

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Implementación de una herramienta computacional para la caracterización de textos y perfiles relacionados con publicaciones sobre vacunación, en la red social Twitter en español

Jhohan Ricardo Franco Sánchez

Universidad Nacional de Colombia

Facultad de ingeniería, Departamento de ingeniería de sistemas y computación

Bogotá, Colombia

2021

Implementación de una herramienta computacional para la caracterización de textos y perfiles relacionados con publicaciones sobre vacunación, en la red social Twitter en español

Jhohan Ricardo Franco Sánchez

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

Magister en ingeniería de sistemas y computación

Director (a):

PhD. Luis Fernando Niño Vázquez

Línea de Investigación:

Sistemas inteligentes

Grupo de Investigación:

LISI

Universidad Nacional de Colombia

Facultad de ingeniería, Departamento de ingeniería de sistemas y computación

Bogotá, Colombia

2021

Dedicatoria

*A mis padres, los padres de mi esposa, esposa
y mi hija en camino. Gracias por el apoyo
incondicional.*

Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

Jhohan Ricardo Franco Sánchez

Fecha 09/04/2021

Fecha

Agradecimientos

Hay muchas personas y entidades a las que quisiera agradecer, sin embargo, sería una lista interminable, pues el presente trabajo se ha ejecutado con un gran esfuerzo y gracias a muchas colaboraciones, mi familia, esposa, amigos y compañeros de trabajo. Aun así, se comenzará agradeciendo al profesor Luis Fernando Niño, PhD, Profesor de la Universidad Nacional de Colombia, director del grupo de Investigación LISI y director del presente trabajo de grado: Gracias por su guía, paciencia y dedicación; así mismo, deseo agradecer a Camilo Pino, Diana Benavides y a todos los miembros del grupo de investigación LISI, pues, con sus consejos, dedicación y tiempo se logró elaborar tan excepcional trabajo.

Por otro lado, agradecer a la facultad de ingeniería de la Universidad Nacional de Colombia, me honra pertenecer a esta, mi alma mater, pues gracias a sus aulas, docentes y estudiantes, mis ganas de seguir trabajando por un país mejor son cada vez más fuertes.

Por último, gracias a mis padres, hermano, familia y esposa, nuevamente, sin ustedes nada sería posible.

Resumen

Implementación de una herramienta computacional para la caracterización de textos y perfiles relacionados con publicaciones sobre vacunación, en la red social Twitter en español

La salud pública y la reticencia a la vacunación, como base problemática del presente trabajo, representan el marco de discusión en épocas de pandemia: virus, bacterias y demás agentes que afectan la salud de los humanos y su entorno. El eje principal que se abordará, se encuentra ligado al apoyo de la salud pública para establecer políticas que busquen garantizar integralmente la salud de la población, por medio de acciones dirigidas, tanto individual como colectivamente; el foco del presente estudio está centrado en la vacunación y publicaciones relacionadas en español en la red social Twitter, por medio del análisis de discurso y la detección de patrones basados en los tweets (trinos), que identifiquen los diferentes perfiles de interés en la red social.

Realizando trabajos de encuesta y obtención de datos, algunos estudios encontraron que la mayoría de las personas usa internet y, específicamente, las redes sociales, para consultar temas de salud, dejando en claro que el 95.8% ha buscado información de salud en Internet y 44.4% temas de vacunación [1], sin consultar fuentes oficiales de esta última, amenazando potencialmente la inmunidad grupal a cierto tipo de enfermedades curables por medio de las vacunas.

Se logró contribuir a suplir la necesidad que existe para la automatización sobre la identificación de características y comportamientos de los diferentes grupos (Pro-vacunación, Anti-vacunación y Neutral) sobre la red social Twitter en español, en términos de salud pública y vacunación, facilitando el análisis de los expertos que fundamentan políticas públicas y regulaciones en pro del correcto manejo de la información.

Palabras clave: Twitter, Vacunas, Salud pública, Automatización, Redes Sociales, Analítica de datos, Sistemas inteligentes.

Abstract

Implementation of a computational tool for the characterization of texts and profiles related to publications on vaccination, in the social network Twitter in Spanish

Public health and reluctance to vaccination, as the problematic basis of this work, represent the framework for discussion in times of pandemic: viruses, bacteria and other agents that affect the health of humans and their environment. The main problem to be addressed is linked to the support of public health to establish policies that seek to fully guarantee the health of the population, through directed actions, both individually and collectively; The focus of this study is centered on vaccination and the social network Twitter in Spanish, through discourse analysis and the detection of patterns based on tweets, which identify the different profiles in the social network.

Carrying out survey work and obtaining data, some studies found that the majority of people use the internet and, specifically, social networks, to consult health issues, making it clear that 95.8% have searched for health information on the Internet and 44.4 % vaccination issues, without consulting official sources of the latter, potentially threatening group immunity to certain types of diseases curable by means of vaccines.

It was possible to contribute to supply the need that exists for automation on the identification of characteristics and behaviors of the different groups (Pro-vaccination, Anti-vaccination and Neutral) on the social network Twitter in Spanish, in terms of public health and vaccination, facilitating the analysis of the experts who develop public policies and regulations in favor of the correct handling of information.

Keywords: Twitter, Vaccines, Public health, Automation, Social Networks, Analytics, Intelligent systems.

XII Implementación de una herramienta computacional para la caracterización de textos y perfiles relacionados con publicaciones sobre vacunación, en la red social Twitter en español

Este Trabajo Final de maestría fue calificado en Junio de 2021 por el siguiente evaluador:

Juan David García Arteaga PhD.
Profesor Departamento de Imágenes Diagnósticas
Facultad de Medicina
Universidad Nacional de Colombia

Contenido

	Pág.
Contenido	
Resumen	IX
Abstract	XI
Lista de ilustraciones	XV
Lista de tablas	XVI
Lista de Símbolos y abreviaturas	XVII
Introducción	1
Objetivo general	6
Objetivos específicos	6
1 Salud pública y reticencia a la vacunación	7
1.1 Reticencia a la vacunación desde las redes sociales	9
2 Análisis de redes sociales y procesamiento de lenguaje natural	12
2.1 Análisis de redes sociales	12
2.2 Análisis de contenido	13
2.3 Procesamiento de lenguaje natural	13
3 Metodología propuesta	15
3.1.1 Arquitectura de la solución.....	16
3.2 Actividades desarrolladas.....	18
4 Preparación del conjunto de datos	20
4.1 Descripción de las fuentes de datos	20
4.2 Características de los datos	23
4.2.1 Calidad y disponibilidad de los datos	24
4.2.2 Limpieza de los datos	25
5 Modelamiento de textos para la caracterización de textos sobre vacunación ..	27
5.1 Modelamiento por Latent Dirichlet Allocation (LDA).....	27
5.2 Reconocimiento de Entidades Nombradas (NER)	29
5.3 Visualización de datos por T-SNE	32

6	Visualización e interpretación.....	35
7	Diseño e implementación de la herramienta computacional.....	41
7.1	Evaluación cualitativa preliminar de la herramienta.....	43
7.2	Vistas y funcionalidades de la herramienta.....	43
7.2.1	Tablero principal.....	43
7.2.2	Reconocimiento de entidades (NER).....	44
7.2.3	Asignación Latente de Dirichlet (LDA).....	46
7.2.4	T-SNE.....	47
8	Conclusiones y recomendaciones.....	49
	Conclusiones.....	49
	Recomendaciones.....	50
	Bibliografía.....	56

Lista de ilustraciones

	Pág.
Ilustración 1 Citas por Tweet ante perfiles pro-vacunas y anti-vacunas [5].....	3
Ilustración 2 Atributos de contenido [23].....	10
Ilustración 3 Atributos de diseño [23].....	11
Ilustración 4 Arquitectura metodológica.....	15
Ilustración 5 Fase 1 de la solución.....	17
Ilustración 6. Fase 2 de la solución.....	17
Ilustración 7. Fase 3 de la solución.....	18
Ilustración 8: Ejemplo de formato JSON sobre los datos descargados de Twitter	20
Ilustración 9 Archivos descargados de los perfiles de Twitter	20
Ilustración 10 Verificación de Twitter vs datos descargados.....	24
Ilustración 11 Diagrama de flujo del proceso de limpieza de datos.....	25
Ilustración 12 Modelo gráfico de LDA: En la caja exterior se representan los documentos M, en la caja interior N se representan los temas y palabras dentro de un documento.[31]	28
Ilustración 13 Estructura básica de un perceptrón	31
Ilustración 14. Ejemplo arquitectura Reconocimiento de entidades Nombradas	32
Ilustración 15 Vista de interacciones en la red social.....	35
Ilustración 16 Vista de recuperación de usuario	36
Ilustración 17 Vista de recuperación de tweet	36
Ilustración 18 Resultados de la evaluación de texto seleccionado para ejecutar LDA	37
Ilustración 19 Vista de resultado de Reconocimiento de Entidades Nombradas (NER) ..	38
Ilustración 20 Bolsa de palabras de los perfiles.....	38
Ilustración 21 Bolsa de palabras del conjunto de Tweets base.....	39
Ilustración 22 Tópicos por texto ingresado	39
Ilustración 23 Contadores bases de datos.....	40
Ilustración 24 Vista global herramientas y lenguajes utilizados.....	41
Ilustración 25 Arquitectura y librerías.....	42
Ilustración 26 Home aplicación Web.....	44
Ilustración 27 Vista del reconocimiento de entidades en la aplicación Web.....	45
Ilustración 28 Vista de Asignación latente de Dirichlet (LDA).....	46
Ilustración 29 Grafica de T-SNE para 10000 documentos con una perplejidad de 50	47

Lista de tablas

	Pág.
Tabla 1 Resultados de la regresión evaluando los posibles factores asociados al rechazo de las vacunas por parte de los padres que respondieron a la encuesta. [1]	8
Tabla 2 Estructura objeto Tweet	21
Tabla 3 Estructura objeto User.....	22
Tabla 4 Estructura objeto Entities.....	23
Tabla 5 Etiquetas de entidades conocidas	30
Tabla 6 Representación vector de documentos de alta dimensionalidad	47
Tabla 7 Representación de vector de documentos de baja dimensionalidad	48

Lista de Símbolos y abreviaturas

Abreviaturas

Abreviatura	Término
NLP	Natural Language Processing
PLN	Procesamiento de Lenguaje Natural
POS	Part of Spech
Tweet	Post o publicación en la red social Twitter
LDA	Latent Dirichlet allocation - Asignación de Dirichlet latente
NER	Named Entity Recognition – Reconocimiento de entidades nombradas
D3	Data-Driven Documents
API	Interfaz de programación de aplicaciones
t-SNE	t-distributed stochastic neighbor embedding

Introducción

Acorde a las búsquedas realizadas sobre bases de datos especializadas y revistas indexadas, sobre temas de ingeniería y salud pública (en las bases de datos ScienceDirect, Scopus, Elsevier y Scielo de salud pública), se puede decir que se habla de salud pública orientada a las vacunas desde su primera aparición en el siglo XVIII con la creación de la vacuna para la viruela. Sin embargo, la presente investigación tiene relevancia por lo ocurrido desde el siglo XXI con el nacimiento de la Web 2.0 como fenómeno social de la conectividad entre usuarios de internet; especificando, para el 2006, la creación de Twitter como red social.

En concordancia con lo mencionado anteriormente, se puede decir que la mayor cantidad de trabajos son relativamente recientes, menores a 15 años. En términos de la unión entre dos áreas de interés, la salud pública y las redes sociales, se encuentra literatura del 2004 hasta el 2020, exaltando trabajos como el realizado en “Internet y vacunas: análisis de su uso por padres de familia, sus percepciones y asociaciones” [1], en el cual se dice:

“(…) Complicando las cosas, la personalización que servicios como Facebook pueden crear burbujas ideológicas, de tal manera que a un usuario puede aparecerle sólo información acorde a su punto de vista, sean estos correctos o erróneos. Con esto en mente, presentamos los resultados de una encuesta, cuyos objetivos fueron conocer el estado de confianza hacia las vacunas en nuestra región y explorar si existe asociación entre el uso de internet y redes sociales y las actitudes de rechazo hacia las mismas (…)”[1].

Realizando trabajos de encuesta y obtención de datos por medio de las percepciones del investigador, los resultados encontraron que la mayoría de las personas usa Internet y, específicamente, las redes sociales para consultar temas de salud, dejando en claro que el 95.8% ha buscado información de salud en internet y 44.4% temas de vacunación [1],

sin consultar fuentes oficiales de esta última, amenazando potencialmente la inmunidad grupal a cierto tipo de enfermedades curables por medio de las vacunas.

De igual forma, se encuentra en otras investigaciones diversos motivos de no vacunación, este es el caso, donde se señala que en los países en vías desarrollo, las poblaciones no tienen un fácil acceso a la información, por lo tanto, su racionalidad, en términos generales, no se encuentra familiarizada con los modelos biomédicos, llevado a concluir que las costumbres, las ideas y la “ignorancia” de las poblaciones operan como barreras para la participación [2]. Sobre la misma línea de investigación y con un enfoque de investigación mixta, encontramos que para un estudio del 2000 al 2015 evaluando los programas de vacunación infantil en América Latina, se presentaron resultados de ineficiencia vacunal del 1% al 23%, dependiendo de la realidad de cada país [3], lo cual exalta la importancia de generar políticas públicas y mayores facilidades para adquirir una inmunidad de rebaño en términos de vacunación sobre enfermedades curables.

Para el caso específico colombiano, no se encontraron estudios investigativos cuyo enfoque sean las vacunas y las redes sociales, específicamente Twitter en español, así mismo, tampoco se evidenciaron estudios sobre salud pública orientada a la reticencia de la vacunación, desde un aspecto técnico de analítica de datos en redes sociales. En 2017 se publicó “Motivos de no vacunación en menores de cinco años en cuatro ciudades colombianas” [4], develando diferentes factores que restringen los hábitos de vacunación: desconocimiento a las reacciones pos-vacunales, condiciones socioeconómicas, geográficas, condiciones laborales del personal de vacunación, problemas administrativos, económicos o el mismo desarrollo regional, ante el olvido estatal sobre precarios sistemas de información.

Por otro lado, se encontraron una gran cantidad de investigaciones relacionadas con la extracción, transformación y carga (ETL, por su sigla en inglés) de datos sobre redes sociales; continuando con el proceso de limpieza, análisis y caracterización de textos, especialmente en inglés, sobre los conjuntos de datos, entre ellos se encontró: “Twitter message types, health beliefs, and vaccine attitudes during the 2015 measles outbreak in California” [5] donde ejecutan un análisis de texto de 3000 tweets sobre 1 millón capturados, evidenciando la cantidad de mensajes relacionados con experiencias

personales, experiencias interpersonales, lecturas científicas, recursos sobre fuentes no certificadas, etc. (ver Ilustración 1, recordar que el p ajustado de 0.003 representa la significancia a nivel estadístico de los resultados) tomando estos resultados como un insumo relevante para el presente trabajo, ya que, a pesar de tener una labor manual de etiquetado de los datos, resulta ser una caracterización acertada en los dos tipos de perfiles, pro-vacunación y anti-vacunación. Del mismo modo, se delimitan aplicaciones específicas sobre países desarrollados (Italia, USA, España y Francia) [6] [7] [8] [9]. Aun así, no se encuentran textos relevantes para Colombia o textos en español.

