PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR
MAESTRÍA EN SISTEMAS DE INFORMACIÓN
MENCIÓN DATA SCIENCE

IDENTIFICACIÓN AUTOMÁTICA DE *TWEETS* DE EMERGENCIA EN LA RED SOCIAL "X": CASO DE ESTUDIO EN ECUADOR

Autor:
Jandry Hernaldo Franco Cantos

Director:
Eduardo José Montero Bermúdez

2024

CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

Resumen

Las redes sociales, en particular la plataforma "X", representan mecanismos potencialmente valiosos para el reporte e identificación oportuna de situaciones de emergencia, gracias a la vasta cantidad de información generada continuamente por los usuarios. No obstante, la disponibilidad de corpus específicos relacionados con situaciones de emergencia es limitada, así como la automatización para la identificación de contenido textual pertinente. En este contexto, el presente estudio tiene como objetivo desarrollar un modelo de inteligencia artificial para la identificación automática de textos que abordan situaciones de emergencia, utilizando como base de aprendizaje el contenido generado por usuarios en Ecuador. El análisis y experimentación contempla una comprensión de la incidencia de las características lingüísticas específicas del país al momento de reportar situaciones de emergencia. Los resultados obtenidos de los experimentos muestran un desempeño satisfactorio en la identificación de textos sobre emergencias mediante el clasificador SVM y el clasificador LR, a su vez, los datos indican que las particularidades lingüísticas del español ecuatoriano tienen una incidencia poco significativa para identificar temas de emergencia, sugiriendo que el modelo desarrollado puede generalizarse eficazmente dentro del contexto ecuatoriano. En conclusión, el estudio confirma que es posible identificar textos relacionados con emergencias utilizando técnicas de procesamiento de lenguaje natural en el contexto específico de Ecuador, y que las características lingüísticas particulares del español ecuatoriano no representan una barrera significativa para la eficacia del modelo. Esta contribución puede ser relevante para mejorar los sistemas de alerta y respuesta ante emergencias, utilizando las redes sociales como una herramienta complementaria en la gestión de crisis.

Palabras clave: Aprendizaje automático, Ecuador, red social, situaciones de emergencia.

1. INTRODUCCIÓN

Las redes sociales son una fuente de información valiosa sobre diversos temas, por su accesibilidad, inmediatez y diversidad. En emergencia, las redes sociales permiten que quienes están en el lugar de los hechos compartan información rápidamente y las autoridades pueden usar esta información para responder de manera eficaz.

La implementación de tecnologías de inteligencia artificial (IA) para la identificación de *tweets* de emergencia se ha convertido en un área crucial, especialmente en el ámbito de la gestión de situaciones críticas. A nivel mundial, diversos estudios han explorado la aplicación de modelos de procesamiento del lenguaje natural (PLN) y aprendizaje automático en redes sociales (Martinez et al., 2018).

En los últimos años, se han desarrollado varios modelos de inteligencia artificial para la identificación automática del tipo de contenido presente en *tweets*. Estos modelos se basan en una variedad de técnicas, incluyendo el procesamiento del lenguaje natural (PLN), el aprendizaje automático y el análisis de redes sociales (Gautam et al., 2019).

Estos modelos han demostrado ser eficaces en la identificación del contexto de los *tweets*; sin embargo, se han desarrollado principalmente en contextos occidentales, y su eficacia en contextos con diversidad lingüística es aún una cuestión abierta. En Ecuador, la diversidad lingüística presenta un desafío adicional para la implementación de estos modelos, ya que las expresiones locales y las variaciones en el uso del lenguaje pueden afectar la precisión de la identificación automática de *tweets* relacionados con emergencias.

Esta investigación propone el desarrollo de un modelo de inteligencia artificial adaptado específicamente al contexto ecuatoriano para la identificación de *tweets* sobre situaciones de emergencia. Se busca explorar cómo las características lingüísticas propias de Ecuador influyen en la identificación automática de estos *tweets* y desarrollar un modelo que sea eficaz en este contexto específico. Este proyecto contribuirá al avance del conocimiento en la inteligencia artificial aplicada a datos sociales y la gestión de emergencias.

2. PLANTEAMIENTO DEL PROBLEMA

En Ecuador, al igual que en otros países, las situaciones de emergencia son una realidad recurrente que afecta la seguridad, bienestar y la infraestructura. La abundancia de información en las redes sociales, específicamente en la red social "X", se revela como una fuente potencialmente valiosa para la identificación de eventos críticos.

La amplia cantidad de información en las redes sociales, sin un filtro automatizado y efectivo, complica la tarea de distinguir rápidamente entre *tweets* ordinarios y aquellos relacionados con eventos de emergencia. La diversidad lingüística de Ecuador constituye uno de los principales desafíos, debido a que las expresiones locales y variaciones en el uso del lenguaje pueden representar obstáculos para la creación de un enfoque universal, por tanto, en el contexto de situaciones de emergencia, resulta indispensable establecer si existe un peso significativo de dichas expresiones.

A partir de ello, el problema central se manifiesta en la necesidad de contar con un modelo de inteligencia artificial capaz de identificar los *tweets* acerca de situaciones de emergencia dentro del contexto ecuatoriano, donde, la pregunta central de investigación se formula de la siguiente manera: ¿Cómo influyen las características lingüísticas propias de Ecuador en la identificación automática de *tweets* relacionados con situaciones de emergencia en la red social "X", y de qué manera puede desarrollarse un modelo de inteligencia artificial que sea eficaz en este contexto específico?

3. JUSTIFICACIÓN

La propuesta de desarrollar un modelo de inteligencia artificial destinado a identificar automáticamente *tweets* sobre situaciones de emergencia en la red social "X" en Ecuador es muy relevante. La presente propuesta busca identificar si existe una incidencia significativa de las características lingüísticas de Ecuador al momento de reportar situaciones de emergencia a través de la red social “X”, a partir de la creación y análisis de una base de datos que contempla el reporte histórico de eventos críticos en el Ecuador en la mencionada red social. De igual manera, se busca desarrollar un modelo de inteligencia artificial capaz de identificar situaciones de emergencia a partir de información proveniente del contexto ecuatoriano. Esta convergencia de enfoques aborda la singularidad de la información generada por la comunidad local en situaciones de emergencia, lo que permitirá que el modelo reconozca patrones y expresiones lingüísticas propias del contexto ecuatoriano.

En términos teóricos, el proyecto contribuye al avance del conocimiento en el campo de la gestión de emergencias y la inteligencia artificial aplicada a datos sociales. Donde, partiendo del análisis de las particularidades lingüísticas, se busca conocer su nivel de incidencia dentro de la identificación de situaciones de emergencia, cuyos resultados permitirán identificar limitaciones y mejorar la capacidad de discernimiento en la identificación de dichos *tweets*.

Los resultados obtenidos con métodos avanzados de procesamiento del lenguaje natural y aprendizaje automático contribuyen un marco metodológico que puede adaptarse y replicarse en contextos similares. La metodología propuesta sienta las bases para futuras investigaciones y desarrollos en el campo de la inteligencia artificial y la gestión de emergencias a nivel global.

4. OBJETIVOS

4.1. OBJETIVO GENERAL

- Desarrollar un modelo de inteligencia artificial capaz de identificar automáticamente *tweets* en la red social "X" referentes a situaciones de emergencia a partir de una base de datos que contemple las características lingüísticas del territorio ecuatoriano.

4.2. OBJETIVOS ESPECÍFICOS

- Constituir un conjunto de datos diversos y representativos de *tweets* relacionados con situaciones de emergencia en Ecuador, proporcionando una base sólida para el entrenamiento del modelo de inteligencia artificial.
- Comprender las particularidades lingüísticas de Ecuador presentes en la comunicación dentro de la red social “X”, al momento de reportar situaciones de emergencia para identificar patrones relevantes que permitan una adaptación efectiva del modelo de inteligencia artificial.
- Generar un modelo eficaz y preciso mediante pruebas y validaciones exhaustivas con conjuntos de datos independientes, garantizando su capacidad de discernir entre *tweets* ordinarios y aquellos vinculados con situaciones de emergencia reportados en la red social "X" dentro del territorio ecuatoriano.

5. MARCO TEÓRICO

5.1. MODELOS DE INTELIGENCIA ARTIFICIAL Y PROCESAMIENTO DEL LENGUAJE NATURAL (PLN)

La inteligencia artificial (IA) y el procesamiento del lenguaje natural (PLN) han transformado significativamente la forma en que se analiza y comprende el lenguaje humano en el ámbito digital. El PLN, es una rama de la IA, que se centra en la interacción entre las computadoras y el lenguaje humano, permitiendo que las máquinas procesen y analicen grandes cantidades de datos textuales de manera eficiente (Clark et al., 2018).

Los antecedentes históricos del uso de modelos de IA para la comprensión de textos se remontan a la década de 1960, cuando se desarrollaron los primeros modelos de PLN, estos modelos se basaban en técnicas sencillas, como el uso de reglas gramaticales o diccionarios, y su capacidad para comprender textos era limitada. Los modelos de IA y PLN han evolucionado considerablemente en los últimos años, estos modelos se han vuelto fundamentales en la automatización de tareas que requieren la comprensión y generación de lenguaje (Minghao Hu et al., 2018).

Uno de los avances más significativos en este campo es la creación de modelos de lenguaje basados en redes neuronales profundas. Modelos como BERT (Bidirectional Encoder Representations from Transformers) y GPT (Generative Pre-trained Transformer) han revolucionado el PLN al permitir una comprensión más profunda y matizada del contexto y los significados subyacentes en los textos (Devlin et al., 2019).

En el contexto de la detección de *tweets* de emergencia, estos modelos de IA y PLN ofrecen una serie de ventajas, debido a que cuentan con capacidades para analizar grandes volúmenes de datos en tiempo real, identificando patrones y tendencias que podrían indicar la ocurrencia de una emergencia. De igual manera, pueden comprender y procesar el lenguaje natural de los usuarios de redes sociales, lo que permite una identificación más precisa y contextual de los *tweets* relevantes (Kumar et al., 2022).

La implementación de estos modelos requiere una combinación de datos etiquetados, técnicas avanzadas de aprendizaje automático y recursos computacionales significativos, el proceso de entrenamiento de los modelos implica la recopilación y etiquetado de grandes conjuntos de datos de *tweets* relacionados con emergencias, seguido por el ajuste y optimización de los modelos para mejorar su precisión y eficiencia (Abubakar et al., 2019).

5.2. ESTUDIOS PREVIOS SOBRE IDENTIFICACIÓN DE *TWEETS* DE EMERGENCIA

La historia de la aplicación de inteligencia artificial (IA) en la identificación de eventos de emergencia a través de redes sociales tiene sus raíces en la última década. Inicialmente, la detección de eventos críticos se centró en la monitorización de palabras clave, pero con la evolución de las técnicas de procesamiento del lenguaje natural (PLN) y el aprendizaje automático, se ha avanzado hacia modelos más sofisticados capaces de comprender contextos (Martinez et al., 2018). Estos avances tecnológicos han mejorado significativamente la precisión y eficacia de los modelos de detección de emergencias, superando las limitaciones de los enfoques basados en palabras clave.

Las tecnologías emergentes ofrecen nuevas oportunidades para mejorar la precisión y la eficacia de estos modelos. En particular, el desarrollo de modelos de aprendizaje automático de última generación, como los modelos de aprendizaje profundo (*deep learning*), están mostrando un gran potencial para la identificación automática de *tweets* de emergencia (Kumar et al., 2022). Estos modelos pueden analizar grandes volúmenes de datos en tiempo real y detectar patrones complejos, lo que es crucial para una respuesta rápida y eficiente en situaciones de emergencia.

La aplicación de IA en la identificación de *tweets* de emergencia ha sido explorada principalmente en contextos como desastres naturales, crisis sanitarias y eventos sociales significativos. Las situaciones de emergencia son eventos que pueden causar daños a la población y la infraestructura (Madichetty, 2021). En Ecuador, como en muchos otros países, estas situaciones son una realidad recurrente que pueden afectar la seguridad, bienestar y el desarrollo.

