ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

DESARROLLO DE UN MODELO "BIAS-AWARE" DE RECOMENDACIÓN DE CONTENIDO PARA USUARIOS DE TWITTER CON USO DE ESTRATEGIAS DE APRENDIZAJE AUTOMÁTICO Y MINERÍA DE TEXTO

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DE GRADO DE MAGÍSTER EN COMPUTACIÓN CON MENCIÓN SISTEMAS INTELIGENTES

CLEOPATRA YOMARA GUERRA ALMEIDA
cleopatra.guerra@epn.edu.ec

DIRECTOR: ING. LORENA RECALDE CERDA, PhD.
lorena.recalde@epn.edu.ec

CODIRECTOR: ING. EDISON LOZA, PhD.
edison.loza@epn.edu.ec

Quito, noviembre 2023

APROBACIÓN DEL DIRECTOR

Certifico que el presente trabajo fue desarrollado por Guerra Almeida Cleopatra Yomara, bajo mi supervisión.

Firmado electrónicamente por:
LORENA KATHERINE
RECALDE CERDA

Ing. Lorena Recalde Cerda, PhD.

DIRECTOR DE PROYECTO

APROBACIÓN DEL CODIRECTOR

Certifico que el presente trabajo fue desarrollado por Guerra Almeida Cleopatra Yomara, bajo mi supervisión.

Firmado electrónicamente por:
EDISON FERNANDO
LOZA AGUIRRE

Ing. Edison Loza, PhD.

CODIRECTOR DE PROYECTO

DECLARACIÓN

Yo, Cleopatra Yomara Guerra Almeida , declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Firmado electrónicamente por:
CLEOPATRA YOMARA GUERRA ALMEIDA

Cleopatra Yomara Guerra Almeida

DEDICATORIA

A mi amada mami,

Quien es mi inspiración y mi guía. Tu inquebrantable fe en mí, es mi motor en los momentos de duda y temor. Gracias por siempre creer en mí y nunca soltarme. Por demostrar tu amor incondicional innumerables ocasiones. Me enseñaste que con esfuerzo y perseverancia, los sueños, aunque avancen a paso lento, pero firme, se cumplen. Admiro tu incansable dedicación y entrega.

AGRADECIMIENTOS

Empezaré expresando mi gratitud a Dios, porque en los momentos que sentí que no podía y quise rendirme. Él me sostuvo impulsándome a seguir adelante.

Quiero extender mi agradecimiento a la Escuela Politécnica Nacional, en particular al Vicerrectorado de Investigación, Innovación y Vinculación por brindarme la oportunidad invaluable de retomar mis estudios.

Mi más profundo agradecimiento se dirige hacia mi directora de tesis, Dra. Lorena Recalde, por su paciencia, conocimiento, orientación y disposición para guiarme a lo largo de este proceso. De igual manera, quiero agradecer al Dr. Edison Loza por acogerme en el Laboratorio ADA, por brindarme su apoyo constante, sus sabios consejos, compromiso, dedicación y apoyo en cada paso de este camino.

CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Como una solución a los inconvenientes generados por la presencia de sesgo en los Sistemas de Recomendación, este proyecto propone el desarrollo de un modelo que permita disminuir el sesgo en las recomendaciones de contenido de Twitter basado en estrategias de análisis de sesgo. El objetivo es que se pueda detectar, simplificar, minimizar y alcanzar cierto grado de imparcialidad durante la ejecución de los algoritmos que permiten identificar el sesgo en los Sistemas de Recomendación. El modelo propuesto debe proporcionar recomendaciones con una mínima presencia de sesgo. Con el desarrollo del modelo *bias-aware*, este trabajo persigue evaluar y demostrar la existencia de sesgo, lo cual ayudará a diseñar mejores sistemas de software para recomendación. De esta manera, el modelo que se desarrollará contribuirá en la implementación del nuevo Sistema de Recomendación. El corpus con el que se trabaja corresponde al Paro Nacional que se dio en junio 2022 y la Muerte Cruzada en junio 2023. El procedimiento empieza con la limpieza de los datos. Aquí se emplea técnicas tales como: tokenización, eliminación de *stopwords*, *tweets* duplicados, entre otras. Cuando los datos se encuentran limpios se emplea *word2vec* para crear vectores de 50, 100 y 200 dimensiones. A continuación, se aplica el algoritmo k-means para identificar clústers. Se calculan los centroides de cada clúster y una vez que se determina el clúster de política se etiqueta los *tweets* que tienen contenido político. Luego se evalúa con diferentes algoritmos de clasificación como: SVM, árboles de decisión, KNN y Naive Bayes para las diferentes dimensiones. Finalmente, se propone un modelo que empleará un nuevo diccionario de política formado a partir de las distancias mínimas de las palabras del clúster de política a los otros clústers. Con este nuevo clúster se repite el procedimiento de evaluación de algoritmos. Las métricas empleadas son: *Accuracy*, *F1-score*, *Recall*, Precisión, F2 y F0.5. Los mejores resultados se obtienen con el nuevo clúster, con una dimensión de 200 para los algoritmos de SVM y árboles de decisión.

Palabras Clave: Sesgo, Twitter, trends, tweets, tokenización,stopwords, word2vec, k-means, SVM, árboles de decisión, KNN, Naive Bayes, Accuracy, F1-score, Recall, Precisión, F2, F0.5.

ABSTRACT

As a solution to the challenges generated by the presence of bias in Recommendation Systems, this project proposes the development of a model that allows reducing bias in Twitter content recommendations based on bias analysis strategies. The goal is to detect, simplify, minimize, and achieve a certain degree of impartiality during the execution of algorithms that identify bias in Recommendation Systems. The proposed model aims to provide recommendations with minimal bias.

With the development of the "bias-aware"model, this work seeks to assess and demonstrate the existence of bias, which will help design more effective recommendation software systems. Thus, the developed model will contribute to the implementation of the new Recommendation System.

The corpus used in this work corresponds to the National Strike that took place in June 2022 and the Crossed Death in June 2023. The procedure begins with data cleaning, employing techniques such as tokenization, removal of stopwords, removal of duplicate tweets, among others. Once the data is clean, word2vec is used to create vectors of 50, 100, and 200 dimensions. Next, the k-means algorithm is applied to identify clusters. The centroids of each cluster are calculated, and once the political cluster is determined, tweets with political content are labeled. Then, evaluation is performed with different classification algorithms such as SVM, decision trees, KNN, and Naive Bayes for the different dimensions. Finally, a model is proposed that will use a new political dictionary formed from the minimum distances of the words from the political cluster to the other clusters. With this new cluster, the algorithm evaluation procedure is repeated. The metrics used include Accuracy, F1-score, Recall, Precision, F2, and F0.5. The best results are obtained with the new cluster, with a dimension of 200 for the SVM and decision tree algorithms.

Keywords: Bias, Twitter, trends, tweets, tokenization, stopwords, word2vec, k-means, SVM, decision trees, KNN, Naive Bayes, Accuracy, F1-score, Recall, Precision, F2, F0.5.

1 INTRODUCCIÓN

El impacto de las redes sociales y su uso ha incrementado en los últimos años. Actualmente, una gran cantidad de datos está disponible para su análisis. Twitter, por ejemplo, constituye una red social en la cual los usuarios intercambian opiniones sobre diversas temáticas. Siendo tan elevado su alcance e influencia que hay preocupaciones que se han levantado en relación a su uso y el contenido que presenta a los usuarios. Bajo esta premisa, en el presente trabajo se realizará un análisis del sesgo que existe en las recomendaciones de contenido para los usuarios que emplean Twitter. Si bien el sesgo está presente implícitamente en acciones que se realizan a diario, como las búsquedas en los navegadores o pueden abarcar temas más complejos como su presencia en aspectos sociales, culturales, de género, raza, etnia, entre otros. Para este estudio se tomará un tipo de sesgo social que corresponde a la política en el país. En el último año el Ecuador vivió acontecimientos políticos de gran relevancia: un (Paro Nacional) y una (Muerte Cruzada). Esta problemática social, así como la definición de conceptos que orientarán el marco teórico de la investigación se explica en el Capítulo uno.

El capítulo dos presenta la metodología, desde la obtención de la información para extraer y almacenar su contenido para su análisis. En los datos recolectados se aplicarán técnicas de procesamiento de lenguaje natural que permiten almacenar un corpus limpio y listo para el estudio. Es importante recalcar el uso de la metodología *Desing Science Research* (DSRM) junto con *Cross Industry Process for Data Mining* (CRISP-DM), puesto que es una base sólida sobre la cual se desarrolla el proyecto. En la fase 1 se plantea las preguntas de investigación. La fase 2, explica la naturaleza de los datos de Twitter. Por otro lado, la fase 3 detalla el desarrollo del modelo inicial. En la fase 4, se explica la vectorización de la data en diferentes dimensiones. Finalmente, la fase 5, da respuestas a las preguntas de investigación planteadas en la fase 1 y adicionalmente, se propone el nuevo modelo que permitirá analizar el sesgo de política.

A continuación, el Capítulo tres muestra los resultados del estudio. Estos resultados se eva-

lúan con cuatro algoritmos de aprendizaje automático, estos algoritmos son: Naïve Bayes, Árboles de Decisión, k-nearest neighbor (KNN) y *Support Vector Machine* (SVM) . Además, se detalla los valores de las métricas que se obtuvieron del análisis de cada algoritmo. Las métricas empleados son: la exactitud (*acuracy*), sensibilidad o *recall*, precisión y *f1-score*, especificidad, precisión positiva, tasa de falsos positivos, el valor f2 y el valor f0.5.
Finalmente, en el Capítulo cuatro se expone las conclusiones del trabajo. De igual manera, en este capítulo se proporciona algunas recomendaciones que permitirán una mejora en el modelo planteado.

1.1 PLANTEAMIENTO DEL PROBLEMA

El crecimiento exponencial de la tecnología contribuye a que cientos de datos alrededor del mundo sean compartidos por miles de usuarios en pocos segundos. Uno de los medios de propagación e intercambio de información son las redes sociales. En Ecuador, el uso de este tipo de aplicaciones Web ha experimentado un notable crecimiento. Así, para el año 2019, apenas el 2 % [1] de la población ecuatoriana empleaba las redes sociales, mientras que para el año 2020 su uso incrementó notoriamente al 10 %. [1]. De las redes sociales existentes, Facebook es la que encabeza la popularidad en el país con mayor concurrencia; a este le siguen YouTube, Pinterest, Twitter, Instagram, Reddit y LinkedIn. Para enero del 2021 el 4.31 % de la población ecuatoriana empleaba Twitter [1], mientras que para octubre del 2021 su uso se incrementó en un 50 % [2].

Por otro lado, las redes sociales implementan sistemas inteligentes, conocidos como sistemas de recomendación (SR). Los SR colaboran a la mejora de la experiencia del usuario, fortalecen las interacciones y, en cierta medida logran fidelizarlo. Además, los SR son una respuesta al usuario respecto a su necesidad de contar con estrategias de personalización de contenido. De esta forma, los SR emplean información proporcionada por el usuario (comentarios, likes, ratings, logs de sesiones, contenido multimedia, etc.) para crear modelos que entienden y predicen sus gustos [3]. Los SR brindan al usuario la posibilidad de acceder a información de forma eficiente y posibilitan el consumo de ítems como: películas, canciones, libros, aplicaciones, sitios web, lugares turísticos, entre otros, en relación con sus preferencias.

Cada modelo de recomendación emplea diferentes algoritmos según los resultados que se espera del sistema de recomendación y al contexto de los datos [4]. Sin embargo, los SR

se comparan a una caja negra, en la cual hay mucho por explorar. De este modo, existe una gran interrogante sobre si los algoritmos de recomendación pueden llegar a ser discriminatorios con respecto a cierto contenido bajo criterios de: género, demografía, estatus socioeconómico, nacionalidad, postura política, actividades de los usuarios, cultura, lengua o etnia [5]. Un ejemplo del uso de los SR son los anuncios de empleos en la publicidad *online* al que acceden indistintamente hombres y mujeres. Se ha demostrado que en estos tipos de SR las mejores ofertas laborales son presentadas a los hombres [6].

El sesgo algorítmico prevalece incluso cuando no hay intención de una discriminación. Este sesgo puede llegar a ser más complejo cuando fluye en cascada a través de los SR. Incluso, puede ser consecuencia involuntaria del trabajo que realiza un desarrollador web al tratar de aislarlo [5].

Al no existir un análisis y control del sesgo, éste continuará presente en los SR en los que actualmente existe discriminación, producto de muchas aplicaciones de aprendizaje automático con un diseño de algoritmos que desde su planteamiento albergan un sesgo.

1.1.1 Objetivo General

Desarrollar un modelo "Bias-Aware"de recomendación de contenido para usuarios de Twitter con uso de estrategias de aprendizaje automático y minería de texto.

1.1.2 Objetivos Específicos

- Realizar una revisión de literatura sobre trabajos relacionados a la presencia de Sesgo en los Sistemas de Recomendación.
- Recolectar un corpus de comentarios en Twitter de la población ecuatoriana para crear la base de datos de estudio.
- Implementar un método de clasificación que permita identificar la presencia de sesgo en el sistema de recomendación de Twitter.
- Desarrollar un modelo de control que permita monitorear, cuantificar el sesgo algorítmico y minimizar su presencia en los Sistemas de Recomendación de contenido.
- Evaluar y validar el modelo desarrollado.

1.1.3 Alcance

El alcance del proyecto contempla la fase de recolección automática de *tweets* en el territorio Ecuatoriano. A continuación, se realiza la limpieza del corpus recolectado empleando técnicas de minería de texto. Como siguiente paso, se emplea algoritmos de aprendizaje supervisado, tales como: árboles de decisión, clasificación de Naïve Bayes, *Support Vector Machine* (SVM) ó *K-Nearest Neighbors* (KNN). El uso de los algoritmos de clasificación ayudará a cuantificar los resultados para explicar de manera numérica el comportamiento de los datos. Los pasos antes descritos buscan establecer una metodología que permita mitigar en cierto grado el sesgo existente en los algoritmos que son usados por los Sistemas de Recomendación en Twitter.

1.2 MARCO TEÓRICO

El impacto que los SR generan en la toma de decisiones y en cómo se consumen productos y servicios digitales por parte de las personas se ha incrementado en los últimos años y es motivo de estudio en la actualidad. La fidelización, personalización y reducción de tiempo de búsqueda contribuyen a que los usuarios mejoren la experiencia al encontrar productos o servicios que se ajustan a sus preferencias individuales. En cuanto al funcionamiento de los SR, son muchos los factores que se consideran como un tema de discusión, ya sea por los algoritmos, técnicas de inteligencia artificial con la que analizan los datos y la presencia de sesgo. En esta sección se detalla el marco teórico que será la base para comprender las diferentes temáticas que aborda este proyecto. Estas herramientas son: Twitter, Sistemas de Recomendación, Minería de Texto, Aprendizaje Automático, Sesgo y MongoDB.

1.2.1 Twitter

Internet contribuye en gran medida en la circulación de datos digitales. Con frecuencia, más usuarios acceden a contenido digital y forman parte de la nueva cultura de la conectividad; parte de esta cultura involucra redes sociales, tales como: Facebook, Twitter o YouTube.

Mantener una conversación, mostrar fotografías y enviar mensajes eran hábitos comunes que han sufrido un notable cambio en los últimos años; pasando de ser compartidos con un

grupo exclusivo a formar parte de una red global con acceso a cientos o incluso miles de usuarios en poco tiempo.

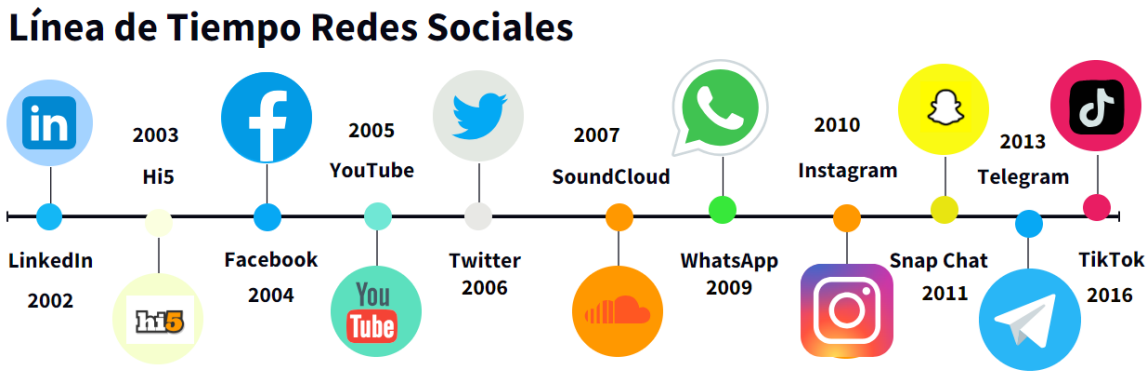


Figura 1.1: Línea del Tiempo de Redes Sociales
Elaborado por: Guerra Cleopatra

En la Figura 1.1, se muestra una línea de tiempo con el desarrollo de las Redes Sociales; de estas redes, se hará énfasis en Twitter.

Twitter es un servcio de microblogging creado por Jack Dorsey, Biz Stone y Evan Williams en el año 2006. Esta plataforma nació con la idea de enviar mensajes cortos, con un límite de 140 caracteres; actualmente, la longitud del mensaje ha incrementado a 280 caracteres, obligando a los usuarios a enviar mensajes concisos y enfocados en temas específicos. A través del tiempo, Twitter ha evolucionado para incluir nuevas funcionalidades como agregar fotos, vídeos ó enlaces.

El uso de microblog se ha vuelto una tendencia en los últimos años debido a la capacidad de permitir una comunicación rápida y concisa. Estudios muestran el papel que tiene Twitter en la divulgación de información, en la vida cotidiana de las personas o como un medio para expresar emociones o dar apoyo social [7].

A continuación, se detalla algunos términos comunes usados en esta red social [8] [9]:

- **Tweet:** Es el mensaje corto que se comparte con un grupo de seguidores en esta red social, puede incluir texto, video, enlaces o fotografías.
- **Hashtags:** Un *hashtag* es usado en un *tweet* para que el mensaje sea descubierto. Si el usuario hace clic en un *hashtag*, puede visualizar los *tweets* que emplean el mismo *hashtag*.
- **Retweets:** Consiste en compartir un *tweet* de otro usuario. Este mecanismo permite a los usuarios difundir información a su elección más allá del alcance de los seguidores originales del *tweet*.

- **Likes:** Es la acción que demuestra la aceptación con el contenido del *tweet*.
- **Respuestas:** Son los mensajes que los usuarios colocan bajo un determinado *tweet* para participar en una conversación.
- **Mensajes Directos:** Son aquellos mensajes que se envían de manera privada a otros usuarios.
- **Listas:** Son empleadas para organizar las cuentas que siguen en categorías específicas.
- **Trending Topic:** Puede ser una palabra, una frase o un tema del momento que se hace popular en poco tiempo.

Gracias a su naturaleza en tiempo real, Twitter es una plataforma popular en la que cada minuto se muestran noticias de última hora, eventos en vivo y discusiones sobre diversos temas. La funcionalidad de Twitter permite que los usuarios que tienen intereses similares puedan conectarse.
En Twitter un usuario puede seguir a cualquier usuario. El uso de esta red social ha evolucionado en la cultura, es así como más usuarios se conectan a Twitter para emplear términos como: RT, para denotar un retweet, @ seguida del de un nombre de usuario se emplea para dirigir un mensaje a un usuario en específico y # seguido de una palabra se refiere a un *hashtag* [9].
Para mejorar su servicio y mantener su tendencia, Twitter busca incorporar nuevas funcionalidades tales como: calidad de los vídeos, una función de búsqueda y la creación de salas de audio en vivo para conversaciones en grupo. Debido a la gran cantidad de información que es trasmitida en esta red social, los científicos y analistas de datos obtienen información valiosa sobre tendencias, comportamientos sociales, culturales y opiniones de los usuarios.

1.2.2 Sistemas de Recomendación

La idea básica de un sistema de recomendación es emplear información de varias fuentes para inferir en los gustos de los usuarios. Dos términos muy comunes en los SR son: los usuarios y los ítems. El usuario es la entidad a la que se proporciona la recomendación y el producto que se recomienda es el ítem, es decir las recomendaciones se basan en la interacción de ítems y usuarios [3].
Un aspecto importante en los SR es la manera cómo se formula el problema de recomendación, siendo muy común [10]:

- **El problema de la predicción:** También conocido como el problema de completar matrices. La idea parte de que m usuarios y n ítems les corresponde una matriz m*n incompleta, donde los valores especificados u observados se emplean para el entrenamiento y los valores faltantes o no observados se predicen mediante los algoritmos de aprendizaje [10].

- **Problema de clasificación:** Este problema está presente cuando un usuario desea recomendar un ítem específico. El usuario da una calificación al ítem. Esta calificación puede generar un error al asignar un mismo ítem a varias categorías. Esto provoca que la clasificación puede no ser precisa con los ítems que se recomienda [10].

Clasificación de los Sistemas de Recomendación

El principio de las recomendaciones es la existencia de dependencias significativas, estas dependencias se pueden aprender a través de matrices de calificaciones que generan un modelo que es empleado para hacer predicciones para los usuarios. Si se usan calificaciones de múltiples usuarios de manera colaborativa para predecir calificaciones faltantes, el filtrado es colaborativo. Si las calificaciones de los usuarios y las descripciones de los atributos de los artículos son aprovechados para hacer predicciones, se trata de un filtrado de contenido. Por otro lado, cuando en lugar de usar calificaciones históricas o datos de compra se emplea bases de conocimiento externo o requisitos de los usuarios especificados explícitamente se trata de Sistemas de Recomendación basados en conocimiento. Algunos SR combinan estos aspectos forman sistemas híbridos. Estos últimos emplean las fortalezas de otros SR lo que proporciona que creen técnicas con mayor solidez en una amplia variedad de entornos [3].

A continuación, en la Figura 1.2 se menciona algunos tipos de SR [3].

- **Colaborativos:** Basados en las actividades que tienen los usuarios al hacer recomendaciones. Los algoritmos se emplean para analizar el registro de los clics o de las calificaciones que el usuario da a un artículo para luego hacer una recomendación basada en similitudes [3].
- **Basados en Contenido:** Los algoritmos aprenden los patrones de las preferencias de los usuarios y realizan las recomendaciones. La información de contenido puede ser de autores, directores, género, etc [3].

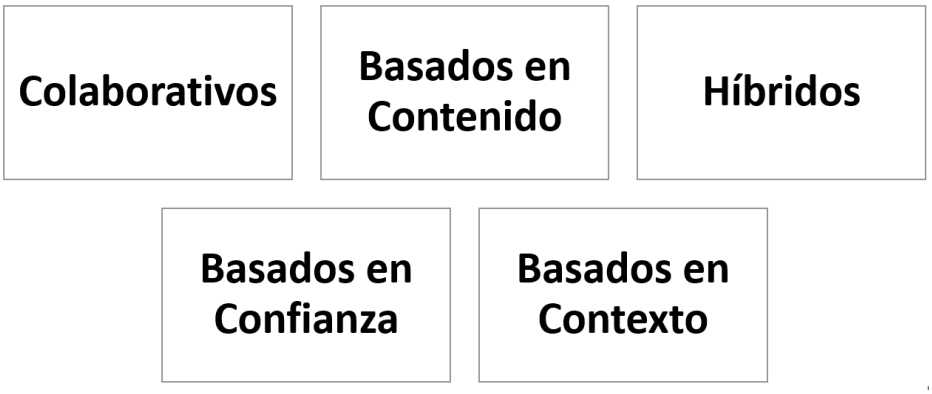


Figura 1.2: Clasificación Sistemas de Recomendación
Elaborado por: Guerra Cleopatra

- **Híbridos:** Son la combinación de un sistema colaborativo con uno basado en contenido, emplean diferentes técnicas de combinación ponderada, de cascada o de características [3].
- **Basados en Confianza:** Emplean información de confianza del usuario, los algoritmos son usados para analizar los datos de actividad del usuario, como registro de calificaciones o críticas de otros usuarios, para luego hacer una recomendación basada en la confianza de un usuario, pero para otro usuario [3].
- **Basados en Contexto:** Emplean información contextual como la ubicación geográfica, el momento del día o una actividad del usuarios, también emplean patrones para aprender preferencias y dar recomendaciones más precisas con la información contextual [3].

Existen algunos escenarios en los cuales los SR son empleados, por ejemplo, las sugerencias de series o películas en Netflix, los productos de Amazon, los amigos en Facebook, las noticias de Google News, entre otros. Los SR están muy inmersos en la vida cotidiana, han llegado a tener un papel importante en las decisiones que el usuario tiene sobre algún tema, gusto o preferencia.

A medida que incrementa la cantidad de datos empleados para el análisis y la capacidad de procesamiento de estos datos, los SR se vuelven más precisos. El uso adecuado de un sistema de recomendación contribuye a la satisfacción de un cliente en cuanto a la personalización del contenido que recibe, se incrementan ventas en un negocio e incluso se ahorra tiempo al convertirse en una herramienta que contribuye a que un usuario encuentre determinado producto o servicio en menos tiempo y con menos esfuerzo. Sin embargo, un punto importante a considerar es la presencia del sesgo, al tener una recomendación específica

hacia cierto tipo de producto o servicio que llega a limitar la diversidad en las recomendaciones. Por otro lado, un sistema de recomendación invade la privacidad de un usuario, debido a que trabaja con la recopilación de gran cantidad de información personal. La privacidad se ve expuesta en ciertos casos en los que los SR no son totalmente trasparentes. Esta transparencia del sistema de recomendación se refiere a cómo hace la recomendación, en el caso de que la recomendación no es correcta, conlleva a una desconfianza del usuario.

1.2.3 Minería de Texto

La minería de texto, ó *Text Mining*, es una técnica de procesamiento de datos con la que se puede extraer información útil y significativa de grandes cantidades de datos en un formato no estructurado. Para el análisis de esta información se emplea técnicas estadísticas y lingüísticas que ayudan a identificar patrones y tendencias dentro del texto.
El proceso empieza con la recopilación de información. A continuación, se prepara el texto, eliminando caracteres no deseados y normalizando los términos empleados. Es común emplear técnicas como la tokenzación, lematización y etiquetado gramatical. Estas técnicas ayudan a comprender las palabras y la relación que existe entre ellas [11].
Algunos estudios muestran que cerca del $80\,\%$ de la información representa a datos no estructurados [11]. En ciertos casos, la minería de texto suele ser un poco más compleja que la minería de datos por trabajar con datos no estructurados.
La minería de texto es un campo multidisciplinario debido a que involucra la recopilación de información, el análisis de texto, la extracción de información, la agrupación, la categorización, la visualización, bases de datos, el aprendizaje automático, y la minería de datos. [11].
En la Figura 1.3, se muestra los pasos del *Text Mining*, comenzando con la recolección de información. En esta parte el conjunto de datos es no estructurado, esta data puede texto, imágenes, video o audio, recolectada de redes sociales, documentos, entre otras. Al referirse a datos no estructurados quiere decir que se trabaja con información que no está organizada y que requiere de técnicas especiales de procesamiento de lenguaje para convertirse en datos estructurados que es la siguiente fase. A continuación, se identifican los patrones que ayudarán a obtener información útil empleando aprendizaje automático. Finalmente, se almacena en una base de datos [12]. Este conjunto de datos es comúnmente conocida como corpus.
El elemento clave del *Text Mining* es la colección de documentos. Estos documentos pue-

Figura 1.3: Descripción de los pasos empleados en *Text Mining*
Elaborado por: Guerra Cleopatra

den ser estáticos en el caso de que no cambian o dinámicos que se caracterizan por tener actualizaciones de periodos de tiempo. Otro elemento básico es el documento en sí, que contiene gran cantidad de información que forma parte de la colección de documentos. Estas colecciones son información no estructurada, puesto que contiene caracteres especiales, números, signos de puntuación o letras mayúsculas, que luego de aplicar técnicas de *Text Mining* se convierten en documentos estructurados que serán parte del corpus.
El correcto análisis de los datos parte de dos objetivos del *Text Mining*, el primer objetivo es la recolección de archivos que cuenten con calidad de contenido y el segundo objetivo es la elección de las características sobre las cuales se realizará el análisis, es decir los patrones sobre los cuales se tomará la información [13]. Estos datos pueden contener:

- **Caracteres:** Son componentes individuales, números o signos especiales. De acuerdo con su análisis dan cierto grado de interpretación, por ejemplo en el barrido de la información se analiza que se encuentra un símbolo como: (:)) que corresponde a un emoji de cara feliz. Este símbolo dentro del análisis da la interpretación de que el usuario está feliz, está de acuerdo con algún comentario, etc.
- **Palabras:** Esta información puede estar dentro de frases, expresiones y palabras simples. Para el análisis de esta información se debe tomar en cuenta los espacios existentes, aquí se ayuda de las *stopwords*. Las *stopwords* son palabras que dentro de un idioma específico no tienen ningún valor dentro del texto.
- **Términos:** son palabras simples y palabras que forman parte de una frase, en esta

parte se debe hacer una normalización de la información que se consigue con la tokenización y lematización. Ejemplo: El presidente Abraham Lincoln experimentó una carrera de éxito en la Casa Blanca. Palabras simples: Lincoln, tomó, carrera, éxito, experimentó Multipalabras: presidente Abrajam Lincoln, Casa Blanca.

- **Conceptos:** Un concepto hace referencia a un conjunto amplio de información que implica la identificación y agrupación de términos relacionados o palabras claves. Las palabras comparten significados similares ó poseen una relación semántica. Para identificar los conceptos se requiere del uso de técnicas avanzadas como: el análisis vectorial que mide la similitud de las palabras en base a la relación semántica existente [14]. Para comprender esto, se toma el siguiente ejemplo, un artículo sobre "deporte extremo" no necesariamente contiene la frase "prueba de manejo", sin embargo este concepto puede estar dentro de otro grupo de palabras. De aquí la importancia de emplear metodologías que usan análisis vectorial, para medir similitudes y asociar las palabras de manera conceptual.

Arquitectura Funcional de *Text Mining*

Como se viene explicando en esta sección, un sistema de minería de texto en su entrada toma varios documentos que pasan por:

- **Tareas de Procesamiento:** Incluye todas las tareas necesarias para preparar los datos para las operaciones que conllevan al descubrimiento del conocimiento. Estas tareas suelen centrarse en la categorización y el pre procesamiento de las fuentes de las cuales se toman los documentos. En esta parte se forman nuevas colecciones de documentos que están formadas por conceptos [15].
- **Operaciones básicas de Minería de Texto:** Estas tareas son la parte central de la arquitectura de funcionamiento. En esta parte se evalúa la frecuencia de las palabras, se identifica patrones y se identifica niveles [15].
- **Técnicas de Refinamiento:** Incluyen tareas que filtran información redundante y agrupan los datos estrechamente relacionados.

La arquitectura se detalla en la Figura 1.4. **Extracción de Información**

A continuación, se explicará la extracción de la información que es una parte elemental del proceso de *Text Mining*. El proceso se muestra en la Figura 1.5. Este proceso empieza con la **tokenización**, también conocida como zonificación. La *tokenización* divide al documento de entrada en bloques, es decir separa el contenido del documento en bloques de

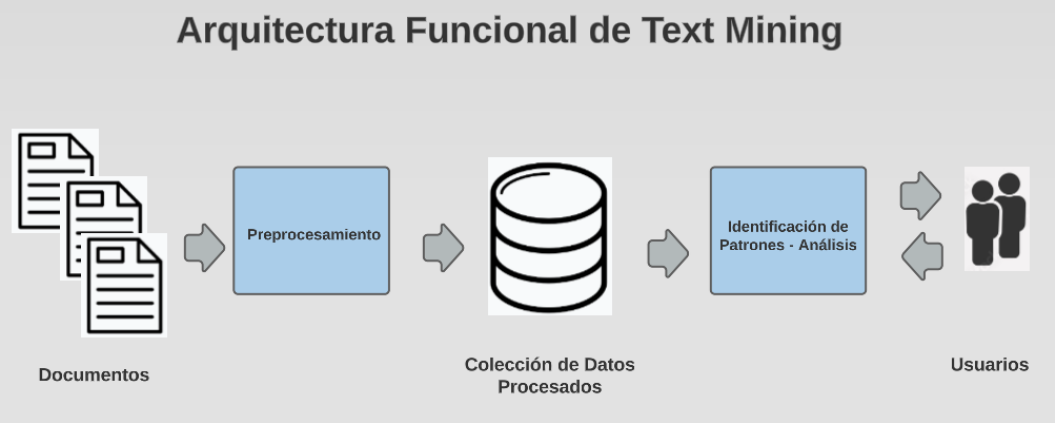


Figura 1.4: Descripción de la Arquitectura Funcional de *Text Mining*
Elaborado por: Guerra Cleopatra

palabras, oraciones y párrafos. El siguiente paso es el **módulo de análisis morfológico y léxico**, que consiste en quitar la ambigüedad de las palabras. En esta parte se crean frases básicas como: frases nominales y frases verbales. El tercer componente es el **análisis sintáctico** que ayuda a establecer la conexión entre las diferentes partes de la oración. Este análisis puede llegar a ser completo o superficial. Finalmente, el **análisis del dominio**, que es una función que combina toda la información recopilada en los pasos antes descritos y crea información completa que muestra la relación que mantienen las palabras dentro del corpus [15]. Ejemplo del proceso:

Oración: Luis ama comer pizza.

Tokenización: ["Luis", “ama”, “comer”, "pizza","."]

Análisis morfológico y léxico: [(Luis, PRON), (ama,VERB), (comer, VERB), (pizza, SUST), (.,PUNT)]

Análisis semántico:

("Luis") es el sujeto.

(“ama comer pizza") es el predicado.

("pizza") es el objeto directo.

1.2.4 Aprendizaje Automático

El aprendizaje automático es conocido como *Maching Learning*, es una rama de la Inteligencia Artificial que tiene como objetivo trabajar con modelos matemáticos y algoritmos que aprenden a partir de datos y realizan tareas específicas. Las tareas que realizan los algoritmos, en lugar de ser programados de manera explícita, usan el entrenamiento para

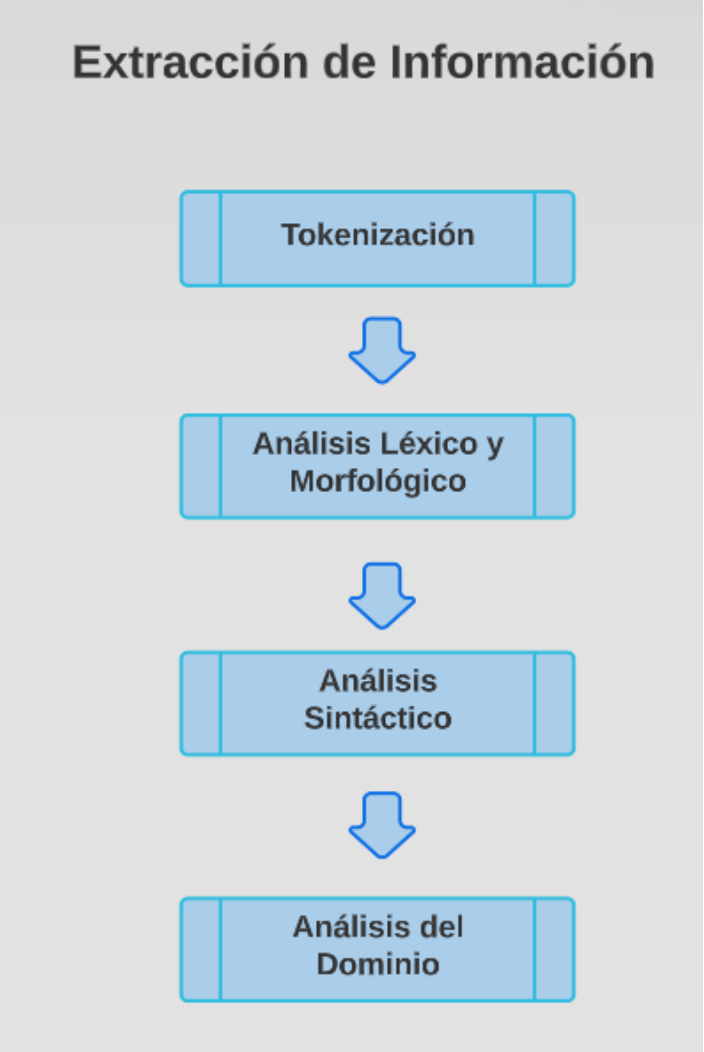


Figura 1.5: Extracción de Información en los Documentos
Elaborado por: Guerra Cleopatra

aprender patrones y relaciones en los datos. El aprendizaje de máquina se basa en la idea de que las computadoras pueden aprender y utilizar esta información para realizar predicciones o tomar decisiones. El aprendizaje de máquina es empleado en una gran variedad de aplicaciones como: la clasificación de imágenes, reconocimiento de voz, detección de fraudes, reconocimiento de productos, visión por computadora, análisis del buró crediticio, entre otros [16].

El aprendizaje de máquina forma parte del aprendizaje automático; que como su nombre lo indica, este tipo de aprendizaje se enfoca en implementar algoritmos y modelos que mejoran su rendimiento automáticamente [17]. Tanto el aprendizaje automático como el aprendizaje de máquina son términos usados de manera indistinta al tratarse de sistemas de computación que aprenden de los datos.

Clasificación del Aprendizaje Automático

El aprendizaje automático se clasifica en aprendizaje supervisado, no supervisado y por refuerzo. A continuación se describe cada uno de ellos.

- **Aprendizaje Supervisado:** Es una técnica en la cual los algoritmos aprenden a mapear las entradas a las salidas correctas, con el objetivo de realizar predicciones precisas sobre nuevas entradas. El algoritmo aprende a partir de ejemplos que son etiquetados y emplea esta información para realizar las nuevas predicciones sobre los nuevos datos [12] [18][19]. Este tipo de aprendizaje es usado para la detección de correo electrónico como spam o no spam y el fraude en las tarjetas de crédito.
- **Aprendizaje No Supervisado:** Es una técnica en la que el algoritmo encuentra patrones y estructuras en los datos, sin que estos se encuentren etiquetados. El algoritmo aprende sin tener una respuesta previa [20][19]. El objetivo de estos algoritmos es encontrar patrones en los datos y agruparlos en categorías [12]. Un ejemplo de este tipo de aprendizaje es la detección de anomalías en los sistemas o la segmentación de clientes.
- **Aprendizaje por Refuerzo:** En este tipo de aprendizaje, existe un agente y un entorno, el agente toma las decisiones en un entorno determinado. El agente recibe una recompensa o un castigo por cada acción que realiza. El agente aprende con prueba y error y ajusta su comportamiento en base a las recompensas y castigos que recibe [12][21][19]. Este aprendizaje es usado en el desarrollo de videojuegos como el ajedrez en el que compite humano - máquina o el control de robots.

Por otro lado, existen otros tipos de aprendizaje automático como:

- **Aprendizaje Semi-Supervisado:** Este tipo de técnica es una mezcla del aprendizaje supervisado con el no supervisado, en el cual los algoritmos trabajan con datos etiquetados y no etiquetados [22].
- **Aprendizaje por Transferencia:** Es una técnica en la que el algoritmo transfiere el conocimiento adquirido en una tarea para mejorar el desempeño de otra tarea relacionada [23]. El objetivo es emplear el conocimiento previo del modelo para mejorar la eficiencia y la precisión del aprendizaje en la nueva tarea [12]. Este aprendizaje es empleado en aplicaciones como reconocimiento de imágenes, procesamiento de lenguaje natural y robótica.