Ilustración 1 Citas por Tweet ante perfiles pro-vacunas y anti-vacunas [5]

Differences in attitudes toward vaccination in terms of cited media sources

	News agency/ newswire service (n = 18)	Television network (n = 10)	Newspaper (n = 173)	Radio (n = 18)	Magazine (n = 89)	News website/ blog (n = 344)	SNS (n = 43)	Other (n = 138)	No media link (n = 329)
Pro-vaccine tweets (N = 1076), n (%)	17 (1.6)	102 (9.5)	168 (15.6)	17 (1.6)	89 (8.3)	266 (24.7)	33 (3.1)	104 (9.7)	280 (26)
Anti-vaccine tweets (N = 186), n (%)	1 (0.5)	8 (4.3)	5 (2.7)	1 (0.5)	0 (0)	78 (41.9)	10 (5.4)	34 (18.3)	49 (26.3)
Adjusted P	.27	.02	<.003*	.27	<.003*	<.003*	.11	<.003*	p = .92

SNS, Social Networking Site.

*Significant results with an adjusted P value of .003.

Trabajos previos sobre salud pública, específicamente de vacunación, han demostrado que existen factores predisponentes para la emergencia y reemergencia de enfermedades infecciosas, que se vienen presentando históricamente debido a situaciones ecológicas, cambios adaptativos de la susceptibilidad de los humanos, cambios microbianos, climáticos, demográficos, de comercio, turismo, desarrollo tecnológico, desarrollo industrial, pobreza e inequidad, conflicto, migraciones, e incluso ineficientes políticas de salud en los países en vías de desarrollo [10].

No obstante, la información para educar a la comunidad sobre los beneficios o los potenciales riesgos sobre las vacunas se ha convertido en un problema de salud pública [11]. Se tiene como objeto de estudio, en este trabajo, las redes sociales, partiendo del hecho de que los grupos anti-vacunación son una de las causas principales de resistencia hacia la vacunación, conectando la gente con la influencia de comentarios sobre la seguridad de las vacunas, la falta de información adecuada y la percepción de que no son eficaces o necesarias [12].

Los temas de salud pública están ganando territorio en las agendas de los países, dado que cada vez surgen más enfermedades que amenazan la tranquilidad de las

comunidades, generando pérdidas económicas, sociales e incluso políticas; son estos escenarios donde surge la necesidad de reducir las brechas en la cobertura de vacunación, lo que hace necesario establecer escenarios de diálogo entre los nuevos movimientos de pseudociencias – anti-vacunación y las ciencias modernas – pro-vacunación, evitando que algunas enfermedades eliminadas o erradicadas puedan volver a ser endémicas [13]. Conforme a esta situación, los cibernautas de las redes sociales o internet pueden adquirir Información errónea o inexacta sobre los riesgos y beneficios de las vacunas [14].

“Las redes sociales trabajan como un poderoso diseminador de información, como una plataforma abierta para grupos en contra de la vacunación; y el impacto de estas publicaciones aún no han sido medidos ni valorados; con el agravante de la personalización de servicios que proveen plataformas como Facebook; que puede crear burbujas ideológicas, de tal manera que a un usuario pueda aparecerle sólo información acorde a su punto de vista, sean estos correctos o erróneos. Representan un área del conocimiento poco estudiada para los salubristas que requieren trabajar en estrategias para contrarrestar miedos y mitos en torno a las vacunas, mediante la evidencia científica de hechos contra la desinformación.” [11]

Durante el presente trabajo, se busca suplir la necesidad de usar herramientas tecnológicas que automaticen la caracterización (tanto de los perfiles como de los textos) y develen aspectos relevantes del comportamiento de los diferentes grupos (Pro-vacunación, Anti-vacunación y Neutrales), dentro del contexto de la salud pública y la vacunación [13]. El proyecto se encontrará fundamentado en el estudio de la red social Twitter, ya que esta, desde su creación en el 2006, se ha vuelto fuente vital de micro información, dado que es una fuente gratuita de acceso a la información, de rápida transmisión en una comunidad amplia y que puede llegar a proveer comunicación directa con el publicador de la información [15], lo que abre la puerta a fuentes confiables de información correcta, o también a fuentes poco fiables de información.

Un aspecto relevante que justifica el presente estudio es la inexistencia de implementaciones y tecnologías enfocadas al análisis de las interacciones de la red social, en términos de diseminadores de información sobre vacunación. Se lograron identificar algunas investigaciones realizadas sobre la red social Twitter, los cuales, en su mayoría,

se centran en los idiomas inglés, francés y alemán, abordados desde una perspectiva espacio temporal (sobre un conjunto de datos, en rangos de tiempo específicos, se hace el análisis puntual). Nuestro objetivo es generar una herramienta tecnológica que identifique de manera continua dichos comportamientos y de al investigador herramientas que lo ayuden a comprender las interacciones de la red social en términos de información sobre vacunación.

Objetivo general

Implementar una herramienta computacional para la caracterización de los textos y perfiles de usuarios de la red social Twitter en español, en la divulgación de información referente a vacunación y la identificación de roles pro-vacunación, anti-vacunación y neutrales.

Objetivos específicos

- Aplicar las herramientas disponibles en la actualidad, para realizar el proceso de descarga, limpieza y análisis exploratorio de un conjunto de textos (en español) sobre la red social Twitter.
- Integrar las herramientas disponibles, en pro de ejecutar las acciones necesarias de configuración e implementación sobre la caracterización de roles y clasificación de textos en español.
- Aplicar un método de clasificación de tópicos jerárquica para para la evaluación y clasificación de los textos sobre el conjunto de publicaciones de Twitter.
- Aplicar un método de aprendizaje no supervisado para la categorización de las clases y elementos clave que caractericen los perfiles pro-vacunación, anti-vacunación y neutrales.

1 Salud pública y reticencia a la vacunación

Para tener un contexto sobre salud pública, es necesario establecer un punto de partida, ya que este campo contiene en sí mismo una gran cantidad de variantes por país, por gobierno y por entidades; se puede encontrar su definición para Colombia en la ley 1122 de 2007 dispuesta por el congreso de la república:

“Artículo 32º. De la salud Pública. La salud pública está constituida por el conjunto de políticas que buscan garantizar de una manera integrada, la salud de la población por medio de acciones de salubridad dirigidas tanto de manera individual como colectiva, ya que sus resultados se constituyen en indicadores de las condiciones de vida, bienestar y desarrollo del país. Dichas acciones se realizarán bajo la rectoría del Estado y deberán promover la participación responsable de todos los sectores de la comunidad.” [16]

Dada la anterior definición, es necesario continuar exponiendo los autores relevantes sobre la temática, orientados a salud pública en términos de vacunación y las redes sociales; Ubaldo Cuesta Cambra, entre otros colaboradores, han ejecutado una serie de estudios relevantes al ámbito español, relacionados a la reticencia a la vacunación y la afectación de esta en la salud pública, arrojando unos resultados que identifican que la gente está preocupada por la influencia de las redes sociales en las tendencias mundiales de vacunación; en su texto, donde habla desde una perspectiva psicosocial, ahondando en las tendencias de los medios tecnológicos y las TIC, Tecnologías de la Información y Comunicaciones [17] como medios predisponentes para dichos escenarios. Así mismo, toman relevancia temas como la "reputación online" de la información de vacunas en Internet [18], vacunas y anti-vacunas en la red social YouTube [19], Comunicación 2.0 y salud pública: social networking [20], Análisis de información pro vacuna y anti vacuna en redes sociales e Internet: patrones visuales y emocionales [12]. Todos los trabajos previos,

ejecutados por medio de análisis del discurso mediante métodos de análisis descriptivo, entrevistas, variables cuantitativas y cualitativas.

Como referente en estudios sobre salud pública y vacunación, se encuentra, así mismo, a una investigadora de la Universidad Vasca/Euskal Herriko Unibertsitatea, Carmen Peñafiel Saiz y colaboradores, en una publicación del 2019 sobre la desconfianza de los medios y las vacunas para el análisis de contenido en titulares de noticias, con el fin de verificar los informes de los medios de comunicación sobre las vacunas y determinar las características clave, el contexto y tono (positivo, negativo o neutro) ejecutando técnicas de análisis de contenido relacionado con vacunas [21].

Pasando al contexto latinoamericano, en México, desde el Hospital pediátrico de Sinaloa, se realizó un estudio transversal que analizó las opiniones y asociaciones de los padres de familia sobre la relación entre el Internet y las vacunas; dando como resultado que la mayoría de las personas utilizan Internet y las redes sociales para informarse sobre temas de salud, identificando un alto porcentaje de personas que se niegan a vacunarse basados en información no científica, escasamente comprobada o rumores sin bases científicas sólidas; concluyendo una posible amenaza a la inmunidad de rebaño de la ciudad, junto con una relación directa entre el uso de las redes sociales y la asistencia en escuelas privadas [1] (ver Tabla 1).

Tabla 1 Resultados de la regresión evaluando los posibles factores asociados al rechazo de las vacunas por parte de los padres que respondieron a la encuesta. [1]

<i>Factor</i>	<i>Razón de momios</i>	<i>IC95%</i>
Acudir a un jardín de niños privado	2.48	1.22 a 5.06
Buscar información sobre vacunas en internet	1.86	1.09 a 3.17
Presionar "Me Gusta" en artículos de salud o nutrición en Facebook	2.51	1.03 a 6.11
Leer completos los artículos de salud en Facebook antes de compartirlos	0.68	0.32 a 1.46
Leer los comentarios de otras personas en artículos de salud en internet	1.80	0.84 a 3.83

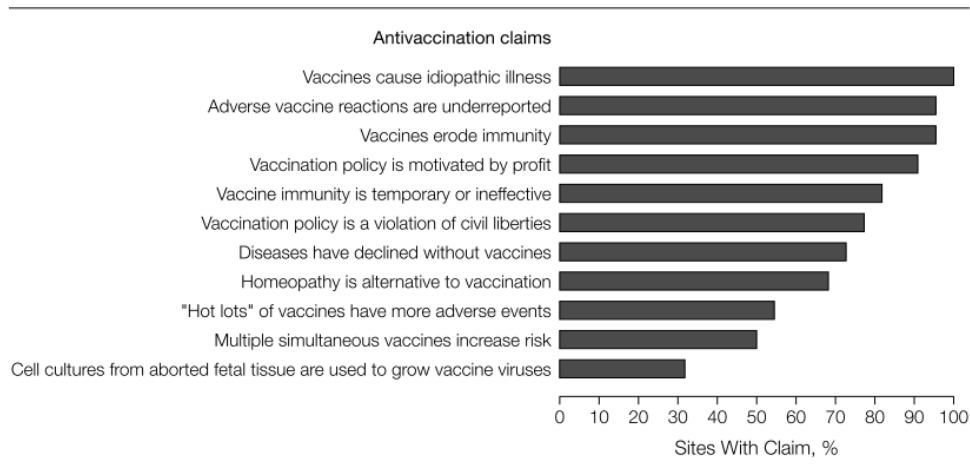
Por último, se tiene el contexto colombiano, en el cual se identificaron estudios sobre encuestas que relacionan la falta de voluntad para vacunarse; en el 2018, se ejecutó una encuesta dirigida por Fabio Escobar Díaz, May Bibiana Osorio Merchán y Fernando De la Hoz, en cuatro ciudades colombianas y dirigida a niños menores de 5 años que no estaban vacunados, los principales factores encontrados incluyen: temor a reacciones pos vacunación, estrato socioeconómico, geográfico y seguridad de la población, e incluso las condiciones laborales de los vacunadores, adicional, problemas administrativos, económicos y el desarrollo inestable de los sistemas de información [4].

Como se puede evidenciar, existe una problemática clara desde el contexto de la salud pública, las redes sociales y su relación con la vacunación; en esta era de la Web 2.0, donde un porcentaje importante de la población mundial se encuentra interconectada, con información y desinformación viajando a miles de kilómetros por segundo, causando pánico entre los agnósticos de las vacunas, incertidumbre entre aquellos cuya posición es neutral y preocupación para los que creen en las ciencias.

1.1 Reticencia a la vacunación desde las redes sociales

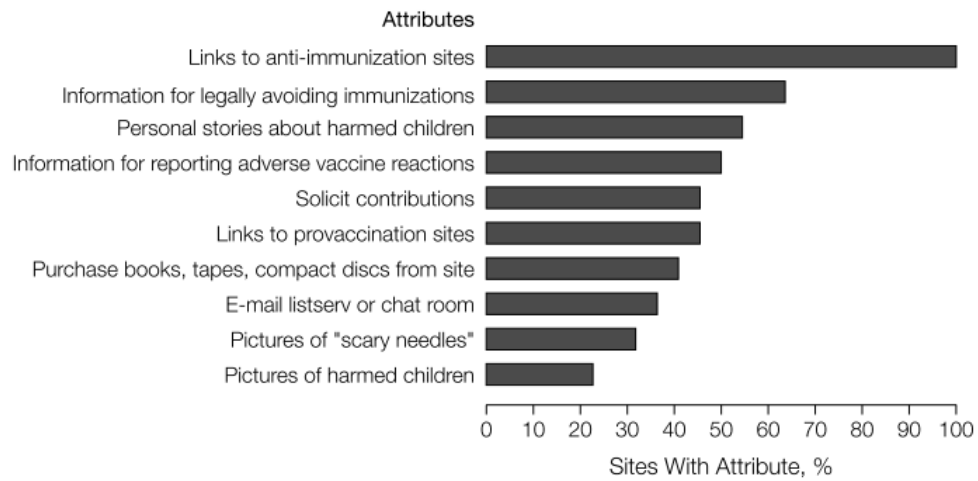
El avance de las redes sociales y la adopción de las mismas es cada vez mayor en todo en todo el mundo; en cifras, se tiene que: hay un aproximado de 7.83 billones de personas en el mundo, de las cuales, 5.22 billones usan teléfonos celulares (equivalente al 66.6% de la población mundial), 4.66 billones tienen internet (para el 2021, internet tiene una penetración mundial del 59.9%) y, por último, las redes sociales abordan un aproximado de 4.20 billones de personas, indicando un crecimiento del 13% respecto al 2020, ocupando un 53% de la población mundial total [22]. Dados estos datos, año tras año cobra más relevancia estudiar la interacción del usuario con la información, planteando nuevos retos en la entrega de las comunicaciones, optimizando los canales disponibles y acelerando su consumo; por esto, se hace necesario valorar la autonomía del usuario en el proceso de generación, uso, difusión y reutilización de los contenidos expuestos en internet. Lo que significa, una redefinición del papel de los expertos de la salud en la intermediación entre información y usuario.

Ilustración 2 Atributos de contenido [23]



Realizando un análisis exploratorio sobre trabajos que ejemplifican el discurso pro vacunación y anti vacunación, y, vinculando estos a los riesgos que conllevan para la salud pública, junto con la inmunidad de rebaño, se identifican estudios que abarcan prácticas desde la extracción de información por medio de cuestionarios, hasta análisis de sitios Web relacionados con la problemática. Para el último punto, existe un trabajo publicado en el 2002 que analizó 22 sitios web de grupos anti-vacunación, para ejecutar análisis de contenido y diseño sobre el sitio web, sus resultados sugieren que el discurso más común es que las vacunas causan enfermedades idiopáticas (100% de los sitios), las vacunas erosionan la inmunidad (95%), las reacciones adversas a las vacunas no se reportan (95%) y la política de vacunación está motivada por las ganancias (91 %) (ver Ilustración 2). Entre los atributos de diseño más comunes, se encontró la presencia de enlaces a otros sitios de anti vacunación (100% de los sitios), información para evitar legalmente las vacunas (64%) y el uso de historias con carga emocional de niños que supuestamente habían sido asesinados o heridos por vacunas (55%) [23] (ver Ilustración 3).

