UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERIA, ARQUITECTURA Y URBANISMO

ESCUELA ACADÉMICO PROFESIONAL DE INGENIERA DE SISTEMAS

TESIS

EVALUACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA EL MINADO DE OPINIÓN EN TWITTER

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

Autor:
Segura Vásquez, Luis Yayir

Asesor:
Mg. Samillán Ayala Alberto Enrique

Línea de Investigación:
Ciencias de la Computación

Pimentel, Perú
2019

EVALUACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA EL MINADO DE OPINIÓN EN TWITTER

Aprobación de la Tesis

Mg. Atalaya Urrutia Carlos William
Presidente del Jurado de Tesis

Mg. Celis Bravo Percy Javier
Secretario del jurado de Tesis

Dr. Ramos Moscol Mario Fernando
Vocal del Jurado de Tesis

Pimentel, Perú
2019

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

INFORMACIÓN GENERAL

1.1. Título del Informe de Investigación:
"EVALUACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA EL MINADO DE OPNIÓN EN TWITTER"

1.2. Línea de Investigación:

Ciencias de la computación

1.3. Autor:

Segura Vásquez, Luis Yayir.

1.4. Asesor Metodólogo:

Mg. Samillán Ayala Alberto Enrique

1.5. Tipo y diseño de investigación.

Tipo experimental, metodología cuantitativa.

1.6. Facultad y Escuela Académico Profesional:

Facultad de Ingeniería, Arquitectura y Urbanismo
Escuela Profesional de Ingeniería de Sistemas

1.7. Periodo: 2017-II

1.8. Fecha de inicio y término del proyecto:
Abril – Diciembre 2017

1.9. Firma de los autores del proyecto:

_____
Segura Vásquez Luis Yayir
AUTOR

1.10. Aprobado:

_____
Mg. Samillán Ayala Alberto Enrique
ASESOR DE INVESTIGACIÓN

1.11. Fecha de Presentación: Diciembre del 2017

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

DEDICATORIA

Es mi deseo como sencillo gesto de agradecimiento, dedicarle mi trabajo plasmado en la siguiente tesis, a mis padres **Luis y Gloria**, quienes permanentemente contribuyen de manera incondicional a lograr mis metas y objetivos propuestos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

AGRADECIMIENTO

Agradecer a todas personas que ofrecieron su ayuda de manera constante y ayudaron al proceso de desarrollo del presente proyecto.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Índice

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Índice de Figuras

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

RESUMEN

La presente investigación denominada "**EVALUACION DE ALGORITMOS DE CLASIFACION PARA EL MINADO DE OPINION EN TWITTER"** tiene como objetivo realizar un análisis de los diversos algoritmos utilizados en el proceso de tratamiento de textos.

La posibilidad de extraer y analizar información de los diferentes medios sociales, ha motivado que en la última década se realicen estudios que van desde la publicidad a temas socio-culturales, por medio del análisis de sentimientos (SA), también conocido como minería de opinión (opinion mining); que para Bing Liu (2016), es un campo de estudio que se centra principalmente en analizar las opiniones que expresan o implican sentimientos positivos o negativos.

Para abordar esta problemática, en la investigación "Clasificación automática de la intención del usuario en mensajes de Twitter" (Martis & Alfaro, 2014), propone un modelo para la clasificación de mensajes de Twitter de forma automática para intentar comprender cuál es la intención que tiene el usuario cuando publica un mensaje. Para este caso, los investigadores definieron un conjunto de 8 categorías, para las cuales utilizaron algoritmos de clasificación supervisada como el Super Vector Machine (SVM) y Naive Bayes; luego de evaluar el comportamiento de los mismos, indicaron que el SVM obtiene una clara ventaja sobre el segundo;

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

concluyendo así, que lo mejor es utilizar Maquinas de Soporte Vectorial para la clasificación automática de los tweets.

Palabras Clave: Minería de Opiniones, Análisis de Textos

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

ABSTRACT

The present research called "EVALUATION OF CLASSIFICATION ALGORITHMS FOR THE MINING OF OPINION IN TWITTER" aims to perform an analysis of the various algorithms used in the word processing process.

The possibility of extracting and analyzing information from different social media has motivated the study of the last decade in studies ranging from advertising to socio-cultural issues, through the analysis of feelings (SA), also known as mining of Opinion (opinion mining); That for Bing Liu (2016), is a field of study that focuses mainly on analyzing the opinions that express or imply positive or negative feelings.

To address this problem, in the research "Automatic classification of user's intention in Twitter messages" (Martis and Alfaro, 2014), proposes a model for the classification of Twitter messages automatically to try what is the intention that You have the user when you post a message.

For this case, the researchers defined a set of 8 categories, for which classification algorithms were used such as Super Vector Machine (SVM) and Naive Bayes; After evaluating the behavior of the same, they indicated that the SVM obtains a clear advantage over the second one; So ending, the best in use Vector Support Machines for automatic sorting of tweets.

Key Words: Mining of Opinions, Analysis of Texts

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

INTRODUCCION

El uso de las redes sociales (Twitter, Facebook, etc.) hoy en día ha dejado de ser solo un punto de integración social, donde individuos desde un dispositivo electrónico interactúan entre sí, compartiendo sus experiencias, ahora las empresas utilizan las redes sociales como un nuevo canal de comunicación y acercamiento con sus clientes, a la vez como centro de captación de nuevos consumidores, es decir, se ha vuelto tan vital posicionarse en las comunidades sociales hoy en día para las empresas, donde un tópico tendencia ("trending topic") permite reunir a un grupo de individuos y generar un "sentimiento" en sus opiniones expresadas, con la cual se puede medir la posición de una marca.

Sin embargo, para cuantificar la opinión de los usuarios de las redes sociales se usan mecanismos básicos (Like, Favoritos, etc.) ya que analizar cada post o comentario realizado utiliza demasiado tiempo, debido a la gran cantidad de post que se genera en estos medios sociales, dificultando esta operación.

En la actualidad se trata de automatizar el proceso de clasificación de opiniones en las redes sociales a través del análisis de los textos, donde el problema principal radica en cómo enseñar a un algoritmo o software a clasificar opiniones que tienen un carácter subjetivo, propio de la naturaleza humana.

Si bien es cierto existen modelos que tratan de dar solución a este problema, sin embargo, dada la complejidad de factores como el idioma, sintaxis, granularidad del texto en sí, entre otros factores, hacen requerir la existencia de un continuo mantenimiento para el aprendizaje y retroalimentación de dichos modelos, como es el caso del Procesamiento de Lenguaje Natural (PLN) y el Análisis Sentimental (AS).

Saber de qué se está hablando en los medios sociales, se traduciría en una ventaja para las empresas o cualquier otra organización que, por ejemplo desean conocer la reputación de su marca, analizar la aceptación de algún producto recientemente lanzado al público, o conocer cuáles son las preferencias de los usuarios a cierto candidato político.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO I

PROBLEMA DE INVESTIGACION

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

I. CAPITULO I: PROBLEMA DE INVESTIGACION

1.1. Situación Problemática

Han pasado más de 25 años desde que se empezaran a proliferar las primeras páginas web por la Word Wide Web, creada por el programador inglés Tim Berners-Lee en 1989. Desde ese entonces se han generado inimaginables volúmenes de datos. Con la aparición de la web 2.0 y el innegable auge de los medios sociales, el análisis de texto se ha convertido en un gran campo de interés en los diferentes ámbitos profesionales. Las redes sociales como medio de difusión son una fuente muy valiosa de información ya que almacena gustos, preferencias y millones de opiniones vertidas a cada instante por usuarios de todas partes del mundo.

La posibilidad de extraer y analizar información de los diferentes medios sociales, ha motivado que en la última década se realicen estudios que van desde la publicidad a temas socio-culturales, por medio del análisis de sentimientos (SA), también conocido como minería de opinión (opinion mining); que para Bing Liu (2016), es un campo de estudio que se centra principalmente en analizar las opiniones que expresan o implican sentimientos positivos o negativos.

En un artículo publicado por Villena (2015), menciona que la tarea de clasificar automáticamente un texto escrito en un lenguaje natural en un sentimiento positivo o negativo, opinión o subjetividad, es a veces tan complicada que incluso es difícil poner de acuerdo a diferentes anotadores humanos sobre la clasificación a asignar a un texto dado. La interpretación personal de un individuo es diferente de la de los demás, y además se ve afectada por factores culturales y experiencias propias de cada persona. Y la tarea es aún más difícil cuanto más corto sea el texto, y peor escrito esté, como es el caso de los mensajes en redes sociales como Twitter o Facebook.