- **Aprendizaje Profundo:** Este tipo de aprendizaje emplea redes neuronales artificiales con múltiples capas para aprender y extraer características complejas de los datos de entrada. Utiliza técnicas de aprendizaje no supervisado para aprender patrones de los datos de entrada sin la necesidad de etiquetas explicitas [12][24]. El aprendizaje profundo se emplea para reconocimiento de voz, imágenes, etc.

La clasificación mencionada en esta sección es la más común. Es importante tener en cuenta que hay otros tipos de aprendizaje que no se encuentran descritos, sin embargo, tienen relevancia, ya que son aplicados para resolver problemas específicos.

Algoritmos de Aprendizaje Automático

Los algoritmos en el aprendizaje automático juegan un papel importante, debido a que gracias a la lógica de estos algoritmos las computadoras aprenden a partir de datos y mejoran el desempeño en cientos de tareas para las cuales están programados. La característica principal de los algoritmos es su capacidad de descubrir patrones en los datos y realizar tareas complejas de manera eficiente y precisa, tareas que para los humanos en ciertos escenarios resultan difíciles de detectar y resolver. En esta sección se describen los algoritmos que serán empleados en el desarrollo del proyecto de investigación. Los algoritmos elegidos están basados en el análisis realizado en la fase de revisión sistemática de la literatura. Además, en la Figura 1.6 se muestra ciertos algoritmos como una clasificación general.

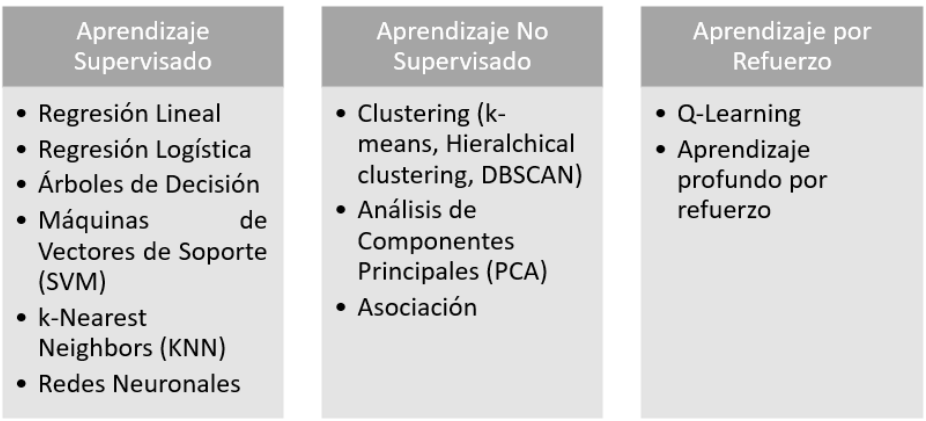


Figura 1.6: Clasificación Algoritmos de Aprendizaje Automático
Elaborado por: Guerra Cleopatra

1.2.4.0.1 Árboles de Decisión

Los árboles de decisión son empleados en clasificación. En esta técnica se construye un árbol en el que cada nodo representa una característica del conjunto de datos y cada rama representa una posible respuesta a la característica. Un ejemplo de este tipo de algoritmos es cómo con su uso se puede mejorar la predicción del riesgo de padecer una enfermedad cardiovascular en pacientes que tienen diabetes [25].

El funcionamiento de este algoritmo es a partir de un conjunto de datos de entrenamiento, estos datos están formados por objetos, atributos y una clase. Un árbol de decisión se forma de manera recursiva dividiendo el conjunto de datos en subconjuntos más pequeños, empleando un atributo como un criterio para la división en los nodos del árbol. A medida que el algoritmo avanza en su ejecución, este busca el atributo que mejor separa los objetos en subconjuntos que sean homogéneos en función de su clase.

Para determinar la homogeneidad dependerá del tipo de problema que se desea resolver, por ejemplo, para clasificación binaria se emplea la entropía de Shannon.

Continuando con el proceso del algoritmo, al seleccionar el atributo se crea el nodo del árbol. Este nodo es la representación de la decisión de dividir el conjunto de datos en los subconjuntos en función de un determinado atributo. Aquellos objetos que cumplan con la condición del nodo se colocan en el subconjunto, sin embargo, aquellos objetos que no cumplan con la condición son colocados en otro subconjunto. Este proceso se repite de manera recursiva hasta que se alcanza un criterio de parada que puede ser la profundidad máxima del árbol o un tamaño mínimo de un subconjunto.

Cuando se termina de construir el árbol, este puede ser empleado para clasificar nuevos objetos siguiendo el camino de la raíz hasta una hoja en donde se encuentra la clase que es asignada a un objeto [26].

A continuación se presenta un ejemplo desde la formación de la raíz y los nodos en un árbol de decisión. En el problema se identifican tres variables: temperatura, humedad y velocidad del viento. Ahora, si el día es soleado, la temperatura es alta, la humedad es baja y la velocidad del viento es moderada, entonces es probable que en este día se pueda salir a pasear. Si el día es nublado, la temperatura es baja, la humedad es alta y la velocidad del viento es alta, entonces es menos probable que en este día se pueda salir a pasear.

El árbol de decisión de este problema se muestra en la Figura 1.7 [26]. Por otro lado, una manera de describir el árbol de decisión sería:

- Si el día es soleado - Si la temperatura es alta - Si la humedad es baja - Si la velocidad del viento es moderada - Se puede salir a pasear.
- Si la velocidad del viento es alta - No se puede salir a pasear.
- Si la humedad es alta - No se puede salir a pasear.
- Si la temperatura es baja - No se puede salir a pasear.
- Si el día es nublado - No se puede salir a pasear.

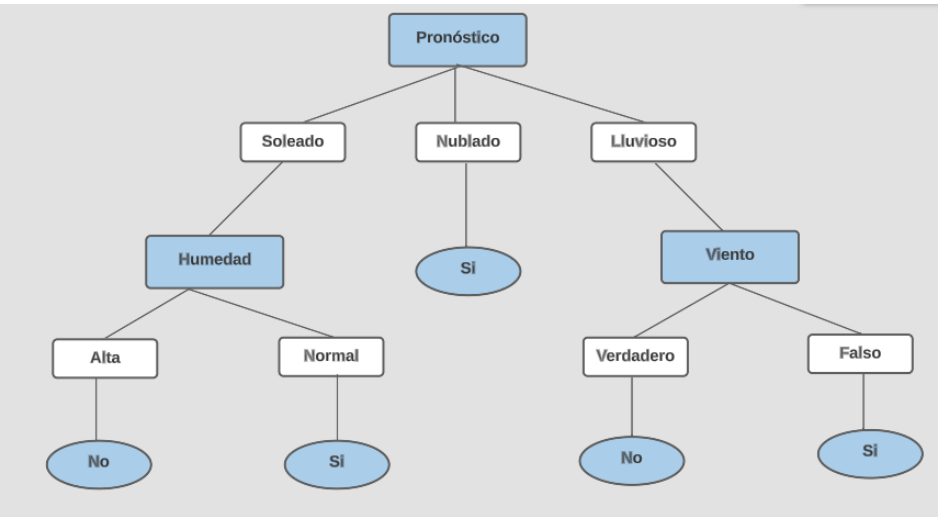


Figura 1.7: Ejemplo algoritmo árbol de decisión
Elaborado por: [26].

1.2.4.0.2 Máquinas de Vector de Soporte

Este tipo de algoritmo es comúnmente conocido como (*Support Vector Machine*), y es empleado para clasificación y regresión. El objetivo de este algoritmo es encontrar un hiperplano que separa dos clases diferentes de datos de manera óptima. Un hiperplano se puede entender como una línea que separa los puntos de manera eficiente en dos clases [27]. La mejor separación es encontrar el margen más amplio entre las dos clases.

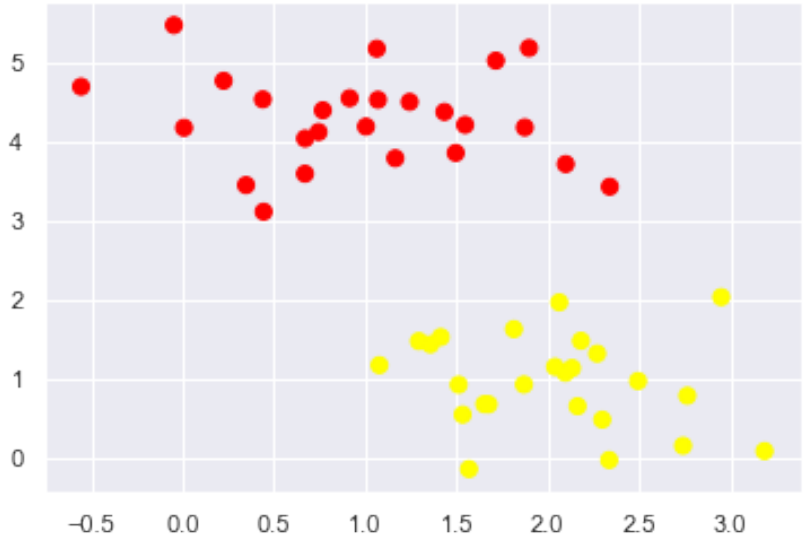
A continuación, en la Figura 1.8a, se observa dos grupos de objetos, de color amarillo y rojo, que corresponden a dos clases. *Support Vector Machine* dibuja una línea recta que separa las dos clases, sin embargo, en este espacio se dibujan algunas líneas, como se muestra en la Figura 1.8b. Ahora, cada línea se puede ampliar por ambos lados hasta que

toque el punto más cercano, de cualquiera de las dos clases iniciales, como se observa en la Figura 1.8c. La principal línea que se toma en cuenta, es aquella que tiene el mayor ancho o margen, esto se muestra en la Figura 1.9.

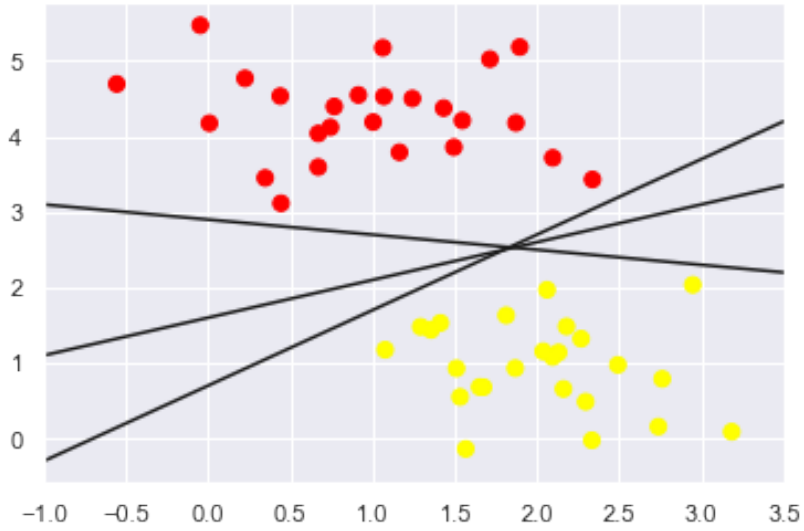
Support Vector Machine es empleado en múltiples disciplinas como: reconocimiento facial, bioinformática, extracción de conceptos de minería de texto, reconocimiento de voz [28], detección de fraudes bancarios [29], entre otros.

Por otro lado, *Support Vector Machine* trabaja con más de dos clases. Para este caso, el algoritmo empleará un kernel, que es una función matemática que da una mayor dimensión a los datos de entrada. El kernel puede ser: radial, sigmoidal, polinómico y lineal [12].

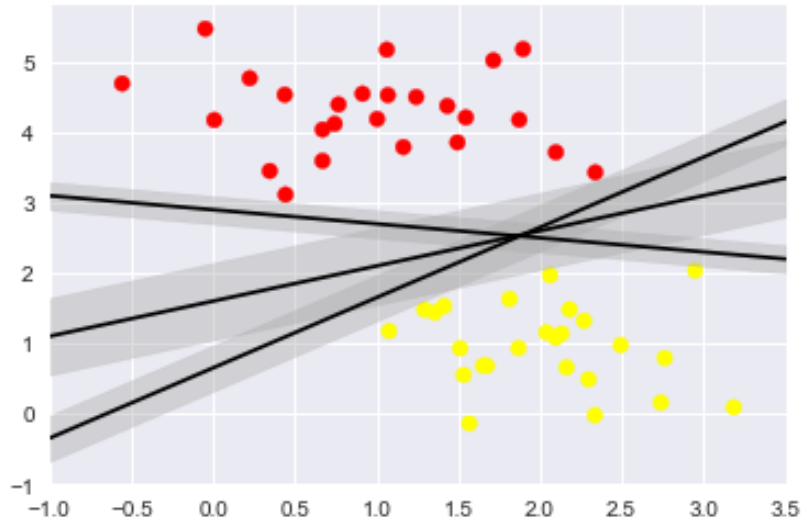
El kernel es una función matemática que se utiliza para transformar los datos de entrada en un espacio de mayor dimensión. El objetivo de esta transformación es encontrar un hiperplano que pueda separar los datos de entrada en diferentes clases. El kernel se utiliza para calcular el producto escalar entre dos vectores en el espacio de mayor dimensión sin tener que calcular explícitamente la transformación de los datos. La elección del kernel depende del tipo de datos y del problema de clasificación específico que se esté abordando. Presenta los documentos en forma de vector con la frecuencia de los términos existentes en el documento, además, busca en un conjunto de datos de entrenamiento los t puntos de datos más cercanos para asignarlos a un nuevo punto de datos.
Para encontrar los puntos más cercanos se emplea una métrica de distancia, esta métrica de distancia ayuda a la determinación de la clase que contiene la mayoría de los puntos cercanos para luego asignar a la nueva instancia. Además, emplea la proximidad para realizar la clasificación. Esta clasificación parte de la suposición de que existen puntos similares que se pueden encontrar cerca uno del otro.
Un ejemplo de su funcionamiento, es al trabajar con texto, una vez que se genera el vector de características el algoritmo busca todos los ejemplos de entrenamiento para realizar la comparación de la similitud que existe en los vectores de características [30].
A continuación, encuentra los t ejemplos de entrenamiento que tengan mayor proximidad y el documento desconocido es asignado a los t vecinos más cercanos a un objeto y emplea la clase de esos ejemplos para clasificar los objetos desconocidos. Un punto importante a considerar en este tipo de algoritmos es la elección del valor t, si este valor es demasiado pequeño la clasificación puede resultar inexacta, por otro lado un valor de t alto puede llevar a una clasificación errónea [12].



(a) Grupo de datos, clasificados en dos clases **[27]**



(b) Identificación de Hiperplanos **[27]**



(c) Identificación del Margen en cada Hiperplano **[27]**

Figura 1.8: Funcionamiento Algoritmo SVM

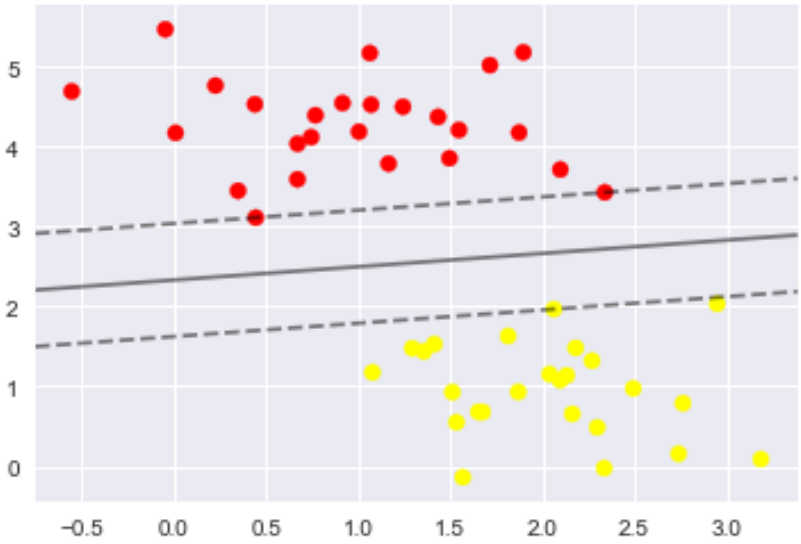


Figura 1.9: Identificación del Hiperplano con Mejor Margen **[27]**

1.2.4.0.3 Naïve Bayes

Este algoritmo basa su funcionamiento en el teorema de Bayes. Como característica, el algoritmo asume que las entradas son independientes ente sí, es decir que la ausencia de alguna característica no afecta la presencia de otra característica durante el análisis.
El algoritmo emplea además la probabilidad condicional para calcular la probabilidad que de una instancia pertenezca a cierta clase. Esta probabilidad es calculada con la fórmula de Bayes 1.1, donde:

$$P(clase|caractersticas) = \frac{P(caractersticas|clase) \cdot P(clase)}{P(caractersticas)} \tag{1.1}$$

1. P(clase | características): Es la probabilidad de que la instancia pertenezca a una clase determinada.
2. P(características | clase): Es la probabilidad de que las características se observan en una instancia de la clase dada.
3. P(clase): Es la probabilidad a priori de la clase.
4. P(características): Es la probabilidad de que se observen las características.

El algoritmo de Naïve Bayes emplea esta ecuación con el objetivo de obtener la probabilidad de que una instancia pertenezca a cada clase, para la asignación de la instancia a la clase se emplea la probabilidad más alta. Además, para el cálculo de las probabilidades

condicionales se usa la frecuencia de las características. El algoritmo se entrena utilizando un conjunto de datos etiquetados y utiliza la frecuencia de las características en cada clase para calcular las probabilidades condicionales. Usualmente este algoritmo es empleado en la clasificación de correos de tipo spam, detección de sentimientos y en el análisis de texto.

1.2.4.0.4 k-means

El algoritmo k-means agrupa objetos en diferentes categorías o clústers. Para realizar la agrupación se minimiza la suma de las distancias entre cada objeto y los centroides de cada grupo. Cada grupo tiene un centroide que corresponde a la media artimética de los puntos asignados al grupo. Su funcionamiento empieza con la selección de k número de grupos, después, se establecen los centroides, luego se asigna los objetos a los centroides más cercanos para que la posición del centroide de cada grupo se actualice tomando como nuevo centroide la posición del promedio de los objetos que pertenecen a un grupo. Este proceso se repite hasta que que no existan cambios durante la asignación de objetos a los clústers [31].
El funcionamiento del algoritmo se describe a continuación.

1. Seleccionar k centroides aleatorios del conjunto de datos.
2. Asignar cada punto de datos al centroide más cercano.
3. Recalcular los centroides como la media de los puntos de datos asignados a cada centroide.
4. Repetir los pasos 2 y 3 hasta que los centroides no cambien o se alcance un número máximo de iteraciones.

El rendimiento del algoritmo depende de la elección del número de centroides y el ruido que pueda existir en los datos que se emplean. Para la elección del número adecuado de centroides se puede emplear el método del codo, comúnmente conocido como *Elbow Method*. Este método consiste en trazar la suma de las distancias al cuadrado de cada punto de datos al centroide más cercano, en la gráfica que se forma en este método se elige el número de centroides en el lugar en donde la curva empieza a aplanarse. Sin embargo, existe otro método conocido como validación cruzada (*cross-validation*), en este método se ajusta el modelo con diferentes números de centroides, se realizan las evaluaciones hasta encontrar el mejor rendimiento [31].
Este algoritmo es empleado en la segmentación de imágenes, detección de anomalías,

agrupación de documentos, entre otros. Un inconveniente que puede presentar el algoritmo es la selección inicial del número de centroides y clústers.

1.2.5 Sesgo

El sesgo se entiende como una tendencia o inclinación hacia un grupo de ideas, opiniones o perspectivas que pueden influir en la toma de decisiones o en algunos casos en la interpretación de la información [32][33]. El sesgo está presente desde circunstancias simples, hasta más complejas como una búsqueda en la web, en los algoritmos que clasifican y presentan información. La presencia de sesgo puede ser causa de ciertos factores como: interacción del usuario, la personalización algorítmica que hace referencia al uso de algoritmos y sistemas que adaptan la presentación de la información al usuario. Estos sistemas emplean los historiales de navegación o clics. Sin embargo, esta personalización algorítmica puede mostrar información con la propia creencia de cada usuario, pero se limita a explorar nuevas ideas. En cuanto a los sistemas de optimización, se puede crear una rivalidad interna en los sistemas cuando un sistema da prioridad a la presentación de contenido que se genera al dar clics o ingresos publicitarios en lugar de presentar información objetiva [5].

El sesgo tiende a ser negativo ya que puede brindar al usuario información errónea o engañosa. Un sesgo negativo da como resultado una degradación en la calidad de las decisiones. El sesgo puede también implicar discriminación y desigualdad debido a que los sistemas pueden marginar a ciertos grupos de personas. Un fenómeno común en la información sesgada es el filtro de burbuja, que hace referencia a cuando un usuario solo obtiene información que se ajusta a sus intereses y preferencias, pero no se expone a otras perspectivas [5].

Ahora bien, en el escenario de las ciencias de la computación, el sesgo se produce cuando un determinado modelo no tiene en cuenta toda la información que tiene el dataset. En general, en los sistemas que existe algún tipo de discriminación existe sesgo [32]. El sesgo puede ser:

1. ***Sesgo preexistente:*** Es aquel que existe de forma independiente antes de la creación del sistema.
2. ***Sesgo Individual:*** Un programador o un cliente que tienen un aporte significativo en el sistema.

3. ***Sesgo social:*** En términos de género.

4. ***Sesgo Técnico:*** Se origina por una limitación de la tecnología informática sea en hardware o en software.

5. ***Sesgo en Algoritmos:*** Cuando el algoritmo no trata a todos los grupos de manera equitativa bajo las mismas condiciones.

6. ***Sesgo emergente:*** Surge después de completar un diseño sea por nuevos conocimientos incorporados, experiencias diferentes o valores diferentes en los usuarios.

El sesgo es muy frecuente en los SR, su presencia influye en las recomendaciones que reciben los usuarios. Es perjudicial desde la perspectiva de equidad al incluir y excluir ciertas recomendaciones de grupos de usuarios. La gran interrogante es cómo combatirlo, para esto se han estudiado algunas medidas, siendo la más simple que tanto usuarios y desarrolladores estén conscientes de su presencia. Se debe, en consecuencia promover la educación digital para que los usuarios tengan conocimiento de cómo funcionan los algoritmos y cómo su funcionamiento afecta la presentación de la información. Otra forma es la diversificación de fuentes, es decir buscar información de varias fuentes e incorporar esta diversidad en los algoritmos para evitar la exclusión o marginación.

1.2.6 MongoDB

MongoDB es un gestor de Base de Datos NoSQL lanzado en el año 2009. Es flexible, escalable y tiene la capacidad de manejar grandes volúmenes de datos. Comúnmente esta orientado a trabajar con documentos [34], su característica principal es que es de código abierto.

Otra característica es que almacena la información. Esta información es una estructura de datos conocida como BSON (*Binary JSON*). Además, guarda registros que son integrados en colecciones equivalentes a tablas SQL (*Structured Query Language*). La estructura de los JSON se manejan por una clave y un valor. Algo importante de esta base de datos es que sus colecciones no necesariamente deben tener los mismos campos, estructuras o tipo de datos [35].

MongoDB es una base de datos orientada a documentos. Esto quiere decir que en lugar de guardar los datos en registros, guarda los datos en documentos, como se muestra en la

Figura 1.10.

Las principales características de esta base de datos son:

```
{
   Nombre: "Miguel",
   Apellidos: "Parada",
   Edad: 39,
   Aficiones: ["Música","Ciclismo","Baloncesto"],
   Amigos: [
       {
         Nombre:"Marie",
         Edad:35    },
       {
         Nombre:"Elsa",
         Edad:42
       }
   ]
}
```

Figura 1.10: Ejemplo Información Almacenada en Mongo

- **Consultas ad hoc:** Se puede realizar varios tipos de consulta, sea por campos, consulta de rangos, expresiones regulares e incluso funciones JavaScript.
- **Idexación:** Cualquier campo puede ser indexado, así como también se puede añadir múltiples índices secundarios.
- **Replicación:** Soporta el tipo de replicación primario-secundario. De este modo, mientras se realizan consultas con el primario, el secundario actúa como réplica de datos en solo lectura a modo copia de seguridad con la particularidad de que los nodos secundarios tienen la habilidad de poder elegir un nuevo primario en caso de que el primario actual deje de responder.
- **Balanceo de carga:** Tiene la capacidad de ejecutarse de manera simultánea en múltiples servidores. Ofrece un balanceo de carga o servicio de replicación de datos en caso de un fallo del hardware.
- **Almacenamiento de archivos:** Es utilizado también como un sistema de archivos que permite manipular archivos y contenido.
- **Ejecución de JavaScript del lado del servidor:** Se realiza consultas utilizando JavaScript, haciendo que estas sean enviadas directamente a la base de datos para ser ejecutadas.

1.3 TRABAJOS RELACIONADOS

El uso de las diferentes plataformas virtuales es parte de los hábitos de los usuarios. Actualmente, redes sociales como Facebook, Twitter, Instagram son la opción perfecta para que miles de usuarios en el mundo compartan contenido y al mismo tiempo se genera una gran base de datos con la información suministrada por los usuarios. En las Redes Sociales en línea, se genera una gran cantidad de sesgo, cuya medición en los SR es difícil e implica un verdadero reto empírico, desde el diseño del algoritmo hasta el aprendizaje que el Sistema de Recomendación dispone [36].

En la literatura se han publicado valiosos aportes en esta materia. Tal es el ejemplo del estudio de la presencia de sesgo en algunos canales de Youtube. En [37], el estudio se realiza en base a términos de análisis de texto, por los comentarios que colocan los usuarios de estos canales. Este trabajo se analiza desde el punto de la violencia y discriminación detectada en los comentarios de una base de datos con más de 7 000 videos y 17 millones de comentarios. Los autores abordan el problema mediante un análisis léxico que emplea los campos semánticos de las palabras, un análisis de tópicos y el sesgo que se encuentra implícito.

Un trabajo similar es el de Deena Abul-Fottouh [38], quién encuentra que la discriminación en los vídeos provocan un incremento en el sesgo, generando un efecto burbuja. Este efecto se disminuye en los videos relacionados a temas de salud. Sin embargo, para Bonchi [39], la presencia de sesgo no se encuentra sólo a partir del texto, sino que también esta presente a nivel individual o grupal, sea por género, grupo étnico, por rango de edad o por orientación sexual dentro de la web. Además, existen casos en los que el sesgo está presente implícitamente a partir del diseño que surge de la interpretación de los desarrolladores de software; de cómo ellos entrenan a los algoritmos [39].

Otros trabajos aluden la presencia de sesgo en el reconocimiento facial [40]. La manera de detectar sesgo al identificar rostros se presenta usualmente al distinguir género, rasgos étnicos, o color de piel. Existen bases de datos que tienen errores sistemáticos en la detección de rostros, lo que inevitablemente introduce sesgo. En estas bases de datos apenas el 10 % corresponde al género masculino africano y el 17 % del género africano femenino, esto genera sesgo por la falta de imágenes de personas de etnia afro-descendiente. Pero, no sólo se tiene una discriminación por raza sino por género. La comunidad LGBTQ, son otro segmento de la población en el cual los resultados del reconocimiento facial arrojados

no suelen ser satisfactorios.

La metodología adoptada en [40], se basó en, crear una nueva base de datos que contenga más imágenes de la comunidad LGBTQ y empezar a realizar el entrenamiento del algoritmo para obtener resultados asertivos y disminuir el sesgo.

Una contribución interesante fue el desarrollado por Joanna Misztal [31], en este trabajo se propone un algoritmo de clustering jerárquico consciente de la presencia de sesgo que detecta aquellos grupos de usuarios potencialmente discriminados en los SR. En este estudio se aplica un modelo post-hoc, que se refiere a un modelo que se emplea después de que se haya entrenado el modelo principal. En los SR, esta aproximación permite explicar el por qué se hacen las recomendaciones a los usuarios. Además, los modelos post-hoc son empleados para mejorar la transparencia y la interpretabilidad de los SR. Esta propuesta, tiene un enfoque que detecta grupos de usuarios discriminados en algoritmos de recomendación basados en sus métricas de rendimiento.

Sin embargo, en todos estos trabajos la pregunta que surge comúnmente es: ¿Cómo mitigar su presencia de sesgo en un nivel práctico?; la respuesta a esta interrogante resulta subjetiva a la metodología empleada en diversas investigaciones. Si bien la eliminación completa del sesgo puede ser un objetivo poco realista, llamar la atención sobre su existencia no sólo puede advertir a los usuarios que el contenido al que se accede está sesgado, sino que también permite que los usuarios evalúen el trabajo de manera objetiva. Esto debido a que los usuarios tienen un escaso conocimiento de cómo los algoritmos en los SR influyen en la información que reciben.

Partiendo de este análisis, en la investigación realizada se adoptó una metodología propia. La recolección minuciosa de *tweets* en Ecuador proporcionó un corpus valioso y estableció las bases para un enfoque integrador que fusiona técnicas de *Text Mining* y algoritmos de aprendizaje automático. Los enfoques previos dependían de datos etiquetados. Esta investigación aborda la falta de etiquetas en el corpus mediante la implementación de técnicas de aprendizaje no supervisado para luego aplicar algoritmos de aprendizaje supervisado y evaluar el comportamiento de cada algoritmo analizado. Se creó un modelo que emplea diccionarios de palabras. Con la implementación del nuevo diccionario se obtuvo una mayor cantidad de *tweets* etiquetados, enriqueciendo el corpus inicial recolectado de acuerdo al tópico estudiado.

2 METODOLOGÍA

En el desarrollo del presente trabajo, se empleará dos metodologías. Por un lado, se utilizará la metodología de investigación basada en la Ciencia del Diseño (*Desing Science Research*, DSRM), y la metodología CRISP-DM (*Cross Industry Process for Data Mining*). DSRM está enfocada en la resolución de problemas mediante el desarrollo de artefactos para de esta manera lograr el entendimiento y la solución de un problema [41]. Se fundamenta en que tanto el diseño de la ciencia como los sistemas de información deben ir de la mano. Sin embargo, lograr una verdadera apreciación de la ciencia del diseño como una investigación en los sistemas de información es un paradigma, puesto que se enfrenta a una dicotomía. La dicotomía relaciona tanto el diseño como un proceso (conjunto de actividades) y el producto (artefacto) [41]. Emplear este paradigma, guiará el proceso desde la fase de definición del problema hasta la producción del artefacto [42].

Combinar DSRM con técnicas de minería de datos, permitirá mejorar paulatinamente los artefactos antes de obtener la solución final. DSRM ayudará a abstraer y comprender situaciones del mundo real. Para el caso en estudio la situación será el sesgo que se genera en los SR de los usuarios que emplean Twitter. De esta manera, se formará una base de datos con la información recolectada. El sesgo se analizará por medio del tópico de política, gracias a su trascendencia Nacional. Por otro lado, el artefacto será el modelo que permitirá entender, verificar y disminuir el grado de la presencia de sesgo en las recomendaciones que la red social haga a los usuarios.

Al trabajar con los datos recolectados de Twitter, se aborda un proceso de minería de datos, por tal motivo se empleará CRISP-DM. Esta metodología está formada por seis fases que representan el proceso de un proyecto de minería de datos [43]. CRISP-DM aborda dos ejes importantes, por un lado, el concepto de metodología que incluye descripciones de las fases del proyecto, las tareas necesarias y una explicación en cada fase. Además, proporciona un modelo de proceso que ofrece un resumen del ciclo de vida de los datos en la minería [44]. En la Figura 2.1, se detalla cada fase.

Tabla 2.1: Relación metodología CRISP-DM y DSRM

Metodología DSRM		Metodología CRISP-DM
Identificación del Problema y Motivación Definir los Objetivos para la solución	➡	Entendimiento del negocio Entendimiento de los datos
Diseño, desarrollo y demostración	➡	Preparación y Modelamiento de los datos
Evaluación	➡	Evaluación
		Implementación

Las dos metodologías empleadas no son excluyentes, por tal motivo, se pueden unir, con

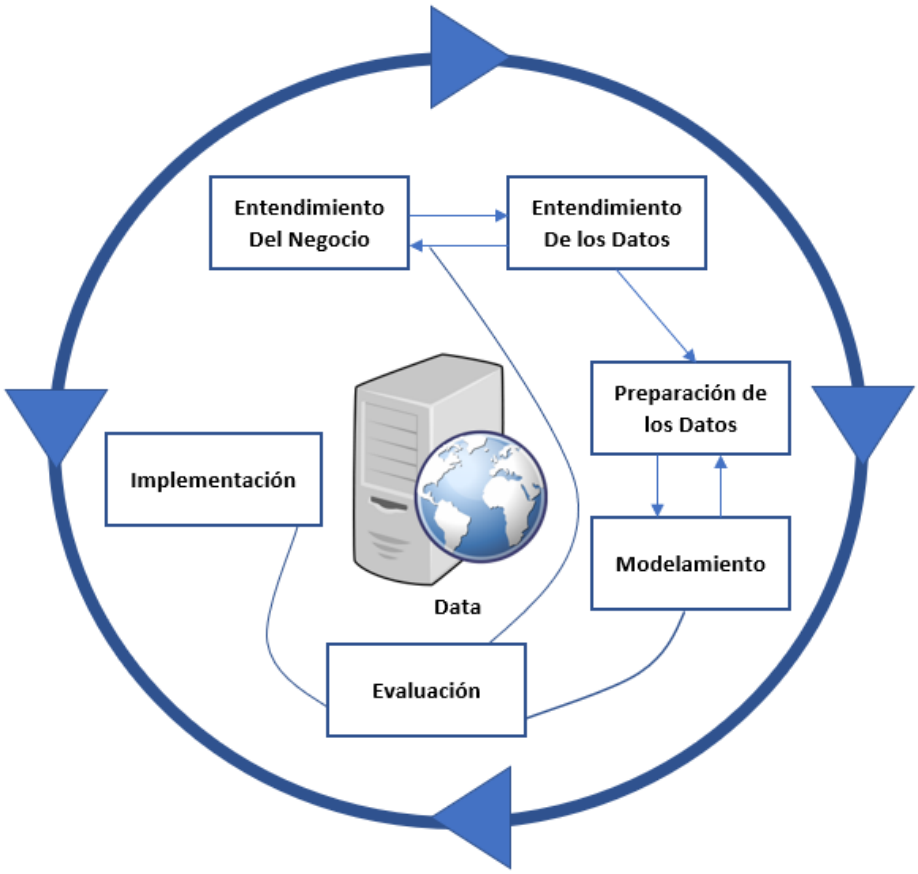


Figura 2.1: Ciclo de vida metodología CRISP-DM
Elaborado por: Guerra Cleopatra

el objetivo de que aquellos procesos comunes sólo se lleven a cabo una vez. Al utilizarse de manera simultánea, la relación existente entre las dos metodologías se puede apreciar en la Tabla 2.1. Cada una de las fases, se detallan a continuación:

- ***Fase 1 - Entendimiento del Negocio:*** Es la fase más importante puesto que aquí se comprende el proyecto como tal, sus objetivos y requisitos. Se evalúa la situación, cuáles son los factores de riesgo involucrados en el proyecto o si existe algún plan para identificar los riesgos del proyecto y conocer cómo solventar estos inconvenientes [45].

- ***Fase 2 - Entendimiento de Datos:*** Esta fase está relacionada con la recolección y preparación de los datos. El objetivo de esta fase es obtener un primer contacto con el problema de estudio. Los datos empleados pueden ser existentes, adquiridos o adicionales. En esta fase, además, se debe tener presente la recolección inicial de datos, la descripción, exploración y la verificación de la calidad de los datos [46].
- ***Fase 3 - Preparación de Datos:*** La fase consiste en selección, limpieza de los datos, construcción, integración y formateo de los datos. Además se puede crear nuevos registros o fusionar campos [47].
- ***Fase 4 - Modelamiento:*** Esta fase consiste en la selección de técnicas de modelado elegidas en función de criterios apropiados para el problema que se persigue resolver. El modelado incluye elegir la técnica de modelado, por ejemplo, árboles de decisión, K – nearest neighbors, etc., según sea el caso [48].
- ***Fase 5 - Evaluación:*** En esta fase se evalúa el modelo tomando en cuenta los objetivos del proyecto, determinando además si el modelo es eficiente para probarlo en el problema real. También, se determina los próximos pasos, es decir, aquí se decide si se avanza a la última fase o si se retorna a las fases posteriores [49].
- ***Fase 6 - Implementación:*** Es la fase final del proyecto, en la cual la solución obtenida se transforma en un producto de producción útil para los usuarios finales. Esta fase debe ser documentada con los resultados obtenidos, la experiencia adquirida y colocar los puntos relevantes que se presentan durante la ejecución del proyecto [50].

2.1 FASE 1 : ENTENDIMIENTO DEL NEGOCIO

La importancia de emplear Twitter como red social se debe a que diariamente miles de datos son compartidos alrededor del mundo. Sin embargo, esta realidad es muy ajena en Ecuador. De aquí la necesidad de realizar un análisis del comportamiento de la población ecuatoriana en base a un tópico específico. El tópico analizado en el trabajo de investigación es el tema político.

El propósito de esta investigación es encontrar una manera de identificar el sesgo a través del análisis de datos recolectados y proponer una metodología que en cierto grado mitigue la presencia del sesgo en la información recolectada.

Al plantear el tema político, el objetivo es encontrar la manera de obtener etiquetas adecuadas con las que se identificará cada *tweet* como político o no político. El tema político es

amplio y en él se puede encontrar sesgo de opinión. El desafío del presente proyecto es mitigar el sesgo, quizá con la inclusión de nuevas características políticas o probar vectores de palabras que permitan realizar una clasificación más justa en el texto.

Una vez establecido el tópico y la fuente que se empleará para obtener la información de la población ecuatoriana, el siguiente paso es analizar la literatura. El objetivo de este análisis es plantear las preguntas de investigación que se detallan a continuación.

2.1.1 Preguntas de Investigación

En esta parte se plantea las preguntas de investigación. Estas preguntas inicialmente serán contestadas en base a la revisión de literatura que se realizó como punto de partida en el presente trabajo de investigación. La revisión de la literatura fue sistemática, parte del planteamiento del objetivo de la investigación y de la definición de las preguntas de investigación. A continuación se estableció un protocolo que incluyó una estrategia requerida para la selección de los artículos [51]. En la búsqueda de artículos se incluyó cadenas con operadores lógicos. Además de usar criterios de inclusión y exclusión con el propósito de obtener artículos que contengan información relevante. Entre las librerías científicas se empleó: Scopus, *Association for Computing Machinery* (ACM), Springer, *Institute of Electrical and Electronics Engineers* (IEEE), entre otras.
A continuación se muestra un ejemplo de una cadena de búsqueda empleada en Scopus:

classification AND algorithm AND improve AND learning AND in AND recommendation AND systems AND (LIMIT-TO (OA , .ªll")) AND (LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017)) AND (LIMIT-TO (DOCTYPE , .ªr") OR LIMIT-TO (DOCTYPE , çp") OR LIMIT-TO (DOCTYPE , re")) AND (LIMIT-TO (SUBJAREA , ÇOMP") OR LIMIT-TO (SUBJAREA, .ENGI")) AND (LIMIT-TO (LANGUAGE , .English"))

Algunos criterios de inclusión fueron:

- Artículos en inglés y español.
- Artículos que muestre una metodología, comparación con algunos algoritmos de clasificación.
- Artículos de Sistemas de Recomendación que empleen la API de Twitter.
- Resultados publicados en revistas y congresos científicos, a partir del año 2016.