Ilustración 3 Atributos de diseño [23]



2 Análisis de redes sociales y procesamiento de lenguaje natural

2.1 Análisis de redes sociales

El análisis de redes sociales proporciona todas las herramientas y metodologías disponibles a nivel técnico y tecnológico, con la finalidad de facilitar el entendimiento de las interacciones entre los miembros de la red o grupo; dado el crecimiento orgánico y constante de dicho entorno, emergen estructuras, comportamientos y perfiles como consecuencia de dichas relaciones [24].

Si se enfocan los estudios en las estructuras sociales emergentes que nacen de la interacción entre los usuarios de las redes sociales, junto con la conducta de los individuos y las relaciones e interacciones que surgen en la red; se pueden encontrar flujos de conocimiento, colaboración, poder, hasta llega a la información y desinformación; desde esta perspectiva se puede aludir a la siguiente cita del texto “El análisis de redes en el desarrollo local”:

“La estructura de la red de relaciones sociales es mejor fuente de explicación de las conductas, que los atributos personales de los individuos. Nótese que se dice ‘mejor’, no única fuente de explicación. Así, por ejemplo, el comportamiento de los jóvenes o de las mujeres, se explicará mejor por las redes en las que se integran que por sus atributos de edad o de género.” [25]

De esta manera, encontramos el análisis de las redes sociales como parte fundamental de la Web 2.0, buscando extraer información, encontrando interacciones y reacciones hasta llegar a una o muchas interpretaciones de múltiples perspectivas. Como mencionan en [26], las redes sociales son cibermundos que merecen la pena analizar, como si fueran un organismo vivo que se transforma en el tiempo y genera sus propias formas de comunicar.

2.2 Análisis de contenido

Como metodología indirecta, el análisis de contenido se centra en el análisis, exploración y explotación de fuentes documentales, más no a los individuos directamente relacionados, tiene la flexibilidad de entender la realidad de las interacciones, pudiendo extraer información en un sentido cuantitativo, cualitativo o mixto. En ese sentido, el análisis de contenido es una metodología inspirada por múltiples corrientes filosóficas, sociológicas y estadísticas, que busca descubrir el significado de una interacción, mensaje, información o perfil, ya sea éste un discurso, una historia de vida, un artículo de prensa, una publicación en internet, un programa televisivo, una película o cualquier tipo de documento [27].

En términos generales, una red social se puede estudiar desde dos perspectivas: desde los atributos particulares de los individuos que interactúan en la red, extrayendo sus características intrínsecas e individuales independientes del contexto; o, desde las relaciones entre los individuos, interpretada como la participación e interacción de los individuos; de tal forma que la relación no es una característica intrínseca, es propiedad emergente de la interacción entre varios individuos. Es necesario considerarlas simultáneamente para que aporten a la visión de conjunto de la realidad social [11]. Al respecto, los individuos en una red pueden tener diferentes roles e influencia según sea su grado de cercanía o poder de intermediación con relación a los demás. Asimismo, los vínculos pueden ser directos o indirectos, direccionales o no direccionales, y tener diferentes intensidades; estas características determinan el tipo de relación existente y el tipo de estructura de red que conforman.

2.3 Procesamiento de lenguaje natural

Para que un lenguaje natural, como el español, pueda ser entendido por la computadora y producir algún resultado o medición deseada, además de ser una tarea desafiante, se deben seguir una serie de pasos con el fin de darle una estructura entendible al texto, para poder ser procesada correctamente por la computadora; a esto se le llama procesamiento de lenguaje natural (NLP, por su sigla en inglés). Una de las mayores complejidades es la posibilidad de obtener múltiples interpretaciones para una misma oración. Las interpretaciones metafóricas son difíciles de aprender por las computadoras, y así mismo,

las ambigüedades en palabras, oraciones, parte del discurso (POS, por su sigla en inglés), sintaxis, significado y, así mismo, se dificulta la comprensión automática del texto. Problemas como la identificación del sarcasmo, el humor, comprender la semántica del texto son tareas que también se abordan en el procesamiento de lenguaje natural [28].

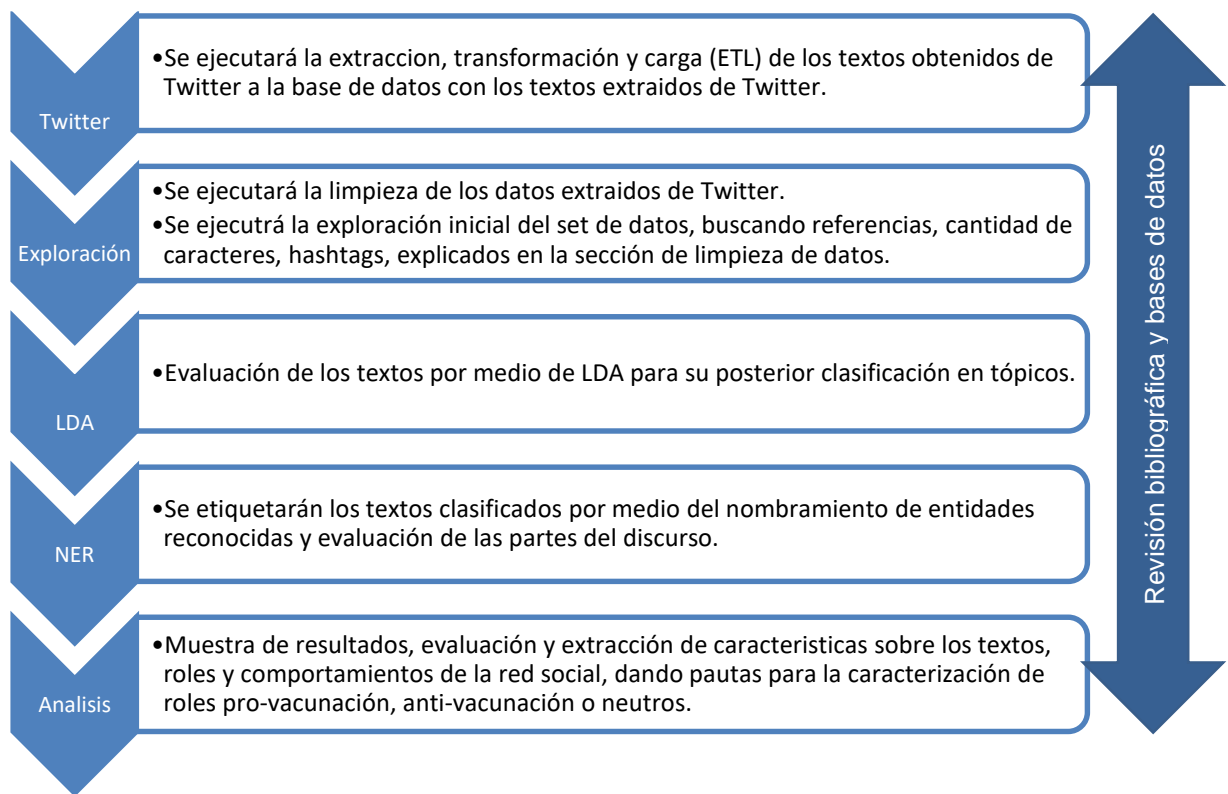
El texto digital que prevalece en las redes sociales brinda una cantidad abundante de información, a partir de la cual se pueden descubrir muchas relaciones emergentes e información relevante que no es visible desde los ojos humanos; algunas de las aplicaciones populares de análisis de texto en las redes sociales son detección de noticias, detección de eventos, detección de sentimientos, respuestas automáticas y clasificación de perfiles.

Las redes sociales como Twitter son conocidas como “micro blogs” que brindan a las personas servicios para publicar mensajes en textos cortos a fin de ser compartidos públicamente. Este tipo de textos retratan la mentalidad de la persona que ha escrito el mensaje y su relación con otros usuarios de la red social. En NLP, el contenido de los textos se analiza para comprender las opiniones del usuario. La fácil disponibilidad de Twitter al ser multiplataforma permite que las interacciones entre los usuarios sean mucho más rápidas y ricas en términos de extracción de información. Twitter también abre las puertas a la comunidad científica por la disponibilidad de su API para la extracción de los datos.

3 Metodología propuesta

La metodología implementada para el proyecto enfatizará esfuerzos en investigación cuantitativa, dado que se busca extraer información de la red social Twitter en español, en pro de caracterizar los perfiles (Pro-vacunación, Anti-vacunación y Neutral) y los tweets para identificar el comportamiento de dichos grupos en el contexto de la salud pública y la vacunación; en ese sentido, el diseño del proyecto, combinará diferentes tipos de implementaciones que se integrarán para la descripción, clasificación y categorización del contenido de Twitter (ver Ilustración 4).

Ilustración 4 Arquitectura metodológica



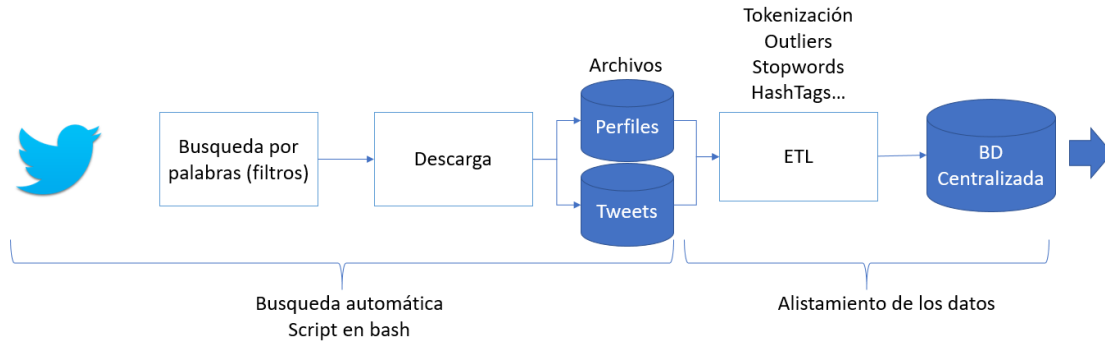
Se ha abordado este proceso metodológico ya que tiene como objetivo la creación de una herramienta sin precedente, no existe una metodología estándar para el desarrollo de este tipo de herramientas, enfocadas a la salud pública y el análisis de las redes sociales, todo el proceso es una propuesta que fija los cimientos para brindar a los investigadores de la rama una herramienta que centraliza algunas librerías para el análisis continuo y focalizado de Twitter. Se debe aclarar que hay múltiples estudios que ejecutan análisis espacio temporales sobre la red social, indicados en las secciones previas, sin embargo, dichos estudios culminan con el análisis y resultados del conjunto de datos obtenidos hasta la finalización del estudio. El presente trabajo desarrollará una herramienta que brinda continuidad y posible evolución en el producto final, acoplando más proyectos para el procesamiento de lenguaje natural y análisis de redes.

Sobre la Ilustración 4 se identifica el plan de trabajo general para la elaboración de la herramienta, y desde allí se segregará la arquitectura de la aplicación (detallada en la sección 3.1.1). Iniciando el proceso con la extracción de los datos desde Twitter, continuando con la exploración y explotación de los datos, hasta finalizar con la muestra de resultados a nivel gráfico dentro de la misma herramienta.

3.1.1 Arquitectura de la solución

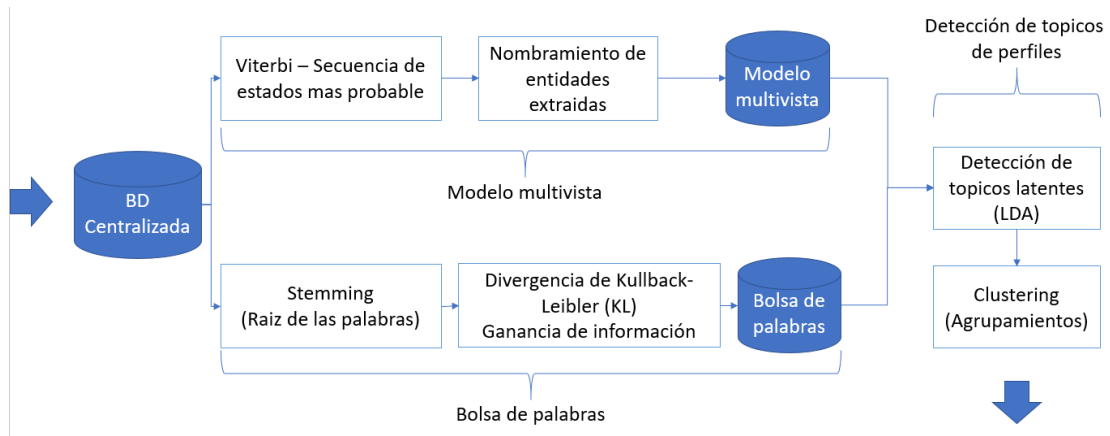
La primera fase de la solución consiste en la captura y manipulación inicial de conjunto de datos a obtener de la red social; se tiene un filtro por coincidencia de palabras a buscar en Twitter; a continuación, se descargan todas las coincidencias y se activan dos procesos paralelos: 1. Se actualiza una base de datos de archivos donde se descarga todo el perfil (también llamado muro) del usuario relacionado al Tweet obtenido y 2. Se actualiza el archivo donde se almacena el Tweet base, obtenido en la búsqueda inicial por palabras relacionadas. Una vez culmina el proceso de descarga de los objetos, se ejecuta la limpieza de los datos descrita en el punto 4.2.2 (Limpieza de los datos) del presente trabajo; para finalmente cargar todos los datos en una base de datos no relacional centralizada (ver Ilustración 5).

Ilustración 5 Fase 1 de la solución



En la segunda fase, partimos de la base de datos centralizada, donde han sido cargados todos los datos obtenidos en la primera fase, con el fin de ejecutar dos procesos paralelos: 1. Generar el modelo multivista para ejecutar el algoritmo de nombramiento de entidades conocidas (NER, Name Entity Recognition), explicado en la sección 5.2 del presente trabajo y 2. Generar la bolsa de palabras (BOW, Bag of words), explicado en la sección 5.1 del presente trabajo. Culminando la fase de entrenamiento con la detección de tópicos latentes (LDA – Latent Dirichlet Allocation) y agrupamiento de los documentos. Hasta este punto se han producido tres resultados relevantes NER, LDA y BOW (ver Ilustración 6).

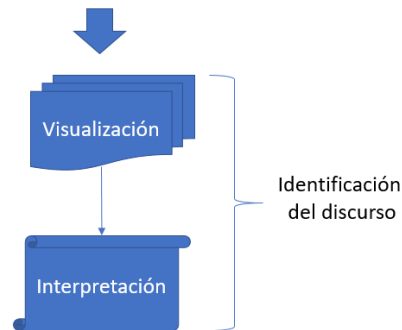
Ilustración 6. Fase 2 de la solución



Como resultado final de todo el procesamiento descrito, se logrará obtener un análisis de discurso sobre los diferentes perfiles, por medio de la visualización de los resultados y una

interpretación con ayuda de expertos que logre acotar dichos resultados en 3 grupos principales: pro-vacunación, anti-vacunación y neutrales (ver Ilustración 7).

