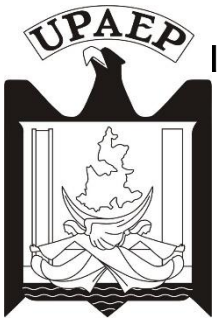

l Popular Autónoma del Estado de Puebla

Decanato de Ingenierías

Facultad de Tecnologías de Información y Ciencia de Datos

Doctorado en

MODELO PARA EL ANÁLISIS DE DATOS Y ANÁLISIS SENTIMIENTOS DENTRO DE LAS CONVERSACIONES EN TWITTER MEDIANTE ALGORITMOS DE PROCESAMIENTO DE LENGUAJE NATURAL Y DICCIONARIOS DE DATOS

Tesis que para obtener el Grado de Doctor

en Tecnologías de la Información y Negocios Electrónicos

Presenta

Juan Ignacio Rivas González

Director

Dr. Miguel Angel Sánchez Acevedo

Co directora

Dra. María del Rocío Guadalupe Morales Salgado

Puebla, México. Mes 2022

CONTENIDO

CAPÍTULO 1. INTRODUCCIÓN

Los trending topics[1] son palabras o frases breves utilizadas, con más frecuencia, por los usuarios de las redes sociales, en un determinado intervalo de tiempo o zona geográfica, reflejando los temas de mayor interés dentro de las redes sociales. El origen de los trending topics comenzó como una clasificación de etiquetas identificadas por almohadillas (#), las cuales nos ayudan a identificar qué tema tiene más flujo de información y proporcionan una radiografía de lo que se habla en redes sociales.

Las tendencias o temas de interés pueden identificarse mediante un algoritmo, que es un conjunto de instrucciones definidas para ejecutar determinadas acciones. El algoritmo, en el universo de redes sociales, conoce los temas que presentan popularidad en un momento dado y los posiciona para que los usuarios los puedan consultar fácilmente; sin embargo, las tendencias necesitan un mejor mecanismo de análisis que nos permita tener datos más precisos y útiles. En este trabajo, diseñar un modelo para el análisis de datos y análisis sentimientos dentro de las conversaciones en Twitter mediante algoritmos de Procesamiento de Lenguaje Natural y diccionario de datos, lo que nos ayudará a proporcionar más información para la interpretación de los resultados y facilitar el consumo de los datos. Para fines de esta investigación, analizaremos el tema del COVID-19, debido a su pertinencia y presencia constante en redes sociales, que ha permanecido en la opinión pública y diariamente tenemos información nueva y relevante del tema que, sin duda, ha afectado los distintos ámbitos sociales de manera global.

1.1 DESCRIPCIÓN DEL PROBLEMA

En diciembre de 2019, un conjunto de casos de neumonía causados por un nuevo coronavirus, COVID-19, fue identificado en Wuhan, China. En los meses siguientes, los casos se fueron expandiendo rápidamente alrededor del mundo, llegando a propagarse a más de 200 países. El 11 de marzo del 2020 la Organización Mundial de la Salud (OMS) declaró el COVID-19 como pandemia, impactando a más de 200 países alrededor del planeta, infectando a más de 20 millones de personas y causando cerca de 750,000 muertes a agosto de 2020 (Worldometer, s.f.).

Esta emergencia de salubridad sin precedente ha resultado en respuestas políticas y sociales nunca antes vistas y en una afectación severa de la economía global. La avalancha de respuesta por parte de la sociedad ha sido facilitada por el flujo de información transmitida por los medios tradicionales, pero en particular por el flujo de información en las redes sociales.

Durante la pandemia, en Twitter se intensificó el flujo de información, sobre todo lo que tuviera que ver con la situación que en ese momento era inédita. Twitter se convirtió en una fuente invaluable de noticias y temas referentes al COVID-

[1] **Trending Topic**: (tendencia, tema de tendencia o tema del momento en español, y TT en forma abreviada) son los temas mas populares de Twitter en cada momento, aquellos que están siendo utilizados por un gran numero de usuarios de la red social. (Berenguer, 2019) (Berenguer, 2019)

19, influenciado también porque ante las medidas de sana distancia y aislamiento, comunicarse mediante redes sociales se convirtió en una herramienta muy útil para las instituciones, organizaciones, gobiernos y para la ciudadanía.

Analizar el tema del COVID-19 en Twitter es de gran utilidad para conocer cómo ha evolucionado la percepción de la pandemia; así como evaluar distintas medidas, situaciones y avances en cuanto al tema. Sin embargo, el proceso de analizar, filtrar, clasificar y resumir de manera manual la gran cantidad de información que proporciona la red social es una tarea que puede resultar tediosa, larga y compleja. Por consiguiente, es necesario plantear un modelo que optimice la precisión del análisis en tendencias y se complemente con análisis de sentimientos. En este sentido nos beneficiaremos de Universal Sentence Encoder (USE_T), del que daremos cuenta en las páginas que siguen, refiriéndonos primero a aquellas tecnologías que sustentan a dicho modelo. En la actualidad, se continúan realizando estudios para la mejora de los mecanismos de automatización de análisis de datos para el procesamiento óptimo de la información, debido a que los resultados que se obtienen hoy en día pudieran mejorar en cuanto a calidad y utilidad.

Con la finalidad de resolver estos problemas, proponemos un modelo para el análisis de datos y análisis sentimientos dentro de las conversaciones en Twitter mediante algoritmos de Procesamiento de Lenguaje Natural y diccionario de datos, que nos facilite la interpretación de lo que les gusta y que no les gusta y con ello lograr una mejor precisión del resultado de tendencias.

1.2 JUSTIFICACIÓN

En la actualidad, las redes sociales como Facebook, Twitter e Instagram están jugando un papel importante al permitir a sus usuarios compartir información y su sentir acerca de situaciones determinadas. Las redes sociales son reconocidas por ser un gran almacén de datos que pueden conducir a la predicción de fenómenos relacionados con algún evento. Por ejemplo, Lampos y Cristiani (2010) demostraron que la información de los micro blogs facilita y mejora la vigilancia de la salud pública, y aporta a la predicción del número de pacientes que sufren de influenza.

En México, de acuerdo al Instituto Nacional de Estadística (INEGI), 39% de los cibernautas utiliza Twitter, lo que corresponde a 31.4 millones de usuarios. Según diversos estudios, estamos dentro de los primeros 10 lugares entre los países con mayores usuarios; estas referencias confirman que México es uno de los países más activos que usa la red social. En efecto, la red social de Twitter es excelente fuente de información que permite ser objeto de estudio por medio del análisis de datos (Martinez, 2020).

A lo largo de los años, Twitter se ha convertido en una herramienta que funciona como un notificador temprano o canal de comunicación de emergencia; también sirve como monitor de percepción y como servicio público de fuente de información de desastres o eventos catastróficos, tales como huracanes, terrorismo, tsunamis, terremotos, la influenza estacional, ébola, entre otros.

Monitorear las conversaciones públicas en Twitter ofrece un barómetro del sentimiento global. Esta información es particularmente valiosa, siempre y cuando tengamos en cuenta que el contexto siempre determinará cualquier resultado. En Twitter no solo encontramos datos, sino distintas realidades que cambian a una gran velocidad, lo cual hace que cualquier actividad en redes sociales sea casi impredecible.

Las conversaciones en Twitter pueden ser analizadas mediante técnicas de análisis de datos para obtener información y conocimiento acerca de los sentimientos que generan los usuarios de esa red (Stieglitz *et al.*, 2018). El análisis de datos se centra en el uso de herramientas de procesamiento de lenguaje natural y lingüística computacional para identificar y extraer información subjetiva de una fuente dada (Shilpa, 2017).

Si bien el análisis de datos en Twitter se ha vuelto muy popular, no deja de haber problemas para analizar las conversaciones no sólo en un contexto y formato dado, sino también al momento de elegir la herramienta de software para analizarlos (Canhoto and Padmanabhan, 2015). Este análisis en las redes sociales proclama combinar, extender y adaptar métodos para mejorar el entendimiento de las publicaciones; sin embargo, todavía existe mucha discusión sobre cuál es el mejor modelo para analizarlos (Stieglitz *et al.*, 2018).

Uno de los mayores retos en el análisis de datos en Twitter es obtener datos de alta calidad. La mayoría de los datos obtenidos en esta red están incompletos o tienen ruido, lo que los hace de baja calidad; además de que, en alguno de los casos, las herramientas de recolección de datos no puedan manejar grandes cantidades de éstos, y en tiempo real se hace más difícil la tarea (Stieglitz et al., 2018).

1.3 OBJETIVOS

1.3.1 OBJETIVO GENERAL

Diseñar un modelo para el análisis de datos y análisis sentimientos dentro de las conversaciones en Twitter mediante algoritmos de Procesamiento de Lenguaje Natural y diccionario de datos.

1.3.2 OBJETIVOS ESPECÍFICOS

- **Documentar** el análisis de datos y sentimientos en Twitter, así como los métodos y formatos que existen para detectar y clasificar temas en las publicaciones para definir las tendencias.

- **Revisar** el contexto en el que se encuentra el análisis de datos y sentimientos, así como propuestas de tratamiento automático de textos en dominios generales y específicos para generar las tendencias.

- **Analizar** la naturaleza de las conversaciones en Twitter, que se generan alrededor de la temática del COVID –19 en cuanto similitud textual semántica de palabras clave y la polaridad de los sentimientos.

- **Diseñar** un método para la extracción de elementos básicos a partir de las publicaciones en Twitter sobre el COVID-19 que permita un registro estructurado de los hallazgos para una posterior explotación.

- **Diseñar** un diccionario de datos para la optimización de modelos de detección de tendencias en Twitter.

1.4 ALCANCE DE LA TESIS

La presente investigación tiene como meta obtener un diccionario aplicable a la red social Twitter, que sirva a las necesidades de análisis de datos y sentimientos dentro de la conversación del COVID-19. Este diccionario queda únicamente para fines demostrativos, para lo cual se ha escogido una muestra de publicaciones de Twitter sobre el COVID-19 en México, dejando la replicabilidad de la solución a otras redes sociales como trabajo futuro.

1.5 METODOLOGÍA

A continuación se presenta el modelo que describiremos en el capítulo 3 y que es la propuesta principal de esta tesis.

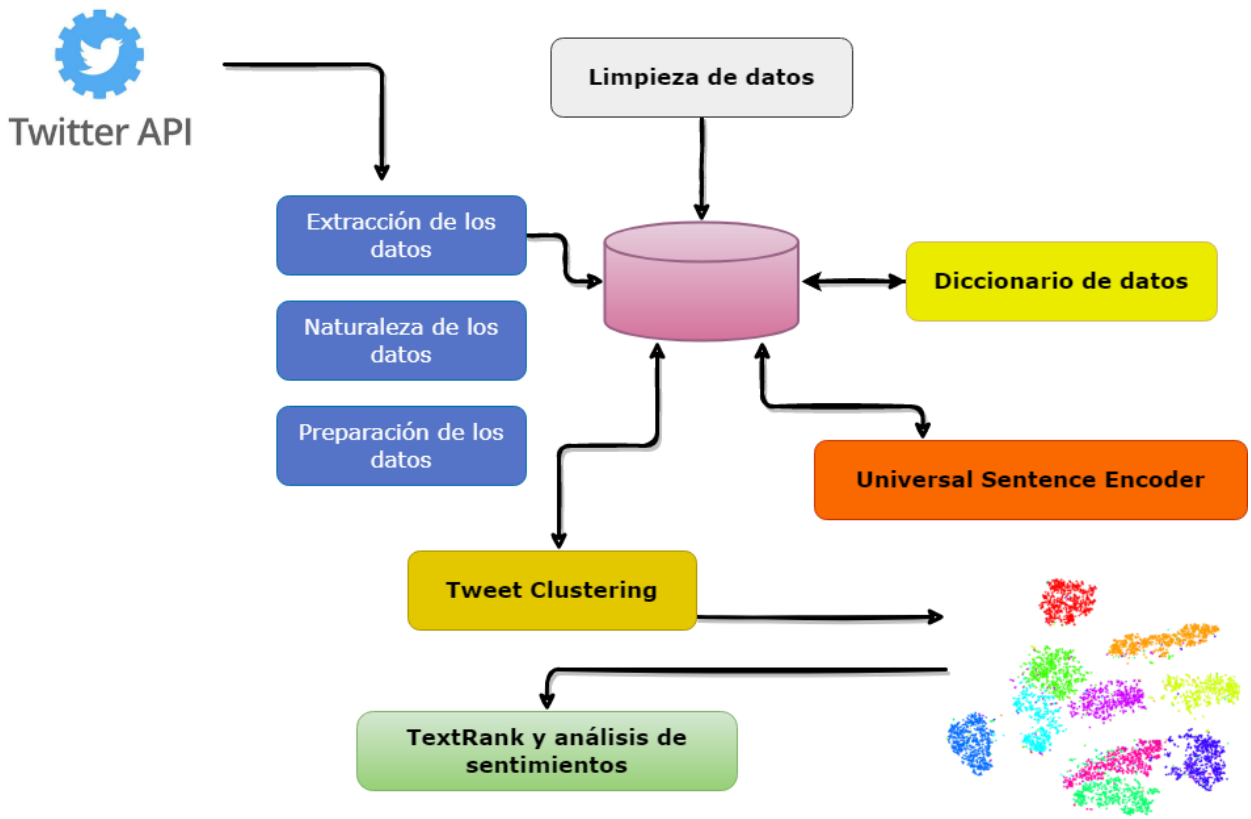

Figura 1. Modelo propuesto en esta tesis

Extracción de datos.- Se realiza la extracción de los datos mediante la API de Twitter.
Limpieza de datos.- Los twitts extraidos se limpian de caracteres especiales, espacios en blanco, emoticones, entre otros.
Universal Sentence Encoder.- se utiliza para convertir el texto en vectores para que se pueda procesar el texto de forma numérica.
Tweet Clustering.- agrupa las palabras o enunciados en común.
TextRank y análisis de sentimientos.- determina la polaridad positiva o negativa del texto.

1.6 ESTRUCTURA DEL DOCUMENTO

En el capítulo 2 se presenta una breve revisión del estado del arte del Procesamiento de Lenguaje Natural para el análisis de tendencias y de sentimientos en las conversaciones en Twitter. Se discuten las principales técnicas de Procesamiento de Lenguaje Natural, se explican sus principales métodos, se mencionan algunas de sus aplicaciones típicas. Finalmente, se documentan los alcances logrados en proyectos dedicados al análisis de la conversación del COVID-19 en Twitter.

En el capítulo 3 se presenta el modelo planteado, se describe cada uno de los pasos a realizar en el modelo y la justificación de la existencia de cada uno de ellos así como la relación que existe entre ellos para el correcto funcionamiento del modelo planteado.

En el capítulo 4 se presentan los resultados de la aplicación del modelo planteado en el capítulo anterior, se explica qué se hizo, cómo se aplicó y finalmente qué se obtuvo al realizarlo.

Por último, en el capítulo 5 se hace un análisis del resultado obtenido de modelo una vez aplicado, así como una comparativa con los modelos ya existentes mostrando las ventajas del modelo que se presenta en el presente documento contra los otros.

CAPÍTULO 2. MARCO TEÓRICO

Para este estudio, se ha realizado un proceso de documentación, que nos ayude a sustentar y conocer el contexto del tema dentro de la ciencia y la investigación. Para partir con esta investigación es fundamental, tener presente los siguientes conceptos:

La web 2.0 En una herramienta tecnológica, donde los usuarios se convierten en los protagonistas al alimentarla de contenidos, videos, imágenes, documentos, audios y todo lo que hoy en día conocemos y con ello va marcando la pauta para la mejora continua y actualización de la misma.

> *Para Marino Latorre, Investigador de la Universidad Marcelino Champagnat: "El uso de la web 2.0 está orientado a interactuar en redes sociales para crear contenido".(Latorre, 2018).*

Las redes sociales son aplicaciones de comunicación en un entorno social, en la que el usuario crea un perfil y exhibe publicaciones con el objetivo de mostrar comportamientos e ideas y expresarse con diferentes niveles de actividad y participación.