- Artículos que sean Open Access.

Por otro lado, los criterios de exclusión fueron:

- Artículos que están no escritos en inglés y español.
- Artículos que tienen años de publicación menores al 2016.
- Artículos que no tengan una metodología.
- Resultados publicados en revistas y congresos científicos, a partir del año 2017.
- Artículos que sólo hablen de Sistemas de Recomendación y no mencionen Bias.

Luego, en una fase posterior, se pretende dar respuesta a cada una de ellas en base al proyecto de investigación efectuado.

❒ **RQ1: ¿Cuál es el tratamiento de los datos recolectados?**

El tratamiento de los datos recolectados empieza con la elección de los datos. Esta elección implica realizar un análisis con el objetivo de proponer sobre qué datos se plantea la investigación. A continuación, se lleva a cabo la recolección de datos. Luego la limpieza de los datos que implica realizar un análisis del léxico, de la semántica, se eliminan símbolos innecesarios, se procede a tokenizar los datos y quitar las palabras conocidas como *stop-words* [52]-[58].
En ocasiones se aplica *word embeddings* y algoritmos LDA (Latent Dirichlet Allocation), [59], [60], [37], [61]; puede ser el caso que se use la librería Selenium de Python para realizar scraping [36].
Existen trabajos que emplean un *Dataset* sintético creado a partir del análisis semantico-lexico realizado en pasos posteriores; esto con el objetivo de contar con datos que serán empleados en la fase de entrenamiento de los algoritmos [31]. Al mencionar datos sintéticos hace referencia a aquellos datos que puede estar formada por diccionarios con palabras tóxicas o palabras clasificadas como positivas - negativas. Por otro lado, hay trabajos que usan un *dataset* creado previamente [62], [40], [63], [64].
Una vez que todos los datos están limpios, se obtiene el corpus final sobre el cual se aplicarán los algoritmos de aprendizaje automático.

Es importante mencionar la fuente de la cual se obtuvieron los datos para ser analizados. Algunos ejemplos de *datasets* empleados en el estado del arte fueron:

- *InfoWars*, [65]

- Twitter
- *Google News*, [66]
- Facebook
- *BBC forum*, [67]
- Digg, [68]
- YouTube
- Artículos de Wikipedia
- Color FERET es una base de datos de 8.5 GB de imágenes. [69]
- *Labeled Faces in the Wild* (LFW), a base de datos con imágenes de celebridades. [70]
- Bases de Datos sintética (creadas por los autores)
- *MovieLens* 100K base de datos de recomendación de películas, [71]
- *Book-Crossing*, [72]
- El Guardián contiene artículos del periódico de Reino Unido, incluido cada artículo publicado en línea entre el 2009 y 2018. [73]
- *TREC Robust 04 test* base de datos de artículos nuevos [74]

❒ **RQ2: ¿Cuál es la metodología empleada en la detección del sesgo en los SR?**

Cada artículo cuenta con una metodología propia. La metodología empieza con la recolección de datos. A continuación, la limpieza, el procesamiento, la aplicación de algoritmos de aprendizaje supervisado, la obtención y el análisis de resultados. Cómo se aplica cada metodología depende del alcance y propósito de la investigación.
Para los autores Sara Hajian, Francesco Bonchi y Carlos Castillo [39], la metodología es:

- Pre-procesamiento mediante la transformación de los datos del origen.
- Proceso de integración de una anti-discriminación.
- Post procesamiento para modificar los resultados de los modelos de minería de datos.
- Se emplea una corrección de los datos en el entrenamiento.

Por otro lado, Raphael Ottoni, Evandro Cunha y Gabriel Magno [37] exponen:

- Limpieza de Datos.
- Identificación de sesgo implícito, con el uso de *Association Implicit Test* (IAT) y *Word Embeddind Association Test* (WEAT)

Armin Mertens y Franziska Pradel [60] emplean una metodología basada en dos pasos.

- Limpieza de Datos.
- Introducción medida de cuantificación de sesgo - uso de diccionarios LIWC.

La metodología propuesta por Giorgi Salvatore, Lynn Veronica, Matz Sandra, Ungar Lyle y Schwartz H. Andrew [56] consiste en:

- Estimación Socio-Demográfica.
- Creación de Factores de pesos.
- Aplicación de Factores de pesos.

Los autores Nathan Bartley y Andrés Abeliuk [36] aplican:

- Crear cuentas de bots que trabajan en pares. Los bots sólo observan los *tweets* no hacen acciones sobre estos. Empiezan a interactuar cuando con existen temas de interés explicito.
- Un bot trabaja con los *top tweets* y el otro con los *latest tweets*.

Por otro lado, Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp y Karsten Donnay [59] emplean:

- Crear diferentes bases de datos, por ejemplo, con el léxico de las palabras.
- Realizar manualmente listas de palabras que describan conceptos y contenidos de temas específicos, que serán usados como semilla. Estas palabras son las palabras sesgadas (se eligen por el mayor valor de coseno de similitud).

Hijazi, Mohd Hanafi Ahmad Libin, Lyndia Alfred, Rayner Coenen y Frans [53] emplean el método *Bias-Aware Thresholding* (BWT) para clasificar palabras positivas y negativas. En el trabajo de Karim y Asim se usa el método *Bias-Aware Thresholding* (BWT) [54], [53]. En este caso se obtiene una lista de palabras calificadas entre valores de menos a más; también para palabras positivas y negativas.

Hube [75], Christoph Fetahu y Besnik crean un diccionario con el léxico de las palabras para emplearlo en el entrenamiento del algoritmo. La creación de un nuevo conjunto de datos como parte de la metodología es esencial. En el trabajo de Wu Wenying, Michalatos y Panagiotis Protopapaps [40], se analiza imágenes de un grupo de personas con el propósito de identificar a personas LGTBI, para esto forman un nuevo conjunto de datos no binario. Estos datos se forman de palabras tóxicas que se identifican en el texto. Gracias a esta identificación se añade al corpus palabras no tóxicas por cada término tóxico durante el

entrenamiento [62]. Otro ejemplo, es el uso de una lista de términos rankeados [63]. O emplear un corpus formado de sólo términos femeninos y todo lo que no se identifica como femenino durante el entrenamiento se clasifica con un sesgo negativo [61].

En el trabajo propuesto por Fabris Alessandro, Purpura Alberto, Silvello Gianmaria y Susto Gian Antonio[55]:

- Por cada consulta se busca el número de palabras femeninas y masculinas.
- Se representa como un cociente para buscar en un radio de m/f.
- Se encuentra una diferencia del sistema con el analizado.
- Basado en esta diferencia se encuentra si es positiva la respuesta como masculino o negativa como femenino.

En el trabajo de Wang Zhu, Yu Zhiwen, Fan Renjie y Guo Bin [58] la metodología es:

- Tomar los datos de Twitter.
- Usar la teoría de Markov Chain.
- Hacer una muestra randmica con los datos proporcionados.
- Hacer una matriz de transición.
- Ajustar los pesos de la matriz.
- Hacer un re-muestreo.
- Muestras tomadas de acuerdo a la matriz de transición.

La metodología puede ser un poco más elaborada como en el trabajo de Misztal-Radecka Joanna y Indurkhy Bipin [31] en la que se detalla:

- Detectar los grupos que son discriminados dentro de la caja negra.
- Describir estos grupos en términos comprensibles para las personas.
- Usar de *bias-aware hirarchical K-Means*.
- Evaluar la detección de grupos discriminatorios para esto se usa el número de grupos con un valor de 1 a -1, donde 1 indica una mejor agrupación. Se emplea el índice de davies-bouldin que indica la similitud media entre cada clúster y su clúster más similar.
- Evaluar la predicción de los clústes, si la calidad de la predicción es baja y la calidad de los clústers es alta significa que existen grupos que no están bien descritos.
- El método es capaz de detectar grupos de usuarios potencialmente discriminados por un algoritmo determinado, se adopta un modelo universal con un enfoque agnóstico a

la métrica que puede emplearse en diferentes aplicaciones comerciales.

❒ **RQ3: ¿Cuáles son los algoritmos de aprendizaje supervisado que se emplean en el análisis del sesgo en los SR?**

De los documentos analizados, apenas 13 de ellos especifican de manera concisa los algoritmos empleados en el análisis de sesgo. Los algoritmos comúnmente usados son:

- Naïve Bayes (NB), [39], [53], [54], [56].
- *Logistic Regresion*, [39], [59], [40], [58].
- *Tree Decision* (TD), [39], [59]
- *Hinge Loss*, [39].
- *Support Vector Machine* (SVM), [39], [53], [59].
- *Adaptative Boosting Clasification*, [39].
- *Linear Discrimination Analysis* (LDA), [59].
- *k-nearest neighbor* (KNN), [53], [59].
- *Multilayer Perceptron* (MLP), [59].
- *Convolutional Network*, [40], [62].
- *K-Means Hierarchical*, [31].

❒ **RQ4: ¿Cuáles son las medidas empleadas para cuantificar el sesgo?**

El sesgo puede existir en los sistemas de manera implícita. Un reto es su mitigación, puesto que no existe una medida específica que determine cuándo su presencia puede ser leve y que esta no afecte al funcionamiento y los resultados que se obtienen en el sistema. Sin embargo, los investigadores han encontrado la manera de medir la presencia de sesgo empleando metodologías propias y en ciertos casos existentes. De esta forma, la manera como se mide el sesgo puede ser:

- Medidas de la distancia de cosenos entre las palabras [61], [63], [57] y también el Coeficiente de Pearson [37].
- *Accuracy*, *Precision*, *Recall*, *F-Score*, *TF-IDF inverse document frequency* (IDF), *Linguistic Inquiry* y *Word Count Additional Lexical features* [59], [62], [56].
- Clasificación de documentos en base a la suma de puntajes positivos y negativos de las palabras, además, de *Polarity Bias Rate* (PBR) que es la relación de falsos

positivos menos falsos negativos sobre el total de documentos [53], [54].

- Medida de sentimiento en los *tweets* en un radio logarítmico entre palabras positivas y negativas[60].
- Medida del impacto de los *tweets* en el *timeline* del contenido de los usuarios[36].
- Medida del sesgo de las cuentas de los audibots respecto a sus amigos y sus posts [36].
- Medida del sesgo de popularidad con el conteo de *likes* y *dislikes* [36].
- Coeficiente Gini, que mide el número de *tweets* que un amigo postea en el *timeline* de la cuenta del bot [36].
- Medir el ranking de acuerdo al volumen de los *tweets* de los amigos generados con la cuenta del bot [36].
- *Naïve post-stratification raking*, usuarios en un país, partición sociodemográfica, corrección de pesos y coeficiente de Pearson [56].
- *Selection Rate*: relación del valor más alto de *accuracy rate* y el valor bajo de *accuracy rate* [40].
- *Gender Stereotype Reinforcement* (GER) y *Mean Average Precision* (MAP) [55].
- Suma y Diferencia entre Falsos Positivos: términos de falsos positivos y términos específicos de falsos positivos [62].
- Suma y Diferencia entre Falsos Negativos: términos de falsos negativos y términos específicos de falsos negativos [62].
- *Pinned Area Under the curve* (pinned AUC) detectar sesgos en una gama más amplia de casos de uso [62].
- Medida de la diferencia del segmento menos la lista de recomendaciones de un grupo de usuarios [31].
- Probabilidad en la matriz y valor estocástico de Markov [58].
- El coeficiente de agrupamiento utiliza la expansión de Taylor para aproximar el sesgo y muestra que la expansión cuadrática es lo suficientemente buena para la aproximación. La expansión cuadrática implica la varianza y la covarianza de la muestra [64].

❒ **RQ5: ¿Cómo se mejora el proceso de aprendizaje de los algoritmos en los SR?**

La mejora en el aprendizaje de los algoritmos depende nuevamente de la metodología adoptada en cada investigación, de esta manera algunas mejoras se listan a continuación:

- Uso de diccionarios de sentimientos basados en el LSD *Lexicode Sentiment Dictionary* [60].
- Identificación de las palabras que están mal clasificadas como no sesgo o falso sesgo, las palabras ambiguas, palabras con errores de anotación o errores aleatorios [59].
- Se trabaja con análisis de sentimientos del lenguaje, luego se normaliza las frases con una intensidad de sentimientos positivo o negativo. Un valor positivo penalizará de manera positiva al aumentar el costo de los errores falsos positivos mientras que un valor negativo penalizará a las predicciones de polaridad negativa al aumentar el costo de los errores falsos negativos [53].
- Base de datos con calificaciones de más a menos para palabras positivas y negativas [54].
- Mejora en las frecuencias de las palabras durante el entrenamiento del modelo para obtener una corrección automática en los pesos de las palabras, además clasifica en categorías al grupo de palabras, por ejemplo para *suicide: mortality, heart disease mortality, number of mentalle unhealthy days* [56].
- Crear listas con todas las palabras que contienen sesgo (lista palabras semilla). Se toma los datos y se compara con la lista de palabras semillas, este conjunto puede contener palabras que no tengan sesgo entonces se obtiene la media de múltiples palabras semillas [75].
- Usar tres modelos:modelo de línea base, *baseline*, *líne base* re definido y VGG16 usando un conjunto logístico [40].
- Usar una lista rankeada, *likelihood* e historial de búsqueda [55].
- Usar un modelo de línea base, un modelo de mitigación de sesgo y un modelo de control [62].
- En el proceso de división de los clústers, si el sesgo máximo de los nuevos clústers es mayor o igual al original, el clúster se divide, de lo contrario se mantiene el clúster original, las iteraciones se hacen hasta que el número de clúster ya no se pueda dividir. Usar algoritmos colaborativos basados en factorización matricial y arquitecturas de *Deep Learning* [31].
- Usar términos específicos de género, términos específicos de ocupaciones de género,

características específicas de género y términos específicos físicos de género [61].

- Hacer que el sesgo sea transparente en el diseño de la interfaz de búsqueda. Hacer que la clasificación sea consciente, para esto se compensa con otras métricas, como la relevancia o la popularidad [63].
- Se divide la edad en grupos más específicos y se separa los grupos en subgrupos [58].
- Se incluye más términos durante el aprendizaje [57].

❒ **RQ6:¿Cómo modelar los SR para disminuir la probabilidad de la presencia de sesgo algorítmico y mejorar el proceso de aprendizaje automático de los algoritmos en los SR?**

En base a los estudios citados anteriormente (artículos referenciados al proporcionar respuesta a las preguntas 1, 2, 3, 4 y 5), para modelar un Sistema de Recomendación y disminuir el sesgo, se exponen diferentes metodologías que son aplicables dependiendo del campo de estudio. Todos los autores realizan un análisis con los datos recolectados que previamente fueron procesados. Además, se emplean diferentes métricas que permitan cuantificar la presencia del sesgo. Se usa nuevos diccionarios, datos sintéticos, palabras calificadas, conteo de palabras, entre otras técnicas que ayudan a disminuir en cierto grado la presencia del sesgo comparando dos escenarios, el antes y el después.

2.2 FASE 2 : ENTENDIMIENTO DE DATOS

En esta fase del proyecto, la principal actividad es la compresión de los datos que fue recolectada desde Twitter.

2.2.1 Territorio de Recolección de Datos

En la Sección 1.1, se define el planteamiento del problema. En esta sección, además, se detalla que el análisis del sesgo se realizará en el tema político en Ecuador. La API de Twitter emplea OIDs geográficas (*Object Identifier*). La ODI empleada en la investigación es: "1.831239,-78.183406". Para este caso, la OID se muestra en coordenadas geográficas (latitud y longitud). Estas coordenadas ayudan a identificar de manera única y precisa a Ecuador como el territorio sobre el cual se recolectará la información.

2.2.2 Limitación en la Recolección de Datos

Emplear la API de Twitter tiene algunas limitaciones como: límite de solicitudes, límites en la búsqueda histórica, acceso a *tweets* protegidos, retraso en la disponibilidad de *tweets* y restricciones de datos sensibles. Estas limitaciones son de manera general, cuando el acceso es gratuito [76]. El uso de la cuenta como desarrollador presenta limitaciones que a continuación son mencionadas.

- ***Acceso a aplicaciones:*** El acceso es a una aplicación por ambiente de desarrollo [77].
- ***Número de Tweets recolectados:*** Mensualmente se puede recoger 1500 *tweets* [77].
- ***Límite de la solicitud:*** El límite de solicitudes por ventana es 900 solicitudes en 15 minutos. Para la búsqueda de palabras claves el límite es de 180 solicitudes en 15 minutos [78].
- ***Métodos:*** Se puede usar *tweets.read*, *users.read* y *bookmrk.write* [79] [77].
- ***Costo:*** El acceso a la API es gratuito [80], sin embargo, existen versiones pagadas en las cuales no existen estas limitaciones.
- ***Búsqueda de históricos:*** Permite acceder a *tweets* de los últimos 7 a 30 días [80].

2.2.3 Recolección de Datos en MongoDB

MongoDB es empleada para almacenar grandes cantidades de datos no estructurados y semi estructurados. Un *tweet* es considerado como un dato semi estructurado. Por un lado, la estructura básica del *tweet* se mantiene constante. Esta estructura se refiere al contenido del mensaje, nombre del usuario, fecha y también a información adicional como imágenes, enlaces, videos y *hashtags*.

La información adicional del *tweet* al no mantenerse constante de un usuario a otro hace que un *tweet* sea semi estructurado.

Por otra parte, el esquema flexible de Mongo no requiere que esta base de datos trabaje con esquema predefinido. Este tipo de esquema proporciona que los *tweets* tengan diferentes estructuras y campos. Además, la interacción con la base de datos facilita la búsqueda y acceso a los datos mediante consultas y actualizaciones realizadas desde el código. Una ventaja de trabajar con MongoDB es su disponibilidad y replicación de las diferentes colec-

ciones almacenadas.

En la Figura 2.2, se muestra un ejemplo de almacenamiento de datos en MongoDB.

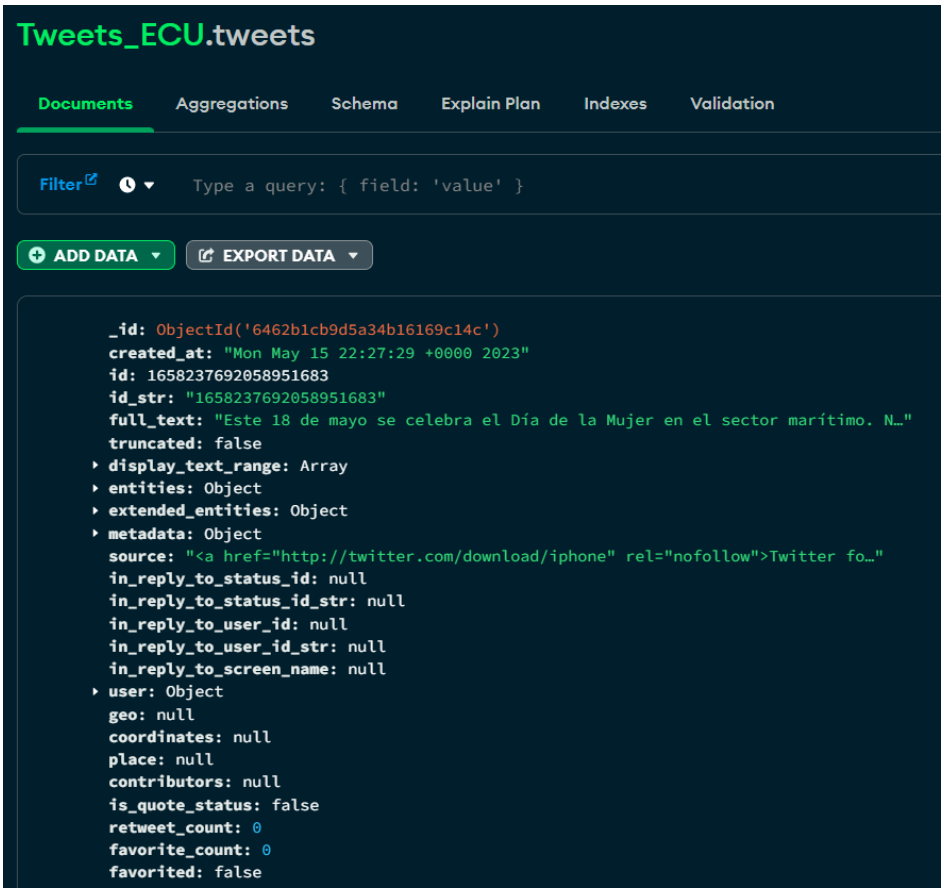


Figura 2.2: Ejemplo de Almacenamiento de un *tweet* en Mongo
Elaborado por: Guerra Cleopatra

❒ Almacenamiento de la estructura del *Tweet* en MongoBD

Los documentos en MongoDB se encuentran en colecciones. Por un lado, la primera colección con la que se trabaja contiene todos los datos del *tweet*. Esta información recopilada se guardó de dos maneras diferentes. La primera forma de almacenamiento corresponde a todos los *tweets* y *trends topics* ocurridos durante el paro nacional de Ecuador que se llevó a cabo del 13 al 30 de junio de 2022 [81]. Estos datos están almacenados en archivos de Excel, tanto los *tweets* como los *trends* fueron recolectados de manera interrumpida durante periodos de 15 minutos con una espera de 10 minutos durante las 24 horas del día.

Por otra parte, lo siguientes datos con la que se trabajó son los hechos ocurridos en mayo 2023, que corresponden a la disolución de la Asamblea Nacional y a la salida del Presidente del Ecuador. Toda esta información se almacenó directamente en dos colecciones de Mongo, una colección para los *tweets* y otra para los *trends*.

En la Figura 2.3 se detalla el almacenamiento de los *trends* correspondientes a muerte cruzada. La figura 2.4 muestra los *tweets* recolectados en muerte cruzada. Finalmente, en la Figura 2.5 se observa los *tweets* recolectados en el paro nacional.

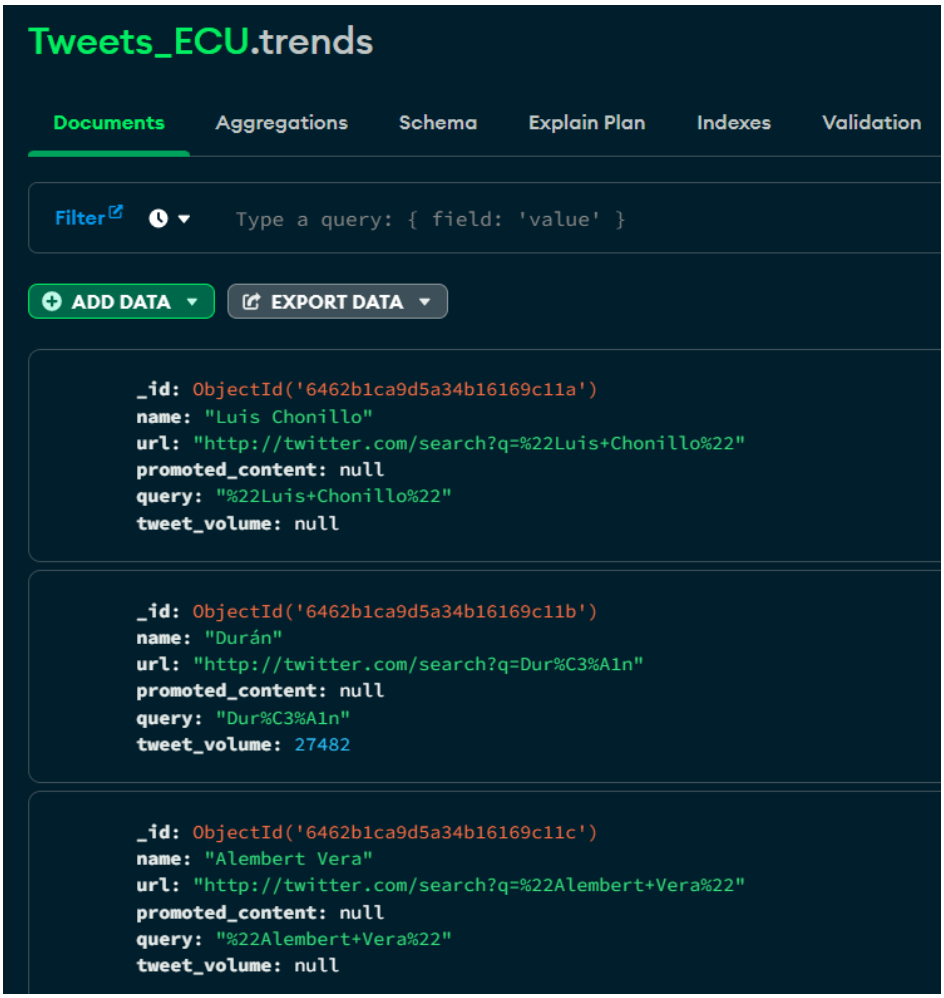


Figura 2.3: Ejemplo de Almacenamiento de un *trend* en Mongo
Elaborado por: Guerra Cleopatra

❒ Metadata de un *Tweet*

La estructura de un *trend*, se aprecia en la Figura 2.3. A detalle, un *trend* se compone de los siguientes elementos:

- ***Nombre del Trend:*** Es una frase o un término que representa la tendencia. Puede contener un *hashtag*. Generalmente estos nombres se actualizan en relación con eventos populares en períodos de tiempo [82].
- ***URL del Trend:*** Con el uso de la url mensualmente se puede recoger 1500 *tweets* [82].

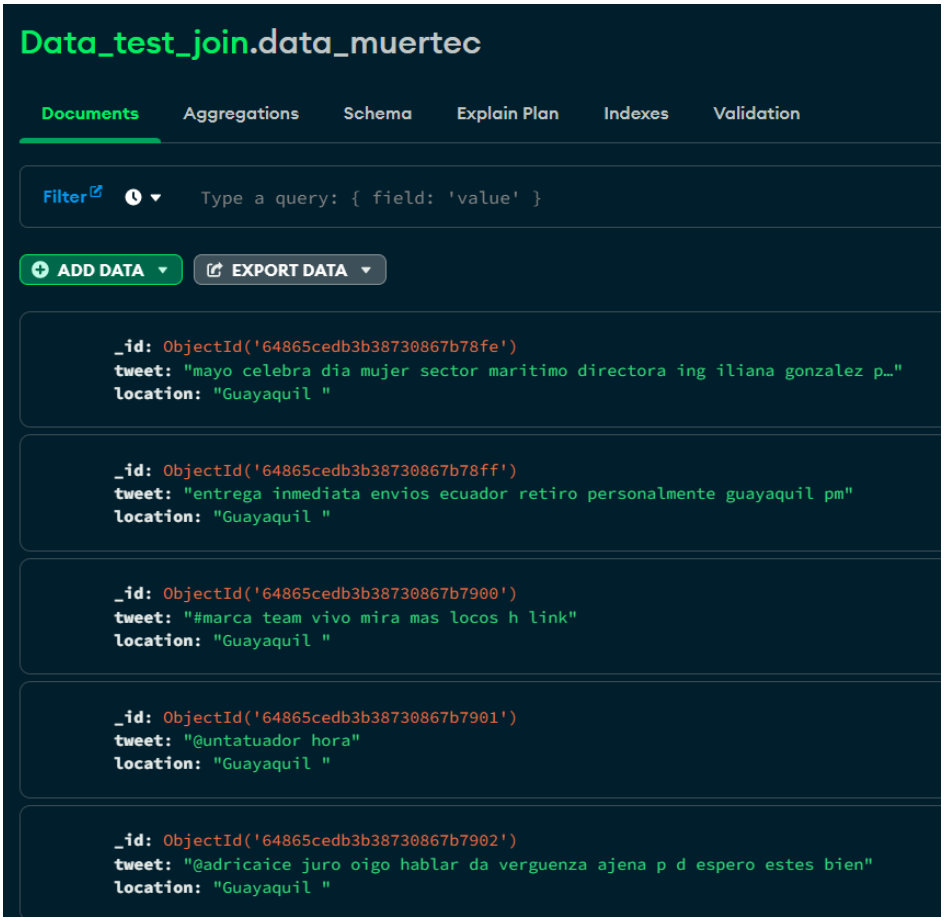


Figura 2.4: Ejemplo Almacenamiento de *tweet* Muerte Cruzada
Elaborado por: Guerra Cleopatra

- ***Contenido de Promotores:*** Usualmente este campo almacena un valor nulo. Cuando el valor es diferente de nulo, es porque la tendencia está acompañada de un promotor o anuncio publicitario [82].
- ***Consulta:*** Se refiere a la consulta o término empleado para realizar el filtro y obtener resultados específicos en cuanto a las tendencias [82].
- ***Volumen:*** Este valor puede contener valores nulos o valores diferentes de 0. El volumen, para valores diferentes de nulo, representa la cantidad de interacción que generó el *tweet* [82].

La estructura de un *tweet* se aprecia en la Figura 2.2. A continuación, se muestra una lista con una breve descripción de los principales componentes de un *tweet*.

- ***Id:*** Es un campo único que identifica de manera inequívoca a cada *tweet* [82].
- ***Texto:*** Es la parte principal del *tweet*. Con el contenido del texto se realiza el análisis de la información. El número de caracteres que el usuario puede emplear es hasta

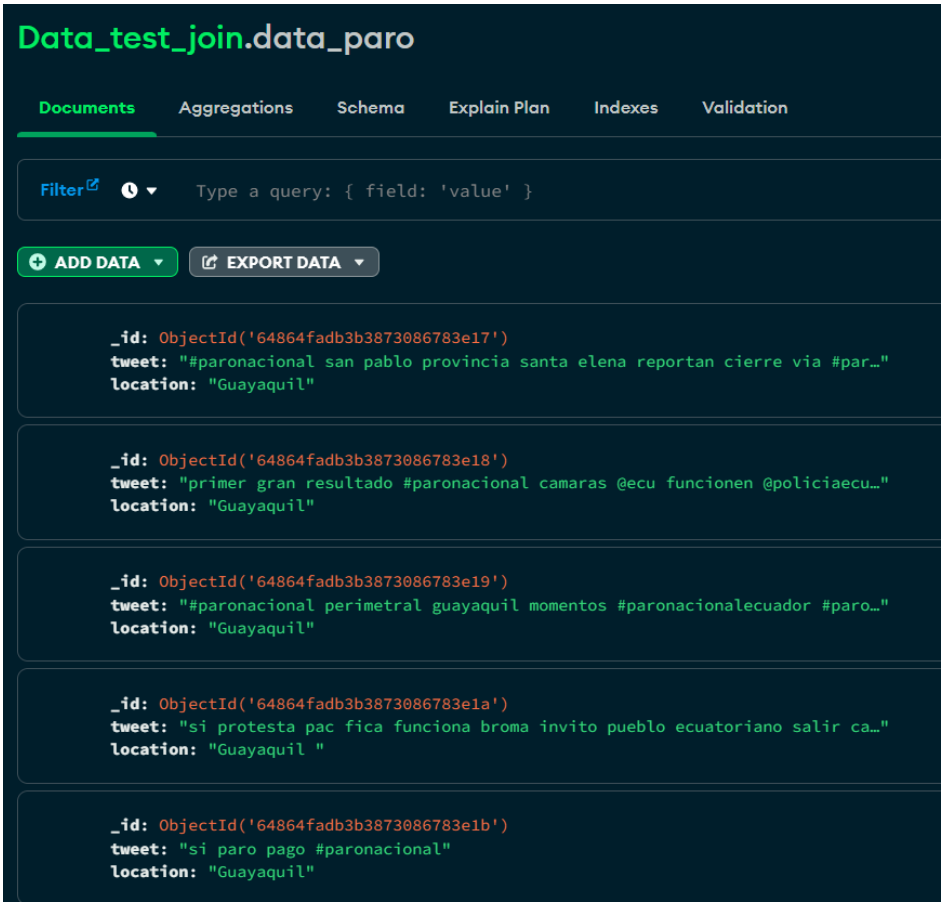


Figura 2.5: Ejemplo de Almacenamiento de *tweet* Paro Nacional
Elaborado por: Guerra Cleopatra

280 caracteres [82].

- ***Fecha y hora:*** La hora y fecha muestra cuando se publicó el *tweet* [82].
- ***Usuario:*** En este campo se encuentra la información relevante del usuario como: nombre, identificador, ubicación, descripción y url [82].
- ***Ubicación:*** Tiene un valor si el usuario proporciona esta información en su perfil. En caso de que este campo esté vacío, la ubicación se obtiene de la metadata del *tweet* [82].
- ***Hashtag:*** Este atributo se encuentra dentro de entidades y corresponde a un solo o más valores que se emplea en un *tweet*. Comúnmente este símbolo categoriza al *tweet* [82].
- ***Media:*** Contiene información de datos multimedia como imágenes, video, tamaño y las url de multimedia [82].

La información que se almacena de un *tweet* es diversa y debe ser analizada a detalle para

encontrar los campos específicos con los que se desea trabajar. En las Figuras 2.6a, 2.6b y 2.7 se muestra el contenido de los datos con detalle.

❒ **Metadata del *Tweet* Empleada**

Una vez que se almacena el *tweet* con toda la información, se procede a guardar en una nueva base de datos y en una nueva colección los *trends* y los *tweets*. Las nuevas colecciones guardarán la información relevante que será empleada en el análisis. Esta información es: el id, el oid, la fecha de creación, la ubicación del usuario, los *hashtag*, el nombre del usuario, el lenguaje, la geolocalización y las coordenadas. Esto se muestra en la Figura 2.8.

2.3 FASE 3 : PREPARACIÓN DE DATOS

En esta fase se describirá las diferentes técnicas empleadas en los datos recolectados. Utilizar técnicas adecuadas contribuirá a que el corpus recolectado sea consistente e íntegro. La cantidad de información redundante puede ser elevada, debido a que los *tweets* se recolectan cada 15 minutos de forma continua. Para disminuir información repetida se debe realizar una limpieza de *tweets* y *trends*. La limpieza es una parte primordial en esta fase. En la Figura 2.9, se muestra un ejemplo de *tweets* recolectados en tiempo real.

En la fase de limpieza en los *tweets* se aplican dos técnicas comunes que son la tokenización y la eliminación de *stopwords*. Durante la tokenización se obtiene palabras y *hashtags* de los *tweets*. Las *stopwords* que se eliminan son del idioma español y corresponden a pronombres, artículos y preposiciones que no contienen información relevante. En esta parte se usó la libraría NLTK (Natural Language Toolkit) de Python [83].

Un proceso adicional durante la limpieza de *tweets* es la extracción de números, símbolos, caracteres especiales y urls. Continuando con el proceso se convierte a minúsculas todo el texto. Las acentuaciones son reemplazadas por vocales sin tildes. Los símbolos que se conservan en el texto son # y '@'. Esto se logra con el uso de expresiones regulares. En la Figura 2.11 se detalla algunas expresiones regulares usadas en la limpieza de los *tweets*.

La Figura 2.10, muestra al *tweet* tokenizado. Además, ya no existen caracteres especiales, excepto # y '@'. En la Figura 2.12 se presenta los *tweets* limpios.

```
{
  "_id": {
    "$oid": "6462b1cb9d5a34b16169c14c"
  },
  "created_at": "Mon May 15 22:27:29 +0000 2023",
  "id": {
    "$numberLong": "1658237692058951683"
  },
  "id_str": "1658237692058951683",
  "full_text": "Este 18 de mayo se celebra el Día de la Mujer en el sector marítimo. Nuestra directora, Ing. Iliana González, participará en el webinar organizado por IMBS y COEMME\n\nEnlace para participar: https://t.co/rODIJTHZPb\n\n#DiaDeLaMujerMaritima #Webinar. https://t.co/Q0ppqToc3D",
  "truncated": false,
  "display_text_range": [
    0,
    246
  ],
  "entities": {
    "hashtags": [
      {
        "text": "DiaDeLaMujerMaritima",
        "indices": [
          215,
          236
        ]
      },
```


(a) Contenido de los datos en un *Tweet* - Parte 1


```
      {
        "text": "Webinar",
        "indices": [
          237,
          245
        ]
      }
    ],
    "symbols": [],
    "user_mentions": [],
    "urls": [
      {
        "url": "https://t.co/rODIJTHZPb",
        "expanded_url": "https://meet.google.com/ndt-gmex-zfn",
        "display_url": "meet.google.com/ndt-gmex-zfn",
        "indices": [
          190,
          213
        ]
      }
    ],
    "media": [
      {
        "id": {
          "$numberLong": "1658237685499142145"
        },
        "id_str": "1658237685499142145",
        "indices": [
          247,
          270
        ],
```


(b) Contenido de los datos en un *Tweet* - Parte 2

Figura 2.6: Datos en un *Tweet* - Parte I
Elaborado por: Guerra Cleopatra

```
  "metadata": {
    "iso_language_code": "es",
    "result_type": "recent"
  },
  "source": "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 1117113762,
    "id_str": "1117113762",
    "name": "ASOTEP",
    "screen_name": "puertosprivados",
    "location": "Guayaquil, Ecuador",
    "description": "Asociación de Terminales Portuarios Privados. Líderes en servicios portuarios de calidad con estándares internacionales de eficiencia.",
    "url": "https://t.co/bkuhQ9aocZ",
    "entities": {
      "url": {
        "urls": [
          {
            "url": "https://t.co/bkuhQ9aocZ",
            "expanded_url": "http://www.asotep.org/",
            "display_url": "asotep.org",
            "indices": [
              0,
              23
            ]
          }
        ]
      },
      "description": {
        "urls": []
      }
```


Figura 2.7: Datos en un *Tweet* - Parte 2
Elaborado por: Guerra Cleopatra

2.3.1 Vectorización de la información

Esta parte hace referencia a la representación vectorial de las palabras del corpus. La librería con la que se trabaja en Python es gensim [84]. En gensim se empleará *word2vec*, que es un algoritmo útil en la generación de representaciones vectoriales de las palabras basado en el contexto del corpus del texto [85]. El algoritmo *word2vec* guarda la semántica y las relaciones entre las palabras. Vectorizar las palabras con *word2vec* tiene como fin que el algoritmo aprenda representaciones vectoriales de las palabras y que pueda predecir otras palabras que se encuentren cerca para un contexto proporcionado.

En la Figura 2.13 se detalla la representación de la palabra #paronacional como vector. Por otro lado, un *tweet* contiene una o más palabras, lo que equivale a tener uno o más vectores. La suma de cada uno de los vectores del *tweet* da como resultado el vector final para un *tweet* específico. La representación vectorial de un *tweet* se muestra en la Figura 2.14.