Ilustración 7. Fase 3 de la solución



3.2 Actividades desarrolladas

Objetivo: Aplicar las herramientas disponibles en la actualidad, para realizar el proceso de descarga, limpieza y análisis exploratorio de un conjunto de textos (en español) sobre la red social Twitter.	
Actividades	<ul style="list-style-type: none"> • Implementar herramientas y librerías para la extracción de objetos de la red social (tweets), utilizando Twitter4J. • Crear la base de datos para almacenar los objetos extraídos de la red social en mongoDB • Realizar la recopilación y búsqueda sobre herramientas disponibles para realizar ETL, utilizando. NetCore y programas ejecutables en consola de comandos. • Realizar limpieza sobre los objetos extraídos. • Realizar análisis exploratorio sobre los datos: Nubes de palabras, análisis de frecuencias.

Objetivo: Aplicar un método de clasificación de tópicos para para la evaluación y clasificación de los textos sobre el conjunto de publicaciones de Twitter.	
Actividades	<ul style="list-style-type: none"> • Se identificó la librería de Catalyst para implementar LightLDA ampliamente utilizada para la detección de tópicos. • Configurar y parametrizar Catalyst para integrar la librería al proyecto y evaluar las clases generadas desde una muestra los objetos (textos) extraídos.

Objetivo: Integrar las herramientas disponibles, en pro de ejecutar las acciones necesarias de configuración e implementación sobre la caracterización de roles y clasificación de textos en español.	
Actividades	<ul style="list-style-type: none"> • Se configuraron e integraron las diferentes herramientas al proyecto, se especificará en capítulos posteriores la metodología abordada y proceso final de esta tarea. • Se ejecutaron pruebas de carga y calidad para garantizar el correcto funcionamiento, así mismo se evaluó la pertinencia de los resultados basado en opinión experta.

Objetivo: Aplicar un método de aprendizaje no supervisado para la categorización de las clases y elementos clave que caractericen los perfiles pro-vacunación, anti-vacunación y neutrales.	
Actividades	<ul style="list-style-type: none"> • Evaluar las implementaciones existentes sobre máquinas de aprendizaje no supervisado que estén acorde a la parametrización de las clases extraídas en el clasificador de tópicos y el análisis exploratorio. • Configurar e implementar la máquina de aprendizaje no supervisado seleccionada, en pro de integrarla al proyecto. • Evaluar y analizar los resultados obtenidos de la caracterización, para describir el comportamiento de los temas sobre los diferentes grupos en la red social.

4 Preparación del conjunto de datos

4.1 Descripción de las fuentes de datos

FORMATO DE LA FUENTE DE DATOS

Los datos para almacenar en la base de datos se encuentran en archivos en formato JSON (ver Ilustración 8), los cuales reflejan fielmente la estructura de los tweets que se pueden encontrar en la red social Twitter descargados por medio de la API de Twitter Developer.

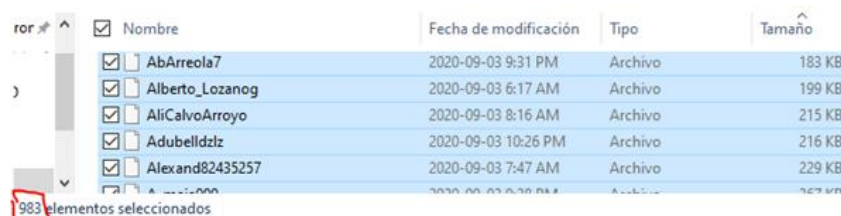
Ilustración 8: Ejemplo de formato JSON sobre los datos descargados de Twitter

```
"user": {  
  "id": 1242410521225302000,  
  "name": "Abraham Arreola",  
  "screenName": "AbArreola",  
  "location": "🇲🇽",  
  "description": "Twitter es mi libreta de apuntes",  
  "descriptionURLEntities": [],  
  "isContributorsEnabled": false,  
  "profileImageUr1": "http://pbs.twimg.com/profile_images/1242414131648626688/Lfw_-Tr6_normal.jpg",  
  "profileImageUr1Https": "https://pbs.twimg.com/profile_images/1242414131648626688/Lfw_-Tr6_normal.jpg",  
  "isDefaultProfileImage": false,  
  "isProtected": false,  
  "followersCount": 12,
```

TAMAÑO DE LOS DATOS

El conjunto de datos principal tiene un peso de 2.84GB y una cantidad aproximada de 1'200.000 de registros divididos en 983 archivos (ver Ilustración 9).

Ilustración 9 Archivos descargados de los perfiles de Twitter



Nombre	Fecha de modificación	Tipo	Tamaño
AbArreola7	2020-09-03 9:31 PM	Archivo	183 KB
Alberto_Lozanog	2020-09-03 6:17 AM	Archivo	199 KB
AliCalvoArroyo	2020-09-03 8:16 AM	Archivo	215 KB
Adubelldzlz	2020-09-03 10:26 PM	Archivo	216 KB
Alexand82435257	2020-09-03 7:47 AM	Archivo	229 KB

983 elementos seleccionados

ESTRUCTURA DE LOS ARCHIVOS FUENTE

Los archivos fuente tienen dentro de su estructura existen tres objetos principales, tweets, usuarios y entidades para los cuales se muestra su descripción a continuación.

Objeto: tweets

Los tweets son el bloque de construcción atómico básico de todo lo relacionado con Twitter. Los tweets también se conocen como "actualizaciones de estado". El objeto tweet tiene una larga lista de atributos de "nivel raíz", incluidos atributos fundamentales como id, created_at y text. Los objetos Tweet también son el objeto "principal" de varios objetos secundarios (ver Tabla 2). Dentro de los objetos secundarios de Tweet se incluyen usuario y entidades.

Tabla 2 Estructura objeto Tweet

Atributo	Tipo	Descripción
created_at	String	Hora UTC cuando se creó el Tweet.
id	Int64	La representación entera del identificador único de este Tweet.
id_str	String	Identificador único del Tweet.
text	String	El texto en UTF-8 del Tweet.
source	String	Dispositivo utilizado para publicar el Tweet, como una cadena con formato HTML.
truncated	Boolean	Indica si el valor del parámetro de texto se truncó
in_reply_to_status_id	Int64	Nullable. Si el Tweet representado es una respuesta, este campo contendrá el ID del Tweet original.
in_reply_to_status_id_str	String	Nullable. Si el Tweet representado es una respuesta, este campo contendrá el texto del Tweet original.
in_reply_to_user_id	Int64	Nullable. Si el Tweet representado es una respuesta, este campo contendrá el ID de autor del Tweet original.
in_reply_to_user_id_str	String	Nullable. Si el Tweet representado es una respuesta, este campo contendrá el nombre del autor del Tweet original.
in_reply_to_screen_name	String	Nullable. Si el Tweet representado es una respuesta, este campo contendrá el nombre de pantalla del autor del Tweet original.
user	Colección	El usuario que publicó el Tweet
coordinates	Coordinates	Nullable. Representa la ubicación geográfica del Tweet según lo informado por el usuario o la aplicación cliente. La matriz de coordenadas tiene el formato geoJSON (primero la longitud y luego la latitud).
place	Places	Nullable. Indica que el tweet está asociado (pero no necesariamente se origina en) un lugar.
quoted_status_id	Int64	Este campo sólo aparece cuando el Tweet es una cita. Contiene el ID de Tweet del Tweet citado.
quoted_status_id_str	String	Este campo sólo aparece cuando el Tweet es una cita. Contiene el texto del Tweet citado.
is_quote_status	Boolean	Indica si se trata de un Tweet citado.
quoted_status	Tweet	Este campo sólo aparece cuando el Tweet es una cita. Este atributo contiene el objeto Tweet del Tweet original que se citó.
retweeted_status	Tweet	Este atributo contiene el Tweet original que se retuiteó.
quote_count	Integer	Nullable. Indica aproximadamente cuántas veces este Tweet ha sido citado por usuarios de Twitter.
reply_count	Int	Número de veces que se ha respondido el Tweet.
retweet_count	Int	Número de veces que se ha retuiteado el Tweet.
favorite_count	Integer	Nullable. Indica aproximadamente cuántas veces le han gustado a este Tweet los usuarios de Twitter.

Fuente: Elaboración propia

Objeto: User

El objeto Usuario contiene metadatos de la cuenta de usuario de Twitter que describen al usuario de Twitter al que se hace referencia. Los usuarios pueden crear tweets, retweets, citar tweets de otros usuarios, responder a tweets, seguir a usuarios, ser @mencionados en tweets y pueden agruparse en listas (ver Tabla 3).

Tabla 3 Estructura objeto User

Atributo	Tipo	Descripción
id	Int64	La representación entera del identificador único del usuario
id_str	String	Representación en cadena del identificador único del usuario
name	String	El nombre del usuario, tal como lo han definido. No necesariamente el nombre de una persona.
screen_name	String	El nombre de pantalla, identificador o alias con el que este usuario se identifica.
location	String	La ubicación definida por el usuario para el perfil de la cuenta
derived	Arrays	Proporciona metadatos de enriquecimiento geográfico del perfil
url	String	Una URL proporcionada por el usuario en asociación con su perfil
description	String	Una cadena UTF-8 definida por el usuario que describe su cuenta
protected	Boolean	Cuando es verdadero, indica que este usuario ha elegido proteger sus Tweets
verified	Boolean	Cuando es verdadero, indica que el usuario tiene una cuenta verificada
followers_count	Int	El número de seguidores que tiene esta cuenta actualmente. Bajo ciertas condiciones de coacción, este campo indicará temporalmente "0"
friends_count	Int	El número de usuarios que sigue esta cuenta (también conocido como sus "seguidores")
listed_count	Int	El número de listas públicas de las que es miembro este usuario
favourites_count	Int	La cantidad de Tweets que le han gustado a este usuario durante la vida de la cuenta
statuses_count	Int	El número de Tweets (incluidos los retweets) emitidos por el usuario
created_at	String	La fecha y hora UTC en que se creó la cuenta de usuario en Twitter.
profile_banner_url	String	La URL basada en HTTPS que apunta a la representación web estándar del banner de perfil subido por el usuario
profile_image_url_https	String	Una URL basada en HTTPS que apunta a la imagen de perfil del usuario.
default_profile	Boolean	Cuando es verdadero, indica que el usuario no ha modificado el tema o el fondo de su perfil de usuario.
default_profile_image	Boolean	Cuando es verdadero, indica que el usuario no ha subido su propia imagen de perfil y se usa una imagen predeterminada en su lugar.
withheld_in_countries	Array of String	Cuando está presente, indica una lista de códigos de país de dos letras en mayúsculas y este contenido se retiene.
withheld_scope	String	Cuando está presente, indica que el contenido que se retiene es un "usuario".

Fuente: Elaboración propia

Objeto: Entities

Las secciones de entidades y entidades extendidas están formadas por matrices de objetos de entidad. Una colección de entidades comunes que se encuentran en los tweets incluye hashtags, enlaces y menciones de usuarios. Este objeto de entidades incluye un atributo de medios, pero su implementación en la sección de entidades solo es completamente precisa para tweets con una sola foto. Para todos los tweets con más de

una foto, un video o un GIF animado, existe otra entidad llamada `extended_entities` (ver Tabla 4).

Tabla 4 Estructura objeto Entities

Atributo	Tipo	Descripción
hashtags	Arreglo de objetos Hashtag	Representa hashtags que se han encontrado del texto del Tweet.
media	Arreglo de objetos multimedia	Representa elementos multimedia subidos con el Tweet.
urls	Arreglo de objetos URL	Representa las URL incluidas en el texto de un Tweet.
user_mentions	Arreglo de objetos de mención del usuario	Representa a otros usuarios de Twitter mencionados en el texto del Tweet.
symbols	Arreglo de objetos de símbolo	Representa símbolos, (por ejemplo: \$ cashtags), incluidos en el texto del Tweet.
polls	Arreglo de objetos de encuesta	Representa las encuestas de Twitter incluidas en el tweet.

Fuente: Elaboración propia

4.2 Características de los datos

Con el fin de ver las características de los datos en términos de las características de Big Data, se definieron 5 características:

Volumen: Para el conjunto de datos principal, se tiene un peso de 2.84GB y un aproximado de 1'200.000 de registros divididos en 983 archivos.

Velocidad: En promedio se generan 200 tweets por día que tienen relación con el conjunto principal. Para el ejercicio de este proyecto, se tomará un conjunto de datos estático de las fechas en su mayoría entre septiembre y octubre de 2020.

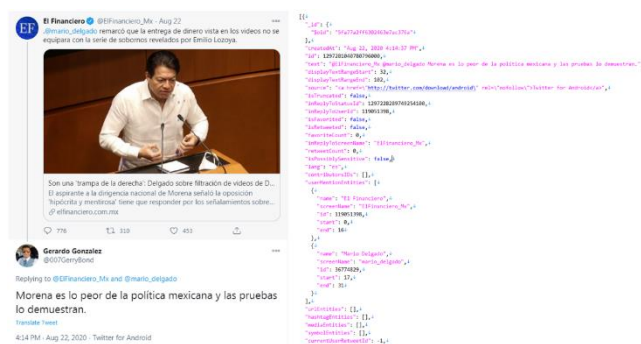
Variedad: Los tweets por sí mismo generan variedad de datos. Por un lado, se tienen los tweets más básicos que se componen de un texto básico. Sin embargo, existen tweets que se componen de videos, audios, fotografías y URL que proporcionan variedad al conjunto de datos.

Valor: El concepto de valor fija que: “entre mayor sean la cantidad de los datos, menos valor tienen”, esto se debe a que pueden llegar a perder el foco de búsqueda entre más cantidad de datos se obtienen, sin embargo, para el conjunto de datos manejados en este

proyecto, se procura tener un balance entre cantidad y calidad para extraer la mejor información posible.

Veracidad: Debido a que la información fue descargada directamente desde Twitter por medio del API, se asume que son datos veraces y que refleja la actividad de Twitter para las fechas antes mencionadas. Para la verificación, se revisaron algunos de los tweets. Un ejemplo de esto se muestra, en Ilustración 10.

Ilustración 10 Verificación de Twitter vs datos descargados



4.2.1 Calidad y disponibilidad de los datos

Disponibilidad: Estos datos se subirán a la base de datos dentro de un servidor de Atlas Mongo, el cual estará habilitado 7 x 24 y por su misma arquitectura de recuperación donde sí existe un punto de fallo en algunos de los nodos, hay nodos secundarios que estarán disponibles para consulta.

Usabilidad: Los datos son extraídos directo de la fuente oficial y sin algún tipo de procesamiento, lo cual los hace ideales para su temprana estructuración, aportando gran usabilidad de estos.

Confiabilidad: Se garantiza exactitud por la fuente oficial, consistencia, integridad y completitud dado el origen de la fuente de los datos.

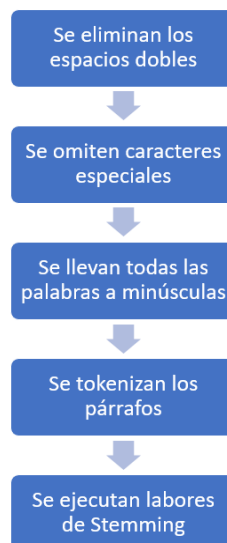
Pertinencia: Dado que se busca establecer patrones de comportamiento de usuarios sobre las redes sociales y temas principales de discusión, específicamente de Twitter, la pertinencia de los datos para el negocio es clara.

Calidad de presentación: Los datos son claros y completos en su totalidad, dado que no ha pasado por ningún tipo de procesamiento o programa.

