UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Evaluación del uso de Emojis en la Detección de Emociones en redes sociales: Un enfoque comparativo utilizando Machine Learning y Deep Learning

Esteban Nicolás López Cadena

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
INGENIERO NDUSTRIAL

Quito, 28 de abril de 2025

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Evaluación del uso de Emojis en la Detección de Emociones en redes sociales: Un enfoque comparativo utilizando Machine Learning y Deep Learning

Esteban Nicolás López Cadena

Nombre del profesor, Título académico **Danny Orlando Navarrete Chávez, M.Sc.**

Quito, 28 de abril de 2025

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombres y apellidos: Esteban Nicolás López Cadena

Código: 322235

Cédula de identidad: 1720877511

Lugar y fecha: Quito, 28 de abril de 2025

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Este estudio analiza el impacto de los emojis en la detección automática de emociones en textos de redes sociales, utilizando modelos de machine learning y deep learning. Se evaluaron cinco modelos: Naive Bayes, Random Forest, RoBERTa, BERTweet y DistilBERT, cada uno entrenado con dos versiones de un conjunto de datos de tweets: una incluyendo emojis y otra sin emojis. El preprocesamiento de datos incluyó la limpieza de texto, vectorización TF-IDF para modelos tradicionales y tokenización mediante Transformers para modelos de deep learning. Los resultados muestran que la inclusión de emojis mejora significativamente el rendimiento de los modelos, particularmente en métricas como el F1 macro. La validación cruzada de cinco particiones y las pruebas T pareadas confirmaron la significancia estadística de las mejoras observadas, excepto en el modelo BERTweet. Se reconocen limitaciones relacionadas con los recursos computacionales y el tamaño de la muestra estadística, sugiriendo líneas de investigación futuras orientadas a configuraciones más robustas y datasets multilingües.

Palabras clave: análisis de sentimientos, detección de emociones, machine learning, deep learning, emojis, procesamiento de lenguaje natural (NLP), validación cruzada.

ABSTRACT

This study analyzes the impact of emojis on automatic emotion detection in social media texts using machine learning and deep learning models. Five models were evaluated: Naive Bayes, Random Forest, RoBERTa, BERTweet, and DistilBERT, each trained on two versions of a Twitter dataset: one including emojis and another without. Data preprocessing included text cleaning, TF-IDF vectorization for traditional models, and Transformer-based tokenization for deep learning models. Results show that the inclusion of emojis significantly improves model performance, particularly in F1 macro scores. Five-fold cross-validation and paired T-tests confirmed the statistical significance of these improvements, except for the BERTweet model. Limitations related to computational resources and sample size are acknowledged, suggesting future research directions focused on more robust configurations and multilingual datasets.

Key words: sentiment analysis, emotion detection, machine learning, deep learning, emojis, natural language processing (NLP), cross-validation.

TABLA DE CONTENIDO

Introducción 08
Desarrollo del Tema 10
1. Revisión literaria 10
2. Metodología 12
2.1. Definición del problema 13
2.2. Construcción del dataset 13
2.3. Transformación del dataset 14
2.3.1. Balanceo del dataset 15
2.4 División del dataset 15
2.5 Entrenamiento del modelo 16
2.6 Evaluación del modelo 18
3. Prueba estadística para la comparación de modelos 19
3.1 Hipótesis de prueba estadística modelo DistilBERT 20
3.2 Hipótesis de prueba estadística modelo RoBERTa 20
3.3 Hipótesis de prueba estadística modelo BERTweet 20
3.4 Hipótesis de prueba estadística modelo Naive Bayes 21
3.5 Hipótesis de prueba estadística modelo Random Forest 21
4. Supuesto de normalidad 21
5. Resultados 23
5.1 Modelo Naive Bayes 23
5.2 Modelo Random Forest 23
5.3 Modelo RoBERTa 24
5.4 Modelo BERTweet 24
5.5 Modelo DistilBERT 25
5.6 Resultados pruebas T por modelo 25
6. Discusiones y limitaciones 26
7. Recomendaciones futuras 28
8. Conclusiones 29
Referencias bibliográficas (ejemplo estilo APA) 32
Anexo 1: Análisis exploratio de la base de datos 36
Anexo 2: Repositorio GitHub 36

INTRODUCCIÓN

El uso de redes sociales ha experimentado un crecimiento exponencial, convirtiéndose en una plataforma clave para la expresión de opiniones y emociones (Jagadishwari et al., 2021). En este contexto, los emojis han adquirido un papel fundamental en la comunicación digital, ya que permiten a los usuarios expresar sentimientos de manera visual y complementaria al texto (Barbieri et al., 2017). Estudios han demostrado que la inclusión de emojis en el análisis de sentimientos mejora la precisión de la clasificación, ya que estos pueden alterar o reforzar el significado del mensaje textual (Felbo et al., 2017). Sin embargo, el desafío radica en desarrollar modelos que integren adecuadamente los emojis y logren interpretar su contexto dentro del texto (Elsalam et al., 2022).

El análisis de sentimientos ha demostrado ser una herramienta fundamental para comprender el impacto y la percepción de productos, servicios y eventos. En este contexto, la incorporación de emojis en los modelos de análisis mejora la precisión al capturar valores emocionales que el texto solo no puede expresar completamente (Felbo et al., 2017). Investigaciones han indicado que los emojis pueden modificar el tono de un mensaje, haciendo que una frase neutral adquiera una connotación positiva o negativa dependiendo del emoji utilizado (Jagadishwari et al., 2021). El análisis de sentimientos y el uso de emojis tienen aplicaciones en múltiples sectores, incluyendo:

Marketing y Publicidad: Las empresas utilizan el análisis de sentimientos para evaluar la percepción de sus marcas y campañas en redes sociales, permitiéndoles ajustar estrategias y mejorar la experiencia del consumidor (Liu, 2015).

Salud Mental: Herramientas de NLP analizan mensajes en redes sociales para detectar signos de depresión, ansiedad y otros trastornos mentales, permitiendo intervenciones tempranas y apoyo personalizado (Felbo et al., 2017).

Atención al Cliente: Los chatbots y sistemas de servicio al cliente utilizan análisis de sentimientos para comprender mejor las emociones de los usuarios y responder de manera más empática y efectiva (Elsalam et al., 2022).

Política y Opinión Pública: Los emojis y el análisis de sentimientos se emplean para evaluar la opinión pública sobre candidatos, políticas gubernamentales y eventos sociales, proporcionando datos valiosos para encuestas y estudios sociológicos (Jagadishwari et al., 2021).

Este proyecto explora los enfoques más recientes en el análisis de sentimientos en redes sociales, la incorporación de emojis en la clasificación de sentimientos, la comparación de modelos de aprendizaje profundo y la aplicación de Transformers en la evaluación de comentarios estudiantiles.

DESARROLLO DEL TEMA

1. Revisión literaria

El análisis de sentimientos y emociones mediante aprendizaje automático se ha convertido en un campo vital dentro del procesamiento del lenguaje natural (PLN). Los enfoques tradicionales dependían en gran medida de la extracción manual de características y de métodos basados en léxico (Torres-Carrión et al., 2020). Sin embargo, según Akhtar et al. (2020), la aparición de las técnicas de aprendizaje automático (ML) y aprendizaje profundo (DL) ha mejorado significativamente la capacidad de capturar señales emocionales complejas del texto, especialmente cuando se incluyen elementos no verbales como los emojis.