Las redes sociales se han convertido en una fuente de información valiosa para la identificación y el análisis de situaciones de emergencia. Estas plataformas permiten a las personas compartir información de manera rápida y sencilla, lo que puede ser crucial para las autoridades responsables de la gestión de crisis (Shrivastava et al., 2021). Además, la geolocalización de los *tweets* y el análisis de sentimientos pueden proporcionar información adicional sobre la ubicación y la gravedad de las emergencias, mejorando aún más la capacidad de respuesta. La investigación y desarrollo en esta área sigue evolucionando, con un enfoque hacia la creación de modelos que comprendan más profundamente el contexto y las necesidades inmediatas de la población afectada.

5.3. PARTICULARIDADES LINGÜÍSTICAS DE ECUADOR

El análisis y la detección de *tweets* de emergencia en Ecuador requieren una comprensión profunda de las particularidades lingüísticas y culturales del país. Ecuador es un país pluricultural y multilingüe, donde coexisten diversas lenguas y dialectos, reflejando la rica diversidad étnica y cultural de su población. Este contexto lingüístico presenta tanto desafíos como oportunidades para la aplicación de modelos de inteligencia artificial (IA) y procesamiento del lenguaje natural (PLN) en la identificación de emergencias a través de redes sociales (Shrivastava et al., 2021).

El español es el idioma oficial y predominante en Ecuador, pero existen variaciones regionales y dialectales significativas. Ecuador es hogar de varias lenguas indígenas, como el kichwa, shuar, y otras lenguas amazónicas y andinas (Izurieta-Brito et al., 2020). Otro aspecto importante es el uso de jerga y modismos locales, que varían mucho entre grupos de edad y regiones, los ecuatorianos pueden utilizar metáforas, referencias culturales y humor en sus comunicaciones (Rojas, 2020).

Tales aspectos pueden complicar la tarea de los modelos de IA para identificar correctamente un *tweet* como relevante en una situación de emergencia. Por ejemplo, expresiones humorísticas o sarcásticas pueden ser malinterpretadas si los modelos no están adecuadamente entrenados para reconocer el contexto cultural específico.

Por ello, una amplia base de datos etiquetada centrada en Ecuador es esencial para que los modelos de IA y PLN puedan diferenciar correctamente el contexto cultural en cuanto a metáforas, referencias y humor se refiere; para establecer la incidencia de las particularidades lingüísticas y culturales de Ecuador. El etiquetado de datos según si se refiere a una emergencia o no, permite a los modelos de PLN diferenciar contextos y determinar la relevancia de un *tweet* (Costa et al., 2022).

A partir de una base de datos etiquetada, se puede establecer la incidencia de las particularidades lingüísticas y culturales de Ecuador en la comunicación de emergencias, a través del análisis de terminología centrada en afirmaciones, negaciones, acciones y vulgarismos porque estos tipos de palabras y expresiones son cruciales al momento de reportar emergencias, ayudando a mejorar la precisión y eficacia de la detección automática de *tweets* de emergencia.

6. METODOLOGÍA

6.1. RECOPILACIÓN Y CONSTRUCCIÓN DEL CONJUNTO DE DATOS

El desarrollo de un modelo de inteligencia artificial para la identificación automática de *tweets* que reporten emergencias en Ecuador requiere una fase inicial de recopilación de datos robusta. Esta fase es fundamental para garantizar que el modelo final esté bien adaptado al contexto específico que se desea analizar.

La fuente principal de información es la red social "X", en la cual, a través de una parametrización precisa de los metadatos permite enfocar la búsqueda y recopilación de *tweets* representativos de las diversas formas en que los usuarios ecuatorianos reportan emergencias en la plataforma.

La parametrización consiste en establecer un rango geoIlustración que abarca las distintas regiones del Ecuador, términos claves de búsqueda, y la definición de períodos de tiempo variados para la recopilación de datos. Esto asegura que los datos recolectados reflejan los patrones de uso de la red social en diferentes temporalidades.

Recopilación preliminar de tweets

Un aspecto fundamental es la selección de los términos de búsqueda. Para determinar términos específicos para los informes de situaciones de emergencia, se crea una base de datos preliminar a partir de términos identificados en los reportes de la página oficial del ECU 911. A partir de esta información, se establecen tres categorías de emergencia con sus correspondientes términos de búsqueda preliminar:

Tabla 1. Distribución de términos según categoría

Categoría	**Términos Claves**
Seguridad Ciudadana	Asesinato, Robo, Secuestro, Agresión
Gestión de Siniestros	Atropello, Choque, Accidente
Gestión de Riesgos	Incendio, Inundación, Derrumbe, Erupción

Cada término incluye sus conjugaciones con la finalidad de considerar diversos contextos en las situaciones de emergencia. Además, a parte de las búsquedas generales, también se establecen búsquedas en cuentas de medios de comunicación y ciudadanos comunes cuyas publicaciones estén enfocadas en el reporte de emergencias, para asegurar una visión amplia y diversa de las emergencias reportadas.

La extracción de datos se realizó mediante el uso de la plataforma "APIFY", la cual ofrece una variedad de herramientas para realizar *web scraping* y permite obtener una gran cantidad de metadatos en los *tweets*.

Se programaron scripts automatizados que recopilan *tweets*, siguiendo los criterios establecidos previamente. Estos *tweets* fueron almacenados en una base de datos estructurada, que permite un fácil acceso y manipulación para etapas posteriores de preprocesamiento. Cada *tweet* extraído incluye metadatos relevantes como la ubicación geográfica y el *timestamp*.

Etiquetado de *Tweets* Preliminares

Para identificar si el contenido del *tweet* se considera reporte de una emergencia, se realizó un etiquetado asistido de cada *tweet* recopilado. El etiquetado fue realizado de forma individual por tres personas: un perfil relacionado a la Gestión de Riesgos, un usuario promedio de "X" y un etiquetador propio del equipo.

Los criterios básicos para catalogar las emergencias fueron:

- **Emergencias consideras para el modelo:** Situaciones que requieran atención y puedan ser gestionadas por personal policial, agentes de tránsito, o el cuerpo de bomberos.
- **Emergencias excluidas para el modelo:** Asuntos relacionados con decomiso de armas, drogas, asuntos políticos.
- **Relevancia:** El contenido debe mostrar un nivel aceptable de contexto o información relevante acerca de las emergencias.

Las emergencias excluidas son asignadas con la etiqueta de no emergencia, y el etiquetado también detalla a que categoría de emergencia corresponde el *tweet*.

Determinación de nuevos términos de búsqueda.

Para la creación del conjunto de datos preliminar se consideró la búsqueda de *tweets*, estableciendo el periodo de búsqueda durante del primer trimestre del 2024. La cantidad resultante fue de 1.041 *tweets*, los cuales una vez etiquetados se dividieron en:

- 744 *tweets* de no emergencia
- 297 *tweets* de emergencia

Dando como resultado las siguientes agrupaciones:

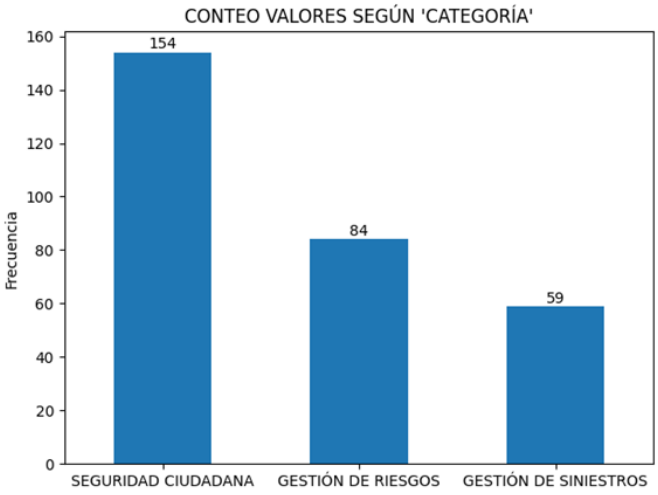

Ilustración 1. Distribución del conjunto de datos preliminar según su categoría

A partir de la revisión de frecuencias de palabras, se establecieron las siguientes palabras de búsqueda en bigramas, sumadas a las unigramas iniciales:

Tabla 2. Agrupación de términos iniciales de búsqueda según categoría

Categoría	**Términos Claves**
Seguridad Ciudadana	[cara delincuentes, delincuentes hirió, hirió asalto, personas heridas, sujetos armados, pillos robaron, balacera momento]
Gestión de Siniestros	[siniestro tránsito, atención siniestro, accidente tránsito, coordinó atención, recuerde maneje, maneje precaución]
Gestión de Riesgos	[fuerte lluvia, inundando varios, incendio forestal, aguacero tormenta, derrumbe tierra, lluvia inundando, desbordado inundando]

Recopilación total de *tweets*

Una vez identificadas las palabras clave, se realizaron las nuevas búsquedas desde el año 2020 hasta el 2023 para abarcar varios contextos históricos, con la finalidad de construir un conjunto de datos robusto y representativo, esencial para el entrenamiento y evaluación del modelo de inteligencia artificial.

La recopilación total de *tweets* y etiquetado se llevó a cabo siguiendo la metodología empleada en el desarrollo de la información preliminar. Se recopiló una mayor cantidad de datos para asegurar que el modelo de inteligencia artificial pudiera generalizar mejor y ser más efectivo en identificar emergencias en diversas circunstancias.

El uso de la plataforma "APIFY" facilitó la recolección automatizada y el almacenamiento estructurado de los metadatos de cada *tweet*, lo que permitió manejar grandes volúmenes de datos de manera eficiente. Sin embargo, las cantidades obtenidas están sujetas a las recopiladas por dicha plataforma acorde a los parámetros establecidos.

De esa búsqueda se obtuvieron 4604 *tweets*. Una vez etiquetados, se dividieron en:

- 2792 *tweets* de no emergencia
- 1,812 *tweets* de emergencia

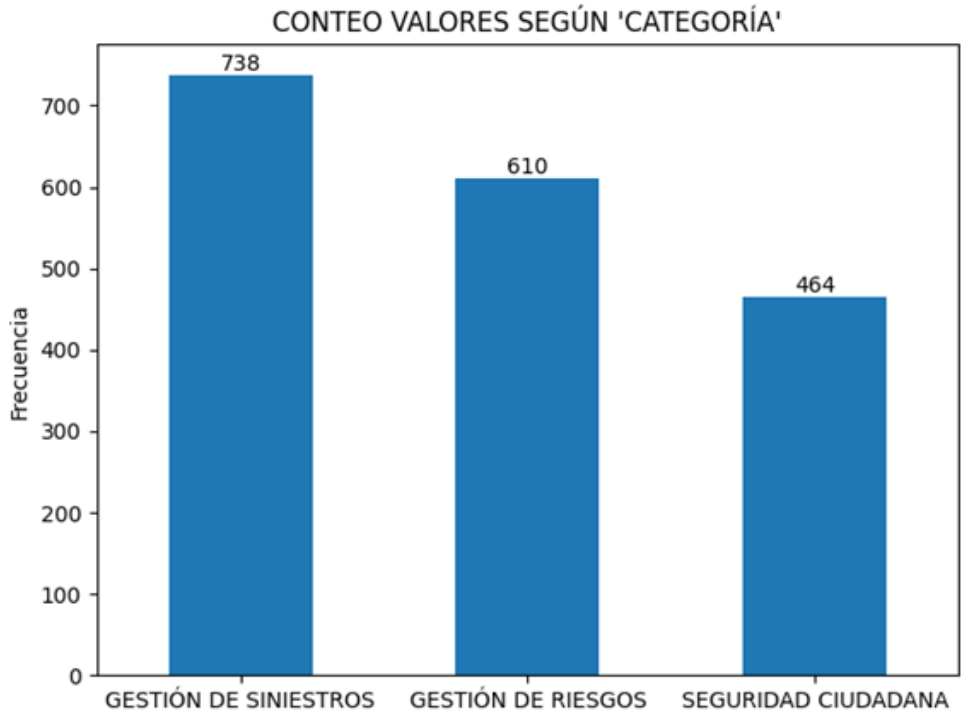

Ilustración 2. Distribución de los tweets de emergencia según su categoría

6.2. ESTRATEGIA PARA LA IDENTIFICACIÓN DE PATRONES LINGÜÍSTICOS

A partir del análisis del estilo de escritura empleado por los usuarios en las situaciones de emergencia y no emergencia, se busca identificar variaciones lingüísticas, el tono, y otras características que pueden influir en la identificación de *tweets* relacionados con emergencias.