> *Para Graciela Padilla, "una red social es una aplicación basada en web 2.0" (Padilla, 2018).*

El **Procesamiento de Lenguaje Natural,** o por sus siglas PLN, es un motor de inteligencia artificial que encontramos en muchas aplicaciones modernas, como los motores de búsqueda en línea como Google o Microsoft Bing, los traductores como Google Translate, los correctores de estilo y ortografía de los procesadores de texto como Word, asistentes virtuales como Siri de Apple o Alexa de Amazon, entre otros. Este procesamiento, sintetiza el lenguaje con el que operan algunas aplicaciones.

> *Para Jackson y Moulinier : "El Procesamiento de Lenguaje Natural (PLN) es la función de software o hardware que modifica el lenguaje humano a lenguaje de computadora (2002)."*

La lingüística computacional es la encargada de diseñar modelos que analicen, interpreten e imiten las habilidades de lenguaje del ser humano a través de tecnologías.

> *Según Rubio López & Bernal Chávez (2016) "La lingüística computacional es el área que se encarga del estudio, diseño y elaboración de modelos y programas capaces de analizar, estudiar, automatizar e imitar las habilidades lingüísticas del ser humano a través de herramientas computacionales"*

El análisis de sentimientos (o minería de opiniones) es parte del estudio en la informática, inteligencia artificial y lingüística que se basa en las interacciones del lenguaje humano y la computadora, su objetivo es clasificar de forma masiva datos automáticamente. Es una tarea que implica básicamente determinar la

clasificación de una opinión u oración con una clasificación de polaridad ya sea positiva, negativa o neutra, sin embargo, podría clasificarse de una forma más avanzada o subjetivamente por ejemplo con estados emocionales como “enfado”, “tristeza”, “felicidad”, etc. por lo tanto, el objetivo del análisis de sentimientos es evaluar la actitud de los interlocutores.

> *Shilpa Balan, menciona en su publicación del 2017, titulada : Mining for social media que: ”que el análisis del sentimiento es el uso de herramientas de Procesamiento del Lenguaje Natural, análisis de texto y lingüística computacional para reconocer y obtener información subjetiva de una fuente determinada.”*
>
> *Para Stefan Stieglitza, “el análisis de las redes sociales implica combinar y ajustar métodos para el análisis de los datos de las redes; sin embargo, todavía existe mucha discusión sobre cuál es el mejor modelo para hacerlo” (2018)*

2.1 ANÁLISIS DE SENTIMIENTOS EN REDES SOCIALES

El sentimiento en redes sociales busca evaluar el nivel de impacto de tendencias en las redes sociales y clasificarlas permitiendo así conocer la opinión de los usuarios con un alto grado de certeza (Shilpa, 2017).

2.1.1. HERRAMIENTAS DE ANÁLISIS DE REDES SOCIALES

El análisis de redes sociales está enfocado en metodologías y tecnologías que transforman datos no estructurados de redes sociales en información de valor con un propósito de negocio (Aguirre, 2011).

Las herramientas de análisis de redes sociales son tecnologías que se alimentan de publicaciones y es necesario conocer su naturaleza para poder seleccionar la herramienta más adecuada.

Elementos que la conforman:

- Minería de datos
- Consultas avanzadas SQL
- Análisis de texto
- Procesamiento de Lenguaje Natural

2.1.1.3 NATURALEZA DE LAS PUBLICACIONES EN REDES SOCIALES

Son todos aquellos componentes, elementos básicos y estructura que forman parte de las publicaciones.

La información presentada aquí son las características particulares de las publicaciones, y los mecanismos de minería de datos son los procesos que permitirán clasificar, medir y presentar información de las publicaciones.

Elementos que lo conforman:

- Tipo de publicación
- Tipo de contenido

2.1.1.4 MECANISMOS DE MINERÍA DE DATOS PARA EL ANÁLISIS DEL SENTIMIENTO

Es el uso de herramientas de Procesamiento de Lenguaje Natural, análisis de texto y lingüística computacional para identificar y extraer información subjetiva de fuente dada. (Shilpa, 2017)

La minería de datos es la secuencia de extracción de información desde varias perspectivas para un uso de valor.

Elementos que lo conforman:

- Público meta
- Sentimientos

2.2 PROCESAMIENTO DE LENGUAJE NATURAL PARA EL ANÁLISIS DE INFORMACIÓN EN REDES SOCIALES

A continuación se presenta el estado del arte de las técnicas de Procesamiento del Lenguaje Natural (PLN), su importancia y el proceso que generalmente se sigue. Se detallan los modelos más destacados, su proceso de aprendizaje, aplicación en el procesamiento de datos y su uso en el análisis de redes sociales. Se describen los recursos que se requieren, los usos que se le pueden dar y los productos que pueden obtenerse. Se presenta, además, el estado del arte que guarda el Procesamiento de Lenguaje Natural en cuanto a proyectos de tratamiento de textos en redes sociales orientados al análisis de datos y sentimientos dentro de las conversaciones del COVID-19 en Twitter. En la Figura 2 se muestran las tres áreas del conocimiento que conforman la investigación y su intersección es el tema a tratar. Finalmente se comentan brevemente los resultados que se alcanzaron.

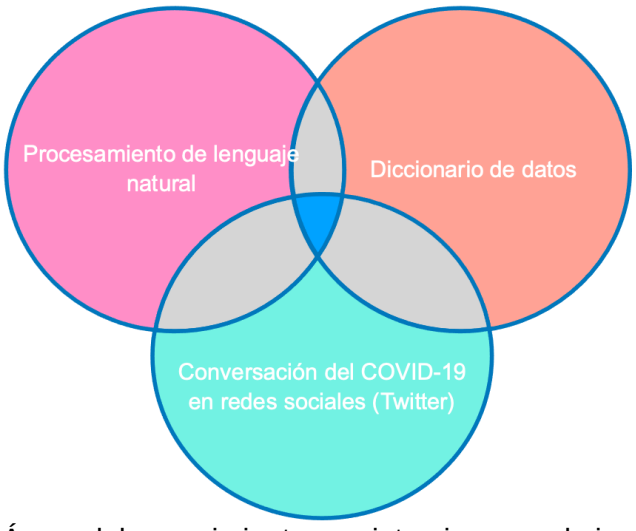

Figura 2. Áreas del conocimiento que intervienen en la investigación

2.2.1 PROCESAMIENTO DE LENGUAJE NATURAL

El Procesamiento de Lenguaje Natural (PLN) consiste de técnicas de computación para el análisis y representación automático del lenguaje humano. Las técnicas de PLN han evolucionado desde la era de procesamiento por lotes, en donde el análisis de una oración podía tomar hasta 7 minutos; actualmente en la era de Google, millones de páginas web pueden ser procesadas en menos de un segundo (Cambria and White, 2014). El PLN permite que una computadora pueda realizar una amplia cantidad de tareas relacionadas con el análisis del lenguaje a todos los niveles, como la clasificación gramatical de palabras hasta motores de traducción y sistemas de diálogo.

Por décadas, los problemas de PLN habían sido abordados con modelos de sombra, como Máquinas de Vectores de Soporte (SVM por sus siglas en inglés) y regresión logística. Las arquitecturas y algoritmos de Aprendizaje Profundo han logrado grandes avances en los campos de visión por computadora y reconocimiento de patrones. Investigaciones recientes de PLN se están enfocando al uso de métodos de Aprendizaje Profundo (Young *et al.*, 2018). Actualmente, son las redes neuronales basadas en representación de vectores densos las que han dado mejores resultados al momento de atender diversas tareas de PLN. Esta tendencia es promovida por el éxito de técnicas como Word Embeddings (Mikolov *et al.*, 2013) y Aprendizaje Profundo (Socher *et al.*, 2013). El Aprendizaje Profundo permite realizar de manera automática y en multiniveles las técnicas de aprendizaje de características. En contraste, los sistemas tradicionales de PLN que usan métodos de aprendizaje automático están fuertemente predeterminados a usar métodos convencionales (hand-crafted) de extracción de características. Estos métodos convencionales consumen demasiado tiempo y frecuentemente son inexactos (Young *et al.*, 2018).

En años recientes se han llevado a cabo numerosos esfuerzos para crear una representación semántica de una secuencia de texto, como podrían ser una oración, un párrafo o, inclusive, un documento entero (Young *et al.*, 2018). Estos métodos incluyen un amplio rango de técnicas como Word Movers Distance (Kusner *et al.*, 2015) Sentence-BERT (Reimers and Gurevych, 2019) y Universal Sentence Encoder (Cer *et al.*, 2018). Todos estos métodos tienen como objetivo obtener un vector que almacene el significado semántico de una oración y a su vez agrupar representaciones similares para oraciones parecidas.

2.2.2 TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL

En el siguiente apartado se presentan las técnicas que representan el estado del arte en el procesamiento de lenguaje natural. Para cada una de las técnicas abordadas se detalla su funcionamiento y las bases que siguen para realizar sus tareas de procesamiento, así como los resultados que ofrecen.

2.2.2.1 WORD MOVERS DISTANCE

Word Movers Distance (WMD) es una técnica novedosa que mide la distancia de disimilitudes entre dos documentos de texto para obtener la distancia mínima

entre palabras con significados/representaciones semejantes de un documento a otro (Kusner *et al.*, 2015).

Representar con exactitud la distancia entre dos documentos tiene múltiples campos de aplicación, como la recuperación de documentos (Salton and Buckley, 1988), categorizar y agrupar noticias (Greene and Cunningham, 2006), identificar canciones (Brochu and Freitas, 2002) y comparar documentos en diferentes lenguajes (Quadrianto, Song and Smola, 2009).

Las dos maneras más comunes en que se representa un documento es a través de una bolsa de palabras (bag of words o BOW) o por frecuencia de términos (term frequency-inverse document frequency oTF-IDF). Sin embargo, estas representaciones frecuentemente no son óptimas para medir la distancia entre documentos debido a su usual casi-ortogonalidad (Schölkopf *et al.*, 2002). Otro problema significativo con estas representaciones es que no se puede obtener la distancia entre palabras de manera individual. Tomemos como ejemplo las siguientes oraciones en dos diferentes documentos: *Obama habla a los medios en Illinois* y *El Presidente saluda a la prensa en Chicago*. Podemos ver que las dos oraciones no tienen palabras en común; sin embargo, nos comparten la misma información, este hecho no puede ser representado por el modelo BOW. En este caso, la cercanía entre las palabras: (Obama, Presidente), (habla, saluda), (medios, prensa) y (Illinois, Chicago) no es procesado entre las distancias obtenidas mediante BOW (Kusner *et al.*, 2015).

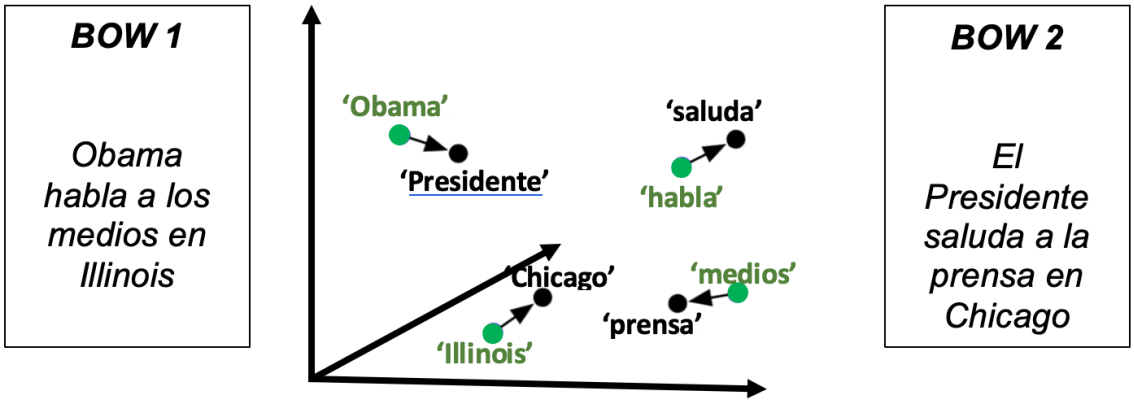

Figura 3. Funcionamiento de WORD MOVER'S DISTANCE

WMD toma provecho de las investigaciones realizadas por Mikolov (Mikolov *et al.*, 2013), quien asegura que su modelo Word2Vec genera significados/representaciones de palabras de gran calidad y puede ser usado de manera natural con grandes cantidades de datos. Los autores demostraron que la relación semántica es casi siempre preservada en operaciones de vectores con vectores de palabras, por ejemplo: vec(Berlin) - vec(Alemania) + vec(Francia) es cercano al vec(Paris). Esto sugiere que las distancias entre la representación/significado de un vector de palabras son en algún grado semánticamente significativas.

WMD usa las propiedades del modelo Word2Vec para representar documentos de texto como una nube de puntos cargados de significados/representaciones de palabras; así pues, la distancia entre dos documentos de texto A y B es la mínima distancia acumulativa que las palabras de la nube del documento A necesitan viajar para llegar al punto exacto de encuentro con sus similares en la nube del documento B (Kusner *et al.*, 2015).

La distancia obtenida por WMD tiene las siguientes propiedades:

1. Es libre de hyper-parámetros y sencilla de entender y usar.
2. Es altamente interpretable como la distancia entre dos documentos y, a su vez, como el espacio entre palabras individuales.
3. Incorpora el conocimiento del modelo Word2Vec y supera en exactitud a otras alternativas de punta en tareas reales de clasificación de documentos.

2.2.2.2 SENTENCE-BERT: SENTENCE EMBEDDINGS USING SIAMESE BERT-NETWORKS

BERT es una técnica basada en redes neuronales para el pre-entrenamiento del procesamiento del lenguaje natural, desarrollada por Google (Devlin *et al.,* 2018). Esta técnica incluye varios procedimientos de tareas de última generación de PLN: responder preguntas, clasificar oraciones y regresión de oraciones-pares, lo cual es parecido a la similitud semántica de textos. Sin embargo, para realizar la tarea de regresión de oraciones-pares se requiere que ambas oraciones sean ingresadas a la red BERT, lo que causa una sobre carga computacional; un ejemplo de esto es que, para encontrar dos oraciones similares en un texto de 10,000 oraciones, BERT requiere cerca de 65 horas de procesamiento en un CPU de última generación (Reimers and Gurevych, 2019). Esta desventaja no hace viable su uso para la búsqueda de semejanzas semánticas, así como para su uso en tareas no supervisadas.

Otra gran desventaja de la estructura de la red BERT es que no hace cálculos independientes de los significados/representaciones de una oración, lo que impide obtener significados/representaciones desde BERT. Para solucionar esta limitación, algunos investigadores ingresan las oraciones de manera individual a través de BERT y con los resultados individuales construyen un vector fijo, similar al que se obtiene al usar WMD (Reimers and Gurevych, 2019).

Sentence-Bert es una modificación de la red BERT que añade el uso de redes siameses y triples para obtener los significados/representaciones de las oraciones. Esto permite que BERT sea usado para tareas que antes eran obsoletas como comparaciones de similitudes semánticas a gran escala, agrupación y recuperación de información a través de búsquedas semánticas. Además, Sentence-Bert agrega una agrupación de operaciones a las salidas/resultados de BERT que permite obtener significados/representaciones de una oración. Con estas modificaciones se reduce el tiempo de procesamiento de un texto de 10,000 oraciones de 65 horas a 5 segundos y se mantiene la exactitud de resultados (Reimers and Gurevych, 2019). Un ejemplo de uso de Sentence-Bert es clasificar las oraciones conforme a la carga de su sentimiento (positivo o negativo) de acuerdo al contenido semántico de la oración; un ejemplo es tomar las opiniones publicadas sobre una película y procesarlas a través de

Sentence-Bert para obtener grupos de oraciones con opinión/sentimiento positivo y opinión/sentimiento negativo.

