UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Análisis de la percepción y los sentimientos sobre el VIH e ITS en Centroamérica expresados en la red social X

Trabajo de graduación presentado por Javier Fernando Aguilar Gramajo para optar al grado académico de Licenciado en Ingeniería en Ciencia de los Datos

Guatemala,

2025

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

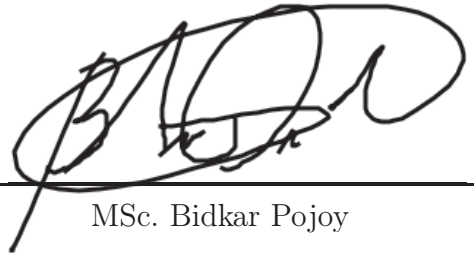
Análisis de la percepción y los sentimientos sobre el VIH e ITS en Centroamérica expresados en la red social X

Trabajo de graduación presentado por Javier Fernando Aguilar Gramajo para optar al grado académico de Licenciado en Ingeniería en Ciencia de los Datos

Guatemala,

2025

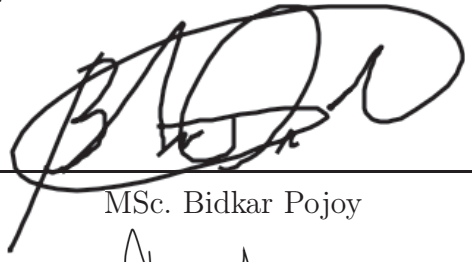
Vo.Bo.:



(f) _____

MSc. Bidkar Pojoy

Tribunal Examinador:

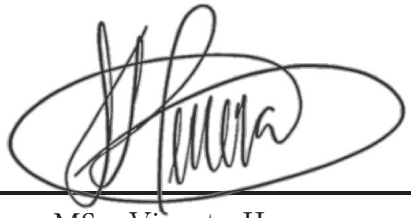


(f) _____

MSc. Bidkar Pojoy

(f) _____

MSc. José Antonio Medrano



(f) _____

MSc. Vicente Herrera

Fecha de aprobación: Guatemala, 20 de junio de 2025.

Prefacio

El presente trabajo de graduación representa la culminación de un esfuerzo académico y profesional enfocado en comprender la percepción y los sentimientos sobre el VIH e ITS en Centroamérica expresados a través de la red social X. Este proyecto ha buscado aplicar herramientas de ciencia de datos y procesamiento de lenguaje natural para generar conocimientos que puedan contribuir a mejorar las estrategias de comunicación y prevención sobre este importante tema de salud pública en la región.

Deseo expresar mi más sincero agradecimiento a todas las personas e instituciones que hicieron posible este trabajo. En primer lugar, a mi asesor MSc. Bidkar Pojoy, por su invaluable guía, dedicación y paciencia durante todo el proceso de investigación. Sus conocimientos, retroalimentación constante y apoyo académico fueron fundamentales para estructurar adecuadamente este trabajo y mantener el rigor científico necesario en cada etapa del proyecto.

Agradezco profundamente al Dr. Silvio Gramajo, cuya orientación metodológica, recomendaciones precisas y revisión minuciosa contribuyeron significativamente a mejorar la coherencia, estructura y calidad de redacción de este documento. Su experiencia y aportes críticos elevaron sustancialmente el nivel académico de esta investigación.

Un reconocimiento especial al Programa de VIH en Centroamérica (VIHCA), a cargo del Centro de Estudios de Salud de la Universidad del Valle de Guatemala (CES), por proporcionar la inspiración inicial para este proyecto y su continuo respaldo. Las reuniones informativas, retroalimentación constante y orientación temática que me brindaron fueron esenciales para contextualizar adecuadamente la investigación y alinearla con las necesidades reales en el campo.

Finalmente, expreso mi gratitud al MSc. José Antonio Medrano, director de carrera, por su apoyo administrativo y orientación durante todo el proceso para el cumplimiento de los requisitos institucionales necesarios para la culminación exitosa de este proyecto académico.

Índice

Lista de figuras

Lista de cuadros

Resumen

El presente estudio tuvo el objetivo de realizar un análisis de sentimientos y percepciones que se publican en la red social "X" sobre el tema del Virus de Inmunodeficiencia Humana (VIH) e Infecciones de Transmisión Sexual (ITS). El propósito de esta investigación fue comprender, identificar y comparar la percepción de los usuarios y cómo esta logra influir en la prevención y los tratamientos de estas enfermedades. Para lo anterior se utilizaron diferentes herramientas de Procesamiento de Lenguaje Natural (PLN), tomando como base la información recolectada en dicha red social.

A pesar de todos los esfuerzos y avances tecnológicos y médicos sobre los tratamientos y prevenciones sobre el VIH e ITS, su tratamiento se sigue viendo afectado debido a que estas enfermedades aún generan incertidumbre e inquietud en la población creando con ello preocupaciones y desafíos. Además, la falta de acceso a servicios e información sobre estas enfermedades conlleva a la desinformación y creación de estigmas.

Para alcanzar este análisis se utilizó una metodología cuantitativa que facilitó incluir las tendencias y niveles de interacción de los usuarios, así como las diferencias o similitudes significativas entre estos, ubicados en la región centroamericana. Por último, se buscaron tipos o niveles de desinformación, así como diferentes exigencias, necesidades y posibles demandas hacia actores políticos, servicios públicos y al Estado respecto a dichas enfermedades.

Tras la investigación de campo se estableció que existe una transformación significativa en el discurso público sobre el VIH e ITS en Centroamérica, caracterizada por una predominancia de contenido neutral (79.41 %) y positivo (19.16 %), con patrones comunicativos diferenciados entre países y demandas específicas hacia actores estatales centradas en acceso a servicios de salud, programas de prevención y educación sexual. Por lo tanto, se concluye que las percepciones sobre el VIH e ITS en la región han evolucionado hacia un enfoque más informativo y menos estigmatizante, aunque requieren estrategias diferenciadas según el contexto geográfico, proporcionando así insumos valiosos para desarrollar programas de divulgación y concientización más efectivos que impacten positivamente en la salud pública centroamericana.

Abstract

This study aimed to analyze sentiments and perceptions published on the social network "Xregarding Human Immunodeficiency Virus (HIV) and Sexually Transmitted Infections (STIs). The purpose of this research was to understand, identify, and compare users' perceptions and how these influence prevention and treatment of these diseases. To achieve this, various Natural Language Processing (NLP) tools were used, based on information collected from the aforementioned social network.

Despite all technological and medical advances in HIV and STI treatments and prevention, their management continues to be affected because these diseases still generate uncertainty and concern in the population, creating worries and challenges. Additionally, the lack of access to services and information about these diseases leads to misinformation and the creation of stigmas.

To conduct this analysis, a quantitative methodology was employed, which facilitated the inclusion of trends and interaction levels among users, as well as significant differences or similarities between them, located in the Central American region. Finally, types or levels of misinformation were identified, as well as different demands, needs, and possible claims directed toward political actors, public services, and the State regarding these diseases.

Following the field research, it was established that there is a significant transformation in public discourse about HIV and STIs in Central America, characterized by a predominance of neutral content (79.41 %) and positive content (19.16 %), with differentiated communication patterns between countries and specific demands toward state actors focused on access to health services, prevention programs, and sexual education. Therefore, it is concluded that perceptions about HIV and STIs in the region have evolved toward a more informative and less stigmatizing approach, although they require differentiated strategies according to the geographical context, thus providing valuable inputs to develop more effective dissemination and awareness programs that positively impact Central American public health.

CAPÍTULO 1

Introducción

Desde el surgimiento del VIH, entre otras infecciones de transmisión sexual, siempre ha existido una serie de circunstancias y contextos que rodean este tema, sus actores involucrados, así como el impacto social que generó desde mediados de los 80´s del siglo pasado. Información, propuestas, pero también miedo, incertidumbre, desconcierto y estigma son solo algunas de las narrativas que se diseminan en los medios de comunicación. En la actualidad, esto corre por otras vías comunicativas y los impactos adquieren otras dimensiones. En esta era digital, las redes sociales han logrado generar un gran impacto en los usuarios, reflejando así sus percepciones y sentimientos en diferentes temas.

Por lo anterior se realizó un estudio en Guatemala acerca de un análisis de datos sobre sentimientos y percepción de usuarios de la red social "X" respecto al tema de VIH y otras Infecciones de Transmisión Sexual (ITS) en la región centroamericana. Debido a las particularidades del tema, así como de las características propias de las redes sociales, se planteó analizar dicha percepción y los sentimientos que generan estas enfermedades debido a sus connotaciones sociales. Se utilizó como unidad de análisis las publicaciones de "X", para lo cual se requiere el uso de herramientas de Procesamiento de Lenguaje Natural (PLN) y algoritmos de aprendizaje automático, con el fin de identificar diferentes patrones de interacción de los usuarios o tendencias de percepciones las mismas en relación con un tema en común.

Este proyecto buscó ir más allá de la sola interpretación de las percepciones de los usuarios y procuró determinar sus necesidades y demandas, así como sus interacciones o respuestas a las políticas de salud establecidos por el Estado u otros actores sociales. Uno de los principales aportes de esta investigación es ofrecer un panorama sobre cómo se abordan y procesan temas sensibles para la sociedad.

La importancia de este estudio gira en torno a la capacidad de generar información relevante y luego ofrecer insumos para potencializar las estrategias para procesos de capacitación, charlas o talleres, así como elaboración de materiales de comunicación y prevención del VIH e ITS. Adicionalmente, se buscó desarrollar intervenciones efectivas y mejor dirigidas a estas comunidades que más afectada tiene la percepción del tema de estudio.

CAPÍTULO II

Antecedentes

El análisis de sentimientos en redes sociales representa un campo de investigación que ha experimentado un notable desarrollo en los últimos años. Sus orígenes pueden rastrearse hasta principios de la década de 2000, cuando los investigadores comenzaron a explorar métodos computacionales para extraer y analizar opiniones de textos en línea. Con el exponencial crecimiento de las plataformas sociales como X (antes llamada esta red social como "Twitter"), Facebook e Instagram, surgió una oportunidad sin precedentes para estudiar la expresión de opiniones, preferencias y estados emocionales en tiempo real. Este fenómeno provocó una transformación metodológica desde enfoques rudimentarios basados en diccionarios hacia técnicas más sofisticadas que incorporan aprendizaje automático y procesamiento de lenguaje natural.

La evolución de este campo ha estado marcada por avances técnicos significativos, partiendo de métodos simples de conteo de palabras con polaridad predefinida, hasta llegar a los actuales modelos de aprendizaje profundo capaces de captar sutiles matices contextuales y lingüísticos en diversos idiomas. Esta progresión refleja no solo el perfeccionamiento de las herramientas computacionales disponibles, sino también una comprensión más refinada de cómo las emociones y opiniones se manifiestan textualmente en entornos digitales, especialmente en plataformas caracterizadas por mensajes concisos como lo es X.

En el contexto académico, diversos estudios han explorado metodologías para el análisis de sentimientos en redes sociales, particularmente en X. Lovera y Cardinale (2023) Lovera y Cardinale, 2023 realizaron un estudio comparativo sobre análisis de sentimientos en Twitter, evaluando diversas técnicas y modelos. Su investigación revela que los algoritmos basados en aprendizaje profundo, específicamente los modelos Long Short-Term Memory (LSTM), alcanzan resultados superiores con una precisión del 88 % y un valor F1 de 89 %, superando significativamente a los modelos clásicos como SVM, que obtuvieron una precisión del 78 % y un valor F1 del 79 %. Este hallazgo subraya la eficacia de las arquitecturas neuronales recurrentes para captar dependencias contextuales a largo plazo en textos cortos como lo son las publicaciones de esta plataforma o tweets.

Por su parte, Tasente y Caratas (2024) Tasente y Caratas, 2024 llevaron a cabo un

exhaustivo análisis bibliométrico sobre el análisis de sentimientos en redes sociales, proporcionando una visión comprehensiva de la evolución del campo, sus principales contribuyentes y temas emergentes. Su investigación destaca el crecimiento significativo en este dominio desde 2011, con una tasa de crecimiento anual cercana al 40 %. Los autores subrayan la naturaleza colaborativa de este campo, donde las colaboraciones internacionales constituyen el 27 % de las investigaciones, con Estados Unidos, China e India como los principales países productores de conocimiento en esta área. Este estudio revela también la diversidad temática dentro del análisis de sentimientos, identificando áreas como la extracción de léxicos, comunicación política, y el impacto del boca a boca digital como temas motores en la investigación actual. Estas observaciones proporcionan un marco valioso para comprender tanto la trayectoria histórica como las tendencias futuras en este campo de estudio.

Duarte-Anselmi et al. (2022) Duarte-Anselmi et al., 2022 realizaron una investigación cualitativa para diseñar una intervención digital de prevención de ITS/VIH y promoción de salud sexual en jóvenes universitarios. A través de grupos focales con 20 estudiantes y entrevistas a 13 informantes clave, exploraron las experiencias y percepciones sobre sexualidad, riesgo y campañas preventivas. Sus resultados revelaron que la educación sexual recibida por los jóvenes era escasa, reducida a aspectos biológicos, cargada de prejuicios y sesgos de género, lo que limitaba significativamente el manejo de información. Identificaron estrategias vacilantes de prevención que no lograban motivar ni ofrecer oportunidades para la toma de decisiones conscientes en salud sexual. Las campañas de ITS/VIH fueron evaluadas como poco inclusivas, lejanas y basadas en el miedo. Los autores concluyeron que las intervenciones en salud sexual han fallado en considerar los aspectos experienciales de la sexualidad juvenil, basándose en modelos de comportamiento ideal y estereotipado, y propusieron la necesidad de innovar con intervenciones fundamentadas en diseños integradores, multidisciplinarios y situados que valoren tanto la teoría como la experiencia de las poblaciones objetivo.

Mejía et al. (2020) Mejia et al., 2020 realizaron un estudio transversal analítico y multicéntrico para caracterizar la percepción de miedo o exageración que transmitieron los medios de comunicación durante la pandemia de COVID-19 en Perú. A través de una encuesta virtual aplicada a 4,009 personas en 17 ciudades peruanas, evaluaron tres factores: la exageración mediática, el miedo generado y la comunicación proveniente del personal de salud, familiares y amigos. Sus resultados mostraron que los participantes percibieron que las redes sociales (64 %) y la televisión (57 %) exageraban la información, mientras que la televisión (43 %) y las redes sociales (41 %) eran los principales medios que aumentaban la percepción del miedo. El análisis multivariado reveló que las mujeres y las personas con mayor nivel educativo tenían un menor puntaje total de miedo y percepción de exageración. Los autores concluyeron que la percepción de exageración y generación de miedo en la población fueron ocasionados principalmente por la televisión y las redes sociales, hallazgos relevantes para entender cómo los medios de comunicación influyen en la percepción de crisis sanitarias.

Arnao y Ramírez (2022) CIRUJANO et al., s.f. realizaron un estudio observacional descriptivo de corte transversal con el objetivo de conocer la percepción y el nivel de información sobre las personas viviendo con VIH en la población adulta con acceso a redes sociales en Lima, Perú. A través de una encuesta virtual aplicada a 623 personas, entre septiembre y noviembre de 2021, los investigadores evaluaron tanto el conocimiento sobre la enfermedad como las actitudes hacia quienes viven con el virus. Sus resultados mostraron que, aunque el 81.38 % de los encuestados identificaba correctamente que VIH y SIDA tienen definiciones

diferentes, existían importantes brechas de conocimiento, pues el 55.8 % desconocía que una persona con carga viral indetectable no transmite el virus. Asimismo, evidenciaron actitudes discriminatorias persistentes: mientras que el 95.67 % manifestó que podría ser amigo de una persona con VIH, solo el 32.26 % consideraría tener una relación de pareja con alguien con esta condición. Los autores concluyeron que la falta de conocimiento en conceptos asociados al VIH resalta la importancia de crear políticas de concientización y campañas informativas para eliminar la discriminación hacia las personas que viven con el virus.

Restrepo (2016) Restrepo-Pineda, 2016 realizó un estudio transnacional comparativo entre Colombia y España para analizar las diferencias en las percepciones sobre el VIH/SIDA de varones homosexuales y bisexuales colombianos con y sin experiencia migratoria. A través de 87 entrevistas en profundidad realizadas entre 2011 y 2013 a varones residentes en diferentes comunidades autónomas españolas (Madrid, Cataluña, Valencia y Andalucía) y departamentos colombianos (Caldas, Quindío, Risaralda y Valle del Cauca), el investigador indagó sobre cómo estas percepciones influyen en la vulnerabilidad social determinada por el desconocimiento de la diversidad cultural y sexual de las personas inmersas en procesos migratorios. El autor concluyó que la relación entre migración y sexualidad debe abordarse desde una visión integral que enriquezca la comprensión tanto en la sociedad de origen como en el país de acogida, considerando aspectos sociales y culturales. Además, enfatizó que los programas de promoción y prevención en salud deben considerar las especificidades de las personas para evitar generalizaciones e instrumentalización, reconociéndolas como sujetos de pleno derecho que opinan, hablan y participan.

Reyes y Vargas (2022) Reyes Lorzo y Vargas Mendoza, s.f. realizaron una investigación cualitativa mediante un estudio descriptivo-interpretativo para analizar la intervención del trabajador social y la importancia de las redes sociales de apoyo en usuarios diagnosticados con VIH/SIDA que acuden al Centro Ambulatorio para la Prevención y Atención en SIDA e Infecciones de Transmisión Sexual (CAPASITS) en el Estado de México. A través de la aplicación del cuestionario historia de vida y el cuestionario MOS de apoyo social a una muestra de ocho usuarios de diferentes municipios, los investigadores encontraron que los participantes se encontraban en la etapa de juventud-adultez (20-59 años), con diferentes vías de contagio y distintas etapas de la enfermedad. El estudio reveló que siete usuarios contaban con un índice global medio de apoyo social (33-60 %), mientras que uno presentaba un índice mínimo (27 %). Los autores identificaron que la intervención profesional del trabajador social presentaba funciones rutinarias y poca vinculación entre la práctica y el conocimiento científico, concluyendo sobre la importancia de fortalecer la intervención profesional y el trabajo multidisciplinario para brindar un tratamiento integral, además de consolidar el apoyo de las redes sociales primarias con el propósito de mejorar la calidad de vida e integración social de los usuarios.

En el ámbito de la salud pública, Prieto, Gómez y Borges (2017) Vasquez et al., 2017 realizaron una investigación sobre el uso e importancia de Twitter para la prevención en salud, centrándose en comunidades hispanohablantes. A través de un análisis de contenido cuantitativo de 3000 mensajes con el hashtag #prevención, los autores encontraron que la mayoría de los mensajes relacionados con prevención en salud eran unidireccionales, con escasa promoción de comunicación y movilización por parte de los usuarios. Sus resultados mostraron que los principales emisores de estos contenidos eran medios de comunicación y agencias gubernamentales, quienes se enfocaban principalmente en temas de salud pública

y seguridad vial. Esta investigación destaca las potencialidades de las redes sociales, específicamente "X", como herramientas para la comunicación y prevención en salud, ofreciendo hallazgos útiles para el diseño e implementación de campañas preventivas en estas plataformas digitales, especialmente dirigidas a comunidades de habla hispana donde este tipo de estudios son escasos.

Un aspecto importante de este tipo de investigación es el enfoque metodológico aplicado. Los estudios contemplan tanto métodos lexicales como modelos de aprendizaje automático y enfoques híbridos. Los métodos lexicales se basan en diccionarios predefinidos con palabras anotadas según su polaridad sentimental, calculando el sentimiento general mediante el conteo de palabras con carga emocional. Aunque estos métodos son interpretables, suelen tener dificultades con sentimientos específicos del contexto y la detección de sarcasmo.

La aplicación de estas metodologías abarca diversos campos de estudio, ofreciendo valiosas perspectivas sobre el sentimiento público. Investigaciones recientes han demostrado su utilidad en análisis de sentimientos sobre la salud pública como lo pueden ser recientemente las vacunas del COVID-19, en la comunicación durante crisis turísticas, y percepciones hacia plataformas de comercio electrónico, revelando tendencias y patrones que pueden informar estrategias de comunicación y marketing.

La revisión de estas investigaciones previas proporciona un marco metodológico y conceptual sólido para el presente estudio sobre la percepción y sentimientos relacionados con el VIH en redes sociales en Centroamérica. Los avances en modelos computacionales, técnicas de preprocesamiento y enfoques analíticos documentados ofrecen una base robusta para desarrollar una metodología de investigación rigurosa y adaptada a las particularidades culturales y lingüísticas de la región centroamericana, así como a las características específicas de la comunicación digital sobre temas de salud pública como el VIH e ITS.

CAPÍTULO III

Justificación

En la actualidad, las redes sociales tienen un gran impacto y desempeñan un papel fundamental en la comunicación alrededor del mundo; la red social X, antes conocida como Twitter, por ejemplo, permite que, de manera pública, los usuarios publiquen sus emociones, sentimientos y opiniones y puedan interactuar con los mensajes y conversaciones de los demás usuarios. Sin embargo, a nivel mundial se ha visto dentro de esta una creciente y preocupante desinformación y falta de información sobre diversos temas tanto sociales, médicos, políticos.

Dentro de la salud pública a nivel global se encuentra el VIH que afecta a millones de personas en el mundo y Centroamérica no es la excepción. A pesar de los tratamientos y avances tanto tecnológicos como en el área de la salud, siguen persistiendo las preocupaciones y los desafíos que engloba toda la enfermedad del VIH, lo cual incluye la percepción y falta de acceso a servicios e información de calidad.

La selección de este tema como trabajo de graduación quiere responder ante la necesidad inherente de lograr percibir y comprender la gran influencia de una plataforma de redes sociales como X. Por consiguiente, esto puede llegar a generar un impacto en el tratamiento y prevención de la enfermedad del VIH y otras ITS, ya que dentro de las redes sociales se manejan opiniones y percepciones además de noticias, ya sean verídicas o falsas. Sin embargo, esto último puede a dificultar la correcta extracción y creación de percepciones y sentimientos precisos para un tema tan sensible como lo es una enfermedad que afecta a millones de personas alrededor del mundo. La desinformación y la falta de información son fenómenos que ayudan a perpetuar y crear falsos estigmas y mitos sobre las mismas, creando desafíos en el área de la salud pública y en sus intervenciones educativas.

