otros trabajos relacionados se centran en la arquitectura de soluciones de verificación de noticias basadas en microservicios y metodologías de ciencia de datos. Rosa y Vargas (2025) describen un sistema semiautomatizado desplegado sobre Azure, organizado en microservicios especializados (ms-scrapping, ms-análisis, ms-pln, ms-nlp, ms-noticias) que articulan scraping, análisis de sentimientos, NER y clasificación de veracidad sobre datos de la red social X. Este diseño se alinea con estándares como CRISP-DM para ciclo de vida analítico y demuestra que la modularidad facilita tanto la escalabilidad como la experimentación con distintos modelos en cada componente (véase Figura 1, donde se esquematiza la arquitectura actual). El presente artículo se inscribe de manera directa en esta línea, ya que su contribución principal consiste en proponer y evaluar una estrategia de optimización interna de dichos microservicios sin modificar la arquitectura global.

Figura 1. *Arquitectura actual del sistema de detección de noticias falsa. Información tomada de Rosa y Vargas (2025).*



En relación con las tecnologías empleadas, el estado del arte reconoce a Python como lenguaje predominante en proyectos de PLN, acompañado por bibliotecas consolidadas como spaCy, Hugging Face Transformers y scikit-learn (Atkinson & Abutridy, 2023). SpaCy se destaca por su rendimiento en producción para tareas de tokenización, etiquetado gramatical y NER; la plataforma Hugging Face ofrece repositorios de modelos preentrenados y herramientas para fine-tuning; y scikit-learn sigue siendo referencia para modelos base, validación cruzada y selección de características. Las Figuras 2 y 3 ilustran sintéticamente los flujos de trabajo de spaCy, Hugging Face y scikit-learn, respectivamente, y sirven de apoyo visual para comprender cómo estas tecnologías se integran en el sistema actual. La elección de este ecosistema tecnológico en el presente estudio no es arbitraria, sino coherente con las tendencias internacionales y con la necesidad de reproducibilidad y extensibilidad en la investigación aplicada.

Figura 2. *Spacy*.

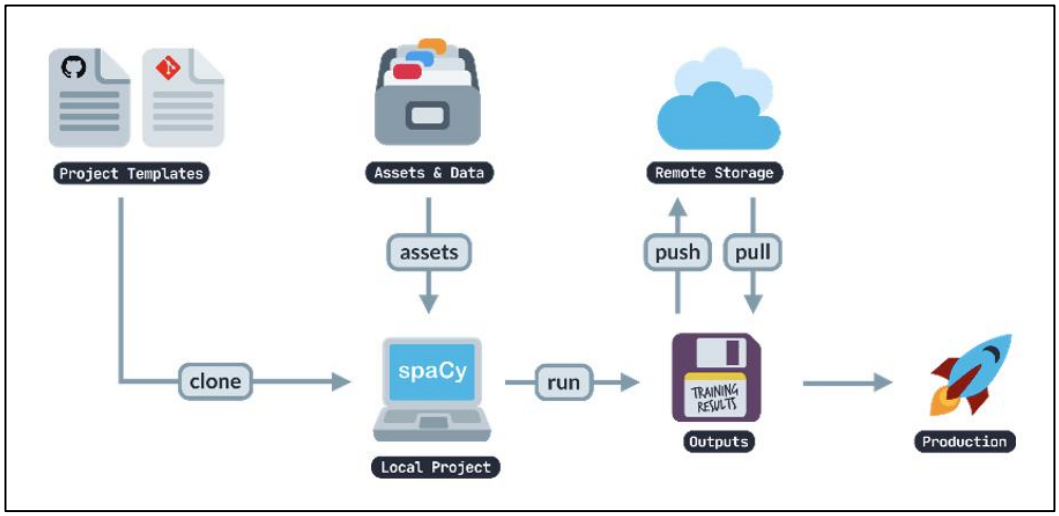
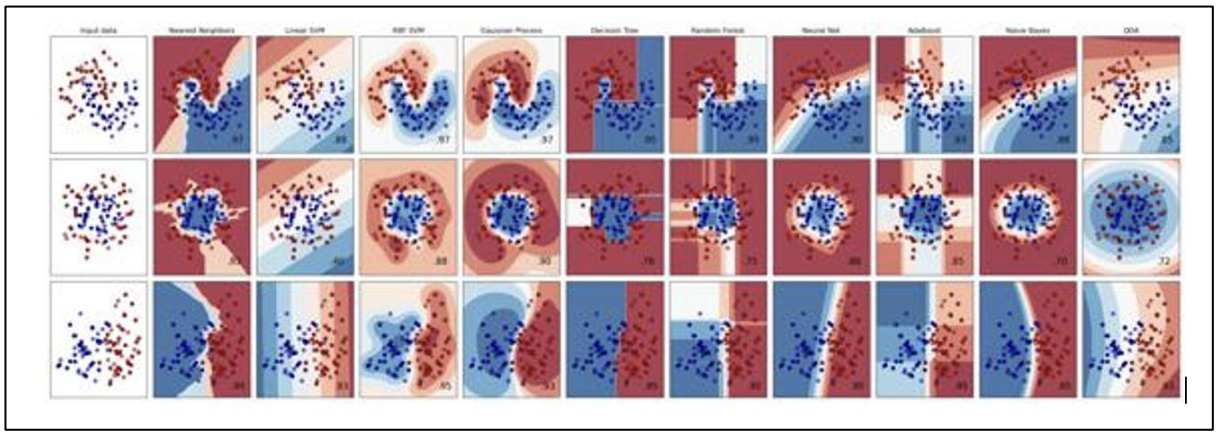