El uso de técnicas avanzadas de procesamiento de lenguaje natural (NLP), permiten descomponer los textos y se extraer características relevantes, aspectos que se emplearán en los datos etiquetados como emergencia y no emergencia, con el fin de identificar patrones específicos en el uso del lenguaje y establecer diferencias en el estilo de escritura para cada situación.

Un punto clave, previo al análisis de los patrones lingüísticos, es conocer la distribución de los datos según las regiones del Ecuador debido a las variaciones culturales entre cada región. Dicho aspecto permite establecer una visión clara respecto al origen de los *tweets* y por ende delimitar el análisis presentado.

En este caso, el conjunto de datos generado previamente presenta un mayor porcentaje de *tweets* la región sierra, seguido de la región costa, y con una muy baja presencia de datos provenientes de la región del Oriente. Por tanto, la data contempla una notoria orientación a textos provenientes de la región sierra.

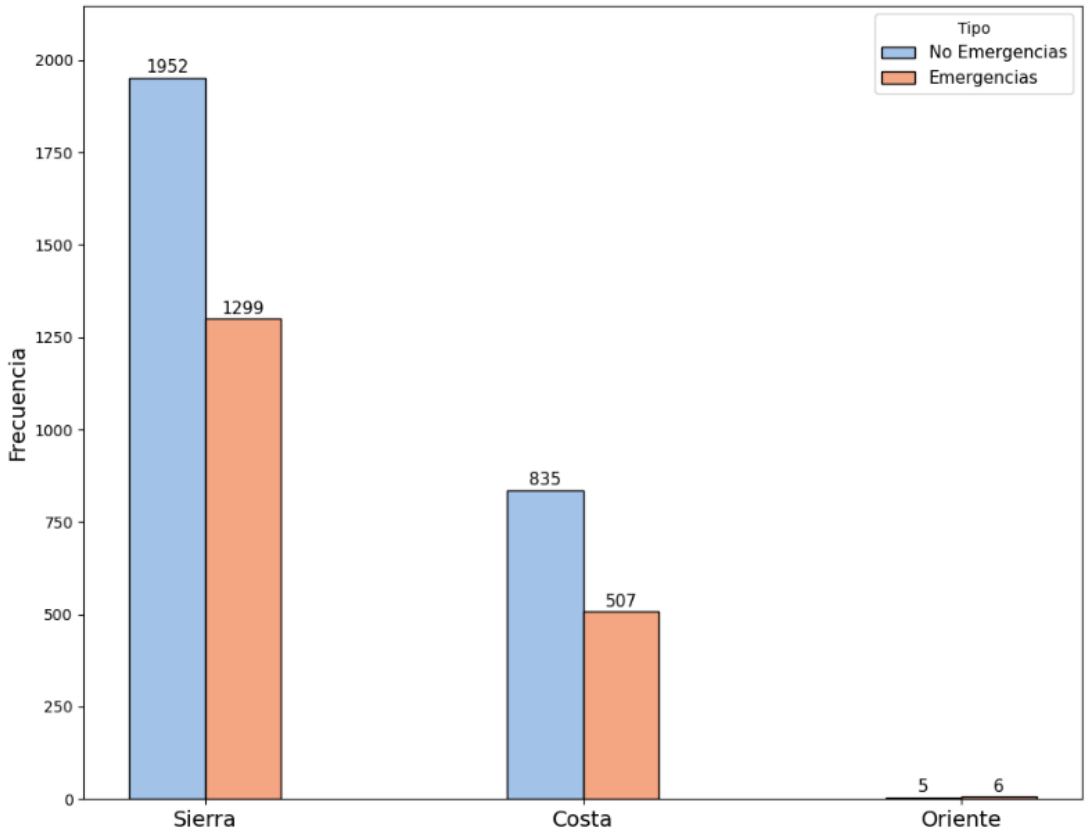

Ilustración 3. Distribución detallada de los tweets por región

La ilustración 3 presenta una distribución detallada de los *tweets* en cada región del Ecuador. Cabe indicar que, dichas cantidades no constituyen como un indicativo de las zonas de mayor incidencia de emergencias, debido a que las cantidades de los datos responden a las obtenidas mediante la plataforma "Apify", tal y como se mencionó previamente, no obstante, son cantidades que resultan representativas para determinar patrones lingüísticos, enfocado principalmente en la región Sierra y Costa.

Comprensión de los Datos Enfocado al Estilo de Escritura

La comprensión del estilo de escritura es un aspecto que permite identificar los patrones lingüísticos empleados por los usuarios en los *tweets* de emergencia y no emergencia. Para ello se analizarán varios elementos claves que reflejan la forma de comunicación en cada contexto:

- **Longitud Promedio de *tweets* (LPT)**: La longitud de las oraciones puede ofrecer referencias sobre la urgencia o complejidad del mensaje, a partir de si son cortas y directas, o más largas y elaboradas.

$$LPT = Total\ de\ palabras\ /\ Cantidad\ total\ de\ tweets$$

- **Longitud Promedio de Palabras (LPP)**: La longitud promedio de las palabras utilizadas en los *tweets* puede proporcionar información sobre la complejidad del lenguaje. Palabras más largas pueden indicar un uso más formal o técnico del lenguaje, mientras que palabras más cortas pueden ser indicativas de un lenguaje más coloquial o directo.

$$LPP = Suma\ total\ de\ la\ longitud\ de\ todas\ las\ palabras\ /\ Total\ de\ palabras$$

- **Diversidad Léxica (DL)**: La cantidad de palabras únicas en un conjunto de datos es una medida de la diversidad léxica. Un alto número de palabras únicas puede indicar una mayor variedad en el vocabulario utilizado. Esta métrica se calcula contando las palabras distintas en cada conjunto de datos.

$$DL = Cantidad\ de\ palabras\ \text{únicas}\ /\ Total\ de\ palabras$$

- **Uso Promedio de Signos de Puntuación (USP)**: El uso de signos de puntuación, como puntos, comas, signos de exclamación e interrogación, puede proporcionar información sobre el tono y la estructura del mensaje.

$$USP = Total\ de\ signos\ de\ puntuación\ /\ Cantidad\ total\ de\ tweets$$

- **Uso Promedio de Palabras en Mayúsculas (UPM)**: Analizar la proporción de palabras en mayúsculas en cada conjunto de datos permite identificar patrones de comunicación relevantes. Las palabras en mayúsculas pueden indicar un esfuerzo por destacar información importante en situaciones críticas.

$$UPM = Total\ de\ palabras\ en\ may\text{úsculas}\ /\ Cantidad\ total\ de\ tweets$$

- **Análisis de Bigramas**: La identificación de bigramas (pares de palabras consecutivas) ayuda a comprender mejor las combinaciones de palabras características de los *tweets*. Al analizar la incidencia de estos bigramas en cada contexto, se pueden identificar frases recurrentes y patrones específicos que actúan como indicadores.

Análisis de la frecuencia de terminología ecuatoriana

Para determinar la frecuencia de terminología ecuatoriana y su incidencia, se establecen dos validaciones para el conjunto de datos, la primera parte de los resultados generados a partir de una nube de palabras, donde se busca identificar la presencia de terminología presente en el territorio ecuatoriano a partir de la visualización de las palabras más significativas en los textos etiquetados como emergencias y no emergencias. La siguiente validación radica en la búsqueda directa de terminología ecuatoriana que reflejen las particularidades lingüísticas locales a partir de un conjunto de palabras que tengan relación con situaciones de emergencia. Estas palabras se definen considerando las investigaciones de Izurieta-Brito et al. (2020) y Rojas (2020).

A partir de ello, los términos escogidos fueron agrupados en tres categorías: afirmaciones, desacuerdos y vulgarismos. En las investigaciones mencionadas, existe una gran variedad de términos ecuatorianos, pero como se ha mencionado, se escogieron aquellos que, por su significado, presentan una mayor incidencia para comprender el contexto de los *tweets*. Este análisis ayudará a determinar la frecuencia de aparición de estas palabras, permitiendo comprender cómo los usuarios ecuatorianos se expresan en diferentes contextos, y especialmente la incidencia de dichos términos.

Tabla 3. Agrupación de términos coloquiales según categoría

CATEGORÍA	TÉRMINOS
Afirmaciones	• chevere • bacan • pilas • simon • posi
Desacuerdos	• Chuta • Vaina • Huevada • Pendejada • Joder • mierda

Vulgarismos	• Verga • Gaver • Conchetu • Chucha • Cabron • Cojudo • Mamaverga • Hijodeputa

Los términos relacionados a las afirmaciones pueden aparecer en *tweets* que describen situaciones controladas o alerta ante situaciones confirmadas. Por ejemplo, un *tweet* que dice "Hay disparos cerca del centro de la ciudad, pilas en no ir cerca de allá" utiliza "pilas" para indicar estar atento a alguna situación. En situaciones de no emergencia, estas palabras pueden aparecer en contextos cotidianos o informales, como "El concierto estuvo bacan".

Por su parte, los términos relaciones a desacuerdos expresan desaprobación, frustración o negaciones. Estas palabras son comunes en descripciones de situaciones negativas o críticas durante emergencias. Por ejemplo, "¡Chuta! La vaina se salió de control, muchos heridos" usa "chuta" y "vaina" para expresar preocupación y gravedad de la situación. En situaciones de no emergencia, estas palabras pueden aparecer en contextos de queja o descontento diario, como "¡Joder! Se me olvidó comprar leche".

Los vulgarismos son palabras fuertes que expresan emociones intensas, como ira, desesperación o sorpresa. Durante emergencias, los usuarios pueden recurrir a estos términos para enfatizar la gravedad de la situación. Por ejemplo, "¡Verga! El ladrón se llevó todo de la tienda" utiliza "verga" para enfatizar la intensidad del evento. En situaciones de no emergencia, estos términos pueden aparecer en contextos de alta tensión emocional o informalidad extrema, como "¡Chucha! Perdí las llaves de nuevo".

6.3. ESTRATEGIA PARA EL DESARROLLO DEL MODELO DE INTELIGENCIA ARTIFICIAL

El desarrollo de un modelo de inteligencia artificial (IA) para la identificación automática de *tweets* que reporten emergencias en Ecuador implica una serie de etapas fundamentales, cada una crucial para garantizar la precisión y efectividad del modelo final.

Preparación de Datos

Se definen varias funciones específicas diseñadas para limpiar y estandarizar el texto de los *tweets*. Las principales funciones utilizadas se encargan de eliminar menciones de usuarios, *retweets*, *URLs*, *hashtags*, caracteres no alfanuméricos, y saltos de línea, además de convertir el texto a minúsculas, eliminar emojis, números y *stopwords*, y normalizar caracteres no ASCII.

Tabla 4. Ejemplificación de tweets preprocesado

***Tweet* Original**	***Tweet P*reprocesado**
#DenunciaCiudadana Atentos con este delincuente!! Roba los retrovisores de los vehículos mientras esperan el cambio del semáforo en el túnel de retorno al mall del fortín. Su especialidad son las mujeres porque sabe que no se van a bajar a seguirlo. Quiero ver esos 3 Doritos Después.	denunciaciudadana atentos delincuente roba retrovisores vehiculos mientras esperan cambio semaforo tunel retorno mall fortin especialidad mujeres sabe van bajar seguirlo quiero ver doritos despues
Accidente de tránsito km 19 vía a la costa, vehículo se impacta contra un árbol https://t.co/WqgdqO7q9r	accidente transito km via costa vehiculo impacta árbol

Nivelación de la Data

El conjunto de datos cuenta con 2874 *tweets* de no emergencias y 1,812 *tweets* de emergencias. Para nivelar los datos, se ajustó el número de *tweets* de no emergencias a 1,812, igualando así el número de *tweets* de ambas clases. Este equilibrio asegura que el modelo representará igual a ambas categorías, fundamental para su entrenamiento y evaluación.