En el trabajo de investigación de García & Azaustre (2014) “Minería de datos aplicadas a las redes sociales”, indica que existen una serie de técnicas para el minado de datos, los cuales provienen de la inteligencia artificial y la estadística, dichas técnicas no son más que algoritmos, más o menos sofisticados, los cuales se aplican a un conjunto de datos para obtener resultados. En la misma investigación se menciona que según el objetivo del análisis, los algoritmos se pueden clasificar en algoritmos supervisados (predictivos) y no supervisados (o del descubrimiento del conocimiento).

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Para abordar esta problemática, en la investigación “Clasificación automática de la intención del usuario en mensajes de Twitter” (Martis & Alfaro, 2014), propone un modelo para la clasificación de mensajes de Twitter de forma automática para intentar comprender cuál es la intención que tiene el usuario cuando publica un mensaje. Para este caso, los investigadores definieron un conjunto de 8 categorías, para las cuales utilizaron algoritmos de clasificación supervisada como el Súper Vector Machine (SVM) y Naive Bayes; luego de evaluar el comportamiento de los mismos, indicaron que el SVM obtiene una clara ventaja sobre el segundo; concluyendo así, que lo mejor es utilizar Maquinas de Soporte Vectorial para la clasificación automática de los tweets.

Saber de qué se está hablando en los medios sociales, se traduciría en una ventaja para las empresas o cualquier otra organización que, por ejemplo, desean conocer la reputación de su marca, analizar la aceptación de algún producto recientemente lanzado al público, o conocer cuáles son las preferencias de los usuarios a cierto candidato político.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

1.2. Formulación del Problema

¿Qué algoritmo de clasificación obtendrá mejores resultados para el minado de opinión en Twitter?

1.3. Delimitación de la Investigación

Se diseñará el modelo teniendo como escenario datos de la red social Twitter, específicamente orientado al análisis de contenido por Hashtag Trendent Topic (Tópicos tendencia de twitter).

El proyecto concluye con el desarrollo de un prototipo del sistema informático clasificador, que consiste en un portal web que consulta los hashtags de twitter y realiza la clasificación de contenidos, así como estadística de los resultados obtenidos.

1.4. Justificación e Importancia de la Investigación

En el sector científico - académico por que realiza investigación sobre sistemas inteligentes, modelos computacionales capaces de simular el comportamiento humano a través de la interpretación de sus acciones o emociones. Además, pretende definir la composición o arquitectura que implicaría la solución tecnológica a diseñar (Metodología de desarrollo de un software y de un sistema inteligente).

Integra los principios de máquina de aprendizaje (Machine Learning) para brindar un motor retroalimentado en los conocimientos del modelo, este motor permitirá optimizar la función de mantenimiento que requiere este tipo de soluciones.

En el sector social-empresarial porque permite el análisis a través de extracción de información de los usuarios de las redes sociales como un elemento o variable de performance que permita realizar un seguimiento por parte de la empresa como oportunidad de valor de su negocio.

Se expandirá el dominio de análisis de los individuos de las redes sociales, la interpretación de los textos que en la actualidad es de rango abierto, ya que es difícil de catalogar, quedará disponible y se podrá mejorar la eficacia del análisis en redes sociales.

1.5. Limitaciones de la Investigación

Se aplicará el estudio delimitando los hashtags de la región Sudamérica, procesando y analizando textos en idioma español (Español Latino).

1.6. Objetivos de la Investigación

Objetivo general

Evaluar algoritmos de clasificación con datos extraídos de la red social Twitter.

Objetivos específicos

a) Analizar el ámbito donde se aplicarán los algoritmos de clasificación.
b) Determinar una estrategia para la extracción y tratamientos de tweets.
c) Seleccionar los algoritmos de clasificación.
d) Diseñar y construir un modelo clasificador para el minado de opiniones en Twitter
e) Evaluar los algoritmos de clasificación seleccionados.
f) Implementar una aplicación para el análisis y visualización de resultados.

CAPITULO II
MARCO TEORICO

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

II. CAPITULO II: MARCO TEORICO

2.1. Antecedentes de Estudios

"Evaluación de reglas de asociación en text mining utilizando métricas semánticas y estructurales" (Gonzales, 2010)

El enfoque de minado de textos basado en Lattice Conceptual Semántico mostro mejores resultados frente al sistema de minería de textos tradicional y el modelo de Toussaint en términos de la correlación y su poder predictivo. Esto podría deberse a que el modelo Lattice Conceptual semántico incorpora un modelo de conocimiento semántico del dominio que permite evaluar el grado de interés de las reglas extraídas mediante una nueva métrica de conformidad semántica.

En relación con la tesis propuesta aporta significancia en la propuesta del algoritmo a utilizar en las técnicas de Minería de textos del modelo a diseñar.

"Un modelo linguistico-semantico basado en emociones para la clasificación de textos según su polaridad e intensidad" (Cuadrado, 2011)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

El trabajo presentado en esta tesis describe una nueva aproximación para la clasificación de textos según su polaridad e intensidad emocional, basada en el análisis semántico del texto, y en el uso de reglas lingüísticas avanzadas. El objetivo es determinar cuándo una oración o documento expresa un sentimiento positivo, negativo o neutral, así como la intensidad del mismo. El método hace uso de un algoritmo de desambiguación semántica para trabajar a nivel de conceptos en lugar de términos, y utiliza el léxico afectivo Senti Sense, desarrollado como parte de esta tesis, para extraer el conocimiento emocional y representar cada texto de entrada como un conjunto de categorías emocionales.

A diferencia de los enfoques anteriores, el texto de entrada es modelado como un conjunto de emociones en lugar de términos o expresiones polares, lo que permite capturar con mayor fidelidad la polaridad e intensidad del texto afectivo. Así mismo, se han desarrollado técnicas lingüísticas avanzadas para la identificación de negaciones, cuantificadores y modales, así como su ámbito de acción y su efecto sobre las emociones a las que afectan.

En relación con la tesis propuesta, la investigación permitió comprender el análisis sentimental de una expresión y añadir el

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

factor de grado de polaridad importante a la hora de que el modelo empiece a clasificar.

"Text mining aplicado a la clasificación y distribución automática de correo electrónico y detección de correo spam" (Echevarria, 2009)

La conclusión de esta tesis es que se optimizó la capacidad de clasificar el contenido de los correos recibidos como SPAM o no, permitiendo así que el espacio de disco no sea ocupado por correos basuras; este problema se lo pudo resolver de manera óptima con el algoritmo Naive Bayes, mediante el filtro bayesiano que se creó el cual se volverá más eficiente al adquirir la información necesaria de correos SPAM y no SPAM que el usuario proporcione.

En relación a la investigación esta tesis presentó el uso del algoritmo Naive Bayes que también se usara para el desarrollo del modelo propuesto en la presente investigación

2.2. Estado del arte

(Kotwal, Fulari, Jadhav & Kad, 2016) en su "Improvement in Sentiment Analysis of Twitter Data Using Hadoop" indican que

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Twitter produce una cantidad extremadamente grande de datos todos los días en forma de tweets. Estos datos están principalmente desestructurados o estructurados y se denominan BigData. Por lo tanto, para poder realizar el análisis de sentimientos, se necesita tecnología avanzada que tenga la capacidad de tratar con grandes cantidades de datos de manera eficiente. Hay varios desafíos al momento de tratar con este gran volumen de datos como el procesamiento de grandes conjuntos de datos, la extracción de información útil de conjuntos de datos generados en línea, etc. El término Big Data se utiliza globalmente para la recopilación de conjuntos de datos que son enormes y complejos; esto hace que sea difícil de procesar mediante la adopción de métodos tradicionales de procesamiento de datos. Este papper propone utilizar la herramienta Hadoop para resolver los desafíos relacionados con el Big Data, proporcionando una oportunidad para entender los patrones de datos y ayudando con la predicción de eventos y resultados.

Para la presente investigación se aplicarán los conceptos básicos de Big Data para el tratamiento masivo de información de la red social.

(Saifa, He, Fernadez & Alania Ene, 2015) en “Contextual semantics for sentiment analysis of Twitter”, presentaron

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

SentiCircles, un enfoque basado en el léxico para el análisis del sentimiento en Twitter. Diferente de los enfoques basados en el léxico típico, que ofrecen una polaridad de palabras preestablecidas y estáticas anteriores sin importar su contexto, SentiCircles toma en cuenta los patrones de co-ocurrencia de palabras en diferentes contextos en tweets para capturar su semántica y actualizar sus asignaciones preasignadas Fuerza y polaridad en los léxicos del sentimiento en consecuencia. El enfoque presentado permite la detección de sentimientos tanto a nivel de entidad como de tweet. El enfoque propuesto fue evaluado en tres conjuntos de datos de Twitter usando tres diferentes léxicos de sentimientos para derivar sentimientos anteriores de palabra. Los resultados mostraron que el enfoque propuesto supera significativamente a las líneas de base en la precisión y la medida-F para la subjetividad de la entidad (neutral frente a polar) y la polaridad (positivo frente a las detecciones negativas). Para la detección de sentimientos a nivel de tweets, el enfoque funciona mejor que el estado de la técnica de SentiStrength por 4-5% de precisión en dos conjuntos de datos, pero cae ligeramente por detrás en un 1% en F-medida en el tercer conjunto de datos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

El aporte de la investigación será la determinación de las propiedades del modelo a diseñar y los algoritmos a implementar para la clasificación.