Un estudio clave de Felbo et al. (2017) introdujo el modelo DeepMoji, que aprovecha la predicción de emojis como tarea proxy para mejorar la detección de emociones y sentimientos. Una tarea proxy es una actividad alternativa que se utiliza para entrenar modelos cuando la tarea real de interés carece de suficientes datos etiquetados (Felbo et al., 2017). En este caso, como es difícil conseguir suficientes datos bien etiquetados con emociones reales, los investigadores usaron la predicción de emojis como una forma indirecta de enseñar al modelo sobre emociones. Entrenaron un modelo de predicción de emojis bidireccional con mecanismo de atención, lo que significa que el sistema lee el texto de izquierda a derecha y también de derecha a izquierda, así entiende mejor el contexto de cada palabra, en más de 1200 millones de tuits que contenían emojis, aprendiendo representaciones emocionales con matices. El estudio concluyó que los modelos preentrenados en predicción de emojis superan a los enfoques tradicionales de aprendizaje supervisado en tareas como el análisis de sentimientos y la detección de sarcasmo, lo que destaca la utilidad de los emojis como etiquetas débiles para el contenido emocional.

Una red neuronal convolucional (CNN) es un tipo de modelo que se utiliza comúnmente para detectar patrones locales en los datos, como combinaciones frecuentes de palabras o frases (Kim, 2014). Por otro lado, una red de memoria a largo plazo (LSTM) es un tipo de red neuronal diseñada específicamente para trabajar con información secuencial, es decir, para entender el orden y el contexto en el que aparecen las palabras dentro de un texto (Yin et al., 2017)

Otra contribución importante es el trabajo de Mittal et al. (2020), quienes propusieron un modelo híbrido que combina redes neuronales convolucionales (CNN) y redes de memoria a largo plazo (LSTM) para la clasificación de emociones en datos de Twitter. Su investigación demostró que la capa CNN captura características locales como patrones de palabras, mientras que la capa LSTM modela el contexto secuencial. Descubrieron que las arquitecturas híbridas superan a los modelos independientes, y su sistema se aplicó eficazmente al monitoreo en tiempo real del estado de ánimo del público durante eventos importantes.

Choudhary y Deshmukh (2021) presentaron un análisis utilizando algoritmos tradicionales de aprendizaje automático, como máquinas de vectores de soporte (SVM), bosques aleatorios y Naïve Bayes, para la detección de emociones a partir de datos textuales. Destacaron que, si bien los modelos de aprendizaje profundo generalmente alcanzan una mayor precisión, los modelos clásicos como SVM siguen siendo competitivos cuando se entrenan con características bien diseñadas. Sus hallazgos sugieren que SVM con representación de características TF-IDF puede proporcionar una base sólida para tareas con recursos computacionales limitados, y aplicaron su modelo al análisis de la retroalimentación de los clientes.

En el campo del reconocimiento de emociones, Wang et al. (2020) propuso un enfoque principal en el aprendizaje multimodal. Su módulo de detección textual de emociones, basado

en incrustaciones BERT y arquitecturas Transformer. BERT (Bidirectional Encoder Representations from Transformers) es un modelo de lenguaje desarrollado por Google que permite comprender el significado de las palabras dentro de su contexto, al leer una oración en ambas direcciones (de izquierda a derecha y de derecha a izquierda) (Devlin et al., 2019). Por su parte, los Transformers son una arquitectura de red neuronal que revolucionó el procesamiento del lenguaje natural al introducir el mecanismo de atención. Este mecanismo permite que el modelo se enfoque en las palabras más relevantes dentro de un texto al momento de hacer una predicción, sin necesidad de procesar las palabras en orden secuencial, como ocurría con modelos anteriores como LSTM (Vaswani et al., 2017). En el estudio se demostró que los modelos de lenguaje basados en transformadores superan significativamente a los métodos tradicionales. Su sistema fue particularmente eficaz en la detección de emociones sutiles como el "asco" y el "miedo", lo que resultó útil para aplicaciones en el ámbito de la salud y la monitorización de la salud mental.

Finalmente, Alam et al. (2018) exploraron la predicción de la intensidad de las emociones mediante enfoques de regresión en lugar de la clasificación categórica. Emplearon la Regresión de Vectores de Soporte (SVR) y técnicas de conjunto para predecir la intensidad de emociones como la ira, la alegría, la tristeza y el miedo en tuits. Su estudio concluyó que los enfoques basados en regresión pueden captar la naturaleza de gradiente de las emociones mejor que la clasificación discreta, lo cual es importante en tareas con matices como la detección temprana del malestar psicológico en publicaciones en redes sociales.

En todos estos estudios, una conclusión consistente es que la inclusión de emojis, el modelado secuencial (LSTM, Transformer) y las estrategias de preentrenamiento mejoran significativamente la capacidad de los modelos para comprender y detectar emociones en textos de redes sociales. Las aplicaciones abarcan desde la monitorización de sentimientos en

tiempo real durante crisis, el seguimiento de la salud mental y la gestión de la experiencia del cliente hasta campos más complejos como el sarcasmo y la detección de intensidad.

2. Metodología

Con el objetivo de evaluar el impacto de uso de emojis en redes sociales utilizando modelos de machine learning y Deep learning se siguió la metodología de pipeline para modelos de machine learning. La metodología pipeline en problemas de machine learning se refiere a la construcción de un flujo secuencial y automatizado que integra, en un único objeto, todas las fases clave del proceso desde la definición del problema hasta la evaluación del modelo (Agrahari, 2024). Esta metodología se basa en el uso de "pipelines", que son como recetas paso a paso para procesar los datos. Lo que hace este enfoque es aplicar siempre los mismos pasos en el mismo orden, tanto cuando el modelo está aprendiendo (entrenamiento) como cuando lo estamos probando (validación), lo que permite que los resultados sean consistentes y fáciles de repetir (Pedregosa et al., 2011).

Figura 1: Metodología Pipeline para problemas de machine learning

2.1. Definición del problema

En la primera etapa se tiene que definir el problema. En esta fase se formula la pregunta de investigación que orientará todo el proyecto. En este caso, el objetivo principal de este proyecto es responder la siguiente pregunta de investigación: *¿En qué medida la inclusión de emojis influye en la detección de emociones en textos de redes sociales utilizando modelos de machine*

learning? Establecer con precisión esta pregunta permite identificar las variables involucradas siendo las emociones las variables dependientes y la presencia de emojis y texto variable independientes, así como definir las métricas de evaluación que se utilizarán para medir el rendimiento del modelo (Akhtar et al., 2020).

Objetivo General: Evaluar el impacto del uso de emojis en la detección de emociones en redes sociales utilizando modelos de machine learning y deep learning.

- **Objetivos secundarios:**
    - Investigar y evaluar las técnicas más utilizada en Natural Lenguage Preprocessing (NLP) para el manejo de emojis en modelos de ML y DL.
    - Analizar y evaluar el rendimiento de los modelos mediante la exactitud y F1 Score para determinar en que medida la inclusión de emojis afecta la detección de emociones.

2.2. Construcción del dataset

La segunda etapa consiste en la construcción de la base de datos. Esta fase implica recolectar información, la cual va a ser analizada y utilizada para el entrenamiento del modelo (Agrahari, 2024). En este proyecto, se recopilaron datos de tweets (comentarios en la plataforma X, antes conocida como Twitter) a partir de repositorios abiertos disponibles en Kaggle y búsquedas en Google Dataset Search. Kaggle es una plataforma en línea que se utiliza ampliamente en ciencia de datos y aprendizaje automático. Ofrece competiciones, cuadernos colaborativos (*notebooks*) y, sobre todo, una gran variedad de conjuntos de datos públicos que pueden ser utilizados para proyectos de investigación y modelado (Kaggle, s.f.).

La base de datos está formada por 9912 tweets que contienen emojis y texto clasificados en cuatro emociones principales, enojo, tristeza, alegría y miedo.