Tabla 5. Distribución de la cantidad de tweets según etiqueta

Detalle	**Data Sin Nivelar**	**Data Nivelada**
Emergencias	1,812	1,812
No Emergencias	2874	1,812
Total	**4686**	**3624**

Técnicas de Vectorización

Para el entrenamiento de los modelos, el conjunto de datos será vectorizado para cada técnica de vectorización establecida, este enfoque asegura que se exploren múltiples representaciones de los datos textuales, maximizando las oportunidades de identificar la técnica de vectorización más efectiva para cada modelo.

Para convertir el texto en un formato numérico procesable por los algoritmos de aprendizaje automático, se establecen las siguientes técnicas de vectorización:

- **TF-IDF (*Term Frequency-Inverse Document Frequency*):** Utiliza el TfidfVectorizer de *sklearn* para transformar los datos textuales en una representación numérica que refleja la importancia de las palabras en el corpus.
- **GloVe (*Global Vectors for Word Representation*):** Carga *embeddings* preentrenados mediante una función personalizada load_glove_model, proporcionando una representación densa y semánticamente rica de las palabras.
- **Word2Vec:** Utiliza modelos preentrenados de *Word2Vec* disponibles en la biblioteca *gensim*, que generan *embeddings* de palabras basados en su contexto de aparición.

Creación de Modelos

Se establecen varios tipos de modelos para explorar diferentes enfoques y seleccionar el más eficaz según métricas de rendimiento, cada uno estará conIlustración do con sus hiperparámetros predeterminados, para comparar sus capacidades en una tarea específica:

- **Regresión Logística:** Un modelo lineal simple adecuado para tareas de clasificación binaria. Este modelo es relevante debido a su interpretación directa y su capacidad para manejar problemas lineales de clasificación de manera eficiente.

- **Máquina de Soporte Vectorial (SVM):** Conocido por su potencia en clasificación y su capacidad para encontrar el hiperplano óptimo que separa las clases. Este enfoque es útil en conjuntos de datos de alta dimensionalidad.

- **Bosque Aleatorio:** Un modelo de ensamble que utiliza múltiples árboles de decisión para mejorar la precisión y reducir el sobreajuste. Este modelo es relevante debido a su capacidad para manejar datos ruidosos y complejos.
- **Red Neuronal:** Este modelo es relevante debido a su flexibilidad y capacidad para modelar relaciones no lineales en el texto.

Entrenamiento y Evaluación de Modelos

Se empleará validación cruzada, específicamente con *KFold cross-validation*, para asegurar que los modelos fueran entrenados y evaluados de manera robusta y en diferentes particiones del conjunto de datos.

La evaluación de los modelos resultantes partirá de la utilizando de un conjunto de métricas de rendimiento (*precision*, *recall*, *F1-score* y *AUC-ROC)*, con la finalidad de realizar una evaluación integral del rendimiento de los modelos, considerando tanto su capacidad para identificar correctamente los *tweets* de emergencia (*precision* y *recall*) como su capacidad para equilibrar los falsos positivos y negativos (*F1-score y AUC-ROC*).

La evaluación de los modelos también considera pruebas con los propios datos de entrenamiento. Si un modelo obtiene un rendimiento muy alto en los datos de entrenamiento, pero bajo en los datos de prueba, esto indica sobreajuste (*overfitting*), donde el modelo se ajusta demasiado a los datos de entrenamiento y no puede generalizar a nuevos datos.

7. RESULTADOS

7.1. PATRONES LINGÜÍSTICOS

Para la identificación de patrones lingüísticos en los *tweets* de emergencia y no emergencia en Ecuador, se ha realizado un análisis de diversas características textuales, análisis frecuencia de términos y análisis de la incidencia de la terminología ecuatoriana. Se generaron Ilustración s para visualizar la distribución de estas métricas, y se calcularon estadísticas descriptivas para proporcionar un resumen cuantitativo de los datos.

Longitud de los *Tweets*

Se realizó un análisis de densidad de la longitud de los *tweets* etiquetados como de emergencia y no emergencia. Los datos se resumen en las siguientes estadísticas:

Tabla 6. Descripción del conjunto de datos a partir de la longitud de tweets

Métrica	***Tweets de Emergencia***	***Tweets* de no Emergencia**
Cantidad	1,812	1,812
Media	223.8913	198.6038
Desviación estándar	72.4630	91.2396
Mínimo	23.0000	10.0000
1er cuartil (25%)	179.7500	121.0000
Mediana (50%)	234.0000	211.0000
3er cuartil (75%)	280.0000	277.0000
Máximo	1178.0000	701.0000

Los *tweets* de emergencia tienen una longitud mayor, con una media de 223.8913 caracteres y una mediana de 234.0000 caracteres. Esto sugiere que, en situaciones de emergencia, los usuarios tienden a proporcionar más detalles y contexto. Los *tweets* de no emergencia presentan una longitud menor, con una media de 198.6038 caracteres y una mediana de 211.0000 caracteres. Esto indica que, en contextos no urgentes, los usuarios pueden ser más breves y menos detallados en su comunicación.

La desviación estándar es mayor en los *tweets* de no emergencia (91.2396) en comparación con los *tweets* de emergencia (72.4630), lo que indica una mayor variabilidad en la longitud de los *tweets* en contextos no urgentes

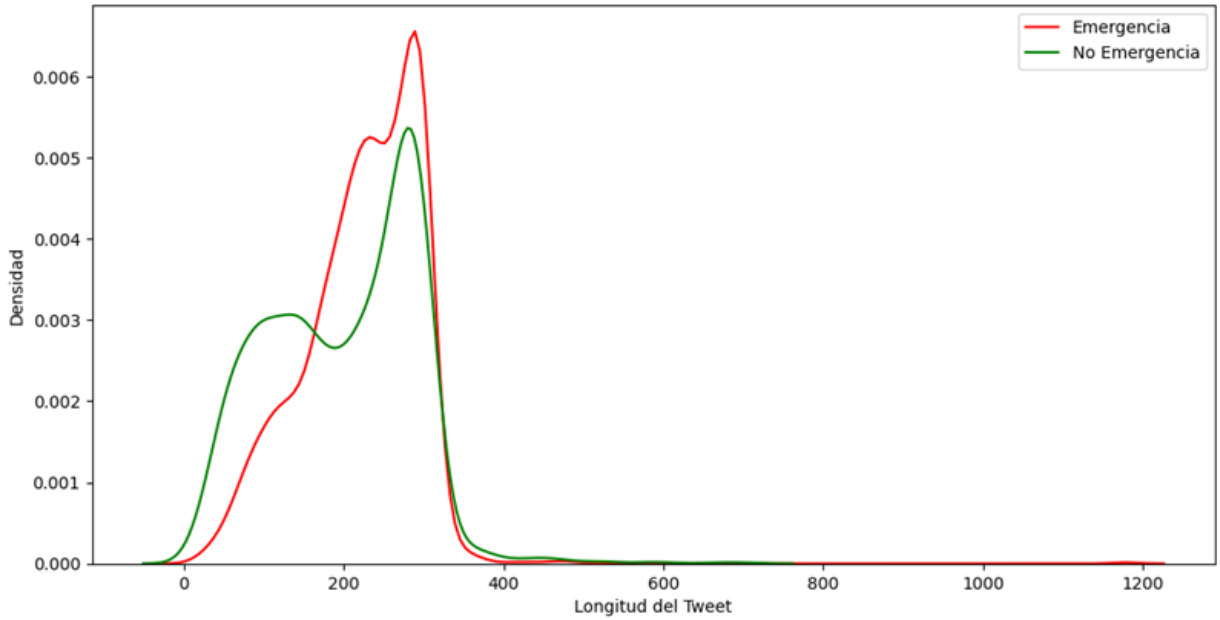

Ilustración 4. Distribución de longitud de tweets

La Ilustración 4 muestra que la mayor densidad para los *tweets* de emergencia es de 200-250 caracteres, mientras que, para los de no emergencia, la densidad alcanza su máximo entre 150-200 caracteres.

Los *tweets* de emergencia tienden a ser más largos y detallados, reflejando la necesidad de comunicar información crucial de manera clara y extensa. En contraste, los *tweets* de no emergencia presentan una mayor variabilidad en la longitud, indicando un estilo de comunicación más relajado y menos estructurado.

Longitud Promedio de Palabras en *Tweets*

Tabla 7. Descripción del conjunto de datos a partir de la longitud de palabras.

Métrica	***Tweets* de Emergencia**	***Tweets* de no Emergencia**
Cantidad	1,812	1,812
Media	7.0209	6.5442
Desviación estándar	0.7088	1.0007
Mínimo	4.2500	2.0000
1er cuartil (25%)	6.5833	5.9231
Mediana (50%)	7.0000	6.5455
3er cuartil (75%)	7.4807	7.1538
Máximo	10.5000	13.5000

Los *tweets* de emergencia tienen una longitud promedio de palabras de 7.0209 caracteres, mayor que la longitud promedio de 6.5442 caracteres en los *tweets* de no emergencia. La variabilidad es menor en los *tweets* de emergencia (0.7088) en comparación con los *tweets* de no emergencia (1.0007). La longitud mínima es mayor en los *tweets* de emergencia (4.2500) en comparación con los *tweets* de no emergencia (2.0000).

Sin embargo, la longitud máxima es mayor en los *tweets* de no emergencia (13.5000) en comparación con los *tweets* de emergencia (10.5000). La longitud promedio de palabras en los *tweets* de emergencia es mayor en los tres cuartiles que en comparación con los de no emergencia.

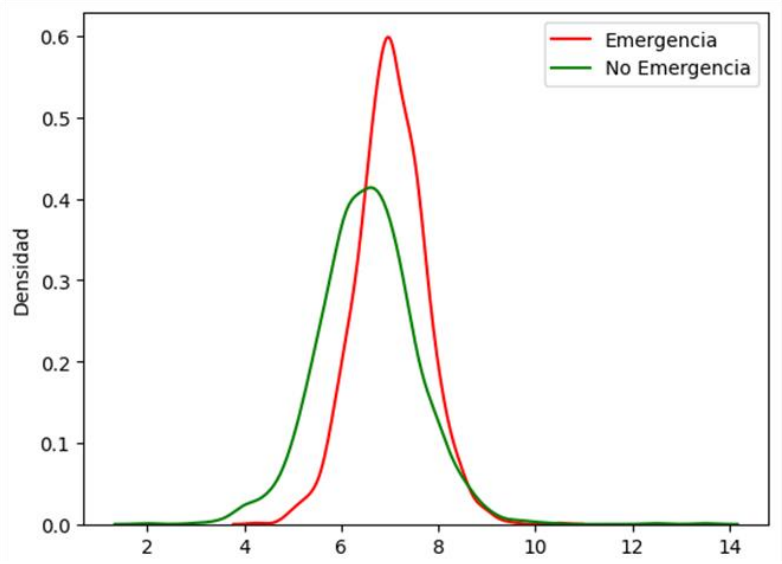

Ilustración 5. Longitud promedio de palabras

La ilustración 5 muestra que los *tweets* de emergencia tienden a tener una longitud promedio de palabras ligeramente mayor que los *tweets* de no emergencia. La mayor densidad para los *tweets* de emergencia se encuentra alrededor de 7 caracteres, mientras que para los *tweets* de no emergencia se encuentra alrededor de 6.5 caracteres. Las distribuciones son bastante similares, pero con una ligera desviación hacia palabras más largas en los *tweets* de emergencia.

Los *tweets* de emergencia utilizan palabras más largas en promedio, lo que refleja un esfuerzo por comunicar información detallada y precisa en situaciones críticas. En contraste, los *tweets* de no emergencia muestran una mayor variabilidad en la longitud de palabras, indicando un estilo de comunicación más flexible y menos estructurado.

Palabras en Mayúsculas en *Tweets*

Tabla 8. Descripción del conjunto de datos a partir de las mayúsculas.