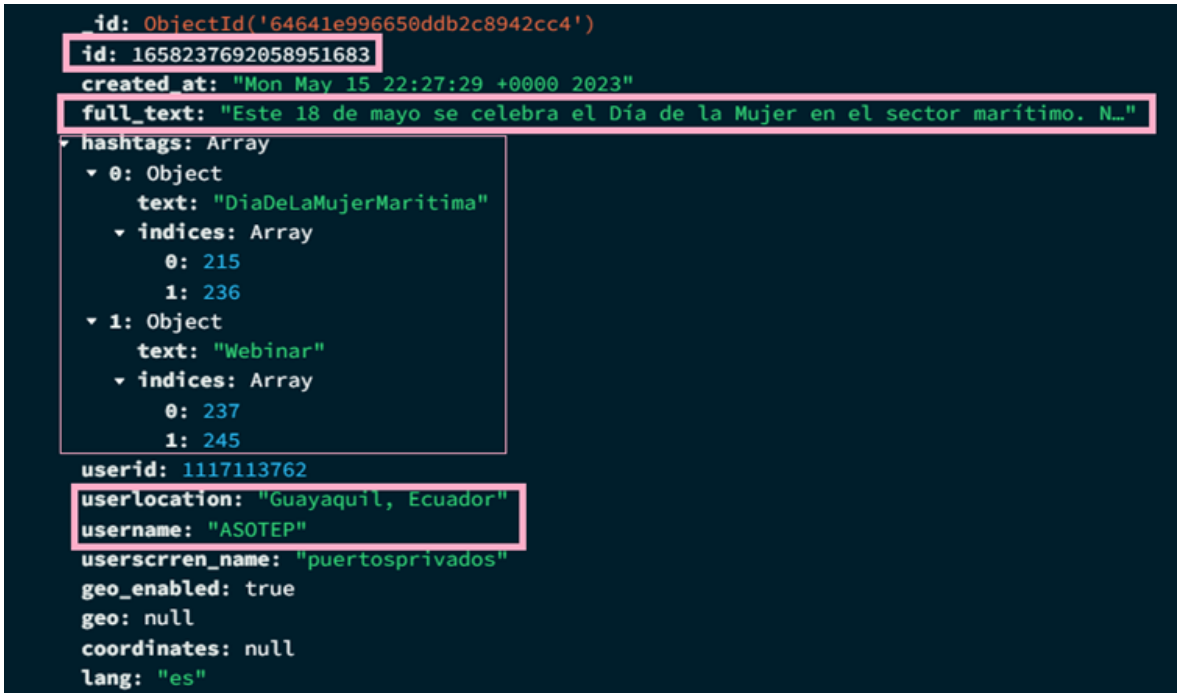


Figura 2.8: Datos Empleados de un *Tweet*.
Elaborado por: Guerra Cleopatra

2.3.2 Elección del valor k

Para determinar el número adecuado de clústers se empleará el algoritmo k-means. K-means inicializa un número (k) de centroides de manera aleatoria, asignando las palabras a un clúster más cercano de acuerdo con la distancia. Luego, nuevamente el algoritmo recalcula los centroides como el centroide promedio de las palabras asignadas a cada clúster. Los dos pasos mencionados anteriormente se repiten hasta que los centroides ya no cambien su posición.

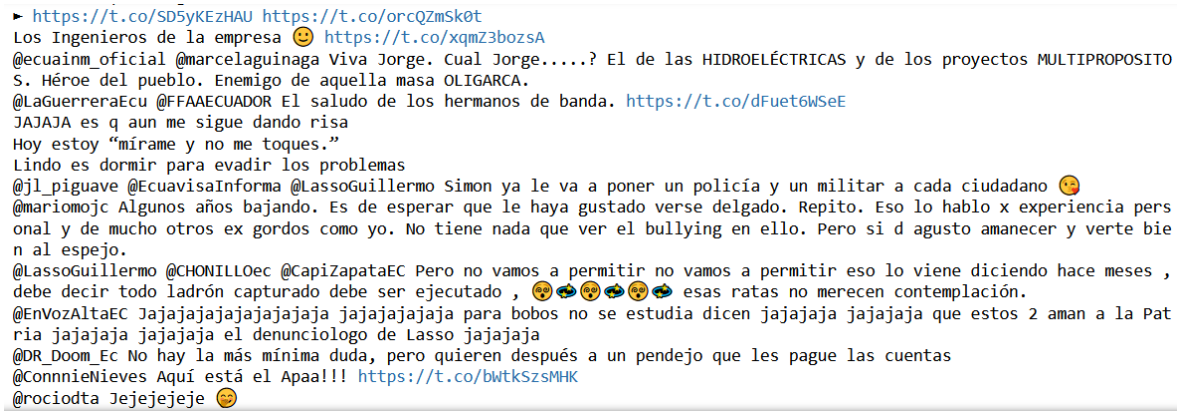


Figura 2.9: Ejemplo de *Tweets* Recolectados.
Elaborado por: Guerra Cleopatra

```
['a', 'conseguirme', 'un', 'gringo', 'que', 'me', 'saque', 'de', 'aqui', 'xq', 'esta', 'vaina', 'se', 'puso', 'feaaa']
['@ladataec', 'par', 'de', 'lacras', '#socialcristiano', 'tapinado', 'a', 'sido', 'el', 'fernando']
['la', 'primera', 'es', 'la', 'unica', 'opcion', 'los', 'demas', 'pasan', 'con', 'la', 'pantalla', 'rota', 'ilogicamente
']
['@pierinaescorrea', 'pero', 'si', 'con', 'eso', 'suenan', 'uds', 'los', 'correistas', 'con', 'la', 'muerte', 'cruzada',
'ojala', 'no', 'se', 'les', 'de']
['el', 'caos', 'para', 'entrar', 'y', 'salir', '@atm', 'transito', '@atmguayaquil', 'parque', 'empresarial', 'colon', 'la
', 'rodrigo', 'de', 'chavez']
['@yesguamani', '@fabriciovelav', '@lassoguillermo', 'y', 'lo', 'que', 'gastamos', 'el', 'pueblo', 'ecuatoriano', 'en', '
pagarles', 'a', 'ustedes', 'y', 'no', 'hacen', 'nada', 'eso', 'tambien', 'es', 'un', 'despilfarro', 'de', 'dinero', '@yes
enaguamani']
['@mariomojc', 'el', 'man', 'fue', 'gordo', 'x', 'mucho', 'tiempo', 'y', 'x', 'experiencia', 'cuando', 'empiezas', 'a', '
verte', 'delgado', 'quieres', 'seguirlo', 'siendo', 'y', 'te', 'persigue', 'el', 'fantasma', 'de', 'la', 'gordura', 'sea
', 'x', 'salud', 'x', 'que', 'te', 'gusta', 'verte', 'asi', 'en', 'el', 'espejo', 'x', 'que', 'la', 'ropa', 'te', 'queda
', 'mejor', 'x', 'que', 'entras', 'mejor', 'en', 'el', 'auto', 'etc', 'y', 'el', 'man', 'ya', 'lleva']
['@', 'hayleypluas', 'a', 'todos', 'nos', 'llega', 'el', 'momento', 'mientras', 'tanto', 'esperando', 'mi', 'momento', 'j
ajajaja']
['@janaravelez', 'ahoraaa']
['juicio', 'politico', 'contra', 'guillermo', 'lasso', 'las', 'claves', 'para', 'entenderlo']
```

Figura 2.10: Ejemplo de *Tweets* Tokenizados.
Elaborado por: Guerra Cleopatra

```python
for tweet in tweets:
    # Limpieza
    tweet_cleaned = tweet['full_text']

    # Quitar URLs
    tweet_cleaned = re.sub(r'http\S+|www\S+|https\S+', '', tweet_cleaned)

    # Quitar números
    tweet_cleaned = re.sub(r'\d+', '', tweet_cleaned)

    # Reemplazar
    tweet_cleaned = tweet_cleaned.replace('ñ', 'n').replace('á', 'a').replace('é', 'e').replace('í', 'i')

    # Convertir a minúsculas
    tweet_cleaned = tweet_cleaned.lower()

    #####
    # Conservar los # junto a las palabras
    tweet_cleaned = re.sub(r'@(\w+)', r'\1', tweet_cleaned)

    # Conservar los @ junto al texto
    tweet_cleaned = re.sub(r'#(\w+)', r'\1', tweet_cleaned)

    # Quitar signos de puntuación y caracteres especiales, excepto '#'
    #tweet_cleaned = re.sub(r'[^\w\s#]', '', tweet_cleaned)
    tweet_cleaned = re.sub(r'[^\w\s]', '', tweet_cleaned)
```

Figura 2.11: Ejemplo de Expresiones Regulares Epleadas.
Elaborado por: Guerra Cleopatra

Los parámetros empleados en k-means son:

- ***maxiter:*** El número de iteraciones es alto para mejorar la clasificación.
- ***init:*** Este parámetro determina la inicialización de los centroides. El valor empleado es *k-means++*. k-means++ realiza una inicialización inteligente que se basa en la distancia para seleccionar los centroides con mayor representividad.
- ***algorithm:*** El valor usado es *auto*.Este parámetro seleccionará automáticamente el algoritmo más adecuado de acuerdo al tamaño de los datos.

La Figura 2.15 detalla los hiperparámetros configurados.

El número de clústers se determinaron con el método del codo (*elbow*). La Figura 2.16 muestra que el número de clústers empleados es de k = 5. Mediante este valor de k=5, los clústers están identificados como: Clúster1, Clúster2, Clúster3, Clúster4 y Clúster5. Sin

```
@untatuador hora
@adricaice juro oigo hablar da verguenza ajena p d espero estes bien
@marciatama sorprendida mija solo cabreada si sacar puta estudiando mas ofrezcan basico
deje historia reves
omar andrade asistente tecnico #emelec momento malo hora levantar sacar adelante @lrodriguezerazo
inicio semana rumbo #bscvsorense detalles #ligaprobet
pon cuerpo feo creerte pues
@acostabogados preferible chiro visa visa dinero sucio
@jpjaramillo @martinminguchi pendejo vende patria
@carlarcentales saludos paso programa sonorama
@arelizavillalba cuidando mios igualitos jajajaja
asi volvieron dias tan normales emocion alguna
@letralive @cupsfire gye @ jnac @emergenciasec @paultutiven desinformado cosa xq borrarlo si leiste bien
@institutoideal @mashirafael tan ignorante pillo siempre
#insolito bobi gano record guinness
ingenieros empresa
@ecuainm oficial @marcelaguinaga viva jorge jorge hidroel ctricas proyectos multipropositos heroe pueblo enemigo aquella
masa oligarca
@laguerreraecu @ffaaecuador saludo hermanos banda
```

Figura 2.12: *Tweets* Limpios
Elaborado por: Guerra Cleopatra

```
Palabra: #paronacional
Vector de la palabra: [ 0.9258977   1.070285    0.9037513  -0.08190376 -1.7233483   0.08413237
 -0.78859276  0.8738786  -1.5956594  -0.46428075  0.28166214 -1.194849
  0.78291994  0.46320778  0.03964065  0.8224424  -0.22202666 -0.9367091
 -2.064562   -1.0708616  -1.3655475  -1.1765631   0.5061171  -0.11954936
  1.5838228  -1.2334507  -1.8683761   1.9676361  -1.7022157   1.772713
  1.4652375   0.18387312 -0.53121775 -0.43198436 -0.2651429   0.21566133
  1.4882003   2.0681796   0.5124962  -1.3594605   0.3452373  -0.30904672
 -1.0957116   0.06862701 -0.61328214  0.12401997 -1.6571629   0.7111968
 -0.5728009  -0.6094555 ]
```

Figura 2.13: Representación Vectorial de una Palabra
Elaborado por: Guerra Cleopatra

embargo, se observó que con este número de clústers las palabras no estaban clasificadas de manera correcta. La clasificación de 5 clústers se observa en la Figura 2.17.

Tomando como precedente esta clasificación, se determina que el nuevo número de clústers sea de k = 4. En esta nueva clasificación la distribución de palabras en cada clúster mejora. La Figura 2.18 detalla una mejor clasificación para las palabras. En esta clasificación se puede observar que las palabras de política pertenecen en su mayoría a un sólo clúster.

2.3.3 Formación de Diccionarios

Al mencionar diccionarios se hace referencia al conjunto de palabras con las que se trabajará en el ámbito político. El diccionario está formado por un grupo específico de palabras que se encuentran vectorizadas. Un diccionario es útil para encontrar la pertenencia de un *trend* con un *tweet*. Se empleará 4 diccionarios, por cada dimensión. En la siguiente lista se menciona el número de palabras y el clúster de política seleccionado. Cada grupo de palabras es almacenado en un archivo .txt, esta documentación forma parte del Anexo I.

- ***Dimensión 50:*** Clúster elegido es el 3 con 356 palabras.
- ***Dimensión 100:*** Clúster elegido es el 3 con 327 palabras.
- ***Dimensión 200:*** Clúster elegido es el 0 con 310 palabras.

```
# Obtener la representación vectorial del tweet
tweet_vector = np.mean([model.wv[word] for word in tweet_tokens if word in model.wv], axis=0)
#print(tweet_vector)

# Imprimir el tweet original y su representación vectorial
print("Tweet original:")
print(tweet)
print("\nRepresentación vectorial:")
print(tweet_vector)
```

```
['mayo', 'celebra', 'dia', 'mujer', 'sector', 'maritimo', 'directora', 'ing', 'iliana', 'gonzalez', 'participara', 'webinar
', 'organizado', 'imbs', 'coemme', 'enlace', 'participar', '#diadelamujermaritima', '#webinar']
Tweet original:
mayo celebra dia mujer sector maritimo directora ing iliana gonzalez participara webinar organizado imbs coemme enlace parti
cipar #diadelamujermaritima #webinar

Representación vectorial:
[ 0.3077577  -0.32887504 -0.53472805  0.27971977  0.05370532 -0.30264363
  0.41190776  0.46921372 -1.1113145  -0.33244586  0.05133514 -0.4507647
 -0.02523629 -0.15325688 -0.26505056  0.10093506 -0.28514355 -0.11784459
  0.29233164 -0.6124688   0.09045453  0.48747495  0.5148788  -0.6192163
  0.69522077  0.23050971 -0.46585557  0.05449793 -0.67891556 -0.1904983
  0.15137199  0.17492089 -0.23766318  1.0027817  -0.42306024  1.0294386
 -0.2801943   0.47568998  0.5207938  -0.31302115  0.31076354 -0.15618391
  0.13299099  0.31058878  1.1426162   0.08807725 -0.3530127   0.05824808
  0.08337342  0.4067111 ]
```

Figura 2.14: Representación Vectorial de un *Tweet*
Elaborado por: Guerra Cleopatra

```
# Especificar los hiperparámetros deseados
num_clusters = 4
random_state = 5
max_iter = 100
init = 'k-means++'
algorithm = 'auto'
```

Figura 2.15: Hiperparámetros k-means configurados
Elaborado por: Guerra Cleopatra

Un ejemplo de algunas palabras que forman parte del clúster 3 de los vectores de 50 dimensiones se presenta en la Figura 2.19.

2.3.4 Etiquetación de *tweets*

Una vez que el clúster de palabras fue elegido, como siguiente paso está la asignación de una etiqueta a cada *tweet* recolectado. La asignación de etiquetas para el clúster de política es de 1 para aquellos *tweets* que son de política y 0 para los *tweets* que no son de política.

Los pasos para asignar la etiqueta a cada *tweet* se detallan a continuación:

- Se almacena los vectores de cada clúster.
- Se calcula los centroides de cada clúster.
- Identificación del clúster de política.
- Se realizan dos pruebas. La primera es medir la similitud que existe desde el centroide del clúster hasta cada oración del *tweet*. Por otro lado, la segunda prueba consiste en determinar la distancia del coseno tomada desde el centroide del clúster hasta cada

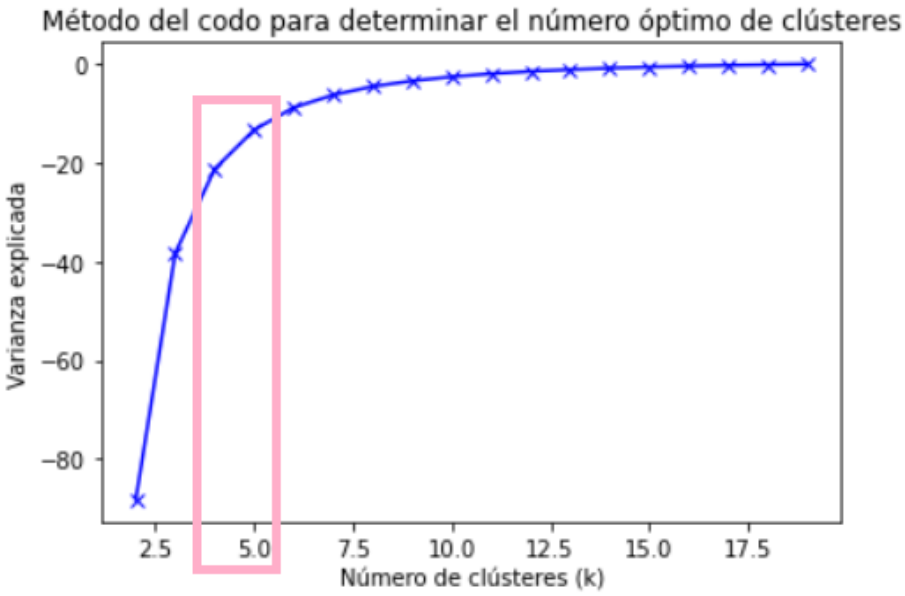


Figura 2.16: *Método Elbow*
Elaborado por: Guerra Cleopatra

oración del *tweet*. Después de analizar los dos resultados, la conclusión es que se trabajará con la distancia del coseno. Es importante acotar que para las comparaciones, en caso de ser la similitud se toma los valores más altos y en el caso de la distancia se toma los valores más pequeños. Esto quiere decir que a valores menores los *trends* están más cercanos a los *tweets*.

- Luego se asigna el valor de 1 para todos aquellos *tweets* que tengan la menor distancia al clúster 3 (para la dimensión de 50), e implícitamente los demás *tweets* tendrán el valor de 0.
- Como siguiente paso se guardan los *tweets* etiquetados con 1 y 0 en archivos de Excel con formato .xlsx, algunas capturas del contenido de estos archivos se muestran en el Anexo II.
- Se almacena los *tweets* etiquetados con el valor de 1 en un archivo diferente también en formato .xlsx. Esto se detalla en el Anexo II.

Tanto los vectores de los *trends* como los vectores de los centroides se almacenan en archivos .txt. Los ejemplos de estos archivos se muestran en el Anexo III.

Además, para los *tweets* etiquetados con 1 se asigna la ubicación geográfica que fue tomada de la metadata del *tweet*. La ubicación ayudará a realizar un análisis de la opinión de los usuarios en las diferentes provincias del país.

A continuación, se muestra en la Figura 2.20 los *tweets* etiquetados con 0 y 1. Mientras que,

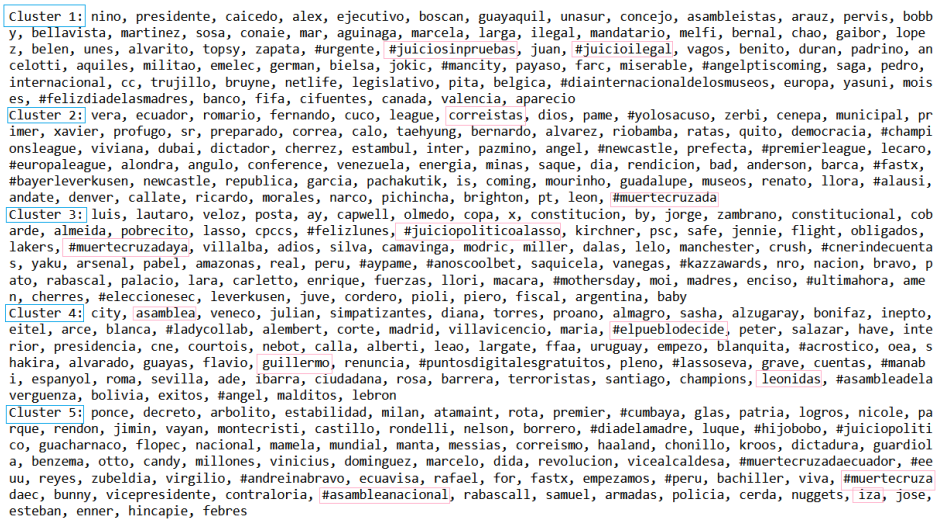
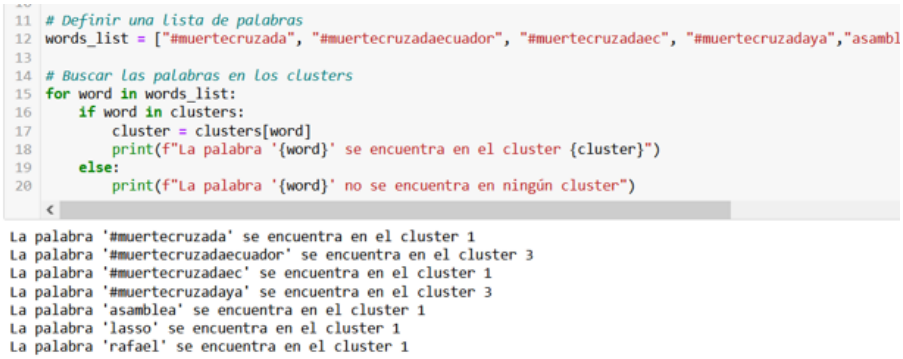


Figura 2.17: Clasificación de *Trends* en Clústers
Elaborado por: Guerra Cleopatra



```
# Definir una lista de palabras
words_list = ["#muertecruzada", "#muertecruzadaecuador", "#muertecruzadaec", "#muertecruzadaya","asambl

# Buscar las palabras en los clusters
for word in words_list:
    if word in clusters:
        cluster = clusters[word]
        print(f"La palabra '{word}' se encuentra en el cluster {cluster}")
    else:
        print(f"La palabra '{word}' no se encuentra en ningún cluster")
```

```
La palabra '#muertecruzada' se encuentra en el cluster 1
La palabra '#muertecruzadaecuador' se encuentra en el cluster 3
La palabra '#muertecruzadaec' se encuentra en el cluster 1
La palabra '#muertecruzadaya' se encuentra en el cluster 3
La palabra 'asamblea' se encuentra en el cluster 1
La palabra 'lasso' se encuentra en el cluster 1
La palabra 'rafael' se encuentra en el cluster 1
```

Figura 2.18: Ejemplo de Pertenencia de un *Trend* a un Clúster
Elaborado por: Guerra Cleopatra

en la Figura 2.21 se observa una lista de algunos *tweets* que son etiquetados como *tweets* políticos.

2.3.5 Entrenamiento y Prueba

El entrenamiento también es conocido como *training*. En el entrenamiento, el modelo aprende a relacionar de manera correcta los *trends* con los *tweets* en la entrada. El entrenamiento empieza desde la fase de limpieza de *tweets* y *trends*, continúa con la extracción de características relevantes, se realiza la vectorización y concluye con el entrenamiento de un modelo de clasificación que emplea algoritmos de clasificación.

Además, los datos etiquetados ayudan a reconocer patrones y aprender de ellos. Al contar con datos etiquetados el modelo analiza las combinaciones que durante el proceso llegan

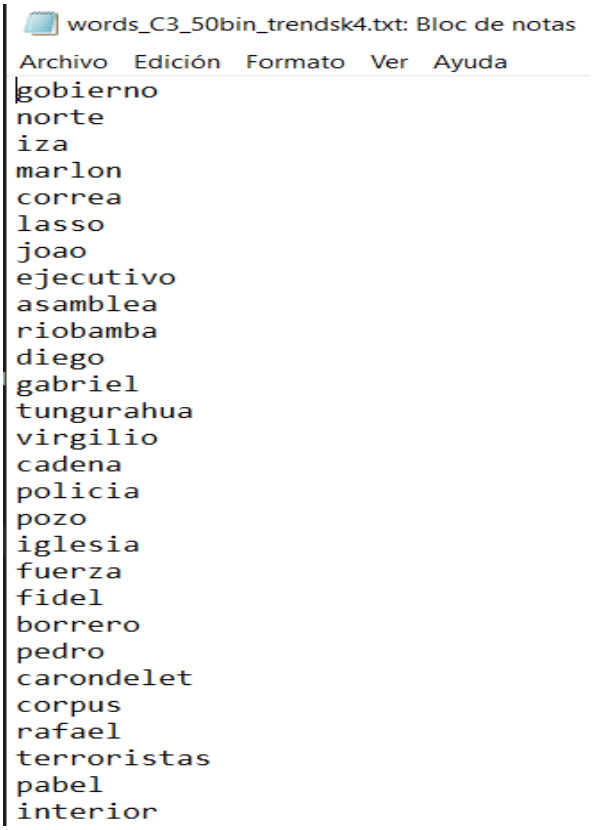


Figura 2.19: Ejemplo de palabras pertenecientes al clúster 3
Elaborado por: Guerra Cleopatra

a influir en una clasificación adecuada. En el proceso de entrenamiento los parámetros son ajustados de manera interna para realizar mejores predicciones.

Por otro lado, la Evaluación o *Testing* evalúa el rendimiento del modelo previamente entrenado. En esta fase, se hacen predicciones sobre los datos de prueba con el objetivo de comparar estos datos con las etiquetas reales. Es importante mencionar que en esta fase se obtienen métricas como la precisión, *recall*, *F1-score*, sensibilidad, entre otras. Hay que prestar atención a las métricas que se deben interpretar, puesto que en el problema planteado existen mayor cantidad de *tweets* que fueron etiquetados como no políticos.

Para el entrenamiento y pruebas se empleará el 80 % y 20 % respectivamente. El propósito que pretende cumplir es obtener una generalización en el modelo cuando se trabaje con nuevos datos. Al no emplear los datos de prueba durante el entrenamiento la evaluación del rendimiento del modelo es más objetiva.

tweet	label
#paronacional san pablo provincia santa elena reportan cierre via #paroecuador #paronacionalecuador	0
primer gran resultado #paronacional camaras @ecu funcionen @policiaecuador aparezcan	0
#paronacional perimetral guayaquil momentos #paronacionalecuador #paroecuador #paronacional	0
si protesta pac fica funciona broma invito pueblo ecuatoriano salir calles	0
si paro pago #paronacional	0
subsidio combustibles fosiles debe ser focalizado especializado direccionado unicamente sector	0
@ssbj @lassoguillermo @lenin anda seguir defendiendo nutella inseguridad falta medicina	0
leonidas iza presidente conaie si dia hoy presidente republica da respuestas entonces l	0
#paronacional armas dice manifestante miembro policia nacional localidad s	0
yuca lasso iza aqui ofende dos igual ambos velan intereses p	1
acuerdo anos socialismo pais jamas sali realizar actos bandalicos ser	0
paro nuevo culo nuevo #paronacional	0
momentos reporta cierre via puente #palmar ruta spondylus provincia	0
@lassoguillermo llevo meses trabajo insostenible oportunidades gracias gobierno	0
movilizacion social derecho constitucional lograr correcciones especificas conduccion politic	0
indigenas siempre poder paralizar pais #paronacional	0
motivacion semana ninguna encima aguantar gente indigenas paro ata	0
densita situacion ecuador #paronacional	0
ahora existe normalidad provincia #santaelena relacion #paronacional sectores transp	1
llegaron indigenas guayaquil #paronacional	0
estan cerca manifestantes estan preparando terreno clima llego lluvia neblina #guayaquil	0

Figura 2.20: *Tweets* etiquetados como 0 y 1
Elaborado por: Guerra Cleopatra

tweet	label
yuca lasso iza aqui ofende dos igual ambos velan intereses p	1
ahora existe normalidad provincia #santaelena relacion #paronacional sectores transp	1
apoyo #paronacional ningun derecho exigido via pacifica vez remover voz	1
dios proteja inocente hoy salga manera democratica alzar voz clamar descontento	1
q siguen d alcahuetes dice #lassorevocatoriaya #paronacional	1
toda latinoamerica sucede mismo ninguna institucion orden publico sirve delincuencia c	1
#urgente #ecuador inicio #paronacionalecuador rechazo gobierno guillermo lasso #paronacional	1
#urgente #ecuador inicio #paronacional rechazo medidas economicas gobierno guillermo lasso	1
va traer mucha cola nivel internacional desgaste gobierno comienza curso exterior	1
#urgente sucede ahora ciudad guaranda cierran vias acceso guaranda riobamba gallo rumi v	1
colegios privados estan optando clases virtuales manana solo ministra educacion dice q	1
unica manera lasso tome cuenta sentir popular #paronacional segun lasso amiga loren	1
gobierno hijueputas #lassorevocatoriaya #paronacional #paroecuador	1
si lasso gestion medianamente aceptable nadie apoyaria #paronacional dieran importancia	1
elites organizan protestas sociales llaman democracia libertad si organizan sino l	1
loco cortez	1
#radioambato #deportes loco sali c rcel abogado gabriel loco cortez informo juga	1

Figura 2.21: *Tweets* con etiqueta 1
Elaborado por: Guerra Cleopatra

2.4 FASE 4 : MODELAMIENTO

La representación vectorial de las palabras se efectuará en tres dimensiones. La necesidad de trabajar con estas dimensiones es para analizar los resultados y la cercanía que las palabras experimentan con diferentes dimensiones. Además, se busca determinar si el número de palabras que se encuentran en un clúster varían de un clúster a otro.

2.4.1 Elección de Dimensiones de Vectores

Las dimensiones elegidas son: 50, 100 y 200 dimensiones. Es decir, una palabra contará numéricamente con 50, 100 y 200 vectores. En la Figura 2.22, se muestra la palabra política como un vector de 100 dimensiones.

```
Vector de la palabra "politica":
[-2.1373401e+00  8.0822986e-01 -2.5153813e+00  1.0919323e+00
 -2.0859928e+00 -1.2548137e+00 -4.5253220e-03  3.4973409e+00
  9.7165883e-01  1.9063934e+00  1.7054582e-02 -1.2354970e+00
  2.0315411e+00  3.3546972e-01 -2.0797117e+00 -1.7853172e+00
 -9.5868021e-01 -1.1481298e+00  1.1711103e+00 -1.5272447e+00
 -5.0536031e-01  3.3708668e-01 -1.5671326e+00 -1.0458575e+00
  2.2372978e+00 -1.2533226e+00  7.4515212e-01  1.4916065e+00
 -1.6937454e+00 -8.2955760e-01  2.8769491e+00 -4.1090322e-01
  8.0936277e-01  7.8173721e-01 -1.2463654e+00 -2.4763210e-02
 -1.3339336e+00  6.8940163e-01 -6.8938339e-01 -8.5649049e-01
 -4.0530500e-01  3.6000013e-01 -1.4219290e+00  8.4977686e-01
 -7.7105004e-01 -4.3288788e-01  1.2493736e+00 -1.8589338e+00
 -1.3688157e+00 -1.1595780e+00 -2.0174139e+00 -1.5047417e+00
 -1.1467468e+00  9.0298003e-01 -2.9103467e-01  1.0544069e+00
  2.3021362e+00  7.6992047e-01  9.5397514e-01 -1.4495195e-01
  2.2115347e+00  2.8613247e-03  1.5057547e-01 -1.7723217e+00
 -1.2421287e+00  6.0494411e-01 -5.0046158e-01  2.9008949e+00
 -1.3077315e+00  8.3285719e-02 -6.2595004e-01 -4.4142503e-01
  2.4692769e+00 -1.4828143e-01  3.4951820e+00 -2.2418914e+00
 -2.8560741e+00 -1.0373375e+00 -2.3489373e+00  9.8921329e-01
  5.7502802e-02  5.3604752e-01 -1.1388010e-01  2.0574296e+00
 -1.0768647e+00 -2.0862372e+00  3.6756217e-01 -1.3678635e+00
  7.7964240e-01  8.0417085e-01  3.1264970e+00 -4.8780513e-01
  1.7701783e+00 -3.8835070e-01 -1.0043026e+00  1.0449930e+00
  1.7532079e+00  1.1105964e+00  8.1888336e-01  3.1192204e-01]
```

Figura 2.22: Representación de una palabra en 100 dimensiones
Elaborado por: Guerra Cleopatra

Por otro lado, el Anexo IV, detalla los resultados en cuanto a similitudes, distancias y analogías que se realizaron para las tres dimensiones de vectores. El objetivo de mostrar las similitudes es para establecer que palabras están más próximas a otras. Para esta comparación es importante mencionar que todo se realiza en las mismas condiciones. En la Figura 2.23 se muestra un ejemplo de la similitud de la palabra política y de una palabra en relación a un conjunto de palabras 2.24. También, en las Figuras se observa algunos ejemplos de analogías encontradas 2.25, 2.26 y 2.27.

```
v1 = 'correa'
v2 = 'iza'
v3 = '#muertecruzada'
v4 = 'madre'
v5 = 'bancario'
v6 = 'asamblea'
v7 = 'aprobacion'
v8 = 'lasso'

s1 = float("{0:.3f}".format(1 - spatial.distance.cosine(model.wv.get_vector(v1), model.wv.get_vector('lasso'))))
s2 = float("{0:.3f}".format(1 - spatial.distance.cosine(model.wv.get_vector(v2), model.wv.get_vector('lasso'))))
s3 = float("{0:.3f}".format(1 - spatial.distance.cosine(model.wv.get_vector(v3), model.wv.get_vector('lasso'))))
s4 = float("{0:.3f}".format(1 - spatial.distance.cosine(model.wv.get_vector(v4), model.wv.get_vector('lasso'))))
s5 = float("{0:.3f}".format(1 - spatial.distance.cosine(model.wv.get_vector(v5), model.wv.get_vector(v8))))
s6 = float("{0:.3f}".format(1 - spatial.distance.cosine(model.wv.get_vector(v6), model.wv.get_vector(v7))))

print('Similitud entre "{}" y "lasso": {}'.format(v1, s1))
print('Similitud entre "{}" y "lasso": {}'.format(v2, s2))
print('Similitud entre "{}" y "lasso": {}'.format(v3, s3))
print('Similitud entre "{}" y "lasso": {}'.format(v4, s4))
print('Similitud entre "{}" y "{}": {}'.format(v5, v8, s5))
print('Similitud entre "{}" y "{}": {}'.format(v6, v7, s6))

Similitud entre "correa" y "lasso": 0.732
Similitud entre "iza" y "lasso": 0.632
Similitud entre "#muertecruzada" y "lasso": 0.662
Similitud entre "madre" y "lasso": -0.161
Similitud entre "bancario" y "lasso": 0.478
Similitud entre "asamblea" y "aprobacion": 0.655
```

Figura 2.23: Ejemplo de similitud de una palabra en relación a otras palabras
Elaborado por: Guerra Cleopatra

2.4.2 Algoritmos de Aprendizaje Supervisado Empleados

En el presente trabajo se evalúan cuatro algoritmos de aprendizaje supervisado. Estos algoritmos son: árbol de decisión, *Support Vector Machine*, KNN y Naïve Bayes. Cada algoritmo es evaluado en condiciones similares.

2.5 FASE 5 : EVALUACIÓN

2.5.1 Preguntas de Investigación Aplicadas al Escenario de Estudio

En esta sección se responderá a las preguntas de Investigación planteadas en la fase inicial, pero, en base al proyecto de investigación planteado.

- ❒ **RQ1: ¿Cuál es el tratamiento de los datos recolectados?**

En la Fase 3 se realizó una descripción a detalle del proceso que siguen los datos para llegar a formar un corpus de *tweets* y *trends* limpios. Es importante acotar que la principal fuente de información fue Twitter en el territorio ecuatoriano. Con los datos limpios, los datos fueron vectorizados con *word2vec* y se trabajó con palabras de 50, 100 y 200 dimensiones.

- ❒ **RQ2: ¿Cuál es la metodología empleada en la detección del sesgo en los SR?**

```
Palabras similares a "politica":
('social', 0.7557496428489685)
('conmocion', 0.7440696954727173)
('crisis', 0.7405933737754822)
('interna', 0.7212983965873718)
('cargo', 0.719679594039917)
('falicidades', 0.7165998220443726)
('grave', 0.7073567509651184)
('responsabilidad', 0.7029790282249451)
('soooooolmano', 0.6948387622833252)
('economica', 0.6933570504188538)
('preparacion', 0.6895236968994141)
('ninguna', 0.687318027019507)
('romantizar', 0.6837320923805237)
('inseguridad', 0.682798445224762)
('capacidad', 0.6769053936004639)
('militancia', 0.6755533814430237)
('mendigos', 0.6728559136390686)
('leonardolatc_', 0.6727458834648132)
('situacion', 0.6710880994796753)
('ideales', 0.66954106092453)
('aptitud', 0.6692067384719849)
('inmersos', 0.6683827042579651)
('publica', 0.6677729487419128)
('casos', 0.6668655276298523)
('fabianandradeo', 0.6598691344261169)
```

Figura 2.24: Ejemplo de similitud de una palabra en relación a un conjunto de palabras
Elaborado por: Guerra Cleopatra

La metodología empleada empieza con realizar una clasificación de los *trends* con el uso del algortimo k-means y con un número de k=4. Una vez que se obtienen las palabras en 4 clústers, se calculan los centroides. A continuación, se identifica el clúster que contenga mayor número de palabras de política. De esta forma, al contar con un clúster de política se pretende analizar el sesgo. En primer lugar, en base a una comparación con todos los *tweets*. Y en segundo lugar, se analiza las frecuencias de todos los *trends* en los datos para luego compararlos con la ubicación de cada *tweet* con las provincias de Ecuador. De esta manera, se podrá determinar el sesgo político en las diferentes regiones del Ecuador. Este sesgo se identifica en base a la interacción que los usuarios tengan en la red social de Twitter. Los análisis de frecuencias y resultados en las provincias del Ecuador se presentan en el Anexo V.

Tomando cómo referencia la metodología expuesta en la fase inicial [59], [40], [75] y [62] se plantea crear un nuevo diccionario de *trends*. Este diccionario se forma a partir de un nuevo cálculo de distancias desde el clúster de política hasta las palabras de los otros clústers. De esta manera, se establecerá un nuevo diccionario con el cual se realiza un nuevo entrenamiento. Cada análisis se efectúa en las tres dimensiones (50, 100 y 200).

```
# Ejemplo de uso
word_a = "presidente"
word_b = "lasso"
word_c = "iza"

analogy_word = find_analogy(word_a, word_b, word_c)
print(f"{word_a} - {word_b} + {word_c} = {analogy_word}")
```

```
presidente - lasso + iza = carcel
```

Figura 2.25: Ejemplo 1 - Analogía
Elaborado por: Guerra Cleopatra

```
# Ejemplo de uso
word_a = "asambleista"
word_b = "almeida"
word_c = "villavicencio"

analogy_word = find_analogy(word_a, word_b, word_c)
print(f"{word_a} - {word_b} + {word_c} = {analogy_word}")
```

```
asambleista - almeida + villavicencio = arda
```

Figura 2.26: Ejemplo 2 - Analogía
Elaborado por: Guerra Cleopatra

Además, se aplican los cuatro modelos de aprendizaje supervisado para los dos escenarios.

- ❒ **RQ3: ¿Cuáles son los algoritmos de aprendizaje supervisado que se emplean en el análisis del sesgo en los SR?**

Los algoritmos de aprendizaje supervisado que se emplean son: Naïve Bayes, Árboles de Decisión, k-nearest neighbor (KNN) y SVM.