4.2.2 Limpieza de los datos

La ejecución de la limpieza de los datos es crucial para garantizar una correcta obtención y extracción de información; para el presente trabajo se ejecutaron cinco actividades sobre cada uno de los textos obtenidos (ver Ilustración 11).

Ilustración 11 Diagrama de flujo del proceso de limpieza de datos.



Fuente: Elaboración propia

1. **Eliminación de espacios dobles:** Se logró identificar una gran cantidad de textos que contienen espacios dobles, saltos de línea, tabulaciones y caracteres invisibles que se deben eliminar en cada iteración de limpieza sobre el conjunto de datos.

2. **Omisión de caracteres especiales:** Los usuarios de la red sociales comparten emoticones, vínculos, menciones, hashtags, los cuales serán omitidos, dado que el enfoque se centra en el texto del tweet.
3. **Palabras en minúsculas:** Con el fin de estandarizar las estructuras de los textos, se deben llevar a minúsculas cada una de las palabras para evitar diferencias entre los contadores de palabras y evaluadores de expresiones, por ejemplo: “Textos de TwITTer”, se transformará en “textos de twitter”.
4. **Tokenización de párrafos:** El proceso de análisis de textos requiere llevar documentos de palabras a tokens que atomicen, evalúen, sumen y procesen, en general, los textos como un conjunto de datos, desde una perspectiva global hasta una individual. En la sección “Modelamiento de textos para la caracterización de textos sobre vacunación” se tendrá una aplicación sobre la relevancia de la tokenización.
5. **Stemming:** En la morfología lingüística y la recuperación de información y el stemming es el proceso de reducir palabras derivadas a sus respectivas raíces, palabras base o raíz, de su forma escrita. La raíz de una palabra no siempre tiene que ser la misma que la palabra o sus derivados; incluso si la palabra relacionada en sí no es una raíz válida, generalmente, es posible asignar la palabra relacionada a la misma raíz. Muchos motores de búsqueda utilizan palabras con la misma raíz en lugar sinónimos como una forma de expansión de consultas. Por ejemplo, la palabra “vendedor” se simplificará a su raíz “vend”, con lo cual se homologará con “vender”, “vendido” y todos sus derivados.

5 Modelamiento de textos para la caracterización de textos sobre vacunación

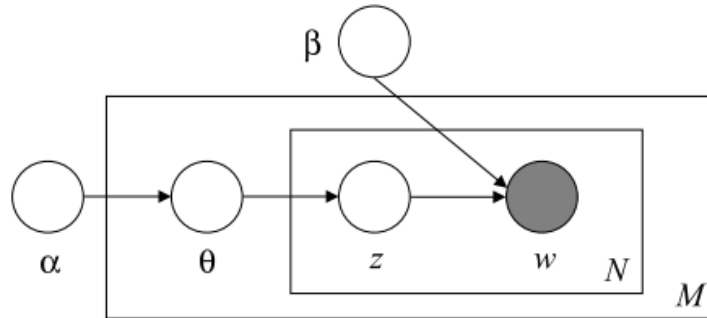
Como parte fundamental del presente trabajo, se busca ejecutar un modelamiento sobre el conjunto de datos obtenidos, después de la limpieza y análisis exploratorio. En esta sección se describe el modelamiento de los datos desde dos perspectivas: modelamiento por Latent Dirichlet Allocation (LDA) y reconocimiento de entidades nombradas (NER) (ver Ilustración 5).

5.1 Modelamiento por Latent Dirichlet Allocation (LDA)

Para esta versión de la herramienta, y dado que se debe contemplar eficiencia a nivel computacional sobre los procesos de extracción, transformación y evaluación de los datos, se utilizó la librería LightLDA [29][30], una versión ligera de LDA que busca optimizar el procesamiento de los datos para ser ejecutados en equipos computacionales sin altas capacidades de procesamiento. Es necesario aclarar que el modelo tradicional de LDA se sigue aplicando, pero con optimizaciones a nivel de máquina con ejecuciones de multi-hilo.

Describimos LDA como un modelo probabilístico generativo para colecciones de datos discretos, como colecciones de textos; es un modelo bayesiano jerárquico de tres niveles, en el que cada elemento de la colección se modela como una mezcla finita de un conjunto subyacente de temas, por otra parte, cada tema se modela como una mezcla sobre un conjunto de probabilidades de temas. En el contexto del procesamiento del lenguaje natural, las probabilidades de cada tema proporcionan una representación de un documento. La interpretación gráfica de la anterior definición se instruye en la Ilustración 12.

Ilustración 12 Modelo gráfico de LDA: En la caja exterior se representan los documentos M , en la caja interior N se representan los temas y palabras dentro de un documento.[31]



- M Cantidad de documentos
- N Cantidad de palabras en un documento dado (documento i tiene N_i palabras)
- α Parámetro del Dirichlet antes de las distribuciones de temas por documento
- β Parámetro de Dirichlet antes de la distribución de palabras por tema
- θ_i Distribución de temas para el documento i
- z_{ij} Tema de la j – ésima palabra en el documento i
- w_{ij} Palabra específica

Es necesario en este punto identificar 3 términos base que ayudarán a comprender la funcionalidad de LDA, en términos del agrupamiento e identificación de tópicos sobre los textos, acorde a Blei Ng y Jordan [31]: 1) Palabra: es la unidad básica de los datos y se definirá dentro de un arreglo de palabras como un vector indexado del tipo $\{1, \dots, V\}$ donde V es la cantidad de palabras en el texto que no se repite (también llamado vocabulario). 2) Documento: es una secuencia de N palabras denotadas por $w = w_1, w_2, w_3, \dots, w_N$ donde w_n es la n -sima palabra en la secuencia y 3) Corpus: es la colección de M documentos denotados por $D = \{w_1, w_2, \dots, w_M\}$.

Dadas las anteriores definiciones, se enmarca en el texto de Blei, Ng y Jordan [31] que la probabilidad total de los tópicos por la colección de textos está dada por la Ecuación 1. Entendiendo que el resultado final será en la probabilidad de que una palabra dentro de un documento se encuentre en una agrupación de palabras llamadas tópicos.

Ecuación 1 Probabilidad total de la colección de documentos.[31]

$$p(D | \alpha, \beta) = \prod_{d=1}^M \int p(\theta_d | \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d\theta_d$$

5.2 Reconocimiento de Entidades Nombradas (NER)

Reconocimiento de entidades nombradas (NER, por su sigla en inglés), tiene dos objetivos: identificar y clasificar entidades sobre texto. Por una parte, la clasificación se basa en categorías predefinidas, tales como ubicaciones, personas u organizaciones y, por otra parte, la identificación de las categorías gramaticales (POS, por su sigla en inglés), que ubican la estructura base, a nivel gramatical, del texto. La ejecución de NER es una tarea crucial en el proceso de análisis de lenguaje natural y extracción de información, dado que sus resultados logran arrojar información relevante sobre los análisis de discurso, como es el caso del presente estudio.

La historia de NER comienza con motores de decisión basados en reglas, para luego evolucionar a enfoques de aprendizaje automático como los perceptrones. En los últimos años, se ha demostrado que los enfoques de aprendizaje profundo logran superar los métodos tradicionales de aprendizaje automático. Las redes neuronales profundas son un conjunto de métodos de aprendizaje automático que se han aplicado en campos como la visión por computadora y reconocimiento de voz [32]. Sin embargo, hasta el momento hay pocas investigaciones sobre el uso del aprendizaje profundo aplicado a NER en español, aun así, para la ejecución de este trabajo, se tomaron dos perspectivas, la primera, en la utilización de una red pre entrenada llamada WikiNER, cuya base son textos previamente etiquetados de Wikipedia y, la segunda, un perceptrón promediado que ejecutará tareas de aprendizaje para la detección y nombramiento de entidades sobre cada una de las palabras (tokens) en los documentos.

El perceptrón promediado es una versión simple de una red neuronal. Sobre este enfoque, al igual que las redes neuronales bayesianas, las entradas se clasifican en dos posibles salidas basadas en una función escalón, para luego combinarse en un conjunto de

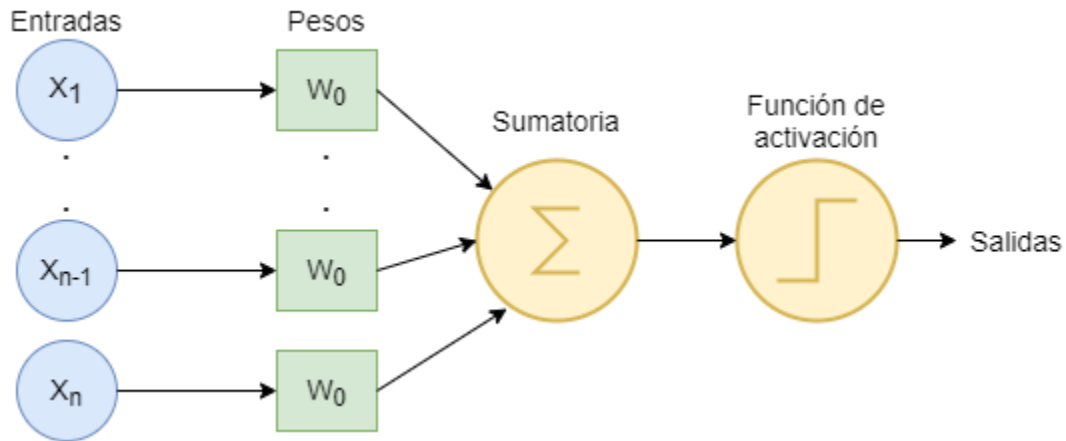
ponderaciones que se derivan del vector de características (pesos); este modelo de perceptrón es adecuado para aprender patrones linealmente separables y, de manera muy eficiente computacionalmente, llegar a una clasificación a modo de entrenamiento continuo, razón por la cual se ha seleccionado el modelo de perceptrón promediado.

Tabla 5 Etiquetas de entidades conocidas

Etiqueta	Descripción
Organization	Etiqueta de anotación para una organización
Person	Etiqueta de anotación para una persona
Location	Etiqueta de anotación para un lugar (ciudad o país)
URLOrEmail	Etiqueta de anotación para una URL o un Email
X	Etiqueta de anotación para una entidad no identificada

Para entender más a profundidad del perceptrón promediado y sus ventajas, revisemos el siguiente ejemplo: supongamos que tenemos 10.000 registros etiquetados que se evaluarán en el perceptrón, se evalúan 9.999 de manera satisfactoria, sin embargo, la última evaluación no se pudo clasificar correctamente y esto causó que todos los 9.999 ejemplos anteriores fallarán; esto se soluciona, normalmente, bajo un sistema de votos, si se entrena en 5 iteraciones con 10.000 registros, se almacenarán 50.000 modelos con sus pesos y evaluaciones, para ser seleccionado el de mejor resultado, aun así, a una escala macro, esto resulta ser un desperdicio de memoria. El mejor método identificado fue el perceptrón promediado donde se acumulan los vectores de peso de los 50.000 modelos para ser promediados en cada iteración, registrando solo el valor acumulativo y optimizando recursos de hardware. Trabajos como el ejecutado por Lars Buitinck y Maarten Marx develan sus ventajas sobre otros tipos de arquitecturas para perceptrones promediados [33].

Ilustración 13 Estructura básica de un perceptrón



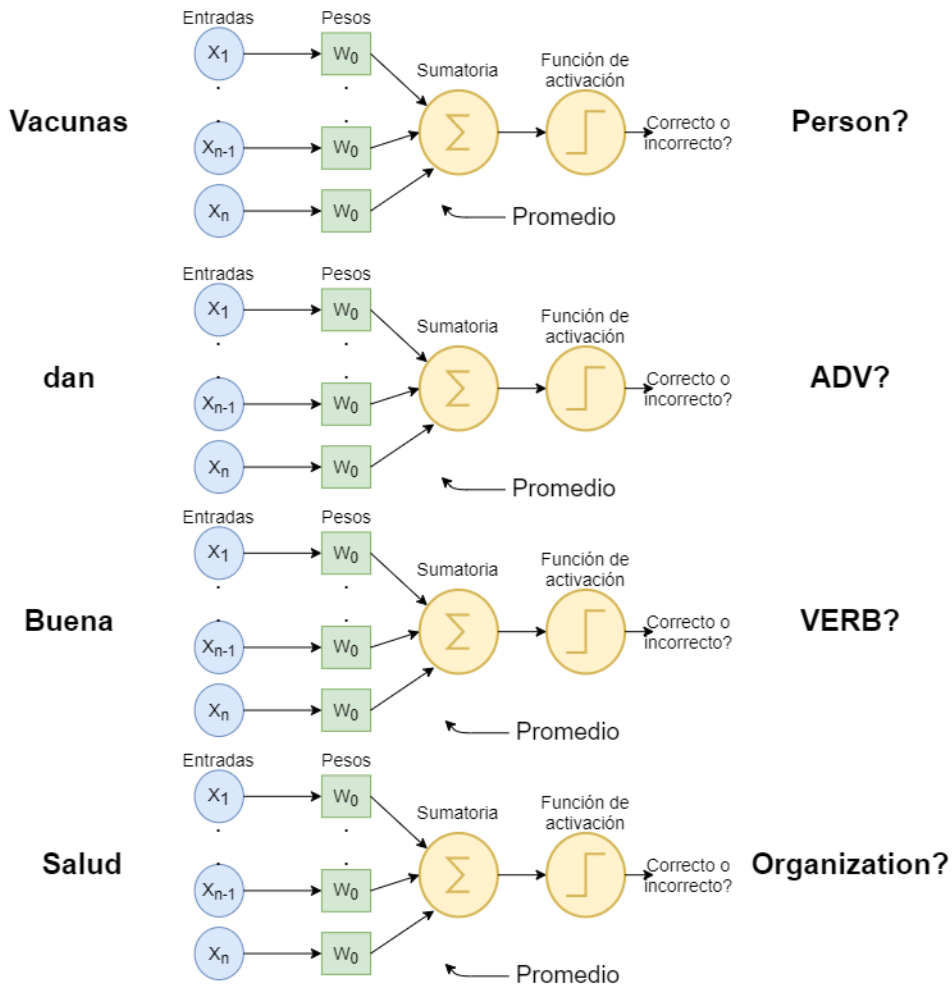
Fuente: Elaboración propia

En la Ilustración 13 Estructura básica de un perceptrón, se puede identificar que, basados en las entradas x_1, x_2 y x_n se calcularán los pesos promediados (w_i) para ser evaluados por la función de activación tipo escalón y llegar a una clasificación, positiva o negativa en la fase de entrenamiento (ver Tabla 5 Etiquetas de entidades conocidas).

Algoritmo de entrenamiento WikiNer [34], [35] (ejemplo de arquitectura, ver Ilustración 14):

- a) Elegir un conjunto de características que serán las entradas del perceptrón (ver Anexo: Tabla de vector de características Spacy).
- b) Crear lista de tags y entidades válidas a partir de los datos de entrenamiento (ver Tabla 5).
- c) Crear un perceptrón promedio a partir de la lista de tags y entidades para controlar los coeficientes de peso del perceptrón.
- d) Entrenar coeficientes de peso en un número dado de iteraciones, para este caso 10 iteraciones.
- e) Aplicar el modelo entrenado en los datos de prueba para cada iteración.
- f) Evaluar la precisión de los datos de prueba para cada iteración.
 - g) Analizar errores. Para el presente trabajo se lograron obtener resultados satisfactorios con un F1 del 94.44%, precisión del 97.66% recordación del 91.44% y un promedio de 1879724 tokens (palabras) por segundo.