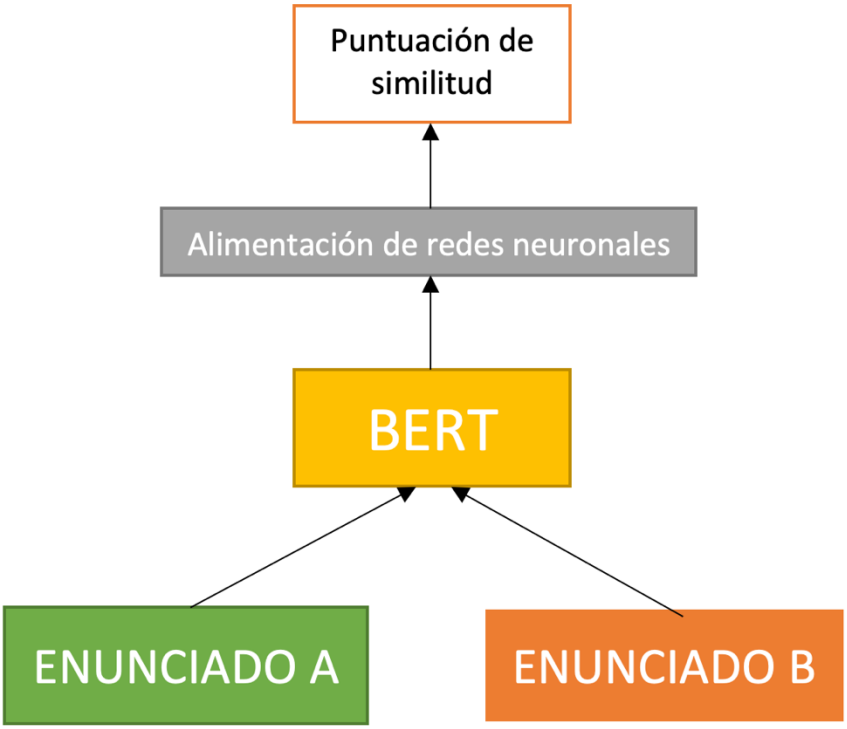

Figura 4. Funcionamiento de BERT

2.2.2.3 UNIVERSAL SENTENCE ENCODER

El modelo de Universal Sentence Encoder es otra técnica para procesar textos y obtener sus significados/representaciones, cuenta con diferentes versiones de acuerdo a los diferentes casos de uso que se tengan. El modelo permite la obtención de los significados/representaciones a nivel de oraciones con una gran calidad, usando recursos computacionales razonables y con muy pocos datos de entrenamiento (Cer *et al.*, 2018). Universal Sentence Encoder está desarrollado con dos modelos para el análisis de oraciones: Deep averaging network (DAN) y Transformers.

El modelo DAN, llamado también USE_DAN, produce significados/representaciones de oraciones de tal manera que primero promedia los significados/representaciones de palabras y bigardas y después aplica una red de retroalimentación neuronal sobre las representaciones promedio (Iyyer, Manjunatha, Boyd-Graber, & Daume III, 2015).

Por otro lado, Transformers, llamada comúnmente USE_T, es una versión más reciente y ofrece una exactitud mayor y de procesamiento más intenso que USE_DAN. Este modelo construye los significados/representaciones de las oraciones usando subgrafos de codificación de caracteres (Vaswani *et al.*, 2017).

La codificación de caracteres presta especial atención en procesar representaciones del conocimiento del contexto de las palabras en una oración tomando en cuenta el orden e identidad de otras palabras. Las representaciones del conocimiento del contexto de las palabras son promediadas en conjunto para obtener los significados/representaciones a nivel de una oración. USE_T puede procesar palabras, oraciones y documentos. La capacitad multitarea y de multi-lenjuage del paradigma de entrenamiento de USE_T hace más factible su uso en tareas como la clasificación/ranqueo semántica de oraciones. Por otra parte, la mayor ventaja de este modelo es su capacidad de encontrar textos similares sin la necesidad de procesos pares (Cer *et al.*, 2018).

2.2.2.4 TEXT RANK

Los algoritmos de ranqueo basados en grafos como el algoritmo HITS (Kleinberg, 1999) o el PageRank (Brin and Page, 1998) de Google han tenido gran éxito cuando son usados en el análisis de menciones, redes sociales y para el análisis de la estructura de vínculos en Internet. De manera sencilla, estos algoritmos deciden la importancia de un nodo/vértice dentro de un grafo tomando en cuenta información global del grafo procesada de manera recursiva, en lugar de solo analizar la información local del nodo específico. La idea básica, detrás de la implementación de estos algoritmos, es la votación o recomendación; cuando un nodo se conecta a otro lo que hace es buscar su recomendación, entre más recomendaciones/votos logre, mas importante llega a ser el nodo.

Text Rank es un modelo que aplica la misma lógica de los algoritmos de ranqueo para grafos semánticos obtenidos de documentos de lenguaje natural, el resultado es un modelo de ranqueo por grafos que puede ser aplicado a tareas del análisis del lenguaje natural. Entre las tareas que puede realizar está la extracción de palabras clave y la extracción de oraciones clave (Mihalcea and Tarau, 2004). La extracción de palabras clave identifica los términos que mejor describen el texto dentro de un documento. Por su parte, la extracción de sentencias consiste en la generación automática de un resumen del contenido más destacado de un texto dentro de un documento.

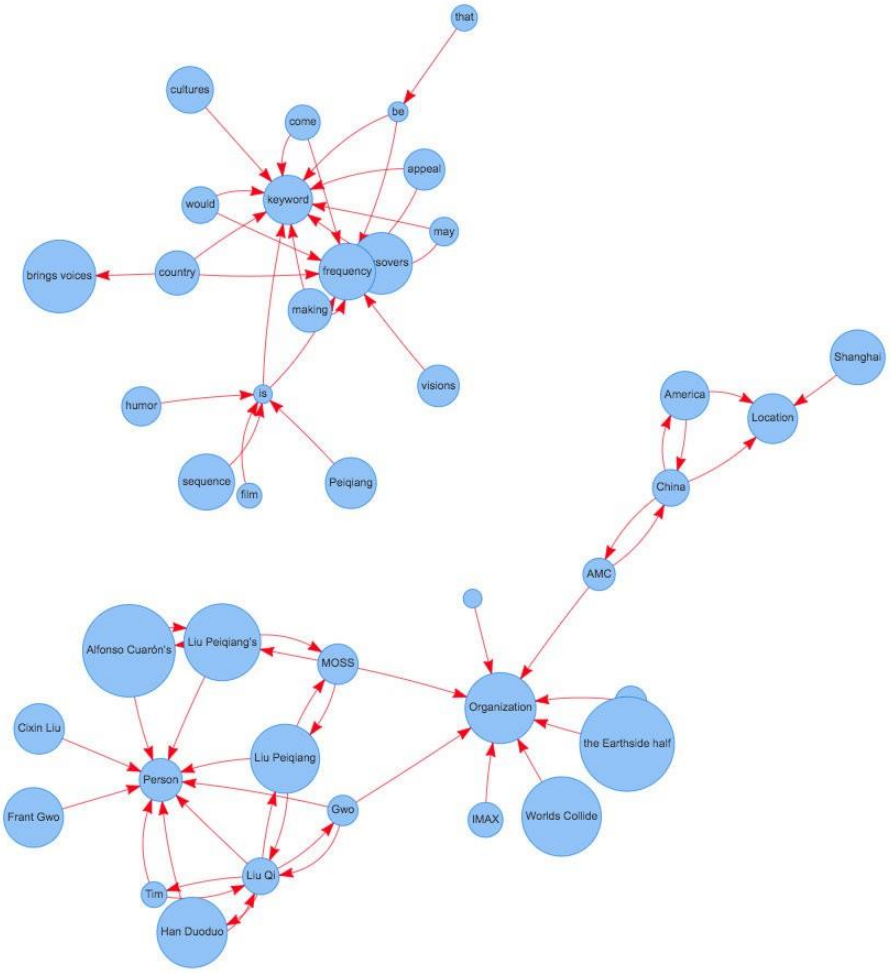

Figura 5. Funcionamiento de TEXT RANK

2.3 DICCIONARIO DE DATOS PARA EL ANÁLISIS DE SENTIMIENTOS

Como una opción más para el autoaprendizaje de analizar tweets y clasificar sentimientos se encuentran los diccionarios y algoritmos más simples que los descritos anteriormente. En este contexto, un diccionario es un conjunto de palabras que se relacionan entre sí y además se clasifican conforme un sentimiento o expresión de una emoción mismas que se evalúan conforme a dos condiciones de: polaridad positiva o negativa y valor de la fuerza de dicho sentimiento o emoción.

Un programa deberá evaluar la condición del sentimiento tomando en consideración el texto del tweet que se presente en los términos que incluye el diccionario. A diferencia de los casos anteriores, el programa no se enfoca en la detección de reglas sintácticas o de aprender a través de corpus categorizados. Su función consiste en encontrar coincidencias en el diccionario y relacionar el sentimiento basado en las coincidencias de polaridad y valor de la fuerza mencionados anteriormente.

Como podemos notar, el objetivo de esta técnica está basada en diccionarios con clasificación de sentimientos. Un factor muy importante es el idioma del texto

de un tweet. En ingles vamos a encontrar varios diccionarios disponibles muy completos con información detallada y de buen nivel que nos servirán para clasificar los sentimientos de las palabras enlistadas.

Es el caso de SentiWordNet (Baccianella, Esuli y Sebastiani, 2010), el conjunto de datos de expresiones subjetivas Multi-perspective Question Answering (MPQA) (Wilson, Wiebe y Hofmann, 2005) o el diccionario LIWC (Pennebaker, Mehl y Niederhofer, 2003). En cambio, los diccionarios en español, se enfocan en la polaridad de las palabras. Los diccionarios de Elhuyar (Saralegi y San Vicente, 2013) y el CRiSOL (Molina González, Martínez Cámara y Martín Valdivia, 2015) clasifican las palabras de acuerdo a su polaridad ya sea positiva o negativa.

Cabe mencionar que la evaluación del sentimiento de una oración no se encuentra solamente por el contenido de la semántica, si no que podemos encontrar negaciones que logran invertirnos su orientación en las palabras, es decir podemos encontrar frases con palabras positivas y negativas en el mismo texto.

De este modo, es básico identificar la sintaxis del mensaje, puesto que identificar las secuencias gramaticales ayuda a clasificar el sentimiento si existen negaciones, esto sin duda mejora el resultado.
Los diccionarios contienen ante todo adjetivos, que expresan información relevante al analizar los sentimientos; sin embargo, también contienen adverbios, verbos y sustantivos.

Se pueden encontrar en la web casi todos los diccionarios (la mayor parte en inglés, pero también en español, sin embargo, no tan completos y precisos). De igual modo nos sirven de base para clasificar si una frase es positiva o negativa, se verifica la oración conforme a los términos que se encuentren en el diccionario y el valor de su sentimiento. ~~A~~lgunos de los diccionarios son:

- Dictionary of Affect in Language (DAL)
- Linguistic Inquiry and Word Count (LIWC)
- SentiWordNet
- General Inquirer
- MPQA Subjectivity Lexicon
- Entre otros

DAL el nuevo Diccionario de afecto en el lenguaje (llamado DAL o Diccionario para abreviar) es un instrumento diseñado para medir el significado emocional de palabras y textos. Hace esto por comparar palabras individuales con una lista de 8742 palabras que han sido calificadas por personas para su activación y evaluación conforme a la polaridad y a la fuerza del mensaje.

LIWC es uno de los diccionarios más completos; su versión más completa está en inglés, aunque posee una versión beta en español. En este caso, las palabras son etiquetadas en una categoría determinada además de darle un peso.

SentiWordNet asigna un valor a cada palabra, asociado a puntuaciones pudiendo ser positiva, negativa o neutra, un valor entre 0 y 1.
General Inquirer tiene 1915 palabras en la categoría: “positivas” y 2291 palabras en la categoría “negativas”, con clasificaciones tales como activa o pasiva, fuerte o débil, placentera o dolorosa, etc. y es gratis para uso en investigación.

MPQA Subjectivity Lexicon incluye cerca de 8000 palabras, con su respectiva polaridad y categoría gramatical.

2.4 HERRAMIENTAS DE ANÁLISIS EN TWITTER

En años recientes, Twitter es una de las plataformas de redes sociales más importante del mundo. Su límite de 280 caracteres y la naturaleza de los millones de tweets que se generan al día hacen que los métodos estándares de recuperación de información y minería de datos sean inadecuados para Twitter, ya que la mayoría de los esfuerzos están basados en tareas específicas y hechas a mano, lo cual las hace ineficientes (Vosoughi, Vijayaraghavan, & Roy, 2016).

2.4.1 TWITTER ENCODER

El uso de modelos basados en técnicas de Aprendizaje Profundo combinados con algoritmos de transferencia de conocimiento y clustering representan el estado del arte para la detección de temas destacados en las conversaciones de Twitter. Esta técnica permite la extracción de los significados/representaciones de las oraciones en los Tweets, logrando obtener la información semántica de la oración; posteriormente se agrupan (clustering) las oraciones similares, basándose en sus significados/representaciones en grupos; estos grupos contienen diferentes temas semánticos que pueden ser resumidos a través del ranqueo de texto para identificar los temas destacados (Asgari-Chenaghlu, Nikzad-Khasmakhi and Minaee, 2020).

La variante del modelo de Universal Sentence Encoder basada en Transformers USE_T permite obtener los significados/representaciones de las oraciones en los tweets. USE_T, en su versión 5 está entrenado con un conjunto de datos monolingual dentro una dimensión de 512 significados/representaciones, esto permite la obtención de las similitudes semánticas entre tweets (Cer, y otros, 2018). La distancia entre el significado/representación de dos tweets representa el grado de disimilitud entre ellos.

Una vez que los significados/representaciones de los tweets son obtenidos se aplican algoritmos de agrupamiento(clustering) para agrupar los tweets de acuerdo a sus significados/representaciones. Pueden aplicarse diferentes algoritmos de agrupamiento como K-means, spectral clusterin, mean-shift o density based spatial clustering (DBSC) (Asgari-Chenaghlu, Nikzad-Khasmakhi, & Minaee, 2020).

La agrupación de tweets de acuerdo a sus significados/representaciones lleva a que cada grupo contenga tweets de temas similares; estos temas pueden ser mostrados para resumir la conversación en cada grupo.

2.4.2 TWITTER-BERT

Para mejorar la comprensión y análisis de los mensajes de Twitter es posible usar un modelo basado en BERT (Müller, Salathé and Kummervold, 2020).

Modelos como BERT, RoBERTa y ALBERT están todos basados en el mismo principio: entrenamiento bidireccional de modelos de transformers. Este proceso es realizado mediante el uso de métodos como el modelado mask language (MLM), next sentence prediction (NSP) y sentence order prediction (SOP). Todos estos métodos de entrenamiento son hechos sin supervisión y generan un modelo de lenguaje que posteriormente es usado para tareas de procesamiento de lenguaje como la clasificación de tweets (Devlin, Chang, Lee and Toutanova, 2018).

Una vez entrenado, Twitter-Bert puede convertir una secuencia de tweets de entrada en un conjunto de tokens basados en vocabulario obtenido en el entrenamiento. Entre más pasos de entrenamiento se realicen mayor será el vocabulario del lenguaje dentro del modelo y permitirá una mayor exactitud al momento de obtener los temas destacados de Twitter (Müller, Salathé and Kummervold, 2020).

2.4.3 TWEET2VEC

Tweet2Vec es un método que permite la representación de tweets en vectores de propósito general que pueden ser usados en cualquier tarea de clasificación. Este método es especialmente útil para el procesamiento de tareas de lenguaje natural en Twitter, como la clasificación de la conversación y la detección de reacciones (Vosoughi, Vijayaraghavan and Roy, 2016).

Tweet2Vec usa un codificador-decodifcador CNN-LSTM (Convolutional Neural Network - Long-Short Term Memory Network) (Zhang and LeCun, 2015) (Hochreiter and Schmidhuber, 1997) que opera a nivel de caracteres para aprender y generar vectores de representación de tweets. El codificador consiste en una red neuronal convolucional (CNN) que extrae las características de los caracteres y palabras de las oraciones de los tweets. La red de memoria de corto y largo plazo codifica la secuencia de características en un vector que contendrá todos los significados/representaciones de cada tweet.

El codificador-decodifcador CNN-LSTM debe ser entrenado con una selección aleatoria de tweets en un mismo idioma, usando técnicas de aumento de datos. El aumento de datos consiste en la sustitución de palabras por sus sinónimos en los tweets que se repliquen (Vosoughi, Vijayaraghavan and Roy, 2016). Tweet2Vec usa WordNet (Fellbaum, 1998), para obtener los sinónimos y decidir que palabras pueden ser o no sustituidas en las oraciones de los tweets.

El modelo Tweet2Vec puede ser usado para obtener la relación semántica entre tweets y para la clasificación del sentimiento de los tweets. La relación semántica

entre tweets es la obtención de la similitud del contenido del mensaje entre dos tweets a través de la comparación de sus vectores; obteniendo la separación de sus elementos, puede determinar si hablan de lo mismo o no (Xu, Callison-Burch and Dolan, 2015).