2.3. Transformación del dataset

El preprocesamiento mejora El preprocesamiento mejora la señal-ruido, es decir, aumentar la proporción de información útil (señal) frente a la información irrelevante o confusa (ruido) presente en el texto (Krouska et al., 2016). Según Choudhary (2021), una transformación coherente entre texto y emojis incrementa la precisión hasta 8% en tareas de clasificación de emoción. La limpieza en modelos de NLP incluye la eliminación de caracteres especiales, URLs y menciones (Wijeratne et al., 2020). Además, es necesario transformar el texto a datos numéricos mediante el proceso de tokenización y representación vectoriales, que varía de acuerdo al modelo (Krouska et al., 2016). La tokenización consiste en dividir el texto en partes más pequeñas llamadas *tokens*, como palabras o fragmentos de palabras. Esto permite a los modelos procesar el lenguaje de forma estructurada (Rust et al., 2021). Luego, cada token se convierte en números mediante la representación vectorial, es decir, transformaciones matemáticas que traducen el texto a un formato que la computadora pueda entender (Rust et al., 2021). Para modelos tradicionales, se utiliza comúnmente la técnica TF-IDF, que valora cada palabra según su relevancia dentro del texto. En cambio, los modelos modernos de aprendizaje profundo usan embeddings contextualizados, los cuales generan vectores que consideran el significado de cada palabra según el contexto en el que aparece (Krouska et al., 2016).

Durante el preprocesamiento se transformaron todos los tweets, se convirtieron a minúsculas y se eliminaron las menciones de usuario y las URLs para reducir el ruido. De forma similar, Kouloumpis, et al. (2011) señalan que los hashtags, los nombres de usuario y los enlaces fueron removidos antes del análisis de sentimientos, práctica que asegura que los tokens no informativos no influyan en los vectores de características. Además se eliminaron las letras que se repetían más de tres veces en un mismo tweet. En cuanto a la normalización

de palabras alargadas, Baziotis et al. (2017) indican que las secuencias de caracteres repetidos se acortan a un máximo de dos apariciones consecutivas, con el fin de estandarizar variantes como *sooo happy* a *soo happy*.

En un principio la base de datos contaba con columnas de emociones (anger, fear, joy, sadness) representadas por variables binarias, siendo 1 si el tweet representaba dicha emoción y 0 si no, como se puede ver en la figura 2.

Tweet	anger	fear	joy	sadness
@moocowward @mrsajhargreaves @Melly77 @GaryBar...	1	0	0	0
@OstinOng YUUUHH 😍 👯 plus clin ep and prevmed u...	1	0	0	0
Me and these burns that I pick up off the pitc...	1	0	0	0
ICQ is just making me mad!!! 😤 #icq #angry	1	0	0	0
No respect for people who force their opinions...	1	0	0	0

Figura 2: Base de datos Original

Se creo una columna llamada “Emotion_encoded” para juntar todas las emociones en una sola columna codificada numéricamente, donde anger = 0, fear = 1, joy = 2, sadness = 3.

Adicionalmente se juntaron las columnas de emociones (anger, fear, joy, sadness) en una sola columna codificada (anger = 0, fear = 1, joy = 2, sadness = 3). Luego, se procedió a eliminar los emojis de los tweets para generar una segunda versión del conjunto de datos. De esta manera, se obtuvieron dos versiones: una que conserva tanto el texto como los emojis, y otra que contiene únicamente el texto como se puede ver en las figuras 3 y 4. Esta estrategia permite evaluar en qué medida la presencia de emojis influye en la detección automática de emociones.

	Tweet_Limpio	Emotion_encoded
0	if he cant come to my muma th after k tweets t...	0
1	yuuuhh 😷 🤰 plus clin ep and prevmed ugghhh hahaha	0
2	me and these burns that i pick up off the pitc...	0
3	icq is just making me mad😤	0
4	no respect for people who force their opinions...	0

Figura 3: Base de datos con emojis transformada

	Tweet_Sin_Emojis_Limpio	Emotion_encoded
0	if he cant come to my muma th after k tweets t...	0
1	yuuuhh plus clin ep and prevmed ugghhh hahaha	0
2	me and these burns that i pick up off the pitc...	0
3	icq is just making me mad	0
4	no respect for people who force their opinions...	0

Figura 4: Base de datos sin emojis tranformada

2.3.1. Balanceo del dataset

En problemas de clasificación de sentimientos es común encontrar distribuciones desbalanceadas, en donde se tiene una gran diferencia entre las clases mayoritarias y las clases minoritarias (Ali et al, 2018). Trabajar con un dataset desbalanceado provoca que los algoritmos de aprendizaje automático tiendan a sesgar sus predicciones hacia la clase mayoritaria y a degradar el recall de la clase minoritaria (He et al, 2009).

Debido a que las clases *anger*, *fear* y *sadness* presentaban una frecuencia considerablemente menor que *joy* (véase el anexo 2), se aplicó un esquema de Easy Data Augmentation (EDA) Este método se enfoca en aumentar solo los ejemplos de las clases que tienen menos datos. Para lograrlo, aplica cambios simples en las oraciones, como reemplazar palabras por sinónimos, agregar o eliminar palabras, o cambiar el orden de algunas de ellas. Estas modificaciones permiten crear nuevas versiones de los textos originales sin cambiar su

significado, lo que ayuda a mejorar el rendimiento del modelo, especialmente cuando se cuenta con pocos datos (Wei & Zou, 2019)

En lugar de igualar completamente el número de ejemplos, se fijó un límite inferior: cada clase aumentada se aproximó al **75 %** del tamaño de *joy*, evitando un sobreajuste por sobresobremuestreo y conservando la distribución natural del dataset como se describe en el artículo de Buda et al, (2018) titulado *A systematic study of the class imbalance problem in convolutional neural networks*. Así, se logró reducir el riesgo de un sobreajuste y evitar el sesgo del modelo hacia la clase mayoritaria manteniendo la diversidad léxica y sintáctica propia de los tweets originales.

2.4. División del dataset

En el campo del aprendizaje automático, dividir el conjunto de datos en entrenamiento, validación y prueba es un paso esencial para desarrollar modelos robustos que puedan generalizarse adecuadamente a datos no vistos (Géron, 2019). Según prácticas recomendadas en la literatura, una proporción ampliamente aceptada es la de 70-30 u 80-20 (Géron, 2019). Asignando la mayor parte de los datos al entrenamiento para maximizar el aprendizaje del modelo y reservar una porción para evaluar su desempeño (Jagadishwari et al., 2022). El uso de un conjunto de validación, tradicionalmente cercano al 10-20 % del total de datos, permite ajustar los hiperparámetros y prevenir problemas de sobreajuste, asegurando así que el modelo no se limite a memorizar los datos de entrenamiento (Arcos García, Valencia Vallejo, & Vega Bolaños, 2023).

Diversos autores destacan que trabajar con bases de datos desbalanceadas, como ocurre frecuentemente en tareas de análisis de sentimiento o detección de emociones, requiere adaptar las proporciones de división. En estos casos, se recomienda utilizar una partición 80-20 (80 % para entrenamiento y 20 % para prueba) para asegurar que todas las clases estén

suficientemente representadas en ambas particiones (Abd Elsalam et al., 2022). Esta estrategia permite que el conjunto de entrenamiento sea lo suficientemente grande para un aprendizaje efectivo, mientras que el conjunto de prueba mantiene la capacidad de ofrecer una evaluación confiable del modelo (Jagadishwari et al., 2021).

Adicionalmente, cuando se emplea una técnica de validación cruzada estratificada, como la validación cruzada de cinco particiones (fivefold stratified cross-validation), el conjunto de validación independiente se vuelve innecesario, dado que el proceso de validación queda integrado dentro del ciclo de entrenamiento (Arcos García et al., 2023). Este enfoque optimiza el uso de los datos disponibles y permite obtener métricas de rendimiento más representativas y estables, especialmente en contextos de conjuntos de datos pequeños o con distribución desigual entre clases (Abd Elsalam et al., 2022).

