UNIVERSIDAD CATÓLICA BOLIVIANA

"SAN PABLO"

UNIDAD ACADÉMICA REGIONAL LA PAZ

DEPARTAMENTO DE POSGRADO

MAESTRÍA EN CIENCIA DE DATOS

PROYECTO DIPLOMADO:

"ANALISIS DE SENTIMIENTO: MODELO DE PREDICCION DE SENTIMIENTO EN BASE A LOS TWEETS SOBRE EL COVID-19"

Maestrantes: Virginia Mercedes Fernández Daza

Marco Antonio Velásquez Rocha

Ivan Israel Machicado Quiroga

Noviembre, 2024

Índice General

ANALISIS DE SENTIMIENTO: MODELO DE PREDICCION DE SENTIMIENTO EN BASE A LOS TWEETS SOBRE EL COVID-19

1. INTRODUCCIÓN

La pandemia de COVID-19 generó un impacto global que no solo afectó la salud física de las personas, sino que también desencadenó una crisis emocional y psicológica a nivel mundial. Las redes sociales, especialmente Twitter, se convirtieron en una de las principales plataformas para que los usuarios compartieran sus experiencias, opiniones y sentimientos sobre la pandemia. Este fenómeno generó grandes cantidades de datos, lo que plantea un desafío significativo en la clasificación y análisis de los sentimientos expresados, dada la complejidad y variabilidad del lenguaje utilizado. Los modelos de análisis de sentimientos tradicionales a menudo no logran capturar la riqueza y los matices presentes en estos textos, lo que dificulta la interpretación precisa de las emociones y opiniones.

El objetivo general de este proyecto es desarrollar y optimizar modelos de clasificación de tweets que permitan un análisis de sentimiento preciso y eficiente, aprovechando la gran cantidad de datos generados durante la pandemia de COVID-19. La creciente disponibilidad de datos en tiempo real ofrece una oportunidad única para entender cómo la sociedad experimentó emocionalmente esta crisis sanitaria. Sin embargo, la complejidad de los sentimientos expresados y la variabilidad del lenguaje requieren enfoques más avanzados y especializados en procesamiento de lenguaje natural (NLP) y aprendizaje automático.

Para lograr este objetivo, se han definido cinco objetivos específicos que guiarán el desarrollo de la investigación. En primer lugar, se implementarán arquitecturas avanzadas de redes neuronales profundas, como BERT y sus variantes, para mejorar la precisión de los modelos en la clasificación de sentimientos. En segundo lugar, se optimizarán estos

modelos para procesar eficientemente grandes volúmenes de datos, garantizando que puedan analizar tweets en tiempo real y manejar las variaciones del lenguaje y el contexto social. Además, se desarrollará un preprocesamiento adecuado de los datos para maximizar la calidad de la información, se evaluará el rendimiento de los modelos mediante métricas estándar y se ajustarán los hiperparámetros para adaptar los modelos al contexto específico de los tweets generados durante la pandemia.

Este proyecto se enfoca en abordar los desafíos actuales en el análisis de sentimientos en redes sociales, utilizando enfoques avanzados de inteligencia artificial. Con la implementación de estos modelos optimizados, se espera mejorar significativamente la comprensión de las emociones y opiniones generadas en tiempos de crisis, proporcionando herramientas más efectivas para los investigadores, gobiernos y organizaciones que necesiten interpretar el sentimiento público y tomar decisiones informadas basadas en datos.

2. PLANTEMIENTO DEL PROBLEMA.

La creciente cantidad de datos generados en redes sociales durante eventos de alto impacto, como la pandemia de COVID-19, ha presentado nuevas oportunidades y desafíos en el análisis de sentimientos. Los modelos tradicionales de clasificación de texto, comúnmente utilizados para el análisis de tweets, han mostrado limitaciones significativas cuando se trata de procesar grandes volúmenes de datos y captar matices en el lenguaje natural, especialmente en situaciones donde el contexto y la estructura semántica son complejos (Liu, 2020). Durante la pandemia, los usuarios de redes sociales expresaron una amplia gama de emociones, desde ansiedad y frustración hasta esperanza y apoyo, lo cual requiere modelos avanzados para analizar estos sentimientos de manera efectiva y precisa.