Métrica	***Tweets* de Emergencia**	***Tweets* de No Emergencia**
Cantidad	1,812	1,812
Media	0.0286	0.0497
Desviación estándar	0.0737	0.1340
Mínimo	0.0000	0.0000
1er cuartil (25%)	0.0000	0.0000
Mediana (50%)	0.0000	0.0000
3er cuartil (75%)	0.0333	0.0435
Máximo	0.9783	1.0000

Los *tweets* de emergencia tienen una proporción promedio de 0.0286 palabras en mayúsculas, mientras que los *tweets* de no emergencia tienen una proporción de 0.0497. La variabilidad es mayor en los *tweets* de no emergencia, con 0.1340 en comparación con 0.0737 para los *tweets* de emergencia. Ambos conjuntos tienen un mínimo de 0, pero los *tweets* de no emergencia tienen un máximo de 1.0000, mientras que los *tweets* de emergencia tienen un máximo de 0.9783. En ambos conjuntos, el 25% y 50% de los *tweets* no contienen palabras en mayúsculas. Sin embargo, el tercer cuartil es ligeramente mayor en los *tweets* de no emergencia (0.0435) en comparación con los *tweets* de emergencia (0.0333).

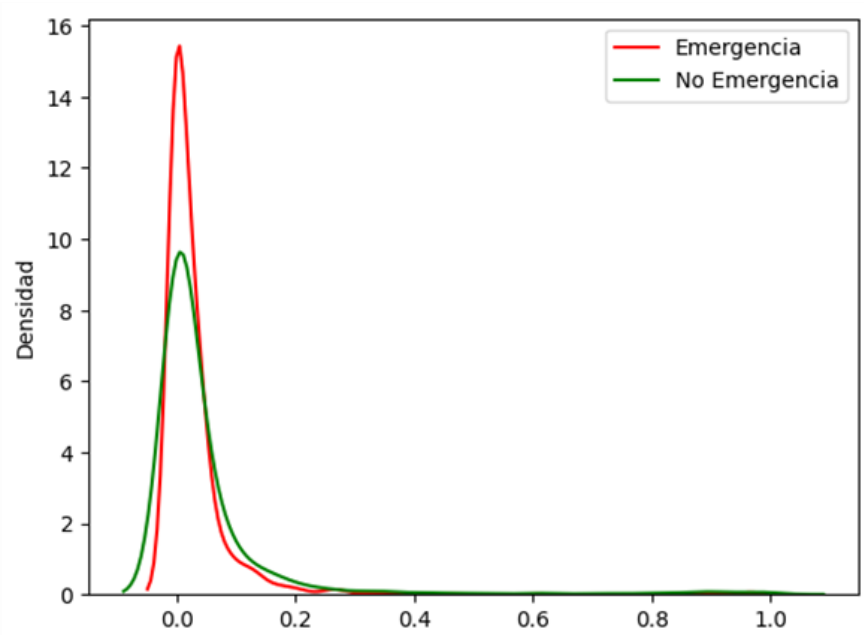

Ilustración 6. Proporción de palabras en mayúscula

La Ilustración 6 muestra que los *tweets* de emergencia tienen una menor proporción de palabras en mayúsculas en comparación con los *tweets* de no emergencia. La mayor densidad de ambos conjuntos se encuentra cerca del 0%, indicando que la mayoría de los *tweets* contienen pocas palabras en mayúsculas. Sin embargo, hay una ligera diferencia donde los *tweets* de no emergencia muestran una mayor densidad en proporciones ligeramente superiores a 0.

Los *tweets* de emergencia muestran un uso más controlado y específico de las mayúsculas, probablemente para resaltar información crítica en situaciones de urgencia. En contraste, los *tweets* de no emergencia presentan una mayor variabilidad en el uso de mayúsculas, reflejando un tono más informal y menos estructurado.

Uso de Signos de Puntuación en *Tweets*

El análisis del uso de signos de puntuación en los *tweets* reveló diferencias notables entre los *tweets* de emergencia y no emergencia.

Tabla 9. Descripción del conjunto de datos a partir del uso de signos de puntuación

Métrica	***Tweets* de Emergencia**	***Tweets* de no Emergencia**
Cantidad	1,812	1,812
Media	2.782	2.392
Desviación estándar	2.213	2.169
Mínimo	0	0

1er cuartil (25%)	1	1
Mediana (50%)	2	2
3er cuartil (75%)	4	3
Máximo	16	16

Los *tweets* de emergencia tienen un promedio de 2.782 signos de puntuación por *tweet*, mientras que los *tweets* de no emergencia tienen un promedio de 2.392 signos. La variabilidad es similar entre los dos conjuntos, con 2.213 para *tweets* de emergencia y 2.169 para *tweets* de no emergencia. Ambos conjuntos tienen un mínimo de 0 y un máximo de 16 signos de puntuación. Los *tweets* de emergencia muestran una mayor concentración de signos de puntuación en los percentiles superiores (75%), con 4 signos de puntuación, en comparación con 3 signos en los *tweets* de no emergencia.

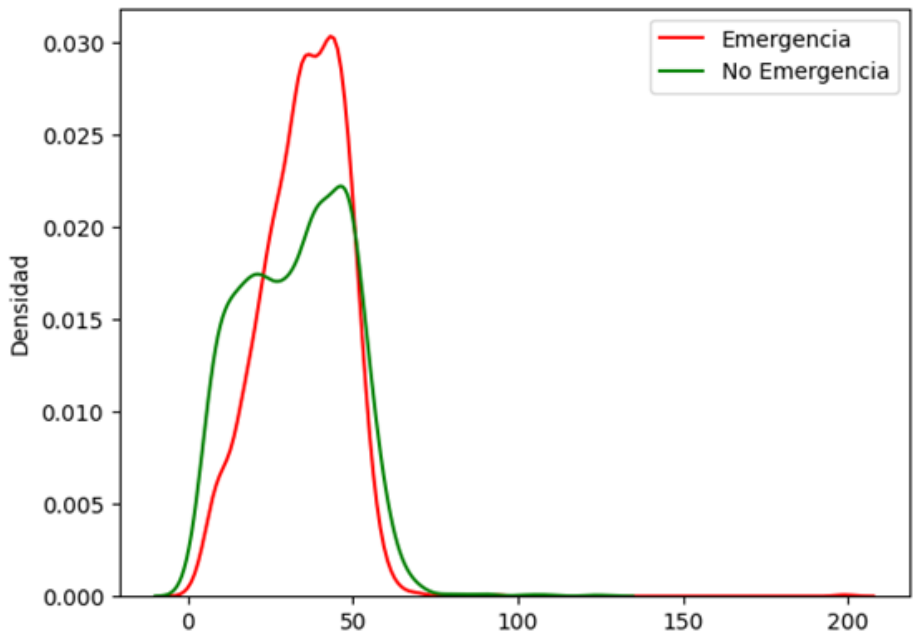

Ilustración 7. Proporción de signos de puntuación

La Ilustración 7 muestra una distribución muy similar para ambos tipos de *tweets*, con una ligera tendencia de los *tweets* de emergencia a usar más signos de puntuación. Los picos principales de densidad se encuentran cerca de 0, indicando que la mayoría de los *tweets* utilizan pocos signos de puntuación. Sin embargo, la cola derecha de la distribución de *tweets* de emergencia es un poco más larga, sugiriendo algunos *tweets* con un uso intensivo de signos de puntuación.

Los *tweets* de emergencia suelen incluir un número algo mayor de signos de puntuación que los de no emergencia, lo que indicaría un mayor esfuerzo en comunicar de manera precisa y clara en situaciones de emergencia.

Análisis de Riqueza Léxica

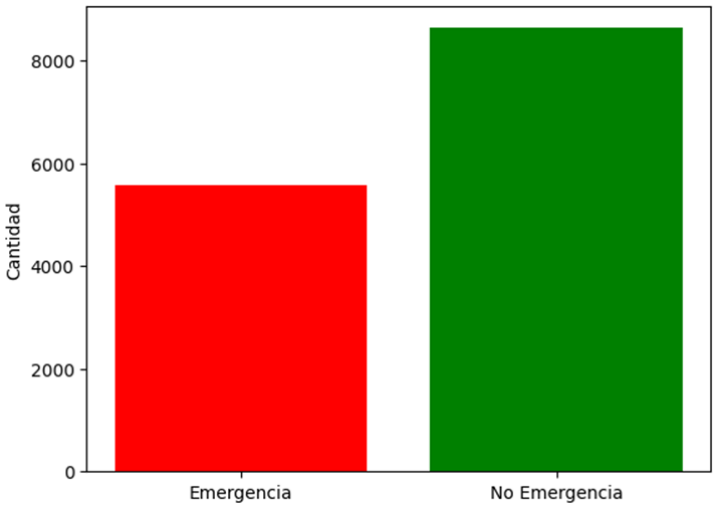

Ilustración 8. Cantidad de palabra únicas

El conjunto de datos de emergencia tiene 5,566 palabras únicas. Esta cifra indica un vocabulario relativamente especializado y restringido, probablemente debido a la necesidad de transmitir información específica y directa en situaciones críticas. La menor riqueza léxica sugiere un lenguaje más homogéneo y estandarizado, probablemente para facilitar la comprensión rápida y efectiva. La repetición de términos específicos relacionados con emergencias (como "herido", "accidente", "incendio") es común.

El conjunto de datos de no emergencia cuenta con 8,634 palabras únicas, lo que sugiere un vocabulario más amplio y variado. Esto es coherente con la naturaleza más diversa de los *tweets* no relacionados con emergencias, que pueden cubrir una amplia gama de temas y contextos. La mayor riqueza léxica indica una comunicación más variada y menos restringida. Los usuarios pueden emplear una gama más amplia de términos y expresiones, reflejando una variedad de temas y estilos de escritura.

Los *tweets* de emergencia tienden a tener un vocabulario más limitado y específico, adecuado para la comunicación rápida y clara en situaciones críticas. Por otro lado, los *tweets* de no emergencia muestran una mayor diversidad léxica, reflejando la variedad de temas y contextos en los que se utilizan.

Frecuencia de términos conjunto de datos *Tweets* de Emergencia

Análisis de la Nube de Palabras

Ilustración 9. Nube de palabras conjunto de datos de emergencias

En la nube de palabras, destacan términos como "pérdida", "transporte", "vehículo", "camioneta", y "herido". Estos términos son indicativos de situaciones comunes en *tweets* de emergencia. La prevalencia de estas palabras sugiere que una gran cantidad de *tweets* de emergencia en Ecuador están relacionados con incidentes vehiculares y sus consecuencias.

En la nube de palabras analizadas no se observan términos específicos de la terminología ecuatoriana. Esto puede deberse a la naturaleza formal y directa de los *tweets* de emergencia, donde los usuarios posiblemente prefieran usar un lenguaje más universal y menos coloquial para asegurar que su mensaje se comprenda inmediatamente y clara.

Análisis de Bigrama

El análisis de bigramas en los *tweets* de emergencia proporciona información valiosa sobre las combinaciones de palabras más frecuentes y características en situaciones de emergencia. Para el conjunto de datos de emergencia, los bigramas más comunes y su porcentaje de incidencia son los siguientes:

Tabla 10. Frecuencia de bigramas en tweets de emergencia

N-grama	Conteo	Porcentaje (%)
(personas, heridas)	267	1.0075
(accidente, tránsito)	241	0,9094
(incendio, forestal)	133	0,5019

(siniestro, tránsito)	116	0,4377
(dos, personas)	88	0,3320
(incendio, estructural)	88	0,3320
(resultaron heridas)	79	0,2981
(personas, resultaron)	69	0,2603
(ataque, armado)	64	0,2415
(cuerpo, bomberos)	58	0,2188

El bigrama más común, (personas, heridas), aparece 267 veces, lo que representa el 1.007585% de todos los *tweets* de emergencia. Bigramas como (accidente, transito) y (siniestro, transito) son muy frecuentes, apareciendo 241 y 116 veces respectivamente.

Los bigramas más frecuentes reflejan una tendencia a reportar accidentes de tránsito, incendios y personas heridas, así como la intervención de cuerpos de bomberos y la ocurrencia de ataques armados.