- ❒ **RQ4: ¿Cuáles son las medidas empleadas para cuantificar el sesgo?**

Las medidas empleadas en la cuantificación del sesgo son: la exactitud (*Acuracy*), sensibilidad o *recall*, precisión y *f1-score*. Sin embargo, se tomará en cuanta otras métricas adicionales debido a que, la clase minoritaria es la clase etiquetada con 1 que representa a los *tweets* de política. Siendo estos *tweets* de tema de interés. Con esta premisa se analizará también: la especificidad, precisión positiva, tasa de falsos positivos, el valor f2 y el valor f0.5 de la clase etiquetada como política.

```
# Ejemplo de uso
word_a = "presidente"
word_b = "lasso"
word_c = "correa"

analogy_word = find_analogy(word_a, word_b, word_c)
print(f"{word_a} - {word_b} + {word_c} = {analogy_word}")

```

```
presidente - lasso + correa = profugo
```

Figura 2.27: Ejemplo 3 - Analogía
Elaborado por: Guerra Cleopatra

❒ **RQ5: ¿Cómo se mejora el proceso de aprendizaje de los algoritmos en los SR?**

Se propone mejorar el proceso de aprendizaje con la creación de un nuevo diccionario que se forma a partir del clúster de política y el cálculo de las distancias desde el centroide del clúster a las palabras de los otros clústers. De esta manera, el nuevo diccionario contendrá las palabras anteriores y a estas se suman aquellas palabras que tengan una mejor relación de distancia al clúster.

2.5.2 Propuesta del Modelo de Sesgo

❒ **RQ6:¿Cómo modelar los SR para disminuir la probabilidad de la presencia de sesgo algorítmico y mejorar el proceso de aprendizaje automático de los algoritmos en los SR?**

Cómo se explicó anteriormente el modelo que se propone se detalla a continuación:

- Vectorizar los *trends* y los *tweets*.
- Aplicar el algoritmo k-means con k=4 para obtener las palabras asociadas a cada clúster.
- Calcular los centroides de cada clúster.
- Identificar el clúster de política.
- Etiquetar los *tweets* con un valor de 1 para aquellos que son de política.
- Entrenar el modelo.
- Obtener las frecuencias de los *trends* en relación a todos los datos.
- Obtener la frecuencia de los *trends* dependiendo de la ubicación del usuario que se obtuvo de la metadata *userlocation*.
- Tomar el clúster de política y calcular las distancias a cada palabra de los tres clústers restantes.
- Identificar las palabras que tiene menor distancia al centroide del clúster y almacenarlas en un nuevo archivo, tanto las palabras como su representación vectorial.
- En otro archivo almacenar las palabras que no tienen distancias pequeñas al clúster con su representación vectorial.
- Unir aquellas palabras con la menor distancia al clúster de política, de esta manera el número de palabras incrementará.

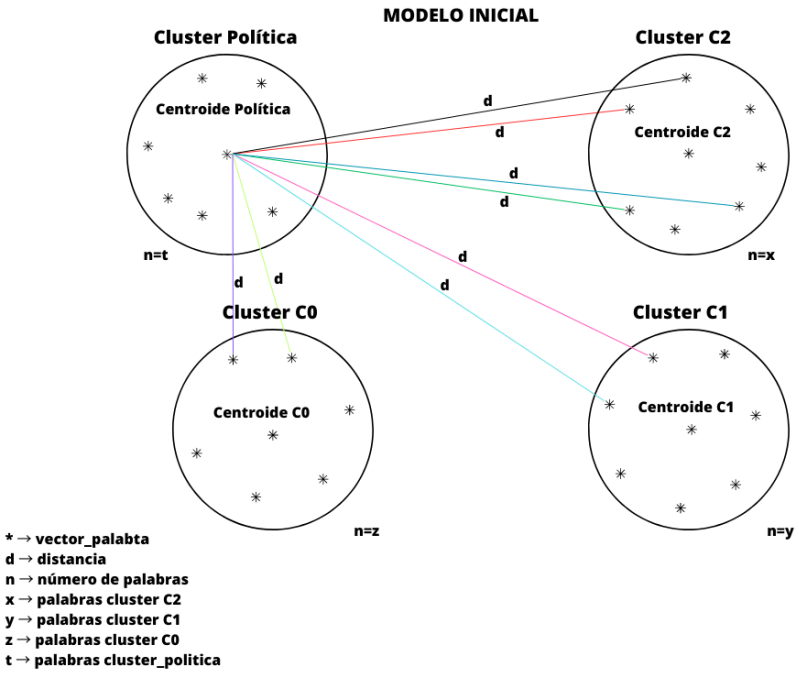
- Para los otros clústers el número de palabras disminuye.
- Recalcular los nuevos centroides de los 4 clústers.
- Con los nuevos centroides calcular las distancias a cada *tweet*.
- Repetir el procedimiento e identificar los nuevos *tweets* etiquetados.

Para tener una idea más clara de la formación de clúster y de la pertenencia de las palabras, así como de la propuesta del nuevo modelo; en la Figura 2.28a se describe el escenario 1, que es el Modelo Inicial y en la Figura 2.28b se detalla el modelo propuesto que es el escenario 2.

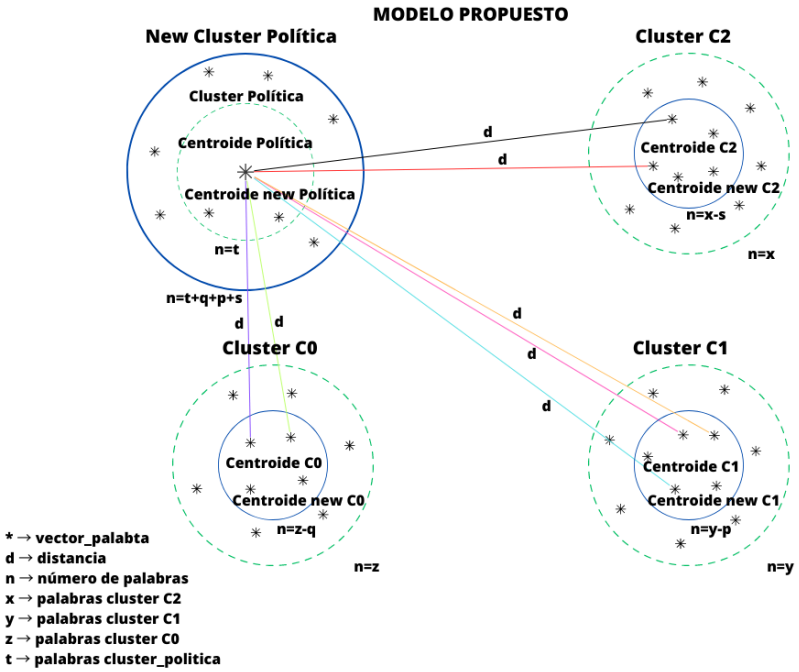
A continuación, en la Figura 2.29a se muestra los *tweets* para el escenario 1. Mientras que, en la Figura 2.29b se muestra el escenario 2 en los que se puede evidenciar que con las nuevas palabras agregadas al clúster nuevos *tweets* son etiquetados. Al igual que en el caso del escenario 1 los *tweets* del escenario 2 son almacenados en archivos .xlsx, algunos ejemplos se muestran en el Anexo VI.

2.6 FASE 6 : IMPLEMENTACIÓN

Esta fase no está dentro del alcance del proyecto de investigación. Sin embargo, para nuevas investigaciones queda abierta a una futura implementación.



(a) Modelo Inicial - Escenario 1



(b) Modelo Propuesto - Escenario 2

Figura 2.28: Comparación Escenario 1 - Escenario 2
Elaborado por: Guerra Cleopatra

tweet	label
#paronacional san pablo provincia santa elena reportan cierre via #paroecuador #paronacionalecuador	0
primer gran resultado #paronacional camaras @ecu funcionen @policiaecuador aparezcan	0
#paronacional perimetral guayaquil momentos #paronacionalecuador #paroecuador #paronacional	0
si protesta pac fica funciona broma invito pueblo ecuatoriano salir calles	0
si paro pago #paronacional	0
subsidio combustibles fosiles debe ser focalizado especializado direccionado unicamente sector	0
@ssbj @lassoguillermo @lenin anda seguir defendiendo nutella inseguridad falta medicina	0
leonidas iza presidente conaie si dia hoy presidente republica da respuestas entonces l	0
#paronacional armas dice manifestante miembro policia nacional localidad s	0
yuca lasso iza aqui ofende dos igual ambos velan intereses p	1
acuerdo anos socialismo pais jamas sali realizar actos bandalicos ser	0
paro nuevo culo nuevo #paronacional	0
momentos reporta cierre via puente #palmar ruta spondylus provincia	0
@lassoguillermo llevo meses trabajo insostenible oportunidades gracias gobierno	0
movilizacion social derecho constitucional lograr correcciones especificas conduccion politic	0
indigenas siempre poder paralizar pais #paronacional	0
motivacion semana ninguna encima aguantar gente indigenas paro ata	0
densita situacion ecuador #paronacional	0
ahora existe normalidad provincia #santaelena relacion #paronacional sectores transp	1
llegaron indigenas guayaquil #paronacional	0
estan cerca manifestantes estan preparando terreno clima llego lluvia neblina #guayaquil	0

(a) Tweets Etiquetados - Escenario 1

tweet	label
#paronacional san pablo provincia santa elena reportan cierre via #paroecuador #	1
primer gran resultado #paronacional camaras @ecu funcionen @policiaecuador a	0
#paronacional perimetral guayaquil momentos #paronacionalecuador #paroecuac	1
si protesta pac fica funciona broma invito pueblo ecuatoriano salir calles	0
si paro pago #paronacional	1
subsidio combustibles fosiles debe ser focalizado especializado direccionado unica	0
@ssbj @lassoguillermo @lenin anda seguir defendiendo nutella inseguridad falta n	1
leonidas iza presidente conaie si dia hoy presidente republica da respuestas entonc	0
#paronacional armas dice manifestante miembro policia nacional localidad s	1
yuca lasso iza aqui ofende dos igual ambos velan intereses p	1
acuerdo anos socialismo pais jamas sali realizar actos bandalicos ser	0
paro nuevo culo nuevo #paronacional	1
momentos reporta cierre via puente #palmar ruta spondylus provincia	0
@lassoguillermo llevo meses trabajo insostenible oportunidades gracias gobierno	1
movilizacion social derecho constitucional lograr correcciones especificas conduc	0

(b) Tweets Etiquetados - Escenario 2

Figura 2.29: Comparación de Tweets Etiquetados
Elaborado por: Guerra Cleopatra

3 RESULTADOS Y DISCUSIÓN

En este capítulo se presenta los resultados obtenidos en el proyecto de investigación. Estos resultados brindan información sobre la presencia de sesgo en la data analizada. En los capítulos anteriores se describe a Twitter como la red social empleada para la extracción de los datos.

En la investigación se busca definir una base que medirá la presencia del sesgo. Anteriormente, se mencionó que el sesgo se encuentra en la información de manera implícita. Existen diferentes tipos de sesgo, sin embargo, para el propósito de análisis y debido a la naturaleza de los tweets recolectados en Ecuador, se analizará el sesgo político.

A lo largo del capítulo se detalla los principales hallazgos.

3.1 ANÁLISIS DE *TRENDS*

En esta sección, se mostrará el impacto de los *trends* en los datos recolectados. Se mide la frecuencia de cada *trend* en relación con el total de los *tweets* almacenados. El motivo principal de detallar estos resultados, es debido a que en base a los *trends* se recoge la información. Los *trends* son palabras que contienen información valiosa.

La representación inicial que se da a los *trends* es en gráficas de palabras. El uso de un diagrama de palabras permite visualmente encontrar de manera precisa, concisa y comprensible el impacto de las palabras. Por otro lado, las palabras más frecuentes se muestran en un tamaño grande y en posiciones prominentes. A diferencia de las palabras menos frecuentes que se ubican en lugares menos destacados.

De esta manera, el diagrama de palabras cumple el propósito de resaltar las palabras clave que encierran un gran significado. Además, de identificar patrones y temas en tendencia.

A continuación, en la Figura 3.1 se muestra un diagrama de palabras del Paro Nacional y en la Figura 3.2 se observa el diagrama de los acontecimientos de Muerte Cruzada en

Ecuador. En estas Figuras se puede determinar que palabras como: nacional, guayaquil. ministro, muerte cruzada, asamblea, Borrero, asambleístas, tienen mayor tendencia que fernando, correa, carlos, calo, el padrino, juiciosinpreubas que su tendencia es menor.

Figura 3.1: Gráfica de Palabras - Paro Nacional
Elaborado por: Guerra Cleopatra

Figura 3.2: Gráfica de Palabras - Muerte Cruzada
Elaborado por: Guerra Cleopatra

3.2 ANÁLISIS DE *TWEETS*

El análisis de sesgo en los *tweets* recolectados como se expuso anteriormente se realiza desde el punto de vista político en Ecuador. Es importante comprender qué provincias interactúan de manera activa en Twitter para distinguir aquellas provincias en las que se presenta una interacción casi nula.

Por otro lado, la información que ayudará a encontrar la ubicación del *tweet* es userlocation.

Sin embargo, este campo no suele presentar siempre un valor. Al no presentar un valor de ubicación el análisis esperado no será real, puesto que existen *tweets* que no presentan ubicación. En las Figuras 3.3, 3.4, 3.5 se muestra un gráfico de barras con el porcentaje de *tweets* que registran ubicación y aquellos que no la registran. El primer gráfico corresponde al Paro Nacional. A continuación, Muerte Cruzada y finalmente, para los datos unificados. En las gráficas se puede destacar que en el total de *tweets* el 91.97 % tienen registrada la ubicación y apenas el 8.03 % no registran ubicación.

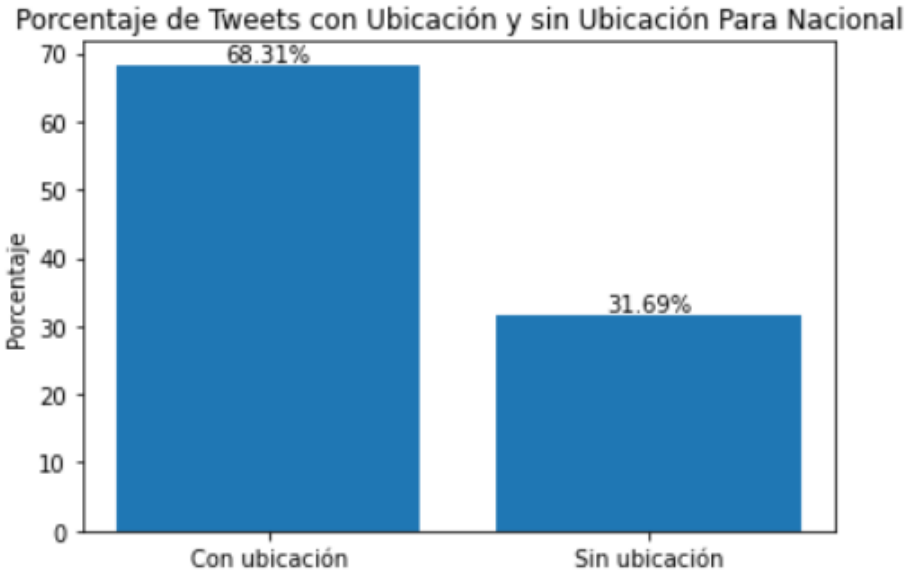


Figura 3.3: Ubicación *Tweets* en los Datos de Paro Nacional
Elaborado por: Guerra Cleopatra

Figura 3.4: Ubicación *Tweets* en los Datos de Muerte Cruzada
Elaborado por: Guerra Cleopatra

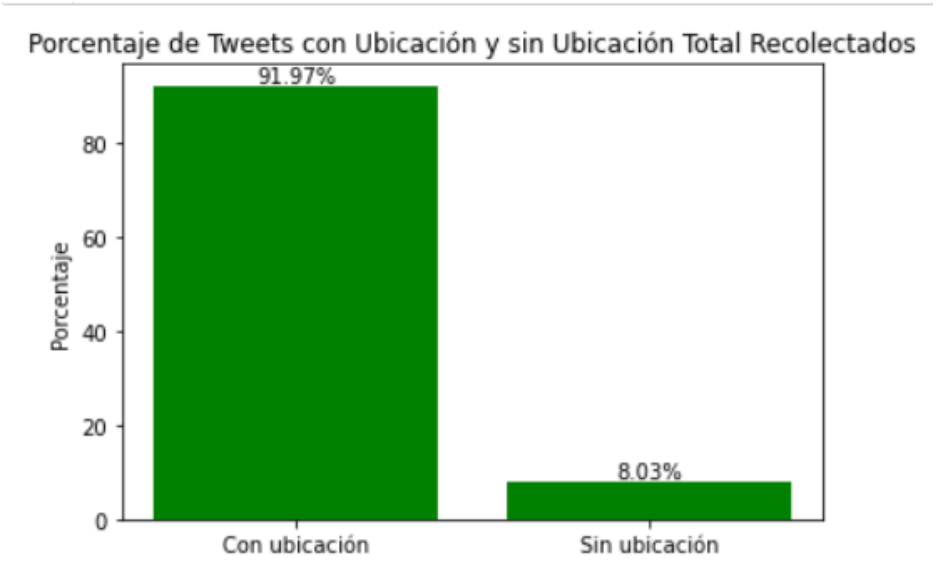


Figura 3.5: Ubicación *Tweets* en los Datos Recolectados
Elaborado por: Guerra Cleopatra

Por otra parte, con las frecuencias de las palabras se realizan histogramas que proporcionan una representación visual de la distribución de los datos. Esta agrupación muestra cómo los datos se extienden a los largo de un rango de valores. En el eje y, se representa la frecuencia, mientras que, para el eje x se muestra los valores. Es decir, cada barra del histograma representa un intervalo con la cantidad de observaciones dada por su altura.

La Figura 3.6 presenta las 50 palabras más frecuentes en un histograma. En el gráfico se aprecia que la palabra con mayor frecuencia es ecu, lasso, guillermo, @asambleaecuador, hasta la menos frecuente como guerra, protesta, entre otras.

Además, se realizó un análisis con el impacto en las diferentes provincias el país. De esta manera, en la Figura 3.7 se observa el ejemplo de un solo *trend* y cómo este tuvo acogida en algunas provincias y en otras no. Por otro lado, hay provincias en las que existe el uso de Twitter y en otras provincias el empleo de esta red social es nulo; esto se detalla en la Figura 3.8 y 3.9 respectivamente.

3.3 IMPACTO DEL TAMAÑO DEL VECTOR DE PALABRAS EN LOS RESULTADOS

En este apartado se mostrará los resultados de la vectorización de las palabras en tres dimensiones. El propósito de emplear las dimensiones de 50, 100 y 200 es para determinar

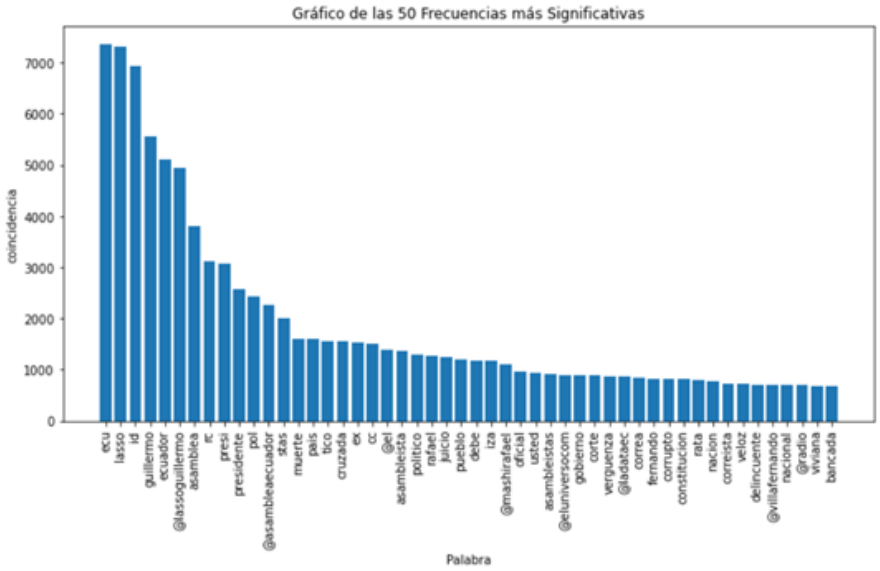


Figura 3.6: Histograma 50 palabras con Mayor Frecuencia
Elaborado por: Guerra Cleopatra

Palabra	Azuay	Bolivar	Cañar	Carchi	Chimborazo	Cotopaxi	El Oro	Esmeraldas	Galapagos	Guayas	Imbabura	Loja	Los Rios	Manabi	Morona Santiago	Napo	Orellana	Pastaza
#paronacional	66	0	0	0	3	4	0	0	0	100	0	1	0	6	1	1	0	1
san	3	1	0	0	0	2	0	0	0	82	0	0	6	9	0	0	0	0
pablo	1	0	0	0	0	0	0	0	0	36	0	0	0	9	1	0	0	0
protesta	15	0	0	0	1	4	0	0	0	65	0	0	0	4	0	0	0	0
conaie	97	1	9	0	14	9	0	1	0	575	0	0	5	29	1	5	0	11
policia	35	1	0	0	8	4	0	0	1	279	0	0	6	22	1	3	0	2
actos	5	0	0	0	2	2	0	0	0	98	0	0	3	8	0	0	0	0
ruta	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0
gracias	11	3	2	0	34	10	1	0	0	1106	1	1	12	79	1	1	1	0

Figura 3.7: Impacto de un *trend* en las provincias
Elaborado por: Guerra Cleopatra

y capturar las mejores relaciones semánticas y sintácticas entre las palabras en función de su contexto.

Estas dimensiones son empleadas en los escenarios descritos en la metodología 2.28a, 2.28b. Durante la experimentación, tanto en el escenario 1 cómo en el escenario 2, se determinó que la mayor cantidad de *tweets* etiquetados fue para la dimensión de 200.

Con estos resultados se puede concluir que existe una mayor capacidad del modelo para capturar las relaciones semánticas de las palabras cuando la dimensión del vector es mayor.

En la Figura 3.10 se observa el número de *tweets* etiquetados como política en el escenario 1. En esta figura se determina que a mayor dimensión del vector mayor es el número de *tweets* etiquetados.

Por otro lado, en la Figura 3.11 se detalla una comparación del escenario 1 y escenario 2 de

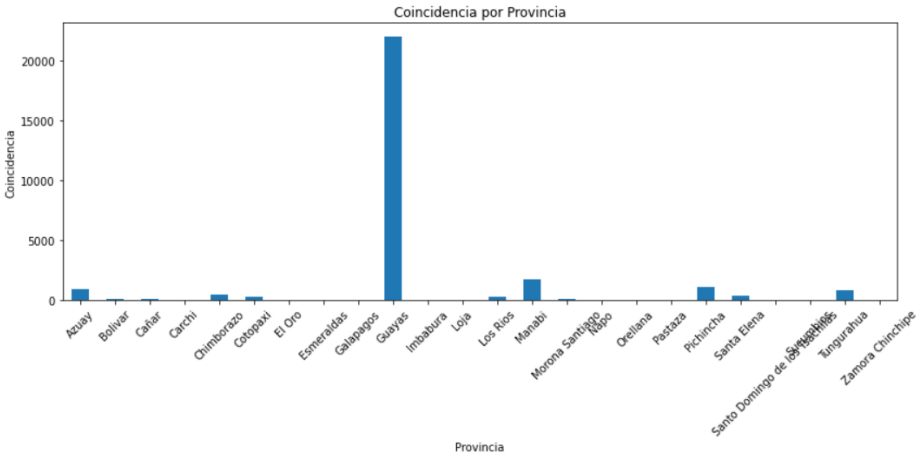


Figura 3.8: Provincias con gran Interacción en Twitter
Elaborado por: Guerra Cleopatra

los *tweets* etiquetados como política. En esta figura además se concluye que con el modelo planteado (escenario 2) el número de *tweets* etiquetados como política es mayor con una dimensión de 200.

3.4 ANÁLISIS DE LOS ALGORITMOS EMPLEADOS EN EL CONTEXTO DEL MODELO

Los algoritmos empleados en el presente trabajo fueron: árbol de decisión, KNN, Naïve Bayes y *Support Vector Machine*. Todos los algoritmos son empleados en los dos escenarios y con dimensiones de 50, 100 y 200 respectivamente. A continuación, se mostrará a detalle los diferentes valores e interpretación de métricas empleadas para evaluar el rendimiento de los modelos de aprendizaje automático.

Es importante en esta sección acotar que las métricas sobre las que se tendrá mayor énfasis son aquellas que están relacionadas directamente con los valores de *"True Positives"* (TP), debido a que la clase minoritaria es la que está etiquetada como los *tweets* de política (label 1). La clase mayoritaria está etiquetando a los *tweets* con 0. Se mostrará los resultados en cuanto a Precisión, *Recall*, *F1-score*, F2 y F0.5. Sin embargo, para la clase mayoritaria también se mostrará resultados como Exactitud y Especificidad.

❒ Dimensión 50

Los resultados de las Tablas 3.1 y 3.2 respectivamente muestran cada métrica tomada

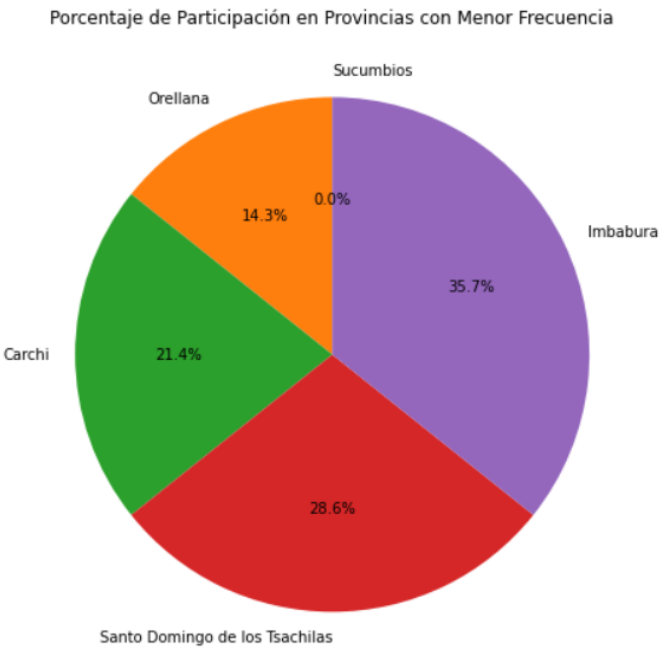


Figura 3.9: Provincias con Interacción casi nula en Twitter
Elaborado por: Guerra Cleopatra

para los cuatro algoritmos propuestos en el trabajo de investigación. La columna *Before* corresponde a las medidas del Escenaro 1. Mientras que, la columna de *After* muestra las métricas del Escenario 2.

Además, los valores de *Recall* son similares a Sensibilidad y Precisión a Precisión Positiva, esto debido a que, para el primer caso se calcularon a partir de la matriz de confusión y en el segundo caso con un reporte de métrica.

Los mejores resultados son para SVM en *recall* 3.12 con un valor de 0,91339135. A este valor sigue el algoritmo de árbol de decisión. Para este caso el *recall* asocia la capacidad del modelo para detectar y clasificar correctamente los *tweets* etiquetados como política.

Por otro lado, la precisión 3.13 muestra la exactitud con la que los *tweets* clasificados como política son realmente de política.

El valor representativo del *F1-score*, es gracias a que muestra una evaluación equilibrada del rendimiento de los casos verdaderos negativos. La métrica de F2 da mayor importancia a *F1-score*. El valor de F0.5 a la precisión.

Algunas gráficas se muestran a continuación, las demás se detallan en el Anexo VII.

❒ **Dimensión 100**

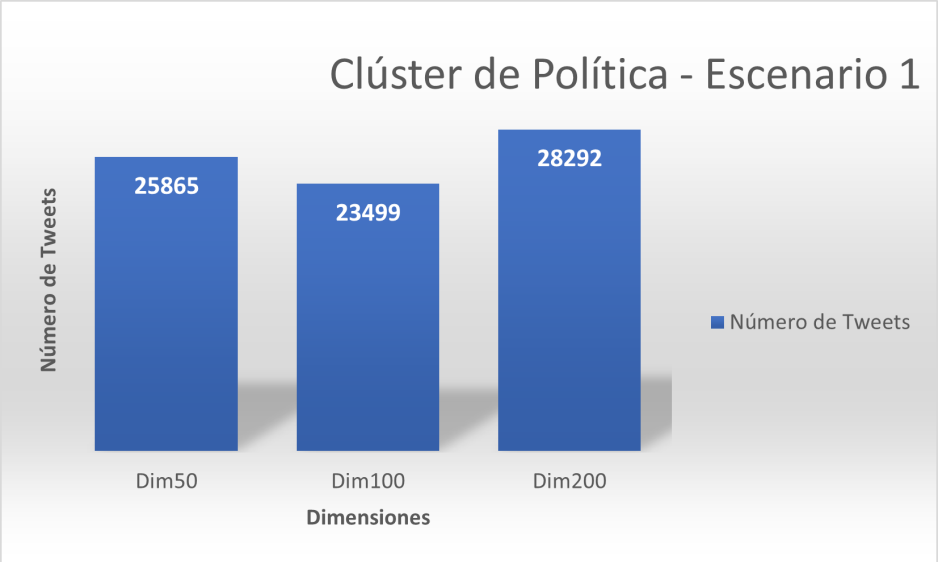


Figura 3.10: Comparación de Dimensiones de *tweets* Etiquetados - Escenario 1
Elaborado por: Guerra Cleopatra

Tabla 3.1: Algoritmos Dimensión 50 - Parte I

	Dim 50			
	Arbol de Decision		**KNN**	
	Before	After	Before	After
Exactitud	0,83320909	0,89562586	0,83339752	0,79807428
Precisión	0,81290909	0,84338457	0,78682978	0,82983193
Recall	0,85454893	0,90103874	0,57779052	0,60991016
F1-score	0,83320909	0,87125891	0,66629932	0,70307443
Otras Métricas				
Sensibilidad	0,85454893	0,90103874	0,57779052	0,60991016
Especificidad	0,92049757	0,89213646	0,93672255	0,91937381
Precisión positiva	0,81290909	0,84338457	0,78682978	0,82983193
Tasa de falsos positivos	0,07950243	0,10786354	0,06327745	0,08062619
Valor F2	0,84588315	0,88888581	0,61021397	0,6440472
Valor F0.5	0,82090922	0,85431750	0,73373786	0,77401311

La información recopilada en cuanto a las métricas para la dimensión de 100, se observan en la Tabla 3.3 y Tabla 3.4. De esta información se determina que al igual que para la dimensión de 50, el mejor algoritmo es para SVM. El segundo lugar es para los árboles de decisión.

En la Figura 3.14 y 3.15 se observa que los mejores resultados se obtiene en el escenario 2. Recodando que F2 se relaciona con *F1-score* y F0,5 con la precisión del modelo.

❒ Dimensión 200

Los mejores resultados continúan siendo para la métrica de Precisión en el algoritmo SVM 3.5 y 3.6.

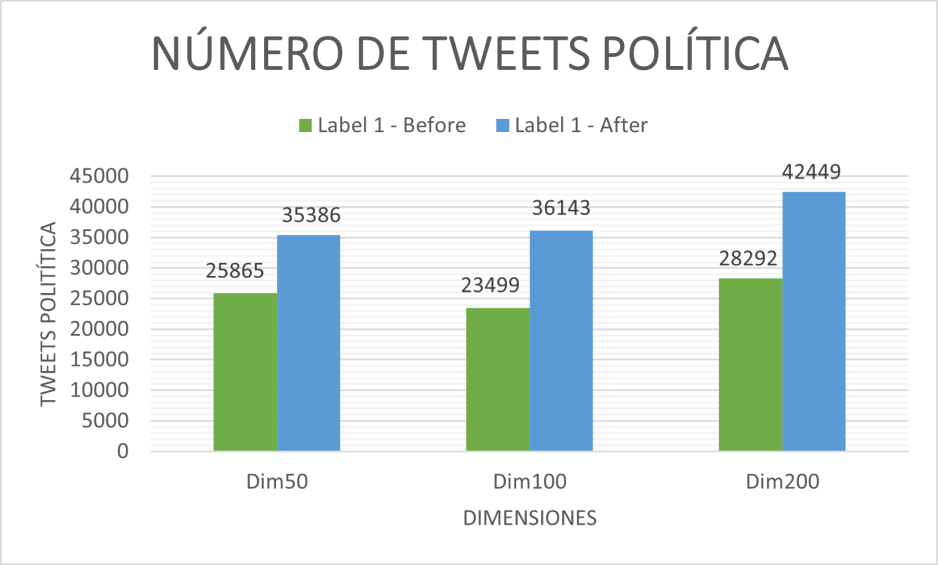


Figura 3.11: Comparación de Dimensión de *tweets* en los Dos Escenarios
Elaborado por: Guerra Cleopatra

Tabla 3.2: Algoritmos Dimensión 50 - Parte II

	Dim 50			
	Naïve Bayes		**SVM**	
	Before	After	Before	After
Exactitud	0,85529574	0,84731774	0,93061898	0,93072902
Precisión	0,84327177	0,82822642	0,87668374	0,91019723
Recall	0,61085627	0,77021336	0,88321865	0,91339135
F1-score	0,70849036	0,79816714	0,87993907	0,91179149
Otras Métricas				
Sensibilidad	0,61085627	0,77021336	0,88321865	0,91339135
Especificidad	0,95410647	0,89702289	0,94977980	0,94190571
Precisión positiva	0,84327177	0,82822642	0,87668374	0,91019723
Tasa de falsos positivos	0,04589353	0,10297711	0,0502202	0,05809429
Valor F2	0,64649243	0,78115657	0,88190389	0,91275074
Valor F0.5	0,78364064	0,81593505	0,87798298	0,91083427

En esta parte se mostrará la figura de una métrica adicional que se obtiene a partir de un reporte detallado. Esta métrica es la Precisión Positiva para los *tweets* que fueron etiquetados como política 3.16.

En esta sección se abordará el por qué SVM brinda mejores resultados sobre los otros algoritmos. SVM es muy empleado en problemas de clasificación que captura patrones y relaciones complejas. Funciona adecuadamente para grandes y pequeñas cantidades de datos gracias a su capacidad para encontrar un hiperplano óptimo de separación.

Sin embargo, los algoritmos de árboles de decisión, KNN y Naive Bayes tienen sus ventajas y para el caso de árboles de decisión en este escenario se obtuvieron resultados que capturaron las relaciones existentes en los datos analizados.

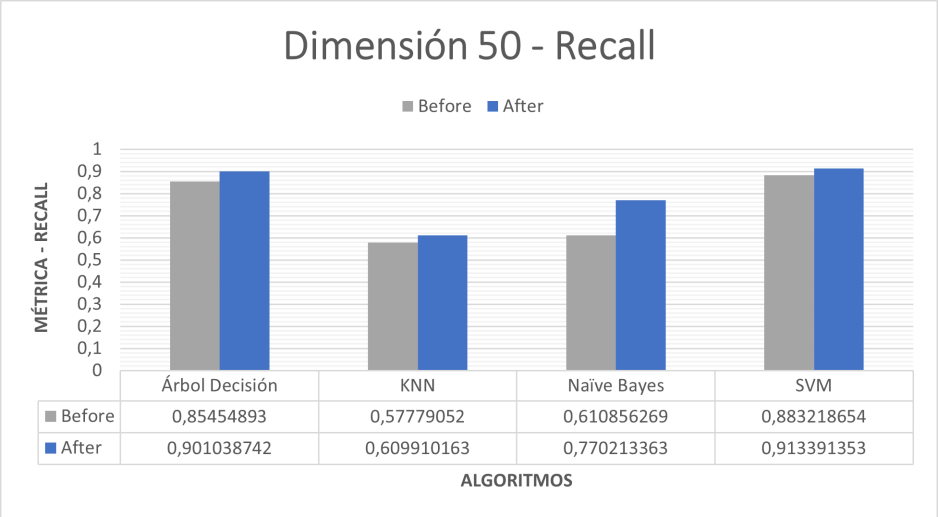


Figura 3.12: Comparación de *Recall* en los Dos Escenarios
Elaborado por: Guerra Cleopatra

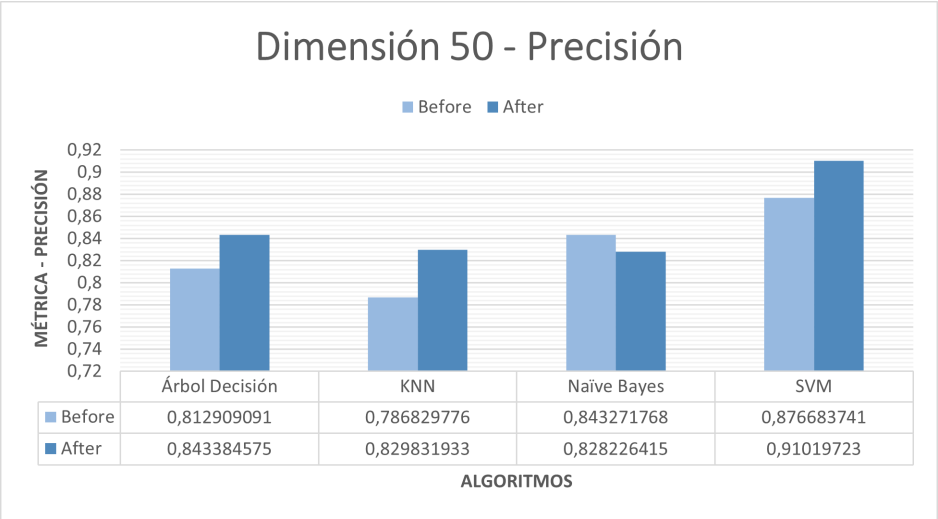


Figura 3.13: Comparación de Precisión en los Dos Escenarios
Elaborado por: Guerra Cleopatra

3.5 DISCUSIÓN

El análisis realizado en el presente trabajo empezó con la recolección de *tweets* y *trends* en Ecuador. En la data almacenada un inconveniente fue que no todos los usuarios tenían habilitada su ubicación, para ciertos casos la ubicación se pudo extraer de la metadata del *tweet* específicamente del campo *userlocation*. La ubicación fue de gran ayuda para establecer un análisis de sesgo por provincias, mediante el uso de Twitter. Se pudo determinar que en el país hay provincias en dónde no existe interacción con Twitter.

En la data existe un sesgo negativo de aquellos datos que han sido discriminados, para

Tabla 3.3: Algoritmos Dimensión 100 - Parte I

	Dim 100			
	Arbol de Decision		KNN	
	Before	After	Before	After
Exactitud	0,90740028	0,89568088	0,84632737	0,79900963
Precisión	0,79209257	0,84235689	0,76194376	0,80160082
Recall	0,87160441	0,90604587	0,59231749	0,65420431
F1-score	0,82994847	0,87304138	0,66650746	0,72044081
Otras Métricas				
Sensibilidad	0,87160441	0,90604587	0,59231749	0,65420431
Especificidad	0,91992869	0,88888889	0,93522989	0,89389800
Precisión positiva	0,79209257	0,84235689	0,76194376	0,80160082
Tasa de falsos positivos	0,08007131	0,11111111	0,06477011	0,10610200
Valor F2	0,85445013	0,89254908	0,61991915	0,67918158
Valor F0.5	0,80681282	0,85436817	0,72066722	0,76703712

Tabla 3.4: Algoritmos Dimensión 100 - Parte II

	Dim 100			
	Naïve Bayes		SVM	
	Before	After	Before	After
Exactitud	0,84440165	0,84770289	0,93837689	0,93155433
Precisión	0,87174428	0,84634642	0,87826453	0,91183391
Recall	0,46880306	0,75177206	0,88497453	0,91563586
F1-score	0,60971571	0,79626086	0,88160677	0,91373093
Otras Métricas				
Sensibilidad	0,46880306	0,75177206	0,88497453	0,91563586
Especificidad	0,97585976	0,91056466	0,95706752	0,94198543
Precisión positiva	0,87174428	0,84634642	0,87826453	0,91183391
Tasa de falsos positivos	0,02414024	0,08943534	0,04293248	0,05801457
Valor F2	0,51655598	0,76895738	0,88362434	0,91487293
Valor F0.5	0,74387123	0,82557465	0,87959838	0,91259177

este caso aquellos datos en los que la ubicación era un campo vacío.