Por ello se generó información pertinente que pueda ser útil a autoridades u otros actores interesados a generar procesos de capacitación y prevención de estas enfermedades e infecciones en la región de Centroamérica, tales como el programa VIHCA del Centro de Estudios en Salud (CES) de la Universidad del Valle de Guatemala (UVG). Creando con esta información campañas estratégicas de educación y comunicación que logren contrarrestar de la manera más optima la desinformación. Del mismo modo, se pueden generar campañas de comunicación que ayuden a mejorar el bienestar y salud de las personas afectadas por

distintas enfermedades de transmisión sexual en cada uno de los países de Centroamérica. Finalmente, con este análisis se espera que logre ser una herramienta útil y valiosa para combatir la falta de información y la desinformación promoviendo una educación y una perspectiva adecuada sobre el VIH.

CAPÍTULO IV

Objetivos

A. Objetivo general

Comprender, por medio de análisis de patrones de interacciones sociales y de comunicación, la percepción y los sentimientos de usuarios de la red social X de Centroamérica, respecto a procesos de prevención, experiencias personales, tratamientos, disponibilidad de servicios en recursos y ubicaciones físicas, sobre el VIH y otras Infecciones de transmisión sexual (ITS), así como sus efectos secundarios sobre cualquier tratamiento médico. Con ello se pretende contribuir a la mejora de los programas de divulgación y concientización que permitirán la creación de estrategias más efectivas y precisas para sensibilizar a la población, mejoras en la efectividad y acercamiento de los programas destinados a abordar estos temas sensibles y cruciales para la salud pública.

B. Objetivos específicos

- Realizar análisis de patrones de texto y de percepción de sentimiento para de establecer patrones, tendencias y niveles de interacción de los usuarios dentro y fuera de su país de origen (Centroamérica), sobre los tópicos de VIH e Infecciones de Transmisión Sexual (ITS) mediante diferentes métodos y algoritmos de aprendizaje automático.

- Comparar dichos patrones, tendencias y niveles de interacción identificados de los usuarios, buscando similitudes o diferencias en su comportamiento y de necesidades de información, mediante distintas visualizaciones y métricas sobre los resultados de los algoritmos.

- Identificar posibles acciones de demandas al Estado como a otros actores por parte de los usuarios, lo cual puede surgir de un análisis de los resultados más relevantes examinados del estudio.

- Identificar y comprar similitudes y diferencias entre las diferentes regiones de Centroamérica, sobre las preocupaciones, actitudes y niveles de conocimiento y desinformación, mediante un análisis de las publicaciones geolocalizados, aplicando diferentes modelos específicos de clasificación de texto para cada región.

CAPÍTULO V

Alcance y limitaciones

A. Alcance

Para efectos de esta investigación se recolectaron y extrajeron los datos de la referida red social durante el periodo de julio a diciembre del 2024. Además, se centró la atención en los patrones de texto como resultado del análisis del contenido de las publicaciones y mensajes de las personas, logrando así establecer y determinar tendencias, patrones y niveles de interacción de los usuarios de "X". El estudio incluyó las comparaciones de estos patrones dentro de los diferentes países de la región centroamericana, lo que permitió identificar diferencias y similitudes significativas en el comportamiento y necesidades sobre la falta de información de los usuarios. Adicionalmente, sé determinó e identificó una serie de diferentes factores o actores políticos y posibles acciones de demanda de recursos de los usuarios hacia el Estado con respecto al tema de estas enfermedades.

B. Limitaciones

Para el desarrollo del presente estudio se enfrentó una limitación: del universo de redes sociales, únicamente se utilizó la red social "X", antes conocida como Twitter, lo cual representa solamente utilizar una proción de la población que tiene acceso a dicha red; sin embargo, se amplió a otros territorios fuera de Guatemala, tal es el caso de la región centroamericana. Otro factor que representó un límite es que a pesar del análisis exhaustivo que se realizó, el análisis de sentimientos y percepción de un sector en específico está sujeto a ambigüedad y errores dado al lenguaje, expresiones y variabilidad utilizada en cada uno de los lugares del estudio.

Otro punto a señalar está en los cambios de emociones que pueden a ver a lo largo de tiempo en las personas dado que se estudió se realizó en periodos muy cortos. Por lo que, la naturaleza dinámica de las emociones y percepciones de los usuarios presentó un desafío significativo durante el estudio, ya que a que estos elementos pudieron evolucionar más allá del

alcance temporal de la investigación. Además, al limitarse geográficamente a Centroamérica, las generalizaciones hacia otras regiones resultaron limitadas en su aplicabilidad.

Además, la identificación de las percepciones de los usuarios y sus demandas hacia el Estado y otros actores políticos estuvo sujeta a posibles malentendidos y sesgos en los datos recolectados de la red social.

Una de las limitaciones más significativas encontradas durante el desarrollo de este trabajo fue la política de monetización implementada por la red social "X"para el acceso a su API. La versión básica de la API, que fue la única económicamente viable para este estudio, presentó restricciones sustanciales en la recopilación de datos. Esta versión únicamente permitió la extracción de publicaciones con una antigüedad máxima de siete días a partir de la fecha de consulta, además de imponer limitaciones en la frecuencia y cantidad en la extracción de datos.

La situación se agravó durante el periodo de recolección, debido a que la red social duplicó el costo de la suscripción básica, lo que impactó significativamente el presupuesto destinado para este estudio. A pesar de estas restricciones, se realizó un esfuerzo económico considerable que permitió obtener datos durante aproximadamente seis meses. Es importante mencionar que esta limitación fue particularmente relevante considerando la naturaleza específica del tema de estudio, ya que no era factible encontrar datos unicamente referidos al VIH, por lo cual se amplió el patrón de búsqueda a enferemedades relacionadas con esta dolencia, especialmente las de transmisión sexual.

La alternativa para superar estas restricciones hubiera sido la suscripción al plan empresarial de la API, el cual ofrecía capacidades más robustas para la extracción de datos históricos y un mayor volumen de consultas, sin embargo, el costo asciende a US$5,000 mensuales, lo cual está fuera del alcance para realizar la investigación.

Para finalizar, también hay que tomar en cuenta las limitantes tecnológicas y componentes propias de las herramientas para hacer dichos análisis, así como las limitaciones inherentes sobre las técnicas y herramientas de PLN, como lo puede llegar a ser la precisión y la validez.

CAPÍTULO VI

Marco teórico

Para el desarrollo de este proyecto se establecieron como variables fundamentales de estudio: los sentimientos, la percepción, VIH y enfermedades de transmisión sexual (ITS), Tecnología - análisis de contenido, ética y privacidad. Esta estructura teórica permitió situar el problema de este trabajo dentro del conocimiento existente y proporciona las bases conceptuales para el análisis de datos en redes sociales aplicado a un tema de salud pública.

A. Sentimientos y percepción

Los sentimientos se derivan de las emociones que hacen referencia a una amplia gama de estados internos que experimentamos como lo son la alegría, tristeza, ira, amor, entre otros. Asimismo, es una respuesta emocional a través de nuestras vivencias y experiencias. Del mismo modo, los sentimientos pueden determinar el estado de ánimo o la disposición del usuario o persona de estudio (Porto, J. P., & Gardey, A., 2010).

La percepción de un individuo es la forma que establece e interpreta una situación o sensación que obtiene y recibe por medio de los cinco sentidos del cuerpo, con eso lograr comprender las señales que vienen desde terceros o del exterior formando así una impresión inconsciente o consiente de lo que está pasando a su alrededor (Editorial, E., & Etecé, 2014).

Además, ambas variables de estudio pueden estar fuertemente relacionadas, existen algunas diferencias significativas entre estos términos. En donde principalmente los sentimientos son los estímulos recibidos que un individuo logra recibir de fuentes externas o internas. Por otro lado, la percepción llega a ser un procesamiento e interpretación de la información sensorial para lograr una compresión de alguna situación o del entorno del individuo (Porto, J. P., & Gardey, A., 2010).

B. VIH e Infecciones de Transmisión Sexual (ITS)

En todo el mundo existen muchas enfermedades e infecciones de transmisión sexual (ITS) de la cuales estas enfermedades presentan grandes desafíos diariamente en la salud pública global y de la misma manera las mismas presentan una enorme resistencia a los mejores esfuerzos para de toda la medicina en la actualidad para lograr combatirlas (ONUSIDA).

Dentro de estas ITS está el VIH (Virus de Inmunodeficiencia Humana) el cual su propósito es atacar y debilitar el sistema inmunológico, lo que permite a demás a otras enfermedades entrar y desarrollarse de manera más eficiente. Asimismo, si no se llega a tratar a tiempo, esto conlleva a la producción y desarrollo del SIDA (Síndrome de Inmunodeficiencia Adquirida). De la misma manera, es importante resaltar que las ITS son infecciones que se transmiten principalmente por el contacto sexual, como lo pueden ser bacterias, parásitos y hongos (CDC, 2022).

1. Prevención

Para combatir y luchar contra el VIH son esenciales los métodos de barreras para la reducción del riesgo de infección y lograr frenar la propagación y el contagio de este (ONUSIDA). Asimismo, estos métodos han logrado demostrar ser una muy efectiva forma de barrera contra estas ITS, no solo por el VIH y del mismo modo funcionar como una protección hacia el mismo (CDC, 2022). Un ejemplo de estos métodos son los preservativos masculinos y femeninos o los medicamentos PrEP y PEP (MedlinePlus).

2. Tratamientos

Desde que se descubrieron estas ITS dentro del área de la medicina y salud pública mundial se han llevado a cabo nuevas formas y medicamentos antirretrovíricos especialmente contra el VIH. Específicamente para las personas que tienen el virus del VIH el tratamiento estándar de este es una combinación de al menos tres antirretrovíricos distintos (PubMed).

3. PrEP (Profilaxis Pre-Exposición) y PEP (Profilaxis Post-Exposición)

La PrEP como su nombre lo indica es una profilaxis, es decir, es la acción o la búsqueda preventiva o de la propagación de enfermedades infectocontagiosas como lo puede ser el VIH. Por lo que llega a ser un medicamento de prevención para las personas que aún no tienen el virus, pero pueden estar en situaciones de alto riesgo de contraer el mismo, como lo pueden ser poblaciones clave tal como: trabajadoras sexuales, personas transgénero, entre otras personas. Este medicamento puede llegar a ser muy efectivo si se toma consciente y responsablemente (CDC, 2022).

De la misma manera, el PEP es un medicamento preventivo, pero en este caso este medicamento tiene personas objetivo diferente, las cuales estas mismas es posible o tienen una alta probabilidad de haber estado expuestas al virus del VIH. Para un funcionamiento

eficiente del medicamento se debe dar inicio de este dentro de las primeras 72 horas después de la posible exposición al virus del VIH (MedlinePlus).

4. Efectos secundarios

Conforme al paso del tiempo y la evolución de la tecnología y la medicina, actualmente los tratamientos ya no son tan tóxicos como eran el inicio de los tratamientos contra el virus, por el momento los medicamentos que pueden llegar a tener algunos efectos secundarios en algunas personas pueden ser los preventivos como lo son el PrEP y PEP que pueden llegar a causar nauseas, mareos o dolores de cabeza (Crook Th, Ferris Sh, Alvarez Xa, Laredo, M., & Moessler, H., 2005).

5. Servicios disponibles

En muchos países se cuenta con estos servicios de asistencia y asesoramiento a las personas sobre estas enfermedades, en las cuales estos mismos pueden llegar a incluir tratamientos, apoyo psicológico, pruebas de VIH y asesoramientos. Todos estos servicios están disponibles para todas aquellas personas que viven, conviven o están con un riesgo alto de contraerlo (AIDSinfo | UNAIDS).

C. Tecnología y análisis de contenido

En el contexto de este proyecto sobre el VIH en Centroamérica, los avances tecnológicos en esta era digital como lo son las herramientas del análisis y la recopilación de datos son vitales para llegar a comprender, por ejemplo, sobre la percepción del público objetivo respecto a este tema. De tal manera, que el análisis de contenido percibe y nos puede llegar a proporcionar una valiosa y amplia visión sobre esta enfermedad que es el VIH en distintas regiones.

1. Análisis de sentimientos en redes sociales

El análisis de sentimientos en redes sociales constituye un proceso sistemático de obtención y análisis de datos provenientes de plataformas de interacción social. Este estudio se centró en específico en la red social “X”, antes llamada “Twitter”, donde se analizó la información recopilada de cada uno de los mensajes, promts o publicaciones, interpretando así lo que las personas dicen, piensan o creen de cualquier tema en discusión. Por lo que, dentro de este análisis lo que permitió examinar son las opiniones y emociones transmitidas en los mensajes publicados por las mismas personas dentro de las discusiones del tema.

2. Estrategias y métricas

Dentro de las diferentes estrategias que existen en el análisis de sentimientos de una red social se puede encontrar el análisis de todos los parámetros de interacción que incluye la publicación de un mensaje en una red social, como lo puede ser el estudio o el análisis de los números de me gusta, compartidos, comentarios, visitas, entre otras métricas tradicionales de la publicación. De la misma manera, puede incluir en seguimiento de un usuario a una cuata o marca, dado que estos análisis es un estudio más avanzado y profundo, debido a que se centra en la calidad de la interacción del público o de las personas hacia un tema u organización en específico.

3. Recopilación y preprocesamientos de datos

El preprocesamiento de datos constituyó uno de los pasos fundamentales para garantizar la efectividad y precisión del análisis. Este proceso comprendió la extracción sistemática de datos de la red social "X", seguida de una fase de limpieza y preparación para el análisis posterior. La limpieza de datos incluyó la eliminación de palabras vacías (stop words), la detección y normalización del idioma, y la estandarización del contenido textual.

4. Procesamiento de Lenguaje Natural (PLN)

El Procesamiento de Lenguaje Natural (PLN) llega a ser el intermediario y encargado de la comunicación entre la maquina y la persona, dado que es una rama de Machine Learning o de la Inteligencia artificial, esto mediante el uso de las lenguas naturales (IBM,2023). Además, para de este proyecto se incluyeron algoritmos y técnicas aplicados logrando así analizar los sentimientos y la percepción expresada en los usuarios en la red social “X”.

Adicionalmente, el análisis de sentimientos automatizado utilizó técnicas de PLN y aprendizaje automático, basándose en reglas predefinidas, aprendizaje supervisado y no supervisado. El proceso, también, incluyó la creación de conjuntos de datos de entrenamiento para desarrollar modelos capaces de clasificar automáticamente los sentimientos y percepciones del contenido analizado.

Asimismo, el análisis que se realizó pasó por un modelo de evaluación sobre la precisión de los modelos aplicados, dado que esto es vital para el progreso y avanece del algoritmo, puesto que con ello se buscó garantizar e identificar si el modelo presenta algún punto de mejora o si está funcionando de manera eficiente y correcta. Por lo que, en esta evaluación se llegó a incluir las siguientes métricas: la precisión, el Recall, la validación cruzada y la F1-score (IBM,2023).

5. Perspectiva

Dentro de las diferentes perspectivas o enfoques que tiene el PLN se encuentra amplios modelos estadísticos, así como el aprendizaje automático o un aprendizaje profundo, esto

basándose en elementos de la lingüística computacional, debido a que el mismo lo que trata es de modelar el lenguaje de las personas basándose principalmente en reglas dadas.

D. Herramientas y metodologías

Igualmente, existen diversas metodologías y herramientas para emplear y aplicar un PLN, como lo puede ser un análisis sintético, análisis semántico o un morfológico (léxico). Las cuales permiten entender, interpretar y generar a las maquinas el lenguaje humano. Dentro de las herramientas se utilizaron para efectos de este proyecto fue el análisis de sentimientos basado en aspectos (ABSA) el cual permitió relacionar opciones especificas e identificarlas, con aspectos particulares de un servicio o producto. Otra herramienta que se utilizó fue el procesamiento de lenguaje natural profundo (Deep PLN), el que se enfoca en analizar y entender de mejor manera el contexto y la ironía, elementos que se pierden principalmente en métodos más simples, esto basándose en redes neuronales y aprendizaje profundo.

Asimismo, la implementación metodológica de dichas herramientas se requiere una estructura definida. En donde el proceso inicia con una exhaustiva extracción y recopilación de los datos, la cual por consiguiente se le debe de aplicar una limpieza y un preprocesamiento correcto, eliminando así exceso de ruido y datos irrelevantes para el estudio. Esto se logró con una herramienta de la red social exclusiva para este proceso que es la API de la plataforma, en donde una API (Interfaz de Programación de Aplicaciones) consiste en diferentes mecanismos o conjunto de protocolos que se utiliza y permite el desarrollo e integración de dos componentes o aplicaciones de software interactuar entre sí según sus requerimientos u objetivos.

Seguidamente, los datos extraídos de la red social se almacenan en una base de datos más robusta para su posterior análisis y consultas. Para su manejo y gestión se utilizan diferentes herramientas como lo pueden ser: Dbeaver y Datagrip, la cuales proporcionan una interfaz gráfica que permite la administración de múltiples sistemas de gestión de bases de datos. Esta misma herramienta es ampliamente utilizada en el ámbito profesional, debido a su amplia capacidad de soportar una variedad de conexiones y motores de datos como lo pueden ser: MySQL, PostgreSQL, MariaDB, SQLite, Oracle, Microsoft SQL Server, entre otros. (DBeaver Community, s.f.)

Para la realización de este estudio se seleccionó el motor de datos PostgreSQL, ya que lleva un desarrollo activo hace más de 35 años, teniendo así un sistema potente, robusto y eficiente de bases de datos relacionales de código abierto, permitiendo así la gestión de grandes volúmenes de datos. Además, utiliza SQL como su lenguaje de manejo de consultas, lo que facilitará la interacción y consulta de los datos almacenados dentro de la misma. Esta elección aseguró la administración fiable y eficiente de la información necesaria y recopilada para este estudio. (Moraguez, E. R., 2023).

Luego, se debe seleccionar las combinaciones de modelos estadísticos o de aprendizaje más adecuados para el conjunto de datos extraído y preprocesado anteriormente. Dentro del preprocesamiento a realizar se encuentra el modelo de "Bolsa de palabras" o más conocida como por su traducción al inglés como "Bag-of-words", en donde el objetivo es poder crear

vectores de frecuencias de las palabras a analizar, lo que permite hacer una eliminación de ruido con eso una limpieza más efectiva y a mayor profundidad de los conjuntos de palabras y crear representaciones vectoriales que pueden ser procesadas eficientemente por algoritmos de aprendizaje automático. La representación vectorial resultante, aunque pierde información sobre el orden de las palabras, ha demostrado ser altamente efectiva para tareas de clasificación de texto, permitiendo capturar la esencia semántica del contenido analizado.

Además, se emplearon dos metodologías complementarias para el análisis específico de sentimientos las cuales fueron: TextBlob y VADER (Valence Aware Dictionary and sEntiment Reasoner). TextBlob se fundamenta en un enfoque léxico que utiliza diccionarios predefinidos y reglas lingüísticas para asignar puntuaciones de polaridad a fragmentos de texto. Este método evalúa cada palabra y frase dentro de su contexto, asignando valores que indican si el contenido es positivo, negativo o neutral. Por su parte, VADER representa un avance significativo en el análisis de sentimientos para redes sociales, debido a que está específicamente optimizado para manejar las particularidades del lenguaje en estas plataformas. Su arquitectura incorpora reglas heurísticas que consideran elementos como emojis, modismos, refuerzos léxicos y puntuación, proporcionando un análisis más matizado de la intensidad emocional en el texto. La combinación de ambas herramientas permitió obtener una evaluación más robusta y multidimensional de los sentimientos expresados en las publicaciones analizadas. Complementariamente, se aplicó la metodología TF-IDF (Term Frequency-Inverse Document Frequency), una técnica estadística que evalúa la importancia relativa de las palabras en un documento dentro de una colección de textos. Esta metodología asigna un mayor peso a los términos que aparecen frecuentemente en un documento específico pero que son relativamente distintos en el conjunto total, permitiendo identificar palabras distintivas y relevantes en cada contexto regional. En este trabajo, TF-IDF resulta fundamental para discernir las diferencias en el vocabulario empleado al hablar sobre este caso de estudio entre los diferentes países centroamericanos, facilitando la identificación de preocupaciones específicas, necesidades de información y particularidades culturales en cada región. Esta técnica permite establecer similitudes y diferencias significativas en las percepciones y discursos en las diferentes publicaciones dentro de la plataforma o red social de estudio.

En cuanto a los modelos de clasificación empleados, se seleccionaron tres algoritmos principales: Random Forest, Support Vector Machine (SVM) y Naive Bayes. Random Forest, como algoritmo de conjunto, construye múltiples árboles de decisión y combina sus predicciones mediante un sistema de votación. Su arquitectura se basa en el principio de "sabiduría de la multitud", donde cada árbol se entrena con un subconjunto aleatorio de características y observaciones, lo que resulta en una reducción significativa del sobreajuste y una mayor capacidad de generalización. Support Vector Machine, por su parte, opera bajo el principio de ampliación del margen, buscando la superficie de separación ideal que separe las diferentes clases en el espacio de características. Su efectividad en espacios de alta dimensionalidad lo hace particularmente adecuado para el análisis de texto, donde la cantidad de características (palabras) puede ser muy grande. Naive Bayes, basado en el teorema de Bayes, asume la independencia condicional entre características, una simplificación que, aunque teóricamente limitante, ha demostrado ser sorprendentemente efectiva en la práctica para la clasificación de texto.