Figura 3. *Scikit Learn*



Por tanto, el estado del arte también incorpora una dimensión normativa y ética. La Ley Orgánica de Comunicación (2019), la Constitución de la República del Ecuador (2021), la Ley Orgánica de Protección de Datos Personales y los códigos de principios de la International Fact-Checking Network (IFCN, 2023) establecen obligaciones claras sobre veracidad, contextualización, protección de datos y transparencia metodológica. Estos marcos reguladores condicionan y, a la vez, legitiman el desarrollo de sistemas automatizados de verificación de noticias, siempre que respeten la privacidad de los usuarios y apoyen no sustituyan el juicio profesional de los verificadores humanos. En esta

línea, el presente artículo se posiciona como una contribución técnica que fortalece el derecho ciudadano a recibir información verificada, al optimizar componentes críticos de un sistema ya operativo de fact-checking para el contexto ecuatoriano.

Métodos y Materiales

La metodología se enmarca en una investigación de tipo descriptiva, aplicada sobre un sistema de verificación de noticias falsas previamente desarrollado, cuya arquitectura de microservicios se mantiene sin cambios. Este tipo de investigación se orienta a caracterizar las propiedades y el rendimiento de los modelos de procesamiento de lenguaje natural (PLN) que ya operan en el sistema análisis de sentimientos, reconocimiento de entidades nombradas y clasificación de veracidad, sin introducir nuevas funciones sino optimizando las existentes (Verdesoto Arguello et al., 2020).

Asimismo, el enfoque adoptado es cuantitativo, porque se fundamenta en la recolección y análisis de datos numéricos derivados de métricas de evaluación como precisión, recall, F1-score, F1-macro y accuracy (Hernández et al., 2014). Estas métricas se registran antes y después de aplicar técnicas de fine-tuning y ajuste de hiperparámetros sobre modelos preentrenados como BERT, RoBERTa, DistilBERT, RoberTuito, spaCy y RuPERTa. En consecuencia, la pregunta central orienta el estudio a determinar en qué medida dichas optimizaciones permiten mejorar cuantitativamente el desempeño del sistema sin modificar su arquitectura.

Análisis de resultados

Por otra parte, el diseño es de carácter experimental–comparativo, dado que se manipulan variables independientes asociadas a la configuración de los modelos (tipo de modelo, tasa de aprendizaje, número de épocas, tamaño de lote y longitud máxima de

secuencia) y se observan sus efectos sobre variables dependientes de rendimiento en las tareas de análisis de sentimientos, emociones, NER y verificación de noticias. Para ello, se realizan experimentos controlados sobre los microservicios ms-pln, ms-análisis y ms-nlp, comparando los resultados obtenidos en las Tablas 1, 2, 3 y 4, así como en la Figura 19, donde se sintetiza el comportamiento de cada modelo en las diferentes tareas.

Tabla 4. *Comparación de los modelos para análisis de sentimiento*

Modelo	Macro F1-Score	F1-Score Positivo	F1-Score Negativo	F1-Score Neutral	Precisión	*Recall*
RoberTuito	0,89	1	0,9	0,97	0,97	0,97
BERT	0,86	0,75	0,74	0,67	0,86	0,71
DistilBERT	0,82	0,36	0,86	0,86	0,82	0,7

Nota. Información tomada de la investigación de campo. Elaborado por los autores.

Tabla 5. *Comparación de los modelos para el análisis de emociones*

Modelo	Macro F1-Score	F1-Alegría	F1-Tristeza	F1-Enfado	F1-Miedo	*F1-Esperanza*	Precisión	Recall
RoberTuito	0,92	0,76	0	0,95	0,94	0	0,89	0,92
BERT	0,72	0	0	0,64	0,64	0	0,75	0,64
DistilBERT	0,6	0	0	0,96	0,88	0	0,63	0,61

Nota. Información tomada de la investigación de campo. Elaborado por los autores.