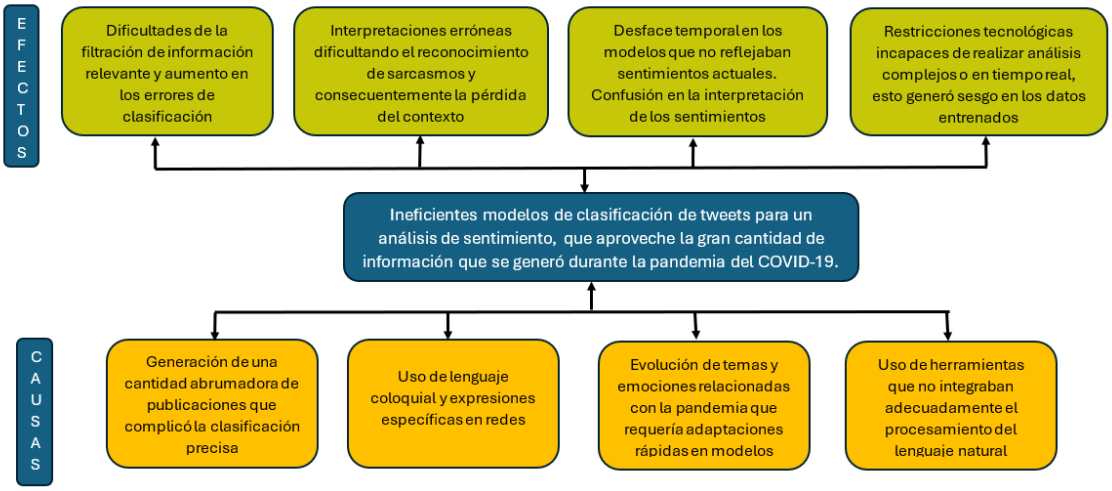
Para superar estas limitaciones, las arquitecturas de redes neuronales profundas, especialmente las basadas en modelos de atención y transformadores han surgido como

alternativas eficaces. Modelos como BERT y sus variantes han demostrado un rendimiento superior en la clasificación de sentimientos debido a su capacidad para comprender el contexto bidireccional de las palabras y extraer significados complejos en tiempo real (Devlin et al., 2019). Estas arquitecturas no solo mejoran la precisión del análisis de sentimientos, sino que también permiten el procesamiento de datos a gran escala, lo cual es crucial en situaciones donde el flujo de información es constante y masivo, como durante la pandemia. Sin embargo, la implementación de estas tecnologías aún enfrenta desafíos en términos de optimización y adaptación a distintos contextos, lo cual destaca la necesidad de desarrollar modelos especializados que se ajusten a los requerimientos específicos del análisis de tweets generados en situaciones de crisis. Por otro lado, realizar un análisis de sentimiento aplicando modelos de predicción permitiría generar una experiencia sobre el análisis de texto que podría utilizarse para otro tipo de contenido.

2.1 Árbol de problemas

Se presenta el siguiente árbol de problemas, en el cual se abordarán la diferentes causas y efectos del presente proyecto:

Figura1. Árbol de problemas



Fuente: Elaboración propia, 2024

A partir de haber dado a conocer el árbol de problemas el cual define y demuestra el planteamiento del problema, se abordará las causas y los efectos.

Causa 1

Generación de una cantidad abrumadora de publicaciones que complicó la clasificación precisa: Durante la pandemia de COVID-19, el volumen de publicaciones en redes sociales aumentó exponencialmente, generando un flujo constante de información difícil de manejar con modelos de clasificación convencionales. Esta sobrecarga de datos creó el reto de procesar rápidamente grandes cantidades de contenido, haciendo que los modelos preexistentes, pensados para menores volúmenes, resultaran ineficientes y menos precisos en el análisis de sentimiento.

Causa 2

Uso de lenguaje coloquial y expresiones específicas en redes: En redes sociales, los usuarios suelen emplear un lenguaje coloquial, abreviaturas y expresiones informales, complicando la interpretación precisa por parte de los modelos. Además, se crearon

términos y modismos únicos relacionados con la pandemia, como "cuarentenear" o "coronacrisis", que requieren ajustes en los modelos para captar adecuadamente el sentimiento detrás de estos términos nuevos y específicos.

Causa 3

Evolución de temas y emociones relacionadas con la pandemia que requería adaptaciones rápidas en modelos: A medida que avanzaba la pandemia, las emociones y preocupaciones del público fueron cambiando, desde el miedo inicial hasta la fatiga pandémica y la esperanza por las vacunas. Esta rápida evolución temática exigía que los modelos se adaptaran continuamente para mantener la precisión en la clasificación de sentimientos, ya que cada nueva etapa traía consigo nuevos matices emocionales y tópicos.