2.3. Base teórica científicas

2.3.1. Minería de Opinión

A. Opinión de internet

Internet se ha convertido en los últimos años como un medio de difusión cada vez más utilizado por las empresas e instituciones que quieren dar a conocer su marca o producto a los miles de millones de usuarios que navegan diariamente en la red, en busca de información que les ayude a tomar una decisión antes de realizar una elección de compra, contratar un servicio, inscribiese en una escuela, etc.

La evolución de Internet hacia la Web 2.0, basada en la participación activa del usuario mediante las redes sociales, ha permitido que la opinión del usuario tenga su importancia en el conjunto global de la red. Merlo, Contreras y Puente (2010).

Las redes sociales, como Facebook y Twitter, han abierto una oportunidad para saber lo que la gente opina sobre

determinados temas de actualidad, o conocer la reacción de los usuarios o potenciales compradores sobre algún producto o servicio. Teniendo en cuenta los datos que muestra Edwin Bardales (2015), el Perú pasará de tener 14 millones de usuarios conectados a internet, a aproximadamente 19 millones para el año 2019; estos datos no hacen más que reflejar una mayor participación de los peruanos en internet en los próximos años, participando con sus opiniones en algún tema de interés o alguna actividad, como por ejemplo algunos comicios electorales.

Actualmente cualquier usuario que navega por el internet tiene la posibilidad de poder opinar, ya sea comentando una publicación o escribiendo alguna entrada para su web o blog. De la misma manera, los anunciantes o vendedores cada vez se preocupan más por que puedan decir los usuarios acerca de su producto o servicio; esto también ocurre en el ámbito político, donde los candidatos son conscientes que la gente los está valuando constantemente a través de sus opiniones que se ven reflejadas en la web.

B. Minería de opinión

La tarea de identificar opiniones en internet se realiza a través de análisis de sentimientos o minería de opinión; esta técnica

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

basada en el procesamiento de lenguaje natural (NLP), análisis de texto y herramientas computacionales, sirve para clasificar comentarios subjetivos vertidos por usuarios sobre diversos temas.

Para Liu (2012), la minería de opinión, es el campo de estudio que analiza las opiniones, sentimientos, evaluaciones, valoraciones, actitudes, y emociones hacia entidades tales como productos, servicios, organizaciones, individuos, temas, eventos, temas y sus atributos. También hay muchos nombres y tareas ligeramente diferentes, por ejemplo, análisis del sentimiento, minería de opinión, extracción de opinión, minería del sentimiento, análisis de subjetividad, análisis de afectos, análisis de emociones, etc. Sin embargo, ahora están bajo el paraguas del análisis del sentimiento o minería de opinión, mientras que en la industria, el término análisis del sentimiento es más comúnmente utilizado, pero en la academia tanto el análisis del sentimiento y la minería de opinión se emplean frecuentemente.

El análisis de sentimientos trata de clasificar los documentos en función de la polaridad de la opinión que expresa su autor. Esta nueva área que combina PNL y minería de textos, incluye una gran cantidad de tareas que han sido tratadas en

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

mayor o menor media, según argumentan Martínez, Martín y Ureña (2011).

Los mimos autores sostienen que existen principalmente dos formas distintas de abordar esta problemática: aplicando aprendizaje automático o aplicando un enfoque semántico. Dos son las aplicaciones más importantes: determinar la polaridad de las opiniones a nivel de documento, frase o característica, y determinar si un documento contiene opiniones.

Al momento de querer analizar y clasificar sentimientos u opiniones en internet, específicamente en Twitter, se deben de tener en cuenta los desafíos que se presenta en esta clase de estudios; los mismos que, Montesinos (2014) los describe en su memoria de grado de la siguiente manera:
En primer lugar, es necesario determinar si existe opinión en el tweet o no, ya que no siempre esto ocurre, pudiendo ser un comentario objetivo, una respuesta a otro usuario, etc.

Determinar el tema sobre el cual se está hablando de manera de saber si es información útil, ya que se puede estar buscando opiniones sobre una empresa determinada y si el

tweet es sobre política no aporta información relevante sobre lo que se está buscando.

Reconocer las abreviaciones y modismos típicos. Al tener Twitter un carácter informal el lenguaje usado no siempre es correcto, ya que normalmente no se ocupan tildes y se ocupan palabras populares que no aparecen en el diccionario (Ej. Ocupar “bn” en vez de “bien”, “x” en vez de “por”, el uso de garabatos, usar expresiones del tipo “pe”,” malooooo”, etc.).

Determinar la polaridad de una oración pudiendo tener palabras positivas y negativas en la misma frase (Ej. “Me alegro que se haya terminado, pésimo el espectáculo”, “La película no fue nada buena”).

Teniendo en cuenta lo anterior, los tweets a analizar vendrían a ser todos aquellos que contengan una opinión, una evaluación o expresen alguna emoción sobre un determinado tema; descartando a las publicaciones objetivas o de carácter informativo. Para solucionar esta tarea, a la opinión se polariza, y se determina si es positiva o negativa con respecto a un tema en específico. Sin embargo, este no es un tema simple de resolver, ya que dependiendo del contexto hay

palabras que pueden expresar tanto una opinión positiva como negativa; como por ejemplo del comentario “no emite sonido alguno”, podría considerarse algo positivo cuando se habla de un auto nuevo, pero también podría indicar algo negativo si se hablase de un equipo de música.

C. **Twitter como caso de estudio**

Según Rodriguez (2011), Twitter es la herramienta social más utilizada en el mundo para hacer “microblogging”; es decir, publicar mensajes cortos de texto para un grupo de seguidores. Su popularidad radica en que los usuarios solo pueden publicar contenidos no mayores a los 280 caracteres; en los cuales se ven reflejados la opinión, los sentimientos, las emociones y sus actitudes hacia ciertos temas o productos. De esta manera, Twitter se ha convertido en una indiscutible fuente de información para realizar estudios de mercado y estudios sociales.

La plataforma Twitter fue lanzado en julio del 2006 por su creador Jack Dorsey, y desde este entonces se estima que tiene más de 320 millones de usuarios registrados alrededor del mundo, generando 65 millones de twees al día.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

El Perú no está exento de estas cifras, y según un estudio realizado por Quantico Trends (2016), afirma que existen cerca de 4,3 millones de usuarios peruanos registrados en esta red social, y se estima que para el cierre del año 2016, esta cifra aumente hasta los 6 millones de usuarios.

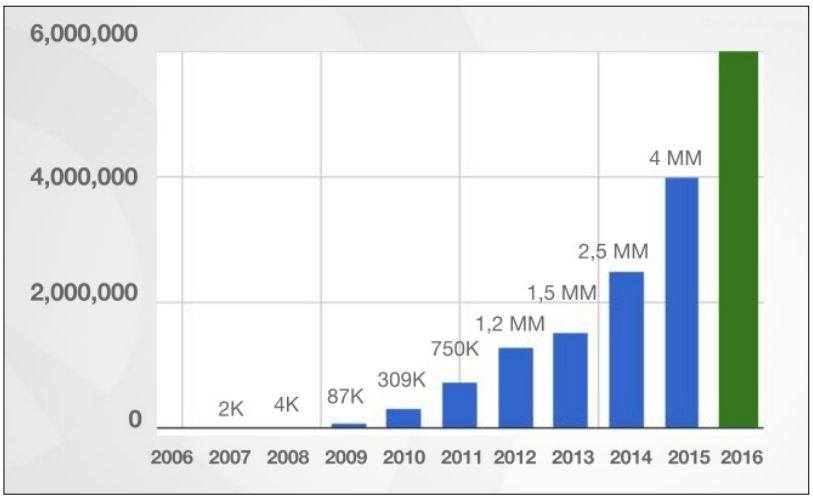


Figure 1. Esta tabla muestra la evolución en cuanto al número de usuarios dados de alta en la red social Twitter desde el año 2006. Fuente: Quantico Trends.

De los 4,3 millones de cuentas en nuestro país, un total de 1,119,624 perfiles pertenecen a empresas, medios, pseudónimos, entre otros; la cantidad restante son personas con un 54% hombres y 46% mujeres. El mismo estudio indica que la mayoría de usuarios peruanos en Twitter son menores de 25 años. De la misma forma, Quantico Trends sostiene que los departamentos desde donde más se tuitean son: Lima (74,11%), La Libertad (6,13%), Arequipa (4,36%), Lambayeque (4,34%), entre otros.