Por su parte, la clasificación del sentimiento de tweets clasifica las oraciones en positivas, negativas o neutras (Rosenthal *et al.*, 2015). Esta tarea la realiza evaluando de manera individual el vector del tweet basado en el entrenamiento previo del modelo.

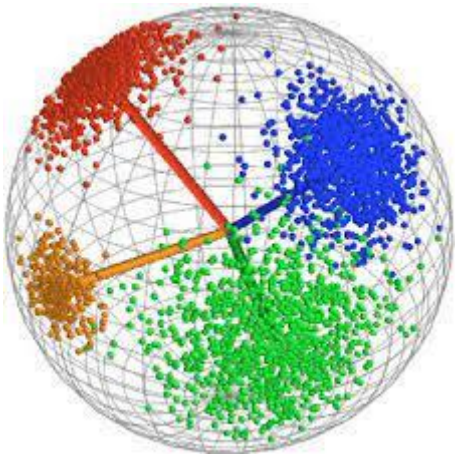

Figura 6. TWEET2VEC

2.5 PROYECTOS DE IDENTIFICACIÓN DE TEMAS DESTACADOS DEL COVID-19 EN TWITTER

Twitter encoder puede detectar temas destacados sobre COVID-19 usando la variante basada en Transformers USE_T que extrae los significados/representaciones semánticas de los Tweets relacionados del COVID-19 para posteriormente en combinación con algún algoritmo de clustering como K-means++ organizar las tendencias en los temas relacionados del COVID-19 (Asgari-Chenaghlu, Nikzad-Khasmakhi and Minaee, 2020).

Una variante de este modelo, llamada COVID-Twitter-BERT (CT-BERT) ha sido desarrollada para analizar y entender los mensajes publicados en Twitter a cerca del COVID-19. CT-BERT ha sido usado para evaluar el sentimiento de temas como COVID-19 y la vacunación (Müller, Salathé and Kummervold, 2020). CT-BERT requiere de un entrenamiento previo bastante extenso para ofrecer resultados confiables.

Utilizando el procesamiento del lenguaje natural utilizando el método LDA para identificar 11 temas y 8 subtemas en los datos de Twitter. El análisis temporal de los temas muestra la sensibilidad del discurso en línea a las noticias estatales significativas y las reacciones del gobierno local a la pandemia. (Shaghayegh Jabalameli, 2022).

Por otro lado, han aplicado el programa de análisis de texto LIWC a los tweets. LIWC es quizás el software líder para capturar información sobre conceptos psicológicos del texto. LIWC utiliza un enfoque de bolsa de palabras basado en la psicología para analizar el texto. Tausczik y Pennebaker brindan una historia de LIWC y el enfoque de la bolsa de palabras, que se deriva de Freud y otros y tiene una larga historia en psicología. En LIWC se representan diferentes conceptos, como "emoción positiva" y "emoción negativa", pero también conceptos relacionados como "ira" y "poder". Para cada concepto, se incluye un diccionario de palabras en LIWC. (Shaghayegh Jabalameli, 2022).

Autor	**Aproximación**	**Enfoque**	**Análisis de sentimientos**
Asgari-Chenaghlu et al., 2020	Transformers USE_T	Organiza las tendencias de los temas destacados	No lo hace
Müler et al., 2020	CT-BERT	Analiza y entiende los mensajes	Requiere entrenamiento previo para llevarlo a cabo
Veda C. Storey, Daniel E. O'Leary, 2022	LDA	Identifica 11 temas y 8 subtemas con los datos en Twitter	Lo hace de forma temporal y solo enfocado en las noticias estatales y reaccione de gobierno local
Shaghayegh Jabalameli, 2022	LIWC	Utiliza un enfoque de bolsa de palabras basado en la psicología para analizar el texto	Se representan diferentes conceptos, como "emoción positiva" y "emoción negativa", pero también conceptos relacionados como "ira" y "poder"
Rivas-Gonzalez	Universal Sentence Encoder	Identifica las principales tendencias, limpia, agrupa y analiza sentimientos	Determina los principales sentimientos encontrados, los clasifica en positivo y negativo y crea un diccionario de datos a partir de ello, dicho diccionario es de auto aprendizaje

CAPÍTULO 3. MODELO PARA MEJORAR EL ANÁLISIS DE DATOS Y SENTIMIENTOS DENTRO DE LAS CONVERSACIONES EN TWITTER MEDIANTE UN DICCIONARIO DE ANÁLSIS DE DATOS UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL.

3.1 METODOLOGÍA

RESUMEN
El objetivo de este trabajo es diseñar un modelo para el análisis de datos y análisis sentimientos dentro de las conversaciones en Twitter mediante algoritmos de Procesamiento de Lenguaje Natural y diccionario de datos. En este capítulo se presentará el modelo utilizado para el logro de dicho objetivo.

INTRODUCCIÓN
La metodología propuesta para el logro del objetivo general planteado se ilustra en la siguiente figura:

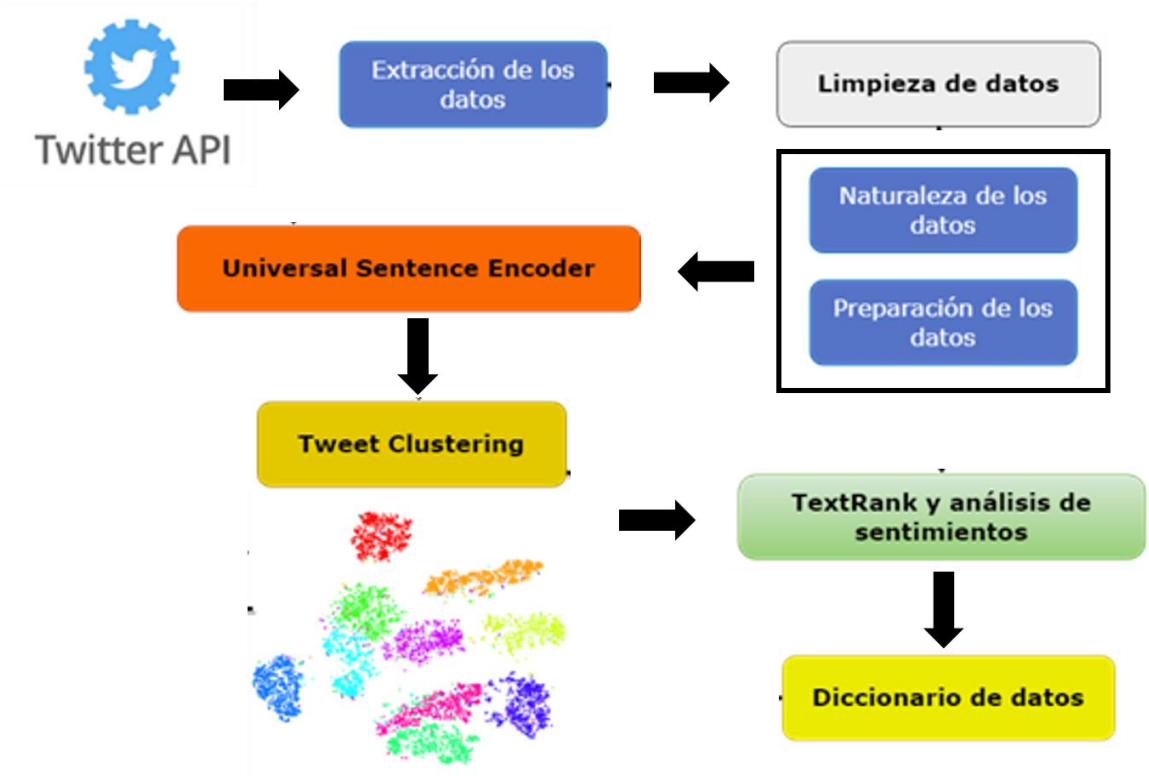

Figura 7: Esquema del modelo propuesto, Universal Sentence Encoder en combinación con un diccionario de datos.

A grandes rasgos, se extraen los datos de Twitter usando la API, se hace una limpieza de los mismos para que pueda ser procesado mediante el modelado de (Universal Sentence Encoder) y posterior a ello, se agrupan las relaciones de oraciones y palabras obtenidas, después se aplica la técnica de resumen de texto (TextRank) a cada grupo y se extraen las palabras más relevantes que son

las de mayor aportación, finalmente con estas palabras claves se les realiza de paso el análisis de sentimiento para identificar su polaridad ya sea negativa o positiva de cada palabra para identificar el sentimiento global del análisis. El diccionario de datos en cada paso de análisis va almacenando estos resultados.

3.1.1 Extracción de datos

En el modelo propuesto la primera capa es la que realiza la extracción de información.

Para el proceso de obtención de los datos utilizaremos la API de Twitter, nuestra principal fuente de información, esta API fue creada por el propio Twitter y se brinda la posibilidad de recuperar información publicada en dicha red social, entre otras funciones, necesitaremos obtener todos los datos acerca de las conversaciones públicas, como los usuarios, su mensaje, fecha de publicación, región e idioma, todos los atributos que permitan analizar, medir y filtrar los tweets que nos interesa, incluso para posteriormente poder validar que la información recopilada son de cuentas verificadas y que no sean automatizadas como las cuentas bots.

Para hacer funcionar la API, primeramente, se empieza por crear una cuenta de Twitter como desarrollador, esto es para que Twitter sepa que haremos uso de su API, aceptando sus políticas de uso y lograr tener los permisos a las funciones de la aplicación.

Logrado esto tendremos acceso a un portal de desarrolladores de Twitter en el cual se crea el proyecto en el que se empezara hacer uso de la API y se generan unas credenciales llamadas KEYS Y TOKENS, que se necesitarán para conectar con la API.

```php
<?php
//Access token & access token secret
define("TOKEN", '1309573073167880192-081hFzhhCAcYm0E5EuKULUhHDG7sTp'); //Access token
define("TOKEN_SECRET", 'Ue3jD9yH75Cxm9ZJOg0mjV3PEzM6CHTQY4iCAGGT5EHvu'); //Access token secret
//Consumer API keys
define("CONSUMER_KEY", 'i9rlKL0XmHUyF0Y6vp16IqQU1'); //API key
define("CONSUMER_SECRET", 'rxpjg0i3c8ZRLoAJj65JG3we29RIn7ZDx2QSY40a1fMZcPRTr0'); //API secret key
```

Posteriormente se desarrolla una aplicación para crear el entorno con toda la lógica del programa que permita conectarnos hacia la API, en nuestro caso se desarrolló un programa en lenguaje PHP para realizar las llamadas HTTP hacia la aplicación.

Una función atractiva es que podemos personalizar las búsquedas y filtrar los tweets por la palabra clave “COVID 19” para extraer los tweets que son de nuestro interés para este trabajo.

Todos los tweets encontrados se almacenan mediante nuestro programa, este es un proceso que configuramos para que trabaje de forma continua y así lograr obtener los tweets más recientes, aunque también, se pudiera personalizar las búsquedas hacia un periodo de tiempo atrás.

Finalmente tenemos como resultado la información de los tweets almacenados que nos servirán en las posteriores etapas.

3.1.2 Limpieza de datos

Una vez recopilada la información, pasará por un filtro de limpieza de texto, que es un proceso obligatorio y lo realizamos seleccionando oraciones más sensatas y utilizables de la pila de tweets. Las oraciones por su estructura y naturaleza de cómo escriben los usuarios contienen algunos elementos además del texto, mismos que se mencionan en la figura 8, los cuales para el modelo son innecesarios y dificultan el procesamiento del lenguaje natural.
Debido a lo anterior la limpieza de datos es uno de los pasos preliminares e importantes, ya que se eliminan los elementos que no aportan mucho sentido ni significado y además facilita el procesamiento al modelo dejando básicamente el texto.

Elementos	**Ejemplos**
Fecha	Sat Aug 06 04:35:43 +0000 2022
Hashtag	#Covid19 #SaludVacunate #AstraZeneca
Menciones	@AndresM @ChicoPerez
Emoticones	🤣😂 ☹
Símbolos	&#%$
URL de sitios web, enlaces de adjuntos como fotos o videos.	https://twitter.com/
Signos de puntuación	(,.¿?¡!"´´)

Figura 8. Lista de elementos a eliminar durante la limpieza de datos

Por mencionar un ejemplo, La API de Twitter nos proporciona el mensaje de un usuario con la siguiente estructura:

- &#%$Sat Aug 06 04:35:43 +0000 2022#%$&Yo sé que no me puede dar COVID por 5ta vez... 🤣 eso dije en la 4ta... y varios amigos me dicen, que como ya es la…

Mediante un proceso en nuestro programa realizamos la limpieza de datos, en el cual básicamente consiste comparar cada palabra del tweet original con una lista de elementos innecesarios que se ilustran en la figura 8 y serán removidos del mensaje, quedando de la siguiente manera:

- Yo sé que no me puede dar COVID por 5ta vez eso dije en la 4ta y varios amigos me dicen que como ya es la

Y así es como tenemos el mensaje limpio y preparado de todos los tweets recopilados para su posterior análisis, cabe mencionar que los datos removidos como fecha y hora, nos sirven para identificar cuando se publicó el mensaje y por su puesto para identificarlo entre periodos de tiempo, por lo tanto, se

almacena por separado al mensaje y este mismo proceso debe realizarse para todo el corpus recopilado.

Estructura del Tweet resultante

Información del tweet sin fecha, sin signos y sin emoticones

3.1.3 Universal Sentence Encoder

El lenguaje natural es un lenguaje utilizado por los humanos, es a través de él como nos podemos comunicar; tenemos un vocabulario de miles de palabras y mediante modelos de procesamiento de lenguaje natural, buscamos que la computadora procese la información tal como nosotros lo hacemos.

Nosotros como seres humanos podemos ver una lista de palabras y sencillamente conocer su significado, de diversas formas, por ejemplo:

decesos
muertos
guerra
casos
contagios
covid19
pandemia
muertes
defunciones
pacientes
síntomas
dosis
secuelas
méxico
semáforo
china
virus

Ahora si planteamos el caso a un texto más largo como “A mí me gusta el café” la situación cambia ya que estamos tratando con una oración y es ahí donde influyen reglas gramaticales, ambigüedades, entre otros matices que nosotros hemos ido aprendiendo para interpretar su significado más aún si lo que se quiso decir es bueno o malo.

Entre otras habilidades que tenemos es que también podríamos ordenar y clasificar las palabras conceptualmente, por ejemplo:

muertes
defunciones
decesos
muertos
casos
contagios
covid19
pandemia
virus
pacientes
síntomas
secuelas
méxico
china
guerra
semáforo

Lo que se busca con los modelos de procesamiento de lenguaje natural es justamente imitar este comportamiento, y analizar grandes corpus de texto, pero el problema es que la computadora únicamente es capaz de procesar números. En el caso de nuestra tesis, los Tweets son texto, por lo que para resolverlo utilizaremos algoritmos integrados al modelo para codificar las oraciones y palabras en forma numérica, la codificación seria una estructura de vectores y estos ya será más fácil para la computadora procesarlos y realizar operaciones.

Cabe mencionar que en los últimos años, se han realizado muchos esfuerzos para crear una representación semántica de la secuencia textual (como una oración, un párrafo o un documento) para lograr analizar grandes cantidades de información.

El algoritmo de Universal Sentence Encoder es un modelo pre entrenado que nos permite codificar las oraciones o palabras textuales en nuestro lenguaje natural a una representación vectorial en la que se representan las características de cada palabra en las oraciones del tweet. Con los vectores podríamos realizar operaciones para clasificar el texto, encontrar similitudes semánticas, realizar agrupaciones en clústeres, entre otros trabajos que son parte del campo del procesamiento del lenguaje natural.

Como se puede ver en la figura, una frase es recibida por el codificador de Universal Sentence Encoder y se vectoriza palabra por palabra para posteriormente crear un único vector de la oración en su conjunto.