Finalmente, la evaluación del rendimiento de estos modelos se realizó mediante un con-

junto comprehensivo de métricas estadísticas. La precisión (precision) mide la proporción de predicciones positivas correctas entre todas las predicciones positivas realizadas, siendo crucial cuando el costo de los falsos positivos es significativo. La exhaustividad (recall) evalúa la capacidad del modelo para identificar correctamente todos los casos positivos reales, siendo especialmente relevante cuando es importante no perder ningún caso positivo. El F1-Score, como media armónica entre precisión y el recall, proporciona una métrica única que equilibra ambos aspectos del rendimiento. La matriz de confusión ofrece una visualización detallada de la distribución de predicciones correctas e incorrectas para cada clase, permitiendo un análisis granular del comportamiento del modelo. La validación cruzada, implementada mediante la técnica de k-fold, proporciona una estimación robusta del rendimiento generalizado del modelo al evaluarlo sobre diferentes subconjuntos de datos. La (Cross-validation), es decir, una validación cruzada, es una herrmienta que permite poder medir el rendimiento de un modelo predictivo. Por lo que, las distribuciones de scores CV los resultantes permiten analizar la estabilidad y consistencia del modelo a través de diferentes particiones de datos, ofreciendo insights sobre su capacidad de generalización y robustez. Con esto poder completar el análisis con diferentes tipos de gráficos para mejorar la interpretación de dichos resultados de los modelos, permitiendo así identificar patrones y tendencias dentro de los hallazgos de los modelos.

E. Análisis e investigaciones anteriores

En relación con los trabajos y proyectos realizados sobre estos temas como lo es el análisis de percepciones y sentimientos, ha transitado desde enfoques predominantemente cualitativos hacia métodos mixtos y cuantitativos potenciados por la tecnología. Esta evolución refleja no solo cambios en las herramientas disponibles, sino también transformaciones en la conceptualización misma de cómo se construyen, expresan y pueden medirse los sentimientos y percepciones sociales.

CAPÍTULO VII

Metodología

El presente estudio se realizó bajo un enfoque de metodología cuantitativa, consistente en técnicas científicas que tuvieron como objetivo recopilar datos en forma numérica, generalmente utilizados en la investigación. Este enfoque implicó estudiar el tema en función de sus características cuantificables, es decir, aquellas que pudieron describirse numéricamente. La metodología cuantitativa permitió colocar datos numéricos a los sentimientos y emociones representados en los tweets, generando así números que hablaron por sí solos. Los resultados fueron objetivos, numéricos, descriptivos y, en ocasiones, incluso predictivos, lo que permitió prever tendencias o patrones sin ningún sesgo subjetivo.

Para la recolección de datos se utilizó la API oficial para desarrolladores de la red social "X"(anteriormente Twitter), implementando la biblioteca Tweepy en Python. Esta herramienta permitió la interacción y extracción del contenido publicado en toda la región de Centroamérica. Se implementó un sistema automatizado de recolección que filtró tweets relacionados con VIH e ITS, utilizando una lista exhaustiva de palabras clave y frases específicas relacionadas con VIH, SIDA, terminología médica asociada, tratamientos, medicamentos, prevención, educación sexual y hashtags relevantes a la temática.

Posteriormente, para la gestión y almacenamiento de los datos y publicaciones extraídas se utilizó el motor de bases de datos PostgreSQL, empleando DataGrip como herramienta de manejador, visualización y gestión de los datos. Esta herramienta permitió un uso más efectivo y visual del motor de datos, dado su interfaz gráfica, potencia y capacidad de consulta mediante el lenguaje SQL. Su compatibilidad con múltiples sistemas de gestión de base de datos la convirtió en una herramienta ideal para la interacción con los datos recopilados y el posterior análisis del estudio.

Durante la manipulación de la información en la red social emergió el ámbito ético del manejo de datos, por ello se implementó un riguroso proceso de anonimización durante la extracción de la información. Este proceso garantizó la eliminación o modificación de

cualquier detalle que pudiera identificar a un usuario individual, incluyendo nombres de usuario, ubicaciones y otra información personal. Este espectro de anonimización se mantuvo incluso durante la presentación de los hallazgos, asegurando que ningún dato publicado pudiera ser rastreado hasta el usuario original. Por lo que, al haber realizado el trabajo de esta manera, se afirma el compromiso de defender la ética, garantizando que, si bien se respetó los derechos y la privacidad de las personas, todavía fue posible extraer conclusiones importantes sobre cómo el público percibe lo relacionado con enfermedades de transmisión sexual. Esto no solo tiene relación con un componente ético, sino también con un componente en materia de derechos humanos, en donde se debe garantizar el derecho de proteger los datos personales sensibles de las personas.

El preprocesamiento y limpieza de los datos se realizó utilizando Python 3.12.3 junto con bibliotecas especializadas como NLTK (Natural Language Toolkit), spaCy y pandas. Este proceso incluyó la eliminación de caracteres especiales y emojis, tokenización del texto, remoción de stopwords en español, normalización del texto y eliminación de duplicados y contenido irrelevante. Todo este proceso se realizó en un entorno de Jupyter Notebook, permitiendo mantener un orden claro en los procesos y facilitando la documentación del análisis.

Para el análisis de contenido se implementaron dos técnicas principales: la Bolsa de Palabras (Bag of Words) y Term Frequency-Inverse Document Frequency (TF-IDF). La primera técnica representó las palabras dentro de un mensaje como una colección única sin considerar el orden, utilizando únicamente el recuento de palabras y eliminando apariciones múltiples. La segunda técnica cuantificó la importancia de las palabras en el conjunto, considerando tanto la frecuencia de la palabra en el contenido como el peso de las palabras menos comunes.

El análisis exploratorio y la visualización de datos se realizaron utilizando bibliotecas especializadas de Python como matplotlib, seaborn y plotly. Se generaron diversos tipos de visualizaciones incluyendo distribuciones temporales de tweets, diagramas de barras para frecuencias de términos, mapas de calor para correlaciones, nubes de palabras y series temporales de actividad. Estas herramientas fueron fundamentales para representar relaciones numéricas y tendencias entre diferentes variables de manera clara y comprensible.

Para el análisis de sentimientos se implementaron dos aproximaciones complementarias: VADER (Valence Aware Dictionary and sEntiment Reasoner) y TextBlob. Se calcularon métricas estadísticas detalladas que incluyeron medias y medianas de sentimientos, correlaciones entre métodos y porcentajes de tweets positivos, neutrales y negativos. Este análisis doble permitió una comprensión más robusta de los sentimientos expresados en los tweets.

En la fase de modelado predictivo, se implementaron y evaluaron tres modelos de aprendizaje automático: Support Vector Machines (SVM), Random Forest y Naive Bayes. Cada modelo se aplicó utilizando una división de datos 80/20 para entrenamiento y prueba, respectivamente. Se realizó validación cruzada y se evaluaron métricas de rendimiento incluyendo accuracy, precision, recall y F1-Score. Los modelos se implementaron utilizando scikit-learn y se evaluaron para tres tareas específicas: clasificación de sentimientos, predicción de país de origen y determinación de relevancia.

Finalmente, como marco de cierre del proyecto, se preparó una serie de entregables para el programa CES de la UVG, buscando maximizar el valor y la aplicabilidad del estudio, lo

cual incluyó un informe detallado sobre los hallazgos, así como recomendaciones estratégicas basadas en los resultados. Además, se preparó una presentación con los hallazgos más importantes para facilitar la divulgación de esta información hacia una audiencia más amplia y acertada.

Esta metodología integral permitió obtener una comprensión profunda y cuantificable de la percepción del VIH y de otras enfermedades de transmisión sexual en Centroamérica, cumpliendo con los objetivos planteados en la investigación y proporcionando resultados estadísticamente significativos y reproducibles. Cualquier cambio u observación que surgió durante el desarrollo y que podría cambiar el alcance del estudio se documentó como recomendación para futuras extensiones de este trabajo.

CAPÍTULO VIII

Presentación de resultados

Como parte esencial del presente trabajo se realizaron diversos análisis preliminares que sirvieron de base para alcanzar los objetivos propuestos. La recolección inicial resultó en 4,954 tweets relacionados con VIH e ITS en la región centroamericana, que después de un riguroso proceso de limpieza y preprocesamiento, se refinó a 3,533 tweets únicos y relevantes. Este proceso incluyó la eliminación 1,421 de tweets duplicados y considerados irrelevantes, asegurando así la calidad y pertinencia de los datos analizados.

El conjunto final de datos presentó características fundamentales que sentaron las bases para los análisis posteriores. El análisis textual del dataset reveló una longitud promedio de tweets de 129.25 caracteres, con una mediana de 139 caracteres, y un promedio de 17.88 palabras por tweet. Estas métricas básicas proporcionaron una comprensión inicial de la naturaleza y estructura de los datos recopilados.

Es importante destacar que estos pasos preliminares, aunque no estén directamente vinculados con los objetivos específicos de la investigación, fueron esenciales para garantizar la solidez del análisis posterior y la validez de los resultados. La rigurosidad en esta etapa inicial permitió establecer una base sólida para los análisis más específicos y detallados que se presentan a continuación, organizados según los objetivos específicos del mismo.

Los resultados que se presentan a continuación se han estructurado de manera que respondan directamente a cada uno de los objetivos específicos planteados en la investigación, permitiendo así una evaluación clara y sistemática del cumplimiento de estos y, por ende, del objetivo general del estudio.

A. Recopilación y preparación de datos

1. Estadísticas de la recolección

Cuadro 1: Distribución de tweets por país antes y después del procesamiento

País	Tweets iniciales	Tweets procesados
Nicaragua	3,515	2,453
Panamá	445	305
Guatemala	291	220
El Salvador	261	184
Costa Rica	237	196
Honduras	193	166
Belice	12	9
Total	4,954	3,533

La recolección inicial de datos fue de 4,954 tweets relacionados con VIH e ITS en la región centroamericana. La distribución geográfica mostró una concentración significativa en Nicaragua (3,515 tweets), seguida por Panamá (445), Guatemala (291), El Salvador (261), Costa Rica (237), Honduras (193) y Belice (12). El período de recolección abarcó desde el 26 de julio de 2024 hasta el 9 de enero de 2025.

Cuadro 2: Top 10 hashtags más frecuentes

Hashtag	Menciones
#VIH	80
#Guatemala	40
#ProgramaDeMigraciónLaboral	29
#Honduras	19
#Nicaragua	17
#ElSalvador	15
#Nacionales	13
#SIDA	11
#Linea1540	10
#PrevenciónVIH	9

Los hashtags más frecuentes en la muestra inicial reflejaron la relevancia temática, siendo #VIH el más común con 80 menciones, seguido por hashtags geográficos como #Guatemala (40 menciones) y temáticos como #PrevenciónVIH (9 menciones).

2. Resultados del proceso de limpieza

El proceso de limpieza y filtrado resultó en una reducción significativa del dataset:

- Se identificaron y eliminaron 1,421 tweets duplicados y considerados irrelevantes
- El dataset final quedó conformado por 3,533 tweets únicos y relevantes

3. Estadísticas descriptivas del dataset procesado

El análisis textual del dataset final reveló:

- Longitud promedio de tweets: 129.25 caracteres
- Mediana de longitud: 139 caracteres
- Promedio de palabras por tweet: 17.88

B. Análisis exploratorio de datos

La recolección de datos resultó en un total de 3,533 tweets distribuidos entre los países de Centroamérica. A continuación, se presentan los principales hallazgos del análisis exploratorio de datos.

1. Distribución geográfica

La Figura 1 muestra la distribución de tweets por país.

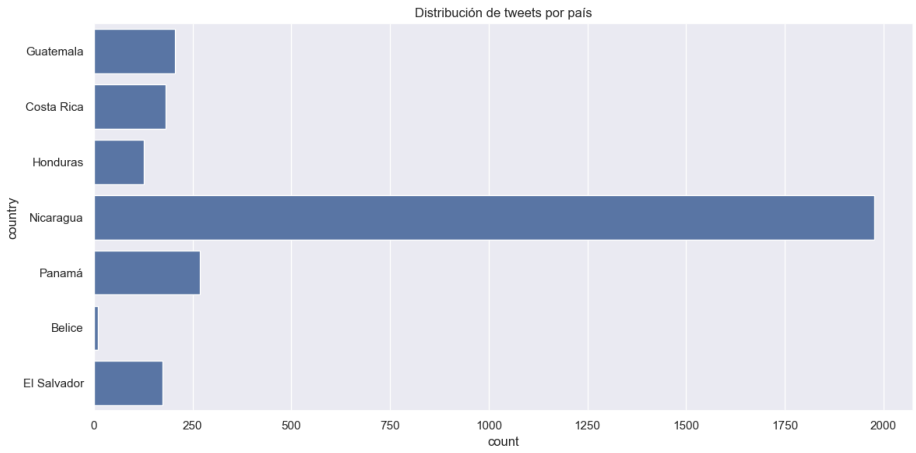


Figura 1: Distribución de tweets recolectados por país en Centroamérica

2. Serie temporal

La distribución temporal de tweets (Figura 2) muestra el inicio y el final de la extracción de los datos.

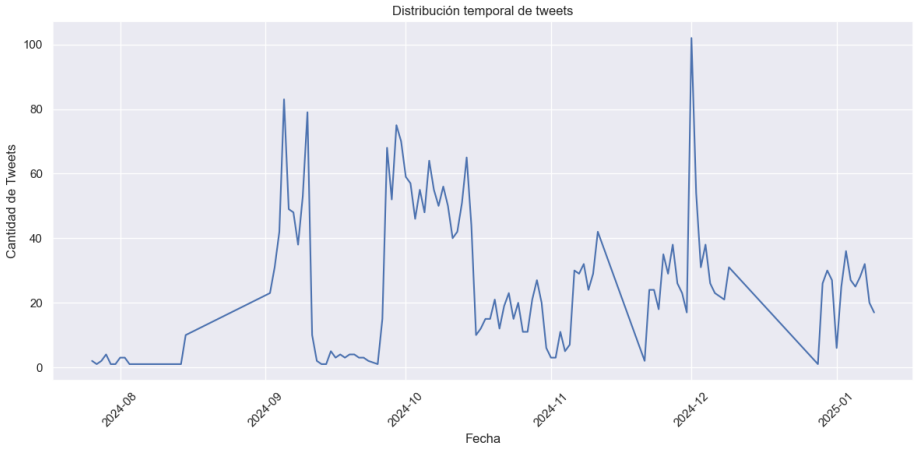


Figura 2: Distribución temporal de tweets

3. Características de los tweets

El análisis de la longitud de los tweets por país (Figura 3) indica variaciones en los patrones de escritura entre países.

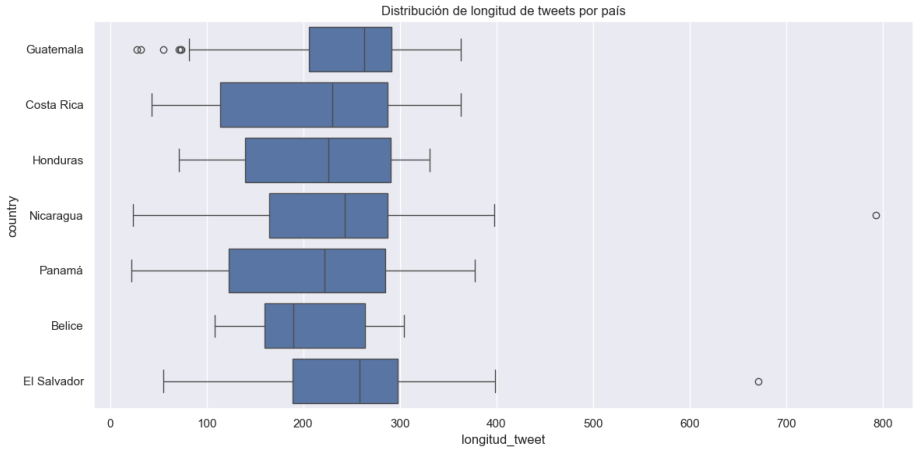
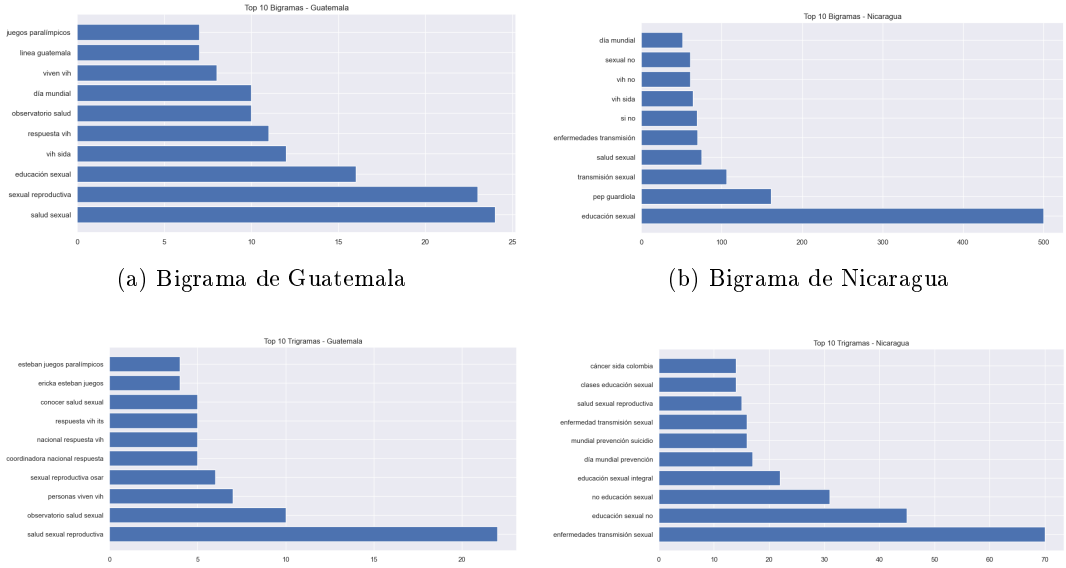


Figura 3: Distribución de longitud de tweets por país

4. Análisis de n-gramas

Los bigramas y trigramas más frecuentes que revelan las frases más comunes utilizadas en las discusiones sobre VIH.



(a) Bigrama de Guatemala

(b) Bigrama de Nicaragua

(c) Trigrama de Guatemala

(d) Trigrama de Nicaragua

Figura 4: Gráficos de n-gramas por país

C. Análisis de patrones de texto y percepción de sentimientos

En esta sección se presentan los resultados correspondientes al primer objetivo, enfocado en el análisis de patrones de texto y percepción de sentimientos en las publicaciones sobre VIH e ITS en Centroamérica. Los resultados se dividen en dos categorías principales: el análisis de sentimientos, que evalúa la polaridad emocional de las publicaciones; y el análisis de percepción, que examina los patrones y relaciones temáticas en el contenido.

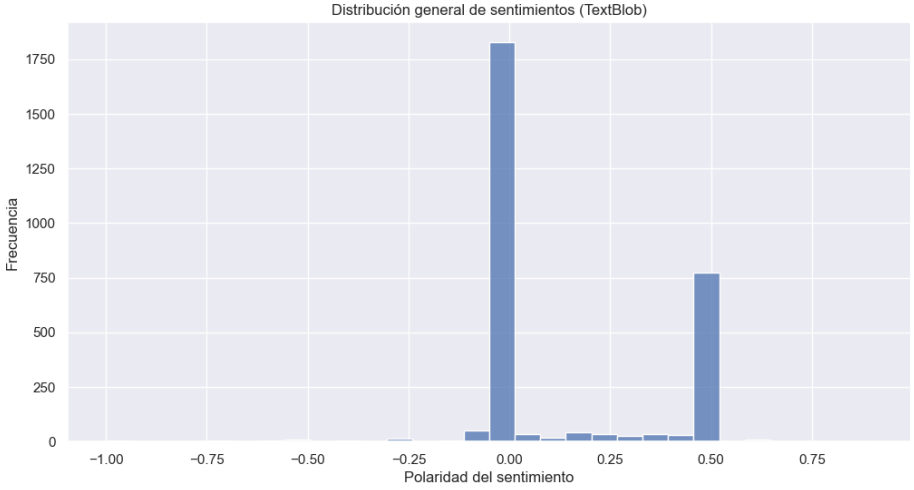


Figura 5: Distribución general de sentimientos (TextBlob)

1. Análisis de sentimientos

Distribución general de sentimientos

La evaluación del sentimiento se realizó utilizando dos metodologías complementarias: TextBlob y VADER. La Figura 5 muestra la distribución general de sentimientos según TextBlob, donde se observa una predominancia de contenido neutral (79.41 %), seguido por contenido positivo (19.16 %) y una menor proporción de contenido negativo (1.43 %). Esta distribución sugiere que la mayoría de las discusiones sobre VIH e ITS en la región mantienen un tono informativo y objetivo.

Comparación de metodologías

Para validar la robustez del análisis, se realizó una comparación entre las metodologías TextBlob y VADER. La Figura 6 presenta las distribuciones mediante diagramas de caja, permitiendo observar la variabilidad en las clasificaciones.

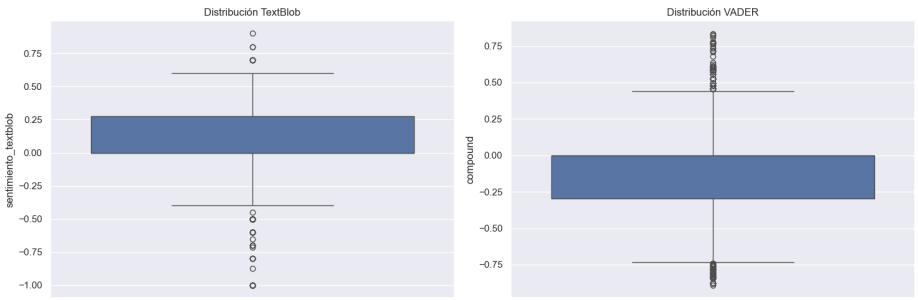


Figura 6: Distribución TextBlob vs sentimiento / distribución VADER vs Compound

Los estadísticos descriptivos del análisis indican una ligera tendencia hacia la neutralidad con una leve inclinación positiva:

Cuadro 3: Estadísticos descriptivos del análisis de sentimientos

Métrica	Valor
Media TextBlob	0.085047
Media VADER	-0.129232
Mediana TextBlob	0.000000
Mediana VADER	0.000000
Correlación entre métodos	0.128623

Distribución por país

El análisis de sentimientos por país revela patrones distintivos en diferentes regiones de Centroamérica. Las figuras 7 y 8 presentan estas distribuciones.