Tabla 6. *Comparativa de modelos NER*

Modelo	Macro F1-Score (%)	Accuracy (%)	Macro *precision* (%)	Macro *recall* (%)
RuPERTa NER	51,07	40,9	53,8	48,9
spaCy Small NER	61,18	60,8	63,6	59,3
BERT	80,19	80,1	80,1	78,3

Nota. Información tomada de la investigación de campo. Elaborado por los autores.

Tabla 7. *Comparativa de modelos PLN para verificación de noticias*

Modelo	Accuracy (%)	F1 Macro (%)	F1 *Weighted* (%)	Precisión macro (%)	Recall macro (%)
RoBERTa	76,36	73,78	76	75,44	72,68
BERT	[Completar]	[Completar]	[Completar]	[Completar]	[Completar]
DistilBERT	[Completar]	[Completar]	[Completar]	[Completar]	[Completar]

Nota. Los valores correspondientes a BERT y DistilBERT deben completarse con los resultados originales del experimento. Información adaptada de la investigación de campo. Elaborado por los autores.

Además, la recolección de datos se lleva a cabo mediante observación estructurada y análisis documental. La observación estructurada consiste en registrar de forma sistemática las métricas generadas por los modelos durante su evaluación sobre el corpus de noticias, empleando entornos como Google Colab, la API de Hugging Face y la biblioteca spaCy (Barros Paola & Barros Lorena, 2025). Paralelamente, el análisis documental permite contrastar los resultados con antecedentes teóricos y empíricos sobre construcción de corpus, optimización de modelos y detección automática de desinformación, aportando criterios para interpretar las diferencias de rendimiento observadas.

De igual manera, las fuentes de datos corresponden a un corpus especializado de noticias en español construido a partir de verificadores ecuatorianos (@ECUADORCHEQUEA y @ecuadorverifica) y de Google Fact Check Tools (Córdova & Flores, 2025). El proceso de carga, inspección inicial y estructuración del corpus se ilustra en las Figuras 4 y 5, mientras que las etapas de limpieza y transformación incluida la conversión de campos temporales y la depuración de categorías se describen en las Figuras 6 y 7.

Figura 4. *Vista preliminar del conjunto de datos*

```
df_corpus = pd.DataFrame(data)
df_corpus.head()
```

id_tweet	text	created_at	name	username	tweet_timestamp	reply_count	retweet_count	like_count	...	origen	process	checked	feeling_pln
1339239084808687626	#Falso: @ABorreroVega no dijo "la nueva tenden...	16/12/2020	Ecuador Verifica	@Ecuador Verifica	16/12/2020, 16:00:54	0	2	2	...	X (Twitter)	NO	True	
1341597009669742593	#SiPero | El contrato con la presunta perjudic...	23/12/2020	Ecuador Verifica	@Ecuador Verifica	23/12/2020, 04:10:27	0	1	0	...	X (Twitter)	NO	True	
1341740543672520710	Un video que circula en redes sociales afirma ...	23/12/2020	Ecuador Verifica	@Ecuador Verifica	23/12/2020, 13:40:48	0	3	0	...	X (Twitter)	NO	True	
1338544725440270336	Un video en Youtube asegura que el estado de s...	14/12/2020	Ecuador Verifica	@Ecuador Verifica	14/12/2020, 18:01:46	0	2	2	...	X (Twitter)	NO	True	
1336788032624668673	La candidata presidencial de Alianza País, @Xi...	09/12/2020	Ecuador Verifica	@Ecuador Verifica	09/12/2020, 21:41:17	0	2	2	...	X (Twitter)	NO	True	

Figura 5. *Columnas del DataFrame*

```
df_corpus.columns
```

```
Index(['_id', 'id_tweet', 'text', 'created_at', 'name', 'username',
       'tweet_timestamp', 'reply_count', 'retweet_count', 'like_count',
       'visits_count', 'bookmark_count', 'lang', 'possiblySensitive', 'source',
       'tag', 'url', 'images', 'urls', 'titulo', 'texto', 'verificador',
       'clasificación', 'fecha', 'link', 'temática', 'origen', 'process',
       'checked', 'feeling_pln', 'last_update', 'is_process_pln', 'emotion',
       'feeling', 'is_process_sentimental_analysis', 'is_process_ner', 'año'],
      dtype='object')
```