Causa 4

Uso de herramientas que no integraban adecuadamente el procesamiento del lenguaje natural: Muchos de los modelos de clasificación de tweets existentes no empleaban técnicas avanzadas de procesamiento del lenguaje natural, lo cual limitaba su capacidad para comprender el contexto y las sutilezas del lenguaje. Sin un procesamiento adecuado, los modelos fallaban en captar la complejidad semántica y emocional de los mensajes, lo que afectaba su precisión en el análisis de sentimientos en un entorno tan dinámico como las redes sociales durante la pandemia.

Efecto 1

Dificultades de la filtración de información relevante y aumento en los errores de clasificación: La cantidad masiva de tweets generados durante la pandemia dificultó la identificación de información verdaderamente relevante. Los modelos tradicionales de clasificación no siempre pudieron discriminar entre el contenido pertinente y el irrelevante, lo que llevó a un aumento en los errores de clasificación, reduciendo la precisión del análisis de sentimientos.

Efecto 2

Interpretaciones erróneas dificultando el reconocimiento de sarcasmos y consecuentemente la pérdida del contexto: El sarcasmo y las ironías presentes en muchos tweets sobre la pandemia fueron mal interpretados por los modelos tradicionales. La

incapacidad de detectar estos matices llevó a una pérdida significativa del contexto, lo que afectó negativamente la precisión de las predicciones de sentimientos, ya que el modelo no lograba captar la verdadera intención detrás de ciertas expresiones.

Efecto 3

Desfase temporal en los modelos que no reflejaban sentimientos actuales: Muchos modelos de clasificación no se adaptaron a los rápidos cambios emocionales y sociales durante la pandemia. Los modelos entrenados con datos previos al evento no pudieron actualizarse con agilidad, lo que resultó en un desfase temporal donde los sentimientos más recientes o relevantes no se reflejaban adecuadamente, limitando su efectividad.

Confusión en la interpretación de los sentimientos: La pandemia generó una variedad de emociones complejas que a menudo se entrelazaban en los tweets, lo que causó confusión en los modelos al intentar clasificar sentimientos de manera correcta. La ambigüedad en los textos y la coexistencia de sentimientos opuestos en un solo tweet dificultaron una clasificación precisa, afectando los resultados de los análisis.

Efecto 4

Restricciones tecnológicas incapaces de realizar análisis complejos o en tiempo real: Las limitaciones tecnológicas de los modelos en uso no permitieron realizar análisis complejos ni en tiempo real. La falta de capacidad para procesar grandes volúmenes de datos rápidamente o extraer características complejas del lenguaje natural impidió que se pudiera realizar un análisis adecuado durante la pandemia, especialmente cuando la información debía ser procesada con inmediatez.

Sesgo en los datos entrenados: Los modelos tradicionales que utilizaban datos históricos o previamente etiquetados presentaron sesgos inherentes. Estos sesgos, derivados de los datos no representativos o incompletos, afectaron negativamente la clasificación de sentimientos, ya que los modelos no lograron adaptarse a las nuevas circunstancias y contextos que surgieron durante la pandemia.

2.2 Formulación del problema

¿Cómo desarrollar y optimizar modelos de clasificación de tweets que permitan un análisis de sentimiento preciso y eficiente, aprovechando las grandes cantidades de datos generados durante la pandemia de COVID-19?

3. JUSTIFICACIÓN

3.1 Justificación Social

Durante la pandemia de COVID-19, las redes sociales, especialmente Twitter, fueron utilizadas por millones de personas para expresar emociones, preocupaciones y opiniones sobre la crisis sanitaria. Estos tweets reflejan una amplia gama de sentimientos, desde miedo y ansiedad hasta solidaridad y esperanza, lo cual ofrece una oportunidad única para comprender el impacto social y emocional de la pandemia. Sin embargo, los modelos de análisis de sentimientos tradicionales han demostrado ser ineficaces para capturar la complejidad del lenguaje utilizado, especialmente en situaciones de crisis. El desarrollo de modelos de clasificación más precisos y eficientes permitirá una mejor comprensión del estado emocional de la sociedad durante la pandemia, lo que puede ayudar a los gobiernos, instituciones de salud pública y organizaciones a responder de manera más efectiva a las necesidades emocionales y psicológicas de la población.