Frecuencia de términos conjunto de datos *Tweets* de No Emergencia

Análisis de la Nube de Palabras

Ilustración 10. Nube de palabras conjunto de datos de no emergencias

En la nube de palabras de no emergencia, destacan términos como "deportiva", "robo", "historia", "primera", "cielo", y "asalto". Estos términos indican que los temas comunes en los *tweets* de no emergencia varían ampliamente, desde eventos deportivos hasta temas de interés general y situaciones de la vida cotidiana.

Como se estableció en la metodología, la búsqueda de *tweets* se definió a partir de conjuntos de palabras relacionadas a emergencias, por ello, aunque robo y asalto son palabras que también pueden aparecer en contextos de emergencia, su presencia en esta nube indica que los usuarios también emplean dichos términos en contextos no críticos, posiblemente como eventos narrativos o informativos, o acusaciones relacionadas a diversos tipos de eventos de incidencia popular como deportes o elecciones.

Al igual que en los *tweets* de emergencia, en esta nube de palabras no se observan términos específicos de la terminología ecuatoriana. Esto sugiere que los usuarios tienden a utilizar un lenguaje más universal y menos coloquial en *tweets* que manejan términos de emergencia, pero el contexto indica situaciones de no emergencia.

Análisis de Bigrama

Para el conjunto de datos de emergencia, los bigramas más comunes y su porcentaje de incidencia son los siguientes:

Tabla 11. Frecuencia de bigramas en tweets de no emergencia

N-grama	Conteo	Porcentaje (%)
(cuerpo, bomberos)	57	0,2288
(varios, sectores)	54	0,2168
(fuerte, lluvia)	47	0,1887
(personas, heridas)	35	0,1405
(incendio, forestal)	34	0,1365
(sala, situacional)	23	0,0923
(accidente, tránsito)	20	0,0803
(ataque, armado)	17	0,0682
(debe, ser)	16	0,0642
(muerte, cruzada)	16	0,0642

El bigrama más frecuente, (cuerpo, bomberos), aparece 57 veces, representando el 0.2288% de todos los *tweets* de no emergencia. Esto sugiere que, aunque no se trata de emergencias críticas, los cuerpos de bomberos son mencionados con cierta frecuencia, posiblemente en contextos informativos o preventivos. Bigramas como **(personas, heridas)** y **(accidente, transito)** aparecen 35 y 20 veces respectivamente.

Aunque estos términos también se encuentran en los *tweets* de emergencia, su presencia en los *tweets* de no emergencia sugiere que se discuten en un contexto más general o retrospectivo. De igual manera, el bigrama **(ataque, armado)** aparece 17 veces (0.0682%), indicando que, aunque se mencionan ataques armados, no se hace con la misma frecuencia e intensidad que en los *tweets* de emergencia.

El análisis de bigramas en los *tweets* de no emergencia revela combinaciones de palabras que están más relacionadas con contextos informativos, preventivos y descriptivos, en lugar de situaciones críticas. Los temas comunes incluyen servicios de emergencia, condiciones meteorológicas, reportes de heridos y accidentes, y menciones a sectores variados y contextos institucionales.

7.2. INCIDENCIA DE LA TERMINOLOGÍA ECUATORIANA

Análisis de la Incidencia de Términos Coloquiales en *Tweets* de Emergencia y No Emergencia

El siguiente análisis se centra en la incidencia de términos coloquiales específicos dentro del conjunto de datos de *tweets* de emergencia y no emergencia en Ecuador.

Los términos analizados incluyen afirmaciones, negaciones y vulgarismos definidos previamente.

Tabla 12. Frecuencia de la incidencia de términos coloquiales

Métrica	***Tweets* de Emergencia**	***Tweets* de no Emergencia**
Cantidad	1,812	1,812
Media	0.0253	0.0237
Desviación estándar	0.1675	0.1627
Mínimo	0.0000	0.0000
1er cuartil (25%)	0.0000	0.0000
Mediana (50%)	0.0000	0.0000
3er cuartil (75%)	0.0000	0.0000
Máximo	2.0000	2.0000

Ambas categorías de *tweets* (emergencia y no emergencia) tienen una media y mediana de incidencia muy bajas (cercanas a cero), lo que indica que la mayoría de los *tweets* no contienen términos coloquiales.

La media de *tweets* de emergencia es ligeramente superior a la de no emergencia. La desviación estándar es similar para ambos conjuntos de datos, lo que sugiere que la variabilidad en la incidencia de términos coloquiales es comparable en ambas categorías. La alta frecuencia de la incidencia cero y la baja frecuencia de incidencias mayores a cero indican una distribución asimétrica hacia la izquierda, con algunos documentos presentando una incidencia significativa (hasta 2) de términos coloquiales.

La baja incidencia de términos coloquiales en *tweets* de emergencia puede deberse a la naturaleza más formal y directa de los mensajes en situaciones críticas, donde los usuarios prefieren utilizar un lenguaje claro y universal. En situaciones de no emergencia, aunque la incidencia de términos coloquiales también es baja, puede haber un uso ligeramente más frecuente de estos términos en contextos narrativos y descriptivos.

Análisis Frecuencia de Términos Coloquiales

La incidencia de términos coloquiales en ambos conjuntos de datos es baja, con una media y mediana cercanas a cero. Sin embargo, es notable que hay una ligera diferencia en la media de incidencia entre *tweets* de emergencia y no emergencia, siendo ligeramente superior en los primeros.