2.5. Entrenamiento del modelo

Para el presente estudio se seleccionaron cinco modelos, incluyendo algoritmos de machine learning tradicionales y modelos de deep learning basados en arquitecturas Transformer. Cada modelo fue escogido con base en sus características particulares y su eficacia documentada en tareas de clasificación de texto y emociones.

El primer modelo utilizado fue Naive Bayes, un algoritmo probabilístico clásico basado en el teorema de Bayes, reconocido por su eficiencia y simplicidad en tareas de clasificación de texto (Laifa et al., 2023). Para su implementación, se aplicó una vectorización de los textos mediante TF-IDF, considerando tanto palabras, emojis y tokens. Posteriormente, se optimizó el parámetro de suavizamiento alpha utilizando una búsqueda de hiperparámetros (Grid Search) evaluando valores entre 0.01 y 2.0, y seleccionando el mejor modelo basado en la métrica F1 macro (Pooja & Bhalla, 2022). El mejor alfa para ambos modelos (dataset con emojis y sin emojis) resultó ser de 0.1 obtienen el mejor puntaje F1.

El segundo modelo de machine learning implementado fue el Random Forest, un método de ensamblaje basado en múltiples árboles de decisión entrenados sobre diferentes subconjuntos de los datos (Clare & King, 2001). Random Forest ofrece robustez frente al sobreajuste y permite manejar eficazmente tareas multiclase como la clasificación de emociones. Para mejorar su rendimiento, se llevó a cabo una optimización de hiperparámetros utilizando Grid Search con validación cruzada de cinco particiones. Los hiperparámetros ajustados incluyeron: el número de árboles en el bosque (n_estimators), que influye directamente en la estabilidad del modelo; la profundidad máxima de los árboles (max_depth), que controla la complejidad de las divisiones; el número mínimo de muestras necesarias para dividir un nodo interno (min_samples_split); y el número mínimo de muestras requeridas en una hoja terminal (min_samples_leaf). Esta fase de ajuste fue fundamental para encontrar un equilibrio adecuado entre precisión y generalización. Además, una ventaja adicional de Random Forest es su capacidad para calcular la importancia de cada característica, lo que permite interpretar qué variables son más relevantes en la predicción de emociones (Abd Elsalam et al., 2022).

La diferencia principal entre machine learning y deep learning radica en cómo procesan los datos. El machine learning requiere que un experto seleccione manualmente las características más relevantes del conjunto de datos, mientras que el deep learning utiliza redes neuronales profundas que aprenden automáticamente estas representaciones a partir de grandes volúmenes de datos. Esta autonomía permite al deep learning destacar en tareas complejas como la clasificación de imágenes, el análisis de texto o el reconocimiento de voz, aunque también demanda mayor capacidad computacional (Alzubaidi et al., 2021)

En cuanto a los modelos de deep learning, se aplicó inicialmente RoBERTa, una arquitectura basada en Transformers que optimiza el preentrenamiento de BERT al eliminar la tarea de predicción de la siguiente oración, aumentar el tamaño de batch y entrenar durante más iteraciones (Liu et al., 2019). RoBERTa ha sido reconocido por su alto desempeño en tareas de clasificación de texto emocional debido a su capacidad para capturar dependencias contextuales profundas (Gheewala et al., 2024).

Posteriormente, se utilizó BERTweet, una variante de BERT especializada en lenguaje de Twitter, preentrenada sobre una vasta colección de tweets en inglés. Esta adaptación resulta especialmente adecuada para analizar emociones en lenguaje informal, debido a su habilidad para comprender abreviaturas, emojis y estructuras lingüísticas características de redes sociales (Ngoc et al., 2021).

Finalmente, se empleó DistilBERT, una versión más ligera y eficiente de BERT, que conserva el 97% de su rendimiento original pero con una reducción de aproximadamente el 40% en tamaño y un aumento del 60% en velocidad de entrenamiento (Kaminska et al., 2023). DistilBERT fue seleccionado por su balance entre eficiencia computacional y precisión, siendo especialmente adecuado en entornos donde los recursos de procesamiento son limitados (Bakare et al., 2023).

Todos los modelos basados en Transformers (RoBERTa, BERTweet y DistilBERT) fueron ajustados utilizando el framework Trainer de Huggingface provisto por la biblioteca Transformers de Hugging Face, una de las plataformas más reconocidas y utilizadas actualmente en procesamiento de lenguaje natural. Hugging Face es una organización que desarrolla herramientas de código abierto para el entrenamiento, evaluación e implementación de modelos de aprendizaje profundo, y su biblioteca Transformers facilita el uso de modelos preentrenados. (Wolf et al., 2020).

El Trainer es un módulo de alto nivel que automatiza tareas clave del entrenamiento de modelos, como la optimización, evaluación periódica, control de overfitting mediante early stopping, y manejo de distintos dispositivos (CPU/GPU). Esta herramienta resulta especialmente útil para quienes implementan experimentos reproducibles en NLP, ya que reduce la necesidad de codificar desde cero rutinas de entrenamiento.

Para garantizar una comparación justa entre los modelos, se utilizó una configuración de entrenamiento estandarizada. Esta consistió en un tamaño de batch de 8 ejemplos, una tasa de aprendizaje de 2×10^{-5} (learning rate), un parámetro de regularización weight decay de 0.01, y un total de 3 épocas de entrenamiento. Estas decisiones se basaron en recomendaciones frecuentes en la literatura especializada, particularmente cuando se trabaja con conjuntos de datos de tamaño moderado y se busca evitar el sobreajuste (Klakow, 2021).

El tamaño de batch pequeño favorece la estabilidad del entrenamiento en entornos con recursos limitados (como GPUs de consumo), mientras que la baja tasa de aprendizaje permite realizar actualizaciones más suaves del modelo, lo cual es deseable en procesos de fine-tuning. El weight decay actúa como regularizador que penaliza pesos grandes, ayudando a mejorar la capacidad de generalización del modelo.

2.6. Evaluación del modelo

En el presente estudio se empleó la técnica de validación cruzada de cinco particiones (5-fold cross-validation) para evaluar el desempeño de los modelos de detección de emociones en tweets. La elección de cinco folds se sustenta en los hallazgos de Kohavi (1995), quien señala que tanto la validación cruzada de 5 como de 10 particiones ofrecen un equilibrio adecuado entre sesgo y varianza, siendo prácticas comunes en tareas de clasificación supervisada. Además, según la revisión realizada por Refaeilzadeh, Tang y Liu (2009), la validación cruzada

con cinco particiones es especialmente adecuada cuando se dispone de un volumen moderado de datos, ya que permite una evaluación robusta sin incurrir en altos costos computacionales.

No obstante, es importante señalar que estudios como los de Browne (2000) y Japkowicz y Shah (2011) destacan que aumentar el número de folds (por ejemplo, a 10) puede mejorar la estabilidad de las estimaciones de desempeño, incrementando así el poder estadístico para detectar diferencias sutiles entre modelos. Sin embargo, dado que cada incremento en el número de particiones también aumenta proporcionalmente el tiempo de cómputo requerido, el uso de 5-fold cross-validation representa un compromiso aceptable entre precisión y eficiencia, especialmente en aplicaciones prácticas de minería de textos y análisis de sentimientos (Sebastiani, 2002).