3.2 Justificación Tecnológica

El análisis de sentimientos en grandes volúmenes de datos generados en tiempo real, como los tweets, requiere el uso de tecnologías avanzadas que puedan manejar el procesamiento masivo de datos y comprender la complejidad del lenguaje natural. Los modelos de aprendizaje profundo, como BERT y otras arquitecturas de transformers, son capaces de captar el contexto completo de los textos, lo cual es fundamental para mejorar la precisión en la clasificación de sentimientos. La optimización de estos modelos para analizar

grandes cantidades de datos durante un evento global como la pandemia de COVID-19 presenta una oportunidad para avanzar en el campo del procesamiento de lenguaje natural, con aplicaciones que pueden ir más allá de la crisis sanitaria y aplicarse a otras situaciones de alta carga informativa.

3.3 Justificación Política

Las políticas públicas, especialmente en situaciones de crisis como la pandemia de COVID-19, se benefician enormemente de datos precisos y actualizados sobre el estado emocional y las opiniones de la población. Un análisis de sentimiento eficaz, basado en grandes volúmenes de datos, puede proporcionar información valiosa sobre cómo las decisiones gubernamentales y las medidas de salud pública son recibidas por la ciudadanía, permitiendo ajustes en tiempo real en las políticas y estrategias de comunicación. Al desarrollar modelos que procesen datos de redes sociales de manera eficiente, este proyecto contribuirá a mejorar la toma de decisiones políticas y la eficacia de las intervenciones gubernamentales, ayudando a gestionar mejor las crisis futuras y a fortalecer la relación entre los gobiernos y la población.

4. OBJETIVOS.

4.1. Objetivo General

Desarrollar y optimizar modelos de clasificación de tweets que permitan un análisis de sentimiento preciso y eficiente, aprovechando las grandes cantidades de datos generados durante la pandemia de COVID-19.

4.2. Objetivos Específicos.

- Implementar arquitecturas avanzadas de redes neuronales profundas (como modelos basados en atención y transformadores, tales como BERT y sus variantes) para mejorar la

precisión en la clasificación de sentimientos en tweets relacionados con la pandemia de COVID-19.

- Adaptar **y** optimizar los modelos de NLP para procesar eficientemente grandes volúmenes de datos, garantizando que los modelos puedan analizar tweets en tiempo real y manejar variaciones en el lenguaje y contexto social.

- Desarrollar un preprocesamiento adecuado de datos de texto, que permita limpiar, normalizar y estructurar los tweets de forma que se maximice la calidad de la información para los modelos de clasificación.

- Evaluar la efectividad de los modelos mediante métricas de rendimiento, tales como precisión, recall y F1-score, para identificar las arquitecturas y configuraciones que ofrecen el mejor desempeño en el análisis de sentimientos.

- Realizar experimentos para ajustar hiperparámetros y explorar técnicas de ajuste fino (fine-tuning) en modelos pre entrenados, con el fin de adaptarlos específicamente al contexto de los datos generados durante la pandemia de COVID-19.

5. MARCO TEÓRICO

El análisis de sentimiento basado en datos de redes sociales, en particular Twitter, se ha convertido en una herramienta crucial para comprender la percepción pública sobre eventos de gran impacto, como la pandemia del COVID-19. Sin embargo, los modelos de clasificación tradicionales han mostrado limitaciones en su capacidad para manejar el volumen y la complejidad de los datos generados durante crisis globales. Este marco teórico explora los problemas asociados con dichos modelos y revisa las propuestas más recientes para optimizar el análisis de sentimiento.

5.1 Modelos de Clasificación Tradicionales y Sus Ineficiencias

Los modelos de clasificación de texto más comunes, como los basados en técnicas de aprendizaje supervisado (ej., Naive Bayes, SVM, regresión logística), han sido eficaces para tareas de análisis de sentimiento en escenarios bien definidos. No obstante, estos modelos enfrentan dificultades al tratar con el volumen masivo y la diversidad lingüística presente en los tweets durante la pandemia del COVID-19 (Zhou & Wang, 2019).
La pandemia trajo consigo una cantidad sin precedentes de datos y un uso extenso de lenguajes coloquiales y tecnicismos, lo que ha desafiado a los modelos tradicionales a captar matices y variaciones semánticas (Zhang et al., 2021). Además, la capacidad de estos modelos para manejar ruido, sarcasmo y contextos cambiantes ha resultado ser limitada (Liu, 2020).