Por otro lado, un punto importante fue establecer diferentes tamaños en los vectores. Se trabajó con 50, 100 y 200 dimensiones. Con cada dimensión el número de *tweets* etiquetados variaba siendo la dimensión 200 la que produjo mejores resultados.

Continuando con el desarrollo del proyecto, se probaron cuatro algoritmos de clasificación. De los cuales SVM proporcionó mejores resultados en cuanto a precisión, *recall*, *F1-score*, F2 y F0.5. Se hace especial énfasis en estas métricas porque la clase etiquetada como *tweets* de política es la clase minoritaria. Cabe destacar que se detectó un sesgo en los datos de entrenamiento, donde la clase minoritaria de verdaderos positivos presentaba un desequilibrio significativo en comparación con la otra clase.

Además, se propuso un nuevo modelo que incorpora un diccionario más amplio para el clúster nuevo de política. Con este nuevo diccionario el número de *tweets* etiquetados aumentó.

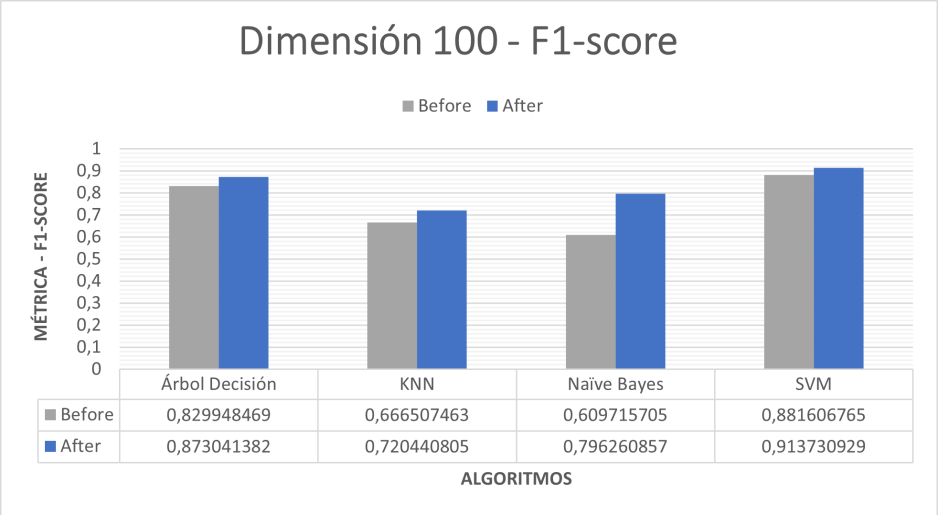


Figura 3.14: Comparación *F1-score* en los Dos Escenarios
Elaborado por: Guerra Cleopatra

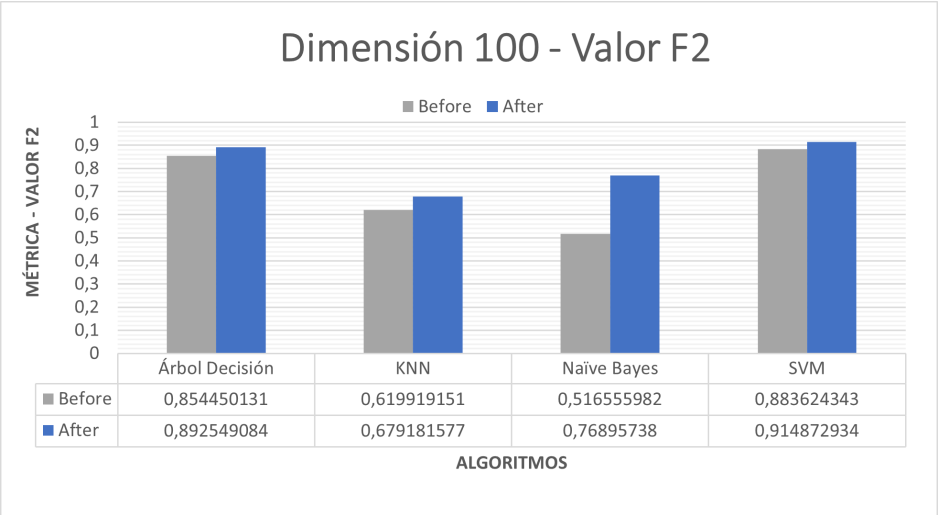


Figura 3.15: Comparación F2 en los Dos Escenarios
Elaborado por: Guerra Cleopatra

De igual manera, para la dimensión de 200 los resultados en comparación al modelo inicial mejoraron para SVM seguidos del algoritmo de árboles de decisión.

Estos hallazgos destacan la importancia de considerar diferentes aspectos, como la dimensión del vector de características y el equilibrio de clases al desarrollar modelos de clasificación de *tweets* más efectivos y precisos.

Es importante mencionar que los valores de *Accuracy* y Especificidad hacen referencia a la clasificación de la clase mayoritaria. La clase mayoritaria está formada por los *tweets* de no política. En esta clase el valor más alto corresponde a la dimensión de 100 con el algoritmo Naïve Bayes. Sin embargo, no es necesariamente cierto que estas métricas garanticen que

Tabla 3.5: Algoritmos Dimensión 200 - Parte I

	Dim 200			
	Arbol de Decision		KNN	
	Before	After	Before	After
Exactitud	0,88876127	0,88429161	0,81772239	0,76165062
Precisión	0,80488650	0,86000224	0,80137348	0,82235835
Recall	0,85135630	0,90037462	0,55608504	0,62865839
F1-score	0,82746949	0,87972548	0,65656784	0,71257962
Otras Métricas				
Sensibilidad	0,85135630	0,90037462	0,55608504	0,62865839
Especificidad	0,90582922	0,87003010	0,93710797	0,87958061
Precisión positiva	0,80488650	0,86000224	0,80137348	0,82235835
Tasa de falsos positivos	0,09417078	0,12996990	0,06289203	0,12041939
Valor F2	0,84163798	0,89199972	0,59234674	0,65973758
Valor F0.5	0,81377015	0,86778445	0,73640777	0,77462351

Tabla 3.6: Algoritmos Dimensión 200 - Parte II

	Dim 200			
	Naïve Bayes		SVM	
	Before	After	Before	After
Exactitud	0,84023431	0,81568088	0,91759031	0,90624484
Precisión	0,81577704	0,79526843	0,87032603	0,89507742
Recall	0,63306452	0,81854367	0,86601906	0,90681339
F1-score	0,71289990	0,80673820	0,86816720	0,90090719
Otras Métricas				
Sensibilidad	0,63306452	0,81854367	0,86601906	0,90681339
Especificidad	0,93476625	0,81314232	0,94112236	0,90574068
Precisión positiva	0,81577704	0,79526843	0,87032603	0,89507742
Tasa de falsos positivos	0,06523375	0,18685768	0,05887764	0,09425932
Valor F2	0,66275232	0,81378026	0,86687704	0,90444164
Valor F0.5	0,77125759	0,79981698	0,86946121	0,89740025

se da una mejor clasificación. Debido a que, en el planteamiento del modelo del escenario 2, con la elección correcta de los hiper-parámetros y la aplicación del nuevo diccionario, se obtiene mejores resultados.

Por otro lado, cabe recalcar que, con el modelo planteado y el uso del nuevo diccionario, el Sistema de Recomendación mejora la recomendación que en modelo inicial. Esto se evidencia en la existencia de mayor número de *tweets* de política que en el modelo inicial.

En cuanto a la interacción de las provincias, es evidente que con el modelo inicial Guayaquil, Pichincha y Manabí acaparan el uso de Twitter. Mientras que en provincias cómo Carchi, Sucumbíos, Zamora Chinchipe la interacción es casi nula. En este escenario algunas provincias tienen valores bajos. Sin embargo, con el nuevo modelo propuesto, si bien es cierto, las tres provincias continuaban encabezando el empleo de esta red social. Otras provincias dejaron de tener valores bajos o nulos y empiezan a tener participación en el ámbito político

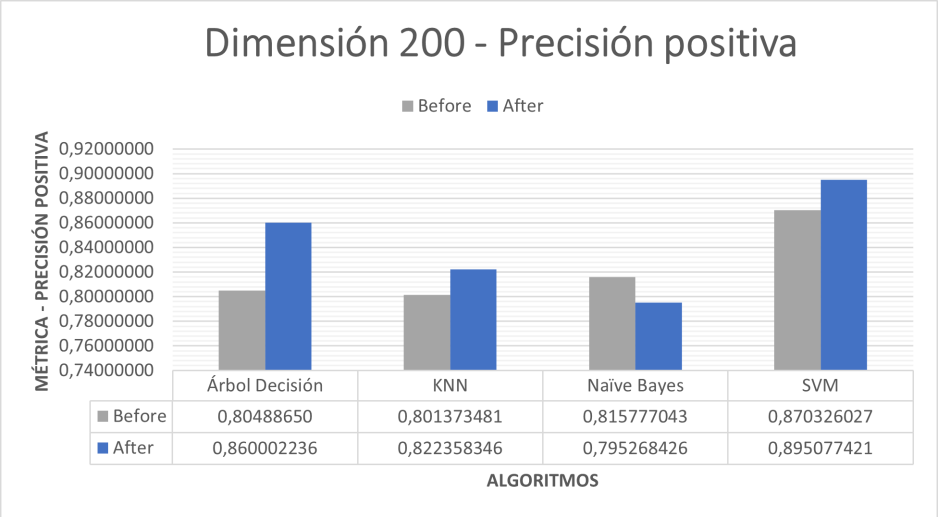


Figura 3.16: Comparación Precisión Positiva en los Dos Escenarios
Elaborado por: Guerra Cleopatra

del país.

4 CONCLUSIONES

- La Revisión de la Literatura fundamentó la base para el planteamiento del modelo propuesto. El análisis de fuentes relacionadas permitió obtener comparaciones significativas en cuanto a metodologías y enfoques propuestos. En investigaciones anteriores la mitigación del sesgo se realizó con el empleo de datos sintéticos, creación de nuevos diccionarios o aumento en los datos de prueba de imágenes por ejemplo. Cuantificar el sesgo es un reto, debido a que existen diferentes tipos de sesgo como: de género, edad, raza, en la selección de datos, implícito, de atributos, algorítmico, social, entre otros.
- La recolección del corpus construyó los cimientos para desarrollar la investigación. Durante la limpieza de los datos se identificó varios *tweets* y *trends* duplicados. La duplicidad ocurrió por la manera de recolectar los datos, se programó que sea durante 24 horas con períodos de 15 minutos. Al no existir tanta recurrencia en el uso de Twitter los datos se duplicaban.
- En el modelo propuesto se empleó un nuevo diccionario, con el propósito de incrementar el número de palabras que guarden una relación con el tópico analizado. Este nuevo diccionario se formó a partir de los clústers iniciales. Se realizó un recalculo de los valores de los centroides en los nuevos clústers. Con el uso de la distancia más corta, aquellas palabras que cumplan con este requisito forman parte del nuevo clúster. De esta manera, en el nuevo clúster se incrementan las palabras. Para el resto de los clústers las palabras disminuyen. Este incremento proporciona que la etiquetación de *tweets* políticos incremente en relación a los *tweets* de política iniciales.
- La mejora con el modelo planteado también se demostró con el uso de vectores de mayor dimensión. En el análisis inicial los *tweets* vectorizados a dimensiones de: 50, 100 y 200 fue menor que el número de *tweets* vectorizados a dimensiones de: 50, 100 y 200 en el modelo final. De las tres dimensiones la que presentó mejores resultados fue en el escenario 2 con la dimensión de 200.

- El modelo propuesto empleó nuevos diccionarios creados a partir de un grupo de datos existentes. Cabe destacar que el uso de nuevos diccionarios proporcionó que la presencia de sesgo disminuya en cierto grado. Fomentando equidad en la recomendación política para los datos analizados.
- Un punto importante es determinar cómo se realizó la medida en la disminución del sesgo en el ámbito político. La manera que el sesgo disminuyó en cierto grado es con el análisis del número de *tweets* etiquetados por provincia en Ecuador. Se observó que en el escenario 1 existía un menor número de *tweets* que después de la aplicación del modelo en el escenario 2. Además, en el escenario 1 existían más provincias que no tenían interacción con Twitter. A diferencia del escenario 2, en el cual provincias cómo Carchi o Sucumbíos que no tenían interacción con Twitter reflejaron una mínima interacción. Pasando, por ejemplo, su estado del número de frecuencias de 1 a 5. Teniendo en cuenta que estas frecuencias están relacionadas con el impacto del *trend* en el número total de *tweets* recolectados.
- Tanto en el escenario 1 y 2, existe un sesgo implícito debido a que la ubicación no estaba de manera implícita en las cuentas de todos los usuarios. En la metadata del *tweet* se observó que ciertas cuentas de Twitter que no tenían lleno el campo de geolocalización en el campo se *userlocation* si registraba la ubicación. Sin embargo, existían *tweets* que no presentaban ningún registro en los dos campos.
- Las métricas para la clase mayoritaria para el caso analizado, no describen la mejor evaluación del modelo en términos del *Accuracy*. El porque el *accuracy* no refleja la mejor evaluación para el modelo propuesto, es debido a que esta medida representa que tan bien los *tweets* no pertenecen al clúster de política y para el problema planteado el interés radica en la clase que fue etiquetada como política. Por tal motivo se analiza las métricas que tengan valor para la clase minoritaria.
- Las métricas empleadas para evaluar los algoritmos están relacionadas directamente con la clase minoritaria. Se emplea *Recall*, *F1-score*, Precisión, F2 y F0.5. Estas métricas dan un buen resultado todas en la dimensión de 200 en el escenario 2 y para el algoritmo de SVM. El segundo algoritmo con mejores resultados es árbol de decisión. *F1-score* proporciona buenos resultados porque es una combinación de la precisión con el *recall*, es útil para encontrar un equilibrio y capturar los resultados de la clase minoritaria. La importancia de f2 pone énfasis en el *recall* y f0.5 en la precisión. Estas dos medidas minimizan los falsos positivos. Por otro lado, SVM presenta un buen

rendimiento en el problema de clasificación en la clase minoritaria; esto debido a su capacidad de encontrar limites de decisión no lineales. Por otro lado, los árboles de decisión muestran resultados que están muy cercanos a SVM gracias a que pueden manejar datos desbalanceados.

4.1 RECOMENDACIONES

La mejora en la clasificación de los *tweets* se puede lograr con el empleo de otras técnicas que clasifiquen los *trends* en los clústers respectivos. Por ejemplo, el uso de DBSCAN, *Agglomerative Hierarchical Clustering* u otros algoritmos que se basan en la densidad como OPTICS.

También, se puede probar con diferentes dimensiones de los vectores como 300 o 400. Esto para capturar de mejor manera la relación entre las palabras.

Una mejora en el modelo sería el uso de redes neuronales convolucionales desde la clasificación. Por otro lado, se puede emplear *Transformers*. Un *Transformers* en un modelo de aprendizaje automático muy útil en la clasificación de palabras y texto. Su funcionamiento es no recursivo y procesa las palabras de manera paralela. Esta forma de trabajo hace que pueda capturar relaciones de mayor alcance y complejidad.

Por otro lado, una mejora en el modelo planteado sería el uso de un corpus de política más amplio. Algo adicional sería la validación de las palabras del clúster seleccionado como político, por un grupo de expertos que garanticen la correcta pertenencia de las palabras al clúster de política.

En el presente trabajo se evidenció que provincias cómo: Carchi, Sucumbíos ó Zamora Chinchipe reflejaron una interacción casi nula con el empleo de Twitter como red social. Sería importante dar a conocer el impacto que el uso de Twitter tiene no sólo en el país, sino en el mundo. Una alternativa sería impartir información en pequeños talleres que eduquen digitalmente a los usuarios. A pesar de que, en muchos lugares de estas provincias existe una brecha tecnológica muy grande en comparación con las grandes provincias y ciudades cómo Guayaquil, Quito, Manta, entre otras.

El diseño de los SR, debe ser transparente de tal manera que el usuario entienda su funcionamiento y en base a que se toma las decisiones. La interfaz que el usuario maneja debe ofrecer controles de preferencias claros y fáciles de usar. Un usuario debe tener la

capacidad de configurar sus preferencias y entender cómo esta configuración afecta a sus recomendaciones.

Un Sistema de Recomendación debe tener implementado un sistema de monitoreo constante con métricas que permitan identificar y corregir el sesgo a medida que surge.

El grupo de desarrolladores de los SR debe estar capacitado sobre la importancia del sesgo en los SR. Además, se debe implementar más filtros que permitan que las recomendaciones sean adecuadas para los usuarios. La diversidad en el grupo de desarrolladores con diferentes perspectivas y experiencias puede contribuir a identificar, comprender y abordar diferentes puntos de vista para reducir la tendencia de sesgos inconscientes.

5 REFERENCIAS BIBLIOGRÁFICAS

[1] P. J. Alcázar Ponce, «Ecuador Estado Digital Ene / 21,» *Mentinno – Innovation Lifetime Value Partners*, pág. 37, 2021.

[2] J. P. Alcazar Ponce, «Ecuador Estado Digital Oct / 21,» *Mentinno Consultores*, pág. 55, 2021. dirección: https://www.mentinno.com/estado-digital-octubre-2021/.

[3] P. Resnick y H. R. Varian, *Recommender systems*. mar. de 1997, vol. 40, págs. 56-58, ISBN: 9783319296579. DOI: 10.1145/245108.245121. dirección: https://dl.acm.org/doi/10.1145/245108.245121.

[4] E. Bárbaro, U. Sust, A. Javier y S. Cuevas, «Sistemas de recomendación semánticos: Una revisión del Estado del Arte Semantic recommendation systems : A State-of-the-Art Survey,» *Revista Cubana de Ciencias Informáticas*, vol. 11, n.º 2, 2017, ISSN: 2227-1899.

[5] R. Baeza-Yates, «Bias on the web,» *Communications of the ACM*, vol. 61, n.º 6, págs. 54-61, 2018, ISSN: 15577317. DOI: 10.1145/3209581.

[6] S. Leavy, «Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning,» págs. 14-16, 2018. DOI: 10.1145/3195570.3195580.

[7] L. Li, Y. Wang y X. Feng, «The psychological effects of microblogging: Examining the relationship between microblogging use and psychological well-being in China,» *Computers in Human Behavior*, vol. 86, págs. 192-200, 2018. DOI: 10.1016/j.chb.2018.04.043.

[8] P. Gupta, P. Kumaraguru, C. Castillo y P. Meier, «Mining Twitter for Trending Topics: A Framework for Temporal Analysis,» *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, n.º 5, pág. 59, 2017. DOI: 10.1145/3078244. dirección: https://doi.org/10.1145/3078244.

[9] H. Kwan, C. Lee, H. Park y S. Moon, «What is Twitter, a Social Network or a News Media?» *In Proceedings of the 19th International Conference on World Wide Web*, vol. 60, n.º 230, págs. 591-600, 2010. DOI: 10.1145/1772690.1772751.

[10] P. Lops, M. De Gemmis y G. Semeraro, «Content-based recommender systems: State of the art and trends,» en *Recommender systems handbook*, Springer, 2011, págs. 73-105.

[11] A.-h. Tan, «Text Mining : The state of the art and the challenges Concept-based,» *Proceedings of the PAKDD 1999 Workshop on*, n.º November 2019, págs. 65-70, 2019. dirección: http://www.mendeley.com/research/text-mining-state-art-challenges-3/.

[12] S. Dang y P. H. Ahmad, «Text Mining : Techniques and its Application Text Mining View project Text Mining: Techniques and its Application,» *IJETI International Journal of Engineering Technology Innovations*, vol. 1, n.º December 2019, 2019, ISSN: 2348-0866. dirección: www.ijeti.com.

[13] R. Feldman y J. Sanger, *The Text Mining Handbook*. New York: United States of America by Cambridge University Press, 2006, ISBN: 9780521836579.

[14] C. D. Manning, P. Raghavan y H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[15] M. Vijaymeena y K. Kavitha, «A Survey on Similarity Measures in Text Mining,» *Machine Learning and Applications: An International Journal (MLAIJ)*, vol. 3, n.º 1, pág. 19, mar. de 2016, M.E. Scholar, Department of Computer Science & Engineering, Nandha Engineering College, Erode-638052, Tamil Nadu, India
Assistant Professor, Department of Computer Science & Engineering, Nandha Engineering College, Erode-638052, Tamil Nadu, India. DOI: 10.5121/mlaij.2016.3103.

[16] A. C. Müller y S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2016.

[17] C. M. P. Pertuz, *Aprendizaje automático y profundo en python*. Ra-Ma Editorial, 2022.

[18] P. C. Sen, M. Hajra y M. Ghosh, «Supervised classification algorithms in machine learning: A survey and review,» en *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, Springer, 2020, págs. 99-111.

[19] B. Mahesh, «Machine learning algorithms-a review,» *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, n.º 1, págs. 381-386, 2020.

[20] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain y A. J. Aljaaf, «A systematic review on supervised and unsupervised machine learning algorithms for data science,» *Supervised and unsupervised learning for data science*, págs. 3-21, 2020.

[21] M. Laskin, D. Yarats, H. Liu et al., «URLB: Unsupervised reinforcement learning benchmark,» *arXiv preprint arXiv:2110.15191*, 2021.

[22] L. Berton, F. Mitsuishi y D. Vega Oliveros, «Analysis of Active Semi-Supervised Learning,» en *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, ép. SAC '23, Tallinn, Estonia: Association for Computing Machinery, 2023, págs. 1122-1129, ISBN: 9781450395175. DOI: 10.1145/3555776.3577621. dirección: https://doi.org/10.1145/3555776.3577621.

[23] Q. Zhang, H. Huang, J. Li, Y. Zhang e Y. Li, «CmpCNN: CMP Modeling with Transfer Learning CNN Architecture,» *ACM Trans. Des. Autom. Electron. Syst.*, vol. 28, n.º 4, mayo de 2023, ISSN: 1084-4309. DOI: 10.1145/3569941. dirección: https://doi.org/10.1145/3569941.

[24] W. Zhang, T. Yao, S. Zhu y A. E. Saddik, «Deep Learning–Based Multimedia Analytics: A Review,» *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, n.º 1s, ene. de 2019, ISSN: 1551-6857. DOI: 10.1145/3279952. dirección: https://doi.org/10.1145/3279952.

[25] V. E. García-Montemayor, «Mortalidad en pacientes en diálisis: importancia y desarrollo de nuevos métodos fiables de predicción,» 2022.

[26] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[27] *¿Qué es el algoritmo de k vecinos más cercanos? | IBM*. dirección: https://la.mathworks.com/discovery/support-vector-machine.html.

[28] *Nodo SVM - Documentación de IBM*, https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-svm-node, Accedido el 8 de mayo de 2023.

[29] J. G. Montalvo, «Investigación económica y datos masivos: mercados, fines sociales y colaboración público-privada* Economic research and big data: Markets, socioeconomic research and,»

[30] *¿Qué es el algoritmo de k vecinos más cercanos? | IBM*. dirección: https://www.ibm.com/mx-es/topics/knn.

[31] J. Misztal-Radecka y B. Indurkhya, «Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems,» *Information Processing and Management*, vol. 58, n.º 3, pág. 102 519, 2021, ISSN: 03064573. DOI: 10.1016/j.ipm.2021.102519. dirección: https://doi.org/10.1016/j.ipm.2021.102519.

[32] B. Friedman y H. Nissenbaum, «Bias in computer systems,» *Computer Ethics*, vol. 14, n.º 3, págs. 215-232, 2017. DOI: 10.4324/9781315259697-23.

[33] K. Orphanou, J. Otterbacher, S. Kleanthous et al., «Mitigating Bias in Algorithmic Systems—A Fish-Eye View,» *ACM Comput. Surv.*, vol. 55, n.º 5, dic. de 2022, ISSN: 0360-0300. DOI: 10.1145/3527152. dirección: https://doi.org/10.1145/3527152.

[34] MongoDB, *MongoDB: La Plataforma De Datos Para Aplicaciones | MongoDB*, https://www.mongodb.com/es.

[35] C. Dasadia y A. Nayak, *MongoDB Cookbook Second Edition*. 2016, pág. 371, ISBN: 9781785289989.

[36] N. Bartley, A. Abeliuk, E. Ferrara y K. Lerman, «Auditing Algorithmic Bias on Twitter,» págs. 65-73, 2021. DOI: 10.1145/3447535.3462491.

[37] R. Ottoni, E. Cunha, G. Magno, P. Bernardina, W. Meira Jr. y V. Almeida, «Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination.,» págs. 323-332, 2018. DOI: 10.1145/3201064.3201081.

[38] D. Abul-Fottouh, M. Y. Song y A. Gruzd, «Examining algorithmic biases in YouTube's recommendations of vaccine videos,» *International Journal of Medical Informatics*, vol. 140, n.º April, pág. 104 175, 2020, ISSN: 18728243. DOI: 10.1016/j.ijmedinf.2020.104175. dirección: https://doi.org/10.1016/j.ijmedinf.2020.104175.

[39] F. Bonchi, S. Hajian, B. Mishra y D. Ramazzotti, «Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining,» *International Journal of Data Science and Analytics*, vol. 3, n.º 1, págs. 2125-2126, 2017, ISSN: 23644168. DOI: 10.1007/s41060-016-0040-z. arXiv: 1510.00552.

[40] W. Wu, P. Michalatos, P. Protopapaps y Z. Yang, «Gender Classification and Bias Mitigation in Facial Images,» *WebSci 2020 - Proceedings of the 12th ACM Conference on Web Science*, págs. 106-114, 2020. DOI: 10.1145/3394231.3397900. arXiv: arXiv: 2007.06141v1.

[41] A. R. Hevner y S. Chatterjee, «Design Research in Information Systems: Theory and Practice,» *Springer*, vol. 2, págs. 1-350, 2010, ISSN: 0742-1222. arXiv: 2003$9.50+0.00. [0742-1222]. dirección: http://link.springer.com/10.1007/978-1-4419-6108-2.

[42] K. Peffers, T. Tuunanen, M. A. Rothenberger y S. Chatterjee, «A design science research methodology for information systems research,» *Journal of Management Information Systems*, vol. 24, n.º 3, págs. 45-77, 2007. DOI: 10.2753/MIS0742-1222240302.

[43] P. Chapman, J. Clinton, R. Kerber et al., «CrispDm 1.0,» Crisp Consortium, inf. téc., 2000, pág. 76.

[44] *Herramienta de proyectos de CRISP-DM - Documentación de IBM | IBM*. dirección: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=modeler-crisp-dm-project-tool.

[45] *Business Understanding - Documentación de IBM | IBM*. dirección: https://www.ibm.com/docs/es/spss-modeler/18.3.0?topic=guide-business-understanding.

[46] *Data Understanding Overview - Documentación de IBM | IBM*. dirección: https://www.ibm.com/docs/es/spss-modeler/18.3.0?topic=understanding-data-overview.

[47] *Data Preparation Overview - Documentación de IBM | IBM*. dirección: https://www.ibm.com/docs/es/spss-modeler/18.3.0?topic=preparation-data-overview.

[48] *Modeling Overview - Documentación de IBM | IBM*. dirección: https://www.ibm.com/docs/es/spss-modeler/18.3.0?topic=modeling-overview.

[49] *Modeling Overview - Documentación de IBM | IBM*. dirección: https://www.ibm.com/docs/es/spss-modeler/18.3.0?topic=evaluation-overview.

[50] *Deployment Overview - Documentación de IBM | IBM*. dirección: https://www.ibm.com/docs/es/spss-modeler/18.3.0?topic=deployment-overview.

[51] M. Petticrew y H. Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide*, Blackwell, ed. 2008, págs. 1-336, ISBN: 1405121106. DOI: 10.1002/9780470754887.

[52] A. F. Godoy Viera, «Machine learning techniques used for text mining,» *Investigación bibliotecológica*, vol. 31, n.º 71, págs. 103-126, 2017, ISSN: 0187-358X.

[53] M. H. A. Hijazi, L. Libin, R. Alfred y F. Coenen, «Bias aware lexicon-based Sentiment Analysis of Malay dialect on social media data: A study on the Sabah Language,» *Proceeding - 2016 2nd International Conference on Science in Information Technology, ICSITech 2016: Information Science for Green Society and Environment*, págs. 356-361, 2017. DOI: 10.1109/ICSITech.2016.7852662.

[54] A. Karim, «Bias-Aware Lexicon-Based Sentiment Analysis,» págs. 845-850, 2016.

[55] A. Fabris, A. Purpura, G. Silvello y G. A. Susto, «Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms,» *Information Processing and Management*, vol. 57, n.º 6, pág. 102 377, 2020, ISSN: 03064573. DOI: 10.1016/j.ipm.2020.102377. dirección: https://doi.org/10.1016/j.ipm.2020.102377.

[56] S. Giorgi, V. Lynn, S. Matz, L. Ungar y H. A. Schwartz, «Correcting Sociodemographic Selection Biases for Accurate Population Prediction from Social Media,» 2019. arXiv: 1911.03855. dirección: http://arxiv.org/abs/1911.03855.

[57] I. Straw y C. Callison-Burch, *Artificial Intelligence in mental health and the biases of language based models*, 2020. DOI: 10.1371/journal.pone.0240376.

[58] Z. Wang, Z. Yu, R. Fan y B. Guo, «Correcting Biases in Online Social Media Data Based on Target Distributions in the Physical World,» *IEEE Access*, vol. 8, págs. 15 256-15 264, 2020, ISSN: 21693536. DOI: 10.1109/aCCESS.2020.2966790.

[59] T. Spinde, L. Rudnitckaia, J. Mitrović et al., «Automated identification of bias inducing words in news articles using linguistic and context-oriented features,» *Information Processing and Management*, vol. 58, n.º 3, pág. 102 505, 2021, ISSN: 03064573. DOI: 10.1016/j.ipm.2021.102505. dirección: https://doi.org/10.1016/j.ipm.2021.102505.

[60] S. Stier, A. Bleier, M. Bonart et al., «As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians Armin,» págs. 193-201, 2018. DOI: 10.4232/1.12992.. dirección: https://dbk.gesis.org/dbksearch/sdesc2.asp?no=6926&db=e&doi=10.4232/1.12992.

[61] L. Boratto y G. S. Eds, *Bias and Social Aspects in Search and Recommendation*. 2020, vol. 1245, pág. 216, ISBN: 978-3-030-52484-5.

[62] L. Dixon, J. Li, J. Sorensen, N. Thain y L. Vasserman, «Measuring and Mitigating Unintended Bias in Text Classification,» *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, págs. 67-73, 2018. DOI: 10.1145/3278721.3278729.

[63] J. Kulshrestha, M. Eslami, J. Messias et al., *Search bias quantification: investigating political bias in social media and web search*. Springer Netherlands, 2019, vol. 22, págs. 188-227, ISBN: 0123456789. DOI: 10.1007/s10791-018-9341-2. dirección: https://doi.org/10.1007/s10791-018-9341-2.

[64] R. Etemadi y J. Lu, «Bias correction in clustering coefficient estimation,» *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, vol. 2018-January, págs. 606-615, 2017. DOI: 10.1109/BigData.2017.8257976.

[65] Jonex Alex, *Infowars*. dirección: https://www.infowars.com/ (visitado 10-07-2022).

[66] Google, *Google News*. dirección: https://news.google.com/topstories?hl=es&gl=ES&ceid=ES:es (visitado 12-05-2022).

[67] BBC World Service, *BBC World Service - The Forum*. dirección: https://www.bbc.co.uk/programmes/p004kln9 (visitado 01-03-2022).

[68] Adelson Jay, *News and Trending Stories Around the Internet | Digg*. dirección: https://digg.com/ (visitado 15-04-2022).

[69] Wechsler Harry y Phillips Jonathan, *color FERET Database | NIST*. dirección: https://www.nist.gov/itl/products-and-services/color-feret-database (visitado 12-07-2022).

[70] Tamara Berg, David Forsyth y the Computer Vision Group at UC Berkeley, *LFW Face Database*. dirección: http://vis-www.cs.umass.edu/lfw/ (visitado 11-07-2022).

[71] G. Lens, *MovieLens 100K Dataset*. dirección: https://grouplens.org/datasets/movielens/100k/ (visitado 13-07-2022).

[72] G. Lens, *Book-Crossing*. dirección: https://grouplens.org/datasets/book-crossing/ (visitado 13-07-2022).

[73] G. M. Group, *The Guardian's News*. dirección: https://www.theguardian.com/uk (visitado 15-07-2022).

[74] T. Program, *TREC Robust Datasets*. dirección: https://ir-datasets.com/trec-robust04.html (visitado 11-07-2022).

[75] C. Hube y B. Fetahu, «Detecting Biased Statements in Wikipedia,» *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, págs. 1779-1786, 2018. DOI: 10.1145/3184558.3191640.

[76] *Twitter Developers*. dirección: https://developer.twitter.com/en/portal/products/free.

[77] *Twitter Developers Search*. dirección: https://api.twitter.com/1.1/search/tweets.json.

[78] *Rate Limit*. dirección: https://developer.twitter.com/en/docs/twitter-api/rate-limits#v2-limits.

[79] *Methods*. dirección: https://developer.twitter.com/en/docs/authentication/guides/v2-authentication-mapping.

[80] *Twitter API Cost*. dirección: https://developer.twitter.com/en/docs/twitter-api.

[81] *Paro Nacional Ecuador Junio 2022*. dirección: https://www.ecuavisa.com/metadatos/-/meta/paro-nacional.

[82] *Twitter Medatada*. dirección: https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet.

[83] S. Bird, E. Klein y E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009, ISBN: 9780596516499.

[84] *Librería Gensim*. dirección: https://pypi.org/project/gensim/.

[85] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado y J. Dean, «Distributed Representations of Words and Phrases and Their Compositionality,» en *Advances in Neural Information Processing Systems*, 2013.

6 ANEXOS

6.1 ANEXO I

El nombre de los archivos en formato .txt que corresponden a las palabras de los 4 diccionarios empleados se muestra en la Figura 6.1. El recuadro negro es para los centroides, el azul para los vectores y el recuadro morado es para las palabras. Este formato se emplea en las otras dos dimensiones.

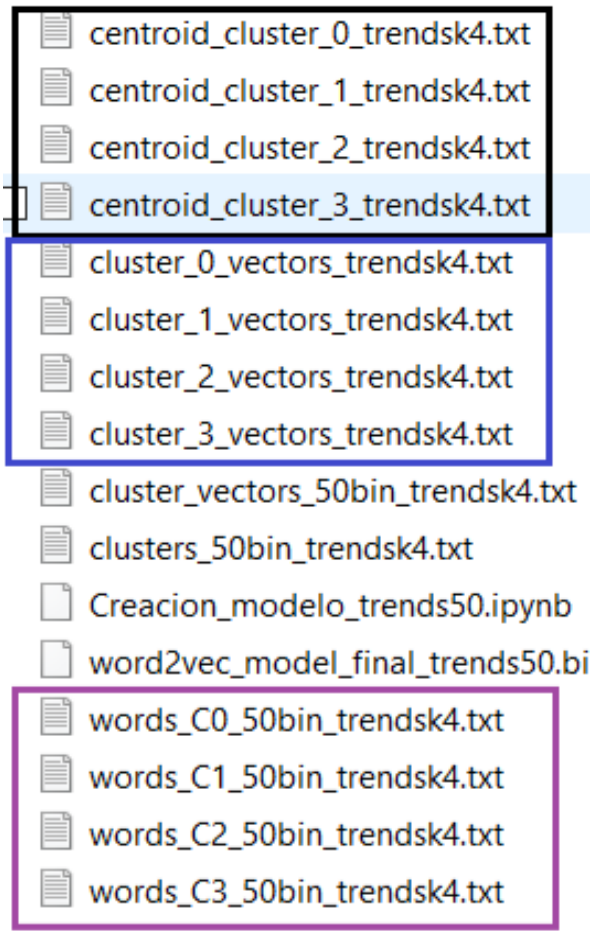


Figura 6.1: Ejemplo Archivos - Dimensión 50
Elaborado por: Guerra Cleopatra

A continuación se muestra una pequeña lista de algunas palabras del clúster 0, 1, 2 y 3.

- ***Clúster 0:*** luis, ecuador, paz, cortez, torres, decreto, martinez, jorge, rosa, alfaro, jefferson, ana, canto, plaza, unidad, parque, suarez, flagrancia, sosa, constitucion, guido, rescalvo, santo, cevallos, joel, sucumbios, basilica, banco, general, tolima, franco, zunio, palacios, hincapie, mundial, belgica, aceptada, hernan, golpismo, domingo, fenocin, madrid, balda, montero, marcela, ortiz, autoatentado, quishpe, jugada, verges, flavio, patricio, etc.
- ***Clúster 1:*** nacional, guayaquil, arbolito, viva, ministro, arboleda, carlos, pichincha, fernando, julio, dia, caicedo, universidad, saquicela, gobernacion, vargas, miguel, imbabura, cuenca, nilson, mejia, pablo, pastaza, dominguez, gracias, cordero, chillos, simon, lara, zubeldia, cobo, majo, flores, ilegal, secuestro, santi, palacio, victor, #muertecruzada, jordy, henry, renuncia, etc.
- ***Clúter 2:*** juan, san, conaie, antonio, jimenez, quezada, valle, atletico, cotopaxi, jose, psc, presidente, catolica, carrillo, maria, patria, garces, alvarado, piero, yaku, ceron, leonidas, defensa, nunez, quito, unes, guillermo, corte, etc.
- ***Clúter 3:*** gobierno, norte, iza, marlon, correa, lasso, joao, ejecutivo, asamblea, riobamba, diego, gabriel, tungurahua, virgilio, cadena, policia, pozo, iglesia, fuerza, fidel, borrero, pedro, carondelet, corpus, rafael, terroristas, pabel, interior, etc.

6.2 ANEXO II

En la Tabla 6.1 se puede observar los tweets de dimensión 50, etiquetados como 1, empleando el diccionario del clúster 3. Considerando el escenario 1.