Ilustración 14. Ejemplo arquitectura Reconocimiento de entidades Nombradas



Fuente: Elaboración propia

5.3 Visualización de datos por T-SNE

T-SNE (t-distributed stochastic neighbor embedding, por sus siglas en inglés) es un método estadístico para la visualización de datos de alta dimensionalidad; teniendo como entrada datos con 3 o más descriptores, también llamadas dimensiones, y, generando como resultado ubicaciones espaciales sobre el plano cartesiano (2D) o en el espacio de 3 dimensiones (3D); este método es

principalmente abordado para la generación de graficas representativas de los datos.

En el texto de Van der Maaten y Hinton [36] se explica con gran precisión el origen de la propuesta y la base estadística del algoritmo, cuyo fundamento se centra en la similitud y distancia euclidiana de los puntos x_i y x_j , con la probabilidad condicional $p_{j|i}$ de que los puntos x sean vecinos acorde a la densidad de probabilidad bajo la curva Gaussiana centrada en x_i , dado por la Ecuación 2 (donde σ_i es la varianza de la Gaussiana).

Ecuación 2 Probabilidad condicional de vecindad para t-SNE [36]

$$p_{j|i} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)}$$

Así mismo como tenemos una distancia de probabilidades para hallar la similitud de los puntos en el espacio de alta dimensionalidad, debemos evaluar las distancias en el espacio de baja dimensionalidad, en este caso, Vander Maaten y Hinton proponen la distribución t de Student con un grado de libertad gracias a sus propiedades para equiparar la ley cuadrática inversa (la probabilidad de vecindad es inversamente proporcional a la distancia del punto de origen) detallada en la Ecuación 3.

Ecuación 3 Distribución t-de Student para mapa de baja dimensionalidad [36]

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_k - y_l\|^2)^{-1}}$$

Por último, tenemos la Ecuación 4, donde se evalúa el gradiente de la divergencia de Kullback-Leibler entre p y q para hallar la diferencia o similitud entre las dos funciones de probabilidad y garantizar que los dos espacios, alta dimensionalidad y baja dimensionalidad, sean similares en su interpretación gráfica.

Ecuación 4 Gradiente de la divergencia de Kullback-Leibler entre p y q [36]

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$$

Previo a evaluar el paso a paso final para la reducción de dimensionalidad, debemos tratar una de las variables más importantes para la parametrización de t-SNE; La perplejidad, que se puede interpretar como una medida aproximada de la cantidad de vecinos sobre un punto, hallando una relación directamente proporcional con la varianza. Normalmente se toma la perplejidad entre 5 y 50 dependiendo de la densidad del conjunto de datos, entre más datos, mayor debe ser la perplejidad. Ahora, paso a paso para la evaluación de t-SNE propuesto por Vander Maaten y Hinton es el siguiente:

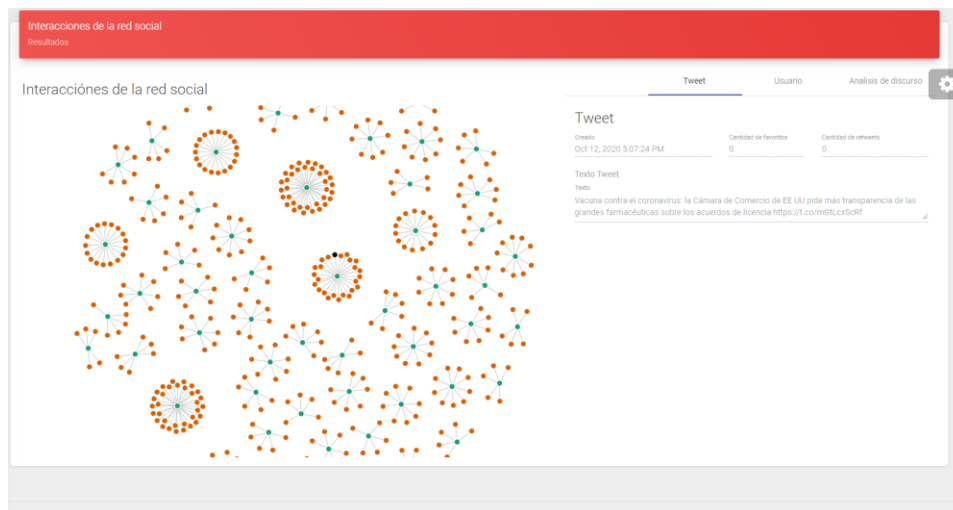
- 1) Parámetros de entrada: Perplejidad, conjunto de datos de alta dimensionalidad ($X = \{x_1, \dots, x_n\}$), número de iteraciones (para nuestro caso se tomaron 1000 iteraciones como máximo, o iterar hasta encontrar la distribución correcta que cumpla con la perplejidad), tasa de aprendizaje (en nuestra experimentación se tomó de $\eta = 0.5$) y momentum (cuyo valor va de α 0.5 hasta 0.8).
- 2) Se calcula la Ecuación 2 y se evalúa la perplejidad con $perp(P_i) = 2^{-\sum_j p_{ji} \log_2 p_{ji}}$
- 3) Se inicializan los parámetros de salida de baja dimensionalidad ($Y = \{y_1, \dots, y_n\}$).
- 4) Se itera desde $t = 1$ hasta el número de iteraciones preestablecido
 - a. Se evalúan las ecuaciones 3 y 4 para actualizar el estado del conjunto de salida, por medio de la ecuación: $Y^{(t)} = Y^{(t-1)} + \eta * \text{gradiente} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$
- 5) Se retorna Y como el conjunto de datos de baja dimensionalidad.

6 Visualización e interpretación

Como se consolida en la última fase del desarrollo de la herramienta (ver Ilustración 7. Fase 3 de la solución se busca mostrar de manera gráfica las interacciones de la red social sobre los usuarios relacionados con vacunación, también llamados, usuarios de interés, a fin de darle instrumentos al investigador para ejecutar la clasificación de perfiles pro-vacunación, anti-vacunación o neutrales, fundamentado en un análisis de discurso del tweet y usuario relacionado.

En la Ilustración 15, se observa un grafo no dirigido, en el que se relacionan los usuarios (puntos azules) con los tweets publicados por dicho usuario (puntos naranja). Se destaca la actividad que tienen algunos usuarios sobre otros; se observan círculos con gran peso que podrían interpretarse como “influenciadores” tanto positivos, negativos o neutrales, en la red social.

Ilustración 15 Vista de interacciones en la red social



Fuente: Elaboración propia

Indagando más a fondo se encuentran las ilustraciones Ilustración 16 e Ilustración 17, donde se muestra cómo se puede recuperar el usuario y tweet evaluando los textos, fechas y cuantificadores relacionados al perfil del usuario. Vale resaltar que es importante para los expertos en salud la recuperación de los atributos relacionados para poder hacer juicios de valor fundamentados.

Ilustración 16 Vista de recuperación de usuario

Tweet	Usuario	Análisis de discurso
Usuario - forosdesalud		
Creado	Cantidad de seguidores	Cantidad de amigos
Sep 9, 2009 8:39:35 PM	564	89
Descripción		
Cuenta de seguimiento de los Foros de Salud del Senado de la República		
Cantidad de tweets	Cuenta verificada	Cuenta principal
4959	false	false

Fuente: Elaboración propia

Ilustración 17 Vista de recuperación de tweet

Tweet	Usuario	Análisis de discurso
Tweet		
Creado	Cantidad de favoritos	Cantidad de retweets
Sep 2, 2020 6:24:07 AM	17	7
Texto Tweet		
¿Cómo se aprobarán las vacunas? Una eficacia aceptada baja, sólo por el afán, sería grave para el manejo de la pandemia -relajación de medidas, falsa seguridad -. Se afectaría la		

Fuente: Elaboración propia

Llegando a la raíz del presente trabajo, se encuentran las ilustraciones Ilustración 18 e Ilustración 19, que muestran los resultados del preentrenamiento de los módulos LDA y NER. Por lo cual, se concluye del análisis del discurso del texto extraído del tweet, que le provee herramientas al investigador para dar juicios de valor sobre la identificación de un perfil pro-vacunación, anti-vacunación o neutral.

Ilustración 18 Resultados de la evaluación de texto seleccionado para ejecutar LDA

Texto	Tópicos
Vacuna contra el coronavirus: la Cámara de Comercio de EE UU pide más transparencia de las grandes farmacéuticas sobre los acuerdos de licencia https://t.co/mGtLcxScRf	[0,700] => global[0] gripe[0] líder[0] saudí[0] servicios[0] verduras[0] match[0] dermatólogo[0] suburbano[0] muerto[0] [0,300] => crisis[0,001] extranjeros[0] aad2018[0] cara[0] nacional[0] banco[0] chicago[0] circus[0] team[0] investigaciones[0]

Fuente: Elaboración propia

En la columna de la izquierda de la Ilustración 18 encontraremos el texto completo del tweet extraído y a su derecha los resultados de los temas, discriminados por las probabilidades de los elementos y temas agrupados. Información similar se encontrará en la Ilustración 19, reflejando los resultados del nombramiento de las entidades y las partes del discurso (POS) relacionadas.

extraído de Twitter; caja inferior izquierda, cantidad de tweets limpios de la base de datos del filtro base; caja inferior derecha, cantidad de tweets limpios de los perfiles extraídos.

Ilustración 23 Contadores bases de datos

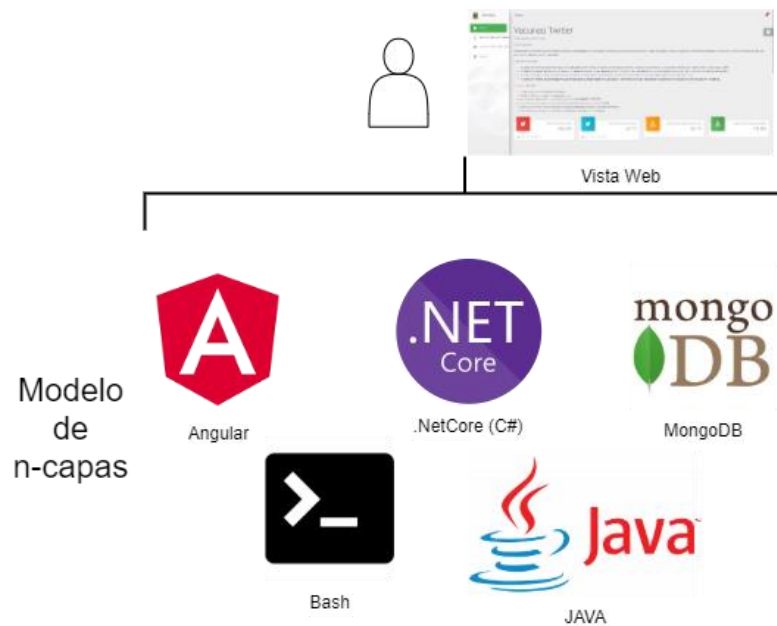


Fuente: Elaboración propia

7 Diseño e implementación de la herramienta computacional

El éxito de una herramienta computacional se basa en su diseño y flexibilidad para evolucionar. En este caso, se ha implementado una arquitectura de n-capas en la cual cada proyecto de analítica, aprendizaje maquina o servicio puede disponer de independencia total sobre el proyecto Web (vista, controlador y modelo de base de datos). Se llegó a un nivel de flexibilidad tal que se ejecutan cuatro integraciones y aplicaciones en cuatro lenguajes de programación diferentes: Java, C#, JavaScript (TypeScript) y Bash (ver Ilustración 24).

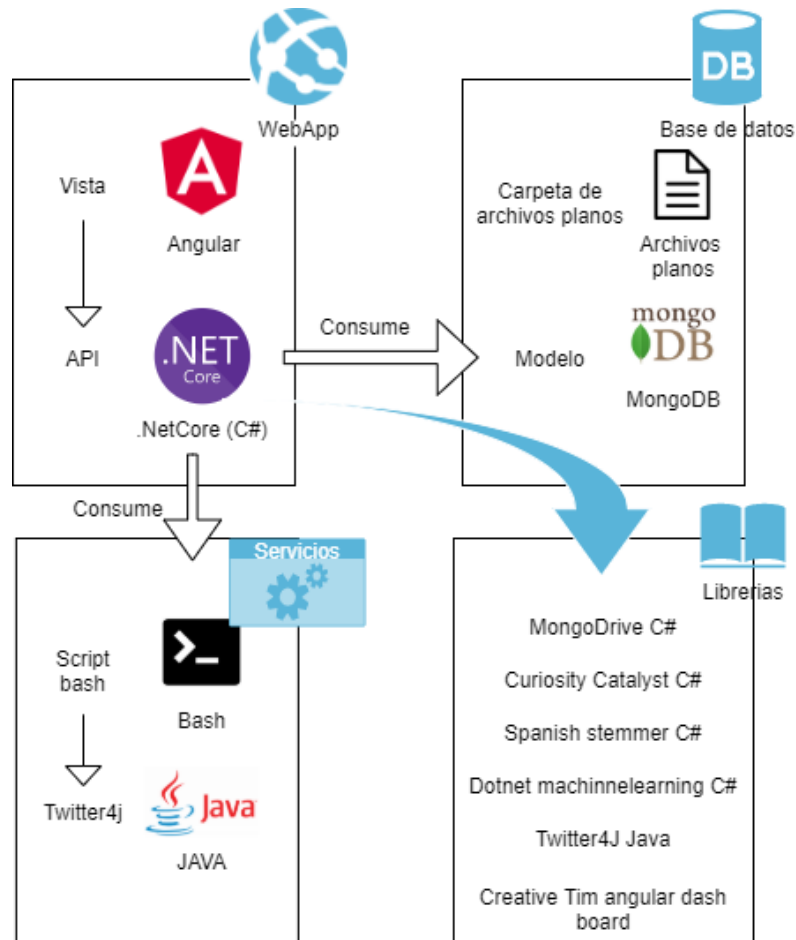
Ilustración 24 Vista global herramientas y lenguajes utilizados



Fuente: Elaboración propia

La arquitectura final implementada se simplifica en la Ilustración 25, donde encontramos el modelo de n-capas descritas en sus componentes: La aplicación Web integrada por el front-end y back-end, la base de datos con los respectivos modelos a nivel de archivos y la base de datos MongoDB, los servicios adicionales compuestos por el Script de Bash para ejecutar el programa en Java con la librería para descargar los tweets por medio de la API de Twitter, y por ultimo las librerías relevantes a la ejecución de la herramienta [30], [37]–[41] (código fuente de la herramienta desarrollada en [42]).

Ilustración 25 Arquitectura y librerías



Fuente: Elaboración propia

7.1 Evaluación cualitativa preliminar de la herramienta

Dado que los objetivos del presente trabajo se enfocan en la utilización e integración de herramientas disponibles, previamente desarrolladas por terceros, y así mismo evaluadas desde los respectivos trabajos; el resultado final ha de ser sometido a juicio de expertos basados en los parámetros que ellos consideran pertinentes para la utilización de esta. Dado el anterior criterio, se mostró la herramienta al grupo de investigación GEESP (Grupo de Epidemiología y Evaluación en Salud Pública de la Universidad Nacional de Colombia – Facultad de Medicina) que reúne tanto expertos como estudiantes, en temas concernientes a salud pública.