El estudio confirma que en el Perú si bien por el momento solamente la cuarta parte de la población en general cuenta con un perfil en Twitter, la proyección del estudio citado anteriormente, nos muestra que cada vez más peruanos se incorporan a esa red social, lo que resulta ideal para realizar un estudio sobre análisis de sentimientos en esta importante plataforma.

D. Minería de opinión en twitter

En la literatura se pueden encontrar estudios de minería de opinión teniendo como fuente de datos a los tweets. Pak y Paraubek (2010) indican que, cada vez son más los usuarios que publican acerca de los productos y servicios que utilizan, o expresan sus puntos de vista tanto en temas políticos como religiosos, y a medida que esto ocurre, el sitio web del microblogging, se convierte en una valiosa plataforma fuente de opiniones y sentimientos de las personas.

Por otra parte, la información publicada en Twitter no sobrepasa de los 280 caracteres, por lo que se puede asumir que los mensajes expresan una única idea, como consecuencia a cada tweet se le asigna una sola opinión, lo que simplifica el problema; esto se diferencia de otras redes

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

sociales, en las cuales la cantidad de texto que se puede publicar es mucho mayor y por lo tanto dentro de este puede haber muchas opiniones o grados de sentimiento.

Según refieren Khan, Atique y Thakare (2015), para determinar si un tweet expresa un sentimiento positivo o negativo, se utilizan principalmente dos enfoques: el enfoque semántico y el enfoque basado en el aprendizaje automático o computacional. El enfoque semántico basado en lexicons determina el sentimiento o la polaridad de la opinión a través de alguna función de las palabras de opinión en el documento o en la oración. El enfoque basado en el aprendizaje de la máquina suele entrenar clasificadores de sentimientos usando características como unigramas o bigrams. La mayoría de las técnicas utilizan algún tipo de aprendizaje supervisado aplicando diferentes métodos de técnicas de aprendizaje tales como Naive Bayes, Máxima Entropía y Support Vector Machines (SMV). Estos métodos requieren un marcado manual de ejemplos de entrenamiento para cada dominio de aplicación. Mientras que la mayoría de los métodos de análisis de sentimientos se propusieron para los documentos de gran opinión (por ejemplo, revisiones, blogs), algunos trabajos recientes se han dirigido a los microblogs.

Las técnicas de aprendizaje supervisado se han utilizado en el desarrollo de sistemas de análisis de sentimiento de Twitter en línea, por lo que los investigadores citados anteriormente indican como un enfoque dominante en la resolución de la compleja tarea analizar sentimientos en Twitter.

E. Metodología para el minado de opinión en Twitter

Determinar si una tweet expresa realmente una opinión o no, conlleva realizar diferentes procesos; los mismos que pueden estar enmarcados dentro de uno de los dos enfoques que se mencionó anteriormente.

Estevez & Almeida (2015) sostienen que los métodos más utilizados para el minado de opinión en Twitter, son las técnicas de clasificación supervisada (descritas en el apartado 3.3.4.) que se encuentran dentro del enfoque de aprendizaje automático o computacional.

Actualmente hay un conjunto de métodos para realizar el minado de opinión en Twitter; sin embargo, estos investigadores manifiestan que no existe una metodología que se muestre superior al resto. Por lo tanto el modelo de

clasificación de twits para el presente trabajo se divide en tres etapas:

Extracción o recolección de datos.

Preprocesamiento o normalización de los datos.

Clasificación supervisada.

2.3.2. Recolección de datos

Para poder acceder y extraer datos de Twitter, la plataforma ofrece diferentes métodos a través de su API pública. Una API REST, que proporciona una interfaz sencilla a las funcionalidades de Twitter, y una API Streaming que es una poderosa API en tiempo real. El acceso a los datos públicos de Twitter es extremadamente limitado con la API REST, y a menor medida para la API de streaming de Twitter. Es importante indicar que ninguno de estos dos métodos proporción el acceso completo a los tweets públicos posteados en esta red social.

A. API Rest

Proporciona gran cantidad de interfaces que engloban las distintas funcionalidades que ofrece Twitter. Entre estas interfaces se encuentra las siguientes: Timeline, Tweets, Search, Streaming, Direct Messages, Friends & Followers,

Users, Suggested Users, Favorites, Lists, Saved Searches, Places & Geo, Trends, Spam Reporting, OAuth, Help.

B. API Streaming

El conjunto de APIs Streaming que proporciona Twitter posibilita el acceso de baja latencia al Stream global de Tweets.

Twitter suministra una serie de streams diferenciados, cada uno de ellos con un propósito distinto. A continuación, se listan los streams existentes:

Public Stream

User Stream

Site Stream

En la documentación oficial de Twitter se recoge la definición del mismo y nos manifiesta que se permite el acceso al Stream de los datos públicos que fluyen a través de Twitter. Por lo que el uso de esta API, se recomienda para seguir usuarios o temas específicos, así como para minería de datos. Debido a esto, el API Stream será el método elegido en la presente investigación para poder conectarse con la plataforma de esta red social.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2.3.3. Normalización de datos

Debido al límite de 280 caracteres que impone twitter, los mensajes generalmente contienen deformaciones del lenguaje; tales como: jergas, urls, palabras repetidas, emoticones, y otros elementos que dificultan de manera notable las tareas minado de tweets.

Según indica Stévez (2015), usualmente la normalización de los textos aumenta la calidad de los algoritmos de aprendizaje automático empleados posteriormente para las tareas de minería de texto. Por este motivo, es común que las aplicaciones de minería de texto sobre tweets cuenten con una fase inicial de pre-procesamiento, con el objetivo de normalizar el contenido de los mensajes.

Del mismo autor citado en el párrafo anterior, se ha tomado una serie de procesos para lograr normalizar los tetes, para que de esta manera el modelo de clasificación tenga mayor precisión en los resultados.

A. Eliminación de etiqueta

En Twitter es relativamente común la utilización de referencias a otros sitios de la Web con frecuencia, como blogs, noticias o reseñas. La dirección de estos sitios,

generalmente acortada usando un servicio específico para ese fin, aparece en el propio mensaje. El problema surge cuando en la URL se encuentra palabras que pueden influir sobre el clasificador de forma no deseada. Por esto es necesario eliminarlas ya que forman parte del texto del mensaje. En un tweet se hace referencia no solo a sitios de Internet, también pueden hacerse referencias a otros usuarios de Twitter, a un tópico en especial o a otros tweets mediante el uso de marcas o etiquetas propias de Twitter que aparecen en el texto y pueden influir de la misma manera que las URL sobre los clasificadores, por lo que también son eliminadas.

B. Identificación de emoticonos

También es frecuente la utilización de emoticones o smileys empleados para darle expresividad al mensaje dado que Internet es un medio frío y a menudo se desean expresar opiniones positivas o negativas. A los emoticones se le da usos más complejos de detectar como es la ironía, de ahí que sean muy importantes al realizar un análisis de sentimientos.

C. Análisis de jerga

La mayor diferencia entre un tweet y un texto en lenguaje natural es el uso indiscriminado de abreviaturas y jerga

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

debido a la restricción en la longitud de 140 caracteres. La jerga, también llamada netspeak, ofrece como resultado más de una abreviatura para una única palabra. Esto provoca que el peso de dicha palabra durante el proceso de entrenamiento, quede repartido entre las variaciones de la misma o que en la clasificación una de las variaciones obtenga un peso mínimo, aunque la palabra esté entre los términos determinantes del mensaje correspondiente. Por tanto, es evidente la necesidad de traducir la jerga a un lenguaje más convencional. Para esto se utilizan diccionarios que permiten convertir expresiones de jerga a su equivalente en lenguaje natural

D. Homogenización

Como parte del pre-procesamiento se hace un tratamiento para evitar las diferencias entre mayúsculas y minúsculas (Ejemplo: “Hola”, “hola”, “hOlA”, “HOLA”). Este paso consiste simplemente en convertir todo a minúsculas.

Una deformación común en Internet es la modificación de la sintaxis de las palabras repitiendo caracteres dentro de la misma con el objetivo de enfatizarla (Ejemplo: “hooooooola”).

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Generalmente, la mayor o menor presencia de este efecto está muy ligado al tipo de contenido con que se trabaja y dentro de este al estilo propio del autor.