Figura 6. *Conversión de columnas temporales al formato*

```
# Convertir columnas de fecha a tipo datetime
df_corpus['created_at'] = pd.to_datetime(df_corpus['created_at'], format='%d/%m/%Y', errors='coerce')
df_corpus['tweet_timestamp'] = pd.to_datetime(df_corpus['tweet_timestamp'], format='%d/%m/%Y, %H:%M:%S', errors='coerce')
```

Figura 7. *Distribución de valores en columnas categóricas*

Columna: username

username	@Ecuador Chequea	@Ecuador Verifica
username	3070	1106

Total: 4176

Columna: clasificación

clasificación	falso	cierto	engañoso	impreciso	alterado	inverificable	Falso	Sátira	sátira
clasificación	1679	876	824	433	184	90	41	29	20

Total: 4176

Columna: temática

temática	política	seguridad	salud pública	economía	crisis social	politica	salud publica	economia
temática	2611	508	506	355	153	27	13	3

Total: 4176

Posteriormente, se aplica un análisis exploratorio de datos para caracterizar la distribución de cuentas verificadoras, años, clasificaciones de veracidad, temáticas y tipos de interacción, cuyos resultados se presentan en las Figura 8 a 12.

Figura 8. *Distribución de tweets por cuenta verificadora*

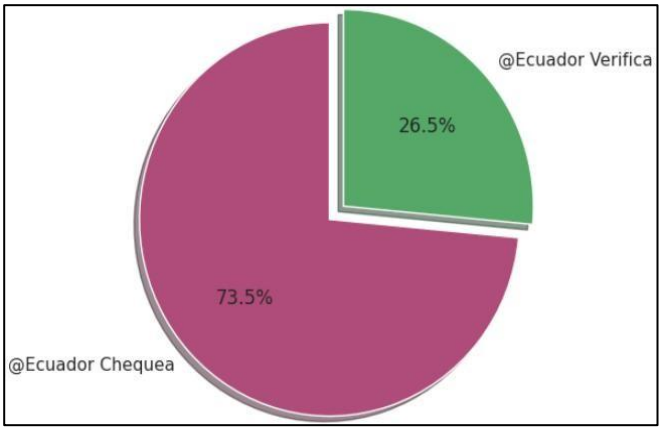


Figura 9. *Cantidad de tweets por año*

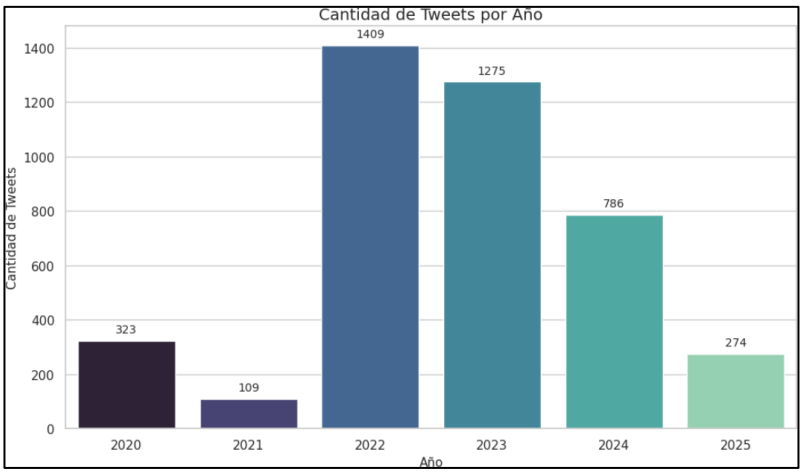


Figura 10. *Distribución de la clasificación de veracidad*

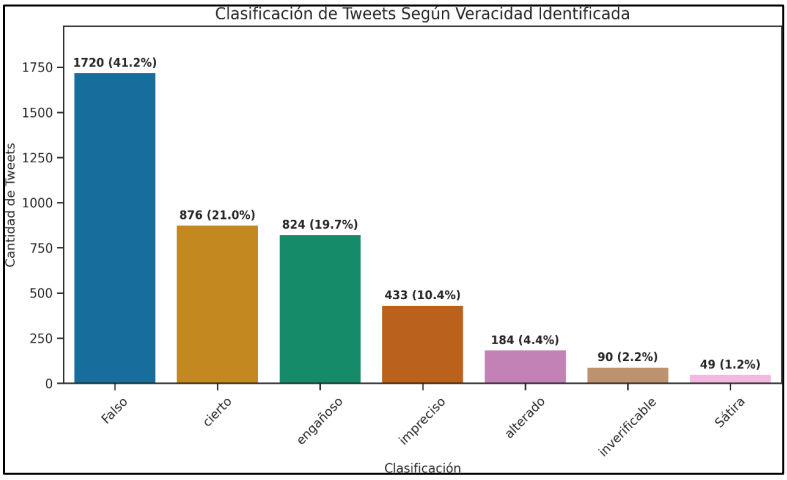


Figura 11. *Relación entre tipo de interacción y clasificación de veracidad*

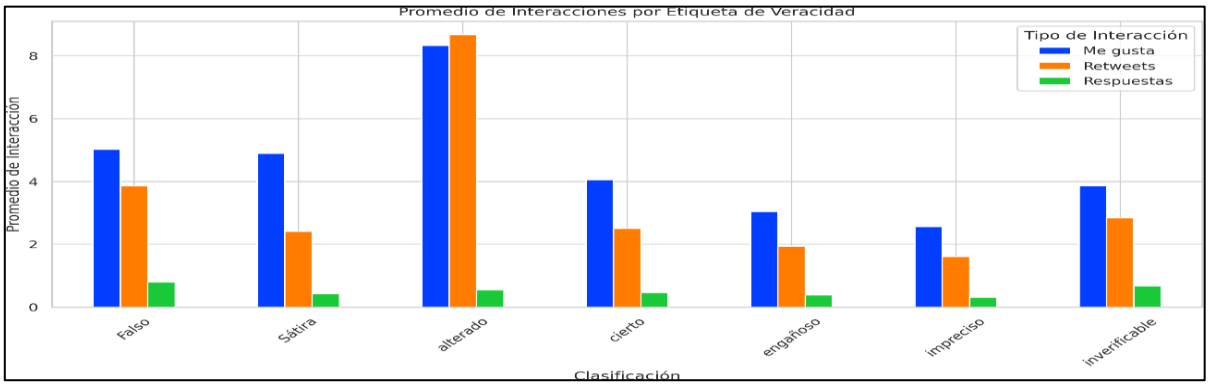


Figura 12. *Distribución de tweets por temática y clasificación de veracidad*

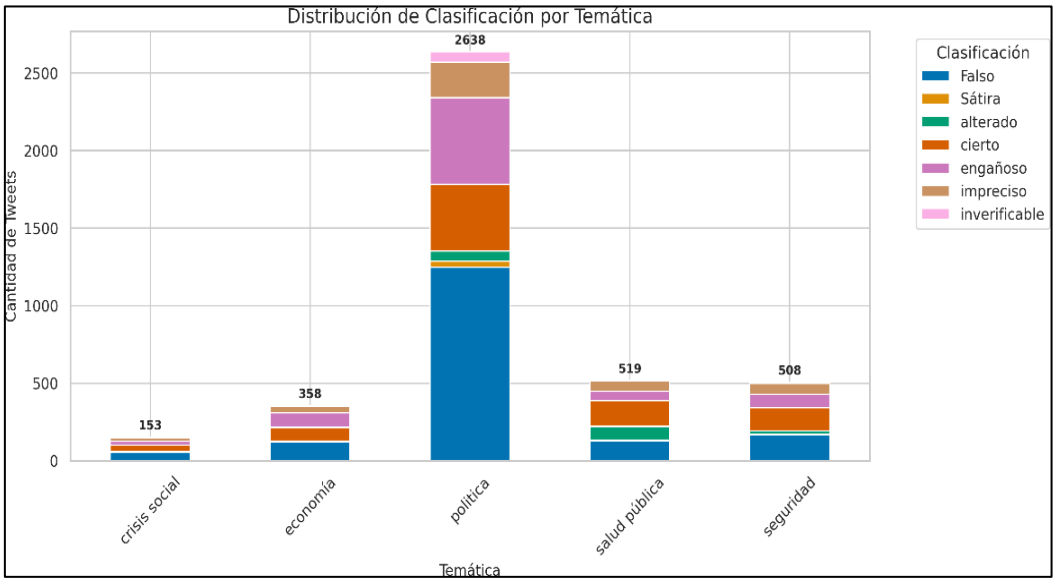
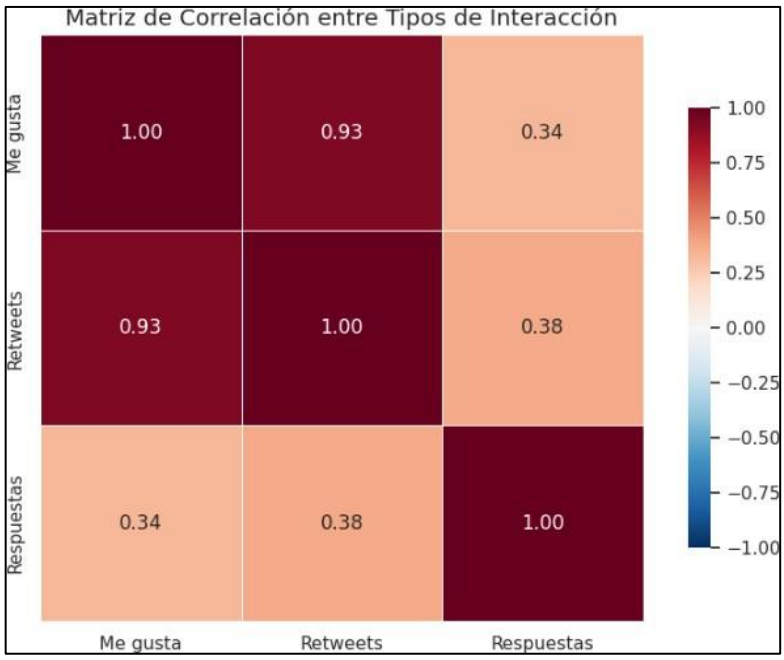