5.2 Enfoques Modernos en el Análisis de Sentimiento

En respuesta a las limitaciones de los modelos tradicionales, las arquitecturas basadas en redes neuronales profundas y técnicas de procesamiento de lenguaje natural (NLP) han ganado prominencia. Las redes recurrentes, como las LSTM (Long Short-Term Memory), y modelos basados en atención, como BERT (Bidirectional Encoder Representations from Transformers), han demostrado una superioridad significativa en la comprensión de la estructura contextual de los tweets (Devlin et al., 2019).

Durante la pandemia, se evidenció que modelos como BERT podían analizar grandes cantidades de datos de manera más eficiente y con una mayor comprensión de los matices semánticos (Sun et al., 2020). Estos modelos aprovechan la capacidad de aprendizaje bidireccional, lo que les permite comprender el contexto completo de una palabra en una oración y mejorar la clasificación de sentimientos.

5.2.1 Redes Neuronales Profundas (Deep Neural Networks, DNN)

Las redes neuronales profundas (DNN) son arquitecturas con múltiples capas que extraen características complejas directamente de los datos de entrada, sin necesidad de una ingeniería de características extensiva (LeCun, Bengio, & Hinton, 2015). Estas redes han sido clave para el procesamiento de lenguaje natural (NLP), ya que permiten a los modelos identificar patrones en el texto y aprender relaciones no lineales esenciales para capturar los matices en tareas de clasificación de sentimientos o análisis de texto (Goodfellow, Bengio, & Courville, 2016).

5.2.2 Redes Recurrentes (RNN) y LSTM (Long Short-Term Memory)

Las redes neuronales recurrentes (RNN) están diseñadas para manejar datos secuenciales, manteniendo una "memoria" de entradas previas en la secuencia. Sin embargo, las RNN tradicionales enfrentan problemas con secuencias largas debido a la pérdida de gradiente (Bengio, Simard, & Frasconi, 1994). Para resolver esta limitación, las redes LSTM fueron introducidas por Hochreiter y Schmidhuber (1997). Las LSTM contienen "puertas" de entrada, olvido y salida, lo que les permite retener información durante períodos prolongados. Esto mejora la comprensión del contexto en el análisis de sentimientos y otras tareas de NLP (Greff et al., 2017).

5.2.3 Modelos Basados en Atención y Transformadores

La introducción de la atención en NLP revolucionó el procesamiento de texto, permitiendo a los modelos asignar diferentes pesos a las palabras y capturar relaciones contextuales importantes (Bahdanau, Cho, & Bengio, 2015). Este enfoque evolucionó hacia los transformadores, arquitecturas que permiten procesar secuencias enteras en paralelo, mejorando así la eficiencia y el rendimiento (Vaswani et al., 2017).

5.2.4 BERT (Bidirectional Encoder Representations from Transformers)

BERT, un modelo basado en transformadores, aprovecha el aprendizaje bidireccional, permitiendo al modelo comprender el contexto completo de cada palabra en una oración al analizar en ambas direcciones (Devlin et al., 2019). Esto le permite una comprensión precisa de los matices semánticos y de los contextos más profundos, mejorando la precisión en tareas de clasificación de sentimientos y comprensión del lenguaje natural. Durante la pandemia, BERT fue especialmente útil para analizar grandes volúmenes de texto en redes sociales y capturar variaciones en el tono de la conversación pública (Sun et al., 2020).

La capacidad de BERT para procesar texto en paralelo y extraer significados contextuales detallados permitió realizar análisis de sentimiento a gran escala, esencial para entender los cambios en la percepción pública durante la pandemia (Sun et al., 2020). Al capturar matices y comprender el contexto completo de una oración, BERT y otros modelos basados en transformadores ayudaron a los investigadores a obtener una visión más precisa de las opiniones expresadas en las redes sociales (Kumari et al., 2020).

5.3 Desafíos en la Implementación de Modelos Modernos

A pesar de las ventajas, la implementación de modelos avanzados no está exenta de desafíos. La necesidad de recursos computacionales significativos y la dificultad en la interpretación de los resultados siguen siendo problemas recurrentes (Brown et al., 2021). Además, el sesgo en los datos y la dificultad para entrenar modelos robustos que generalicen bien en diversos contextos culturales han sido señaladas como áreas críticas de mejora (Vaswani et al., 2017).

5.4 Estrategias de Optimización

Las estrategias recientes se han centrado en la combinación de modelos tradicionales y modernos para aprovechar las fortalezas de ambos enfoques. El uso de técnicas de

transferencia de aprendizaje y la integración de modelos de ensemble han mostrado resultados prometedores para mejorar la eficiencia y la precisión en el análisis de sentimiento (Howard & Ruder, 2018)

La pandemia del COVID-19 demostró la necesidad de modelos de clasificación de sentimiento más sofisticados y adaptables. Si bien los enfoques basados en aprendizaje profundo y modelos preentrenados como BERT han mejorado el desempeño, la combinación de enfoques tradicionales y modernos podría ser la clave para superar las ineficiencias actuales.