E. Tokenizacion

Para analizar el texto es necesario encontrar una unidad básica de información. En este trabajo se utiliza la palabra como elemento básico. Se entenderá por palabra inicialmente el resultado de dividir el tweet por espacios. Luego cada palabra es verificada contra el diccionario de jerga (después de convertirla a minúsculas), y si no se considera jerga se divide la palabra en los grupos continuos de puntuación y de caracteres alfabéticos. Estos últimos continúan siendo procesados como posibles palabras nuevas. Las nuevas palabras se vuelven a comparar con jerga y se les suprimen los caracteres repetidos hasta dos ocurrencias para luego verificar la forma correcta de la palabra mediante los algoritmos de corrección ortográfica.

F. Stop Words

Los stop words son palabras que no suelen aportar información y serán eliminadas una vez que sean encontradas en el texto. Los stop words cambian en función del problema que se esté analizando, pues lo que es

relevante para uno, puede ser prescindible para otro. En algunas ocasiones se puede relacionar el concepto de stop words con la función gramatical que realiza el término en la oración. Existe un conjunto general de stop words que casi siempre son eliminados. Este está compuesto fundamentalmente por artículos, preposiciones, conjunciones y pronombres. Además, se agregan otros términos según la depuración que exija el problema. La eliminación de stop words permite a los algoritmos concentrarse en los conceptos fundamentales y reduce la dimensión del problema. Esto mejora el rendimiento del algoritmo, ya que tiene que lidiar con una menor cantidad de términos.

2.3.4. Clasificación

Como se mencionó en el apartado 3.3.2.4, existen dos enfoques con los cuales se puede realizar la clasificación de tweets; siendo las técnicas de aprendizaje computacional supervisado las dominantes en esta clase de tareas.

La clasificación de los tweets se divide en dos etapas:

Clasificación en Objetivo – Subjetivo

Clasificación de los mensajes subjetivos en positivo – negativo – neutro.

En la primera etapa se separa a los tweets que contienen una opinión (subjetivos) de los que presentan un hecho (objetivos). Luego de obtener los tweets subjetivos, se procede a la clasificación de los mensajes en Positivo – Negativo – Neutro. En ambas etapas se utilizan métodos de aprendizaje supervisados.

A. Clasificadores Supervisados

Para Márquez, Maurelia y Mella (2013), las técnicas basadas en aprendizaje automático computacional, específicamente los métodos supervisados son los más utilizados, los mismos que constan de un proceso de entrenamiento en base a ejemplos entregados humanos en los cuales se le indica explícitamente al sistema a que clase pertenece cada ejemplo. Cada texto es definido en base a sus features (características), las cuales pueden ser unigramas (1 palabra), bigramas (conjunto de 2 palabras), trigramas (conjunto de 3 palabras), etc. En general, ciertos sentimientos son expresados con dos o más palabras, lo cual abarca características importantes como la negación de una frase, donde anteponiendo un “NO” a una frase, es posible cambiar el significado total de ésta. En general, los métodos supervisados consisten en dos etapas: el entrenamiento del sistema y la clasificación de nuevos datos entregados, en

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

cada una de estas etapas es necesario la selección y extracción de las features asociadas a cada dato.

Basándonos en las investigaciones de estos autores, los cuales indican que el Support Vector Machine (SVM) y Naive Bayes (NB), son dos de los métodos supervisados más utilizados; estos se utilizaran en el presente trabajo para poder realizar la evaluación y posteriormente publicar los resultados.

a. Support Vector Machine

Según Montesinos (2014), el Support Vector Machine (SVM) es altamente usado en la clasificación y detección de sentimientos. SVM se basa en métodos kernel, los cuales toman los datos y los ponen dentro de un espacio de características apropiado. De esta manera usan algoritmos lineales para determinar patrones no lineales. El método se basa principalmente en vectores donde, usando aprendizaje computacional, logra tomar decisiones de límite entre dos categorías separándolas lo más posible.

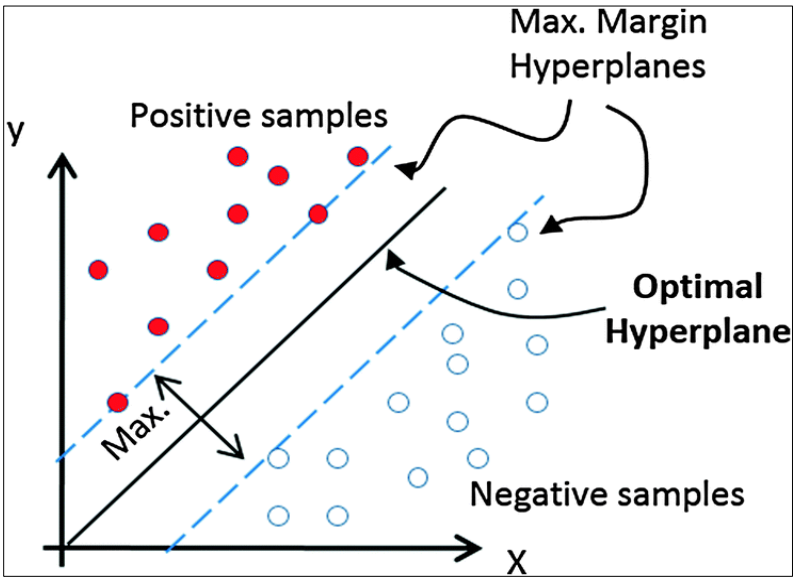


Figure 2. Esta imagen muestra la representación gráfica del Support Vector Machine. Fuente: http://pubs.rsc.org/.

SVM fija el criterio de separación entre clases que esté lo más lejos posible de cualquier dato. Esta distancia, del punto de decisión, al punto más cercano es el margen del clasificador. Es así, como el método queda definido por una función de decisión que involucra un subconjunto de características o datos (support vectors) que definirán la posición del separador. De esta manera, la decisión del límite o margen es bastante importante ya que los datos que queden en torno a este tendrán una menor probabilidad de ser catalogados correctamente.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Algebraicamente, se puede definir un vector perpendicular al híper-plano $\vec{\omega}$ que es conocido como el vector de peso (weight vector). Para determinar un solo híper-plano se especifica un término de intersección b. Así, todos los términos del híper-plano $\vec{x}$ satisfacen $\vec{\omega}\, T\, \vec{x} = -b$, ya que el híper-plano es perpendicular al vector normal $\vec{\omega}$.

Luego, para tomar las decisiones entre ambas clases, generalmente las clases se pueden definir con +1 y -1, se calcula $\vec{\omega}\, T\, \vec{x}$ y se compara con b para determinar a qué lado del híper-plano se encuentra $\vec{x}$, de manera que $f(\vec{x}) = sign(\vec{\omega}\, T\, \vec{x} + b)$, nos da la clasificación esperada (+1 o -1). Por otro lado, si el nuevo dato $\vec{x}$ está muy cerca del híper-plano de separación, suele no asignarse ninguna de las dos categorías, lo cual se hace fijando un límite se distancia. Finalmente ($\vec{x}$) puede ser transformada en una probabilidad de clasificación de manera de tomar decisiones entre las clases.

Este método fue actualizado y usado para clasificación de texto por Joachims en 1999. De esta manera, se tiene un set de entrenamiento donde cada muestra tiene un peso y un vector asociado que separa lo más posible los casos positivos de los negativos. Generalmente, los datos

usados son palabras (unigramas) a las cuales se les asigna un peso durante la fase de aprendizaje con el valor $\delta \geq 0$. Cada palabra etiquetada que cumpla que su peso $\delta > 0$ es llamado support vector. De esta manera los support vectors separa el híper-plano entre la clasificación positiva y negativa. Así, las palabras que aún no han sido entrenadas, son asignadas a los support vectors más cercanos de acuerdo a una ecuación que incluye la función kernel apropiada

Para seleccionar las características a ocupar en SVM correctamente hay varios métodos. Usualmente se ocupan palabras solas que se usen una cierta cantidad de veces en el texto a analizar. También es posible seleccionar bi-grams (dos palabras juntas), tri-grams (3 palabras juntas), la categoría gramatical de la palabra, etc

Este método es bastante usado en la clasificación de sentimientos, el cual ha tenido excelentes resultados tanto en Twitter como en otras plataformas en la web, logrando un acierto en más del 70% de los casos

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

b. Naive Bayes

Naive Bayes (NB) es uno de los métodos más usados en análisis de sentimientos, debido a su fácil implementación y a los buenos resultados obtenidos en la mayoría de los casos. NB es un método probabilístico de aprendizaje donde la probabilidad de que un documento d pertenezca a la clase c está dada por:

$$\mathrm{P(c|d)} \propto \mathrm{P(c)} \prod_{1 \le \mathrm{k} \le n_d} P\ (t_k \mid c)$$

Donde $(tk|c)$ es la probabilidad condicional del término tkocurra en un documento de clase c y nd el número de términos en el documento d. El objetivo principal de NB es obtener la mejor clase que se adapte al documento o conjunto de palabras, para lo cual se calcula el maximun a posteriori (MAP) de la clase c, donde:

$$C_{map} = arg \max \hat{P}(\mathrm{c|d})$$

$$= arg \max \hat{P}(c) \prod_{1 \le \mathrm{k} \le n_d} \hat{P}\ (t_k \mid c)$$

Donde no se conoce las probabilidades de c ni de $t_k \mid c$, pero se pueden estimar a partir de un set de

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

entrenamiento. Una mejor manera de calcular las probabilidades anteriores es a través del logaritmo, de manera sumar en vez de multiplicar las probabilidades condicionales, así la probabilidad $\hat{P}\ (t_k|\ c)$ es un peso que indica que tan buena es la palabra tk para predecir una clase c.