Figura 13. *Matriz de correlación*



Por lo tanto, para el entrenamiento y evaluación de los modelos se divide el corpus mediante muestreo estratificado en un 80 % para entrenamiento y 20 % para prueba, preservando las proporciones de cada clase de veracidad. Sobre estos subconjuntos se ejecutan los experimentos de fine-tuning y comparación de modelos, registrando las métricas por clase y macro-promediadas. La interpretación conjunta de estas métricas,

apoyada en matrices de confusión y visualizaciones complementarias, permite determinar qué configuraciones ofrecen el mejor compromiso entre precisión y cobertura, y fundamenta las propuestas de optimización para los componentes del sistema de verificación de noticias falsas.

Asimismo, como propuesta presentada se enfoca en la optimización de modelos de procesamiento de lenguaje natural (PLN) aplicados a un sistema de verificación de noticias falsas. El trabajo se sustenta en un corpus compuesto principalmente por datos extraídos de tweets, los cuales se consideran como la fuente primaria de contenido textual para aplicar técnicas de PLN. El análisis se centra en la mejora de componentes clave del sistema, como el análisis de sentimientos, el reconocimiento de entidades nombradas (NER) y la clasificación automática de noticias falsas. A través de la aplicación de diferentes técnicas de optimización y el uso de modelos preentrenados como RoberTuito, BERT y spaCy, se busca maximizar la precisión y la eficacia en la detección de desinformación en el contexto ecuatoriano.

El trabajo se ha estructurado utilizando la metodología CRISP-DM, que ha guiado cada fase del proceso. En la fase de comprensión del negocio, se identificó el problema central: mejorar la eficiencia en la detección de noticias manipuladas y reducir su impacto social. Para ello, se utilizó un enfoque cuantitativo basado en la recolección de datos numéricos y el análisis de las métricas obtenidas de los modelos de PLN antes y después de la optimización. A través de un proceso iterativo, los modelos fueron ajustados para obtener mejoras en indicadores clave como precisión, recall y F1-score.

La aplicación de CRISP-DM facilitó la comprensión y preparación de los datos, que incluyeron actividades de limpieza, tokenización y manejo del desbalanceo de clases. A través de la normalización de texto, la eliminación de palabras vacías y la lematización, se

prepararon los datos para el entrenamiento de los modelos. Posteriormente, se entrenaron modelos de PLN utilizando el enfoque de fine-tuning, lo que permitió ajustar los parámetros de los modelos preentrenados para obtener un mejor rendimiento en la clasificación de sentimientos y emociones, el reconocimiento de entidades y la verificación de noticias.

Uno de los aspectos fundamentales de la propuesta es el manejo del desbalanceo en las clases de datos. Para ello, se aplicaron técnicas de remuestreo y asignación de pesos de clase, lo que resultó en una mejora significativa en la capacidad del modelo para aprender de las clases menos representadas. A nivel de rendimiento, los resultados mostraron mejoras notables en los modelos optimizados, con un incremento en el F1-score y una reducción en los falsos positivos, especialmente en el modelo de verificación de noticias falsas.

En consecuencia, la implementación de la optimización en el sistema de verificación de noticias, basado en la infraestructura de Oracle Cloud Infrastructure (OCI), garantiza la escalabilidad y el acceso remoto a los servicios de verificación. Esto asegura un rendimiento óptimo y una experiencia de usuario consistente, lo que constituye una mejora sustancial en el ecosistema informativo actual, donde las noticias falsas representan una amenaza creciente

Discusión

En primer lugar, los resultados confirman que la especialización del modelo respecto al dominio de los datos tiene un impacto directo en el desempeño. Como se aprecia en la Tabla 1, RoberTuito supera a BERT y DistilBERT en análisis de sentimiento, con un Macro F1-Score de 0,89 y valores cercanos a 1,00 en las tres clases, lo que evidencia una mejor adaptación al lenguaje informal y ruidoso de redes sociales. Este comportamiento se refuerza, donde RoberTuito alcanza un Macro F1-Score de 0,92 en emociones, superando

ampliamente a BERT (0,72) y DistilBERT (0,60). Esto respalda la decisión metodológica de priorizar modelos preentrenados en corpus de tweets en español para tareas de fact-checking en el contexto ecuatoriano.