Como conclusiones de dicha sesión, se generó la tabla de evaluación de pertinencia sobre la herramienta (ver Anexo: Tabla de evaluación de pertinencia sobre la herramienta) la cual agrupa todos los puntos a tener en consideración desde la vista de un experto en salud pública. Como tal, fue una sesión con resultados satisfactorios, dado que este trabajo fija un precedente en la intervención en salud pública sobre redes sociales y vacunación, fijando también, trabajos futuros para la evolución de la herramienta y puesta en marcha, para trabajos en salud pública per se.

7.2 Vistas y funcionalidades de la herramienta

7.2.1 Tablero principal

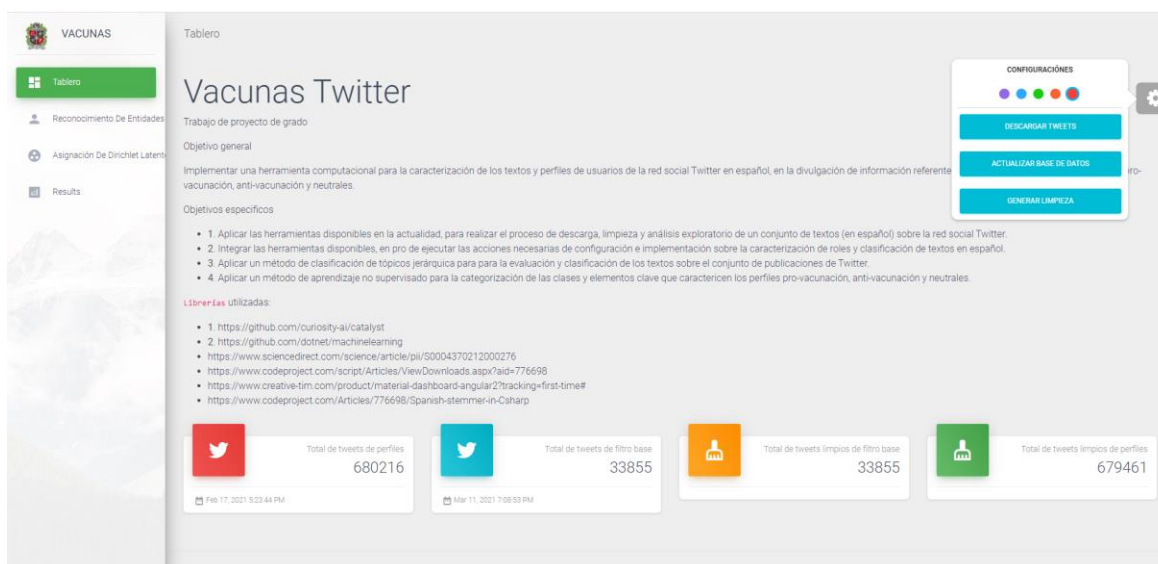
En esta sección, se encontrará, a la izquierda la barra de navegación de toda la aplicación; al centro la página principal de cada vista, en el tablero principal se ubicará la vista general de los objetivos del proyecto, así mismo, contadores de estado de la base de datos (fechas de actualización y cantidad de documentos almacenados). En las configuraciones a la derecha, se encontrarán 3 funcionalidades (ver Ilustración 26):

1. Descargar tweets: Cumple con la fase 1 de la metodología y la sección “Descarga automática - Script en Bash”. Descargar desde Twitter los tweets relacionados con los filtros preestablecidos por los expertos (ver tabla de palabras clave en Anexo:

Tabla de filtros Twitter) generando la base de datos de archivos separados por perfiles y por tweets base, relacionados con los perfiles obtenidos.

2. Actualizar base de datos: Cumple con la fase 1 de la metodología y la sección “Alistamiento de datos”. Obtener los tweets desde los archivos generados y cargar los datos en la base de datos.
3. Generar limpieza: Cumple con la fase 1 de la metodología sobre la limpieza de los datos basados en la sección 4.2.2 Limpieza de los datos.

Ilustración 26 Home aplicación Web



Fuente: Elaboración propia

7.2.2 Reconocimiento de entidades (NER)

En esta sección, se encuentran las bolsas de palabras resultantes de los cálculos sobre las bases de datos generadas, cumpliendo con la Fase 2 - Sección bolsa de palabras de

la metodología, así mismo, el botón para ejecutar el entrenamiento de NER y una tabla de clarificación sobre el etiquetado de las POS (partes del discurso) (ver Ilustración 27).

Ilustración 27 Vista del reconocimiento de entidades en la aplicación Web

ID	Simplificación	Nombre	Significado
1	ADJ	Adjetivo	una cualidad de un sustantivo
2	ADP	edición	es un término de cobertura para preposiciones y posposiciones
3	ADV	Adverbio	describe un verbo, un adjetivo u otro adverbio
4	AUX	Auxiliar	es una palabra funcional que acompaña al verbo léxico de una frase verbal y expresa distinciones gramaticales que no lleva el verbo léxico

Fuente: Elaboración propia

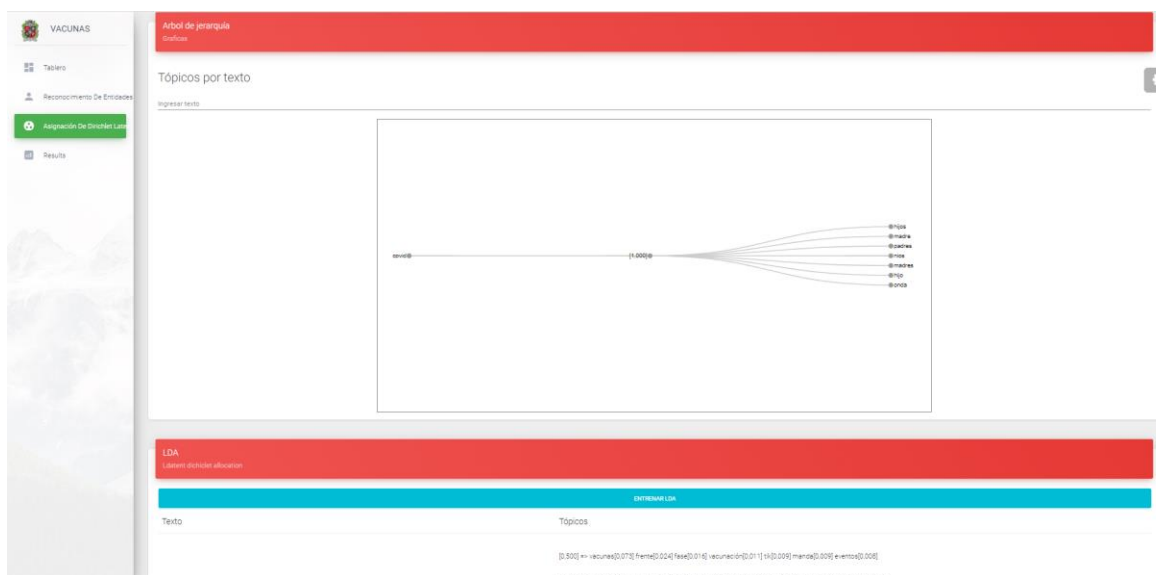
El entrenamiento preliminar en esta sección es vital para la presentación de resultados, ya que, por medio del botón “entrenar NER”, se ejecuta el entrenamiento del perceptrón multicapa para el nombramiento de entidades conocidas, en principio, reconoce fielmente lugares y URLs, sin embargo, no presenta gran estabilidad en el nombramiento de personas y se deben ejecutar varios entrenamientos hasta evidenciar una estabilidad importante.

Del lado de las bolsas de palabras, se representa fielmente la dispersión de las palabras para la gráfica “Word cloud perfiles”, ya que, recordando la metodología, en esta base de datos se almacenan los perfiles completos de los usuarios que participaron de opiniones sobre vacunación, encontrados en la base de datos “base”, extraída de Twitter. Por otro lado, en la gráfica “Word cloud base”, encontramos palabras como “covid”, “AstraZeneca”, “Bill Gates”, “europa”, “Jhonson”, etc. Que evidencian temáticas claras, relacionadas con la pandemia en curso, las vacunas, las fábricas de vacunas e incluso los personajes relevantes.

7.2.3 Asignación Latente de Dirichlet (LDA)

En esta sección, se encuentran el árbol dinámico de detección de tópicos resultantes del entrenamiento de LDA, explicado en la sección de modelamiento de textos, cumpliendo con la Fase 2 - Sección bolsa de palabras de la metodología, así mismo, se muestra un botón para ejecutar el entrenamiento de LDA y una tabla de resultados de prueba sobre el algoritmo (ver Ilustración 28).

Ilustración 28 Vista de Asignación latente de Dirichlet (LDA)



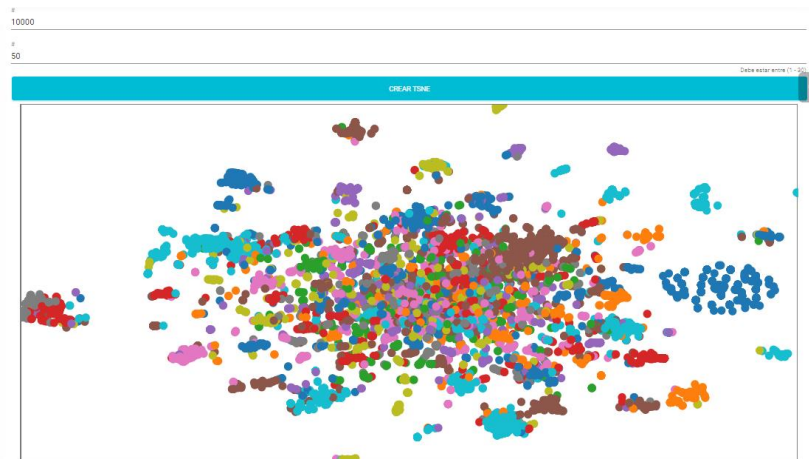
Fuente: Elaboración propia

Como resultados preliminares, encontramos agrupaciones de tópicos latentes, que resultaron interesantes para el grupo de investigación relacionado con el presente trabajo, se probó el algoritmo de LDA con el texto “covid19 afecta familias vulnerables” detectando como tópicos los textos: “mundial”, “salud”, “enfermedades”, “director”, “organización”, “evidencia”, “familias”; e incluso personajes o entidades relevantes, tales como: “Colombia”, “Duque” y “BluRadio”. En cuanto a la evaluación de los perfiles pro vacunación anti vacunación o neutrales se hizo referencia a ser una herramienta útil para para la detección de estos, simplificando el análisis de discurso a nivel cualitativo.

7.2.4 T-SNE

En la Ilustración 29 se identifica el resultado de la evaluación de 10000 documentos con una perplejidad de 50, para la visualización de baja dimensionalidad (en 2D) de los temas relacionados, donde cada documento se representa como un vector de 300 dimensiones sobre las probabilidades de los temas relacionados al texto.

Ilustración 29 Grafica de T-SNE para 10000 documentos con una perplejidad de 50



Fuente: Elaboración propia

La transformación de los datos de alta dimensionalidad a baja dimensionalidad se representa por la Tabla 6 y la Tabla 7, donde, respectivamente tenemos N documentos con M dimensiones, que representan las probabilidades de los temas por documento. Y se reduce a vectores con 3 características, la posición del documento en x y y con el color de su tema más probable.

Tabla 6 Representación vector de documentos de alta dimensionalidad

	PROBABILIDAD TOPICO 1	PROBABILIDAD TOPICO 2	...	PROBABILIDAD TEMA M
DOC 1	$P(D1 \mid \alpha, \beta)$ para tema 1	$P(D1 \mid \alpha, \beta)$ para tema 2	...	$P(D1 \mid \alpha, \beta)$ para tema M
DOC 2	$P(D2 \mid \alpha, \beta)$ para tema 1	$P(D2 \mid \alpha, \beta)$ para tema 2	...	$P(D2 \mid \alpha, \beta)$ para tema M
...
DOC N	$P(DN \mid \alpha, \beta)$ para tema 1	$P(DN \mid \alpha, \beta)$ para tema 2	...	$P(DN \mid \alpha, \beta)$ para tema M

Tabla 7 Representación de vector de documentos de baja dimensionalidad

	DIM 1	DIM 2	COLOR
DOC 1	Punto x para doc 1	Punto y para doc 1	tema mas probable para DOC 1
DOC 2	Punto x para doc 2	Punto y para doc 2	tema mas probable para DOC 2
...
DOC N	Punto x para doc N	Punto y para doc N	tema mas probable para DOC N

8 Conclusiones y recomendaciones

Conclusiones

Se ha logrado desarrollar con éxito la metodología propuesta. La herramienta computacional y el resultado de los análisis fueron expuestos al grupo de investigación GEESP, quienes señalaron la relevancia que lograrán tener los modelos previsualizados en la herramienta para la detección de perfiles pro-vacunación, anti-vacunación y neutrales en la red social Twitter en español. El resultado del conversatorio con este grupo de investigación se expone en el anexo: “tabla de evaluación pertenencia sobre la herramienta”, con el que, cooperativamente, se ha logrado evaluar la herramienta desde la perspectiva de la salud pública.

Durante el establecimiento del modelo computacional, se identificó, en la fase de limpieza de los tweets, una reducción del 43% de los datos extraídos de la red social, se inició con aproximadamente 1'200.000 de registros y culminamos con 680.216, lo cual confirmó la relevancia en la limpieza continua del conjunto de datos/documentos, dado que estos tweets eliminados del estudio representaban una gran cantidad de palabras vacías o documentos que no representarían ganancia de información por parte de los modelos propuestos, llegando a afectar la medición y rendimiento final de la herramienta.

Uno de los retos que se presentó a nivel técnico fue la integración de diferentes herramientas desde el backend de la aplicación, dado que, a pesar de utilizar herramientas previamente probadas, ejecutadas e implementadas, cada una de ellas contaba con maneras diferentes de parametrizar e inicializar los modelos, e incluso, en muchas ocasiones no eran parametrizables, sino obedecían a modelo fijos sin acceso público al código fuente o escasa documentación. La librería principal para los temas asociados a NLP, llamada Catalyst [30] fue la más completa a nivel de funcionalidades y

parametrización para el framework de .NET, sin embargo, la documentación es muy escasa y fue necesario entrar al código fuente para entender el funcionamiento interno.

Como conclusión adicional, se han logrado abordar diferentes tipos de retos durante el diseño y desarrollo de la herramienta computacional, no solo a nivel técnico, sino funcional, fue necesario elaborar varias sesiones con usuarios finales, que llegarán a usar la herramienta, para ejecutar controles de salud pública en términos de red sociales; en especial, y sumando la pandemia del covid-19 en curso, se logró evidenciar comportamientos en la red social fácilmente identificables desde la interacción de los usuarios, sin embargo, la herramienta debe seguir evolucionando conforme al volumen de datos obtenidos y las necesidades de los usuarios finales.

Recomendaciones

Se recomienda darle continuidad a la herramienta, dado que no se identificó ninguna otra implementación similar para ejecutar planes de acción, clasificación o identificación de perfiles de vacunación en términos de salud pública sobre las redes sociales, llegando a ser estos estudios vitales para la creación de políticas públicas fundamentadas en estadística, aprendizaje de máquina y análisis profundos no perceptibles para el ojo humano, tales como la identificación de tópicos u entidades relacionadas con los estudios.