Para calcular $\hat{P}\ (c)$ y $\hat{P}\ (t_k|\ c)$ se ocupa el maximun likelihood estimate (MLE) que tiene relación con la frecuencia con la que aparecen dichas características en el set de entrenamiento. Así, $\hat{P}(\mathrm{c}) = \frac{N_C}{\mathrm{N}}$, donde NC es el número de documentos en la clase c y N es el número total de documentos. Por otro lado, $\hat{P}\ (t_k|c)$ = $\frac{T_C}{\Sigma T_C}$ donde T_C es el número de presencia del término t en la clase c. Suele sumarse 1 a T_C tanto en el numerador como denominador para que no existan probabilidades 0, dado un set de entrenamiento, ya que no siempre es posible cubrir el total de términos usados.

Este método considera el texto como un conjunto de palabras, donde la frecuencia de cada una de ellas es esencial para clasificarlas, por lo cual es importante tener

un set de entrenamiento de gran tamaño ya que de este set dependerá la precisión de los resultados.

El método de clasificación para el caso de análisis de sentimientos se puede reducir a:

$$\mathrm{P(c|d)} = \frac{\mathrm{P(c)P(d|c)}}{\mathrm{P(d)}}$$

En este caso, $(c|d)$ es la probabilidad de que dada una palabra d, ésta corresponda a una clase específica c, ya sea bueno, malo o cualquier otra clase que se quiera determinar (miedo, sorpresa, felicidad, enojo, etc). Es así, que obteniendo esta probabilidad para cada una de las palabras dadas en un texto, es posible determinar su polaridad final.

La probabilidad $(d|c)$ determina la probabilidad de que la palabra esté, dada una cierta clase, la cual es extraída directamente del set de entrenamiento, tal como se explicó anteriormente, donde las clases ya están determinadas y basándose en la frecuencia de las palabras se puede obtener la probabilidad correspondiente. En este caso, se asume que la probabilidad que ocurra una palabra es

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

independiente de otra, de manera que sea más simple realizar el cálculo, con lo cual (d) es simplemente un factor de normalización. Si bien, esta es una asunción fuerte, los resultados obtenidos son buenos, superando el 65% de certeza en la mayoría de los casos, al ocupar 2 categorías de clasificación (a favor y en contra). Generalmente, antes de ocupar Naive Bayes se obtiene el conjunto de entrenamiento con el uso del diccionario léxico, asegurando un conjunto lo más preciso posible. Otras veces, se crea un set de entrenamiento "a mano", clasificando cada texto de forma manual, sin métodos computacionales. Montesinos (2014).

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2.4. Definición de la terminología

Minería de datos. - La minería de datos o exploración de datos es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.

Minería de Textos. - La minería de textos se refiere al proceso de derivar información nueva de textos.

Análisis Sentimental. - hace referencia a la tarea de análisis, identificación y clasificación de todo tipo de contenido emocional, subjetivo u opinado

Extracción Transformación y Carga (ETL).- es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart, o data warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

Corpus Lingüístico. - es un conjunto, habitualmente muy amplio, de ejemplos reales de uso de una lengua. Estos ejemplos pueden ser textos (lo más común) o muestras orales (generalmente transcritas).

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Subjetividad. - es la propiedad de las percepciones, argumentos y lenguaje basados en el punto de vista del sujeto, y por tanto influidos por los intereses y deseos particulares del mismo. Su contrapunto es la objetividad, que los basa en un punto de vista intersubjetivo, no prejuiciado, verificable por diferentes sujetos.

KDD. - es la extracción automatizada de conocimiento o patrones interesantes, no triviales, implícitos, previamente desconocidos, útiles y predictivos de la información de grandes Bases de Datos

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO III

MARCO METODOLOGICO

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

III. CAPITULO III: MARCO METODOLOGICO

3.1. Tipo y Diseño de Investigación

Tipo

El presente trabajo corresponde a una investigación de tipo Cuantitativa, aplicada y tecnológica porque interviene en los conocimientos científicos dando el apoyo en la ciencia de la computación en que sus resultados es resolver los problemas reales en la ciencia tecnológica.

Diseño

De acuerdo al tipo de investigación del diseño utilizado es Cuasi Experimental, debido al generar interrogantes mediante de las hipótesis se permite en resolver la circunstancia por efecto de su naturaleza y de no conocer una selección aleatoria

$$M \leftarrow XY$$

Donde:

X : Causa

Y : Efecto

M : Muestra

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.2. Población y Muestra

Población

Está definida por todas las opiniones expresadas en algún tópico tendencia, denominado hashtag (etiqueta), según la web oficial de Twitter, las etiquetas (escritas con el signo "#" antepuesto) se usan para indexar palabras claves o temas en Twitter. Esta función es una invención de Twitter y permite que los usuarios puedan seguir fácilmente los temas que les interesan.

Al consultar una etiqueta o tópico en red social, esta devolverá un número determinado de comentarios estructurados en:

ÍTEM	DESCRIPCIÓN
ID	Código de entrada
RED SOCIAL	Identificador de red social
TÓPICO	Tópico o tema de interés
ENTRADA (COMENTARIO)	Comentario realizado
USUARIO	Usuario
FECHA	Fecha
CLASE	Positivo / Neutro / Negativo (Entrenamiento) Desconocido (Validación)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Muestra

Está definida por los tópicos y la cantidad de comentarios usados para la investigación. El protocolo de muestra es aplicar muestreo estratificado a una serie de tópicos elegidos al azar por el investigador, tendiendo como parámetros el nombre de tópico, categoría del mismo (clasificada por el autor de la investigación) y la fecha de consulta, obteniendo una cantidad de comentarios, para la investigación se seleccionó 10 Hashtag del tipo top (Tendencia), por generar mucho tráfico de comentarios, se realizó en dos fechas diferentes y se aplicó el muestreo estratificado aplicando factores de reducción proporcionales a cada tópico, donde cada tópico es un estrato.

Muestreo estratificado proporcionado

Suponiendo que hay estratos con cantidades N

$$N = N_1 + N_2 + \ldots + N_k$$

En cada estrato se toman n muestras

$$n = n_1 + n_2 + \ldots + n_k$$

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Donde:

$$n_i = n \cdot \frac{N_i}{N}$$

siendo N el número de elementos de la población, n el de la muestra, N_i el del estrato i

Red Social	Tópico	Comentarios	(Aplica Factor)	(Aplica Factor)	(Factor 1.45%)
Twitter	Tópico 1	14351	0.04	583	8
Twitter	Tópico 2	25896	0.07	1900	27
Twitter	Tópico 3	24787	0.07	1741	25
Twitter	Tópico 4	35896	0.10	3650	55
Twitter	Tópico 5	12253	0.03	425	6
Twitter	Tópico 6	15869	0.04	713	10
Twitter	Tópico 7	87456	0.25	21669	278
Twitter	Tópico 8	65241	0.18	12059	153
Twitter	Tópico 9	35684	0.10	3607	51
Twitter	Tópico 10	35541	0.10	3579	52
TOTAL		352974	1	49927	665

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Al aplicar el muestreo se obtiene un total de 49927 comentarios, por lo que se ha aplicado un segundo factor de reducción para reducir la cantidad de población.

3.3. Hipótesis

El algoritmo de clasificación que tendrá mejores resultados para el minado de opinión en twitter es el árbol de regresión.

3.4. Operacionalización

Variable Independiente

Evaluación de algoritmos de clasificación.

Variable Dependiente

Minado de opinión en Twitter.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Operacionalización de Variables

Indicador	Medida o técnica	Formula	Frecuencia
Exactitud	Matriz de confusión	$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ TP=True Positive (Verdadero Positivo) TN=True Negative (Verdadero Negativo) FP=False Positive (Falso Positivo) FN=False Negative (False Negativo)	Por escenario de prueba
Confiabilidad	Promedio	$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{N}$ P=SUMATORIA(EP)/TEP P=PROMEDIO EP=ESCENARIO DE PRUEBA TEP=TOTAL ESCENARIO DE PRUEBA	Una sola vez aplicado al promedio de todos los escenarios
Procesamiento	Promedio	$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{N}$ P=SUMATORIA(EP)/TEP P=PROMEDIO EP=ESCENARIO DE PRUEBA TEP=TOTAL ESCENARIO DE PRUEBA	Una sola vez aplicado al promedio de todos los escenarios

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.5. Métodos, técnicas e instrumentación de recolección de datos

El método de evaluación será de tipo experimental, para el cual se establece dos frentes, la validación por parte de un grupo de individuos en laboratorio de tipo juez experto, y las validaciones realizadas por el modelo con el uso de técnicas propias de la metodología de desarrollo de modelos de minería de datos.