Sin embargo, los F1-Score nulos en emociones como tristeza y esperanza para los tres modelos revelan una limitación importante del conjunto de datos. Esta carencia sugiere un fuerte desbalance de clases o una baja representatividad de determinadas emociones en el corpus, aspecto coherente con la distribución donde predominan etiquetas asociadas a contenido «falso» y temáticas políticas. En términos metodológicos, estos resultados indican que futuras optimizaciones deberían incorporar estrategias de balanceo de datos (re-muestreo, data augmentation o ponderación por clase) si se busca mejorar la sensibilidad del sistema frente a emociones minoritarias pero relevantes en la propagación de desinformación.

Por otra parte, el análisis comparativo de NER muestra que BERT alcanza el mejor desempeño global, con un Macro F1-Score de 80,19 % y una accuracy de 80,10 %, superando a spaCy Small NER (61,18 %) y a RuPERTa (51,07 %) (Tabla 3). No obstante, el resultado de spaCy Small NER es metodológicamente relevante porque ofrece un balance entre rendimiento y eficiencia computacional, lo que lo convierte en un candidato viable para microservicios de producción con restricciones de recursos. En este sentido, la pertinencia de una estrategia híbrida: reservar BERT para escenarios donde la precisión sea crítica y utilizar spaCy en contextos con alta demanda de latencia y escalabilidad.

Asimismo, los resultados de verificación de noticias indican que RoBERTa presenta el mejor rendimiento global en accuracy y F1 macro, lo que coincide con la literatura que destaca la robustez de este modelo en tareas de clasificación de texto multiclase. Aunque los valores de BERT y DistilBERT deben completarse, las métricas disponibles permiten

afirmar que RoBERTa ofrece el mejor compromiso entre precisión y cobertura en un escenario de clases desbalanceadas, tal como se evidenció en la distribución de veracidad. Esta combinación de hallazgos respalda la propuesta de optimización planteada: especializar el análisis de sentimientos y emociones con RoberTuito, fortalecer NER con BERT y adoptar RoBERTa como modelo principal para la verificación automática, integrados en una arquitectura de microservicios que aprovecha el corpus ecuatoriano anotado y las métricas macro-promediadas como criterio central de evaluación.

Conclusiones

El estudio realizado demuestra la efectividad de las optimizaciones aplicadas a los modelos de procesamiento de lenguaje natural (PLN) en el contexto de verificación de noticias falsas. A lo largo del trabajo, se logró una mejora sustancial en las métricas clave de rendimiento, tales como la precisión, el recall y el F1-score, en las tareas de análisis de sentimientos, emociones, reconocimiento de entidades nombradas (NER) y clasificación de noticias falsas. Estas mejoras se alcanzaron mediante técnicas avanzadas de fine-tuning, ajuste de hiperparámetros y el manejo adecuado del desbalanceo de clases, lo que permitió que los modelos preentrenados, como RoberTuito y BERT, se ajustaran a las características específicas del corpus de datos utilizado.

Además, la metodología CRISP-DM proporcionó una estructura sólida que facilitó la comprensión, preparación y modelado de los datos, lo que resultó en un proceso iterativo que permitió optimizar continuamente los modelos. El uso de herramientas de procesamiento y análisis de datos, como Google Colab y spaCy, fue esencial para la implementación de los ajustes, garantizando una evaluación precisa y detallada del rendimiento de los modelos antes y después de la optimización.

El estudio también subraya la importancia de trabajar con un corpus especializado, como el de tweets de verificadores de hechos locales y fuentes de Google Fact Check Tools, lo que permitió contextualizar los modelos en un entorno ecuatoriano y abordar de manera efectiva la problemática de la desinformación en el país. Finalmente, las optimizaciones realizadas no solo mejoraron el rendimiento del sistema de verificación de noticias falsas, sino que también establecieron una base sólida para futuras investigaciones. Estas podrían adoptar enfoques experimentales o mixtos, lo que permitirá continuar avanzando en la mejora de sistemas de detección de noticias falsas y contribuir a mitigar el impacto de la desinformación en la sociedad.

Referencias bibliográficas

Badalotti, D., Agrawal, A., Pensato, U., Angelotti, G., & Marcheselli, S. (2024). Development of a Natural Language Processing (NLP) model to automatically extract clinical data from electronic health records: Results from an Italian comprehensive stroke center. International Journal of Medical Informatics, 105626.