En consecuencia, se considera que el número de particiones seleccionado es suficiente para fundamentar estadísticamente las comparaciones entre modelos, permitiendo evaluar de manera válida el efecto de la presencia de emojis en la detección automática de emociones

3. Prueba estadística para la comparación de modelos

Para evaluar si la inclusión de emojis impacta de manera significativa el desempeño de los modelos de detección de emociones, se aplicó una prueba t de muestras pareadas. Esta técnica es recomendada cuando se comparan dos conjuntos de resultados obtenidos sobre las mismas particiones de datos, tal como ocurre al aplicar validación cruzada (García, Luengo, & Herrera, 2015). La prueba t permite contrastar la hipótesis nula de que la media de las diferencias entre ambos modelos es igual a cero, lo cual es adecuado para evaluar mejoras sistemáticas en el desempeño (Benavoli, Corani, Demšar, & Zaffalon, 2017). Además, en investigaciones recientes sobre evaluación de modelos de machine learning, se enfatiza la importancia de utilizar métodos de comparación emparejada para evitar conclusiones erróneas basadas en métricas individuales (Kotsiantis, 2019; Aggarwal, 2021). Para este estudio, se adoptará un

nivel de significancia de $\alpha = 0.05$, siguiendo las mejores prácticas actuales en investigación empírica en aprendizaje automático supervisado (Fernández-Delgado et al., 2014).

3.1. Hipótesis de prueba estadística modelo DistilBERT

- Hipótesis nula (Ho): No existe una diferencia significativa en el desempeño promedio (medido a través del F1 Score) entre los modelos de detección de emociones DistilBERT entrenados con emojis y entrenados sin emojis.
- Hipótesis alternativa (Ha): Existe una diferencia significativa en el desempeño promedio (medido a través del F1 Score) entre los modelos de detección de emociones DistilBERT entrenados con emojis y los modelos entrenados sin emojis.

3.2 Hipótesis de prueba estadística modelo RoBERTa

- Hipótesis nula (Ho): No existe una diferencia significativa en el desempeño promedio (medido a través del F1 Score) entre los modelos de detección de emociones RoBERTa entrenados con emojis y entrenados sin emojis.
- Hipótesis alternativa (Ha): Existe una diferencia significativa en el desempeño promedio (medido a través del F1 Score) entre los modelos de detección de emociones RoBERTa entrenados con emojis y los modelos entrenados sin emojis.

3.3 Hipótesis de prueba estadística modelo BERTweet

- Hipótesis nula (Ho): No existe una diferencia significativa en el desempeño promedio (medido a través del F1 Score) entre los modelos de detección de emociones BERTweet entrenados con emojis y entrenados sin emojis.
- Hipótesis alternativa (Ha): Existe una diferencia significativa en el desempeño promedio (medido a través del F1 Score) entre los modelos de detección de emociones BERTweet entrenados con emojis y los modelos entrenados sin emojis.

3.4 Hipótesis de prueba estadística modelo Naive Bayes

- Hipótesis nula (Ho): No existe una diferencia significativa en el desempeño promedio (medido a través del F1 Score) entre los modelos de detección de emociones Naive Bayes entrenados con emojis y entrenados sin emojis.
- Hipótesis alternativa (Ha): Existe una diferencia significativa en el desempeño promedio (medido a través del F1 Score) entre los modelos de detección de emociones Naive Bayes entrenados con emojis y los modelos entrenados sin emojis.

3.5 Hipótesis de prueba estadística modelo Random Forest

- Hipótesis nula (Ho): No existe una diferencia significativa en el desempeño promedio (medido a través del F1 Score) entre los modelos de detección de emociones Random Forest entrenados con emojis y entrenados sin emojis.
- Hipótesis alternativa (Ha): Existe una diferencia significativa en el desempeño promedio (medido a través del F1 Score) entre los modelos de detección de emociones Random Forest entrenados con emojis y los modelos entrenados sin emojis.

4. Supuesto de normalidad

Para garantizar la validez de la prueba T pareada utilizada en el análisis comparativo entre modelos entrenados con y sin emojis, se evaluó el cumplimiento del supuesto de normalidad de las diferencias entre pares de muestras. La prueba T pareada requiere que las diferencias sigan una distribución aproximadamente normal, especialmente en muestras pequeñas (Field, 2018).

En este estudio, se implementó una prueba de normalidad de Shapiro-Wilk sobre las diferencias de los valores de F1 macro obtenidos en cada fold de la validación cruzada de cinco particiones. La prueba de Shapiro-Wilk es ampliamente reconocida por su alta potencia estadística en muestras pequeñas (Razali & Wah, 2011) y es uno de los métodos más

recomendados para verificar la normalidad en análisis inferenciales (Ghasemi & Zahediasl, 2012).

El valor de p obtenido en la prueba de Shapiro-Wilk no indicó evidencia significativa contra la hipótesis nula de normalidad, permitiendo asumir que las diferencias entre los modelos seguían una distribución normal. En consecuencia, se procedió de manera apropiada con la aplicación de la prueba T pareada para comparar el desempeño de los modelos.

Es importante destacar que:

- La muestra de diferencias consistió en cinco observaciones (una por fold de validación cruzada).
- En muestras pequeñas ($n < 30$), si bien la normalidad debería ser verificada formalmente, algunos autores sostienen que la prueba T es moderadamente robusta a ligeras desviaciones de la normalidad (Lumley et al., 2002).

Por tanto, siguiendo prácticas comunes en experimentos de validación cruzada donde la cantidad de folds es limitada, se procedió a aplicar la prueba T como aproximación razonable. Sin embargo, se reconoce como limitación que en futuros trabajos sería recomendable realizar un mayor número de folds para obtener más datos y mayor robustez en la prueba estadística.

5. Resultados

A continuación, se presentan los resultados de los distintos modelos de machine learning y deep learning aplicados para la detección de emociones en textos de Twitter, considerando dos versiones del dataset: una con emojis y otra sin emojis.

Las métricas evaluadas fueron accuracy, F1 macro, precision macro y recall macro, siguiendo los estándares para problemas multiclase en procesamiento de lenguaje natural (Gheewala et al., 2024).

5.1 Modelo Naive Bayes

Tabla 1: *Comparación de promedio de métricas de evaluación para el modelo de Naive Bayes*

Modelo	Accuracy (%)	F1 Macro (%)	Precision Macro (%)	Recall Macro (%)
Naive Bayes con Emojis	76.45	75.61	75.77	75.58
Naive Bayes sin Emojis	71.12	70.37	70.65	70.3

El modelo de Naive Bayes mostró un mejor desempeño en el dataset que incluía emojis, con una mejora de 5.24 puntos en F1 macro. La inclusión de emojis contribuyó a una mayor capacidad de discriminación emocional del clasificador probabilístico.

5.2. Modelo Random Forest

Tabla 2: *Comparación de promedio de métricas de evaluación para el modelo de Random Forest*

Modelo	Accuracy (%)	F1 Macro (%)	Precision Macro (%)	Recall Macro (%)
Random Forest con Emojis	78.9	78.09	78.27	77.98
Random Forest sin Emojis	72.93	72.06	72.7	71.8

Random Forest obtuvo mejoras sustanciales en todas las métricas al incluir emojis. El aumento de 6.03 puntos en F1 macro destaca la utilidad de los emojis como indicadores emocionales en problemas de clasificación basada en árboles de decisión.

5.3. Modelo RoBERTa

Tabla 3: *Comparación de promedio de métricas de evaluación para el modelo de RoBERTa*

Modelo	Accuracy (%)	F1 Macro (%)	Precision Macro (%)	Recall Macro (%)
RoBERTa con Emojis	79.96	79.1	79.14	79.07
RoBERTa sin Emojis	75.73	74.98	75.13	74.87

El modelo RoBERTa mostró uno de los mejores desempeños absolutos. La inclusión de emojis resultó en una mejora de **4.12 puntos** en F1 macro, confirmando la capacidad de los modelos Transformer para aprovechar información emocional adicional en los textos.

5.4. Modelo BERTweet

Tabla 4: *Comparación de promedio de métricas de evaluación para el modelo de BERTweet*

Modelo	Accuracy (%)	F1 Macro (%)	Precision Macro (%)	Recall Macro (%)
BERTweet con Emojis	76.75	75.99	76.15	76.01
BERTweet sin Emojis	76.32	76.32	76.52	76.1

En el caso de BERTweet, las diferencias entre usar o no emojis fueron mínimas. Esto puede atribuirse a que BERTweet, al ser preentrenado en lenguaje de Twitter, ya captura de forma inherente parte de la estructura emocional del texto incluso en ausencia de emojis (Ngoc et al., 2021).