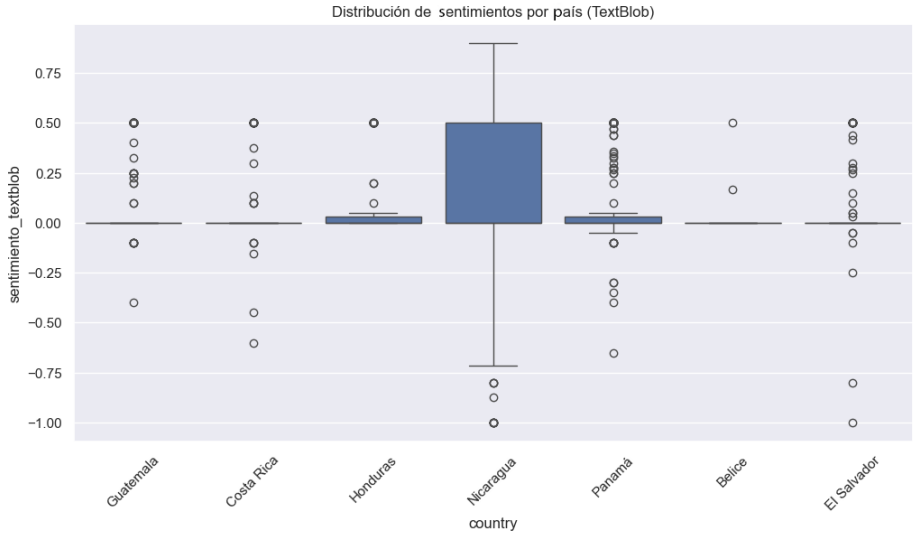


Figura 7: Distribución de sentimientos por país

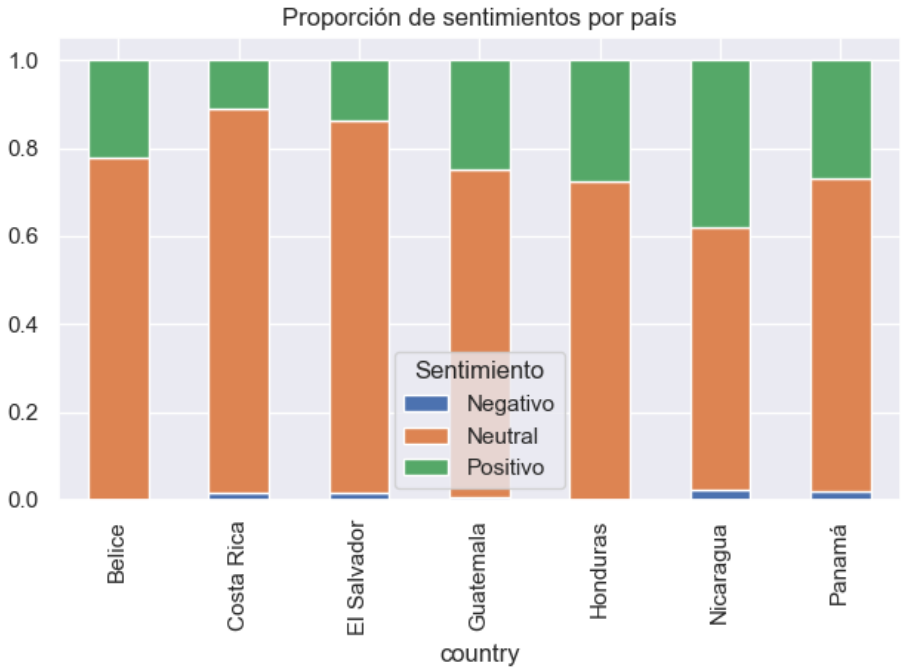


Figura 8: Proporción de sentimientos por país

2. Análisis de percepción

Análisis de frecuencia, menciones y patrones de interacción

El análisis de la percepción se centra en las relaciones entre diferentes temas asociados al VIH. Los niveles de interacción se manifestaron principalmente a través de los términos más

frecuentes en las discusiones. Estos patrones de frecuencia sugieren un enfoque significativo en la educación y prevención dentro de las discusiones sobre VIH e ITS en la región.

Cuadro 4: Términos más frecuentes en tweets procesados

Término	Menciones
sexual	941
vih	724
sida	684
educación	674
prevención	181

La frecuencia de menciones proporciona información sobre la prominencia de diferentes temas y su distribución geográfica. Las figuras 9 y 10 muestran estos patrones.

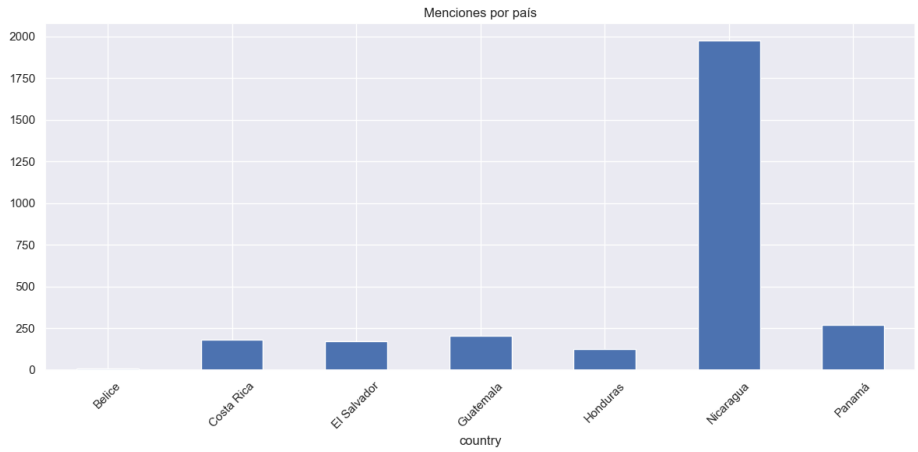


Figura 9: Menciones por país

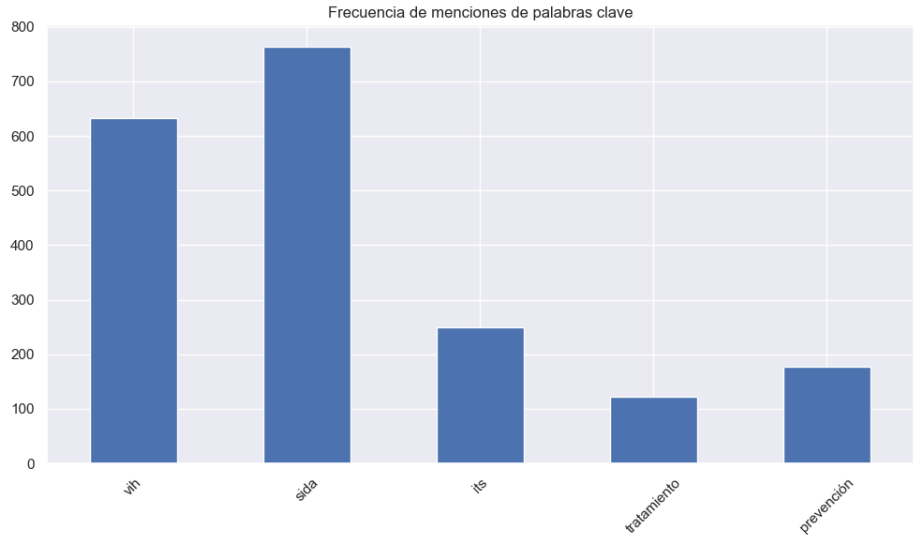


Figura 10: Frecuencia de palabras clave

3. Validación mediante modelos de clasificación de sentimientos

Modelo Random Forest para sentimientos

El modelo Random Forest aplicado a la clasificación de sentimientos mostró un rendimiento general sólido, alcanzando una precisión global del 95 %. En la Figura 11 se presenta la matriz de confusión del modelo, donde se puede observar un desempeño particularmente robusto en la clasificación de publicaciones neutrales, con 514 predicciones correctas de esta categoría.

Los resultados detallados del modelo muestran las siguientes métricas:

Cuadro 5: Métricas de evaluación - Random Forest para sentimientos

Clase	**Precisión**	**Recall**	**F1-Score**
Negativo	0.75	0.14	0.23
Neutral	0.94	0.99	0.96
Positivo	0.99	0.90	0.94
Accuracy		0.95	

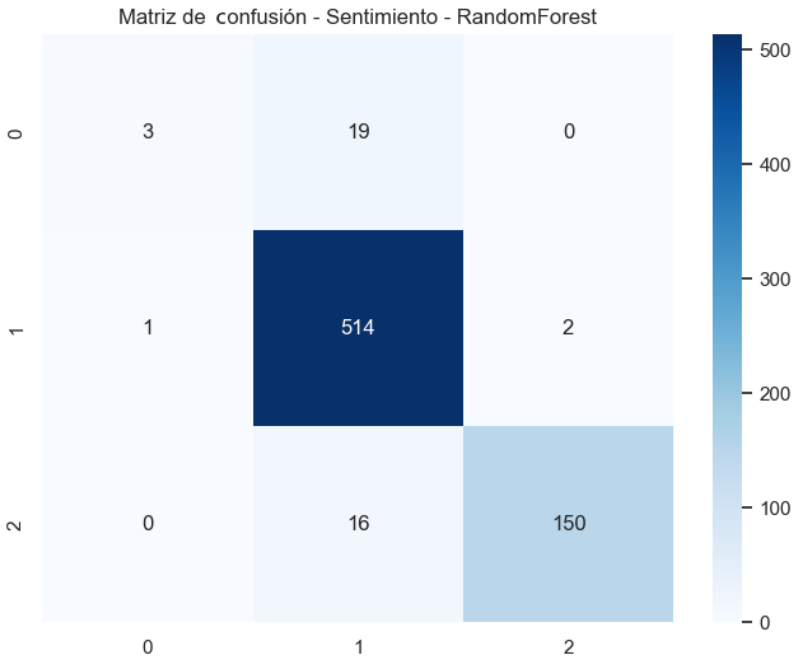


Figura 11: Matriz de confusión - Random Forest para sentimientos

La Figura 12 muestra la distribución de los scores de validación cruzada, indicando una estabilidad considerable en el rendimiento del modelo con una exactitud de 0.95.

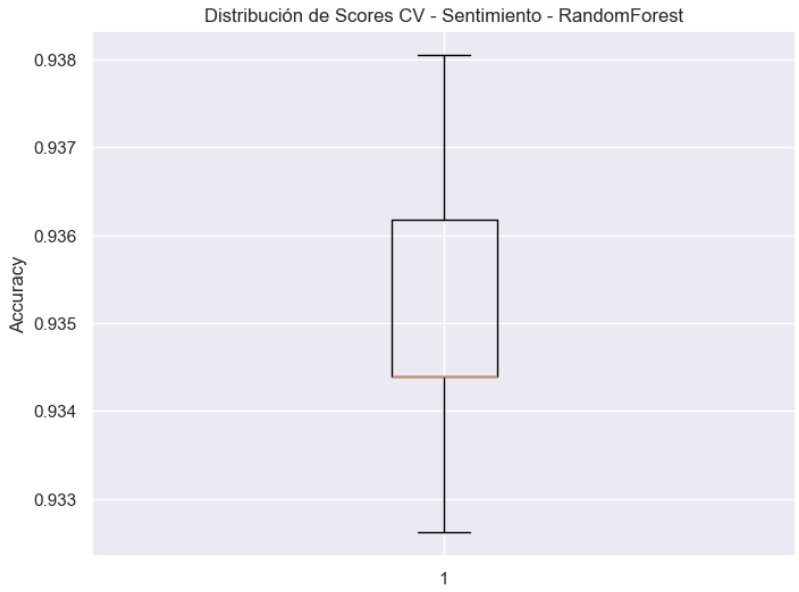


Figura 12: Distribución de Scores CV - Random Forest para sentimientos

Modelo SVM para sentimientos

El modelo Support Vector Machine (SVM) aplicado a la clasificación de sentimientos demostró un rendimiento sobresaliente, alcanzando una precisión global del 95 %. En la Figura 13 se presenta la matriz de confusión del modelo, donde se puede observar un desempeño particularmente notable en la clasificación de publicaciones neutrales, con 515 predicciones correctas de esta categoría.

Los resultados detallados del modelo muestran las siguientes métricas:

Cuadro 6: Métricas de evaluación - SVM para sentimientos

Clase	**Precisión**	**Recall**	**F1-Score**
Negativo	0.90	0.41	0.56
Neutral	0.94	1.00	0.97
Positivo	0.99	0.89	0.94
Accuracy		0.95	

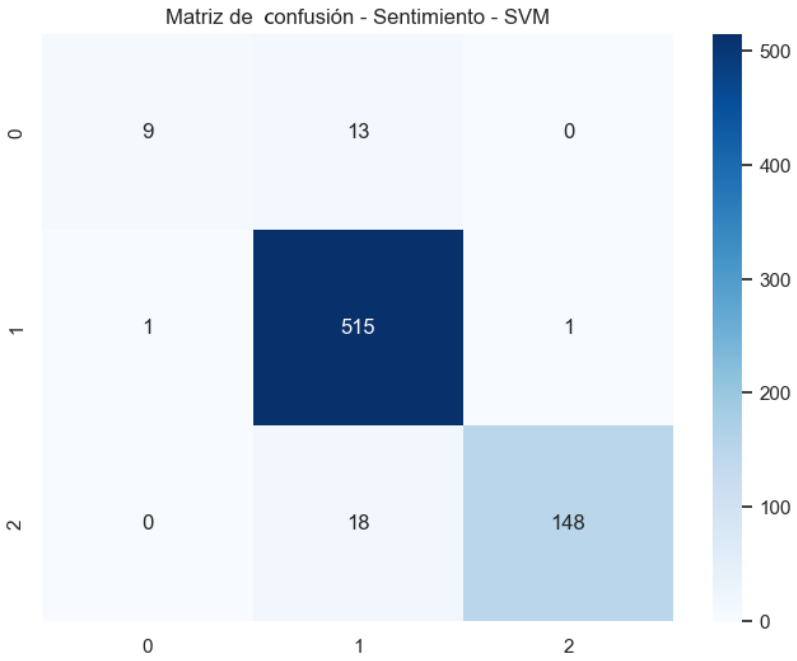


Figura 13: Matriz de confusión - SVM para sentimientos

La Figura 14 muestra la distribución de los puntuajes de validación cruzada, indicando una estabilidad considerable en el rendimiento del modelo con una exactitud de 0.95. La distribución de los puntuajes sugiere una consistencia robusta en el rendimiento del modelo a través de diferentes subconjuntos de datos.

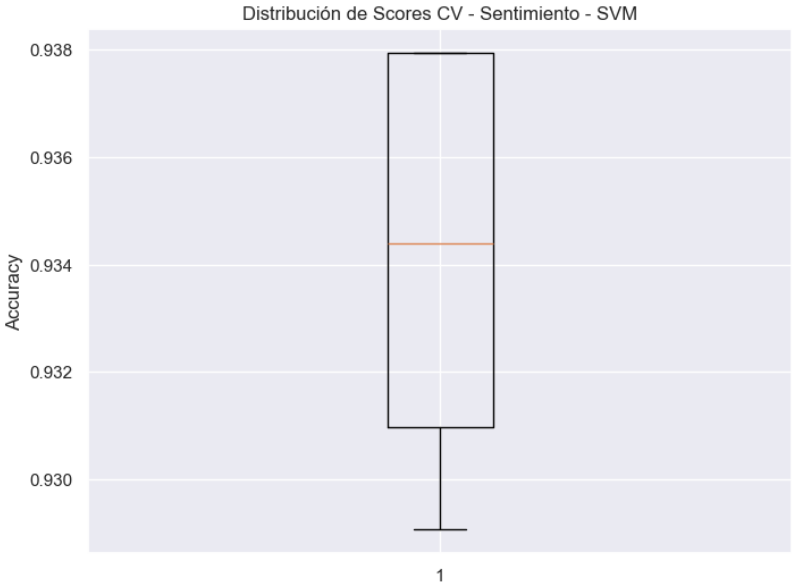


Figura 14: Distribución de Scores CV - SVM para sentimientos

Modelo Naive Bayes para sentimientos

El modelo Naive Bayes aplicado a la clasificación de sentimientos mostró un rendimiento diferenciado entre las distintas categorías, alcanzando una precisión global del 89 %. En la Figura 15 se presenta la matriz de confusión del modelo, donde se puede observar un desempeño notable en la clasificación de publicaciones neutrales, con 514 predicciones correctas, aunque con limitaciones significativas en la identificación de sentimientos negativos.

Los resultados detallados del modelo muestran las siguientes métricas:

Cuadro 7: Métricas de evaluación - Naive Bayes para sentimientos

Clase	**Precisión**	**Recall**	**F1-Score**
Negativo	0.00	0.00	0.00
Neutral	0.87	0.99	0.93
Positivo	0.97	0.67	0.80
Accuracy		0.89	

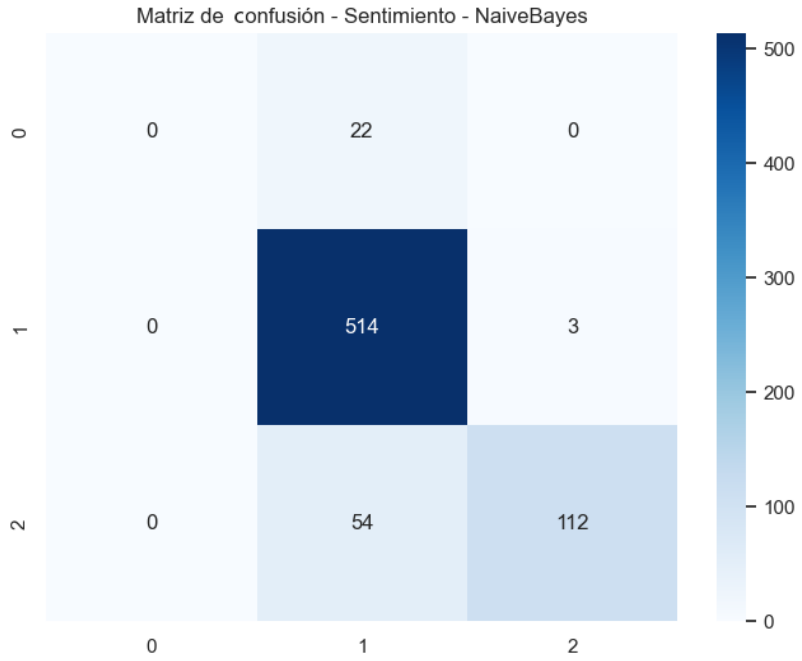


Figura 15: Matriz de confusión - Naive Bayes para sentimientos

La Figura 16 muestra la distribución de los puntuajes de validación cruzada, indicando una estabilidad moderada en el rendimiento del modelo con una exatitud de 0.89. La distribución de los scores presenta una variabilidad más pronunciada en comparación con los modelos anteriores, sugiriendo una menor consistencia en el rendimiento a través de diferentes subconjuntos de datos.

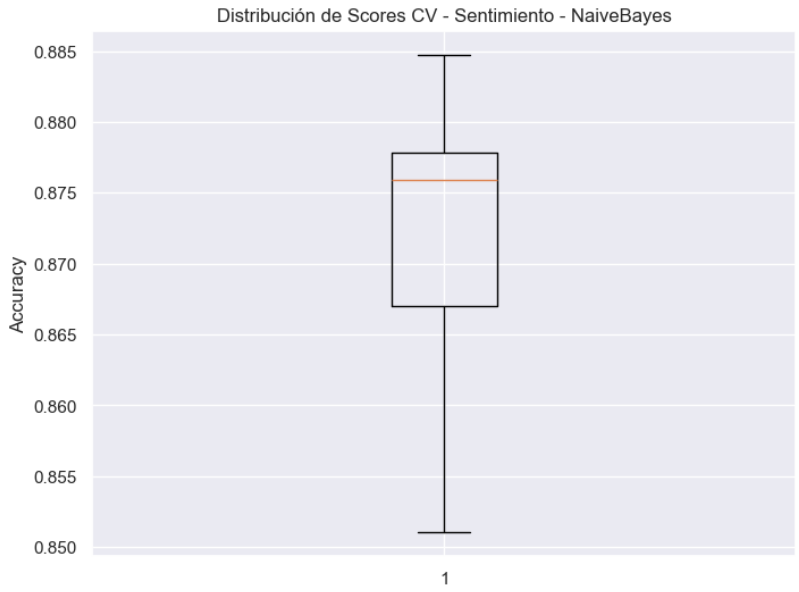


Figura 16: Distribución de Scores CV - Naive Bayes para sentimientos

4. Comparación de modelos para clasificación de sentimientos

La comparación de los tres modelos implementados para la clasificación de sentimientos revela patrones interesantes en su rendimiento. En la Figura 17 se presenta una visualización comparativa de las principales métricas de evaluación para cada modelo, donde se puede observar un rendimiento consistentemente alto en los modelos SVM y Random Forest, con algunas diferencias notables en aspectos específicos.