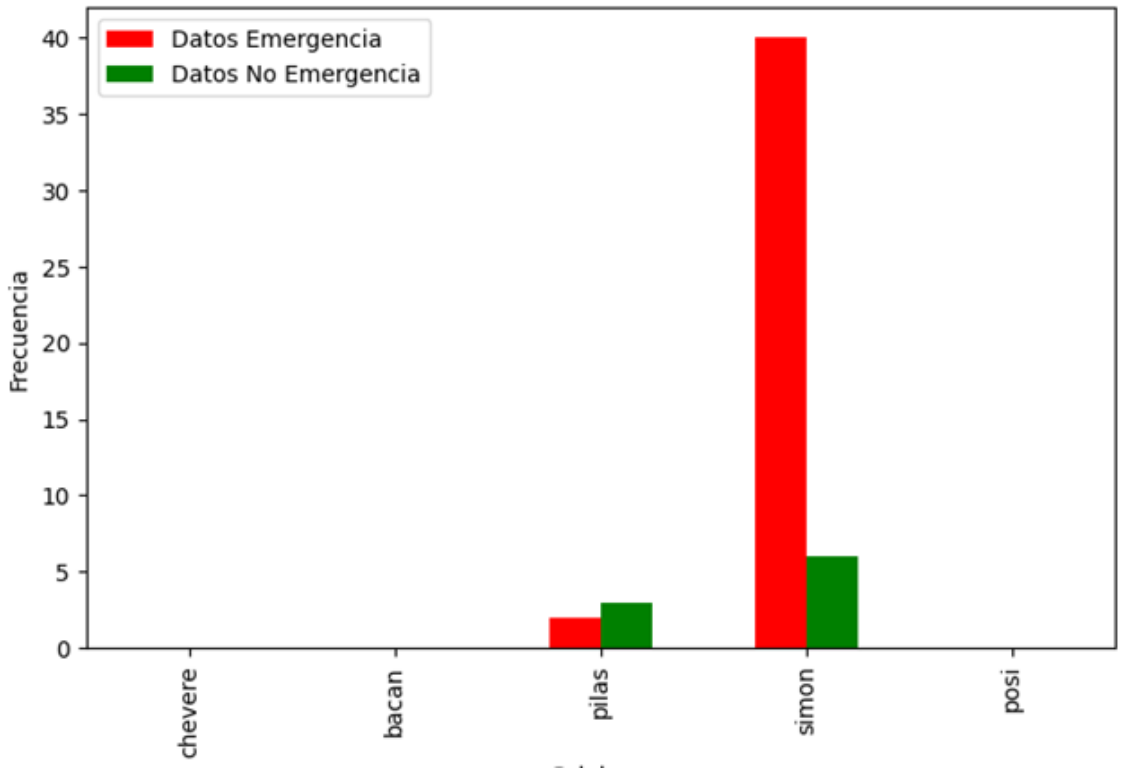

Ilustración 11. Incidencia del conjunto de palabras de afirmaciones

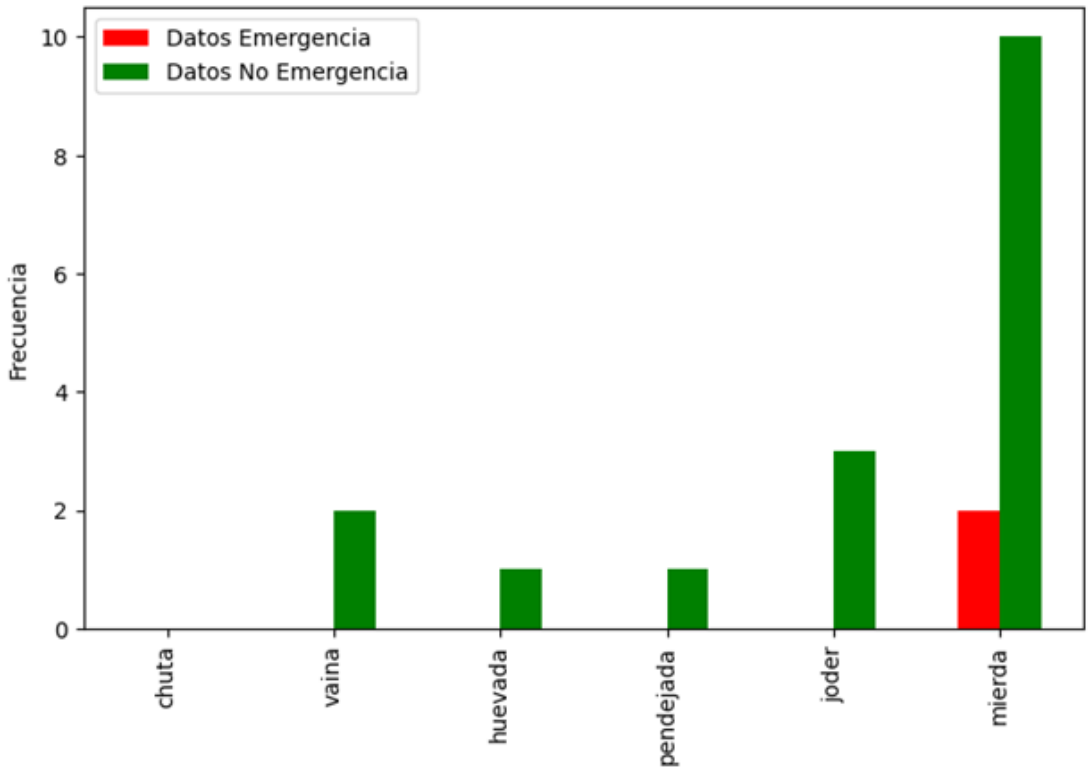

Ilustración 12. Incidencia del conjunto de palabras de Desacuerdo

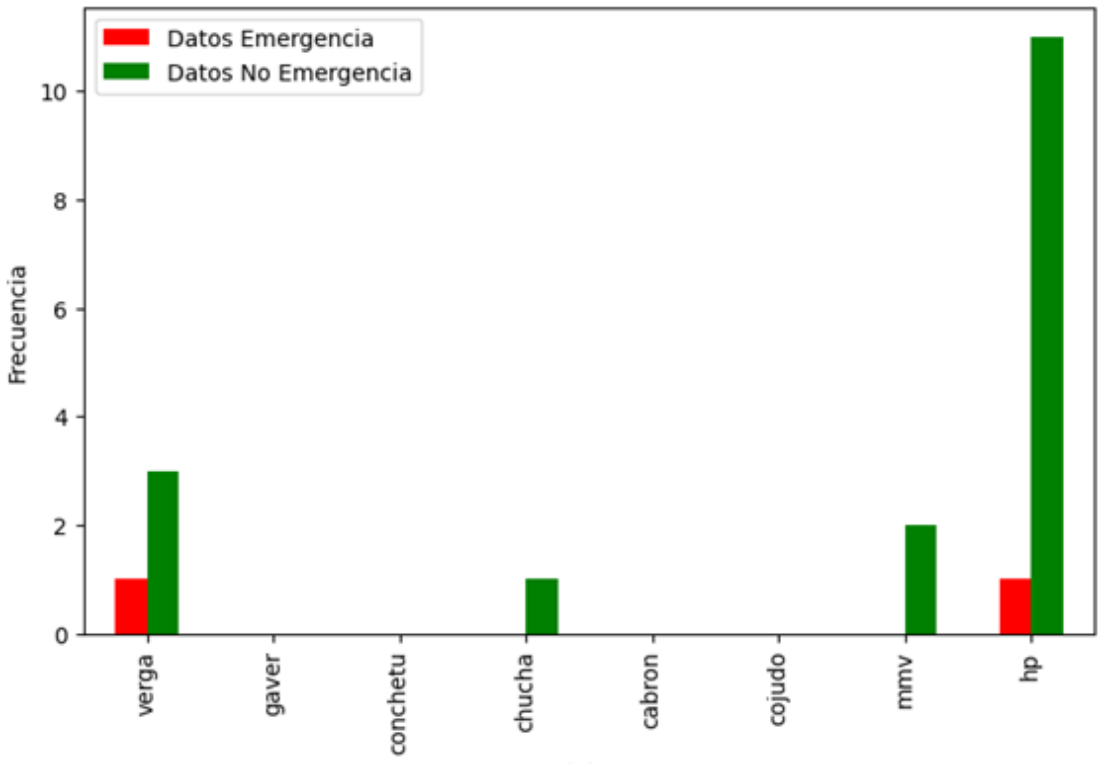

Ilustración 13. Incidencia del conjunto de palabras de vulgarismos

***Tweets* de Emergencia:**

- El término "simon" tiene una incidencia notable con 40 apariciones, lo que indica su uso en situaciones de emergencia, posiblemente en contextos de afirmación o conformidad ante eventos críticos.
- Palabras como "pilas", "mierda", y "verga" tienen una baja pero significativa aparición, reflejando la naturaleza emocional y urgente del contexto de emergencia.
- La mayoría de los otros términos no tienen aparición alguna, sugiriendo que los usuarios prefieren un lenguaje menos coloquial en situaciones críticas.

***Tweets* de No Emergencia:**

- El término "hp" aparece 11 veces, mientras que "mierda" tiene 10 apariciones. Estos términos pueden indicar una expresión de frustración o intensidad emocional en situaciones no críticas.
- Términos como "simon", "joder", y "verga" tienen una presencia moderada, indicando un uso más coloquial y emocional en contextos no urgentes.
- Similar a los *tweets* de emergencia, muchos términos coloquiales tienen una baja incidencia, pero hay una mayor variedad en el uso de términos como "vaina", "huevada", y "pendejada".

Los términos "simon" y "pilas" aparecen tanto en *tweets* de emergencia como en no emergencia, aunque "simon" es significativamente más frecuente en emergencias. Palabras como "mierda", "verga", y "hp" son más comunes en *tweets* de no emergencia, lo que sugiere un uso más coloquial y emocional en situaciones no críticas.

La baja incidencia de muchos términos coloquiales en ambos conjuntos de datos refuerza la hipótesis de que los usuarios tienden a utilizar un lenguaje más directo y menos coloquial en situaciones de emergencia.

7.3. EVALUACIÓN DE LOS RESULTADOS DEL MODELO

Análisis general

Tras una comparación de los resultados, mediante técnicas de validación cruzada y las métricas de rendimiento, incluyendo *accuracy*, *recall*, *precision* y *F1-score*, tanto en los conjuntos de entrenamiento como en los de prueba, se obtuvo que, los resultados de los modelos entrenados con *TF-IDF* presentaron un mejor rendimiento general y menor tendencia al *overfitting* en comparación con las técnicas de vectorización *GloVe* y *Word2Vec*.

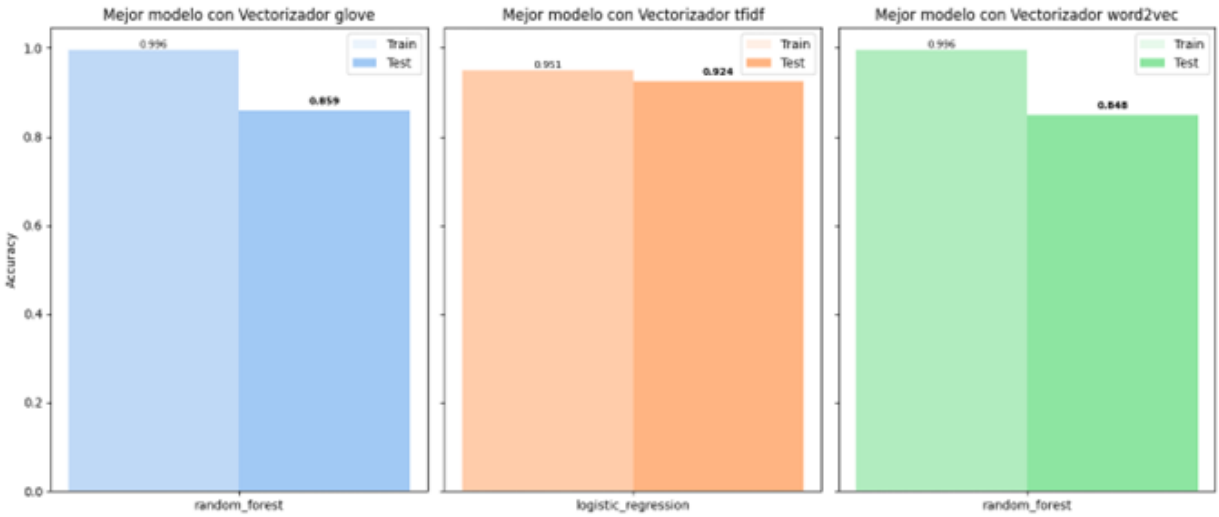

Ilustración 14. Resultados de los mejores modelos por vectorizador

Como se aprecia en la ilustración 14, el modelo que presenta los mejores resultados es el de regresión logística, mediante el vectorizador *tfidf*. Con la finalidad de generar un análisis más preciso del mejor modelo, se filtran los mejores resultados por modelo a partir de la métrica *accurracy* considerando el vectorizador TFIDF, quien presenta el mejor resultado en general, esta selección es considerando los resultados del conjunto de datos de prueba.

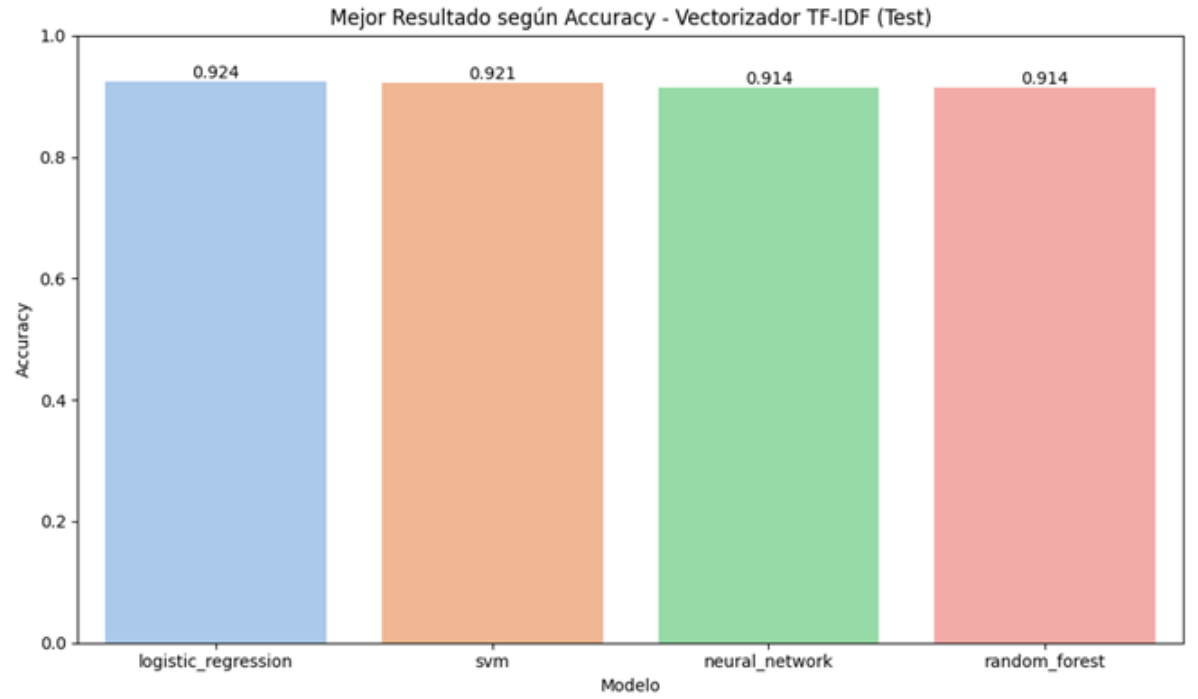

Ilustración 15. Mejor resultado por modelo mediante vectorizador tfidf

Tabla 13. Comparaciones mejores resultados por modelo mediante vectorizador tfidf

	model	vectorizer	accuracy	recall	precision	f1
0	logistic_regression	tfidf	0.9241	0.9059	0.9439	0.9245
1	svm	tfidf	0.9213	0.9237	0.9373	0.9232
2	neural_network	tfidf	0.9144	0.9252	0.9305	0.9153
3	random_forest	tfidf	0.9144	0.8951	0.9353	0.9148

Para el análisis de resultados se prioriza la detección de emergencias (*recall*).

- La Regresión Logística muestra un desempeño robusto con un *accuracy* de 92.41%, lo que indica que el modelo clasifica correctamente la mayoría de los *tweets*. Su alta precisión (94.40%) refleja una excelente capacidad para minimizar falsas alarmas, mientras que un *recall* de 90.59% asegura que una proporción significativa de emergencias es detectada. El *F1-score* de 92.46% destaca el balance óptimo entre *precision* y *recall*, lo que lo hace un modelo confiable para aplicaciones prácticas.
- El modelo SVM presenta un *accuracy* de 92.14%, ligeramente inferior al de Regresión Logística. No obstante, muestra un *recall* superior de 92.37%, lo que indica una mejor capacidad para identificar *tweets* de emergencia. La *precision* de 93.73% es alta, aunque ligeramente menor que la de la Regresión Logística. El *F1-score* de 92.33% refleja un buen balance entre *precision* y *recall*, destacándose como un modelo muy competitivo.
- Las Redes Neuronales alcanzan un *accuracy* de 91.44%, menor en comparación con los modelos anteriores. Sin embargo, su *recall* de 92.53% es el más alto entre los modelos evaluados, lo que indica una excelente capacidad para detectar emergencias. La *precision* de 93.06% es alta, y el *F1-score* de 91.53% muestra un buen equilibrio entre *precision* y *recall*. Aunque el *accuracy* es inferior, el alto *recall* sugiere que este modelo es muy efectivo en la detección de emergencias.
- El modelo de Bosque Aleatorio también presenta un *accuracy* de 91.44%, igual al de la Red Neuronal. Su *precision* es alta (93.54%), lo que indica que el modelo es eficaz en minimizar falsas alarmas. Sin embargo, el *recall* de 89.52% es el más bajo entre los modelos evaluados, lo que sugiere una menor capacidad para identificar todas las emergencias. El *F1-score* de 91.48% refleja un balance razonable entre *precision* y *recall*, aunque no tan alto como en los otros modelos.

Análisis detallado de los mejores modelos.

Dado que los modelos de Regresión Logística y SVM muestran un rendimiento muy similar, se procede a un análisis más detallado. Cabe indicar que el presente análisis constituye para los resultados obtenidos a partir del conjunto de datos de prueba.