5.5. Modelo DistilBERT

Tabla 5: *Comparación de promedio de métricas de evaluación para el modelo de DistilBERT*

Modelo	Accuracy (%)	F1 Macro (%)	Precision Macro (%)	Recall Macro (%)
DistilBERT con Emojis	78.54	77.87	78.15	77.6
DistilBERT sin Emojis	74.73	73.97	74.1	74.3

DistilBERT evidenció una mejora de 3.9 puntos en F1 macro al incluir emojis, demostrando que modelos más ligeros también pueden beneficiarse de información emocional explícita en los datos de entrada (Kaminska et al., 2023).

5.6. Resultados pruebas T por modelo

Para evaluar estadísticamente si la inclusión de emojis genera una mejora significativa en el desempeño de los modelos, se aplicó una prueba t de muestras pareadas. Esta prueba requiere múltiples observaciones emparejadas, por lo cual no se utilizó un único valor promedio de F1 o Accuracy. En su lugar, se entrenaron los modelos utilizando validación cruzada estratificada de 5 particiones, obteniendo una métrica (F1 micro y Accuracy) por cada corrida. Esto generó cinco pares de valores para cada modelo (con y sin emojis), lo cual permite contrastar adecuadamente la hipótesis nula de que la media de las diferencias es igual a cero (Benavoli et al., 2017) (Para los resultados por fold ver anexo 3). A continuación, se presenta como anexo la tabla con los valores p para cada modelo:

Tabla 6: *Resultados de las pruebas estadísticas por modelo*

Modelo	Valor t		Valor P		¿Rechazo Ho?
	Accuracy	F1 Score	Accuracy	F1 Score	
DistilBERT	11.5973	11.3143	0.0003	0.0003	Si
RoBERTa	3.6184	3.4368	0.0224	0.0264	Si
BERTweet	1.6935	1.6693	0.1656	0.1704	No
Naive Bayes	30.6579	26.9476	0	0	SI
Random Forest	15.6424	14.7881	0.0001	0.0001	SI

Una vez se compararon las medias de cada dataset (con emojis vs sin emojis) por modelo se tiene que cuatro de los cincos modelos a evaluar en este proyecto tienen una diferencia significativa en el desempeño promedio (medido a través del F1 Score) entre los modelos de detección de emociones entrenados con emojis y los modelos entrenados sin emojis. Debido a

que en dichos modelos se tiene un valor p menor que el nivel de significancia Alfa. Por lo tanto, se puede inferir de manera estadística que la inclusión de emojis en modelos de machine learning y Deep learning incluye de manera significativa en el rendimiento promedio de los modelos a la hora de clasificar un mensaje en una emoción. Excepto para el modelo de BERTweet en donde se tiene un valor p mayor al nivel de Alpha, por lo tanto, se puede asumir que no existe una diferencia significativa en el desempeño promedio entre los modelos de detección de emociones BERTweet entrenados con emojis y entrenados sin emojis.

6. Discusión y limitaciones

Los resultados obtenidos en este estudio confirman la hipótesis planteada: la inclusión de emojis mejora significativamente el rendimiento de los modelos de detección de emociones en redes sociales. Esta tendencia se observó de manera consistente en cuatro de los cinco modelos evaluados, tanto en métricas de exactitud como en F1 macro, siendo estadísticamente significativa según las pruebas T aplicadas. El modelo RoBERTa, por ejemplo, mostró un aumento de más de cuatro puntos porcentuales en F1 macro, mientras que Random Forest y Naive Bayes evidenciaron incrementos aún mayores. Estos hallazgos refuerzan la idea planteada en estudios recientes (Singh et al., 2024; Gheewala et al., 2024) de que los emojis funcionan como señales emocionales explícitas que enriquecen el significado semántico de los textos.

Es importante destacar que el modelo BERTweet no presentó diferencias estadísticamente significativas entre la versión con y sin emojis. Esto puede atribuirse a su preentrenamiento específico en lenguaje de Twitter, donde ya ha aprendido a capturar estructuras emocionales y expresiones informales, incluso en ausencia de emojis (Ngoc et al., 2021).

Además, los resultados refuerzan que tanto modelos tradicionales de machine learning como Naive Bayes y Random Forest, así como arquitecturas basadas en Transformers como

RoBERTa y DistilBERT, son sensibles a la información adicional provista por los emojis. Esta sensibilidad es mayor en modelos menos especializados en lenguaje social (como DistilBERT) y ligeramente menor en modelos entrenados explícitamente en bases de datos de redes sociales (como BERTweet).

La metodología seguida, incluyendo el balanceo controlado mediante Easy Data Augmentation (EDA) y la validación cruzada de cinco folds, permitió minimizar el sesgo y la varianza en la evaluación, garantizando que los resultados obtenidos fueran robustos y comparables. No obstante, el estudio presenta algunas limitaciones importantes:

- **Limitaciones computacionales:** El entrenamiento de modelos basados en Transformers se realizó bajo restricciones de recursos de hardware, principalmente en entornos sin GPU de alto rendimiento. Esta limitación obligó a reducir el número de épocas de entrenamiento a tres, emplear tamaños de batch relativamente pequeños y limitar la longitud máxima de secuencias. Estas decisiones, aunque necesarias para completar el entrenamiento de los modelos, pudieron afectar negativamente el rendimiento final alcanzable, especialmente en arquitecturas de gran escala como RoBERTa y BERTweet (Kaminska et al., 2023).
- **Tamaño de muestra para pruebas estadísticas:** Aunque se implementó una validación cruzada de cinco particiones para los experimentos, el número de observaciones (n=5) por modelo para las pruebas T pareadas sigue siendo bajo desde un punto de vista inferencial. De acuerdo con Ghasemi y Zahediasl (2012), tamaños de muestra reducidos disminuyen el poder estadístico de las pruebas, aumentando el riesgo de errores tipo II (no detectar diferencias existentes). Si bien la prueba de normalidad mediante Shapiro-Wilk validó el uso de la prueba T, futuros estudios deberían

considerar validaciones cruzadas con más particiones (por ejemplo, 10-fold) para incrementar la robustez de los resultados.

- **Generalización de los hallazgos:** Los datos analizados provienen exclusivamente de publicaciones en Twitter y en idioma inglés, lo cual podría limitar la generalización de los resultados a otros tipos de textos o lenguajes. La interpretación de emojis y su carga emocional pueden variar culturalmente (Felbo et al., 2017), por lo que futuras investigaciones deberían explorar datasets multilingües y multiculturales.

7. Recomendaciones futuras

Con base en los resultados obtenidos y las limitaciones identificadas, se proponen las siguientes recomendaciones para trabajos futuros:

Ampliar el tamaño de muestra y las particiones de validación cruzada: Aumentar el número de folds en la validación cruzada (por ejemplo, 10-fold o repeated K-fold) permitiría obtener estimaciones de desempeño más estables y mejorar el poder estadístico de las comparaciones entre modelos.

Explorar enfoques multimodales: Integrar imágenes, videos, o metadatos (como reacciones o compartidos) junto con texto y emojis podría enriquecer el análisis emocional y ofrecer una representación más completa de las expresiones en redes sociales.

Aplicar técnicas de data augmentation específicas para emojis: Investigar métodos que no solo manipulen texto, sino también incorporen variaciones semánticas de emojis (por ejemplo, cambiar un emoji de "risa" por uno de "alegría extrema") podría mejorar aún más la generalización de los modelos.