3.6. Procedimiento para la recolección de datos

La recolección de datos para el presente trabajo se realizará básicamente de la siguiente manera:

- Darse de alta en la red social Twitter.
- Crear una aplicación para poder obtener la clave y credenciales de acceso.
- Conectarse con el API Stream.
- Seleccionar un hashtag o nombre de usuario para poder acceder a sus datos.
- Guardar los tweets en una base de datos; previamente normalizados.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.7. Análisis Estadístico e interpretación de resultados

Como parte de un enfoque cuantitativo los datos serán evaluados con la estadística descriptiva donde se le aplicara la media aritmética (promedio), la cual consiste en el valor obtenido al sumar todos los datos y dividir el resultado entre el número total de datos. Esta fórmula se utiliza para calcular la confiabilidad y rendimiento, las cuales utiliza la fórmula de promedio.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{N}$$

También se aplica el cálculo para la exactitud o precisión del modelo, evaluando asi los algoritmos clasificadores, la técnica, es aplicar la matriz de confusión, según (Piehadrita, 2013) "*La matriz de confusión, propuesta por Kohavi y Provost en 1998, contiene información sobre los índices de clasificación realizado por un sistema de reconocimiento.",* cuya fórmula es:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.8. Criterios éticos

Los criterios éticos que se respetan en el presente proyecto de tesis es el Código Deontológico del Colegio de Ingenieros de Perú en su Capítulo II “De la Relación con el Público” en su artículo 106 expresa:

Los ingenieros, al explicar su trabajo, méritos o emitir opiniones sobre temas de ingeniería, actuarán con seriedad y convicción, cuidando de no crear conflictos de intereses, esforzándose por ampliar el conocimiento del público a cerca de la ingeniería y de los servicios que presta a la sociedad.

3.9. Criterios de rigor científico

La presente propuesta de investigación se realiza siguiendo los juicios científicos establecidos, estos permiten garantizar la calidad de la propuesta de investigación.

Así, seguimos la coherencia metodológica durante el desarrollo de la propuesta de la investigación, según el muestreo de datos, los cuales son al azar para ser totalmente imparcial en el recojo de datos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO IV

ANALISIS E INTERPRETACION DE LOS RESULTADOS

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

IV. CAPITULO IV: ANALISIS E INTERPRETACION DE LOS RESULTADOS

Generalidades

Muestra

De la muestra de 661 comentarios, se ha procedido a utilizarlos en la siguiente composición aleatoria:

Comentarios		
Entrenamiento	Negativos	40 %
	Positivos	40 %
	Sub Total	**80 %**
Validación	Negativos	10 %
	Positivos	10 %
	Sub Total	**20 %**

Condiciones

Las condiciones de prueba del modelo se realizaron en un computador y red con las siguientes características.

CPU	Intel Core i7 2 núcleos virtualizado
RAM	6 GB RAM
Disco	Disco Duro 100 GB
Red	15 Mbps download 1.5 Mbps upload

4.1. Resultados en tablas y gráficos

Indicador Exactitud

a. Seleccionar los algoritmos de clasificación.
b. Diseñar y construir un modelo clasificador para el minado de opiniones en Twitter
c. Evaluar los algoritmos de clasificación seleccionados.

Escenario 1

Escenario						Matriz	P	N
1	IdComentario	Real	NB	SVM	Tree	P	VP	FP
1	127	1	1	1	1	N	FN	VN
1	128	1	1	1	1	MC		
1	131	1	1	1	2			
1	136	1	1	1	1	NB	P	N
1	138	1	1	1	2	P	9	8
1	139	1	1	1	1	N	1	2
1	140	1	1	1	1	MC	55%	
1	142	1	1	1	1			
1	143	1	2	1	1	SVM	P	N
1	144	1	1	1	1	P	10	0
1	87	2	2	2	2	N	0	10
1	88	2	2	2	2	MC	100%	
1	91	2	1	2	2			
1	92	2	1	2	2	TREE	P	N
1	93	2	1	2	2	P	8	0
1	94	2	1	2	2	N	2	10
1	95	2	1	2	2	MC	90%	
1	96	2	1	2	2			
1	97	2	1	2	2			
1	98	2	1	2	2			

En el escenario 01 se obtiene un 55 % de exactitud para la red bayesiana, un 100 % para SVM y un 90 % para el árbol de regresión.

Escenario 02

Escenario						Matriz	P	N
	IdComentario	Real	Naive E	SVM	Tree	P	VP	FP
2	145	1	1	1	1	N	FN	VN
2	146	1	1	1	1	MC		
2	148	1	1	1	1			
2	153	1	1	1	1	NB	P	N
2	154	1	1	2	1	P	9	1
2	155	1	1	1	1	N	10	0
2	158	1	2	1	1	MC	45%	
2	159	1	1	1	1			
2	162	1	1	1	2	SVM	P	N
2	168	1	1	1	1	P	9	1
2	99	2	1	2	2	N	1	9
2	100	2	1	2	2	MC	90%	
2	101	2	1	2	2			
2	102	2	1	2	2	TREE	P	N
2	103	2	1	2	2	P	9	1
2	105	2	1	1	1	N	1	9
2	106	2	1	2	2	MC	90%	
2	107	2	1	2	2			
2	108	2	1	2	2			
2	110	2	1	2	2			

En el escenario 02 se obtiene un 45 % de exactitud para la red bayesiana, un 90 % para SVM y un 90 % para el árbol de regresión.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Escenario 03

Escenario 03						Matriz	P	N
	IdComentario	Real	Naive E	SVM	Tree	P	VP	FP
3	169	1	1	1	1	N	FN	VN
3	170	1	1	1	2	MC		
3	171	1	1	1	1			
3	172	1	1	1	2	NB	P	N
3	174	1	1	1	1	P	9	1
3	175	1	1	1	2	N	9	1
3	181	1	1	1	2	MC	50%	
3	184	1	1	2	2			
3	186	1	2	1	1	SVM	P	N
3	188	1	1	1	1	P	9	1
3	111	2	1	2	2	N	6	4
3	115	2	1	2	2	MC	65%	
3	116	2	1	2	2			
3	117	2	1	2	2	TREE	P	N
3	119	2	1	1	1	P	9	1
3	120	2	1	1	2	N	2	8
3	121	2	1	1	1	MC	85%	
3	123	2	1	2	2			
3	134	2	1	1	2			
3	135	2	2	2	2			

En el escenario 03 se obtiene un 50 % de exactitud para la red bayesiana, un 65 % para SVM y un 85 % para el árbol de regresión.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Escenario 04

Escenario 04						Matriz	P	N
	IdComentario	Real	Naive E	SVM	Tree	P	VP	FP
4	189	1	1	1	2	N	FN	VN
4	190	1	1	1	1	MC		
4	191	1	1	1	1			
4	192	1	1	2	2	NB	P	N
4	194	1	1	1	2	P	10	0
4	196	1	1	2	1	N	9	1
4	202	1	1	2	2	MC	55%	
4	203	1	1	1	2			
4	209	1	1	2	2	SVM	P	N
4	210	1	1	1	2	P	4	6
4	149	2	1	1	1	N	9	1
4	150	2	1	1	2	MC	25%	
4	152	2	1	1	2			
4	183	2	1	1	2	TREE	P	N
4	193	2	2	2	2	P	3	7
4	206	2	1	1	2	N	2	8
4	207	2	1	1	2	MC	55%	
4	216	2	1	1	2			
4	246	2	1	1	1			
4	252	2	1	1	2			

En el escenario 04 se obtiene un 55 % de exactitud para la red bayesiana, un 25 % para SVM y un 55 % para el árbol de regresión.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Escenario 05

Escenario 05						Matriz	P	N
	IdComentario	Real	Naive B	SVM	Tree	P	VP	FP
5	211	1	1	1	1	N	FN	VN
5	213	1	1	1	1	MC		
5	214	1	1	1	1			
5	226	1	1	1	1	NB	P	N
5	229	1	1	1	1	P	10	0
5	231	1	1	1	1	N	10	0
5	232	1	1	1	1	MC	50%	
5	233	1	1	1	1			
5	234	1	1	1	1	SVM	P	N
5	235	1	1	2	1	P	9	1
5	258	2	1	2	2	N	5	5
5	277	2	1	2	2	MC	70%	
5	299	2	1	1	1			
5	317	2	1	2	2	TREE	P	N
5	323	2	1	1	2	P	9	1
5	327	2	1	1	1	N	2	8
5	344	2	1	1	2	MC	85%	
5	349	2	1	2	2			
5	350	2	1	1	2			
5	359	2	1	2	2			

En el escenario 05 se obtiene un 50 % de exactitud para la red bayesiana, un 70 % para SVM y un 85 % para el árbol de regresión.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Indicador Confiabilidad

a. Seleccionar los algoritmos de clasificación.
b. Diseñar y construir un modelo clasificador para el minado de opiniones en Twitter
c. Evaluar los algoritmos de clasificación seleccionados.
d. Analizar el ámbito donde se aplicarán los algoritmos de clasificación.