Cuadro 8: Resumen comparativo de modelos para clasificación de sentimientos

Métrica	**Mejor Modelo**	**Score**	**Diferencia***
Accuracy	SVM	92.87 %	+3.87 %
F1-Score	SVM	92.09 %	+5.09 %
Precision	SVM	92.56 %	+5.56 %
Recall	SVM	92.87 %	+3.87 %
CV-Score Mean	RandomForest	93.99 %	+5.99 %
CV-Score Std	NaiveBayes	1.52 %	–

*Diferencia respecto al modelo de peor rendimiento

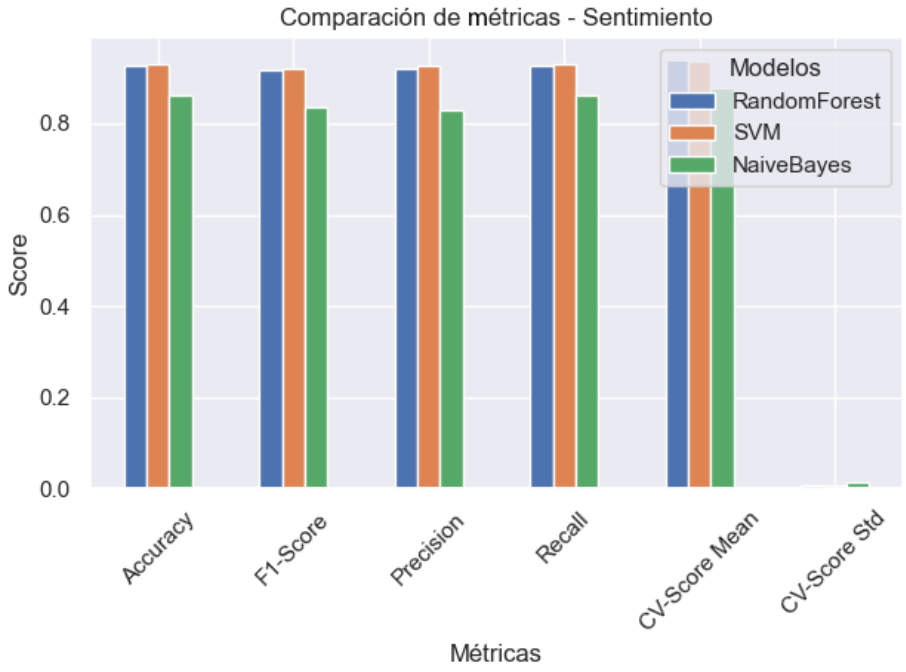


Figura 17: Comparación de métricas - Modelos de clasificación de sentimientos

El modelo SVM demostró un rendimiento superior en varias métricas clave, alcanzando los mejores resultados en accuracy (92.87 %), F1-Score (92.09 %), precisión (92.56 %) y recall (92.87 %). Sin embargo, el modelo Random Forest mostró la mayor estabilidad en la validación cruzada, con un score medio de 93.99 % y una desviación estándar de apenas 0.69 %, como se puede observar en la Figura 18.

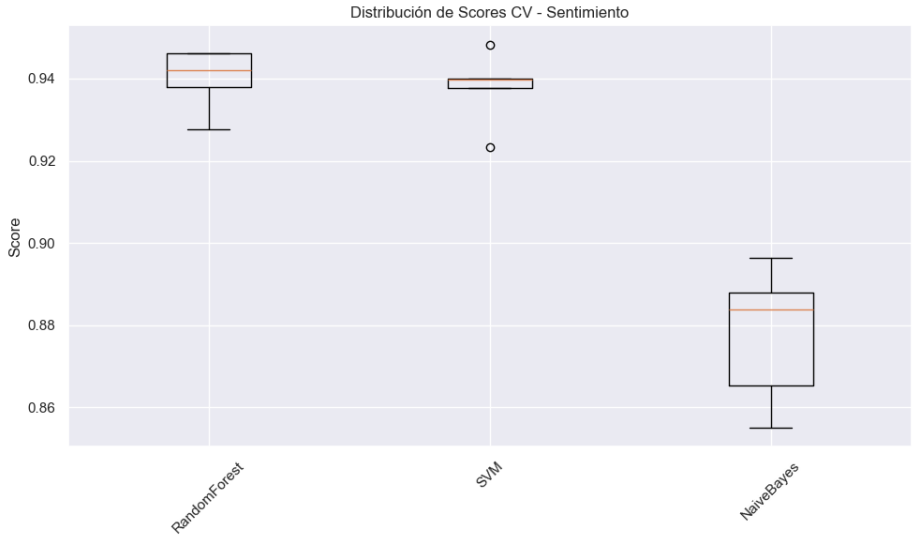


Figura 18: Distribución de Scores CV - Comparativa de modelos para sentimientos

Es notable que el modelo Naive Bayes, aunque presentó el rendimiento más bajo en

términos generales, mostró la mayor variabilidad en sus predicciones con una desviación estándar de 1.52 % en los scores de validación cruzada. Esta característica sugiere una menor confiabilidad en sus predicciones en comparación con los otros dos modelos.

La evaluación integral de los resultados sugiere que el modelo SVM es la opción más adecuada para la clasificación de sentimientos en este contexto, confirmando la fiabilidad de los patrones identificados, combinando un alto rendimiento en todas las métricas con una estabilidad robusta en sus predicciones. El modelo Random Forest se presenta como una alternativa muy competitiva, especialmente en situaciones donde la estabilidad de las predicciones es prioritaria.

D. Comparación de patrones y tendencias entre países

Este apartado correspondientes al segundo y cuarto objetivo específico se presentan de manera integrada, dado que ambos se centran en el análisis comparativo entre las diferentes regiones de Centroamérica. La investigación logró no solo identificar patrones y tendencias significativas entre los países, sino también establecer similitudes y diferencias en cuanto a preocupaciones, actitudes y niveles de conocimiento sobre el VIH e ITS en la región.

Distribución regional del contenido

El análisis de la distribución geográfica reveló patrones distintivos en el volumen y naturaleza de las discusiones sobre VIH e ITS entre los países de la región. Esta distribución desigual sugiere diferentes niveles de participación y compromiso con la temática en cada país.

Cuadro 9: Distribución geográfica de tweets recolectados por país

País	Cantidad de tweets
Nicaragua	2,453
Panamá	305
Guatemala	220
Costa Rica	196
El Salvador	184
Honduras	166
Belice	9

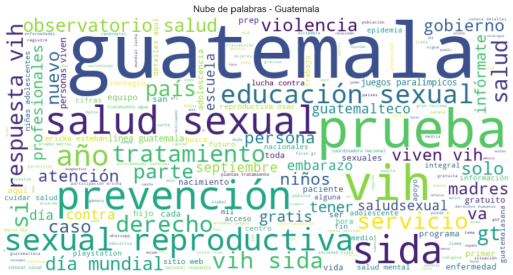
Análisis de contenido textual

El análisis de contenido reveló diferencias significativas en las necesidades de información entre países. Las nubes de palabras y el análisis TF-IDF mostraron patrones distintivos en cada región, destacando variaciones en:

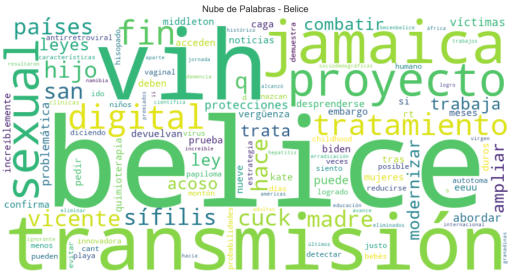
- Términos relacionados con prevención

- Referencias a servicios de salud
- Menciones de tratamientos específicos
- Discusiones sobre educación sexual

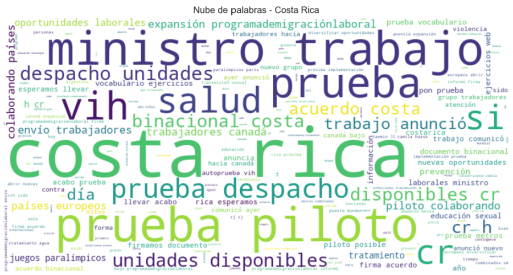
El análisis de términos frecuentes por país se realizó mediante nubes de palabras, revelando los términos más utilizados en las discusiones sobre VIH e ITS en cada país.



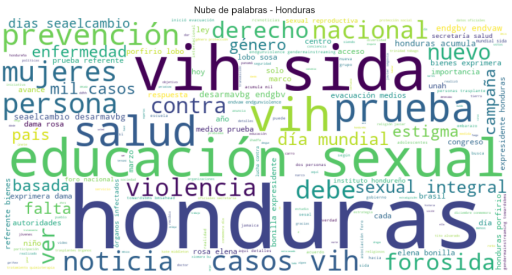
(a) Guatemala



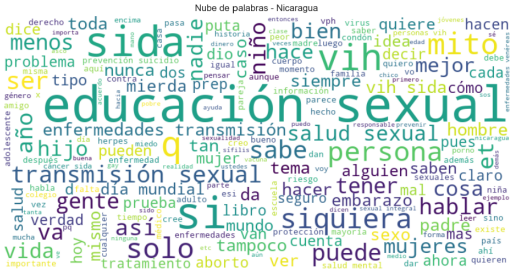
(b) Belice



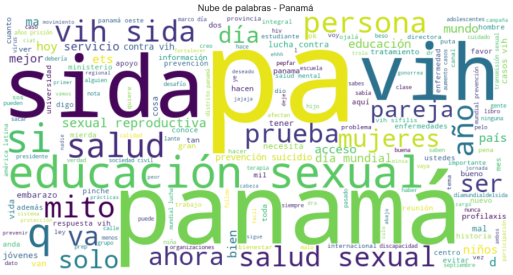
(c) Costa Rica



(d) Honduras



(e) Nicaragua



(f) Panamá

(g) El Salvador

Figura 19: Nubes de palabras por país

La importancia relativa de los términos se evaluó mediante el análisis TF-IDF (Figura 20), identificando las palabras más distintivas en el conjunto de datos.

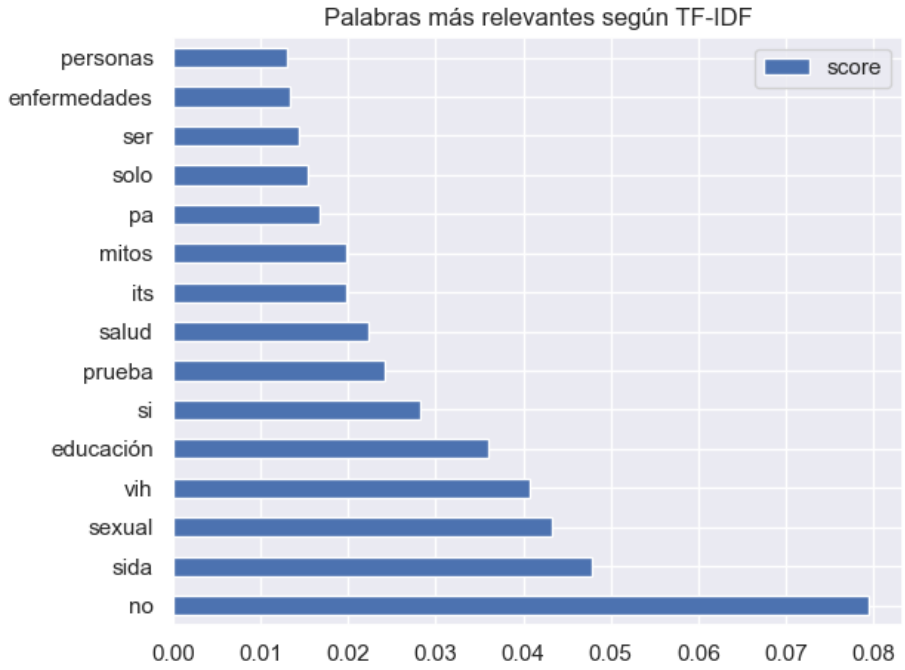


Figura 20: Palabras relevantes según TF-IDF

Esta diversidad en el contenido sugiere diferentes prioridades y necesidades de información en cada país de la región centroamericana.

1. Validación mediante modelos de clasificación por país

El modelo Random Forest aplicado a la clasificación por país demostró un rendimiento robusto, alcanzando una precisión global del 91 %. En la Figura 21 se presenta la matriz de confusión del modelo, donde se destaca particularmente su capacidad para clasificar publicaciones de Nicaragua, con 480 predicciones correctas, demostrando una alta precisión en la identificación de contenido por región geográfica.

Los resultados detallados del modelo muestran las siguientes métricas:

Cuadro 10: Métricas de evaluación - Random Forest para clasificación por país

País	**Precisión**	**Recall**	**F1-Score**
Costa Rica	0.97	0.71	0.82
El Salvador	0.85	0.69	0.76
Guatemala	1.00	0.70	0.82
Honduras	0.89	0.70	0.78
Nicaragua	0.91	0.99	0.95
Panamá	0.90	0.91	0.91
Accuracy		0.91	

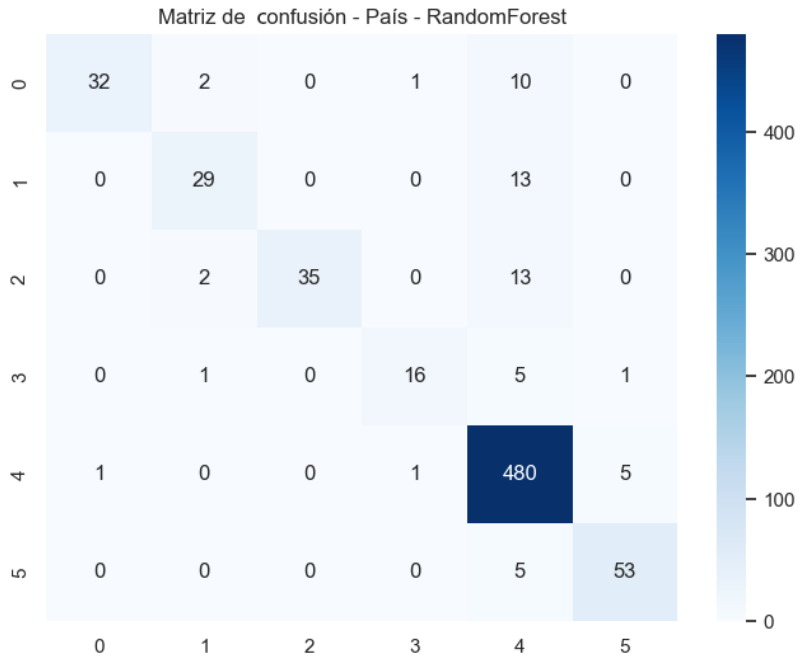


Figura 21: Matriz de confusión - Random Forest para clasificación por país

La Figura 22 muestra la distribución de los puntajes de validación cruzada, indicando una estabilidad notable en el rendimiento del modelo con una exactitud promedio de 0.901. La distribución de los scores sugiere una consistencia robusta en el rendimiento del modelo a través de diferentes subconjuntos de datos, particularmente en la identificación de patrones regionales específicos.

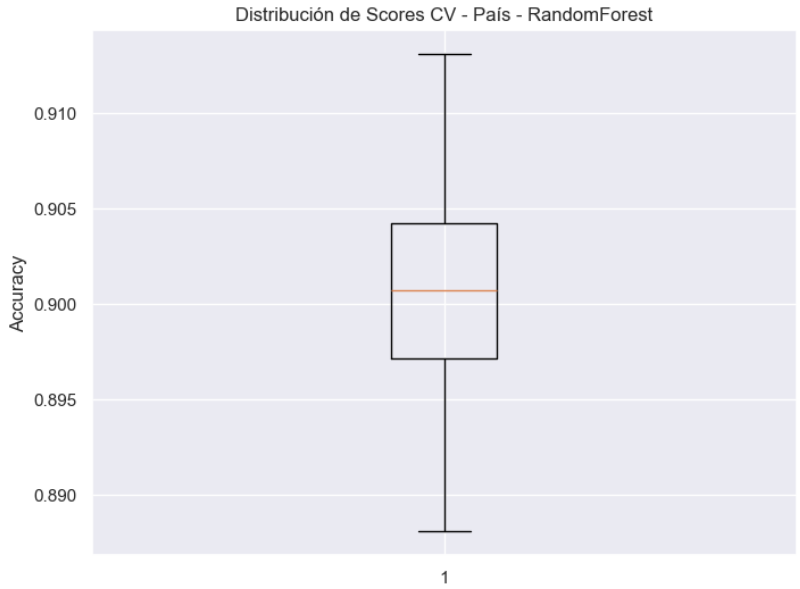


Figura 22: Distribución de Scores CV - Random Forest para clasificación por país

Modelo SVM para clasificación por país

El modelo Support Vector Machine (SVM) aplicado a la clasificación por país exhibió un rendimiento sobresaliente alcanzando una precisión global del 90 %. En la Figura 23 se presenta la matriz de confusión del modelo, donde se destaca su efectividad particularmente en la clasificación de publicaciones de Nicaragua logrando identificar correctamente 480 instancias, así como un rendimiento notable en la clasificación de publicaciones de Panamá con 50 predicciones correctas.

Los resultados detallados del modelo muestran las siguientes métricas:

Cuadro 11: Métricas de evaluación - SVM para clasificación por país

País	**Precisión**	**Recall**	**F1-Score**
Costa Rica	1.00	0.64	0.78
El Salvador	0.82	0.64	0.72
Guatemala	1.00	0.70	0.82
Honduras	0.89	0.70	0.78
Nicaragua	0.90	0.99	0.94
Panamá	0.91	0.86	0.88
Accuracy		0.90	

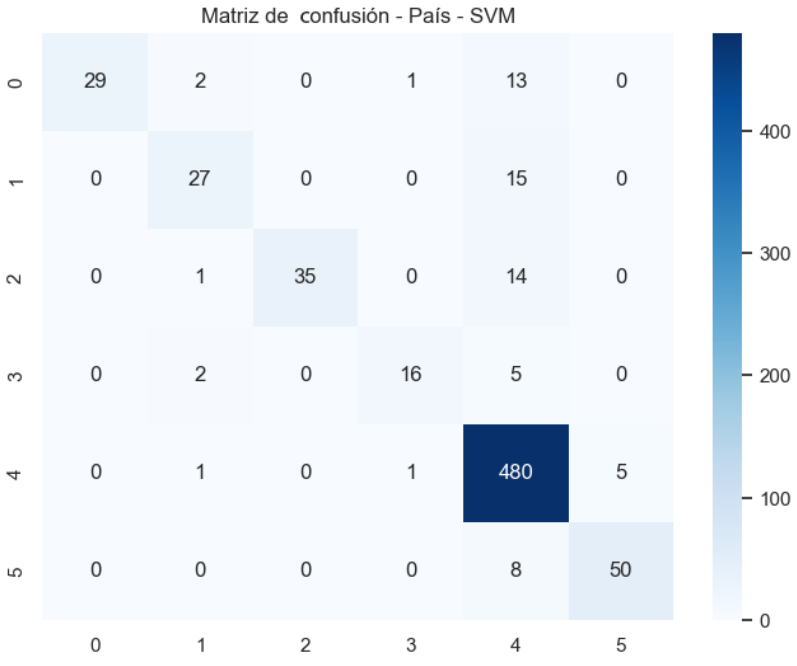


Figura 23: Matriz de confusión - SVM para clasificación por país

La Figura 24 muestra la distribución de los puntajes de validación cruzada, indicando una estabilidad considerable en el rendimiento del modelo con una exactitud de 0.883. La distribución de los scores sugiere una consistencia sólida en el rendimiento del modelo a través de diferentes subconjuntos de datos, aunque con una variabilidad ligeramente mayor

que la observada en el modelo Random Forest.

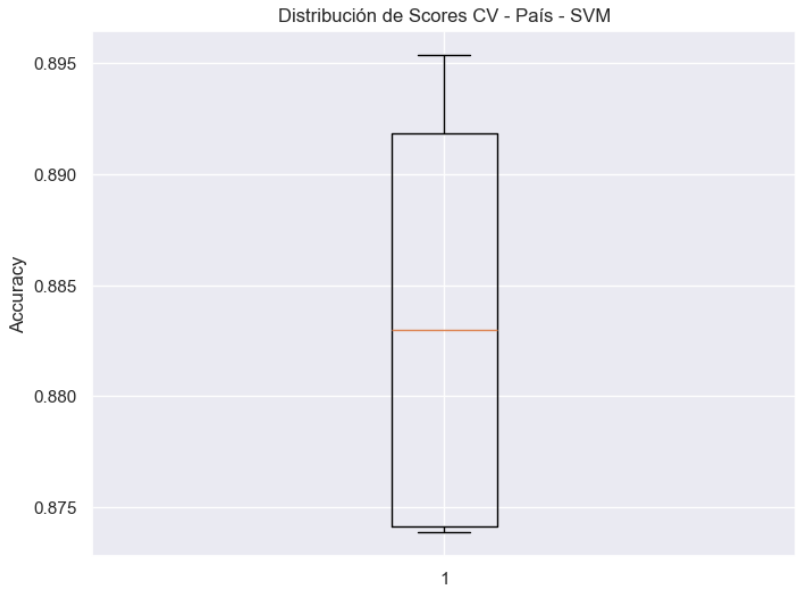


Figura 24: Distribución de Scores CV - SVM para clasificación por país

Modelo Naive Bayes para clasificación por país

El modelo Naive Bayes aplicado a la clasificación por país mostró un comportamiento particular, alcanzando una precisión global del 73 %. En la Figura 25 se presenta la matriz de confusión del modelo, donde se observa un patrón distintivo en la clasificación: una alta concentración de predicciones para Nicaragua, con 487 identificaciones correctas, pero con limitaciones significativas en la identificación de publicaciones de otros países.