Tabla 6.1: Tweets Etiquetados Dimensión 50 - Escenario 1

tweet	**label**
#paronacional san pablo provincia santa elena reportan cierre via #paroecuador #paronacionalecuador	0
primer gran resultado #paronacional camaras @ecu funcionen @policiaecuador aparezcan	0
#paronacional perimetral guayaquil momentos #paronacionalecuador #paroecuador #paronacional	0
si protesta pac fica funciona broma invito pueblo ecuatoriano salir calles	0
si paro pago #paronacional	0
subsidio combustibles fosiles debe ser focalizado especializado direccionado unicamente sector	0
@ssbj @lassoguillermo @lenin anda seguir defendiendo nutella inseguridad falta medicina	0
leonidas iza presidente conaie si dia hoy presidente republica da respuestas entonces l	0
#paronacional armas dice manifestante miembro policia nacional localidad s	0
yuca lasso iza aqui ofende dos igual ambos velan intereses p	1
acuerdo anos socialismo pais jamas sali realizar actos bandalicos ser	0
paro nuevo culo nuevo #paronacional	0
momentos reporta cierre via puente #palmar ruta spondylus provincia	0
@lassoguillermo llevo meses trabajo insostenible oportunidades gracias gobierno	0
movilizacion social derecho constitucional lograr correcciones especificas conduccion politic	0
indigenas siempre poder paralizar pais #paronacional	0

Continúa en la siguiente página

Tabla 6.1 – Continuación de la página anterior

tweet	**label**
motivacion semana ninguna encima aguantar gente indigenas paro ata	0
densita situacion ecuador #paronacional	0
ahora existe normalidad provincia #santaelena relacion #paronacional sectores transp	1
llegaron indigenas guayaquil #paronacional	0
estan cerca manifestantes estan preparando terreno clima llego lluvia neblina #guayaquil	0
dirigencia @conaie ecuador momento mas crucial vivio republica prefirieron nulo odio	0
apoyo #paronacional ningun derecho exigido via pacifica vez remover voz	1
#paro paro #paro #paronacional	0
@radio sucre @wqradio ec asi amanece via aurora salitre altura casino #paronacionalecuador	0
hacen paros va trabajar #paronacional	0
dios proteja inocente hoy salga manera democratica alzar voz clamar descontento	1
si marchas limpiara pais delincuentes narcotraficantes apoyaria ir marchar salida #paronacional	0
si indigenas miembros @conaie ecuador derecho protestar carajos quedan derechos l	0
precisamente refiero tuit previo #paronacional	0
#paronacional	0
sueno #paronacional	0
q siguen d alcahuetes dice #lassorevocatoriaya #paronacional	1
distintos manifestantes exigen combustible si sales trabajar preparado #paronacionalecuador	0
toda latinoamerica sucede mismo ninguna institucion orden publico sirve delincuencia c	1
presidente guillermo lasso pronuncio inicio #paronacional mensaje menciono permi	0
#urgente cerrada via #cuenca #loja altura shina nabon comuneros bloquean paso	0

Continúa en la siguiente página

Tabla 6.1 – Continuación de la página anterior

tweet	**label**
#paronacional @ecuarauz hace mencion dia paro cruces agua gesta heroica prole	0
#paronacional falta d medicinas subida d precios canasta basica despidos d medico	0
#lasso genocida camuflado democrata persona q llamaba movilizarse x intereses ahora conden	0
#urgente #ecuador inicio #paronacionalecuador rechazo gobierno guillermo lasso #paronacional	1
#urgente #ecuador inicio #paronacional rechazo medidas economicas gobierno guillermo lasso	1
#urgente #ecuador mas organizaciones sociales suman #paronacional @cidh @onu @defensoriaec	0
bonito decir puede paralizar sufre escasez medicamentos hospitales cua	0
@crudoecuador obvio #paronacional si tragas recurso publico	0
inicio paro #paronacional	0
buenas intenciones lleno camino infierno sirven putas intenciones presidente	0
va traer mucha cola nivel internacional desgaste gobierno comienza curso exterior	1
vamos aguantar anos mas rectifica v @lassoguillermo @aborrerovega @panchojimenezs #paronacional	0

La Tabla 6.2 se puede observar los tweets con dimensión 100, etiquetados como 1, empleando el diccionario del clúster 3. Considerando el escenario 1.

Tabla 6.2: Tweets Etiquetados Dimensión 100 - Escenario 1

tweet	**label**
#paronacional san pablo provincia santa elena reportan cierre via #paroecuador #paronacionalecuador	0
primer gran resultado #paronacional camaras @ecu funcionen @policiaecuador aparezcan	0
#paronacional perimetral guayaquil momentos #paronacionalecuador #paroecuador #paronacional	0
si protesta pac fica funciona broma invito pueblo ecuatoriano salir calles	0
si paro pago #paronacional	1
subsidio combustibles fosiles debe ser focalizado especializado direccionado unicamente sector	0
@ssbj @lassoguillermo @lenin anda seguir defendiendo nutella inseguridad falta medicina	0
leonidas iza presidente conaie si dia hoy presidente republica da respuestas entonces l	0
#paronacional armas dice manifestante miembro policia nacional localidad s	0
yuca lasso iza aqui ofende dos igual ambos velan intereses p	1
acuerdo anos socialismo pais jamas sali realizar actos bandalicos ser	0
paro nuevo culo nuevo #paronacional	1
momentos reporta cierre via puente #palmar ruta spondylus provincia	0
@lassoguillermo llevo meses trabajo insostenible oportunidades gracias gobierno	0
movilizacion social derecho constitucional lograr correcciones especificas conduccion politic	0
indigenas siempre poder paralizar pais #paronacional	0
motivacion semana ninguna encima aguantar gente indigenas paro ata	0
densita situacion ecuador #paronacional	1

Continúa en la siguiente página

Tabla 6.2 – Continuación de la página anterior

tweet	**label**
ahora existe normalidad provincia #santaelena relacion #paronacional sectores transp	0
llegaron indigenas guayaquil #paronacional	0
estan cerca manifestantes estan preparando terreno clima llego lluvia neblina #guayaquil	0
dirigencia @conaie ecuador momento mas crucial vivio republica prefirieron nulo odio	1
apoyo #paronacional ningun derecho exigido via pacifica vez remover voz	1
#paro paro #paro #paronacional	1
@radio sucre @wqradio ec asi amanece via aurora salitre altura casino #paronacionalecuador	0
hacen paros va trabajar #paronacional	1
dios proteja inocente hoy salga manera democratica alzar voz clamar descontento	0
si marchas limpiara pais delincuentes narcotraficantes apoyaria ir marchar salida #paronacional	1
si indigenas miembros @conaie ecuador derecho protestar carajos quedan derechos l	1
precisamente refiero tuit previo #paronacional	1
#paronacional	1
sueno #paronacional	1
q siguen d alcahuetes dice #lassorevocatoriaya #paronacional	1
distintos manifestantes exigen combustible si sales trabajar preparado #paronacionalecuador	0
toda latinoamerica sucede mismo ninguna institucion orden publico sirve delincuencia c	0
presidente guillermo lasso pronuncio inicio #paronacional mensaje menciono permi	0
#urgente cerrada via #cuenca #loja altura shina nabon comuneros bloquean paso	0
#paronacional @ecuarauz hace mencion dia paro cruces agua gesta heroica prole	0

Continúa en la siguiente página

Tabla 6.2 – Continuación de la página anterior

tweet	**label**
#paronacional falta d medicinas subida d precios canasta basica despidos d medico	1
#lasso genocida camuflado democrata persona q llamaba movilizarse x intereses ahora conden	1
#urgente #ecuador inicio #paronacionalecuador rechazo gobierno guillermo lasso #paronacional	0
#urgente #ecuador inicio #paronacional rechazo medidas economicas gobierno guillermo lasso	0
#urgente #ecuador mas organizaciones sociales suman #paronacional @cidh @onu @defensoriaec	0
bonito decir puede paralizar sufre escasez medicamentos hospitales cua	0
@crudoecuador obvio #paronacional si tragas recurso publico	1
inicio paro #paronacional	1
buenas intenciones lleno camino infierno sirven putas intenciones presidente	0
va traer mucha cola nivel internacional desgaste gobierno comienza curso exterior	0
vamos aguantar anos mas rectifica v @lassoguillermo @aborrerovega @panchojimenezs #paronacional	1

La Tabla 6.3 se puede observar los tweets con dimensión 200, etiquetados como 1, empleando el diccionario del clúster 3. Considerando el escenario 1.

Tabla 6.3: Tweets Etiquetados Dimensión 200 - Escenario 1

tweet	**label**
#paronacional san pablo provincia santa elena reportan cierre via #paroecuador #paronacionalecuador	0
primer gran resultado #paronacional camaras @ecu funcionen @policiaecuador aparezcan	1
#paronacional perimetral guayaquil momentos #paronacionalecuador #paroecuador #paronacional	1
si protesta pac fica funciona broma invito pueblo ecuatoriano salir calles	0
si paro pago #paronacional	1
subsidio combustibles fosiles debe ser focalizado especializado direccionado unicamente sector	0
@ssbj @lassoguillermo @lenin anda seguir defendiendo nutella inseguridad falta medicina	0
leonidas iza presidente conaie si dia hoy presidente republica da respuestas entonces l	1
#paronacional armas dice manifestante miembro policia nacional localidad s	1
yuca lasso iza aqui ofende dos igual ambos velan intereses p	0
acuerdo anos socialismo pais jamas sali realizar actos bandalicos ser	0
paro nuevo culo nuevo #paronacional	1
momentos reporta cierre via puente #palmar ruta spondylus provincia	0
@lassoguillermo llevo meses trabajo insostenible oportunidades gracias gobierno	0
movilizacion social derecho constitucional lograr correcciones especificas conduccion politic	0
indigenas siempre poder paralizar pais #paronacional	0
motivacion semana ninguna encima aguantar gente indigenas paro ata	0
densita situacion ecuador #paronacional	0

Continúa en la siguiente página

Tabla 6.3 – Continuación de la página anterior

tweet	**label**
ahora existe normalidad provincia #santaelena relacion #paronacional sectores transp	1
llegaron indigenas guayaquil #paronacional	0
estan cerca manifestantes estan preparando terreno clima llego lluvia neblina #guayaquil	0
dirigencia @conaie ecuador momento mas crucial vivio republica prefirieron nulo odio	0
apoyo #paronacional ningun derecho exigido via pacifica vez remover voz	0
#paro paro #paro #paronacional	1
@radio sucre @wqradio ec asi amanece via aurora salitre altura casino #paronacionalecuador	0
hacen paros va trabajar #paronacional	1
dios proteja inocente hoy salga manera democratica alzar voz clamar descontento	0
si marchas limpiara pais delincuentes narcotraficantes apoyaria ir marchar salida #paronacional	1
si indigenas miembros @conaie ecuador derecho protestar carajos quedan derechos l	0
precisamente refiero tuit previo #paronacional	1
#paronacional	1
sueno #paronacional	1
q siguen d alcahuetes dice #lassorevocatoriaya #paronacional	1
distintos manifestantes exigen combustible si sales trabajar preparado #paronacionalecuador	0
toda latinoamerica sucede mismo ninguna institucion orden publico sirve delincuencia c	0
presidente guillermo lasso pronuncio inicio #paronacional mensaje menciono permi	1
#urgente cerrada via #cuenca #loja altura shina nabon comuneros bloquean paso	0
#paronacional @ecuarauz hace mencion dia paro cruces agua gesta heroica prole	0

Continúa en la siguiente página

Tabla 6.3 – Continuación de la página anterior

tweet	**label**
#paronacional falta d medicinas subida d precios canasta basica despidos d medico	1
#lasso genocida camuflado democrata persona q llamaba movilizarse x intereses ahora conden	0
#urgente #ecuador inicio #paronacionalecuador rechazo gobierno guillermo lasso #paronacional	1
#urgente #ecuador inicio #paronacional rechazo medidas economicas gobierno guillermo lasso	1
#urgente #ecuador mas organizaciones sociales suman #paronacional @cidh @onu @defensoriaec	1
bonito decir puede paralizar sufre escasez medicamentos hospitales cua	0
@crudoecuador obvio #paronacional si tragas recurso publico	1
inicio paro #paronacional	1
buenas intenciones lleno camino infierno sirven putas intenciones presidente	1
va traer mucha cola nivel internacional desgaste gobierno comienza curso exterior	0
vamos aguantar anos mas rectifica v @lassoguillermo @aborrerovega @panchojimenezs #paronacional	1

6.3 ANEXO III

En la Figura 6.2 se puede observar el vector del centroide del clúster 3. Este vector se obtiene a partir de los vectores de las palabras que se pueden vizualizar en la Figura 6.3. Para los dos casos se muestra con dimensión 50. Estos cálculos se realizan para las dimensiones de 100 y 200.

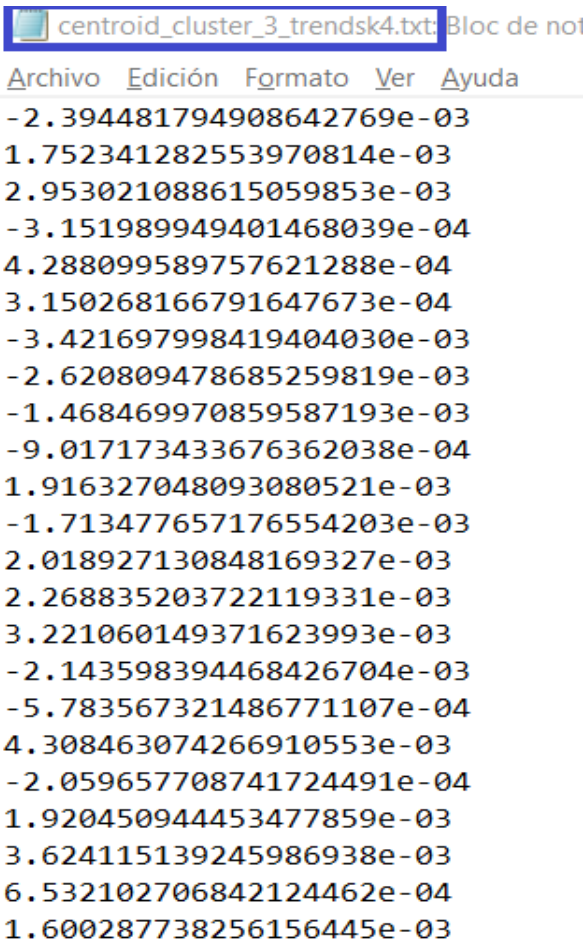


Figura 6.2: Ejemplo Vector Centroide Clúster 3 - Dimensión 50
Elaborado por: Guerra Cleopatra

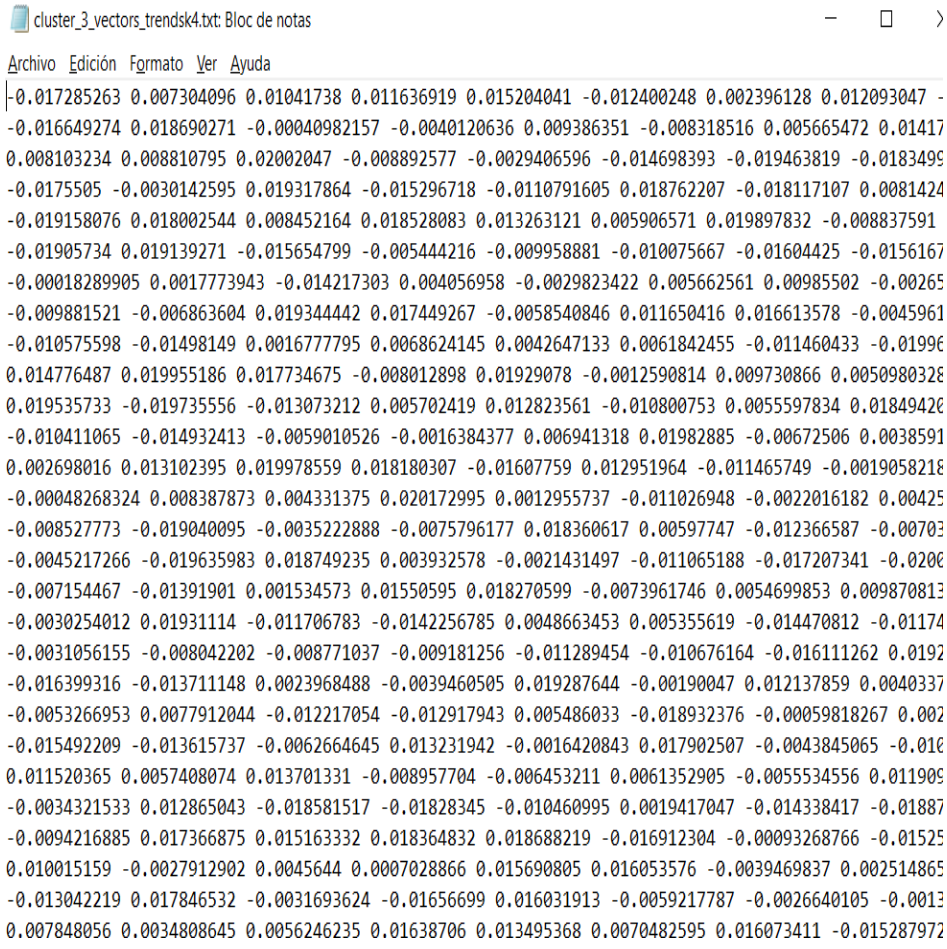


Figura 6.3: Ejemplo Vectores de Palabras Clúster 3 - Dimensión 50
Elaborado por: Guerra Cleopatra

6.4 ANEXO IV

Algunos ejemplos de analogías se muestran a continuación en las Figuras 6.4, 6.5, 6.6 y 6.7.

```python
# Ejemplo de uso
word_a = "vicepresidente"
word_b = "borrero"
word_c = "lasso"

analogy_word = find_analogy(word_a, word_b, word_c)
print(f"{word_a} - {word_b} + {word_c} = {analogy_word}")
```

```
vicepresidente - borrero + lasso = inepto
```

(a) Ejemplo Analogía

```python
# Ejemplo de uso
word_a = "asambleista"
word_b = "almeida"
word_c = "villavicencio"

analogy_word = find_analogy(word_a, word_b, word_c)
print(f"{word_a} - {word_b} + {word_c} = {analogy_word}")
```

```
asambleista - almeida + villavicencio = @adriancoronel
```

(b) Ejemplo Analogía

```python
# Ejemplo de uso
word_a = "presidente"
word_b = "lasso"
word_c = "iza"

analogy_word = find_analogy(word_a, word_b, word_c)
print(f"{word_a} - {word_b} + {word_c} = {analogy_word}")
```

```
presidente - lasso + iza = conaie
```

(c) Ejemplo Analogía

Figura 6.4: Ejemplo Analogías
Elaborado por: Guerra Cleopatra

```python
# Ejemplo de uso
word_a = "presidente"
word_b = "lasso"
word_c = "otto"

analogy_word = find_analogy(word_a, word_b, word_c)
print(f"{word_a} - {word_b} + {word_c} = {analogy_word}")
```

```
presidente - lasso + otto = laso
```

(a) Ejemplo de Analogía

```python
print(model.wv.most_similar('lasso',topn=10))
```

```
[('correismo', 0.8060368895530701), ('votos', 0.764366626739502), ('lenin', 0.7608745694160461), ('banquero', 0.7582812905311584), ('carondelet', 0.7580767273902893), ('destitucion', 0.756513237953186), ('dictador', 0.7535768747329712), ('nebot', 0.740649521350860
6), ('putrefacto', 0.736994743347168), ('presidencia', 0.7337113618850708)]
```

(b) Ejemplo Analogía

```python
vector = model.wv["#muertecruzada"]
# Calcula cosine similarities
similitud = model.wv.cosine_similarities(vector, model.wv.vectors)
# busca los indices mas similares
#obtiene los índices de los 10 vectores más similares con los valores de similitud de coseno más alto
most_similar_indices = similitud.argsort()[:-20:-1]

# Imprime las palalbras mas similares y su cosine
print("Vecinos Proximos:")
for index in most_similar_indices:
    word = model.wv.index_to_key[index]
    similarity = similitud[index]
    print(word, similarity)
```

```
Vecinos Proximos:
#muertecruzada 1.0
#juiciopolitico 0.9536774
decreto 0.95244485
ejecutivo 0.9283466
disolver 0.9258356
#lasso 0.9250083
derogar 0.91811025
legislativo 0.9169778
destitucion 0.91543514
congreso 0.915263
censurar 0.9147578
#juiciopoliticoalasso 0.9144146
decretos 0.9138317
constitucional 0.9110544
#muertecruzadaec 0.9069113
#guillermolasso 0.902106
disolucion 0.90139544
#asambleanacional 0.9009158
firmaste 0.900595
```

(c) Ejemplo Analogía

Figura 6.5: Ejemplo Analogías
Elaborado por: Guerra Cleopatra

```
# Imprime las palalbras mas similares y su cosine
print("Vecinos Proximos:")
for index in most_similar_indices:
    word = model.wv.index_to_key[index]
    similarity = similitud[index]
    print(word, similarity)
```

```
Vecinos Proximos:
lasso 1.0000001
eduardo 0.41692266
lautaro 0.4061287
escobar 0.40305668
ayrton 0.39137018
millones 0.3912559
bogum 0.36718524
magallanes 0.36706853
samuel 0.3445615
duran 0.34252128
abre 0.33977985
#mashirafael 0.33893168
iglesia 0.32762852
vayan 0.32675216
diesel 0.32482666
provincializacion 0.323106
viva 0.31483164
enner 0.308142
bono 0.30623844
```

(a) Ejemplo de Analogía

```
# Imprime las palalbras mas similares y su cosine
print("Vecinos Proximos:")
for index in most_similar_indices:
    word = model.wv.index_to_key[index]
    similarity = similitud[index]
    print(word, similarity)
```

```
Vecinos Proximos:
#paronacional 1.0000001
armas 0.41455555
aceptada 0.4013456
courtois 0.3954042
cynthia 0.39360553
boston 0.393133
tatum 0.37160847
rescalvo 0.37148684
gallese 0.36953485
america 0.3600577
paiva 0.35898542
#juicioilegal 0.34568587
grabois 0.34542757
nutella 0.34113792
dennis 0.34058374
asqueroso 0.33422062
tambillo 0.32931837
villa 0.32494867
hervas 0.3073669
```

(b) Ejemplo Analogía

Figura 6.6: Ejemplo Analogías
Elaborado por: Guerra Cleopatra

```python
# Imprime las palalbras mas similares y su cosine
print("Vecinos Proximos:")
for index in most_similar_indices:
    word = model.wv.index_to_key[index]
    similarity = similitud[index]
    print(word, similarity)
```

```
Vecinos Proximos:
correa 1.0
bunny 0.4337977
pleno 0.4057607
vandalos 0.37683794
eitel 0.37295157
moises 0.36843124
bangtan 0.35883826
bedon 0.34572795
errores 0.34348917
#ecuadorsos 0.3409757
colegio 0.33553675
independiente 0.33355567
covid 0.3230678
nunez 0.31335473
nacion 0.31250328
presidente 0.30979353
parecen 0.30840036
presidencia 0.30645123
loor 0.30430955
```

(a) Ejemplo Analogía

Figura 6.7: Ejemplo Analogías
Elaborado por: Guerra Cleopatra

6.5 ANEXO V

En la Tabla 6.4 de este anexo se muestra un ejemplo de las frecuencias de algunos trends en relación al total de tweets recolectados.

Tabla 6.4: Palabras y Frecuencias

Palabra	Frecuencia
iza	4317
ecuador	9811
ecu	16690
rc	10574
verguenza	1573
id	31059
cree	1269
abogado	302
patria	635
@martinminguchi	496
pendejo	422
vende	473
@mashirafael	2587
rafael	3049
ignorante	321
pillo	297
pueblo	2261
oficial	1991
@ecuainm	700
viva	493
@marcelaguinaga	250
@ffaaecuador	474
banda	363
@laguerreraecu	119
ffaa	567
@lassoguillermo	7392
lasso	11642
guillermo	8262

Continúa en la siguiente página

Tabla 6.4 – continuación de la página anterior

Palabra	Frecuencia
@ecuavisainforma	724
policia	1193
@ecuavisa	861
militar	348
pol	6396
ex	9136
debe	3501
ratas	710
rata	2020
ladron	685
@capizapataec	246
@chonilloec	219
permitir	167
quieren	759
seguridad	992
justicia	597
@dianasalazarm	410
@fiscaliaecuador	238
fiscal	701
delito	221
fiscalia	382
diana	805
nido	683
estan	2141
asamblea	5800
don	1971
borrero	346
plan	1035
pais	4537
politica	1483
nuevas	322
dignidad	198
politicas	168
funciones	159
alcaldes	403

Continúa en la siguiente página

Tabla 6.4 – continuación de la página anterior

Palabra	Frecuencia
@policiaecuador	640
juicio	2097
@oea	92
calo	1340
gobierno	1851
irse	461
puso	258
@ladataec	1271
social	1217
fernando	1493
lacras	114
lacra	1533

6.6 ANEXO VI

En la Tabla 6.5 se puede observar los tweets con dimensión 50, etiquetados como 1, empleando el nuevo diccionario del clúster 3. Considerando el escenario 2.

Tabla 6.5: Tweets Etiquetados Dimensión 50 - Escenario 2

tweet	**label**
#paronacional san pablo provincia santa elena reportan cierre via #paroecuador #paronacionalecuador	0
primer gran resultado #paronacional camaras @ecu funcionen @policiaecuador aparezcan	0
#paronacional perimetral guayaquil momentos #paronacionalecuador #paroecuador #paronacional	0
si protesta pac fica funciona broma invito pueblo ecuatoriano salir calles	0
si paro pago #paronacional	0
subsidio combustibles fosiles debe ser focalizado especializado direccionado unicamente sector	0
@ssbj @lassoguillermo @lenin anda seguir defendiendo nutella inseguridad falta medicina	0
leonidas iza presidente conaie si dia hoy presidente republica da respuestas entonces l	1
#paronacional armas dice manifestante miembro policia nacional localidad s	1
yuca lasso iza aqui ofende dos igual ambos velan intereses p	1
acuerdo anos socialismo pais jamas sali realizar actos bandalicos ser	0
paro nuevo culo nuevo #paronacional	0
momentos reporta cierre via puente #palmar ruta spondylus provincia	0
@lassoguillermo llevo meses trabajo insostenible oportunidades gracias gobierno	1
movilizacion social derecho constitucional lograr correcciones especificas conduccion politic	0
indigenas siempre poder paralizar pais #paronacional	0

Continúa en la siguiente página

Tabla 6.5 – Continuación de la página anterior

tweet	**label**
motivacion semana ninguna encima aguantar gente indigenas paro ata	0
densita situacion ecuador #paronacional	0
ahora existe normalidad provincia #santaelena relacion #paronacional sectores transp	1
llegaron indigenas guayaquil #paronacional	0
estan cerca manifestantes estan preparando terreno clima llego lluvia neblina #guayaquil	1
dirigencia @conaie ecuador momento mas crucial vivio republica prefirieron nulo odio	1
apoyo #paronacional ningun derecho exigido via pacifica vez remover voz	0
#paro paro #paro #paronacional	0
@radio sucre @wqradio ec asi amanece via aurora salitre altura casino #paronacionalecuador	0
hacen paros va trabajar #paronacional	0
dios proteja inocente hoy salga manera democratica alzar voz clamar descontento	1
si marchas limpiara pais delincuentes narcotraficantes apoyaria ir marchar salida #paronacional	0
si indigenas miembros @conaie ecuador derecho protestar carajos quedan derechos l	0
precisamente refiero tuit previo #paronacional	0
#paronacional	0
sueno #paronacional	0
q siguen d alcahuetes dice #lassorevocatoriaya #paronacional	1
distintos manifestantes exigen combustible si sales trabajar preparado #paronacionalecuador	0
toda latinoamerica sucede mismo ninguna institucion orden publico sirve delincuencia c	1
presidente guillermo lasso pronuncio inicio #paronacional mensaje menciono permi	1
#urgente cerrada via #cuenca #loja altura shina nabon comuneros bloquean paso	1

Continúa en la siguiente página

Tabla 6.5 – Continuación de la página anterior

tweet	**label**
#paronacional @ecuarauz hace mencion dia paro cruces agua gesta heroica prole	0
#paronacional falta d medicinas subida d precios canasta basica despidos d medico	0
#lasso genocida camuflado democrata persona q llamaba movilizarse x intereses ahora conden	0
#urgente #ecuador inicio #paronacionalecuador rechazo gobierno guillermo lasso #paronacional	1
#urgente #ecuador inicio #paronacional rechazo medidas economicas gobierno guillermo lasso	1
#urgente #ecuador mas organizaciones sociales suman #paronacional @cidh @onu @defensoriaec	1
bonito decir puede paralizar sufre escasez medicamentos hospitales cua	0
@crudoecuador obvio #paronacional si tragas recurso publico	0
inicio paro #paronacional	0
buenas intenciones lleno camino infierno sirven putas intenciones presidente	0
va traer mucha cola nivel internacional desgaste gobierno comienza curso exterior	1
vamos aguantar anos mas rectifica v @lassoguillermo @aborrerovega @panchojimenezs #paronacional	0

La Tabla 6.6 se puede observar los tweets con dimensión 100, etiquetados como 1, empleando el nuevo diccionario del clúster 3. Considerando el escenario 2.

Tabla 6.6: Tweets Etiquetados Dimensión 100 - Escenario 2

tweet	**label**
#paronacional san pablo provincia santa elena reportan cierre via #paroecuador #paronacionalecuador	0
primer gran resultado #paronacional camaras @ecu funcionen @policiaecuador aparezcan	0
#paronacional perimetral guayaquil momentos #paronacionalecuador #paroecuador #paronacional	0
si protesta pac fica funciona broma invito pueblo ecuatoriano salir calles	0
si paro pago #paronacional	1
subsidio combustibles fosiles debe ser focalizado especializado direccionado unicamente sector	0
@ssbj @lassoguillermo @lenin anda seguir defendiendo nutella inseguridad falta medicina	0
leonidas iza presidente conaie si dia hoy presidente republica da respuestas entonces l	0
#paronacional armas dice manifestante miembro policia nacional localidad s	0
yuca lasso iza aqui ofende dos igual ambos velan intereses p	0
acuerdo anos socialismo pais jamas sali realizar actos bandalicos ser	0
paro nuevo culo nuevo #paronacional	1
momentos reporta cierre via puente #palmar ruta spondylus provincia	0
@lassoguillermo llevo meses trabajo insostenible oportunidades gracias gobierno	0
movilizacion social derecho constitucional lograr correcciones especificas conduccion politic	0
indigenas siempre poder paralizar pais #paronacional	0
motivacion semana ninguna encima aguantar gente indigenas paro ata	0
densita situacion ecuador #paronacional	1

Continúa en la siguiente página

Tabla 6.6 – Continuación de la página anterior

tweet	**label**
ahora existe normalidad provincia #santaelena relacion #paronacional sectores transp	0
llegaron indigenas guayaquil #paronacional	0
estan cerca manifestantes estan preparando terreno clima llego lluvia neblina #guayaquil	0
dirigencia @conaie ecuador momento mas crucial vivio republica prefirieron nulo odio	1
apoyo #paronacional ningun derecho exigido via pacifica vez remover voz	1
#paro paro #paro #paronacional	1
@radio sucre @wqradio ec asi amanece via aurora salitre altura casino #paronacionalecuador	0
hacen paros va trabajar #paronacional	1
dios proteja inocente hoy salga manera democratica alzar voz clamar descontento	0
si marchas limpiara pais delincuentes narcotraficantes apoyaria ir marchar salida #paronacional	1
si indigenas miembros @conaie ecuador derecho protestar carajos quedan derechos l	1
precisamente refiero tuit previo #paronacional	1
#paronacional	1
sueno #paronacional	1
q siguen d alcahuetes dice #lassorevocatoriaya #paronacional	1
distintos manifestantes exigen combustible si sales trabajar preparado #paronacionalecuador	0
toda latinoamerica sucede mismo ninguna institucion orden publico sirve delincuencia c	0
presidente guillermo lasso pronuncio inicio #paronacional mensaje menciono permi	0
#urgente cerrada via #cuenca #loja altura shina nabon comuneros bloquean paso	0
#paronacional @ecuarauz hace mencion dia paro cruces agua gesta heroica prole	1

Continúa en la siguiente página

Tabla 6.6 – Continuación de la página anterior

tweet	**label**
#paronacional falta d medicinas subida d precios canasta basica despidos d medico	1
#lasso genocida camuflado democrata persona q llamaba movilizarse x intereses ahora conden	1
#urgente #ecuador inicio #paronacionalecuador rechazo gobierno guillermo lasso #paronacional	0
#urgente #ecuador inicio #paronacional rechazo medidas economicas gobierno guillermo lasso	0
#urgente #ecuador mas organizaciones sociales suman #paronacional @cidh @onu @defensoriaec	0
bonito decir puede paralizar sufre escasez medicamentos hospitales cua	1
@crudoecuador obvio #paronacional si tragas recurso publico	1
inicio paro #paronacional	1
buenas intenciones lleno camino infierno sirven putas intenciones presidente	0
va traer mucha cola nivel internacional desgaste gobierno comienza curso exterior	0
vamos aguantar anos mas rectifica v @lassoguillermo @aborrerovega @panchojimenezs #paronacional	1

En la Tabla 6.7 se puede observar los tweets con dimensión 200, etiquetados como 1, empleando el nuevo diccionario del clúster 3. Considerando el escenario 2.

Tabla 6.7: Tweets Etiquetados Dimensión 200 - Escenario 2

tweet	**label**
#paronacional san pablo provincia santa elena reportan cierre via #paroecuador #paronacionalecuador	1
primer gran resultado #paronacional camaras @ecu funcionen @policiaecuador aparezcan	1
#paronacional perimetral guayaquil momentos #paronacionalecuador #paroecuador #paronacional	0
si protesta pac fica funciona broma invito pueblo ecuatoriano salir calles	1
si paro pago #paronacional	1
subsidio combustibles fosiles debe ser focalizado especializado direccionado unicamente sector	0
@ssbj @lassoguillermo @lenin anda seguir defendiendo nutella inseguridad falta medicina	1
leonidas iza presidente conaie si dia hoy presidente republica da respuestas entonces l	1
#paronacional armas dice manifestante miembro policia nacional localidad s	0
yuca lasso iza aqui ofende dos igual ambos velan intereses p	0
acuerdo anos socialismo pais jamas sali realizar actos bandalicos ser	1
paro nuevo culo nuevo #paronacional	1
momentos reporta cierre via puente #palmar ruta spondylus provincia	1
@lassoguillermo llevo meses trabajo insostenible oportunidades gracias gobierno	1
movilizacion social derecho constitucional lograr correcciones especificas conduccion politic	1
indigenas siempre poder paralizar pais #paronacional	0
motivacion semana ninguna encima aguantar gente indigenas paro ata	0
densita situacion ecuador #paronacional	0

Continúa en la siguiente página

Tabla 6.7 – Continuación de la página anterior

tweet	**label**
ahora existe normalidad provincia #santaelena relacion #paronacional sectores transp	0
llegaron indigenas guayaquil #paronacional	0
estan cerca manifestantes estan preparando terreno clima llego lluvia neblina #guayaquil	1
dirigencia @conaie ecuador momento mas crucial vivio republica prefirieron nulo odio	0
apoyo #paronacional ningun derecho exigido via pacifica vez remover voz	0
#paro paro #paro #paronacional	1
@radio sucre @wqradio ec asi amanece via aurora salitre altura casino #paronacionalecuador	1
hacen paros va trabajar #paronacional	1
dios proteja inocente hoy salga manera democratica alzar voz clamar descontento	1
si marchas limpiara pais delincuentes narcotraficantes apoyaria ir marchar salida #paronacional	0
si indigenas miembros @conaie ecuador derecho protestar carajos quedan derechos l	1
precisamente refiero tuit previo #paronacional	0
#paronacional	1
sueno #paronacional	1
q siguen d alcahuetes dice #lassorevocatoriaya #paronacional	1
distintos manifestantes exigen combustible si sales trabajar preparado #paronacionalecuador	1
toda latinoamerica sucede mismo ninguna institucion orden publico sirve delincuencia c	1
presidente guillermo lasso pronuncio inicio #paronacional mensaje menciono permi	0
#urgente cerrada via #cuenca #loja altura shina nabon comuneros bloquean paso	1
#paronacional @ecuarauz hace mencion dia paro cruces agua gesta heroica prole	0

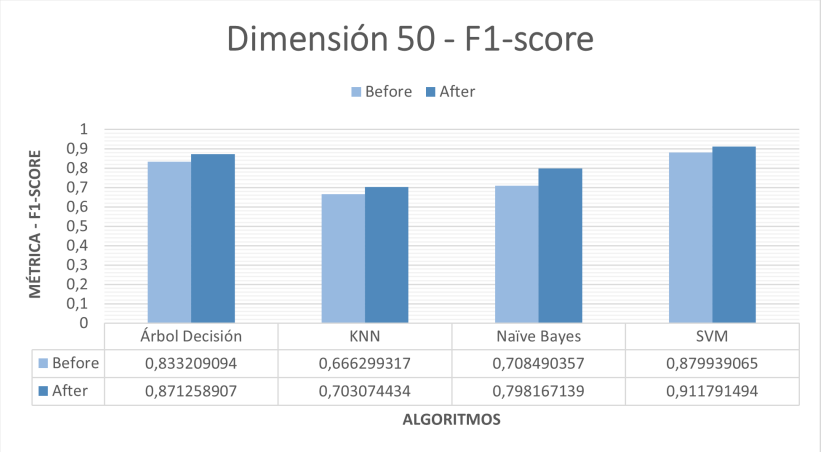
Continúa en la siguiente página

Tabla 6.7 – Continuación de la página anterior

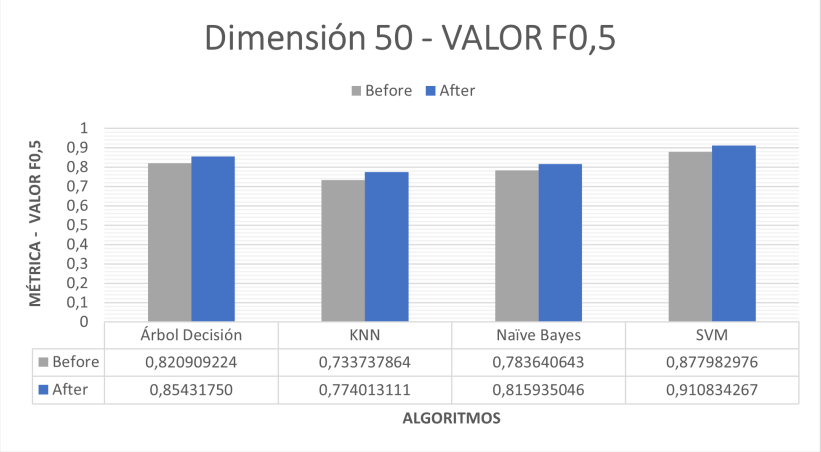
tweet	**label**
#paronacional falta d medicinas subida d precios canasta basica despidos d medico	1
#lasso genocida camuflado democrata persona q llamaba movilizarse x intereses ahora conden	0
#urgente #ecuador inicio #paronacionalecuador rechazo gobierno guillermo lasso #paronacional	0
#urgente #ecuador inicio #paronacional rechazo medidas economicas gobierno guillermo lasso	1
#urgente #ecuador mas organizaciones sociales suman #paronacional @cidh @onu @defensoriaec	0
bonito decir puede paralizar sufre escasez medicamentos hospitales cua	1
@crudoecuador obvio #paronacional si tragas recurso publico	1
inicio paro #paronacional	1
buenas intenciones lleno camino infierno sirven putas intenciones presidente	1
va traer mucha cola nivel internacional desgaste gobierno comienza curso exterior	1
vamos aguantar anos mas rectifica v @lassoguillermo @aborrerovega @panchojimenezs #paronacional	0

6.7 ANEXO VII

Este Anexo esta formado por los resultados de las gráficas obtenidas a partir de las medidas de los algoritmos comparando el escenario 1 y el escenario 2. En primer lugar se colocan las gráficas de la clase minoritaria 6.8, 6.9. Finalmente, se coloca las gráficas de la clase mayoritaria 6.10.