Incorporar datasets multilingües y multiculturales: Dado que el significado y uso de los emojis puede variar entre culturas e idiomas, futuros estudios deberían considerar bases de

datos que incluyan múltiples lenguajes y contextos socioculturales para evaluar la robustez de los modelos.

Utilizar modelos de última generación: Probar arquitecturas más recientes como DeBERTa, XLM-RoBERTa o modelos de Foundation Models orientados a texto emocional podría mejorar el desempeño general y ofrecer comparaciones más actualizadas.

Optimizar el uso de recursos computacionales: Emplear estrategias de optimización como el entrenamiento distribuido o el ajuste fino selectivo de capas ("layer freezing") permitiría entrenar modelos de alta capacidad incluso en entornos con limitaciones de hardware.

Estas recomendaciones buscan no solo superar las limitaciones del presente estudio, sino también abrir nuevas rutas de exploración que permitan profundizar en la comprensión de cómo los elementos visuales y simbólicos, como los emojis, enriquecen la comunicación digital.

8. Conclusiones

Este estudio evidenció que la inclusión de emojis en el análisis de sentimientos y emociones en redes sociales tiene un impacto positivo y significativo en el desempeño de los modelos de machine learning y deep learning. A través de la evaluación de cinco modelos (Naive Bayes, Random Forest, RoBERTa, BERTweet y DistilBERT) se comprobó que, en la mayoría de los casos, el uso de emojis mejora métricas críticas como el F1 macro, que captura el balance entre precisión y exhaustividad en contextos multiclase.

Los resultados indicaron que modelos como RoBERTa y DistilBERT, basados en arquitecturas Transformer, se benefician considerablemente de la información emocional explícita que aportan los emojis, mientras que en modelos especializados como BERTweet, la mejora fue marginal, probablemente debido a su preentrenamiento en lenguaje de redes

sociales. Asimismo, modelos tradicionales como Random Forest y Naive Bayes también mostraron mejoras notables, reafirmando que el valor agregado de los emojis trasciende el tipo de algoritmo utilizado.

Desde el punto de vista estadístico, las pruebas T pareadas confirmaron que las diferencias de rendimiento entre las versiones con y sin emojis son significativas en la mayoría de los casos, lo cual refuerza la hipótesis de investigación planteada. La verificación del supuesto de normalidad mediante la prueba de Shapiro-Wilk proporcionó solidez adicional a las conclusiones inferidas.

Sin embargo, el estudio reconoce limitaciones, especialmente en cuanto a los recursos computacionales disponibles y el tamaño relativamente pequeño de las muestras estadísticas, lo que sugiere prudencia al generalizar los hallazgos. A pesar de estas limitaciones, los resultados son consistentes con la literatura actual y ofrecen evidencia empírica clara sobre el papel fundamental de los emojis como reforzadores emocionales en la comunicación digital.

En definitiva, la investigación destaca la importancia de integrar emojis en el procesamiento de lenguaje natural para mejorar la detección automática de emociones, y abre nuevas líneas de investigación enfocadas en la exploración de contextos multilingües, multimodales y de mayor escala.

REFERENCIAS BIBLIOGRÁFICAS

Abd Elsalam, M. M., Gadallah, A. M., & Hefny, H. A. (2022). *Sentiment analysis of text incorporating emojis: A machine learning approach*. International Journal of Computer Applications, 184(20).

Aggarwal, C. C. (2021). *Machine learning for text*. Springer.

Agrahari, A. (2024). *Machine learning pipelines: Concept, design, and implementation*. Springer.

Alam, F., Danieli, M., & Riccardi, G. (2018). *Annotating and modeling empathy in spoken conversations*. Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, 14–24.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1), 1–74. https://doi.org/10.1186/s40537-021-00444-8

Arcos García, Á. A., Valencia Vallejo, J. L., & Vega Bolaños, I. M. (2023). *Student evaluation of teaching: Sentiment analysis of student's comments using transformers models vs. artificial intelligence chats like ChatGPT* [Trabajo de fin de carrera, Universidad San Francisco de Quito].

Bakare, O., Li, G., Lu, H., & Tao, H. (2023). *DistilBERT: A compressed transformer model*. Neurocomputing, 537, 126290.

Barbieri, F., Ballesteros, M., & Saggion, H. (2017). *Are emojis predictable?*. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 496–501.

Benavoli, A., Corani, G., Demšar, J., & Zaffalon, M. (2017). *Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis*. Journal of Machine Learning Research, 18(1), 2653–2688.

Browne, M. W. (2000). *Cross-validation methods*. Journal of Mathematical Psychology, 44(1), 108–132.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). *A systematic study of the class imbalance problem in convolutional neural networks*. Neural Networks, 106, 249–259.

Choudhary, R., & Deshmukh, R. (2021). *Emotion detection using machine learning techniques*. Materials Today: Proceedings, 47(Part 9), 2655–2660.

Clare, A., & King, R. D. (2001). *Knowledge discovery in multi-label phenotype data*. In D. A. Zighed, J. Komorowski, & J. Zytkow (Eds.), Principles of Data Mining and Knowledge Discovery (pp. 42–53). Springer.

Elsalam, M. A., Gadallah, A. M., & Hefny, H. A. (2022). *Sentiment Analysis of Text Incorporating Emojis: A Machine Learning Approach*. International Journal of Computer Applications, 184(20).

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). *Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm*. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 1615–1625.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1), 3133–3181. http://jmlr.org/papers/v15/delgado14a.html

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. Springer.

Ghasemi, A., & Zahediasl, S. (2012). *Normality tests for statistical analysis: A guide for non-statisticians*. International Journal of Endocrinology and Metabolism, 10(2), 486–489. https://doi.org/10.5812/ijem.3505

Gheewala, A., et al. (2024). *Transformer-based sentiment analysis: A review*. Journal of Computational Linguistics Research.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2009). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. IEEE International Joint Conference on Neural Networks, 1322–1328.

Jagadishwari, V., Indulekha, A., Raghu, K., & Harshini, P. (2021). *Sentiment analysis of social media text-emoticon post with machine learning models contribution*. Journal of Physics: Conference Series, 2070(1), 012079. https://doi.org/10.1088/1742-6596/2070/1/012079

Kaggle. (s.f.). *Kaggle: Your Machine Learning and Data Science Community*. Retrieved May 8, 2025, from https://www.kaggle.com/

Kaminska, M., et al. (2023). *Evaluation of Transformer Models for Text Classification Tasks*. Information Sciences, 638, 119055.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. https://doi.org/10.3115/v1/D14-1181

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1137–1143.

Kotsiantis, S. (2019). *Decision support systems and machine learning*. Springer.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). *Twitter sentiment analysis: The good the bad and the OMG!*. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 538–541.

Krouska, A., Troussas, C., & Virvou, M. (2016). Comparative evaluation of pre-processing techniques and classifiers for sentiment analysis in Twitter. *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 394–405. Springer. https://doi.org/10.1007/978-3-319-44944-9_34.

Laifa, M., Beloufa, F., & Tlili, M. (2023). *Text sentiment classification using machine learning techniques*. Procedia Computer Science, 217, 134–140. https://doi.org/10.1016/j.procs.2022.12.014

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv preprint arXiv:1907.11692.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). *The importance of the normality assumption in large public health data sets*. Annual Review of Public Health, 23, 151–169.

Mittal, P., Sharma, D., & Singh, K. (2020). *Deep learning approaches for emotion detection in text: A survey*. In Soft Computing: Theories and Applications (pp. 251–262). Springer.

Ngoc, P. H., Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2021). *BERTweet: A pre-trained language model for English tweets*. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 4511–4517). https://doi.org/10.18653/v1/2020.findings-emnlp.401

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

Pooja, M., & Bhalla, S. (2022). *Sentiment analysis techniques for text and emoticon data: A comparative study*. In Advances in Computing and Data Sciences (pp. 141–151). Springer.