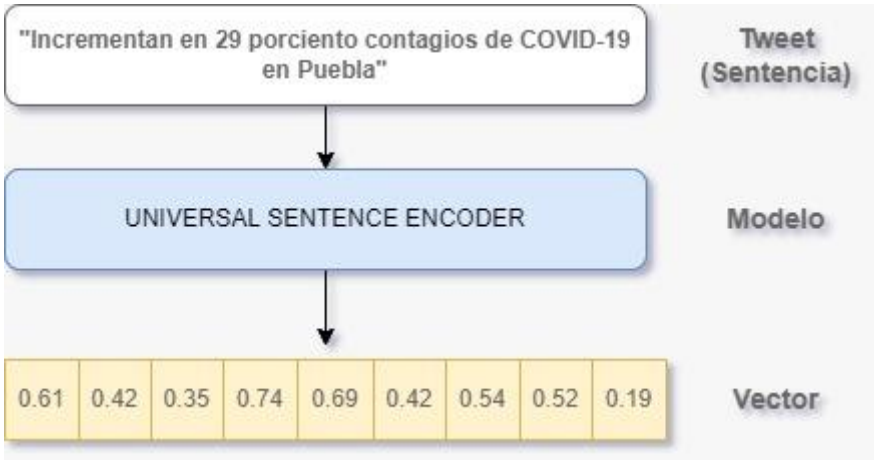

Figura 9: Proceso de vectorización utilizando Universal Sentence Encoder

El vector que se genera de la oración o Tweet, tiene valores llave de identificación que permitirán evaluar la compatibilidad o cercanía con otros vectores, es decir, con otras oraciones, todo esto observándose en un plano multidimensional, a continuación, se muestra el gráfico de similitud textual semántica de las palabras clave extraídas del modelo:

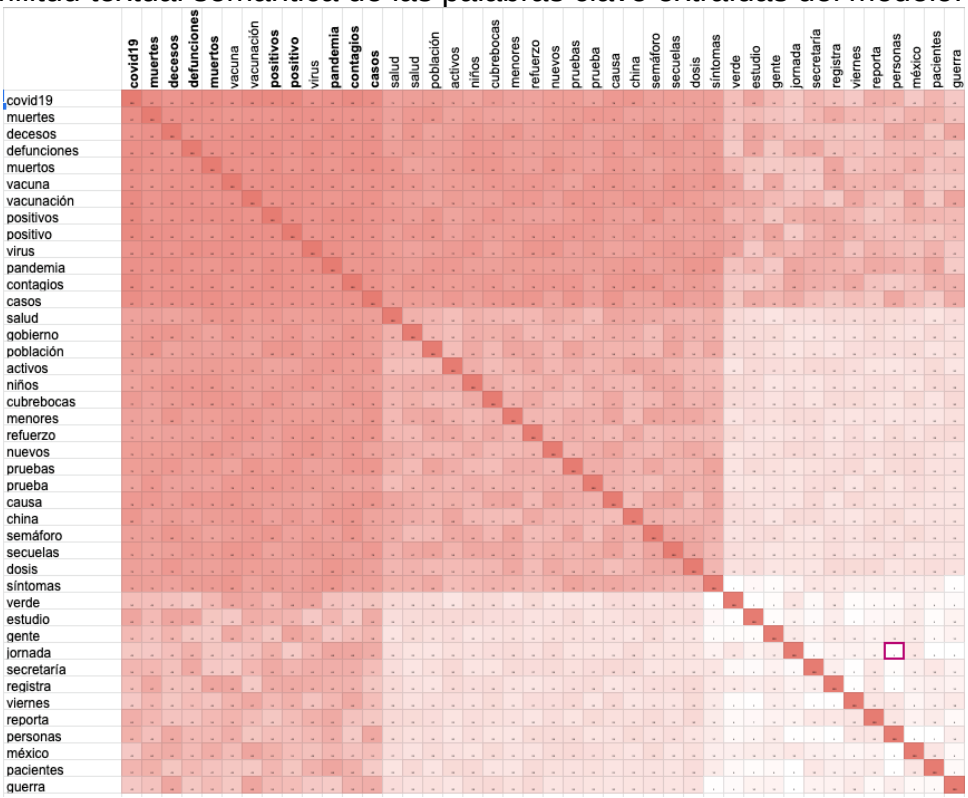

Figura 10: Gráfico de similitud textual semántica de palabras clave

El proceso se realiza para cada una de las oraciones o Tweets que forman parte de nuestro caso de estudio, de tal forma que evaluamos el corpus completo de cientos de Tweets vectorizados y podemos verlos en el espacio multidimensional de la siguiente manera:

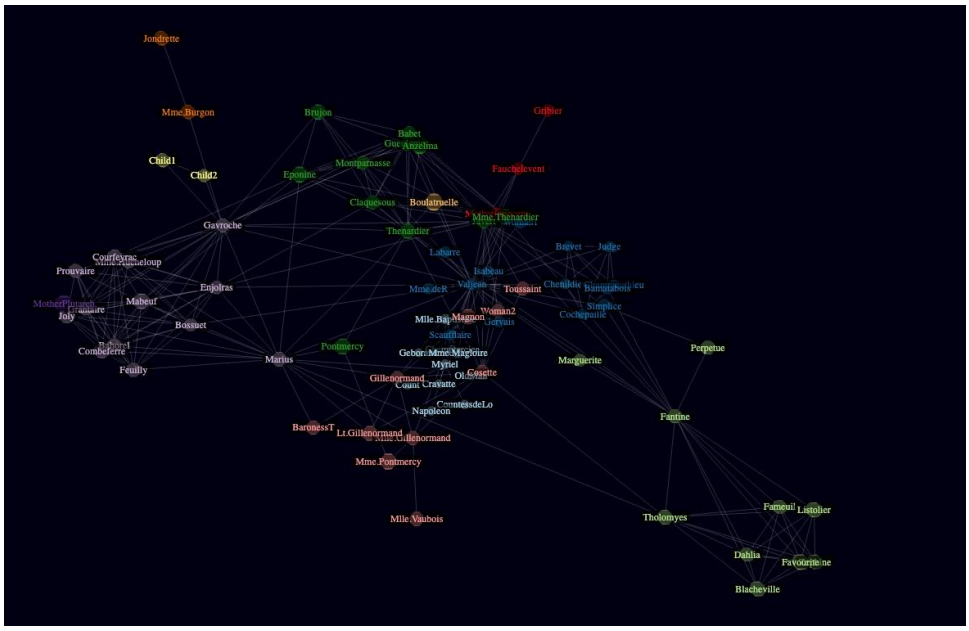

Figura 11: Visualización de vectores por cercanía y compatibilidad

Donde cada uno de los puntos es cada oración convertida en vector multidimensional, la salida es un vector del tamaño 512, es decir cuenta con 512 valores diferentes que aportan información sobre variables que el codificador identificó, estos valores permiten al computador realizar operaciones como el producto punto entre vectores para así crear una matriz de similitud, esta matriz representa una comparativa relacional del significado semántico. Como podemos ver en el gráfico cada punto ya tienen una posición y visualmente podemos ver la distancia o cercanía entre otros puntos y entre más cerca están los puntos de otros más similar son las oraciones en cuanto unos con otros o viceversa.

Internamente y a manera de ejemplo, el código pudiera verse así para hacer la vectorización del texto:

```
embed = hub.Module("../sentence_wise_email/module/module_useT")# Compute
a representation for each message, showing various lengths supported.
word = "Elephant"
sentence = "I am a sentence for which I would like to get its embedding."
paragraph = (
    "Universal Sentence Encoder embeddings also support short paragraphs.
"
    "There is no hard limit on how long the paragraph is. Roughly, the
longer "
    "the more 'diluted' the embedding will be.")
messages = [word, sentence, paragraph]# Reduce logging output.
tf.logging.set_verbosity(tf.logging.ERROR)with tf.Session() as session:
    session.run([tf.global_variables_initializer(),
tf.tables_initializer()])
    message_embeddings = session.run(embed(messages))for i,
message_embedding in enumerate(np.array(message_embeddings).tolist()):
        print("Message: {}".format(messages[i]))
        print("Embedding size: {}".format(len(message_embedding)))
        message_embedding_snippet = ", ".join((str(x) for x in
message_embedding[:3]))
        print("Embedding[{},...]\n".
                    format(message_embedding_snippet))
```

Figura 12: Código para vectorizar texto

El resultado de la vectorización del texto se podría ver de la siguiente manera:

```
Message: Elephant
Embedding size: 512
Embedding: [0.04498474299907684, -0.05743394419550896,
0.002211471786722541, ...]

Message: I am a sentence for which I would like to get its embedding.
Embedding size: 512
Embedding: [0.05568016692996025, -0.009607920423150063,
0.006246279925107956, ...]

Message: Universal Sentence Encoder embeddings also support short
paragraphs. There is no hard limit on how long the paragraph is. Roughly,
the longer the more 'diluted' the embedding will be.
Embedding size: 512
Embedding: [0.03874940797686577, 0.0765201598405838, -
0.0007945669931359589, ...]
```

Figura 13: Texto vectorizado

Como se menciona anteriormente el Universal Sentence Encoder es un componente clave para el análisis. Es capaz de codificar el texto en vectores de alta dimensión y estos vectores se utilizan para calcular la distancia entre dos puntos y con ello definir si son similares a partir de la distancia. Su capacidad de codificar las palabras asociadas a otras sobre su mismo campo semántico facilita encontrar la relación entre palabras u oraciones. Lo cual nos ayuda enormemente.

Para nuestro modelo se empleó la versión de Universal Sentence Encoder (USE), el cual nos permite manejar palabras, oraciones y hasta documentos como entrada. El paradigma de entrenamiento multitarea y multilingüe del USE lo hace más adecuado para tareas como la recuperación de pares de oraciones semánticas. Por otro lado, la implementación de USE en nuestro modelado es más crucial con respecto a la relación de tweets semánticamente cerrados. El lado más poderoso de esta arquitectura es su capacidad para encontrar textos semánticamente similares sin necesidad de muchos cálculos. Proporciona una representación vectorial para cada palabra o unidad de texto, y estos vectores se usan para calcular la distancia o similitud entre diferentes tweets u oraciones.

3.1.4 Tweet clustering / Text Rank

Usamos algoritmos de agrupamiento para agrupar oraciones o términos similares (según sus características) en los mismos grupos. Idealmente, diferentes grupos contienen diferentes campos semánticos. Se aplica la técnica de resumen de TextRank en las oraciones para generar un resumen de cada agrupación, que contiene el tema más representativo.

Se pueden utilizar diferentes algoritmos de agrupamiento para este propósito, como K-means, agrupamiento espectral, desplazamiento medio, agrupamiento espacial basado en densidad (DBSC). Aquí utilizamos el algoritmo de agrupación en clústeres K-means, por su simplicidad, velocidad y la capacidad de predefinir el número de clústeres.

El paso de agrupamiento proporciona varios grupos de oraciones semánticamente similares. En un nivel alto, los tweets en los mismos grupos deben contener temas más similares que los de diferentes grupos. Aunque el

centroide de cada grupo debe contener la inserción promedio (por lo tanto, el tema / concepto promedio de ese grupo), no necesariamente capturará todos los temas de ese grupo. Pero podría servir como una simple línea de base. Una mejor solución para encontrar el tema de cada grupo es utilizar una técnica de resumen de texto para proporcionar un resumen significativo y sensible de ese grupo, capturando los temas clave. Aquí usamos el algoritmo de TextRank. (Federico Barrios, 2016).

3.1.5 Diccionario de datos

Se incorpora el uso de un diccionario de datos de base para realizar el análisis de sentimiento, este se conforma de términos positivos, negativos, verbos, adjetivos y sustantivos. Inicialmente se incluyen los términos de una polaridad fija, ignorando los vocablos que sean de carácter ambiguo o neutro. El diccionario de datos está conformado con información de diversas fuentes:

- En primer lugar, se incluye el diccionario léxico de SentiWordNet 3.0 al idioma inglés. El de mayor cantidad de palabras clasificadas con un polaridades positivo, negativo o neutral, así como categorizada por clase cada palabra. Está conformado por un total de 8,427 palabras es muy común y utilizado idealmente para el procesamiento del lenguaje natural. https://raw.githubusercontent.com/rmaestre/Sentiwordnet-BC/master/data/sentiwordnet.tsv

- En segundo lugar, se usa un diccionario por Pérez Rosas el que abarca 1000 palabras clasificadas igualmente en los términos de positivo y negativo. (Veronica Perez Rosas, 2012.)

- En tercer lugar, se escogió el top de las 100 palabras claves más frecuentes que nos resultan del análisis de los tweets, esto para asegurarnos de abarcar en lo que más les interesa a los usuarios en la red social en relación al tema de investigación, las palabras previamente encontradas se evaluaran con el diccionario, y únicamente las palabras no encontradas, serán las palabras nuevas, el cual se tendrá que hacer la clasificación de su polaridad y se agregaran al diccionario. (Fuente propia).

- En último lugar, se tomaron las palabras del recurso Spanish Emotion Lexicon (SEL) conformado por 2036 palabras en español clasificadas por categorías, las cuales son: miedo, enojo, alegría, tristeza, sorpresa y repulsión, el cual indica el porcentaje de probabilidad de uso afectiva, determinado a un nivel de asociación con valores de (alta, media, baja o nula) relación hacia la categoría evaluada, de forma que para incluir al diccionario se clasifico “alegría” como palabras positivas y las demás como “negativas”, exceptuando la categoría de “sorpresa” que

invariablemente puede ser de ambas polaridades. (Grigori Sidorov, 2012) (Ismael Diaz Rangel, 2014)

En total integrando a todas las fuentes utilizadas se juntaron 11,563 palabras, cabe mencionar que algunos recursos léxicos funcionan para el idioma inglés sin embargo se realizó la traducción al idioma español con la herramienta de Google Traductor para poder darle uso en nuestro idioma, el diccionario se desarrolla a medida que surjan palabras nuevas, o el campo de investigación sea distinto, sin embargo, el diccionario a partir de aquí cuenta con una base de palabras muy amplia para uso general.
Con lo anterior buscamos:

- Contar con una solución genérica con un amplio rango de aplicación para el análisis de tendencias en redes sociales, específicamente en Twitter.
- Realizar entrenamientos confiables de los modelos de análisis de lenguaje natural para ambientes en redes sociales, específicamente en Twitter.

A continuación, se presenta como es el proceso del análisis de sentimientos y de forma más detallada como se estructura el diccionario de datos.

3.2 ANÁLISIS DE SENTIMIENTOS

En este punto, se presenta el proceso y sus tareas subsecuentes para resolver los problemas y desafíos para poder clasificar la polaridad de los sentimientos en un tweet o por una frase en particular. Luego describimos el proceso de creación de diseño de nuestro diccionario de datos y anotación de mensajes de texto cortos de redes sociales.

Para llevar a cabo este proceso primero se hace la recopilación de datos y anotación para el conjunto de datos de mensajes de texto o frases, para ello usamos la API de Twitter de transmisión pública para descargar los tweets. Los tweets recopilados deben ser del mismo periodo de tiempo y centrarse en un tema de interés o relevancia.

A continuación, describimos las cinco subtareas sobre análisis de sentimientos en Twitter

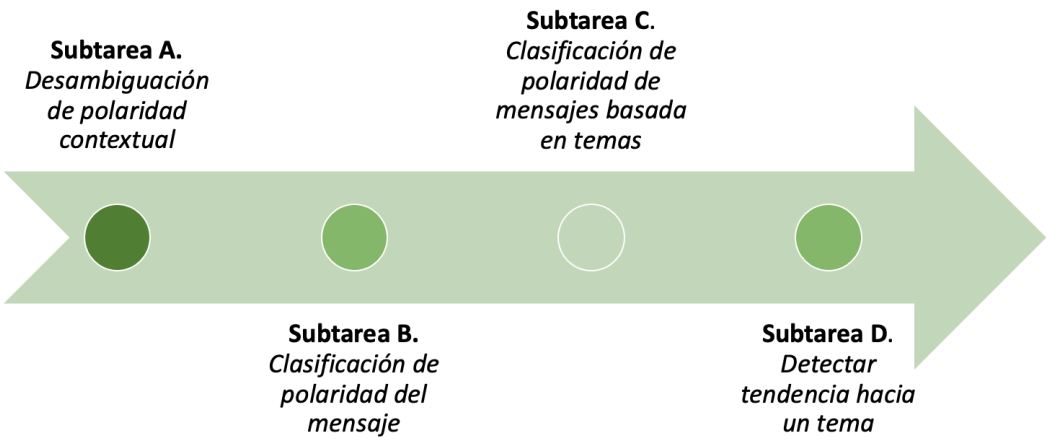

Figura 14: Subtareas para el análisis de sentimientos

- **Subtarea A. Desambiguación de polaridad contextual:** Dada una instancia de una palabra/frase en el contexto de un mensaje, se determina si expresa un sentimiento positivo, negativo o neutral en ese contexto.