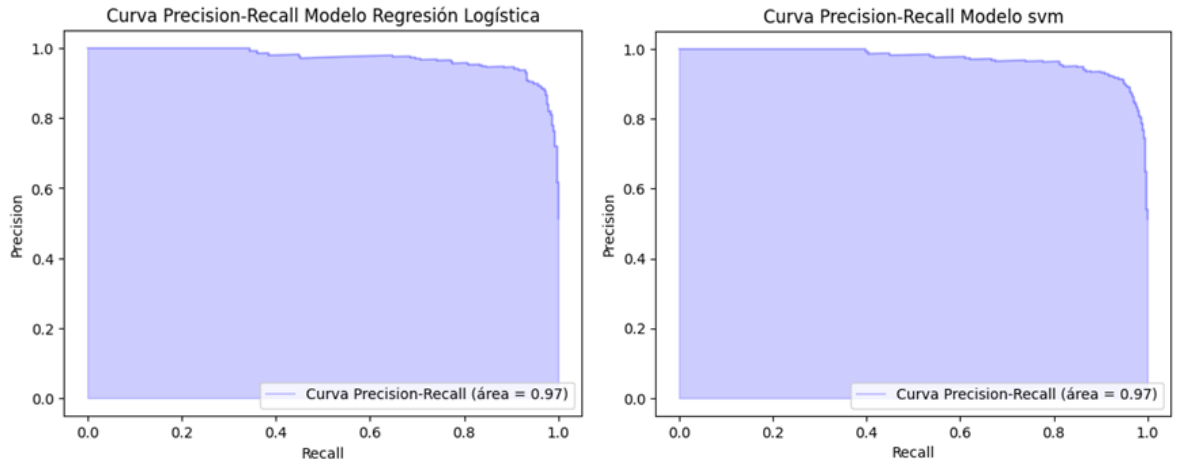

Ilustración 16. Comparativa Curva Precision-Recall

La curva Precision-Recall del modelo de Regresión Logística tiene un AUC de 0.97, lo que indica un excelente rendimiento general del modelo.

El modelo mantiene una alta precisión incluso cuando el *recall* es elevado, lo que significa que puede identificar la mayoría de los *tweets* de emergencia (alto *recall*) mientras minimiza los falsos positivos (alta *precision*). La curva es relativamente plana y se mantiene cerca del valor de precisión de 1.0 en un amplio rango de valores de recall, indicando que el modelo ofrece un desempeño consistente y confiable en la detección de emergencias.

La curva *Precision-Recall* del modelo SVM también tiene un AUC de 0.97, lo que indica un excelente rendimiento general del modelo, comparable al de la Regresión Logística. El modelo SVM mantiene una alta precisión incluso cuando el *recall* es elevado, similar al modelo de Regresión Logística.

Sin embargo, la curva del SVM es ligeramente más plana hacia el final, lo que indica un mejor manejo de los falsos positivos. La curva de SVM se mantiene cerca del valor de precisión de 1.0 en un amplio rango de valores de *recall*, indicando que este modelo también ofrece un desempeño consistente y confiable en la detección de emergencias.

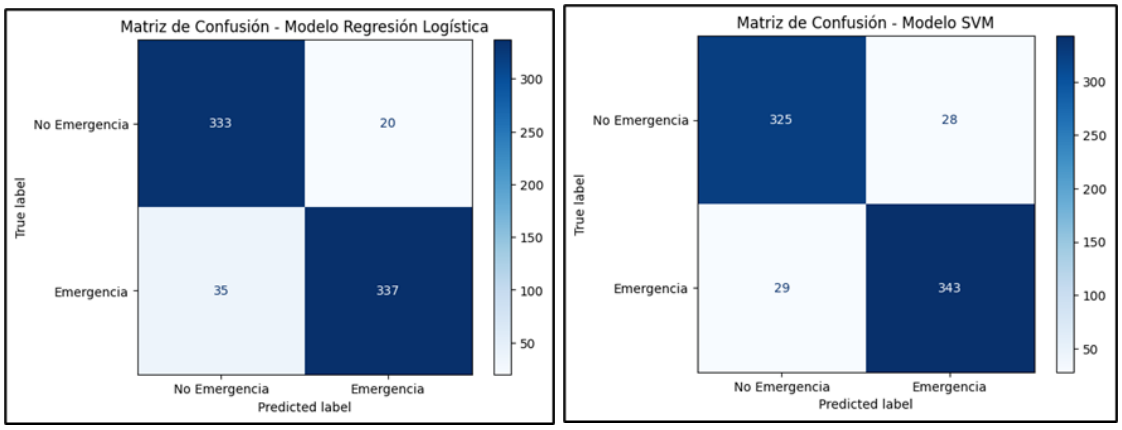

Ilustración 17. Comparativas matrices de confusión

El modelo de regresión logística presenta 333 verdaderos negativos y 337 verdaderos positivos, con 20 falsos positivos y 35 falsos negativos. Esto indica que el modelo es muy preciso, con una alta capacidad de minimizar las falsas alarmas, aunque pierde algunas emergencias (35 falsos negativos). Por su parte, el modelo SVM presenta 325 verdaderos negativos y 343 verdaderos positivos, con 28 falsos positivos y 29 falsos negativos. Este modelo tiene una mayor capacidad para detectar emergencias (343 verdaderos positivos) con menos falsos negativos (29), aunque produce más falsos positivos (28).

Ambos modelos tienen un *AUC* de 0.97, lo que indica un excelente rendimiento en general. Sin embargo, al comparar las matrices de confusión y considerando la prioridad de detectar la mayor cantidad de emergencias posibles (minimizando los falsos negativos), el modelo de SVM es superior.

8. CONCLUSIONES Y RECOMENDACIONES

8.1. CONCLUSIONES

En el presente estudio, se han analizado diversas características lingüísticas de *tweets* de emergencia y no emergencia en el contexto ecuatoriano, así como la efectividad de varios modelos de clasificación utilizando técnicas de vectorización. A continuación, se resumen las principales conclusiones derivadas de los hallazgos:

Análisis Lingüístico de los *Tweets*

Los *tweets* de emergencia tienen una longitud promedio mayor (223,89 caracteres) en comparación con los *tweets* de no emergencia (198,60 caracteres). Esto indica una tendencia a utilizar más caracteres para describir situaciones de emergencia, proporcionando información detallada y precisa. La menor desviación estándar en la longitud de los *tweets* de emergencia (72.46) sugiere una mayor consistencia en la longitud de estos mensajes.

El conjunto de datos de no emergencia presenta una mayor riqueza léxica, con 8,634 palabras únicas, en comparación con el conjunto de datos de emergencia, que tiene 5,566 palabras únicas. Esto refleja una mayor diversidad de vocabulario en los *tweets* de no emergencia, posiblemente debido a la variedad de temas tratados.

En los *tweets* de emergencia, los bigramas más frecuentes incluyen términos relacionados con incidentes específicos, como "personas heridas", "accidente de tránsito" e "incendio forestal", con porcentajes de incidencia de 1.01%, 0.91% y 0.50% respectivamente. En los *tweets* de no emergencia, los bigramas más comunes, como "cuerpo de bomberos", "varios sectores" y "fuerte lluvia", tienen una menor incidencia en el conjunto de datos, con porcentajes de 0,23%, 0,22% y 0,19% respectivamente.

Considerando la baja cantidad de palabras coloquiales, y el enfoque general de las palabras definidas previamente, se tiene como resultado que palabras como "simon" y "mierda", aparecen con más frecuencia en los *tweets* de emergencia, mientras que otras como "mierda" y "hp" son más comunes en los *tweets* de no emergencia. Esto sugiere que, aunque el lenguaje coloquial se utiliza en ambos contextos, su incidencia varía en función de la situación descrita.

La distribución de la incidencia de palabras coloquiales en ambos conjuntos de datos muestra una concentración significativa en valores bajos, lo que indica que la mayoría de los *tweets* contienen pocas terminologías coloquiales. Estas diferencias sutiles en la distribución entre los *tweets* sugieren que el modelo desarrollado puede generalizarse.

Desempeño de los Modelos

En la evaluación de modelos de clasificación, se utilizaron varias técnicas de vectorización, destacando TF-IDF por su rendimiento superior, debido a que, los modelos entrenados con TF-IDF presentaron mejores resultados en términos de equilibrio entre *precision* y *recall*, lo cual es crucial para la detección de emergencias, ya que permite minimizar tanto los falsos positivos como los falsos negativos.

Tras comparar los modelos de Regresión Logística y SVM, ambos utilizando TF-IDF, se optó por el modelo SVM debido a sus métricas ligeramente superiores y su comportamiento más robusto frente a la identificación de emergencias. Aunque ambos modelos mostraron un AUC de 0.97 en sus respectivas curvas *Precision-Recall*, el SVM presentó un *recall* superior, lo cual es esencial para identificar la mayor cantidad posible de emergencias.

8.2. RECOMENDACIONES

A pesar de los resultados satisfactorios, es recomendable aumentar la calidad y cantidad del corpus de entrenamiento, incluyendo una mayor diversidad de situaciones de emergencia. Además, la inclusión de datos de diferentes regiones del Ecuador puede ayudar a robustecer el modelo y mejorar su capacidad de generalización.

Para el presente proyecto, se empleó un etiquetado manual realizado por tres personas, lo que aseguró un alto grado de precisión en la categorización de los datos. Sin embargo, para ampliar la cantidad de datos etiquetados de manera eficiente, se recomienda considerar el uso de modelos basados en Transformers que pueden automatizar el proceso de etiquetado con gran precisión.

Aunque las características lingüísticas del español ecuatoriano no representan una barrera significativa, sería útil ajustar las técnicas de preprocesamiento para abordar variaciones locales en el lenguaje, como el uso de coloquialismos y regionalismos. Esto podría incluir un diccionario especializado que ayude a filtrar mejor las expresiones específicas del contexto ecuatoriano.

Los hallazgos sugieren que el modelo desarrollado puede ser una herramienta útil en sistemas de gestión de crisis. Se recomienda la integración del modelo con plataformas de gestión de emergencias para facilitar la identificación automática de situaciones críticas a partir de redes sociales y mejorar la eficiencia de la respuesta.

9. REFERENCIAS BIBLIOGRÁFICAS

1. Kumar, A., Singh, JP, Dwivedi, YK et al. Una red neuronal multimodal profunda para la clasificación de contenido informativo de Twitter durante emergencias. Ann Oper Res 319 , 791–822 (2022). https://doi.org/10.1007/s10479-020-03514-x
2. V. Shrivastava, S. Karsoliya, B. Verma and N. K. Gupta, "Social *Información* Analysis: Cyber Recruitment Analysis Spam Detection over Twitter *Información*set Using SVM & ARIMA Model," 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2021, pp. 1-7, doi: 10.1109/ICAECT49130.2021.9392543.
3. Madichetty, S., M., S. A stacked convolutional neural network for detecting the resource *tweets* during a disaster. Multimed Tools Appl 80, 3927–3949 (2021). https://doi.org/10.1007/s11042-020-09873-8
4. Izurieta-Brito, Daniel & Barrera, Helder & Bonilla, María. (2020). Vicios del lenguaje en los estudiantes universitarios del Ecuador. 10.47212/Tendencias2020.7.
5. Rojas Tulcanazo, V. L. (2020). Jerga en la comunicación oral de los jóvenes del sector urbano de Quito, Parroquia Cotocollao (Tesis de pregrado). Universidad Central del Ecuador. Quito, Ecuador.
6. K. Gautam, L. Misra, A. Kumar, K. Misra, S. Aggarwal and R. R. Shah, "Multimodal Analysis of Disaster *Tweets*," 2019 IEEE Fifth International Conference on Multimedia Big *Información* (BigMM), Singapore, 2019, pp. 94-103, doi: 10.1109/BigMM.2019.00-38.
7. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
8. Abubakar, U., Bashir, S. A., Abdullahi, M. B., Adebayo, O. S.(2019 Comparative Study of Various Machine Learning Algorithms for Tweet Classification,i-manager's Journal on Computer Science, 6(4),12-24. https://doi.org/10.26634/jcom.6.4.15722
9. Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In Proceedings of the 2018 Conference onEmpirical Methods in Natural Language Processing, pages 19141925.

10. Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In IJCAI.
11. María Martínez-Rojas, María del Carmen Pardo-Ferreira, Juan Carlos Rubio-Romero, Twitter as a tool for the management and analysis of emergency situations: A systematic literature review, International Journal of Information Management, Volume 43, 2018, Pages 196-208, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2018.07.008.