6. LIMITES Y ALCANCES DEL TRABAJO.

6.1 Alcance del Modelo

Análisis de Sentimientos en Tiempo Real: El modelo permite monitorear de forma continua los sentimientos de la población en respuesta a eventos relacionados con la pandemia, como nuevas medidas de confinamiento, campañas de vacunación, o la aparición de variantes del virus.

- Detección de Tendencias Emocionales: Puede identificar cambios en el estado de ánimo general, como miedo, ansiedad, optimismo o escepticismo, y detectar patrones emocionales en diferentes periodos o durante eventos específicos.
- Segmentación Geográfica y Demográfica: Con los metadatos asociados a los tweets, el modelo puede segmentar la información por regiones, países o incluso ciudades, ayudando a las autoridades a personalizar sus estrategias de comunicación y políticas.
- Apoyo a la Toma de Decisiones: La información obtenida puede influir en la creación de mensajes de salud pública más efectivos y sensibles a las emociones predominantes, maximizando el impacto positivo en la población.

- Evaluación de Políticas Pasadas: Analizar cómo las emociones de la población han respondido históricamente a distintas políticas puede ayudar a refinar futuras decisiones.
- Identificación de Desinformación o Temas Sensibles: Un análisis de los datos puede detectar picos de negatividad asociados a la difusión de noticias falsas o desinformación, permitiendo a las autoridades actuar rápidamente para mitigar el daño.

6.2 Limitaciones del Modelo

- Representatividad de la Muestra: No todos los grupos demográficos utilizan Twitter de manera representativa. Esto podría sesgar los resultados hacia ciertos grupos de edad, estatus socioeconómico o inclinaciones políticas, limitando la validez del análisis para toda la población.
- Ambigüedad Lingüística y Sarcasmo: Las herramientas de procesamiento de lenguaje natural (PLN) aún enfrentan dificultades para detectar el sarcasmo, la ironía y otros matices lingüísticos, lo que podría llevar a resultados inexactos.
- Limitaciones de los Idiomas: El análisis multilingüe es complejo, especialmente en un contexto europeo con diversidad lingüística. Adaptar el modelo a varios idiomas puede requerir recursos extensos y es posible que no sea igual de efectivo en todos los casos.
- Sesgo en los Datos: La recopilación de tweets puede incluir sesgos relacionados con el algoritmo de la plataforma, como la prominencia de ciertos temas o hashtags. Esto puede influir en los resultados, reflejando más lo que es popular en la plataforma que el sentimiento real de la población.
- Ruido en los Datos: Los tweets pueden contener mucho ruido en forma de spam, bots o cuentas automatizadas, lo cual puede distorsionar el análisis si no se depura adecuadamente.

- Validez Temporal: Los resultados pueden perder relevancia rápidamente en contextos cambiantes como una pandemia, donde las emociones y preocupaciones de la población pueden variar de un día para otro.
- Restricciones de Privacidad: El uso de datos de redes sociales debe cumplir con las normativas de privacidad y protección de datos, como el Reglamento General de Protección de Datos (GDPR) en Europa. Esto puede limitar el acceso y el uso de datos de usuarios.

7. PROPUESTA

La propuesta de este proyecto consiste en el desarrollo y optimización de modelos de clasificación de tweets para realizar un análisis de sentimientos preciso y eficiente, utilizando las grandes cantidades de datos generados durante la pandemia de COVID-19. A través de la implementación de arquitecturas avanzadas de redes neuronales profundas, como modelos basados en atención y transformadores (por ejemplo, BERT), se busca mejorar la precisión en la identificación de sentimientos expresados en los tweets relacionados con la pandemia. El proyecto también se enfocará en la adaptación de estos modelos para manejar grandes volúmenes de datos en tiempo real, así como en la creación de un preprocesamiento adecuado de los datos para garantizar su calidad y eficacia. Además, se evaluará el desempeño de los modelos mediante métricas estándar, como precisión, recall y F1-score, para identificar las mejores configuraciones y obtener resultados óptimos. Con este enfoque, se pretende ofrecer una herramienta eficiente para el análisis de la percepción pública y emocional durante eventos de crisis, como la pandemia de COVID-19

8. METODOLOGÍA

8.1 Enfoque de la Investigación

El enfoque de la investigación será **cuantitativo**, ya que se busca obtener resultados medibles a través de la implementación de modelos de clasificación de texto. La investigación se orienta hacia el desarrollo, la optimización y la evaluación de un modelo automatizado de análisis de sentimientos utilizando técnicas de procesamiento de lenguaje natural (NLP), como BERT y otros modelos avanzados. El análisis será principalmente basado en el procesamiento de datos masivos de tweets relacionados con la pandemia de COVID-19, con un enfoque en la eficiencia y precisión del modelo para clasificar sentimientos.