Los resultados detallados del modelo muestran las siguientes métricas:

Cuadro 12: Métricas de evaluación - Naive Bayes para clasificación por país

País	**Precisión**	**Recall**	**F1-Score**
Costa Rica	1.00	0.38	0.55
El Salvador	1.00	0.05	0.09
Guatemala	1.00	0.08	0.15
Honduras	1.00	0.13	0.23
Nicaragua	0.72	1.00	0.84
Panamá	1.00	0.03	0.07
Accuracy		0.73	

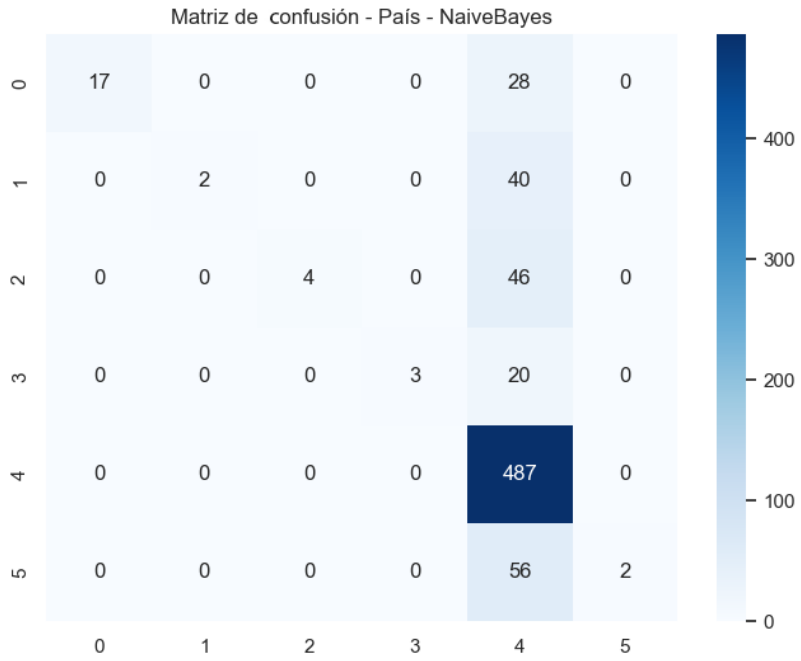


Figura 25: Matriz de confusión - Naive Bayes para clasificación por país

La Figura 26 muestra la distribución de los scores de validación cruzada, indicando una estabilidad moderada en el rendimiento del modelo con una exactitud promedio de 0.729. La distribución de los scores revela una variabilidad notablemente mayor en comparación con los modelos Random Forest y SVM, sugiriendo una menor consistencia en el rendimiento a través de diferentes subconjuntos de datos.

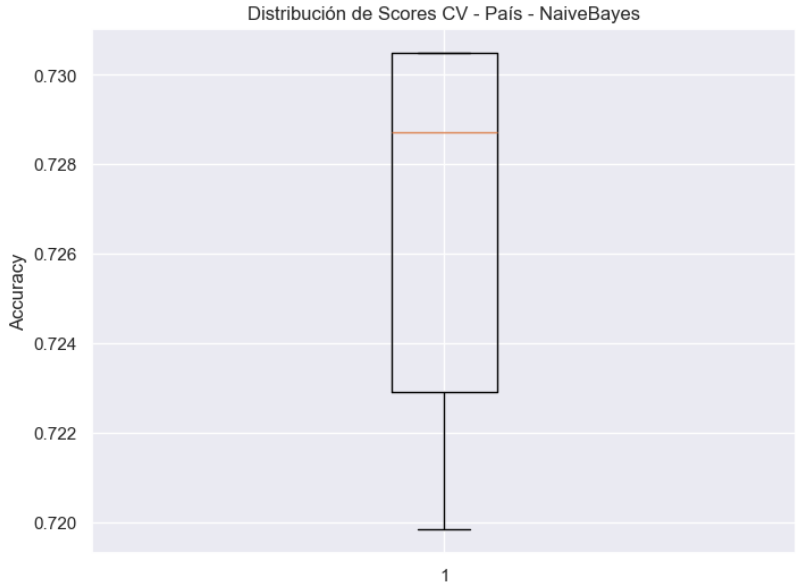


Figura 26: Distribución de Scores CV - Naive Bayes para clasificación por país

2. Comparación de modelos para clasificación por país

El análisis comparativo de los tres modelos implementados para la clasificación por país revela un claro dominante en términos de rendimiento. A continuación, se presenta un resumen detallado de las métricas más relevantes:

Cuadro 13: Resumen comparativo de modelos para clasificación por país

Métrica	**Mejor Modelo**	**Score**	**Diferencia***
Accuracy	RandomForest	90.22 %	+17.22 %
F1-Score	RandomForest	89.75 %	+25.75 %
Precision	RandomForest	90.45 %	+12.45 %
Recall	RandomForest	90.22 %	+17.22 %
CV-Score Mean	RandomForest	89.97 %	+17.67 %
CV-Score Std	SVM	1.22 %	–

*Diferencia respecto al modelo de peor rendimiento

En la Figura 27 se presenta una visualización comparativa de las principales métricas de evaluación para cada modelo. El modelo Random Forest destaca notablemente liderando en todas las métricas de rendimiento principales.

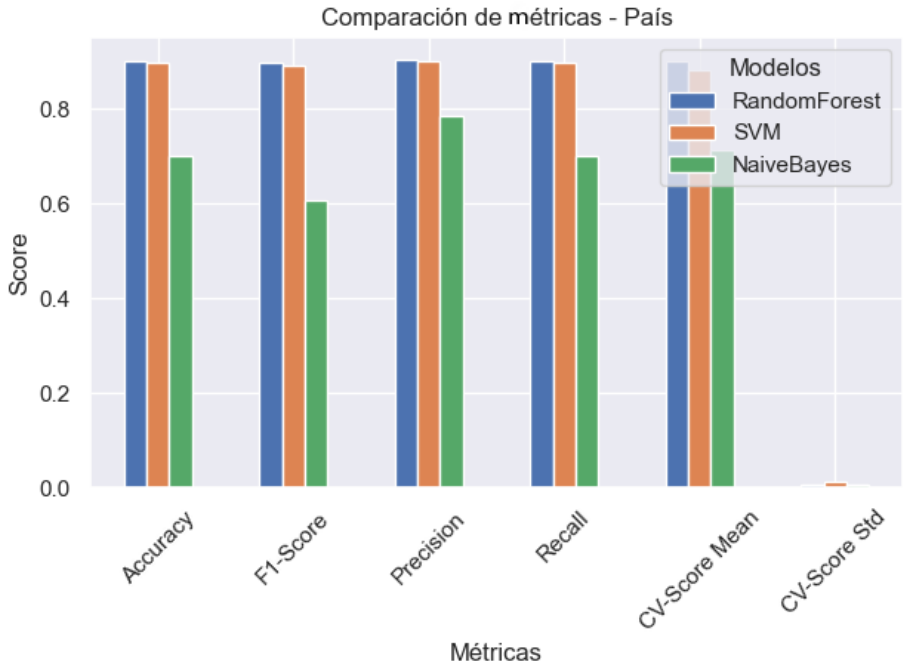


Figura 27: Comparación de métricas - Modelos de clasificación por país

El modelo Random Forest demostró un rendimiento sobresaliente alcanzando la mejor puntuación en accuracy (90.22 %), F1-Score (89.75 %), precisión (90.45 %), recall (90.22 %) y score medio de validación cruzada (89.97 %). La Figura 28 muestra la distribución de los scores de validación cruzada, donde se puede observar la superioridad consistente del Random Forest.

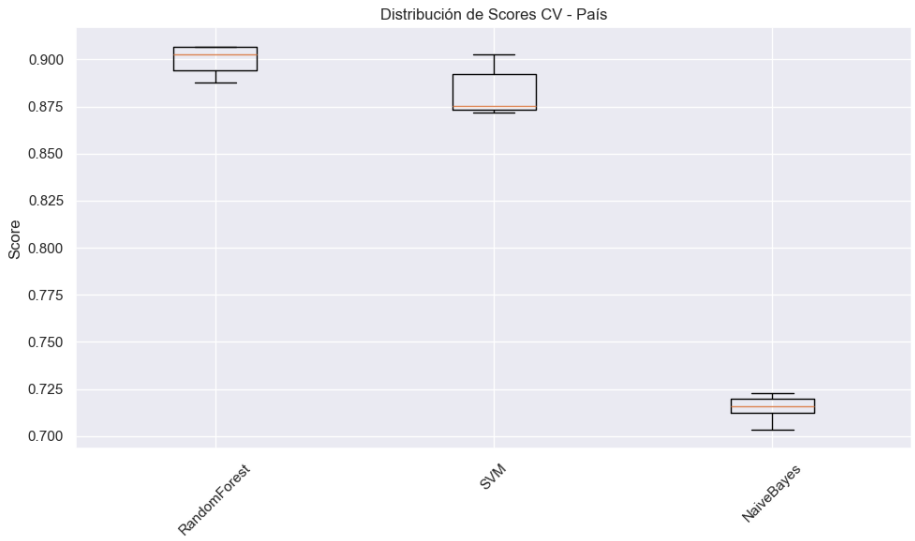


Figura 28: Distribución de Scores CV - Comparativa de modelos para clasificación por país

Es interesante notar que, aunque el modelo SVM mostró la menor desviación estándar en los scores de validación cruzada (1.22 %), el modelo Naive Bayes resultó ser el más consistente en términos generales con una desviación estándar de 0.67 %. Sin embargo, el rendimiento significativamente inferior del Naive Bayes en todas las demás métricas lo hace menos adecuado para esta tarea específica.

La evaluación integral de los resultados indica claramente que el modelo Random Forest es la opción óptima para la clasificación por país demostrando superioridad en prácticamente todas las métricas evaluadas y manteniendo un nivel de consistencia satisfactorio en sus predicciones.

La integración de estos hallazgos proporciona una comprensión más profunda de las dinámicas regionales en torno al VIH e ITS en Centroamérica, cumpliendo así con los objetivos de comparar patrones entre países e identificar similitudes y diferencias regionales. Esta información resulta fundamental para el desarrollo de estrategias de comunicación y programas de salud pública que consideren las particularidades y necesidades específicas de cada región.

E. Identificación de demandas hacia actores estatales

Este apartado corresponde al tercer objetivo específico, enfocado en identificar posibles acciones de demanda hacia el Estado y otros actores.

1. Análisis de interacciones temáticas

La Figura 29 de interacciones temáticas reveló patrones significativos en las demandas de los usuarios.

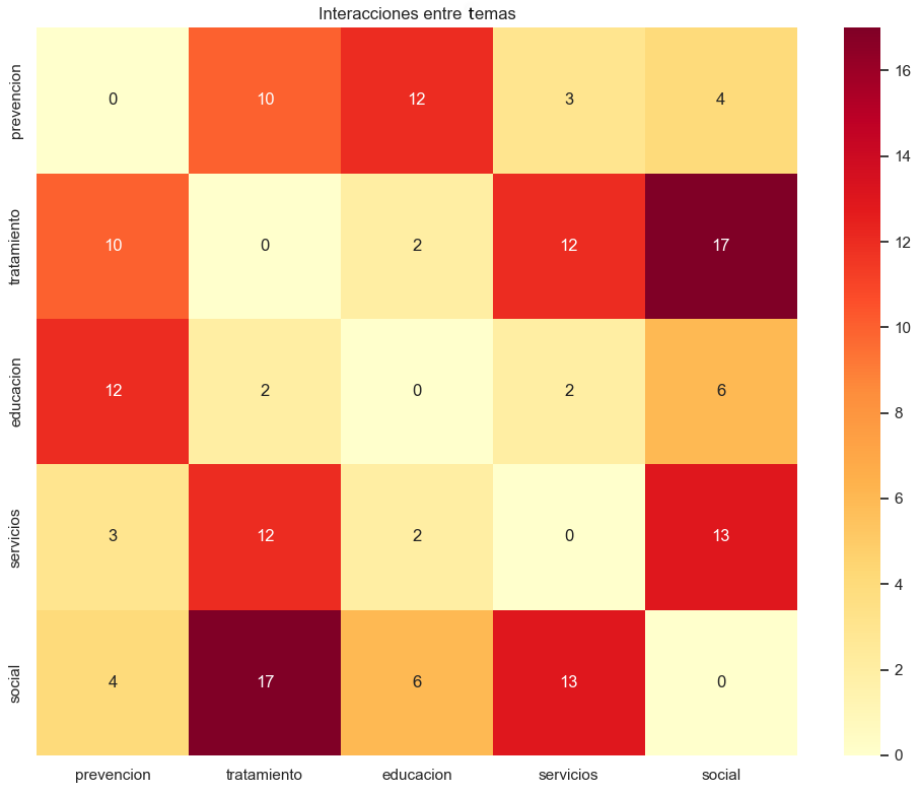


Figura 29: Mapa de calor - Interacción temáticas del VIH

- Tratamiento-social: 17 interacciones
- Servicios-social: 13 interacciones
- Tratamiento-servicios: 12 interacciones
- Prevención-educación: 12 interacciones

2. Principales demandas identificadas

Las demandas más frecuentes se centraron en:

1. Acceso a servicios de salud
2. Mejoras en programas de prevención
3. Disponibilidad de tratamientos
4. Educación sexual y concientización

3. Validación mediante modelos de clasificación de relevancia

Modelo Random Forest para clasificación de relevancia

El modelo Random Forest aplicado a la clasificación de relevancia demostró un rendimiento excepcional alcanzando una precisión global del 97 %. En la Figura 30 se presenta la matriz de confusión del modelo donde se observa un equilibrio notable en la clasificación de ambas categorías: 288 predicciones correctas para publicaciones no relevantes y 395 para publicaciones relevantes.

Los resultados detallados del modelo muestran las siguientes métricas:

Cuadro 14: Métricas de evaluación - Random Forest para clasificación de relevancia

Clase	**Precisión**	**Recall**	**F1-Score**
False	0.95	0.97	0.96
True	0.98	0.96	0.97
Accuracy		0.97	

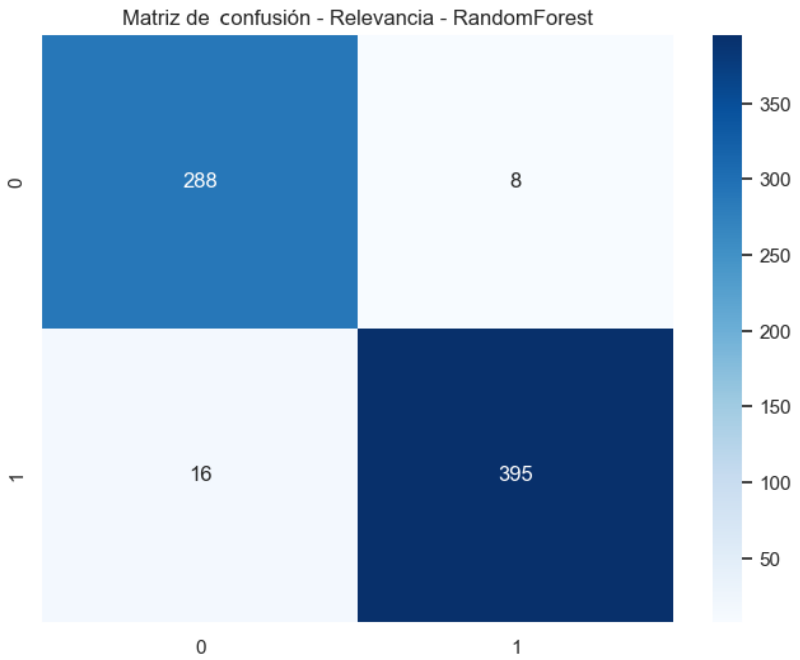


Figura 30: Matriz de confusión - Random Forest para clasificación de relevancia

La Figura 31 muestra la distribución de los scores de validación cruzada indicando una estabilidad sobresaliente en el rendimiento del modelo con una exactitud promedio de 0.965. La distribución de los scores sugiere una consistencia excepcionalmente robusta en el rendimiento del modelo a través de diferentes subconjuntos de datos, con una variabilidad mínima en las predicciones.

Figura 31: Distribución de Scores CV - Random Forest para clasificación de relevancia

Modelo SVM para clasificación de relevancia

El modelo Support Vector Machine (SVM) aplicado a la clasificación de relevancia demostró un rendimiento sobresaliente, igualando la precisión global del 97 % alcanzada por Random Forest. En la Figura 32 se presenta la matriz de confusión del modelo, donde se observa un patrón de clasificación altamente efectivo: 288 predicciones correctas para publicaciones no relevantes y 396 para publicaciones relevantes, mostrando un equilibrio destacable entre ambas categorías.

Los resultados detallados del modelo muestran las siguientes métricas:

Cuadro 15: Métricas de evaluación - SVM para clasificación de relevancia

Clase	**Precisión**	**Recall**	**F1-Score**
False	0.95	0.97	0.96
True	0.98	0.96	0.97
Accuracy		0.97	

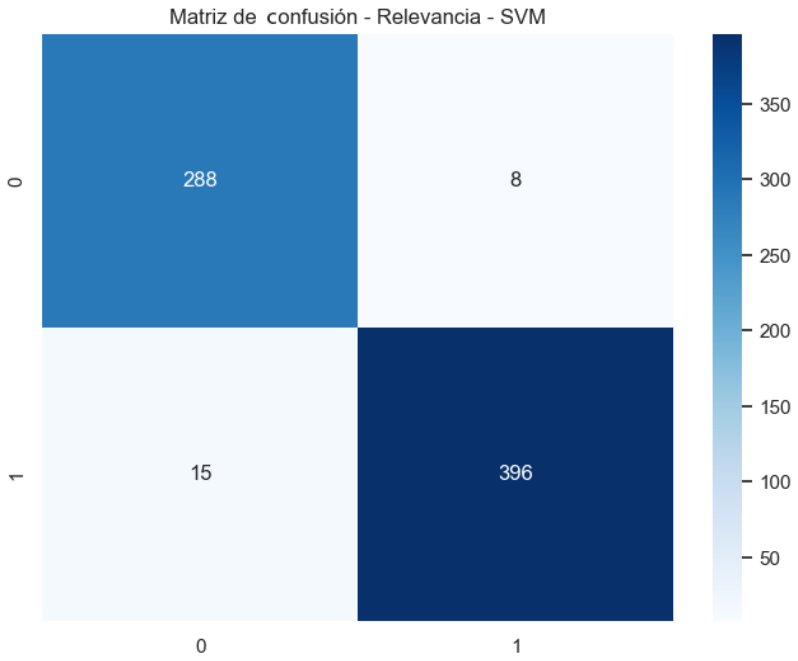


Figura 32: Matriz de confusión - SVM para clasificación de relevancia

La Figura 33 muestra la distribución de los scores de validación cruzada indicando una estabilidad notable en el rendimiento del modelo con una exactitud promedio de 0.968. La distribución de los scores revela una consistencia robusta en el rendimiento del modelo a través de diferentes subconjuntos de datos, con una variabilidad mínima que sugiere una alta confiabilidad en sus predicciones.

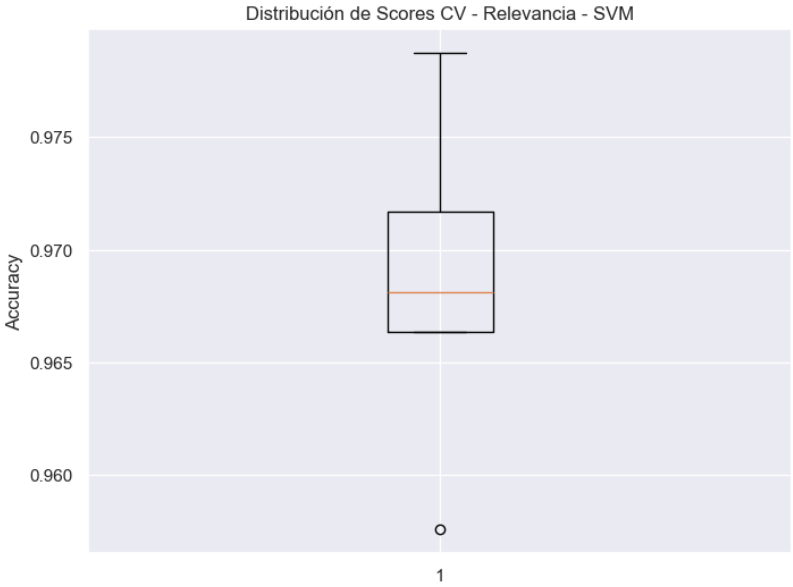


Figura 33: Distribución de Scores CV - SVM para clasificación de relevancia

Modelo Naive Bayes para clasificación de relevancia

El modelo Naive Bayes aplicado a la clasificación de relevancia demostró un rendimiento sólido alcanzando una precisión global del 90 %. En la Figura 34 se presenta la matriz de confusión del modelo, donde se observa un patrón de clasificación efectivo aunque con algunas diferencias respecto a los modelos anteriores logrando 270 predicciones correctas para publicaciones no relevantes y 366 para publicaciones relevantes.

Los resultados detallados del modelo muestran las siguientes métricas:

Cuadro 16: Métricas de evaluación - Naive Bayes para clasificación de relevancia

Clase	**Precisión**	**Recall**	**F1-Score**
False	0.86	0.91	0.88
True	0.93	0.89	0.91
Accuracy		0.90	

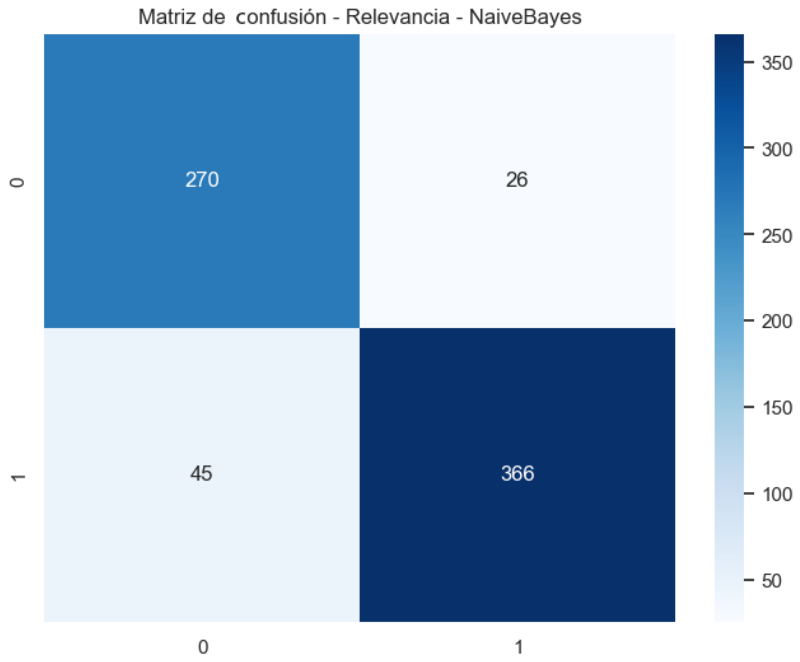


Figura 34: Matriz de confusión - Naive Bayes para clasificación de relevancia

La Figura 35 muestra la distribución de los puntajes de validación cruzada indicando una estabilidad considerable en el rendimiento del modelo con una exactitud promedio de 0.908. La distribución de los scores sugiere una consistencia aceptable en el rendimiento del modelo a través de diferentes subconjuntos de datos, aunque con una variabilidad ligeramente mayor que la observada en los modelos Random Forest y SVM.