  En este paso dada la frase/palabra se tiene que identificar si el mensaje presenta varios sentidos de polaridad y determinarlos para el contexto en el que se dio.

  Por ejemplo:

  Los decesos de personas enfermas se redujeron, los científicos realizan muchos esfuerzos por la salud en el desarrollo de la vacuna.

  Existe más de una polaridad por un lado las palabras subjetivas “decesos” y “enfermas” nos dan un sentimiento negativo, pero otro “esfuerzos” y “salud” nos arrojan el positivo y “desarrollo” o “científicos” tienen un sentido más neutral, es así que de esta manera se identifica dado el mensaje que palabras y fragmentos de oración tienen una polaridad en particular.

  Esto no siempre tiene que ser así, pero es importante como paso principal considerar las ambigüedades que se puede presentar y hacer las acciones para que esta se pierda y sea más fácil de comprender.

- **Subtarea B. Clasificación de polaridad del mensaje:** Dado un mensaje, se determina si expresa un sentimiento positivo, negativo o neutral. Si se expresan sentimientos tanto positivos como negativos, se debe elegir el más fuerte.

Retomando el mismo ejemplo anterior, con las tres polaridades negativo, neutral y positivo se determina de forma general el mensaje completo la polaridad que tenga mayor intensidad.

Se enlista cada una de las oraciones que representan una polaridad del mensaje.

Los decesos de personas enfermas se redujeron. Negativo

Los científicos realizan muchos esfuerzos por la salud. Positivo

El desarrollo de la vacuna. Positivo

Como se puede ver es el sesgo se inclina más al lado positivo por lo que este mensaje se puede clasificar como positivo.

- **Subtarea C. Clasificación de polaridad de mensajes basada en temas:** Dado un mensaje y un tema, se decide si el mensaje expresa un sentimiento positivo, negativo o neutral hacia el tema. Si se expresan sentimientos tanto positivos como negativos, se debe elegir el más fuerte.

  Aquí esta etapa se agrega un factor que puede influir en la clasificación y es el del tema de estudio para nuestro caso es el COVID 19.

  El mensaje ya está anteriormente clasificado como positivo de forma general y en el contexto original del mensaje, en caso de que el mensaje no esté relacionado al tema de estudio, en esta tarea implica evaluar nuevamente la polaridad, pero en el contexto del tema de investigación.

  El ejemplo claramente está en el contexto del COVID-19 ya que trata sobre un tema sobre la salud de las personas, por lo que la polaridad permanece.

  En la recopilación de datos, los tweets ya vienen filtrados sobre el tema de estudio por lo que todos deberían en el contexto del tema.

- **Subtarea D. Detectar tendencia hacia un tema:** Dado un conjunto de mensajes sobre un tema en el mismo período de tiempo, se clasifica el sentimiento general hacia el tema en estos mensajes como (a) fuertemente positivo, (b) débilmente positivo, (c) neutral, (d) débilmente negativo, o (e) fuertemente negativo.

  En esta tarea está estrechamente relacionada con la anterior para resolverla de manera obvia hay que obtener el resultado de C y con ello hacer los cálculos. Donde aquí se debe cuantificar el total del conjunto de mensajes por tipo de polaridad, determinar el sentido hacia donde se inclina más el sentimiento. Como resultado se debe de obtener el sentimiento general del conjunto de mensajes de un periodo de tiempo y sobre el tema de estudio.

Como se describió anteriormente, el paso de cada una de las tareas debe ir acompañada de anotaciones que permitan adjuntar la información sentimiento de cada tweet metadatos, que entablan con el siguiente diseño de diccionario de datos.

Tabla 3. Estructura del diccionario de datos para el análisis de sentimiento

NOMBRE DEL CAMPO	DESCRIPCIÓN	TIPO DE VALOR	VALORES
PALABRA INGLES	Palabra al idioma ingles	Texto alfanumérico	"Words"
PALABRA AL ESPAÑOL	Palabra al idioma español	Texto alfanumérico	"palabras"
POLARIDAD	Clasificación de polaridad de la palabra	Texto alfanumérico	"1,0,-1"
POSITIVO	porcentaje de asociación para la polaridad positiva	Numérico [0.0,1.0]	0.5
NEUTRO	porcentaje de asociación para la polaridad positiva	Numérico [0.0,1.0]	0.3
NEGATIVO	porcentaje en relación para la polaridad positiva	Numérico [0.0,1.0]	0.2
CLASE	Clasificación de la clase de palabra	Texto alfanumérico	a = “adjetivo”
			v= "verbo"
			n= "sustantivo"
			r= "adverbio"
POLARIDAD SUBJETIVA	Clasificación de polaridad en termino subjetivo.	Texto alfanumérico	"negativo"
			"positivo"
			"neutro"
			"alegría”
			“repulsión”
			“enojo"
			"Miedo"
			"Sorpresa"
			"Tristeza"

Tabla 4. Estructura donde se almacenan los tweets recopilados

NOMBRE DEL CAMPO	DESCRIPCIÓN	TIPO DE VALORES
ID	La identificación del tweet	Numero secuencial. (1-9999999)
FECHA PUBLICACIÓN	Fecha de publicación	DD/MM/AAAA HH:MM: SS
USUARIO	Cuenta de usuario que twitteo	Texto alfanumérico
NOMBRE_USUARIO	Nombre del usuario	Texto alfanumérico
FUENTE_ORIGEN	Dispositivo o medio por el cual se publicó el mensaje	Texto alfanumérico
IDIOMA	Idioma del mensaje	Texto alfanumérico
MENSAJE	El texto del mensaje del tweet	Texto alfanumérico
TENDENCIA_TEMA	Relación en tendencia a un tema de investigación. “COVID-19”	A = fuertemente relacionado B = débilmente relacionado C = neutral D = débilmente relacionado E = no relacionado.

CAPÍTULO 4. RESULTADOS

En este capítulo se presentan los descubrimientos logrados por el uso del modelo propuesto. Respondiendo a los objetivos, primero se identifican las palabras más frecuentes, temas más relevantes, los hashtags más mencionados e incluso la referencia de usuarios populares más activos en la red social; en segundo término, se determinará la polaridad de los mensajes a través de un diccionario de datos para evaluar el sentimiento, tanto de forma específica como de manera general.

4.1 MODELO IMPLEMENTADO

En este estudio se propuso un modelo para el análisis de datos y análisis sentimientos dentro de las conversaciones en Twitter mediante algoritmos de Procesamiento de Lenguaje Natural y diccionario de datos. De manera que se puedan analizar grandes cantidades de texto de forma inteligente aplicado a los tweets relacionados con el tema del COVID 19. Este modelo puede proveernos de temas resultantes en forma de oraciones concisas que son más fáciles de leer y comprender para el ser humano.

Para la evaluación del modelo desarrollado, se realizó el caso práctico de estudio sobre el término clave del “COVID 19”, logrando almacenar un conjunto de datos con información de un tamaño alrededor de los 6.1 millones de tweets en el idioma español, que corresponden al periodo comprendido entre el 1 de enero de 2021 al 31 de diciembre de 2021 (etapa intermedia de la pandemia).

Se toma como referencia la figura 7 mencionada en el capítulo anterior para describir los resultados de cada proceso.

4.1.1 Extracción de datos

Se hace una búsqueda para extraer los datos utilizando la API de Twitter, misma que nos arroja los resultados de la Figura 15.

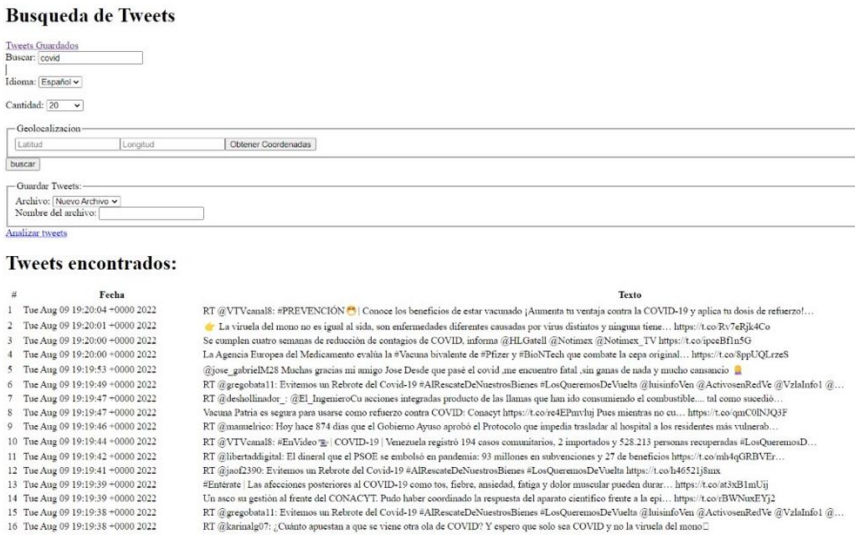

Figura 15. Ejemplo de búsqueda de Tweets

4.1.2 Limpieza datos

Se limpian para dejarlo sin caracteres especiales y otros. La Figura 16 muestra un ejemplo del tweet original y del tweet posterior a la limpieza.

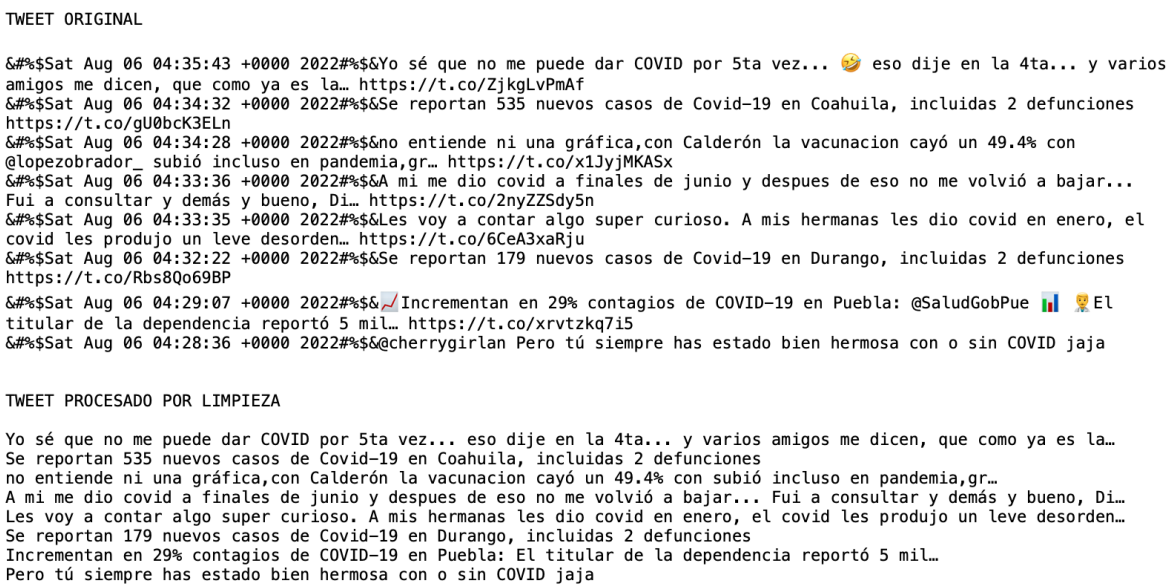

Figura 16. Ejemplo de limpieza de datos

4.1.3 Tweet clustering / Universal Sentence Encoder

Utilizando los algoritmos para el agrupamiento y utilizando el Universal Sentence Encoder, podemos ya agrupar los tweets relacionados; en la Figura 17 se muestra un agrupamiento del tema COVID. A partir de la figura 19, todas las gráficas se realizaron con base en la información obtenida de la extracción, limpieza y clustering, utilizando Excel, con la finalidad de que sea puedan visualizar de forma gráfica los resultados.

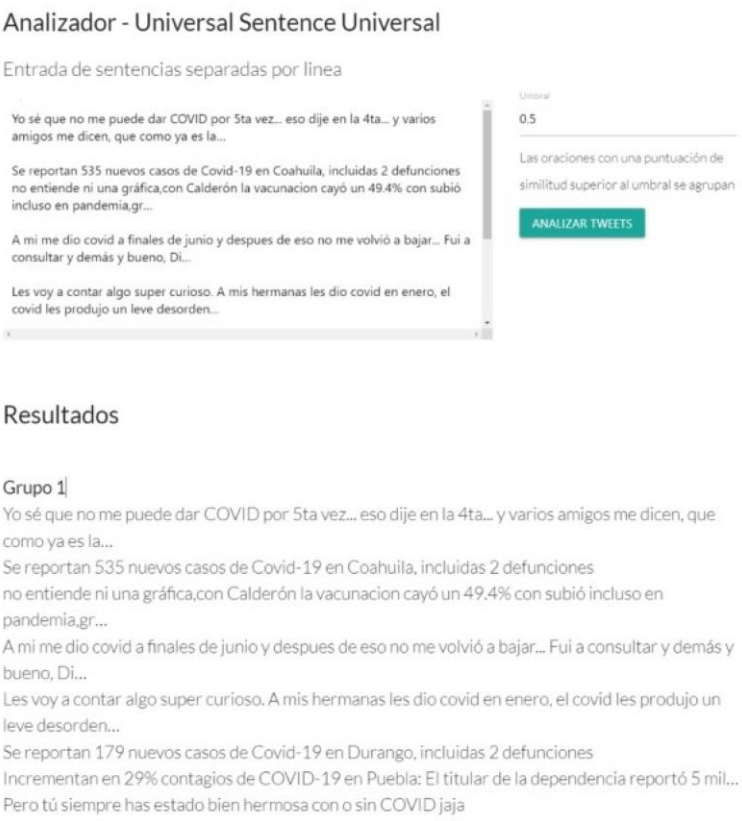

Figura 17. Ejemplo de uso de Universal Sentence Encoder

La Figura 18 muestra, a manera de ejemplo, el agrupamiento de temas genéricos que nos permita ver varios grupos.

Figura 18. Ejemplo de Tweet Clustering

4.2 ANÁLISIS DE DATOS CON MODELOS DE PROCESAMIENTO DE LENGUAJE NATURAL

Con la creciente escala de COVID-19 (durante el periodo del año 2021), ha habido un cambio en las distribuciones de tweets publicados en Twitter, lo que refleja el hecho de que ha aumentado la incidencia de los casos por el mundo, pero sobre todo los de nuestro país, el surgimiento y desarrollo de diferentes vacunas a fin de prevenir la enfermedad del COVID-19, así como los sucesos de muertes a causa de la enfermedad, fueron en una de las principales preocupaciones de las personas en todo el mundo.

Para ilustrar esto, en la Figura 19 y 20, mostramos las palabras más frecuentes utilizadas en Twitter durante todo el año 2021. Como se puede ver, palabras como ***"México", "casos", "vacunas", "dosis", "muertes***" se encuentran entre las palabras populares.

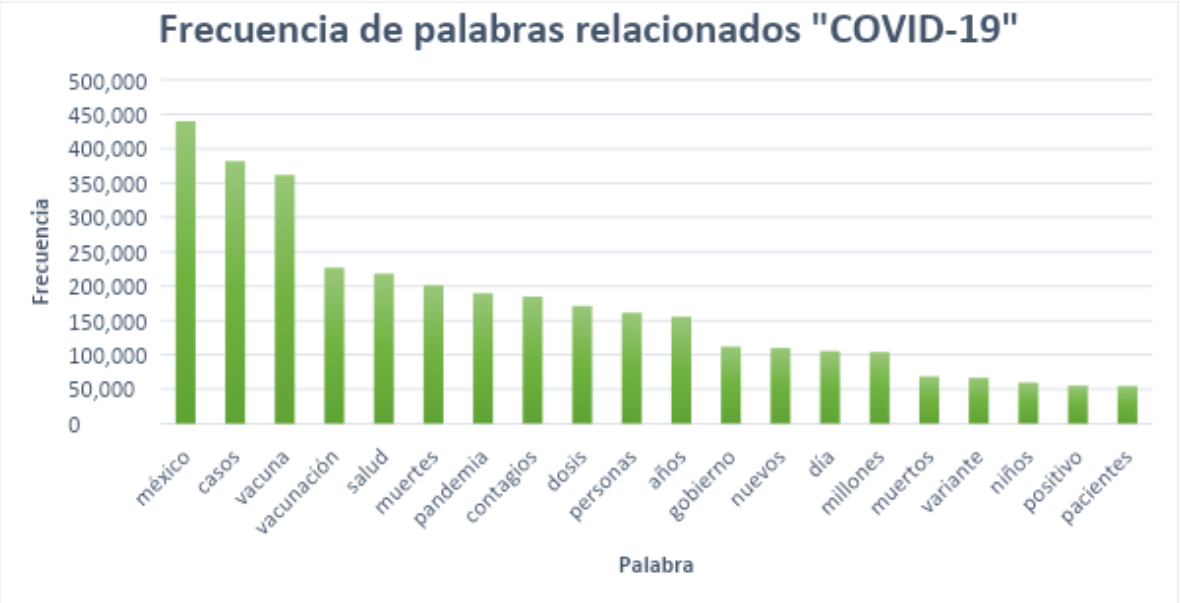

Figura 19. Gráfica con la frecuencia de palabras durante el periodo del 2021

Figura 20. La nube de palabras populares en Twitter durante los meses del año 2021 de la pandemia.