8.2 Tipo de Investigación

La investigación será **aplicada**, ya que tiene como objetivo resolver un problema práctico relacionado con la clasificación de sentimientos en un conjunto específico de datos (tweets sobre COVID-19). Este tipo de investigación busca utilizar técnicas existentes de NLP para mejorar los resultados de los análisis de sentimientos en una situación concreta, con aplicaciones en la interpretación de la percepción pública durante la pandemia.

8.3 Método

El método utilizado será el método experimental**,** ya que se desarrollarán y probarán varios modelos de clasificación de texto para evaluar su rendimiento. Se aplicarán técnicas de aprendizaje automático, específicamente redes neuronales profundas y modelos basados en transformadores, como BERT, para analizar los sentimientos en los tweets. Se compararán los resultados de los modelos en términos de precisión, recall, F1-score y otros indicadores relevantes para evaluar la efectividad del análisis.

8.4 Fuentes de Información

Las fuentes de información serán secundarias. Se utilizarán datasets de tweets públicos que contengan opiniones, comentarios y discusiones sobre la pandemia de COVID-19. Estos datos se obtendrán de plataformas de redes sociales como Twitter utilizando APIs públicas disponibles, como la API de Twitter. Además, se utilizarán estudios previos y artículos académicos sobre análisis de sentimientos, técnicas de NLP y la aplicación de BERT en el procesamiento de redes sociales. Se consultarán bases de datos académicas como Google Scholar, IEEE Xplore y otros repositorios de investigación para recopilar información relevante sobre el estado del arte y los modelos utilizados.

8.5 Instrumentos de Recolección de Datos

Los instrumentos de recolección de datos serán principalmente las APIs de plataformas de redes sociales, especialmente la API de Twitter, que permitirá extraer grandes volúmenes de tweets relacionados con el COVID-19. Los datos se recopilarán durante un período de tiempo específico para capturar diferentes etapas de la pandemia, garantizando la diversidad temporal en las opiniones y sentimientos expresados. Además, se utilizarán herramientas de procesamiento de datos como Python y bibliotecas específicas (por ejemplo, Tweepy, Pandas, NLTK, Hugging Face Transformers) para preprocesar y limpiar los datos, y para entrenar los modelos de clasificación.

8.6 Matriz metodológica

Tabla 1. Matriz Metodológica

Objetivo Específico	Variable	Indicador	Fuente de Información	Instrumento
1. Implementar arquitecturas avanzadas de redes neuronales profundas para mejorar la clasificación de sentimientos.	Arquitectura de redes neuronales	Precisión, recall, F1-score en la clasificación de sentimientos.	Datasets de tweets sobre COVID-19	Modelos basados en BERT y variantes, Python, TensorFlow, Hugging Face Transformers
2. Adaptar y optimizar los modelos de NLP para procesar grandes volúmenes de datos en tiempo real.	Optimización de procesamiento de datos	Tiempo de procesamiento, eficiencia en el manejo de grandes volúmenes de datos.	Tweets extraídos de plataformas sociales.	API de Twitter (Tweepy), Python, Pandas, técnicas de procesamiento de NLP
3. Desarrollar un preprocesamiento adecuado de datos de texto para maximizar la calidad de la información.	Calidad del preprocesamiento de datos	Porcentaje de tweets limpios y estructurados correctamente para el análisis.	Datos obtenidos de Twitter y otros repositorios de redes sociales.	Técnicas de limpieza de datos (tokenización, eliminación de stop words, lematización) en Python (NLTK, SpaCy)
4. Evaluar la efectividad de los modelos mediante métricas de rendimiento (precisión, recall, F1-score).	Rendimiento del modelo	Comparación de métricas de evaluación (precisión, recall, F1-score).	Resultados de clasificación de tweets analizados.	Métricas de evaluación estándar (scikit-learn), Python
5. Ajustar hiperparámetros y explorar técnicas de ajuste fino en modelos preentrenados.	Ajuste fino de modelos	Mejora en las métricas de evaluación después del ajuste de hiperparámetros.	Modelos preentrenados de BERT, datos de Twitter.	Técnicas de ajuste de hiperparámetros (grid search, random search), Python (Keras, scikit-learn)