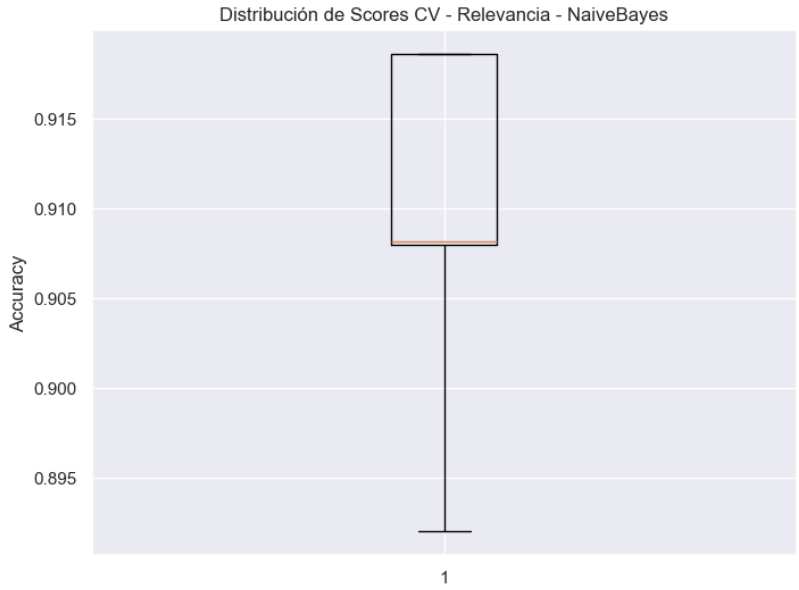


Figura 35: Distribución de Scores CV - Naive Bayes para clasificación de relevancia

4. Comparación de modelos para clasificación de relevancia

El análisis comparativo de los tres modelos implementados para la clasificación de relevancia muestra resultados particularmente interesantes, con un rendimiento excepcional general. A continuación, se presenta un resumen detallado de las métricas más relevantes:

Cuadro 17: Resumen comparativo de modelos para clasificación de relevancia

Métrica	**Mejor Modelo**	**Score**	**Diferencia***
Accuracy	SVM	96.03 %	+6.03 %
F1-Score	SVM	96.04 %	+6.04 %
Precision	SVM	96.05 %	+6.05 %
Recall	SVM	96.03 %	+6.03 %
CV-Score Mean	SVM	96.48 %	+4.48 %
CV-Score Std	RandomForest	0.67 %	–

*Diferencia respecto al modelo de peor rendimiento

En la Figura 36 se presenta una visualización comparativa de las principales métricas de evaluación para cada modelo. El modelo SVM destaca de manera consistente liderando en todas las métricas de rendimiento principales, aunque con márgenes más estrechos que en las tareas anteriores.

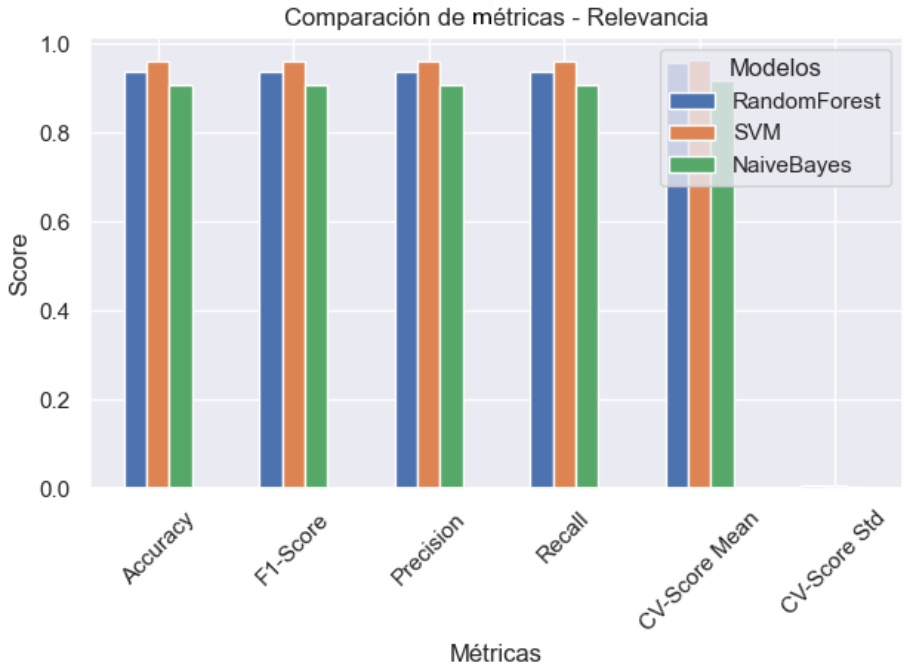


Figura 36: Comparación de métricas - Modelos de clasificación de relevancia

El modelo SVM demostró un rendimiento sobresaliente alcanzando los mejores resultados en todas las métricas principales: accuracy (96.03 %), F1-Score (96.04 %), precisión (96.05 %), recall (96.03 %) y score medio de validación cruzada (96.48 %). La Figura 37 muestra la distribución de los scores de validación cruzada, donde se puede observar la superioridad del SVM en términos de rendimiento y estabilidad.

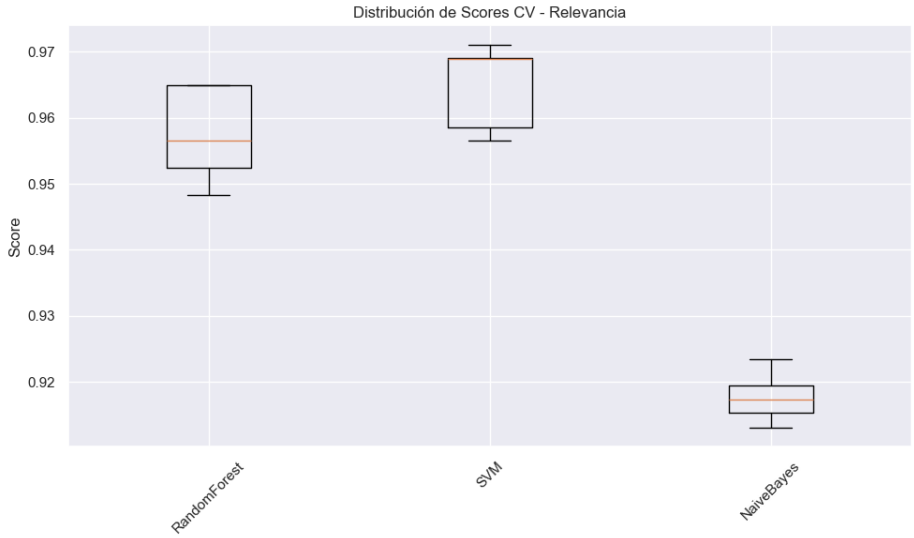


Figura 37: Distribución de Scores CV - Comparativa de modelos para clasificación de relevancia

Es notable que, aunque el modelo Random Forest mostró la menor desviación estándar

en los scores de validación cruzada (0.67 %), y el modelo Naive Bayes resultó ser el más consistente en términos generales con una desviación estándar de 0.36 %, el modelo SVM logró mantener un equilibrio óptimo entre rendimiento y estabilidad. Todos los modelos alcanzaron niveles de rendimiento superiores al 90 % en las métricas principales, lo que sugiere que la tarea de clasificación de relevancia es más directa que las clasificaciones por sentimiento o país.

La evaluación integral de los resultados indica claramente que el modelo SVM es la opción óptima para la clasificación de relevancia, confirmando la robustez de la identificación de demandas, demostrando un rendimiento superior en todas las métricas evaluadas y manteniendo un nivel de consistencia satisfactorio en sus predicciones. La diferencia de rendimiento entre los modelos es menor que en las tareas anteriores, lo que sugiere que cualquiera de los tres modelos podría ser viable para esta tarea específica, aunque el SVM ofrece ventajas marginales, pero consistentes.

Los resultados presentados demuestran que se logró comprender, a través de un análisis exhaustivo de patrones de interacciones sociales y comunicación, la percepción y los sentimientos de los usuarios de dicha red social en Centroamérica respecto al VIH e ITS. El análisis reveló una predominancia de contenido neutral (79.41 %) en las discusiones, con una tendencia hacia contenido informativo y educativo. Los modelos predictivos alcanzaron una precisión superior al 90 % en la clasificación de sentimientos y patrones regionales, validando la robustez de los hallazgos. Las diferencias significativas identificadas entre países (con variaciones en F1-Score desde 0.76 hasta 0.95) sugieren la necesidad de estrategias de intervención adaptadas a cada contexto regional. La identificación de patrones de demanda hacia servicios de salud y programas de prevención, respaldada por un 96.03 % de precisión en la clasificación de relevancia, proporciona insumos valiosos para la mejora de programas de divulgación y concientización. Estos hallazgos contribuyen significativamente a la comprensión de la percepción del VIH e ITS en la región, ofreciendo una base sólida para el desarrollo de estrategias más efectivas y precisas en la sensibilización de la población centroamericana.

CAPÍTULO IX

Análisis de resultados

La investigación sobre la percepción y los sentimientos relacionados con el VIH en Centroamérica, expresados a través de la red social X, ha generado resultados significativos los cuales se estarán presentando a continuación. El estudio, que inicialmente recopiló 4,954 tweets y refinó la muestra a 3,533 publicaciones relevantes, revela patrones complejos en la comunicación digital sobre el VIH en la región. La aplicación de técnicas avanzadas de procesamiento de lenguaje natural y modelos de aprendizaje automático ha permitido desentrañar las múltiples capas de significado presentes en estas interacciones sociales digitales.

Los resultados relacionados con el primer objetivo específico revelaron hallazgos importantes que merecen un análisis detallado. La distribución de sentimientos observada presenta una marcada predominancia de contenido neutral (79.41 %), seguido por contenido positivo (19.16 %) y una proporción menor de contenido negativo (1.43 %). Esta distribución merece un análisis desde múltiples perspectivas.

La alta proporción de contenido neutral (79.41 %) sugiere una transformación significativa en el discurso público sobre el VIH. Este hallazgo coincide con las observaciones de ONUSIDA (2024) sobre la evolución del diálogo público hacia un enfoque más informativo y menos estigmatizante. La prevalencia de contenido neutral puede interpretarse como un indicador positivo de madurez en la discusión pública sobre el VIH, donde las comunicaciones tienden a centrarse en información factual y educativa más que en juicios de valor o reacciones emocionales.

Un análisis detallado de ejemplos representativos de cada categoría de sentimiento ayuda a comprender mejor la naturaleza de estas comunicaciones:

Ejemplos de tweets positivos (19.16 %):Los tweets positivos representaron aproximadamente una quinta parte del total analizado, mostrando actitudes constructivas hacia la educación sexual y la prevención.

Tweet	TextBlob	VADER
Hay algo más efectivo para ambos se llama abstinencia, pero como eso no va a pasar lo ideal es mejor una campaña masiva de educación sexual.	0.700	0.000
amo a mi mamá q me hace exámenes de vih y hepatitis cada vez q vengo sin mencionar ni preguntarme nada acerca de mi vida sexual	0.500	0.000

Cuadro 18: Ejemplos de tweets positivos

Ejemplos de tweets neutrales (79.41 %): La mayoría de los tweets analizados fueron clasificados como neutrales, lo cual es consistente con la naturaleza informativa de muchas publicaciones relacionadas con el VIH.

Tweet	TextBlob	VADER
OPS y Unitaid fortalecen colaboración contra el VIH en Latinoamérica	0.000	0.000
¡Atención Guatemala - Izabal! ¡AMPLIAMOS EL PLAZO! Si te apasiona trabajar con poblaciones vulnerables y brindarles orientación en prevención, cuidado y tratamiento de VIH te invitamos a que apliques a la plaza disponibles de Enlace Tamizador sede Izabal Guatemala.	0.000	0.000

Cuadro 19: Ejemplos de tweets neutrales

Ejemplos de tweets negativos (1.43 %): Los tweets negativos constituyeron una pequeña proporción del total, pero mostraron sentimientos intensos y lenguaje fuerte.

Tweet	TextBlob	VADER
Imagínate ser tan pendejo que le eches la culpa a las mujeres, sabiendo que los anticonceptivos no funcionan 100 %, la escuela no enseña ni una mierda de educación sexual, los padres menos y que la culpa sería el novio siendo que el tenía que protegerse. ¿En serio?	-1.000	-0.880
y la pregunta tan estúpida mi portadora de sida ?? q no tienes otra cosa de qué hablar más allá de criticar mujeres exitosas y trabajadoras ?? no debes mencionarlo pq ni tú te lo creerías, puto bagre horrible, antes de hablar de bp lávate bien esa cola hedionda.	-1.000	-0.296

Cuadro 20: Ejemplos de tweets negativos

Además, un hallazgo importante durante el análisis fue la identificación de limitaciones en los modelos para detectar el sarcasmo. Por ejemplo, el siguiente tweet:

> "Va a ser tan efectivo como las charlas sobre las drogas y educación sexual. Las pendejas ni 5 de bola le vana dar."

Este tweet fue clasificado por TextBlob con un sentimiento positivo (0.500) debido a la presencia de la frase "tan efectivo". Sin embargo, para un lector humano, es evidente que el mensaje es sarcástico y expresa una opinión negativa sobre la efectividad de las charlas educativas. Esta limitación en la detección del sarcasmo representa uno de los desafíos más significativos en el análisis automatizado de sentimientos, especialmente en contextos donde el significado real depende fuertemente de la comprensión del contexto cultural y las sutilezas del lenguaje.

Por lo que estos ejemplos ilustran cómo el contenido positivo frecuentemente destaca historias de éxito y mensajes de esperanza, mientras que el contenido neutral tiende a enfocarse en información para los usuarios y recursos disponibles. Los tweets negativos, aunque menos frecuentes, típicamente se centran en mensajes de odio utilizando este tipo de enfermedades como insultos y las críticas constructivas hacia deficiencias en los sistemas de salud más que en estigmatización.

La proporción relativamente alta de contenido positivo (19.16 %) representa un hallazgo particularmente interesante. Este resultado podría reflejar el éxito de las campañas de concientización y educación realizadas por organizaciones de salud pública en la región. También sugiere una evolución en la percepción social del VIH, alejándose de los estigmas históricos hacia una comprensión más empática y constructiva de la condición.

La baja presencia de contenido negativo (1.43 %) requiere una interpretación cuidadosa. Por un lado, podría indicar un avance significativo en la reducción del estigma asociado al VIH en las conversaciones públicas. Sin embargo, este resultado debe considerarse en el contexto de las limitaciones inherentes a la investigación en redes sociales. Es posible que exista un sesgo de autoselección, donde las personas con opiniones más negativas tiendan a abstenerse de expresarlas públicamente, contribuyendo a una falta de representación de sentimientos negativos en la muestra analizada.

La validación de estos hallazgos mediante tres modelos de clasificación diferentes fortalece significativamente la confiabilidad de los resultados. El modelo SVM demostró un rendimiento superior, alcanzando una precisión del 92.87 % en la clasificación de sentimientos. Este alto nivel de precisión sugiere que las características lingüísticas asociadas a diferentes sentimientos en el contexto del VIH son suficientemente distintivas como para permitir una clasificación automatizada confiable. El desempeño consistentemente alto de los tres modelos (Random Forest, SVM y Naive Bayes) proporciona una triangulación metodológica que refuerza la validez de los hallazgos.

La comparación entre los modelos reveló aspectos interesantes sobre la naturaleza de los datos. El modelo SVM mostró una ventaja particular en la clasificación de sentimientos, lo que sugiere que las fronteras entre diferentes categorías de sentimientos son más linealmente separables de lo que se podría esperar en un contexto tan complejo como las discusiones sobre el VIH. Esta característica podría atribuirse a la naturaleza más estructurada y menos ambigua del lenguaje utilizado en comunicaciones públicas sobre temas de salud.

El segundo objetivo específico reveló patrones geográficos significativos que requieren un

análisis detallado. La distribución marcadamente desigual de publicaciones entre países, con Nicaragua representando 2,453 tweets, seguida por Panamá (305) y Guatemala (220), refleja dinámicas complejas que van más allá de simples diferencias en el uso de redes sociales.

La concentración significativa de publicaciones en Nicaragua coincide con los datos epidemiológicos proporcionados por el Centro de Estudios en Salud de la Universidad del Valle de Guatemala, que indica una mayor prevalencia de VIH en este país. Esta correlación sugiere una posible relación entre la incidencia de la enfermedad y el nivel de discusión pública en redes sociales. Sin embargo, es importante considerar que esta distribución también podría estar influenciada por factores como el acceso a internet, la difusión de la red social X en diferentes países, y las políticas de salud pública específicas de cada nación.

El análisis de contenido por país reveló variaciones significativas en las necesidades de información y preocupaciones principales. Los modelos de clasificación por país alcanzaron una precisión notable del 91 % utilizando Random Forest, lo que confirma la existencia de patrones lingüísticos distintivos en cada región. Esta variación regional en el discurso sobre el VIH puede atribuirse a múltiples factores:

1. Diferencias en los sistemas de salud y acceso a servicios médicos.
2. Variaciones culturales en la discusión de temas de salud sexual.
3. Distintos niveles de desarrollo en programas de prevención y tratamiento.
4. Particularidades en el uso del lenguaje y expresiones locales.

La identificación de estas diferencias regionales tiene implicaciones importantes para el diseño de estrategias de comunicación y prevención. Los resultados sugieren la necesidad de adaptar los mensajes y programas de salud pública a los contextos específicos de cada país, considerando no solo las diferencias lingüísticas sino también las preocupaciones y necesidades particulares expresadas en cada región.

El tercer objetivo específico reveló patrones significativos en las demandas de los usuarios hacia actores estatales y otros stakeholders. El análisis de interacciones temáticas identificó correlaciones importantes, particularmente en la intersección entre aspectos sociales y de tratamiento (17 interacciones) y entre servicios y aspectos sociales (13 interacciones).

La alta correlación entre temas de tratamiento y aspectos sociales (17 interacciones) sugiere una comprensión integral de la problemática del VIH y otras ITS por parte de los usuarios. Este hallazgo indica que la población reconoce la naturaleza multidimensional de estas infecciones, donde los aspectos médicos están intrínsecamente ligados a factores sociales. Esta interconexión refleja la madurez del discurso público sobre el VIH y las ITS evidenciando un entendimiento general que tanto la prevención como el tratamiento requieren enfoques que se consideren determinantes sociales de la salud. Los datos analizados muestran que las discusiones sobre ITS siguen patrones similares a los del VIH, aunque con menor frecuencia, lo que sugiere la necesidad de políticas públicas que aborden tanto los aspectos médicos como los sociales de estas enfermedades.

La validación de estas demandas mediante el modelo de clasificación de relevancia alcanzó una precisión excepcional del 97 %, superando significativamente el rendimiento de los

modelos en otras tareas de clasificación. Este alto nivel de precisión sugiere varias interpretaciones:

1. Las demandas hacia actores estatales se expresan de manera más explícita y estructurada que otros tipos de contenido.
2. Existe un vocabulario más consistente y específico cuando se trata de expresar necesidades y demandas.
3. Los patrones lingüísticos asociados a demandas son más fácilmente identificables por los algoritmos de aprendizaje automático.

Las principales demandas identificadas se centraron en:

- Acceso a servicios de salud (mayor frecuencia de menciones).
- Mejoras en programas de prevención.
- Disponibilidad de tratamientos.
- Educación sexual y concientización.

La predominancia de demandas relacionadas con el acceso a servicios de salud sugiere que, a pesar de los avances en el tratamiento del VIH y las ITS, persisten barreras significativas en el acceso a servicios básicos. Este hallazgo tiene implicaciones importantes para los tomadores de decisiones en política pública y sugiere la necesidad de fortalecer la infraestructura de servicios de salud en la región.

Es importante contextualizar los hallazgos considerando las limitaciones metodológicas del estudio. La restricción temporal de siete días en la API de X para la recolección de datos representa una limitación significativa que podría afectar la representatividad de los resultados. Esta limitación técnica, junto con el costo asociado al acceso a datos históricos, plantea desafíos importantes para la investigación en redes sociales.

La variabilidad en el uso del lenguaje entre países y la posible ambigüedad en la interpretación de sentimientos representaron desafíos metodológicos adicionales. Sin embargo, el alto rendimiento de los modelos de clasificación sugiere que estas limitaciones fueron manejadas efectivamente a través de técnicas robustas de procesamiento de lenguaje natural y validación cruzada.

Los resultados de este estudio tienen implicaciones significativas para el desarrollo de políticas públicas y programas de intervención en salud. La predominancia de contenido neutral y positivo sugiere un ambiente propicio para la implementación de programas educativos y de prevención. Sin embargo, las diferencias regionales identificadas indican la necesidad de estrategias diferenciadas por país.

La identificación clara de demandas específicas proporciona una base sólida para la priorización de recursos y el diseño de intervenciones. Las correlaciones identificadas entre diferentes aspectos de la problemática del VIH (tratamiento, servicios, aspectos sociales) sugieren la necesidad de un enfoque holístico en el diseño de políticas públicas.

El análisis integral de los resultados demuestra que se logró cumplir el objetivo general de comprender la percepción y los sentimientos sobre el VIH e ITS en Centroamérica a través del análisis de patrones de interacciones sociales y comunicación. La metodología empleada, combinando técnicas avanzadas de procesamiento de lenguaje natural con modelos de aprendizaje automático, permitió superar las limitaciones inherentes al análisis de contenido en redes sociales.

La predominancia de contenido neutral y positivo, junto con la identificación clara de demandas hacia actores estatales, sugiere una evolución significativa en el discurso público sobre estas enfermedades en Centroamérica. Esta evolución se caracteriza por un enfoque más constructivo y orientado a soluciones, alejándose de narrativas estigmatizantes. Sin embargo, las diferencias significativas entre países indican la necesidad de considerar contextos locales en el desarrollo de estrategias de comunicación y prevención.