Analizar la opinión y las preocupaciones de las personas en las redes sociales puede ayudarnos a comprender mejor sus intereses y expectativas; permitiendo al gobierno y a funcionarios de salud tener una mejor planificación para el manejo

de la situación. Es ahí la importancia del trabajo de investigación y análisis relacionados con COVID-19.

En un segundo momento, se eliminaron las palabras vacías, dejando solo las más frecuentes. Un interesante hecho de esta trama es que, algunas palabras siempre están presentes entre las populares durante un tiempo. Para mejor visualización eliminamos todas las palabras repetidas de los meses anteriores en la Figura 21.

Figura 21. Nube de palabras del conjunto de datos sobre el COVID 19 en Twitter por mes; se muestran los primeros 3 meses del año.

Como podemos ver en la figura 22, las líneas interpretan la variabilidad de la frecuencia de las 8 palabras más populares en el mes de febrero, donde las palabras "**México", "vacuna"** y **"salud"** son de las que alcanzaron máximos más altos por lo tanto se afirma que el tema de la vacunación era un tema latente y con la mayor atención; sin embargo, se puede observar también que la tendencia temporal de las palabras está en constante cambio, es decir la frecuencia de palabras no siempre será la misma.

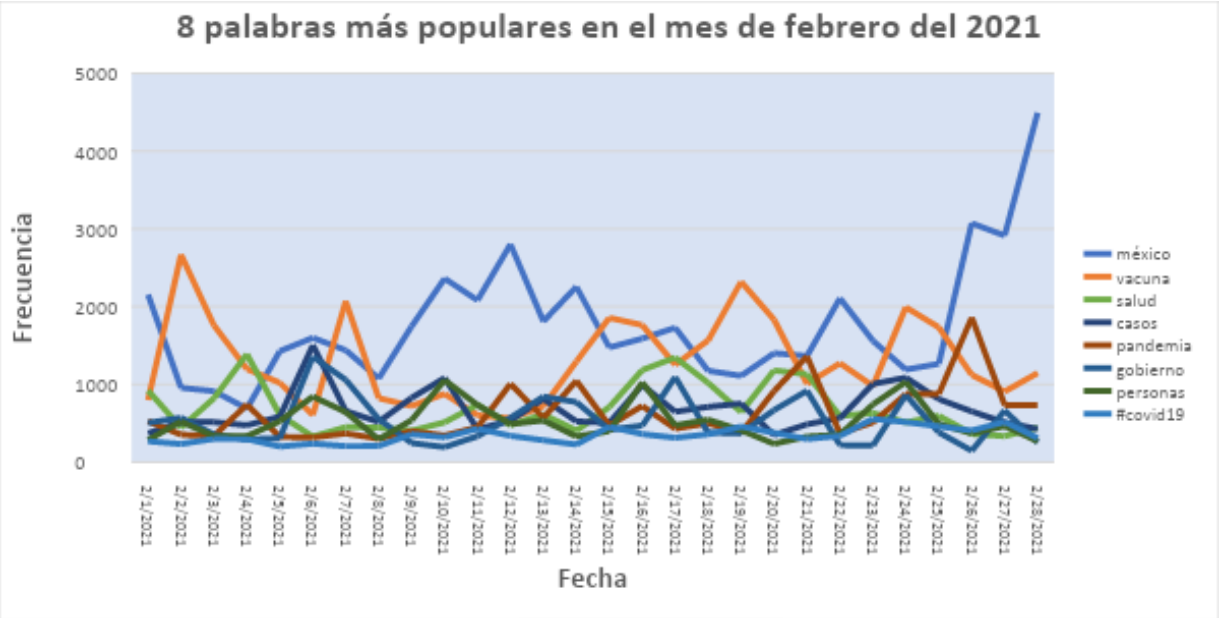

Figura 22. Las 8 palabras más populares en el mes de febrero del 2021.

4.3 TRENDING TOPICS DEL COVID-19 Y SU EVOLUCIÓN

Los Trending Topics de la muestra arrojan una perspectiva de emergencia y miedo a partir de la preponderancia que se le da a las palabras negativas de casos, muertes y pandemia. Esto aumenta considerando palabras con significados similares**,** como contagios, crisis y fallecidos**.** En caso inverso, otras palabras se centran en acciones y medidas de salud, como vacunas, dosis, medidas, y pacientes, son de baja tendencia. Todo lo anterior descrito se muestra en las figuras 23, 24, 25 y 26.

Un punto importante se matiza un agrupamiento que describe el COVID-19 con las palabras de carácter de salud, como vacunación, pruebas, población, medidas y mundo, que describe un paso importante en el control de la pandemia, mejorando en medida la salud y recuperando la confianza en la población y a la gente.

4.4 ANÁLISIS DE SENTIMIENTOS

El análisis de sentimientos es la parte del análisis en el que se busca un mayor contraste de cuál es el estado anímico de las personas en la manera que expresan sus mensajes, qué sentido tiene su orientación emocional hacia nuestro tema de estudio del COVID 19.

A través de un diccionario de datos de palabras etiquetadas y clasificadas con su valor emocional en tres categorías polaridad negativa, neutral o positiva, se establece para cada palabra una intensidad de rango que va del -1 siendo el negativo, 0 para el neutral y +1 el positivo.

El diccionario cuenta con una base de palabras conformadas de diversas fuentes, pero a medida que se va analizando más a fondo el tema de estudio se han ido encontrando nuevas palabras claves que agregar al diccionario, se etiqueta su polaridad y ponderación de su valor, para así mismo ir alimentando y ampliando el diccionario, con esto lograremos un diccionario que se sujeta al tema de estudio, pero que además pueda ir ampliándose a otros campos de estudio.

Ahora evaluando a las palabras claves encontradas en la primera parte del análisis, anteriormente mencionadas que ya por sí solas nos dan una perspectiva del mayor enfoque e interés del momento de estudio hacia el tema, ahora complementaremos a estos resultados, con las polaridades para cada palabra clave encontrada y realizaremos un cálculo de puntajes, para lograr medir el sentimiento global.

En las figuras 23 y 24 se muestran nube de palabras identificadas que favorecen el aumento de los cálculos tanto positivos, como negativos del conjunto de datos analizado. En las gráficas 25 y 26 se muestran las palabras con mayor frecuencia y con la cantidad de veces mencionadas.

Figura 23. Nube de palabras de términos negativos.

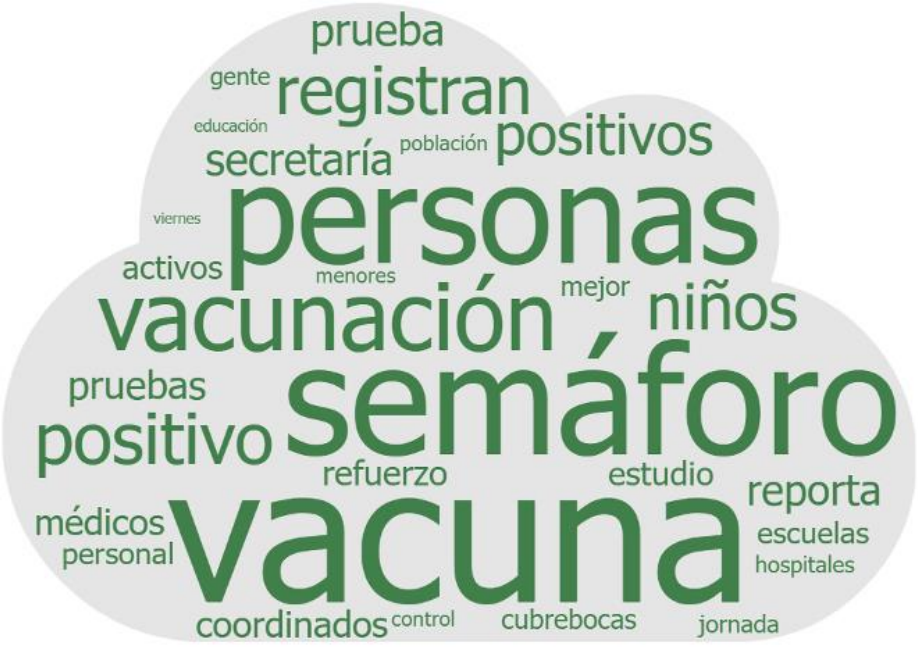

Figura 24. Nube de palabras de términos positivos.

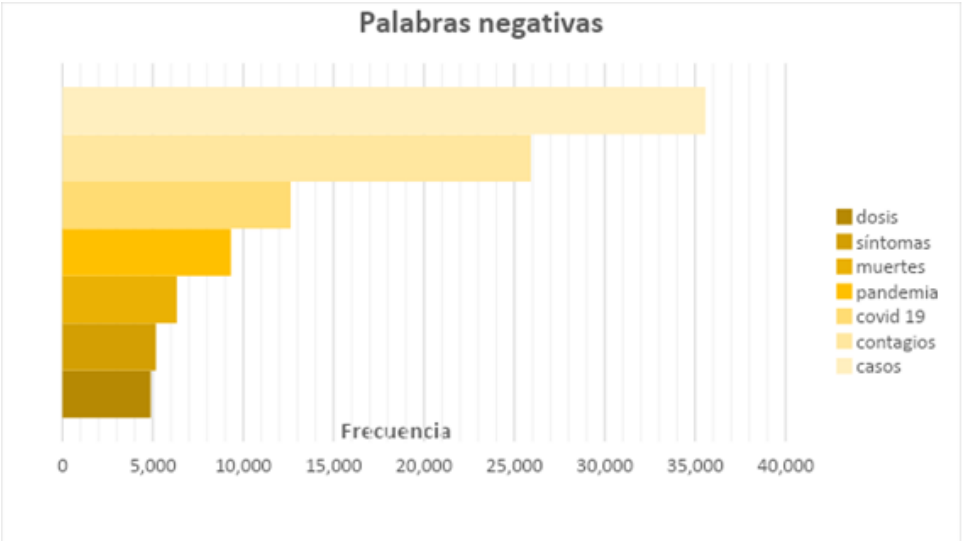

Figura 25. La frecuencia de palabras negativas.

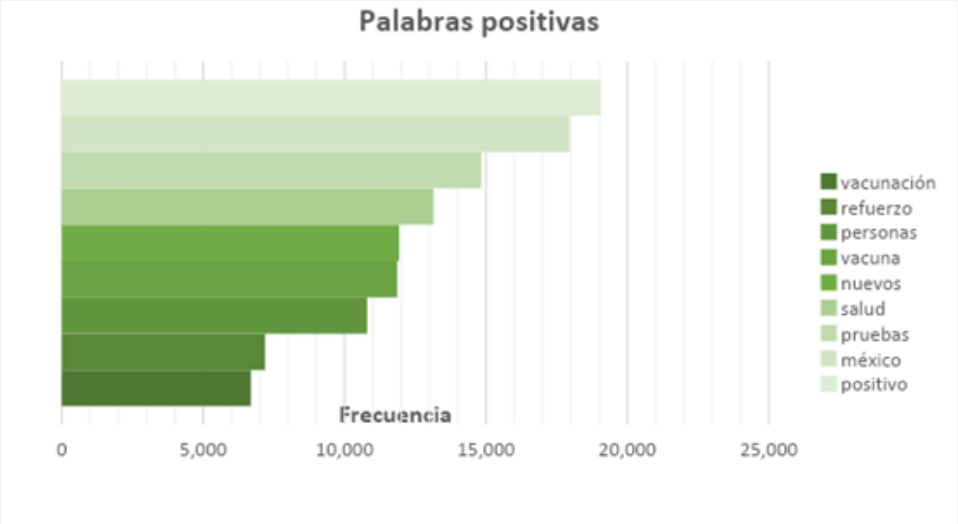

Figura 26. La frecuencia de palabras positivas.

De esta manera, se logra ver una perspectiva más semántica. Según la clasificación, las palabras que más negativamente suman son muerte, guerra y COVID 19, mientras que los positivos, con mucha menos repercusión total, son medidas, personas y México.
Una vez identidades las polaridades y los valores de las palabras claves encontradas y seleccionadas, realizamos las operaciones de suma de puntajes negativos y los positivos, omitiendo los valores neutros que no afectan al cálculo.

$$\sum(Vp\ x\ Fp)$$

Donde la formula representa la sumatoria de “Vp” representa el valor de su carga negativa de la palabra multiplicada por ‘Fp’ que es la cantidad de frecuencia de la palabra.

De la misma manera se utiliza la formula, para el cálculo de puntajes positivos, pero en el sentido opuesto solo las palabras positivas.

Los resultados de los puntajes identificamos se inclinó más al lado negativo: La suma de todas las palabras negativas, alcanzó un valor de – 270,915 puntos negativos, frente a un valor de +241,262 puntos con la suma de todas las palabras positivas.

Finalmente, si entre los puntos positivos le contrarrestamos los puntos negativos obtenemos un valor negativo de - 29,653 ya que supera ligeramente el valor de los negativos a los positivos, aproximadamente un - 10% sobre el valor de la suma de los positivos, **concluyendo con el resultado sentimiento global lamentable pero ligeramente más negativo**.

CAPÍTULO 5. CONCLUSIONES

El virus del COVID-19 se ha convertido en uno de los principales temas constantes en la opinión pública. La pandemia ha impactado a más de 200 países en todo el mundo, infectando a más de 510.9 millones de personas y causando más de 6,2 millones de muertes al 24 de abril de 2022. Actualmente existen bastantes investigaciones sobre la creciente evolución del COVID-19, en diferentes campos de investigación, que abonan a poder analizar el gran impacto que ha tenido no solo en la salud pública, si no en la economía así como en la manera en la que nos desenvolvemos al momento de comunicarnos con otros y percibimos nuestra cotidianidad. Durante este transcurso, la discusión en Twitter también ha cambiado. Las distribuciones de tweets publicados, se ha convertido en una de las principales fuentes para percibir las principales preocupaciones respecto al COVID, las opiniones hacia la gestión de la pandemia y sobre todo el avance y/o modificaciones en los sentimientos en pandemia que comparten las personas alrededor del mundo.

El monitorear las conversaciones públicas en Twitter, permite conocer aquellos temas sobresalientes, que abonen en un determinado momento a la incidencia en distintas problemáticas. Esto es particularmente valioso, a pesar de que la situación del COVID 19 cambia todos los días, gracias a estudios y al análisis pueden ser predecibles ciertos comportamientos. Twitter ha sido utilizado como un notificador de alerta temprana, comunicación de emergencia canal, monitor de percepción pública y de salud. Además, se considera una gran fuente de datos con la que se puede dar seguimiento no solo al tema del COVID 19, si no a otra variedad de enfermedades, desastres naturales, problemas ambientales incluso temas políticos.

El objetivo de esta tesis apunta al uso precisamente de Twitter, usando en amplia medida, métodos automáticos para la recopilación y análisis, que nos da como resultado, las referencias de los Twitts sobre temas relevantes, palabras más frecuentes, hashtags e incluso se apunta hacia los usuarios populares activos en la red social y con lograr evaluar en base a estos indicadores la opinión y el sentimiento de las personas.

Hasta ahora existen varios trabajos que han abordado el problema del análisis de información en redes sociales, más específicamente en la red social Twitter y la tendencia del COVID-19. Cada uno de ellos está centrado en distintas técnicas y herramientas.