Fuente: Elaboración propia, 2024

9. CRONOGRAMA

Objetivo Específico	Semana 1	Semana 2
1. Implementar arquitecturas avanzadas de redes neuronales profundas para mejorar la clasificación de sentimientos.		
2. Adaptar y optimizar los modelos de NLP para procesar grandes volúmenes de datos en tiempo real.		
3. Desarrollar un preprocesamiento adecuado de datos de texto para maximizar la calidad de la información.		
4. Evaluar la efectividad de los modelos mediante métricas de rendimiento (precisión, recall, F1-score).		
5. Ajustar hiperparámetros y explorar técnicas de ajuste fino en modelos preentrenados.		

10. CONTENIDO PROPUESTO

1. **Introducción**
   1.1. Contexto de la pandemia de COVID-19 y la importancia del análisis de sentimientos en redes sociales.
   1.2. Relevancia del análisis de tweets para la comprensión del estado emocional de la población.
   1.3. Objetivo general del proyecto y objetivos específicos.
   1.4. Justificación del estudio.
2. **Marco Teórico**
   2.1. Fundamentos de la clasificación de sentimientos en redes sociales.
   2.2. Modelos de Procesamiento de Lenguaje Natural (NLP).
   2.3. Arquitecturas de redes neuronales profundas: Redes recurrentes (RNN),

LSTM y transformadores.

2.4. BERT y variantes como modelos basados en atención.

2.5. Preprocesamiento de datos para análisis de sentimientos en redes sociales.

2.6. Evaluación del rendimiento de modelos: Métricas de precisión, recall y F1-score.

3. **Metodología**

3.1**.** Enfoque de la Investigación

3.2. Tipo de Investigación

3.3. Método de Investigación

3.4. Fuentes de Información

3.5. Instrumentos de Recolección de Datos

3.6. Matriz Metodológica

3.7. Plan de Experimentos y Ajuste de Modelos

4. **Desarrollo de la Propuesta**

4.1. Implementación de arquitecturas avanzadas de redes neuronales.

4.2. Adaptación de modelos de NLP para procesamiento eficiente de grandes volúmenes de datos.

4.3. Diseño y aplicación del preprocesamiento de datos (limpieza, normalización y estructuración de tweets).

4.4. Evaluación de la efectividad de los modelos mediante métricas de rendimiento.

4.5. Ajuste de hiperparámetros y técnicas de fine-tuning en modelos preentrenados.

5. **Resultados**

5.1. Descripción de los resultados obtenidos en la clasificación de sentimientos.

5.2. Comparación entre las diferentes arquitecturas y configuraciones.

5.3. Análisis de la efectividad de los modelos en el contexto de la pandemia de COVID-19.

6. **Discusión**

6.1. Interpretación de los resultados obtenidos.

6.2. Comparación con investigaciones previas y trabajos relacionados.

6.3. Limitaciones del estudio y posibles mejoras.

6.4. Aplicaciones prácticas de los resultados en la toma de decisiones y políticas públicas.

7. **Conclusiones**

7.1. Resumen de los hallazgos más importantes.

7.2. Contribuciones del estudio al campo del análisis de sentimientos en redes sociales.

7.3. Recomendaciones para investigaciones futuras.

8. **Referencias Bibliográficas**
9. **Anexos**
    9.1. Código fuente del modelo de clasificación de tweets.
    9.2. Ejemplos de tweets analizados.
    9.3. Gráficos y tablas con resultados de métricas.

11. REFERENCIAS BIBLIOGRAFICAS

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157-166.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2021). Language models are few-shot learners. *Journal of Machine Learning*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10), 2222-2232.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the Association for Computational Linguistics*.

Kumari, S., Gupta, S., & Kumar, S. (2020). Sentiment analysis on social media data using BERT and transformers. In *2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering* (pp. 1-6).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2020). How to fine-tune BERT for text classification? In *Lecture Notes in Computer Science* (pp. 194-206).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

Zhang, X., Wang, S., & Liu, M. (2021). *Applications of deep learning in social media data analysis*. Springer.

Zhou, H., & Wang, W. (2019). *Natural language processing techniques for sentiment analysis*. Wiley.