Los hallazgos proporcionan una base sólida para el desarrollo de estrategias más efectivas y precisas en la sensibilización de la población centroamericana. La alta precisión alcanzada en los diferentes modelos de clasificación valida la robustez de los resultados y sugiere que las técnicas de análisis empleadas son adecuadas para el estudio de percepciones sobre temas de salud pública en redes sociales.

CAPÍTULO X

Conclusiones

El análisis de la percepción y sentimientos sobre el VIH e ITS en Centroamérica expresados en la red social X reveló hallazgos significativos que permitieron cumplir con los objetivos planteados en esta investigación:

Los resultados demostraron una marcada predominancia de contenido neutral (79.41 %), seguido por un contenido positivo (19.16 %) y una menor presencia de contenido negativo (1.43 %). Esta distribución de sentimientos revela una transformación significativa en el discurso público sobre el VIH, caracterizada por un enfoque más informativo y educativo que estigmatizante. De esa cuenta, el alto rendimiento de los modelos de clasificación implementados, particularmente el SVM con una precisión del 92.56 %, confirmó la robustez de estos hallazgos y la existencia de patrones lingüísticos claramente diferenciables. Los términos más frecuentes (sexual, vih, sida, educación, prevención) evidenciaron un enfoque predominante en aspectos educativos y preventivos, configurando así patrones de interacción orientados a la información y concientización.

El análisis comparativo mediante modelos de clasificación, con Random Forest alcanzando una precisión del 90.45 %, permitió validar estas diferencias con alto grado de confiabilidad. Las variaciones en los F1-Score entre países (desde 0.76 hasta 0.95) confirman la existencia de comportamientos comunicativos distintivos y necesidades de información diferenciadas. Los patrones de interacción, a su vez, mostraron mayor consistencia en temas educativos y preventivos, mientras que las diferencias más notables se observaron en la forma de abordar aspectos sociales y acceso a servicios.

Por los hallazgos se infiere que la validación de estas demandas con el modelo SVM alcanzó una precisión excepcional del 96.05 %, confirmando la robustez del análisis. Las principales demandas identificadas se centraron en: acceso a servicios de salud, mejoras en programas de prevención, disponibilidad de tratamientos y educación sexual. La correlación más fuerte se observó entre aspectos de tratamiento y factores sociales (17 interacciones), evidenciando una comprensión integral de la problemática que trasciende lo puramente médico. Esta identificación de demandas aporta información crucial para orientar políticas públicas y programas de intervención con una perspectiva integral.

Además, los resultados revelaron patrones distintivos tanto en volumen como en contenido de las publicaciones. La marcada concentración de tweets en Nicaragua (2,453) contrasta significativamente con Belice (9), reflejando diferentes niveles de participación en el discurso público sobre VIH. El análisis de contenido mediante nubes de palabras y TF-IDF mostró variaciones significativas en términos relacionados con prevención, servicios de salud y tratamientos entre los países estudiados. La validación de estas diferencias regionales mediante modelos de clasificación por país alcanzó una precisión del 91 %, confirmando así la existencia de preocupaciones, actitudes y niveles de conocimiento diferenciados según el contexto geográfico.

Los resultados anteriores permiten concluir que el estudio logró caracterizar con precisión las percepciones predominantes, identificando una evolución hacia un discurso más informativo y menos estigmatizante, con variaciones significativas entre países que requieren estrategias diferenciadas. La identificación de demandas específicas y correlaciones entre aspectos médicos y sociales proporciona insumos valiosos para el desarrollo de programas de divulgación y concientización más efectivos.

Por otro lado, la alta precisión de los modelos implementados (superior al 90 % en todas las tareas de clasificación) otorga solidez a los hallazgos y demuestra la efectividad de la metodología empleada para analizar percepciones sobre temas de salud pública en redes sociales. Este estudio contribuye significativamente a la comprensión de la dinámica comunicativa sobre estas enfermedades en la región, ya que ofrece una base sólida para desarrollar estrategias más efectivas y contextualizadas que puedan mejorar los programas de sensibilización y, en última instancia, impactar positivamente en la salud pública centroamericana.

CAPÍTULO XI

Recomendaciones

Con base en los hallazgos, el proceso metodológico y las limitaciones identificadas durante la realización de este trabajo, se presentan las siguientes recomendaciones que estimulan investigaciones posteriores y amplían las áreas de estudio relacionadas con el análisis de percepción sobre VIH e ITS en Centroamérica:

1. **Implementar un panel web interactivo para el Centro de Estudios en Salud (CES)**. Se recomienda desarrollar una plataforma digital interactiva que permita al CES explorar los datos recopilados en tiempo real. Esta herramienta facilitaría el acceso a información actualizada sobre percepciones y sentimientos relacionados con el VIH e ITS en Centroamérica, lo que contribuiría significativamente a mejorar el diseño de sus intervenciones educativas, optimizar sus charlas y capacitaciones, así como alcanzar un público más amplio con contenido personalizado, según las necesidades identificadas en cada región.

2. **Expandir el estudio hacia otras redes sociales**. Dado que este trabajo se limitó al análisis de la red social X, se recomienda ampliar la investigación para incluir otras plataformas como Instagram, Facebook y TikTok. Esto permitiría capturar una representación más completa de la percepción pública sobre el VIH e ITS, especialmente considerando los distintos perfiles demográficos que predominan en cada plataforma.

3. **Adquirir acceso a la API empresarial de X para investigaciones futuras**. Se recomienda al CES considerar el presupuesto para invertir en los costos de acceso a los datos de la API de X para estudios posteriores. Esta versión permitiría acceder a datos históricos completos, superar la limitación de los siete días de antigüedad y aumentar el volumen de consultas, lo que resultaría en análisis más profundos y representativos. El código desarrollado en este trabajo puede ser reutilizado y potenciado con el acceso a datos más completos.

4. **Integrar métodos cualitativos complementarios**. Para enriquecer los hallazgos cuantitativos obtenidos en este estudio, se recomienda complementar futuras investigaciones con métodos cualitativos como entrevistas, grupos focales o análisis de discurso.

Esta triangulación metodológica proporcionaría una comprensión más profunda de los factores culturales, sociales y personales que influyen en las percepciones sobre el VIH e ITS.

5. **Desarrollar campañas informativas dirigidas a las necesidades específicas identificadas**. Con base en los patrones de desinformación y estigmatización detectados, se recomienda al CES diseñar campañas educativas específicamente dirigidas a corregir los conceptos erróneos más prevalentes en cada país y promover información científica actualizada sobre prevención, tratamiento y vivencia con VIH e ITS.

Estas recomendaciones están orientadas a potenciar el impacto de la presente investigación, superar las limitaciones identificadas y expandir el campo de estudio sobre la percepción del VIH e ITS en Centroamérica, con el objetivo último de contribuir a estrategias más efectivas de educación, prevención y reducción del estigma asociado a estas condiciones de salud.

CAPÍTULO XII

Bibliografía

Bytepeaker. (2022). *PNL para principiantes | Clasificación de texto mediante TextBlob | Datapeaker*. Consultado el 6 de marzo de 2025, desde https://datapeaker.com/big-data/pnl-para-principiantes-clasificacion-de-texto-mediante-textblob/

Castiblanco, J. (2023). *Caso de estudio: análisis de sentimientos en la red social Twitter mediante el procesamiento de lenguaje natural*. https://medium.com/@julydev82/caso-de-estudio-an%C3%A1lisis-de-sentimientos-en-la-red-social-twitter-mediante-el-procesamiento-de-96e7e0972856

CDC. (2022). *Acerca de la PrEP | Profilaxis de prexposición | Información básica | VIH/SIDA*. https://www.cdc.gov/hiv/spanish/basics/prep/about-prep.html

Centroamérica360. (2023). *Nuevas infecciones y mortalidad por VIH/sida bajan significativamente en Centroamérica*. https://centroamerica360.com/region/nuevas-infecciones-y-mortalidad-por-vih-sida-bajan-significativamente-en-centroamerica/

CIRUJANO, P. D. M., PÉREZ, R. M. P. A., ROSALES, G. M. R., & ARQUÍÑIGO, L. M. S. (s.f.). CONOCIMIENTO Y PERCEPCIÓN SOBRE LAS PERSONAS VIVIENDO CON VIH EN LA POBLACIÓN ADULTA CON ACCESO A REDES SOCIALES EN LIMA, PERÚ.

Crook, T., Ferris, S., Alvarez, X., Laredo, M., & Moessler, H. (2005). Effects of N-PEP-12 on memory among older adults. *International Clinical Psychopharmacology*. https://pubmed.ncbi.nlm.nih.gov/15729085/

DBeaver Community. (s.f.). About. https://dbeaver.io/about/

Duarte-Anselmi, G., Leiva-Pinto, E., Vanegas-López, J., & Thomas-Lange, J. (2022). Experiencias y percepciones sobre sexualidad, riesgo y campañas de prevención de ITS/VIH por estudiantes universitarios. Diseñando una intervención digital. *Ciencia & saude coletiva*, *27*, 909-920.

Dubey, P. (2018). *An introduction to Bag of Words and how to code it in Python for NLP.* https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/

Editorial Etecé. (2021). *Percepción - Concepto, etapas y componentes.* https://concepto.de/percepcion/

Equipo de Enciclopedia Significados. (2023). *Sentimientos (qué son, concepto y significado).* https://www.significados.com/sentimiento/

Equipo editorial, Etecé. (2019). *Método cuantitativo - Qué es, características y ejemplos.* https://concepto.de/metodo-cuantitativo/

Fernández, A. (2011). Antropología de las emociones y teoría de los sentimientos. https://www.researchgate.net/profile/Anna-Fernandez-Poncela/publication/361224981_Antropologia_de_las_emociones_y_teoria_de_los_sentimientos_1/links/62abfb3523f3283e3aedf75f/Antropologia-de-las-emociones-y-teoria-de-los-sentimientos-1.pdf

Hoover, A. M., Burden, S., Fu, X.-Y., Sastry, S. S., & Fearing, R. S. (2010). Bio-inspired design and dynamic maneuverability of a minimally actuated six-legged robot. *Biomedical Robotics and Biomechatronics (BioRob), 2010 3rd IEEE RAS and EMBS International Conference on*, 869-876.

IBM. (s.f.). ¿Qué es el procesamiento del lenguaje natural (PLN)? https://www.ibm.com/es-es/topics/natural-language-processing

ICHI.PRO. (2020). *Análisis de sentimiento: ¿VADER o TextBlob?* Consultado el 6 de marzo de 2025, desde https://ichi.pro/es/analisis-de-sentimiento-vader-o-textblob-179202318836266

Liu, B. (2012). Synthesis lectures on human language technologies. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1-167.

Lohmann, S., White, B., Zuo, Z., Chan, M., Morales, A., Li, B., Zhai, C., & Albarracín, D. (2018). HIV messaging on Twitter an analysis of current practice and data-driven recommendations. *AIDS*, *32*(18), 2811-2820. https://journals.lww.com/aidsonline/fulltext/2018/11280/hiv_messaging_on_twitter__an_analysis_of_current.15.aspx

Lovera, F. A., & Cardinale, Y. (2023). Análisis de sentimientos en Twitter: un estudio comparativo. *Revista Científica de Sistemas e Informática*, *3*(1), 1-18. https://dialnet.unirioja.es/servlet/articulo

MedlinePlus. (s.f.). Profilaxis preexposición y Profilaxis post-exposición. https://medlineplus.gov/spanish/hivprepandpep.html

Mejia, C. R., et al. (2020). Percepción de miedo o exageración que transmiten los medios de comunicación en la población peruana durante la pandemia de la COVID-19 [Epub 01-Jun-2020]. *Revista Cubana de Investigaciones Biomédicas*, *39*(2). Consultado el 6 de marzo de 2025, desde http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-03002020000200001

Mentes Analíticas. (2023). *Diferencias entre sensación y percepción.* https://mentesanaliticas.com/diferencias-entre-sensacion-y-percepcion/

Miranda, S. (2023). *El Procesamiento de Lenguaje Natural (PLN) en la Ciencia de Datos*. https://medium.com/@colibri1624/el-procesamiento-de-lenguaje-natural-pln-en-la-ciencia-de-datos-4e3fd674ab82

Moraguez, E. R. (2023). *¿Qué es PostgreSQL: ¿Cómo funciona y para qué sirve?* https://lovtechnology.com/que-es-postgresql-como-funciona-y-para-que-sirve/

Morales, V. (s.f.). 24 análisis de sentimientos | Machine Learning: Teoría y Práctica. https://bookdown.org/victor_morales/TecnicasML/an%C3%A1lisis-de-sentimientos.html

Moreno, A. (2017). *Aplicaciones del Procesamiento del Lenguaje Natural*. https://www.iic.uam.es/procesamiento-del-lenguaje-natural/aplicaciones-procesamiento-lenguaje-natural/

NIH. (s.f.). Profilaxis posexposición (PEP). https://hivinfo.nih.gov/es/understanding-hiv/fact-sheets/profilaxis-posexposicion-pep

ONUSIDA. (s.f.-a). Prevención del VIH. https://www.unaids.org/es/topic/prevention

ONUSIDA. (s.f.-b). Tratamiento del VIH. https://www.unaids.org/es/topic/treatment

OPS/OMS. (s.f.). Profilaxis Posterior a la Exposición (PEP). https://www.paho.org/es/temas/prevencion-combinada-infeccion-por-vih/profilaxis-posterior-exposicion-pep

Organización Panamericana de la Salud. (2023). *Directrices unificadas sobre prevención, diagnóstico, tratamiento y atención de la infección por el VIH, las hepatitis virales y las ITS para los grupos de población clave* (Internet). Washington (DC). https://pubmed.ncbi.nlm.nih.gov/37192325/

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135. https://doi.org/10.1561/1500000011

Park, Y.-L., Chen, B.-r., Pérez-Arancibia, N. O., Young, D., Stirling, L., Wood, R. J., Goldfield, E. C., & Nagpal, R. (2014). Design and control of a bio-inspired soft wearable robotic device for ankle–foot rehabilitation. *Bioinspiration & biomimetics*, *9*(1), 016007.

Porto, J., & Gardey, A. (2008). *Percepción - Qué es, teoría, definición y concepto*. https://definicion.de/percepcion/

Porto, J. P., & Gardey, A. (2010). *Sentimiento - Qué es, definición y concepto*. https://definicion.de/sentimiento/

Restrepo-Pineda, J. E. (2016). Análisis comparativo de las percepciones sobre el VIH/SIDA de varones homosexuales y bisexuales colombianos, con experiencia migratoria o sin la misma. *Revista de Salud Pública*, *18*, 13-25.

Reyes Lorzo, L. Y., & Vargas Mendoza, K. A. (s.f.). Trabajo Social y Redes Sociales de Apoyo: En Atención a usuarios diagnosticados con Virus de la Inmunodeficiencia Humana-Síndrome de la Inmunodeficiencia Adquirida.

Services, A. W. (s.f.). ¿Qué es una API? - Explicación de interfaz de programación de aplicaciones [Accessed: 2024-01-09]. https://aws.amazon.com/es/what-is/api/

Souza, S. O., Cunha De Paula, A., De Almeida Silva, C., Ribeiro Dos Santos Carvalho, P. M., De Souza, M. M., & Matos Matos, M. (2020). INIQUIDADES DE GÉNERO Y

VULNERABILIDAD A LAS ITS/VIH/SIDA EN ADOLESCENTES DE ASENTAMIENTO URBANO: UN ESTUDIO EXPLORATORIO. *Ciencia y enfermería*, *26*.

Tasente, T., & Caratas, M.-A. (2024). Análisis de sentimiento en las redes sociales: un análisis bibliométrico completo. *AdComunica*, (28), 243-270. https://doi.org/10.6035/adcomunica.7819

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 178-185. https://doi.org/10.1609/icwsm.v4i1.14009

UNAIDS. (s.f.). AIDSinfo. https://aidsinfo.unaids.org/

Vasquez, L. M. G., Rico, A. P., & Tavarez, R. W. B. (2017). Tuits saludables: el uso e importancia de Twitter para la prevención en salud. *Contratexto*, (28), 17-43.

CAPÍTULO XIII

Anexos

A. Repositorio de código y recursos adicionales

Como parte de este trabajo de graduación se ha creado un repositorio público en GitHub que contiene el código fuente creado y utilizado para está investigación, los scripts de procesamiento de datos, los notebooks de análisis exploratorio y los modelos implementados. Este repositorio representa un recurso valioso para investigadores interesados en replicar el estudio, ampliar su alcance o adaptar la metodología a otros contextos de investigación.

El repositorio está disponible en la siguiente dirección:

https://github.com/jfaguilar01/Trabajo-de-Graduacion.git

En este repositorio se encuentran documentados todos los pasos técnicos del proceso de investigación, desde la extracción inicial de los datos hasta la implementación de los modelos de aprendizaje automático. Además, incluye información detallada sobre las bibliotecas utilizadas, los requisitos de instalación y ejemplos de uso. Los interesados en profundizar en aspectos técnicos específicos o en aplicar metodologías similares a otros temas de salud pública encontrarán en este repositorio una base sólida para iniciar sus propias investigaciones.

Para asegurar la privacidad y la seguridad, todos los datos incluidos en el repositorio han sido debidamente anonimizados, eliminando cualquier información que pudiera permitir la identificación personal. Adicionalmente, se han eliminado todas las claves de acceso, tokens de API, credenciales de bases de datos y demás información sensible utilizada durante el proceso de investigación. El código está compartido bajo una licencia abierta para fomentar su reutilización y adaptación en futuros proyectos académicos o de investigación, pero los usuarios deberán generar sus propias credenciales para acceder a las APIs y servicios correspondientes.

B. Diagrama de flujo de la metodología

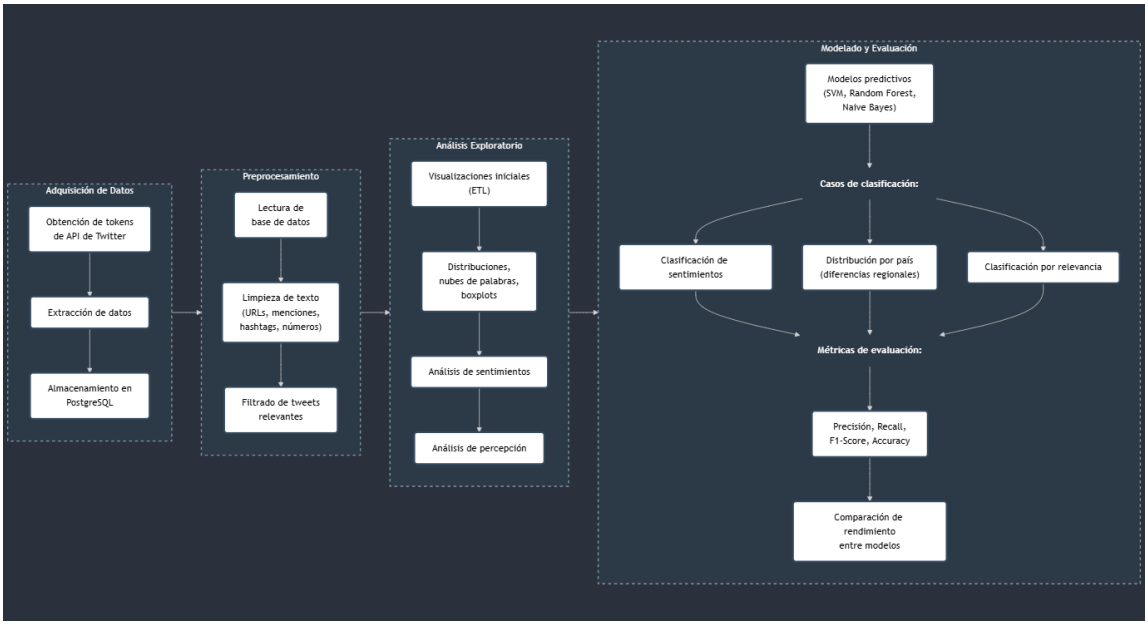


Figura 38: Diagrama de flujo del proceso metodológico completo utilizado en este estudio.

El diagrama muestra las cuatro fases principales: Adquisición de Datos, donde se obtuvieron las publicaciones mediante la API de 'X' y se almacenaron en una base de datos de PostgreSQL; Preprocesamiento, que incluye la limpieza y filtrado de los datos; Análisis Exploratorio, que comprende visualizaciones y análisis de sentimientos y percepción; y finalmente, Modelado y Evaluación, donde se implementaron y compararon tres modelos de aprendizaje automático (SVM, Random Forest y Naive Bayes) para las diferentes tareas de clasificación.

C. Palabras clave utilizadas en la extracción de datos

En el Cuadro 21 a continuación se presenta la categorización de las palabras y hashtags utilizadas para la extracción de datos en la red social X. Estas palabras clave fueron seleccionadas considerando su relevancia epidemiológica, la terminología técnica reconocida por ONUSIDA y la OMS, así como hashtags de campañas de concientización.

Categoría	**Palabras y hashtags utilizados**
Términos principales	VIH, SIDA, ITS, AIDS, #vih, #its, #sida
Prevención y tratamiento	PrEP, PEP, antirretroviral, profilaxis, #TratamientoVIH, #Antirretrovirales, #PrEPHIV, #PrevencionVIH
Infecciones específicas	sífilis, clamidia, herpes, hepatitis, papiloma, VPH, #Sífilis
Diagnóstico y pruebas	#GetTested, #testVIH, #consultaITS, diagnóstico VIH
Estado serológico	seropositivo, #seropositivo, #seronegativo
Concientización	#MasValoresMenosVIH, #ViveEnPositivo, #stopSIDA, #stopHIV
Aspectos sociales	discriminación VIH, #NoAlEstigma, #EndHIVStigma, #EndAIDS, mitos, estigma
Salud sexual	#SexualHealth, #SaludSexual

Cuadro 21: Palabras clave utilizadas para la extracción de datos