Después de realizar un análisis exhaustivo de los trabajos más relevantes, se han identificado las siguientes carencias en estas propuestas y por lo tanto, las ventajas de nuestro modelo propuesto en capítulos anteriores:

- No existe un framework sólido que explote el estado del arte de las técnicas de Procesamiento de Lenguaje Natural y permita detectar de manera orgánica temas tendencia en redes sociales.
- No hay una definición estándar de diccionario de datos, lo cual causa que los resultados obtenidos no sean confiables al no tener diccionarios de datos confiables durante el entrenamiento de los modelos.
- La verificación de los modelos se realiza en lenguajes y formalismos específicos lo que dificulta comprobar su funcionalidad en otros ambientes.

Nuestro modelo de diccionario de datos con análisis de sentimientos:

- Permite contar con una solución genérica con un amplio rango de aplicación para el análisis de tendencias en redes sociales.
- Teniendo una definición estandarizada de diccionario de datos se podrán realizar entrenamientos confiables de los modelos de análisis de lenguaje natural para ambientes en redes sociales.

Durante la documentación de esta tesis se realizó una investigación en la que sin duda hubo diversos trabajos que tratan de forma genérica el análisis de datos en redes sociales, en los cuales se utilizan distintas técnicas y herramientas. Sin embargo, ninguno de ellos ofrece un estándar o modelo que permita la obtención de datos confiables y de manera orgánica los temas en tendencia en las redes sociales.

Es así que se propuso un modelo para la detección de temas en tendencias, este modelo está basado en algoritmos de Procesamiento de Lenguaje Natural llamados “Transformers” que pertenecen a una rama de la Inteligencia Artificial, el cual nos ofrece grandes posibilidades, con un enorme potencial y se encuentra en constante evolución para la mejora en el procesamiento de datos.

Además, también se integró al modelo un diccionario de datos para clasificar la polaridad de las palabras si son negativas o positivas, con la finalidad de enriquecer y reforzar, los resultados obtenidos, los cuáles aportaron una mayor interpretación y permitieron medir el sentimiento en general, de los usuarios del Twitter sobre el tema de estudio.

Se examinó a detalle la naturaleza y componentes de las conversaciones de Twitter, así como las estructuras usadas, el tipo de publicaciones que se generan por los usuarios, que permitió establecer los criterios de recopilación de información de una forma organizada y limpia.

Llegando a este punto, se desarrolló una aplicación con métodos y criterios definidos que permitieran la recopilación de información para su posterior explotación, logrando almacenar alrededor de 10,3 millones de tweets entre el periodo del año 2021 hasta abril del 2022, relacionados al tema del COVID-19.

A partir de esa información y con la implementación del modelo es como se logró obtener en este último capítulo los resultados detección de temas de tendencia en forma de palabras claves que provienen del análisis de datos masivo y que

se resume a un nivel de tan solo algunas palabras claves, que fácilmente podemos leer y comprender por los humanos, es decir, diccionario de datos que permita mejorar el análisis de datos y sentimientos en Twitter.

Se ha logrado cumplir los objetivos hasta llegar a realizar las pruebas con el producto obtenido de esta tesis, el procesamiento de los datos por la aplicación fue realmente rápido, sin embargo, un factor que implico una limitación fueron los recursos de cómputo, como CPU y GPU más potentes para ampliar la capacidad de procesamiento con la mayor cantidad de datos, es por ello que dejamos abierto a discusión o replicación del modelo para otros temas de estudio, como trabajo futuro.

Para fortalecer y mejorar los estudios sobre el COVID-19 a través de redes sociales como Twitter es crucial generar bases datos públicas que contengan publicaciones o microblogging relacionado del COVID-19. Estas bases de datos harán a su vez de diccionarios de datos que permitirán comprender mejor el fenómeno actual.

Es por esto que un diccionario de datos aportará una base de información con los indicadores emocionales que permitirá identificar la polaridad de una oración o párrafo con más facilidad, llegado a este punto es un gran avance logrado para hacer un estudio sobre nuestro tema de interés del COVID-19 y utilizaremos las palabras y términos relacionados al caso. El diccionario deberá ser de auto aprendizaje para que vaya agregando nuevos términos con el tiempo y así refuerce su capacidad de interpretación, incluso pudiera lograr ir incursionando otros campos o temas diferentes o en otros idiomas.

El uso del diccionario de datos es para trabajar en conjunto con los métodos de análisis con inteligencia artificial y lograr obtener los temas relevantes, pero además con la implementación del diccionario de datos, clasificar la polaridad de los mensajes si son positivos o negativos para enriquecer aún más los resultados con información más apropiada y confiable.

En análisis de sentimientos en Twitter es un campo que promete mucho aún y es posible explotar las capacidades de los algoritmos y técnicas existentes en combinación con otras, para mejorar la obtención de información valiosa del mundo que nos rodea, a través del análisis de grandes cantidades de datos generados por los usuarios diariamente.

En el presente trabajo se presentó un modelo que aprovecha las ventajas de usar los instrumentos de inteligencia artificial para el procesamiento de información, así como el uso de técnicas avanzadas de análisis y además de un diccionario de datos de auto aprendizaje, este guarde en memoria y aporte más información a la hora de analizar las tendencias y sentimientos, para que la clasificación sea más precisa y con ello reducir el esfuerzo humano y computacional. De esto depende tener más confiabilidad con los resultados.

Consideramos que este trabajo puede ser la base para contar con diccionarios de datos no solo para temas de COVID-19 sino para trending topics de importancia general de tal forma que permita establecer de una forma más real

las tendencias, lo sentimientos y la situación que se vive respecto a dichos temas.

Además, permitirá ser la base para llevar a cabo este análisis utilizando otras redes sociales y como hemos dicho antes, será importante contar con base de datos e información pública que permita no solo generar el diccionario, sino lograr el auto aprendizaje utilizando técnicas de inteligencia artificial.

BIBLIOGRAFÍA

Aguilar, L. J. (2016). Big Data, Análisis de grandes volúmenes de datos en organizaciones. Alfa Omega Grupo Editor, S.A de C.V., Mexico ISBN 978-607-707-689-6.

Chan , H., Wang, X., Lacka, E., & Zhang, M. (2016). A Mixed-Method Approach to Extracting the Value of Social Media Data. *Production and Operations Management Society*, 568-583.

Padilla, G. (2018). Instagramers e influencers. El escaparate de la moda que eligen los jóvenes menores españoles. *Revista Internacional de Investigación en Comunicación a DResearch ESIC*, 42-59.

R. K. (2016). A survey of data mining and social network analysis based anomaly detection techniques. *Egyptian Informatics Journal*, 199-216.

Criado, J. (2018). Comunicando Datos Masivos del Sector Público Local en Redes Sociales. Análisis de Sentimiento en Twitter. *El profesional de la información*, 614-624.

A. C. (2015). 'We (don't) know how you feel' – a comparative study of automated vs. manual analysis of social media conversations. *Journal of Marketing Managemen*, 1141-1157.

A. P. (2018). Analysis and Visualisation of Geo-Referenced Tweets for Real- Time Information Diffusion. *Procedia Computer Science*, 1138-1146.

F. M. (2016). We Are What We Generate - Understanding Ourselves Through Our Data. *Procedia Computer Science*, 335-344.

G. P. (2018). Instagramers e influencers. El escaparate de la moda que eligen los jóvenes menores españoles. *Revista Internacional de Investigación en Comunicación a DResearch ESIC*, 42-59.

J.-P. H. (2018). Leveragung Social Media Metrics In Improving Social Media Performances Through Organic Reach: A Data Mining Approach. *Review Of Economics & Business Studies*, 33-48.

Muthiah, C. (2017). Performance of Sentimental Analysis by Studying and Mining Social Media using Parsing Technique. *First International Conference on Information Technology, Communications and Computing.* India.

Navarra, P. L., Borrull, A. L., Navarro, J. S., & P. Y. (2018). Medición de la influencia de usuarios en redes sociales: propuesta socialengagement. *El profesional de la información* , 899-908.

Poecze, F. (2018). Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts effectiveness of social media posts. *Procedia Computer Science*, 660-666.

R. H. (2019). Facebook posting activity and the selective amplification of earnings disclosures. *China Journal of Accounting Research*, 135-155.

R. K. (2016). A survey of data mining and social network analysis based anomaly detection techniques. *Egyptian Informatics Journal*, 199-216.

S. B. (2017). Mining for Social Media: Usage Patterns of Small Businesses. *Business Systems Research*.

S. S. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 156-168.

T. T. (2017). Personality Prediction System from Facebook Users. *Procedia Computer Science*, 604-611.

Tricco, A., Zarin, W., Lillie, E., Jeblee, S., Warren, R., Khan, P., . . . Straus, S. (2018). Utility of social media and crowd- intelligence data for pharmacovigilance: a scoping review. *BMC Medical Informatics and Decision Making*.

Yuvaraj, N. (2015). A High End Sentimental Analysis in Social Media Using Hash Tags. *Journal of Applied Science and Engineering Methodologies*, 137-143.

Patel A, Jernigan DB, (2020) 2019-CoV CDC Response Team. Initial Public Health Response, States.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 48–57.

Lampos , V., & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. *2nd International Workshop on Cognitive Informa- tion Processing*, 411-416.

Latorre, D. M. (2018). *HISTORIA DE LAS WEB, 1.0, 2.0, 3.0 y 4.0.* Santiago de Surco: Universidad Marcelino Champagnat.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St John, R., . . . Chris, T. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Allen, J. (1995). *"Natural Language Understanding".* Redwood City.: Benjamin/Cummings.

A. C. (2015). 'We (don't) know how you feel' – a comparative study of automated vs. manual analysis of social media conversations. *Journal of Marketing Managemen*, 1141-1157.

Aguilar, L. J. (2016). *Big Data, Análisis de grandes volúmenes de datos en organizaciones.* Alfa Omega Grupo Editor, S.A de C.V., Mexico ISBN 978-607-707-689-6.

Aguilar, L. J.-6.-7.-6.-6. (n.d.).

Asgari-Chenaghlu, M., Nikzad-Khasmakhi, N., & Minaee, S. (2020). Covid-Transformer: Detecting COVID-19 Trending Topics on Twitter Using Universal Sentence Encoder. *arXiv:2009.03947*.

Ashish Vaswani, N. S. (2017). *Attention is all you need. In Advances in neural information processing systems pages 5998–6008.*

Balan, S., & Rege, J. (2017). Mining for social media: Usage patterns of small businesses. Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy, 8(1), 43-50.

Berenguer, J. A. (2019). *Redes sociales y marketing 2. COMM092PO.* Malaga: IC Editorial.

Bode, L., Kawintiranon, K., Chi, G., & Vraga, E. (2020). A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv:2003.13907v1*.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hyper- textual Web search engine. *Computer Networks and ISDN Systems*, 1-7.

Brochu, E., & Freitas, N. (2002). Name that song! *NIPS*, 1505-1512.

Daniel Cer, Y. Y.-y. (2018). *Universal sentence encoder.* arXiv preprint arXiv:1803.11175.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Fellbaum, C. (1998). WordNet. *Wiley Online Library*. Retrieved from Wiley Online Library.

Federico Barrios, F. L. (2016). *Variations of the similarity function.* arXiv preprint arXiv:1602.03606,.

Gao, Z., Yada, S., Wakamiya, S., & Aramaki, E. (2020). NAIST COVID: Multilingual COVID-19 Twitter and Weibo Dataset. *arXiv:2004.08145*.

Greene, D., & Cunningham, P. (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering. *ICML*, 377-384.

Grishman, R. (1986). *"Computational Linguistics: an introduction". .* Cambridge, Cambridge University Press. .

Grigori Sidorov, Sabino Miranda-Jimenez, Francisco Viveros-Jimenez, Alexander Gelbukh, Noah Castro-Sanchez, Francisco Velasquez, Ismael Diaz-Rangel, and John Gordon . Estudio Empírico de Minería de Opinión en Tweets Españoles . LNAI 7629, 2012, págs. 121-12. 1-14. 2.Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 1735–1780.

Hui DS, A. E. (2020). *The continuing 2019-nCoV threat of novel coronaviruses to global health-the latest 2109 novel coronavirus outbreak in Wuhan, China. Int J Infect Dis. 2020;91:264.* . Retrieved from https://doi.org/10.1016/j.

Iglesia, J. L. (2010). *Web 2.0: una descripción muy sencilla de los cambios que estamos viviendo.* España: Netbiblo.

Ismael Diaz Rangel, Grigory Sidorov, Sergio Suarez-Guerra. Creación y evaluación de un diccionario marcado y ponderado de emociones para el español . Onomazeína, 29, 23 p. , 2014 , DOI 10.7764/onomazeína.29.5

Iyyer, M., Manjunatha, V., Boyd-Graber, J., & Daume III, H. (2015). Deep unordered composition rivals syntactic methods for text classification. *Proceedings of ACL/IJCNLP*.

J.-P. H. (2018). Leveragung Social Media Metrics In Improving Social Media Performances Through Organic Reach: A Data Mining Approach. *Review Of Economics & Business Studies*, 33-48.

Jackson, M. y. (2002). *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization.* John Benjamins Pub.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 604-632.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. *International conference on machine learning*, 957-966.

Müller, M., Salathé, M., & Kummervold, P. (2020). COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv:2005.07503*.

Martinez, C. (2020, 06 05). *El Universal*. Retrieved from https://www.eluniversal.com.mx/cartera/en-mexico-39-de-los-cibernautas-utilizan-Twitter

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404-411.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111-3119.

Moreno Sandoval, A. (1998). *Lingüística Computacional. Introducción a los modelos simbólicos, estadísticos y biológicos.* Madrid. Editorial Sintesis.

Navarra, P. L., Borrull, A. L., Navarro, J. S., & P. Y. (2018). Medicion de la influencia de usuarios en redes sociales: propuesta socialengagement. *El profesional de la información* , 899-908.

OMS. (2020). *Coronavirus disease (COVID-19) outbreak. Geneva: WHO*. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019

Ordun, C., Purushotham, S., & Raff, E. (2020). Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs. *arXiv:2005.03082*.

Padilla, G. (2018). Instagramers e influencers. El escaparate de la moda que eligen los jóvenes menores españoles. *Revista Internacional de Investigación en Comunicación a DResearch ESIC*, 42-59.

Patel A, J. D.-C. (2022, Enero 10). CDC Response Team. Initial Public Health Response and Interim Clinical Guidance for the 2019 Novel Coro-navirus Outbreak.

Quadrianto, N., Song, L., & Smola, A. (2009). Kernelized sorting. *NIPS*, 1289-1296.

R. K. (2016). A survey of data mining and social network analysis based anomaly detection techniques. *Egyptian Informatics Journal*, 199-216.

Reimers, N., & Gurevych, I. (2019). Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., & Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in Twitter. . *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Rubio López, R. Y., & Bernal Chávez, J. A. (2016). *Introducción a la Lingüística Computacional.* Bógota, Colombia: Ediciones de la U.

Schölkopf, B., Weston, J., Eskin, E., Leslie, C., & Noble, W. (2002). A kernel approach for learning from almost orthogonal patterns. *ECML*, 511-528.

S. B. (2017). Mining for Social Media: Usage Patterns of Small Businesses. *Business Systems Research*.

S. S. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 156-168.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & man- agement*, 513-523.

Socher, A., Perelygin, J., Wu, J., Chuang, J., Manning, C., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of* the conference on empirical methods in natural language processing (EMNLP), 1642.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. International journal of information management, 39, 156-168.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., . . . Polosukhin, I. (2017). Attention is all you need. *Proceedings of NIPS.*

Veronica Perez Rosas, Carmen Banea, Rada Mihalcea, "Learning Sentiment Lexicons in Spanish" Proceedings of the international conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 2012.

Vosoughi, S., Vijayaraghavan, P., & Roy, D. (2016). Tweet2Vec: Learning Tweet Embeddings Using Character-level CNN-LSTM Encoder-Decoder. *SIGIR '16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 1041–1044.

Worldometer. (n.d.). Retrieved from COVID-19 CORONAVIRUS PANDEMIC: www.worldometers.info/coronavirus/

Xu, W., Callison-Burch, C., & Dolan, W. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in Twitter. *Proceedings of the 9th International Workshop on Semantic Evaluation*.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational intelligenCe magazine*, 55–75.

Zhang, X., & LeCun, Y. (2015). Text understanding from scratch. *arXiv:1502.01710*.