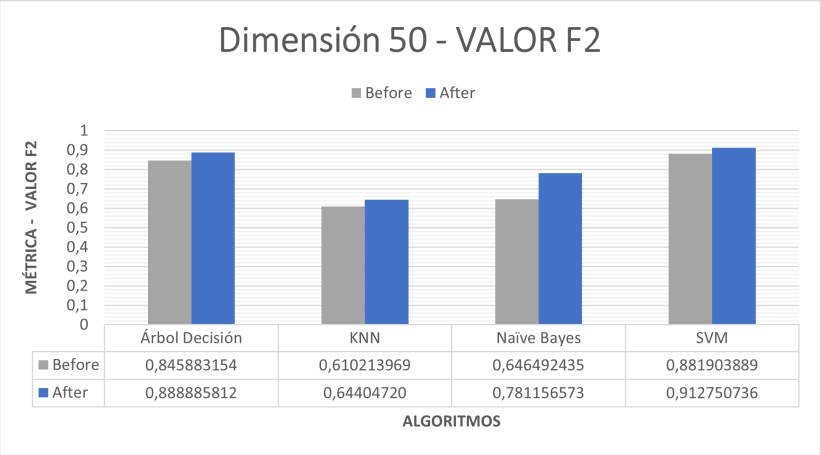
(a) F1-score - Dimensión 50



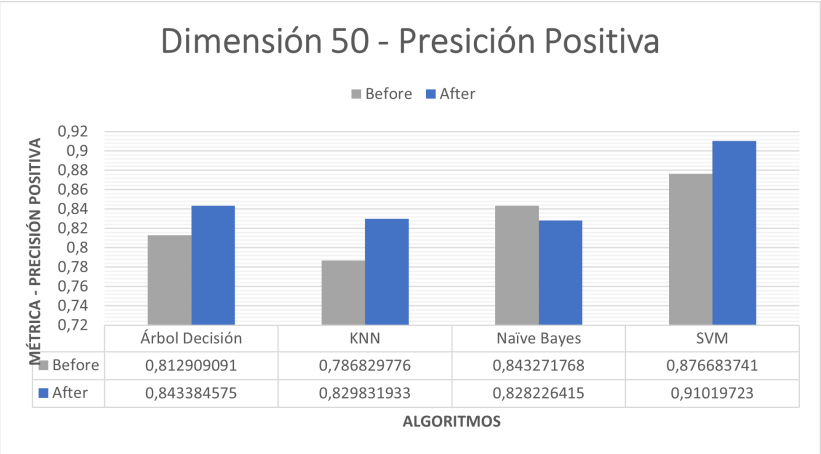
(b) F0.5 - Dimensión 50

Figura 6.8: F1-score y F0.5 - Dimensión 50
Elaborado por: Guerra Cleopatra

A continuación, se detalla los resultados de las gráficas obtenidas a partir de las medidas de los algoritmos comparando el escenario 1 y el escenario 2. Se puede observar las gráficas de la clase minoritaria en las Figuras 6.11, 6.12. Mientras que, en las gráficas 6.13 la clase mayoritaria.



(a) F2 - Dimensión 50

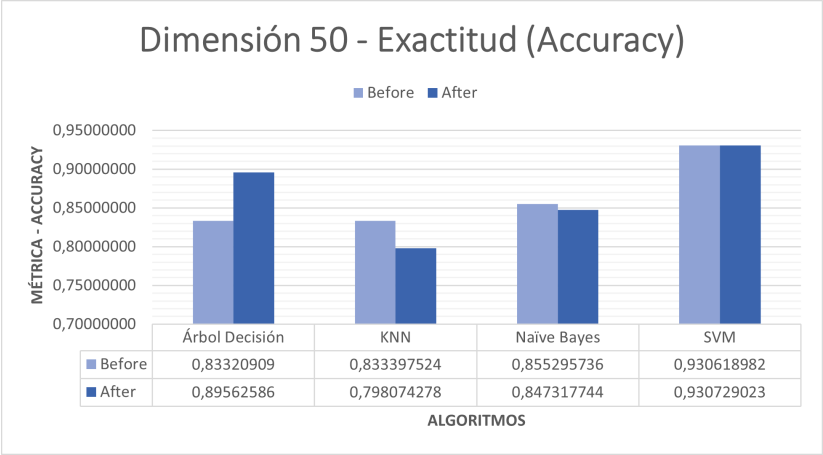


(b) Precisión Positiva - Dimensión 50

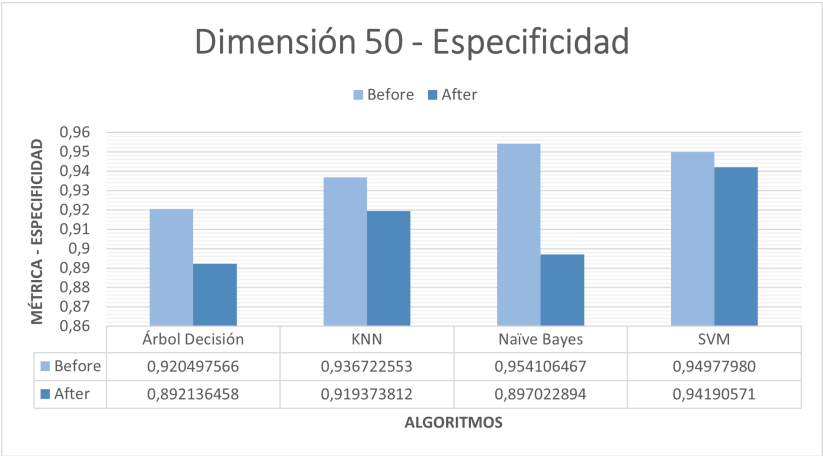
Figura 6.9: F2 y Precisión Positiva - Dimensión 50
Elaborado por: Guerra Cleopatra

Además, para el escenario 1 y 2. Se observa las gráficas de la clase minoritaria en 6.14, 6.15. Mientras que, en las gráficas 6.16 la clase mayoritaria.

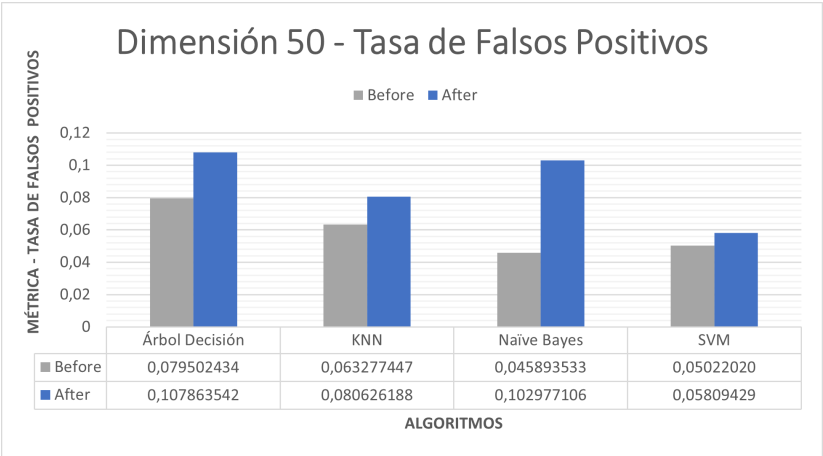
Para concluir el Anexo, se muestra los resultados de las métricas en el escenario del modelo propuesto para las dimensiones de 50, 100 y 200 6.17, 6.18, 6.19.



(a) Exactitud (Accuracy) - Dimensión 50

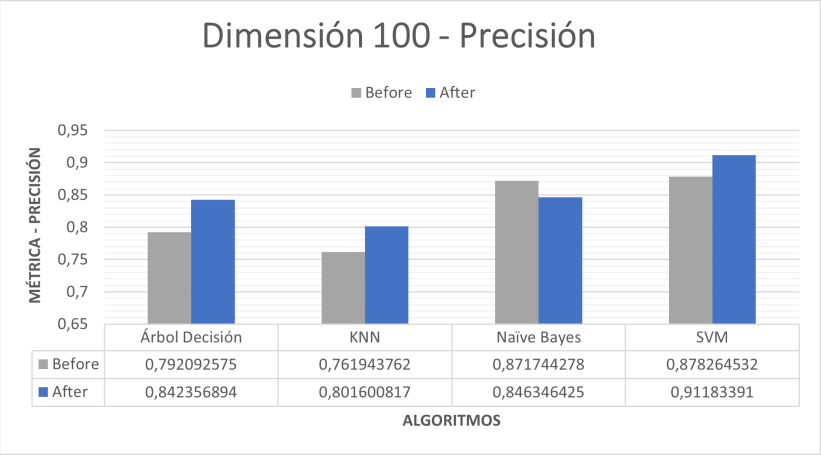


(b) Especificidad - Dimensión 50

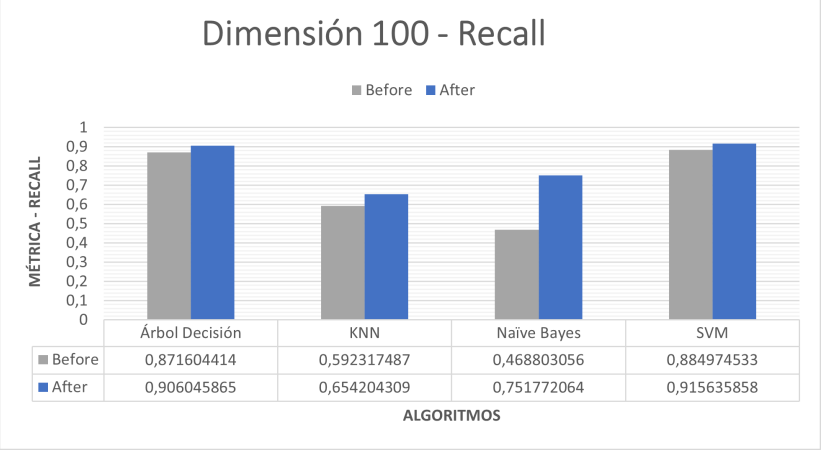


(c) Falsos Positivos - Dimensión 50

Figura 6.10: Exactitud, Especificidad y Falsos Positivos - Dimensión 50
Elaborado por: Guerra Cleopatra

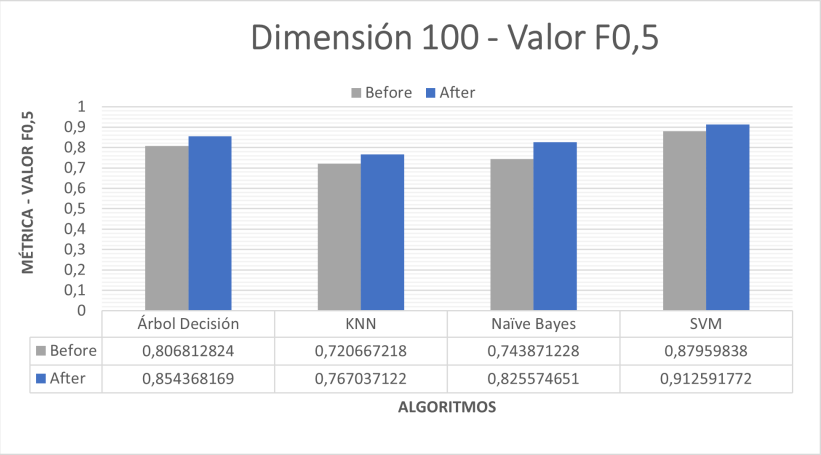


(a) Precisión - Dimensión 100

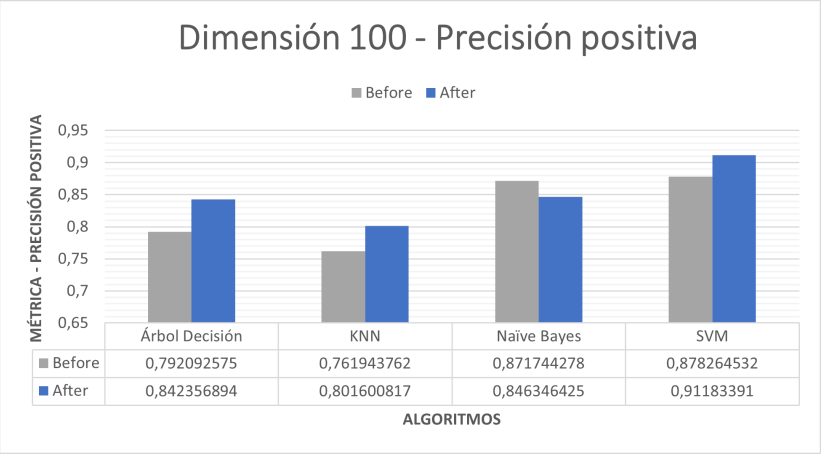


(b) Recall - Dimensión 100

Figura 6.11: Precisión y Recall - Dimensión 100
Elaborado por: Guerra Cleopatra

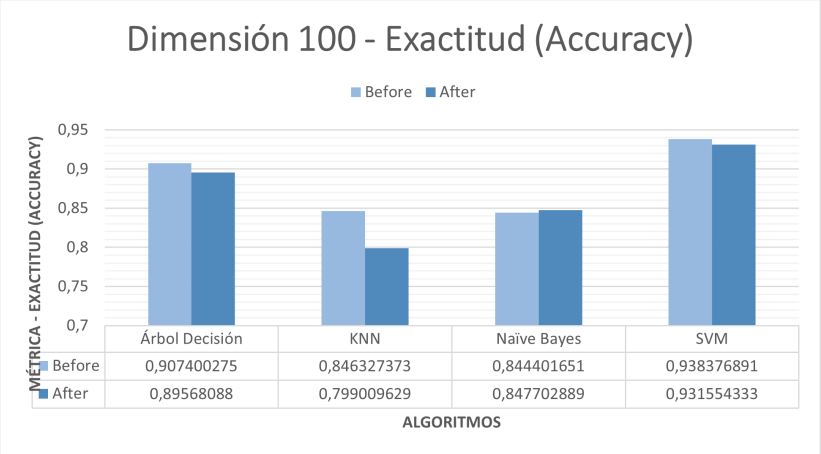


(a) F0.5 - Dimensión 100

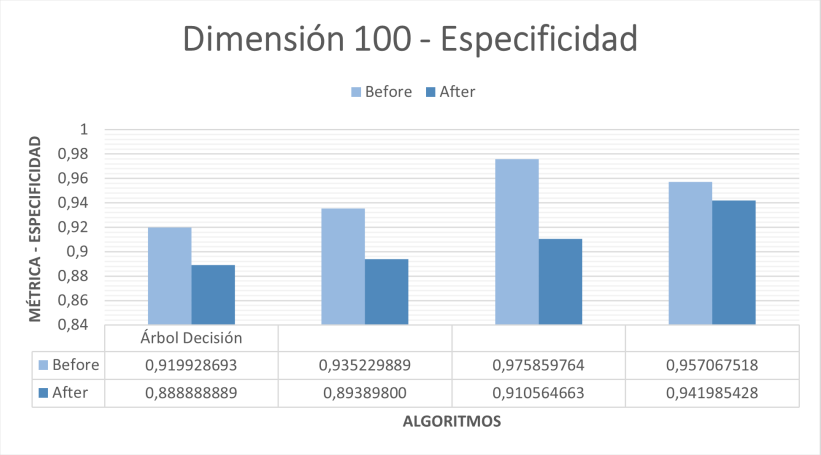


(b) Precisión Positiva - Dimensión 100

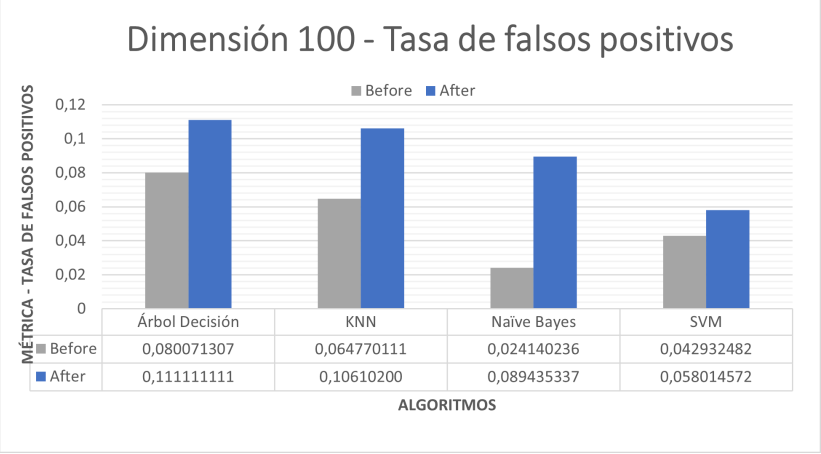
Figura 6.12: F0.5 y Precisión Positiva - Dimensión 100
Elaborado por: Guerra Cleopatra



(a) Exactitud (Accuracy) - Dimensión 100

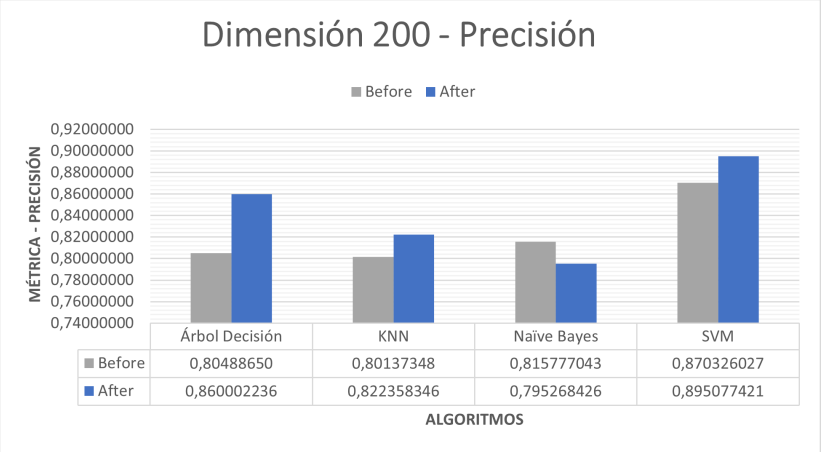


(b) Especificidad - Dimensión 100

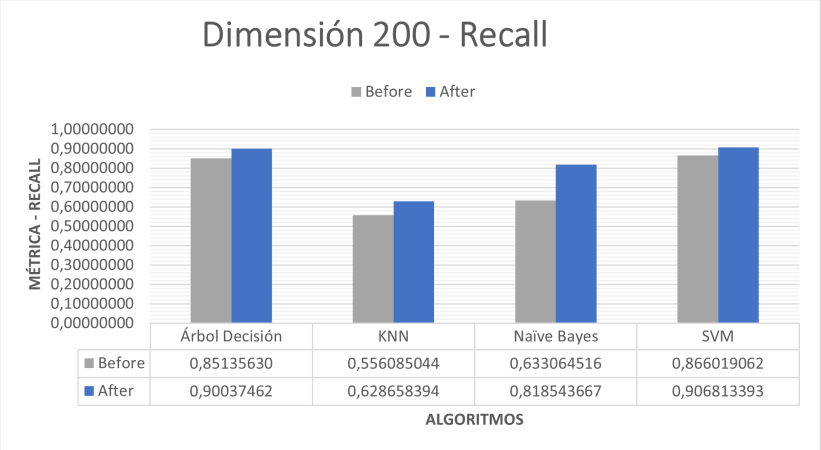


(c) Falsos Positivos - Dimensión 100

Figura 6.13: Exactitud, Especificidad y Falsos Positivos - Dimensión 100
Elaborado por: Guerra Cleopatra

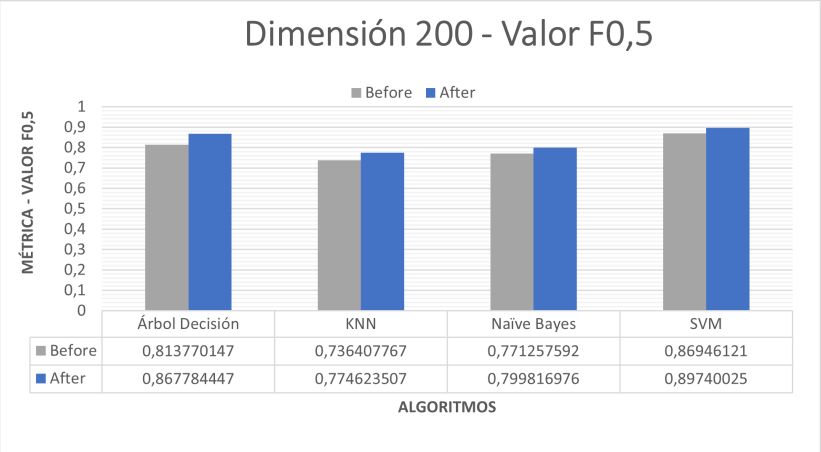


(a) Precisión - Dimensión 200

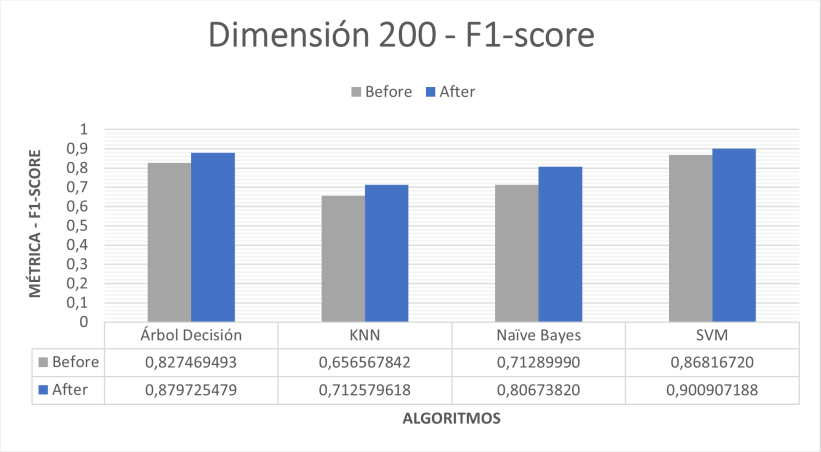


(b) Recall - Dimensión 200

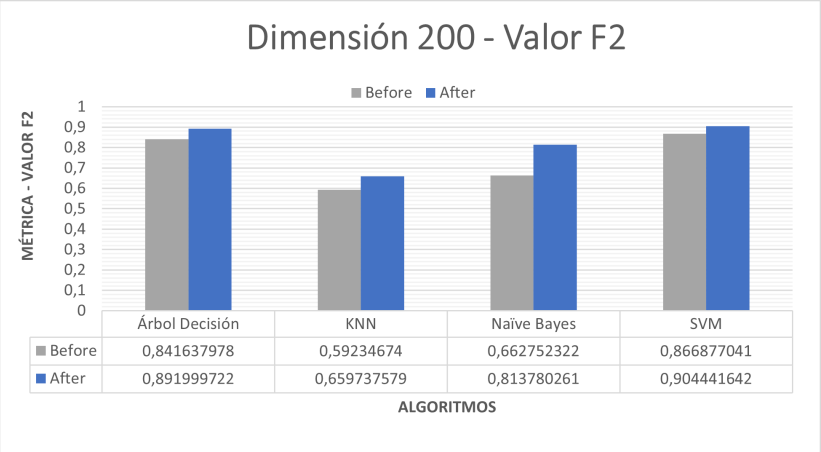
Figura 6.14: Precisión y Recall - Dimensión 200
Elaborado por: Guerra Cleopatra



(a) F0.5 - Dimensión 200

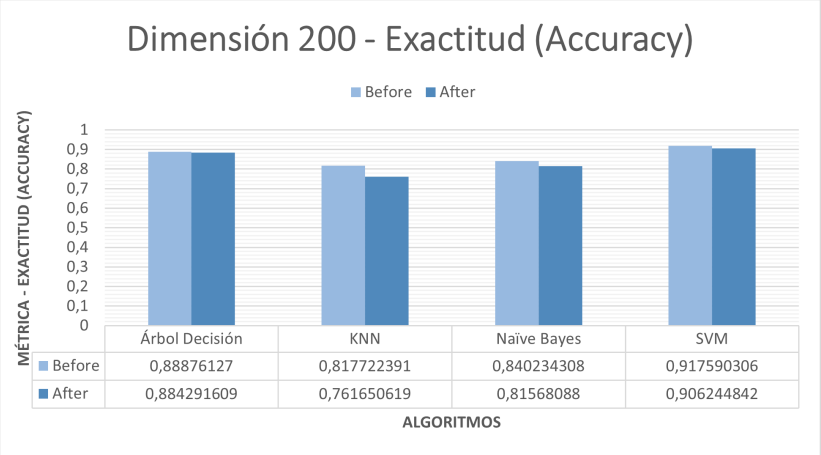


(b) F1-score - Dimensión 200

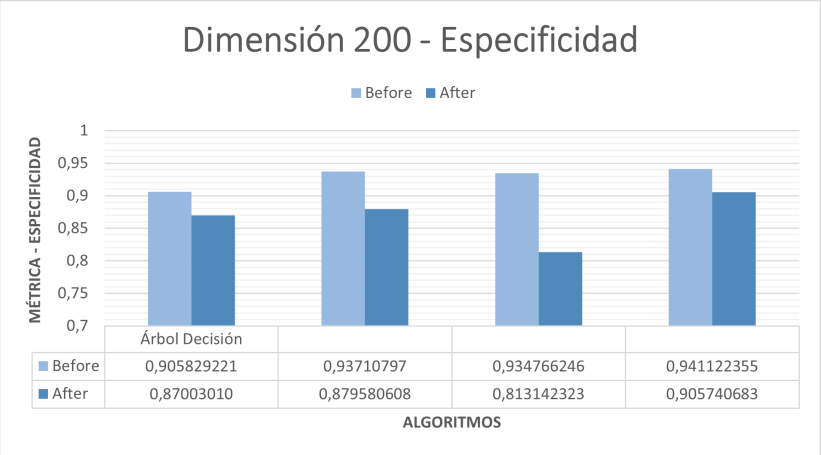


(c) F2 - Dimensión 200

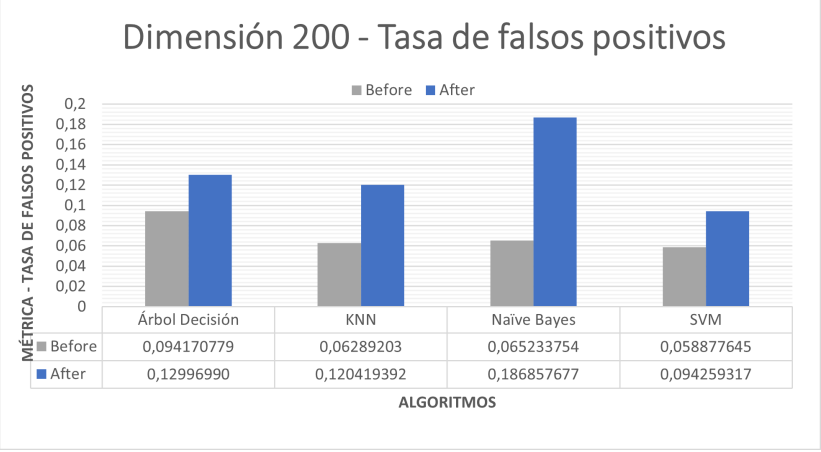
Figura 6.15: F0.5, F2 y F1-score - Dimensión 200
Elaborado por: Guerra Cleopatra



(a) Exactitud (Accuracy) - Dimensión 200

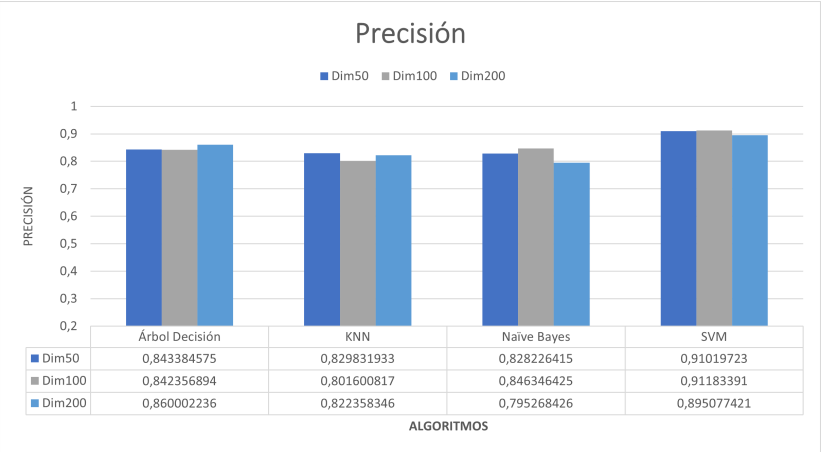


(b) Especificidad - Dimensión 200



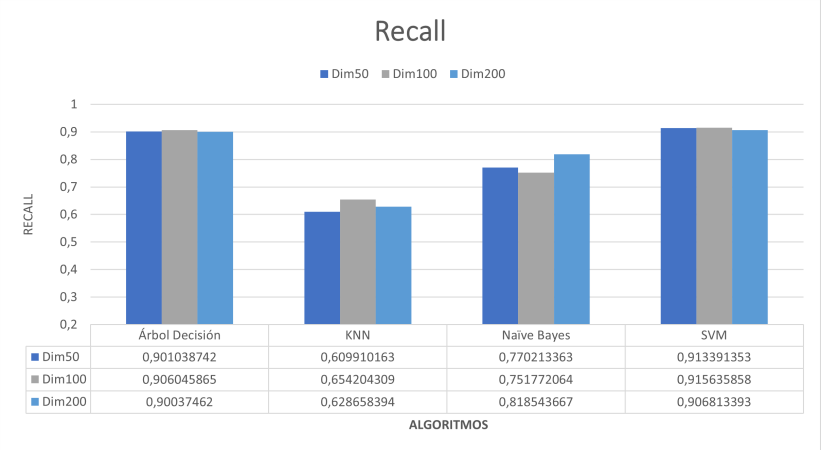
(c) Falsos Positivos - Dimensión 200

Figura 6.16: Exactitud, Especificidad y Falsos Positivos - Dimensión 200
Elaborado por: Guerra Cleopatra



	Árbol Decisión	KNN	Naïve Bayes	SVM
Dim50	0,843384575	0,829831933	0,828226415	0,91019723
Dim100	0,842356894	0,801600817	0,846346425	0,91183391
Dim200	0,860002236	0,822358346	0,795268426	0,895077421

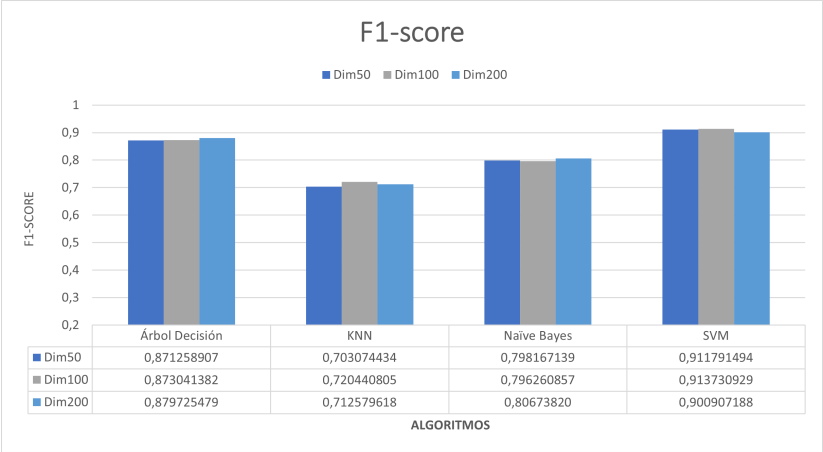
(a) Precisión - Modelo Propuesto



	Árbol Decisión	KNN	Naïve Bayes	SVM
Dim50	0,901038742	0,609910163	0,770213363	0,913391353
Dim100	0,906045865	0,654204309	0,751772064	0,915635858
Dim200	0,90037462	0,628658394	0,818543667	0,906813393

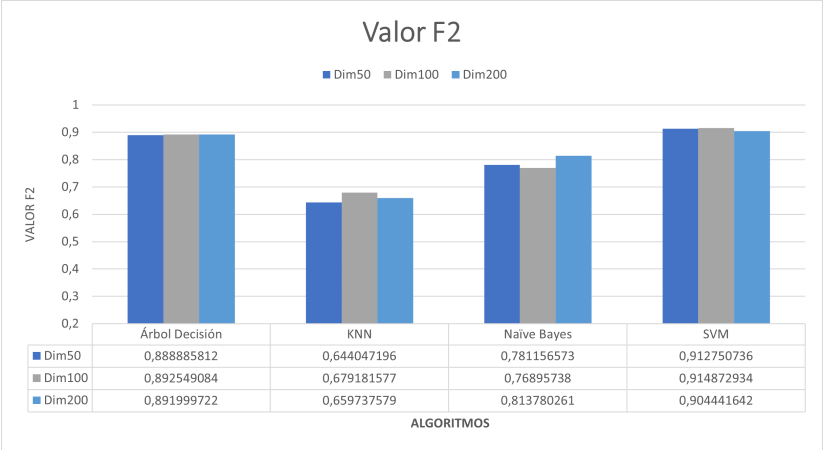
(b) Recall - Modelo Propuesto

Figura 6.17: Comparación de métricas Modelo Propuesto - Parte I
Elaborado por: Guerra Cleopatra



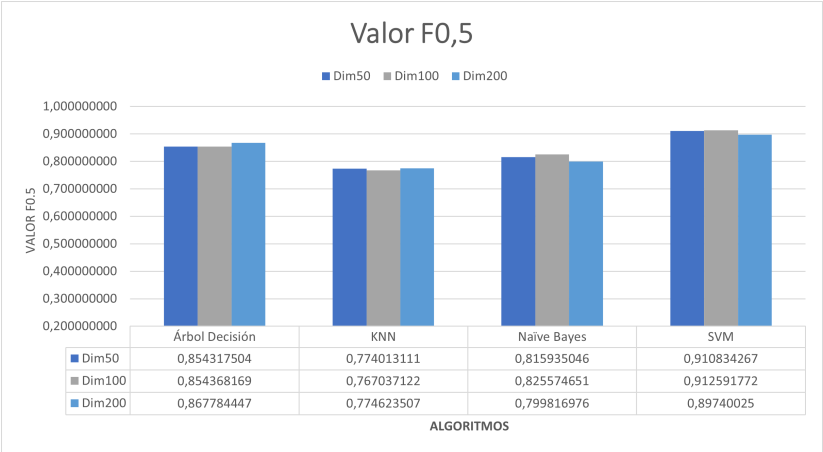
	Árbol Decisión	KNN	Naïve Bayes	SVM
Dim50	0,871258907	0,703074434	0,798167139	0,911791494
Dim100	0,873041382	0,720440805	0,796260857	0,913730929
Dim200	0,879725479	0,712579618	0,80673820	0,900907188

(a) F1-score - Modelo Propuesto



	Árbol Decisión	KNN	Naïve Bayes	SVM
Dim50	0,888885812	0,644047196	0,781156573	0,912750736
Dim100	0,892549084	0,679181577	0,76895738	0,914872934
Dim200	0,891999722	0,659737579	0,813780261	0,904441642

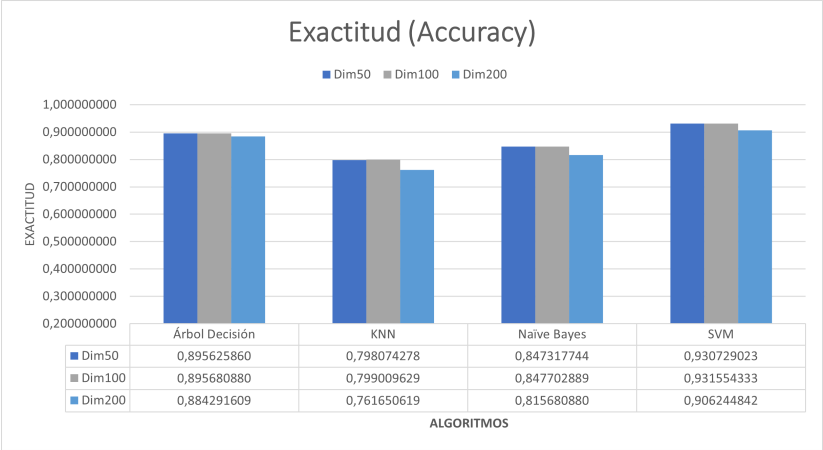
(b) F2 - Modelo Propuesto



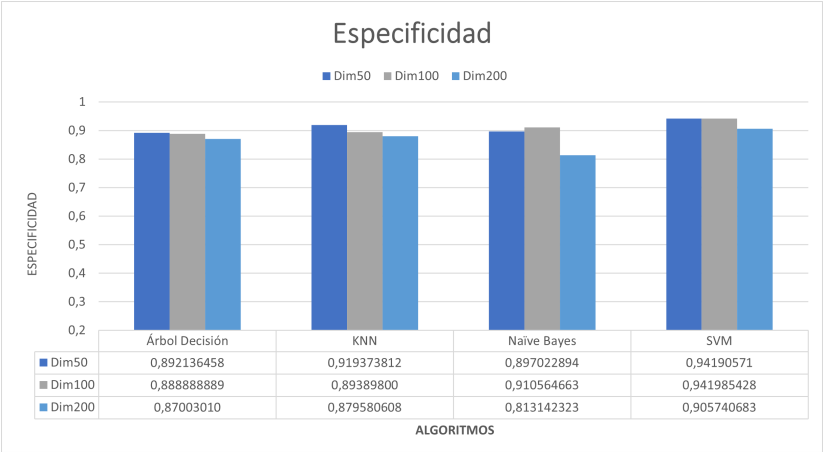
	Árbol Decisión	KNN	Naïve Bayes	SVM
Dim50	0,854317504	0,774013111	0,815935046	0,910834267
Dim100	0,854368169	0,767037122	0,825574651	0,912591772
Dim200	0,867784447	0,774623507	0,799816976	0,89740025

(c) F0.5 - Modelo Propuesto

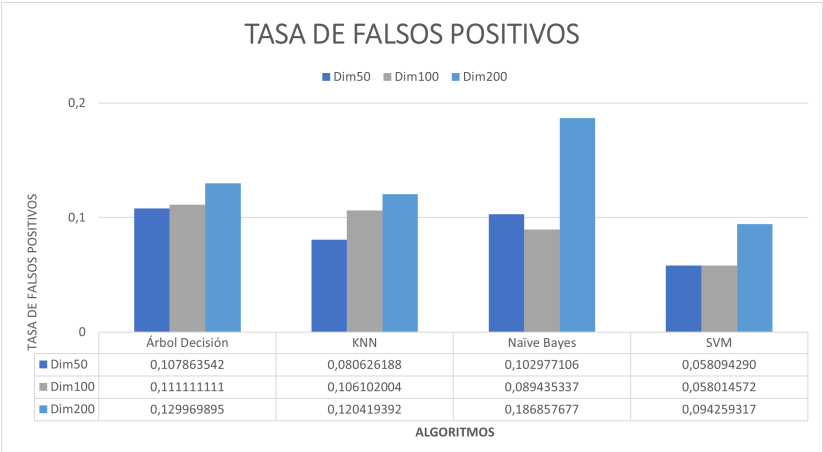
Figura 6.18: Comparación de métricas Modelo Propuesto - Parte II
Elaborado por: Guerra Cleopatra



(a) Exactitud (Accuracy) - Modelo Propuesto



(b) Especificidad - Modelo Propuesto



(c) Falsos Positivos - Modelo Propuesto

Figura 6.19: Comparación de métricas Modelo Propuesto - Parte III
Elaborado por: Guerra Cleopatra