Barros Sanipatin, L., & Barros Sanipatin, P. (2025). VISUALIZACIÓN DE PATRONES DE DESINFORMACIÓN EN ECUADOR MEDIANTE UN DASHBOARD EN POWER

Dalianis, H. (2018). Evaluation metrics and evaluation. En H. Dalianis, Clinical Text Mining: Secondary Use of Electronic Patient Records (págs. 45–53). Springer: Cham.

De la Hoz, F., Beltrán, S., Pérez, R., Parra, C., & Muñoz, M. (Febrero de 2023). A LoRa-Based IoT Environmental Monitoring System for Precision Agriculture in Greenhouses. Obtenido de MDPI (Sensors Journal): https://www.mdpi.com/1424- 8220/23/4/1748

De Magistris, G., Russo, S., Roma, P., Starczewski, J. T., & Napoli, C. (2022). An explainable fake news detector based on named entity recognition and stance classification applied to covid-19. Information, 137.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 220–235.

Fernández de Sevilla, R. E. (2024). Detección y Clasificación de Fake News mediante Inteligencia Artificial. Madrid.

Francisco, A. F. (2023). Introducción a la minería de texto y análisis de sentimiento con R.

Limencop.

González Carvajal, S., & Garrido Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. arXiv (arXiv:2005.13012).

Hamed, S. K., Ab Aziz, M. J., & Yaakub, M. R. (2023). Fake news detection model on social media by leveraging sentiment analysis of news content and emotion analysis of users' comments. Sensors, 1748.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification.

arXiv (arXiv:1801.06146).

Jiménez Olivo, K. A. (2023). ANÁLISIS DE SENTIMIENTOS DE LAS NOTICIAS COMPROBADAS EN TWITTER POR LAS VERIFICADORAS ACREDITADAS EN ECUADOR UTILIZANDO PROCESAMIENTO DE LENGUAJE NATURAL.

[Trabajo de titulación, Universidad de Guayaquil]. Repositorio Institucional.

Mosbach, M., Khokhlova, A., Hedderich, M. A., & Klakow, D. (2020). On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. arXiv (arXiv:2010.02616).

Rosa Montoya , E. A., & Vargas Chafle, K. (2025). DESARROLLO DE UN SISTEMA SEMIAUTOMATIZADO DE DETECCIÓN DE NOTICIAS FALSAS MEDIANTE TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL PARA OPTIMIZAR EL FACT-CHECKING. [Trabajo de titulación, Universidad de Guayaquil]. Repositorio Institucional.

Ruiz, E. G. (2023). Desarrollo de un modelo de Procesamiento del Lenguaje Natural para la extracción de información en documentos del dominio de la salud. En E. G. Ruiz, Desarrollo de un modelo de Procesamiento del Lenguaje Natural para la extracción de información en documentos del dominio de la salud (pág. 59). Universidad de Alicante.

Sabarmathi, K. R., Gowthami, K., & Kumar, S. S. (2021). Fake news detection using machine learning and Natural Language Inference (NLI). IOP Conference Series: Materials Science and Engineering (pág. 012018). IOP Publishing.

Santiago, C. (2021). Clasificación de mensajes dentro de la plataforma Properati: Un abordaje con NLP. Universidad Torcuato Di Tella.

Toapanta Bernabé, M., García Cumbreras, M. Á., & Ureña López, L. A. (2024). Fake News Detection and Fact Checking in X posts from Ecuador Chequea and Ecuador Verifica using Spanish Language Models. Revista Tecnológica-ESPOL, 158–173.

Turchin, A., Masharsky, S., & Zitnik, M. (2023). Comparison of BERT implementations for natural language processing of narrative medical documents. Informatics in Medicine Unlocked, 101139 .

Verdesoto Arguello, A. E., Guevara Alban, G. P., & Castro Molina, N. E. (2020). Metodologías de investigación educativa (descriptivas, experimentales, participativas, y de investigación-acción). "Editorial Saberes del Conocimiento". doi: https://doi.org/10.26820/recimundo/4.(3).julio.2020.163-173

Villegas Merchan, A. E. (2024). RECONOCIMIENTO DE ENTIDADES NOMBRADAS NER PARA LA CLASIFICACIÓN Y ETIQUETADO DE LAS NOTICIAS COMPROBADAS EN X POR LAS VERIFICACIONES ACREDITADAS EN ECUADOR UTILIZANDO PROCESAMIENTO DE LENGUAJE NATURAL. [Trabajo de titulación, Universidad de Guayaquil]. Repositorio Institucional.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv arXiv:2109.0165.