En un futuro, se sugiere llevar esta implementación de la herramienta computacional a la nube, para crear un entorno abierto al público que sea accesible, modificable, de código abierto y evolutivo a través del tiempo; dado que actualmente funciona de manera local y cerrado al grupo de investigación que nos ha apoyado en la elaboración del presente trabajo.

A. Anexo: Tabla de evaluación de pertinencia sobre la herramienta

Vacunas Twitter				
Categorías de Evaluación	Excelente	Satisfactorio	Mejorable	Insuficiente
Utilización de filtro o lista de palabras	El filtro es coherente y logra depurar contenido acorde con la lista de palabras proporcionada por el experto.	El filtro es lógico, depura gran cantidad de contenido similar a la lista de palabras proporcionada por el experto.	El filtro relaciona algunas palabras con la lista de palabras proporcionada por el experto.	El filtro capta pocas palabras de la lista de palabras proporcionada por el experto.
Perfiles	Los perfiles tienen toda la información del usuario y permite identificar características de la cuenta	Los perfiles tienen información del usuario y algunas características de la cuenta	Los perfiles tienen información del usuario, pero no tiene características de la cuenta	Los perfiles no tienen información ni del usuario ni de la cuenta confiable
Tuits	Los tuits contienen las palabras clave del tema de vacunación	Los tuits refieren el tema de la vacunación	Los tuits tocan el tema de la vacunación periféricamente	Los tuits no tienen ninguna de las palabras clave del tema de vacunación
Modelo multivista red neuronal, nube palabras	El modelo proporciona insumos suficientes para hacer un análisis del discurso y realizar inferencias sobre las publicaciones	El modelo proporciona insumos para hacer una aproximación al análisis del discurso sobre las publicaciones	El modelo hace leve una aproximación a temas relacionados con la lista de palabras proporcionada por el experto	El modelo no logra captar ninguna de las palabras de palabras proporcionada por el experto

Bolsa de palabras cuantitativo	Alta frecuencia de palabras relacionadas con las palabras clave proporcionadas por el experto	Media Frecuencia de palabras relacionadas con las palabras clave proporcionadas por el experto	Baja frecuencia de palabras relacionadas con las palabras clave proporcionadas por el experto	Ningún porcentaje de palabras relacionadas con las palabras clave proporcionadas por el experto
Visualización	Es visualmente atractiva y de fácil manejo	Es visualmente agradable y el manejo es amigable	Visualmente es aburrida y no es fácil de manejar	No es atractiva visualmente y no se puede manejar
Interpretación	La datos que extrae la aplicación son relevantes y de alto impacto para realizar análisis cualitativo y cuantitativo para temas de Salud Pública	La datos que extrae la aplicación son útiles para realizar análisis cualitativo y cuantitativo para temas de Salud Pública	La datos que extrae la aplicación son de difícil entendimiento y es difícil realizar análisis cualitativo y cuantitativo para temas de Salud Pública	La datos que extrae la aplicación no son útiles ni comprensibles para realizar análisis cualitativo y cuantitativo para temas de Salud Pública
Utilidad	Identifica elementos completos y explícitos de publicaciones sobre vacunación en la red social Twitter y conduce al análisis sobre las preferencias de los usuarios frente a la vacunación y el impacto en la Salud Pública	Reconoce palabras relacionadas con la vacunación en la red social Twitter y sugiere las tendencias sobre las publicaciones de vacunas y el impacto en la Salud Pública	Encuentra algunas publicaciones aisladas relacionadas con vacunación	Percibe pocas publicaciones relacionadas con vacunación

B. Anexo: Tabla de filtros Twitter

Vacunas
Vacuna
Epidemia
Cura
Vacunacion
farmaceuticas
implantar chip
causan autismo
Vacunas causan
Vacunacion libre
Vacunacion Informada
vacunación infantil
Seguridad vacunas
Sarampión
Virus
Resistencia vacunacion
rechazo vacunacion
antivacunas
antivacuna
anti-vacuna
anti-vacunas
programas vacunacion
Maternidad vacunas
peligros vacunas
inmunizacion

Inmunidad rebaño
Inmunidad grupo
creencias vacunas
Estigma vacunacion
escepticismo vacunacion
escepticismo cientifico
Pandemia
Efectos secundarios vacuna
Efectos devastadores vacuna
Efectos colaterales vacuna
Efectos adversos vacuna
Efectividad vacuna
dudas vacuna
indecision vacuna
Beneficio vacuna
Afectado vacuna
Activismo vacuna

C. Anexo: Tabla de vector de características Spacy

<code>int _HashBias = GetIgnoreCaseHash("bias");</code>
<code>int _HashISuffix = GetIgnoreCaseHash("i suffix");</code>
<code>int _HashIPrefix = GetIgnoreCaseHash("i pref1");</code>
<code>int _HashIShape = GetIgnoreCaseHash("i shape");</code>
<code>int _HashIm1Suffix = GetIgnoreCaseHash("i-1 suffix");</code>
<code>int _HashIp1Suffix = GetIgnoreCaseHash("i+1 suffix");</code>
<code>int _HashIm1Shape = GetIgnoreCaseHash("i-1 shape");</code>
<code>int _HashIp1Shape = GetIgnoreCaseHash("i+1 shape");</code>
<code>int _HashIm1TagIword = GetIgnoreCaseHash("i-1 tag i word");</code>
<code>int _HashIm2Word = GetIgnoreCaseHash("i-2 word");</code>
<code>int _HashIp1Word = GetIgnoreCaseHash("i+1 word");</code>
<code>int _HashIWord = GetIgnoreCaseHash("i word");</code>
<code>int _HashIm1Word = GetIgnoreCaseHash("i-1 word");</code>
<code>int _HashIp2Word = GetIgnoreCaseHash("i+2 word");</code>
<code>int _HashIm1Tag = GetIgnoreCaseHash("i-1 tag");</code>
<code>int _HashIm2Tag = GetIgnoreCaseHash("i-2 tag");</code>
<code>int _HashITagIm2Tag = GetIgnoreCaseHash("i tag i-2 tag");</code>
<code>int _HashIPOS = GetIgnoreCaseHash("i pos");</code>
<code>int _HashIm1POS = GetIgnoreCaseHash("i-1 pos");</code>
<code>int _HashIm2POS = GetIgnoreCaseHash("i-2 pos");</code>
<code>int _HashIp1POS = GetIgnoreCaseHash("i+1 pos");</code>
<code>int _HashIp2POS = GetIgnoreCaseHash("i+2 pos");</code>

Bibliografía

- [1] G. Pérez-Gaxiola, G. V. Castrejón-García, N. León-Sicairos, and C. A. Cuello-García, "Internet y vacunas: análisis de su uso por padres de familia, sus percepciones y asociaciones," *Salud publica de Mexico*, vol. 58, no. 6, Instituto Nacional de Salud Pública, pp. 586–587, Nov. 2016.
- [2] G. Nigenda-López, E. Orozco, and R. Leyva, "Motivos de no vacunacion: Un analisis critico de la literatura internacional, 1950-1990," *Rev. Saude Publica*, vol. 31, no. 3, pp. 313–321, Jun. 1997, doi: 10.1590/s0034-89101997000300015.
- [3] A. Mendoza-Mendoza, K. Cervantes De La Torre, and E. De La Hoz Domínguez, "Programas de vacunación infantil en América Latina, 2000-2015," *Rev. Cuba. Salud Pública*, vol. 45, 2019.
- [4] O.-M. M. D. la H.-R. F. Escobar-Díaz F, "Motivos de no vacunación en menores de cinco años en cuatro ciudades colombianas," *Rev Panam Salud Publica*, vol. 41, 2017, doi: 10.26633/RPSP.2017.123.
- [5] C. Zhang Meadows Phd, L. Tang Phd, and W. Liu, "Twitter message types, health beliefs, and vaccine attitudes during the 2015 measles outbreak in California," *AJIC Am. J. Infect. Control*, vol. 47, pp. 1314–1318, 2019, doi: 10.1016/j.ajic.2019.05.007.
- [6] L. Tavošchi *et al.*, "Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy," *Hum. Vaccines Immunother.*, 2020, doi: 10.1080/21645515.2020.1714311.
- [7] G. J. Kang *et al.*, "Semantic network analysis of vaccine sentiment in online social media," *Vaccine*, 2017, doi: 10.1016/j.vaccine.2017.05.052.
- [8] G. Bello-Orgaz, J. Hernandez-Castro, and D. Camacho, "Detecting discussion communities on vaccination in twitter," *Futur. Gener. Comput. Syst.*, vol. 66, pp. 125–136, 2017, doi: 10.1016/j.future.2016.06.032.
- [9] M. S. Deiner *et al.*, "Facebook and Twitter vaccine sentiment in response to

- measles outbreaks,” *Health Informatics J.*, vol. 25, no. 3, pp. 1116–1132, Sep. 2019, doi: 10.1177/1460458217740723.
- [10] C. Cabezas-Sánchez, “Emerging and re-emerging infectious diseases and their determinants,” *Revista Peruana de Medicina Experimental y Salud Publica*, vol. 32, no. 1, pp. 7–8, 2015, doi: 10.17843/rpmesp.2015.321.1567.
- [11] Benavides-Arias, “Evaluación del rol de las redes sociales en la promoción de la salud y el fenómeno de rechazo a la vacunación. Tesis para optar al grado de magister en salud pública. Universidad Nacional de Colombia – Sede Bogotá.” 2020.
- [12] U. Cuesta Cambra, L. Martínez Martínez, and J. I. Niño González, “Análisis de la información pro vacunas y anti vacunas en redes sociales e internet. Patrones visuales y emocionales,” *El Prof. la Inf.*, vol. 28, no. 2, 2019.
- [13] F. Godlee, “What should we do about vaccine hesitancy?,” *BMJ*, vol. 365, Jun. 2019, doi: 10.1136/bmj.l4044.
- [14] S. E. Williams, “What are the factors that contribute to parental vaccine-hesitancy and what can we do about it?,” *Hum. Vaccines Immunother.*, vol. 10, no. 9, pp. 2584–2596, Sep. 2014, doi: 10.4161/hv.28596.
- [15] K. Vance, W. Howe, and R. P. Dellavalle, “Social Internet Sites as a Source of Public Health Information,” *Dermatol. Clin.*, vol. 27, no. 2, pp. 133–136, Apr. 2009, doi: 10.1016/j.det.2008.11.010.
- [16] CONGRESO DE LA REPUBLICA DE COLOMBIA, *LEY NÚMERO 1122 DE 2007*. 2007.
- [17] C. C. Ubaldo, “Las TICs y la salud desde una perspectiva psicosocial,” *Rev. Comun. y Salud RCyS*, vol. 2, 2012.
- [18] C. C. Ubaldo and G. H. Sandra, “La ‘reputación online’ de la información de vacunas en internet,” *Hist. y Comun. Soc.*, vol. 19, 2014.
- [19] C. C. Ubaldo, C. D. Victoria, and G. H. Sandra, “Vacunas y anti vacunas en la red social Youtube,” *Opción Rev. Ciencias Humanas y Soc.*, vol. 9, 2016.
- [20] J. Francisco Ávila de Tomás, F. Benito Justel Rafael Fernando Beijinho do Rosario, and M. España, “El e-paciente,” 2013.
- [21] C. P.-S. D. Catalán-Matamoros, “Medios y desconfianza en vacunas: un análisis de contenido en titulares de prensa,” *Rev. Lat. Comun. Soc.*, vol. 74, pp. 786–802, 2019.

- [22] Digital 2021, “Digital 2021: Global Overview Report — DataReportal – Global Digital Insights,” 2021. <https://datareportal.com/reports/digital-2021-global-overview-report>.
- [23] R. M. Wolfe, L. K. Sharp, and M. S. Lipsky, “Content and design attributes of antivaccination Web sites,” *J. Am. Med. Assoc.*, vol. 287, no. 24, pp. 3245–3248, Jul. 2002, doi: 10.1001/jama.287.24.3245.
- [24] S. Wasserman and K. Faust, *Social Network Analysis*. Cambridge University Press, 1994.
- [25] F. J. Garrido, “El análisis de redes en el desarrollo local,” in *Prácticas locales de la creatividad social*, no. 1, 2001, pp. 49–62.
- [26] J. M. F. Vivar, “Nuevos modelos de comunicación, perfiles y tendencias en las redes sociales,” *Comunicar*, vol. 16, no. 33, pp. 73–81, 2009, doi: 10.3916/c33-2009-02-007.
- [27] J. Guix Oliver, “El análisis de contenidos: ¿Qué nos están diciendo?,” *Rev. Calid. Asist.*, vol. 23, no. 1, pp. 26–30, Jan. 2008, doi: 10.1016/S1134-282X(08)70464-0.
- [28] D. Ramachandran and R. Parvathi, “Analysis of Twitter Specific Preprocessing Technique for Tweets,” in *Procedia Computer Science*, Jan. 2019, vol. 165, pp. 245–251, doi: 10.1016/j.procs.2020.01.083.
- [29] J. Yuan *et al.*, “LightLDA: Big Topic Models on Modest Compute Clusters.” [Online]. Available: www.petuum.org.
- [30] curiosity-ai, “Repositorio codigo fuente curiosity-ai/catalyst.” GitHub, 2021, [Online]. Available: <https://github.com/curiosity-ai/catalyst>.
- [31] D. M. Blei, A. Y. Ng, and J. B. Edu, “Latent Dirichlet Allocation Michael I. Jordan,” 2003.
- [32] A. I. Domingues Fernandes, “A Deep Learning Approach to Named Entity Recognition in Portuguese Texts,” FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO, 2018.
- [33] L. Buitinck and M. Marx, “Two-stage named-entity recognition using averaged perceptrons,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7337 LNCS, pp. 171–176, doi: 10.1007/978-3-642-31178-9_17.
- [34] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, “Learning

- multilingual named entity recognition from Wikipedia,” *Artif. Intell.*, vol. 194, pp. 151–175, Jan. 2013, doi: 10.1016/j.artint.2012.03.006.
- [35] J. Votrubec, “Morphological Tagging Based on Averaged Perceptron,” *Proc. Contrib. Pap.*, vol. 1, pp. 191–195, 2006, [Online]. Available: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.186.6832&rep=rep1&type=pdf>.
- [36] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” 2008. [Online]. Available: <https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>.
- [37] MongoDB, “Documentación de MongoDB C#/.NET Driver.” 2021, [Online]. Available: <https://docs.mongodb.com/drivers/csharp/>.
- [38] CodeProject, “Codigo fuente para: Spanish stemmer in C#.” 2021, [Online]. Available: <https://www.codeproject.com/Articles/776698/Spanish-stemmer-in-Csharp>.
- [39] Microsoft, “Documentación de ML.NET | Aprendizaje maquina para .NET.” 2021, [Online]. Available: <https://dotnet.microsoft.com/apps/machinelearning-ai/ml-dotnet>.
- [40] Twitter4J, “Documentación de Twitter4J - A Java library for the Twitter API.” 2021, [Online]. Available: <http://twitter4j.org/en/>.
- [41] Creative Tim, “Codigo fuente de Material Dashboard Angular: Admin Template.” Web page, 2021, [Online]. Available: <https://www.creative-tim.com/product/material-dashboard-angular2?tracking=first-time#>.
- [42] J. R. Franco Sánchez, “Repositorio del codigo fuente del proyecto desarrollado durante el trabajo de grado (VacunasTwitter).” GitHub, 2021, [Online]. Available: <https://github.com/jhhfrancos/VacunasTwitter>.