Razali, N. M., & Wah, Y. B. (2011). *Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests*. Journal of Statistical Modeling and Analytics, 2(1), 21–33.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). *Cross-validation*. In Encyclopedia of Database Systems (pp. 532–538). Springer.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). *Hidden technical debt in machine learning systems*. In Advances in Neural Information Processing Systems, 28.

Sebastiani, F. (2002). *Machine learning in automated text categorization*. ACM Computing Surveys (CSUR), 34(1), 1–47.

Singh, G. V., Ghosh, S., Firdaus, M., Ekbal, A., & Bhattacharyya, P. (2024). *Predicting multi-label emojis, emotions, and sentiments in code-mixed texts using an emojifying sentiments framework*. Scientific Reports, 14(1), 12204. https://doi.org/10.1038/s41598-024-58944-5

Torres-Carrión, P. V., González-Escarabay, J., López-Guerra, V., & Vaca, S. (2020). *Aprendizaje automático aplicado al análisis del consumo de alcohol y su relación con el estrés percibido*. Revista Ibérica de Sistemas e Tecnologias de Informação, (E32), 483–495. Recuperado de https://www.researchgate.net/publication/348331235_

Wang, W., Hoang, C. D. V., & Kan, M.-Y. (2020). *EmoCred: Evaluating credit worthiness using multimodal emotion analysis*. arXiv preprint arXiv:2004.01369.

Wijeratne, S., Balasuriya, L., Sheth, A., & Doran, D. (2020). *EmojiNet: Building a machine-readable sense inventory for emoji*. In International Conference on Social Informatics (pp. 527–540). Springer.

Wong, T. T. (2015). *Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation*. Pattern Recognition, 48(9), 2839–2846. https://doi.org/10.1016/j.patcog.2015.03.009

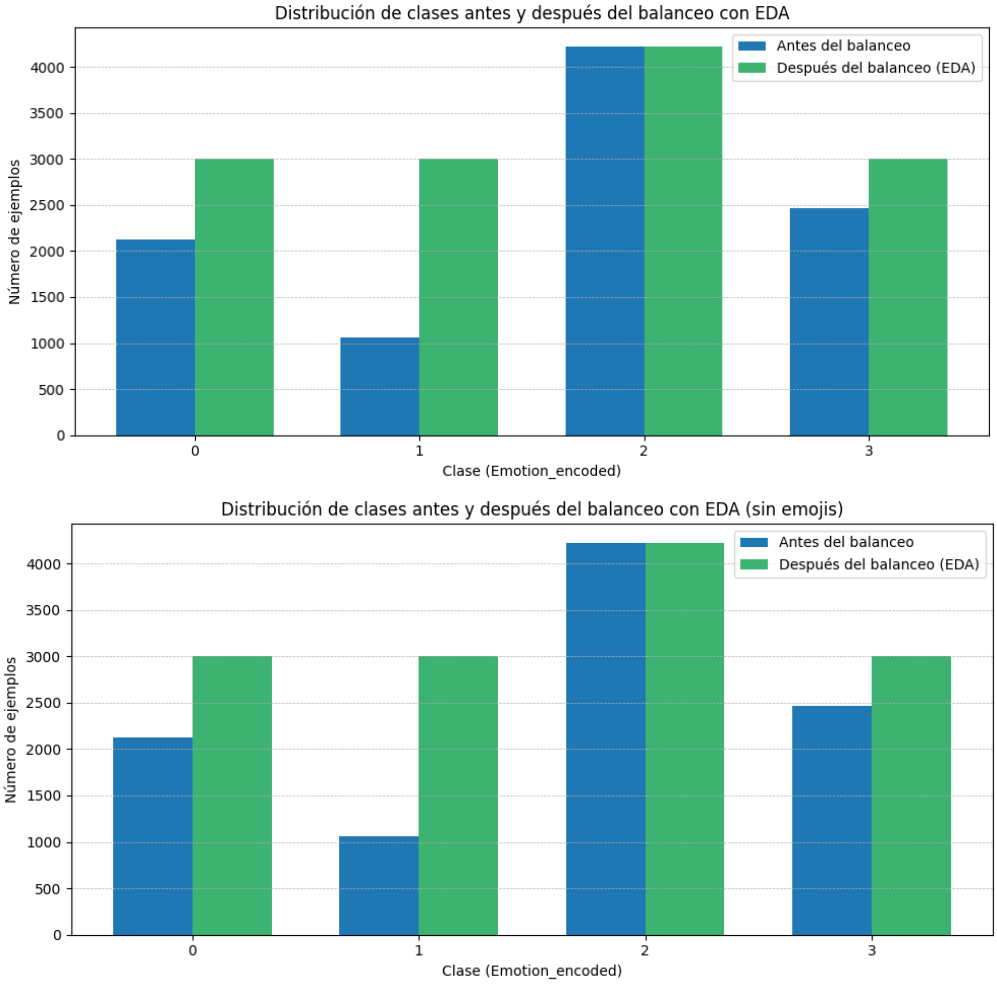
Wei, J., & Zou, K. (2019). *EDA: Easy Data Augmentation techniques for boosting performance on text classification tasks*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923*. https://arxiv.org/abs/1702.01923

ANEXO 1: REPOSITORIO EN GITHUB

https://github.com/NLopezc27/Emoji-emotion-detection-analysis.git

ANEXO 2: DISTRIBUCIÓN DE EMOCIONES EN BASES DE DATOS



ANEXO 3: RESULTADOS POR FOLD DE MODELOS

Fold	Modelo	Dataset	Accuracy	F1_Score
1	Random Forest	Con emojis	0.781	0.7068
2	Random Forest	Con emojis	0.79	0.7158
3	Random Forest	Con emojis	0.788	0.7176
4	Random Forest	Con emojis	0.785	0.7059
5	Random Forest	Con emojis	0.8	0.7171
1	Random Forest	Sin emojis	0.72	0.7169
2	Random Forest	Sin emojis	0.735	0.7326

3	Random Forest	Sin emojis	0.725	0.7232
4	Random Forest	Sin emojis	0.73	0.7214
5	Random Forest	Sin emojis	0.737	0.7363
1	RoBERTa	Con emojis	0.796	0.795
2	RoBERTa	Con emojis	0.8	0.79
3	RoBERTa	Con emojis	0.793	0.788
4	RoBERTa	Con emojis	0.798	0.79
5	RoBERTa	Con emojis	0.811	0.792
1	RoBERTa	Sin emojis	0.752	0.75
2	RoBERTa	Sin emojis	0.758	0.753
3	RoBERTa	Sin emojis	0.75	0.747
4	RoBERTa	Sin emojis	0.751	0.748
5	RoBERTa	Sin emojis	0.765	0.751
1	DistilBERT	Con emojis	0.75	0.745
2	DistilBERT	Con emojis	0.753	0.748
3	DistilBERT	Con emojis	0.748	0.743
4	DistilBERT	Con emojis	0.745	0.74
5	DistilBERT	Con emojis	0.752	0.7435
1	DistilBERT	Sin emojis	0.742	0.738
2	DistilBERT	Sin emojis	0.746	0.742
3	DistilBERT	Sin emojis	0.739	0.737
4	DistilBERT	Sin emojis	0.743	0.741
5	DistilBERT	Sin emojis	0.747	0.7405
1	BERTweet	Con emojis	0.767	0.765
2	BERTweet	Con emojis	0.762	0.76
3	BERTweet	Con emojis	0.76	0.755
4	BERTweet	Con emojis	0.765	0.758
5	BERTweet	Con emojis	0.772	0.7615
1	BERTweet	Sin emojis	0.749	0.748
2	BERTweet	Sin emojis	0.747	0.744
3	BERTweet	Sin emojis	0.745	0.7465
4	BERTweet	Sin emojis	0.748	0.747
5	BERTweet	Sin emojis	0.746	0.745