Se implementa la matriz para promediar todos los escenarios de prueba

Escenario	NB	SVM	TREE
1	55%	100%	55%
2	45%	90%	90%
3	50%	65%	85%
4	55%	25%	55%
5	50%	70%	85%
Promedio	51%	70%	74%

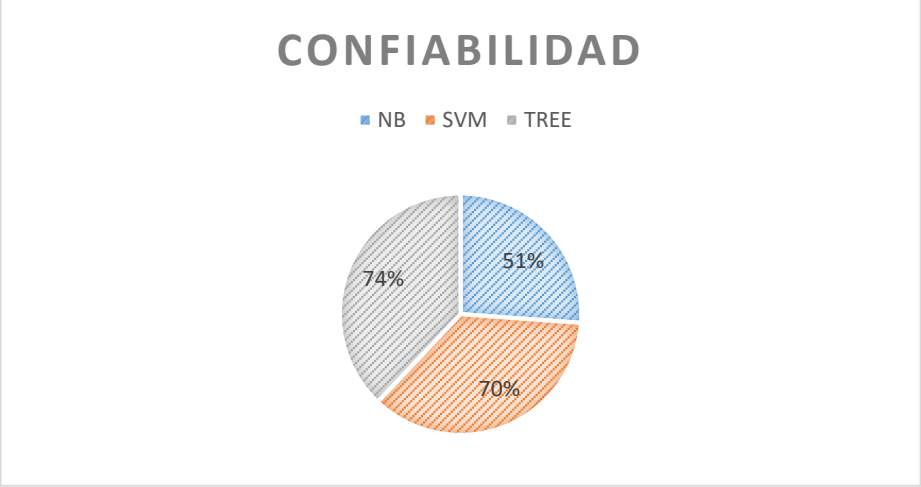


Figure 3. Porcentaje de confiabilidad de los tres algoritmos evaluados
Fuente: Elaboración propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Indicador Procesamiento

a. Diseñar y construir un modelo clasificador para el minado de opiniones en Twitter
b. Evaluar los algoritmos de clasificación seleccionados.
c. Determinar una estrategia para la extracción y tratamientos de tweets.
d. Implementar una aplicación para el análisis y visualización de resultados.

Naive Bayes

NB	ESCENARIO	user	system	elapsed
1a	1	20.17	0.00	21.27
5a	5	21.60	0.00	21.89
4a	4	21.17	0.00	21.93
3a	3	21.91	0.00	22.17
2a	2	22.21	0.00	22.69
	Promedio	21.41	0.00	21.99

SVM

SVM	ESCENARIO	user	system	elapsed
1b	1	26.30	0.00	27.20
5b	5	26.55	0.00	27.65
4b	4	26.50	0.00	27.10
3b	3	26.39	0.00	26.90
2b	2	26.56	0.00	27.25
	Promedio	26.46	0.00	27.22

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

TREE

TREE	ESCENARIO	user	system	elapsed
1c	1	32.59	0.14	33.78
5c	5	33.23	0.05	33.89
4c	4	32.80	0.11	34.03
3c	3	33.37	0.07	34.58
2c	2	33.21	0.19	34.60
	Promedio	33.04	0.11	34.18

4.2. Discusión de resultados

De lo resultados se puede discutir lo siguiente:

En el indicador exactitud, sometido a 05 escenarios de prueba, se obtienen diversos resultados, esto implica, que la naturaleza de los textos tratados, genera una distorsión o ruido que los algoritmos deben tratar. En la mayoría de los casos se observa una red bayesiana con baja exactitud, mientras que SVM y TREE mantienen valores cercanos.

Esto se puede corroborar en el indicador de confiabilidad, que analiza los datos de las 05 iteraciones o escenarios de prueba, obteniendo en promedio que la red bayesiana consigue un 51 % de confiabilidad, mientras que SVM un 70 % no muy alejado de un árbol de regresión con 74 %.

Desde el punto de vista del rendimiento, se obtiene que el árbol, genera mayor tiempo y consumo de recursos que la red bayesiana y el SVM.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO V

DESARROLLO DE LA PROPUESTA

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

V. CAPITULO V: DESARROLLO DE LA PROPUESTA

Desarrollo basado en objetivos

Generalidades

a) **Analizar el ámbito donde se aplicarán los algoritmos de clasificación.**

La propuesta de investigación trata sobre realizar un análisis de algoritmos de clasificación para minado de opiniones en redes sociales.

En las redes sociales, la extracción, transformación y carga designa el conjunto de técnicas que se utilizan para mapear los datos del sistema de información existentes en los modelos de redes sociales.

Entidades presentes en los sistemas deben ser normalizados y resueltas, y las interacciones entre ellos seleccionados se transforman en relaciones.

Dominio Contexto

Se aplicará la investigación a la data generada de tipo no estructurada, definido por el contenido de las redes sociales:

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

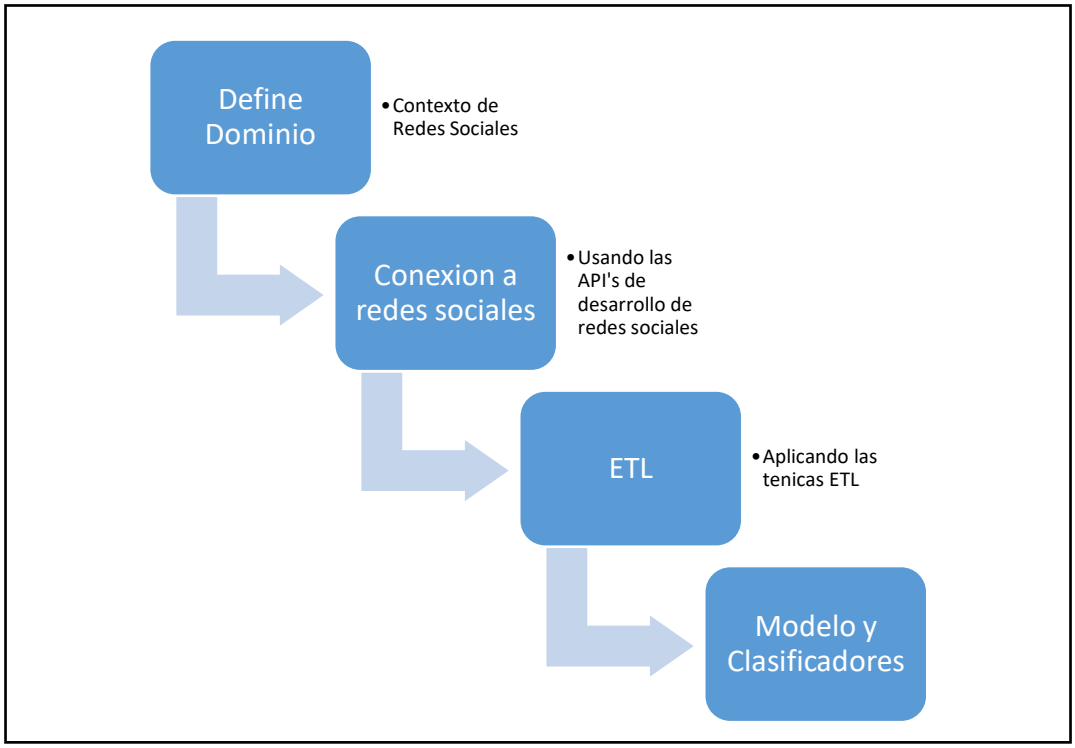


Figure 4. Flujograma de proceso en investigación partiendo del dominio

Elaboración Propia

ETL Orientado a Procesamiento Social Media	Tipo de registro o entrada	Característica	Ámbito
Twitter	Tweets	Micro texto	Múltiple Tópico
Facebook	Post	Texto	Múltiple Tópico
Instagram	Post	Micro texto – Imagen	Múltiple Tópico
Linkedin	Post	Micro texto - interés	Tópico Laboral

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

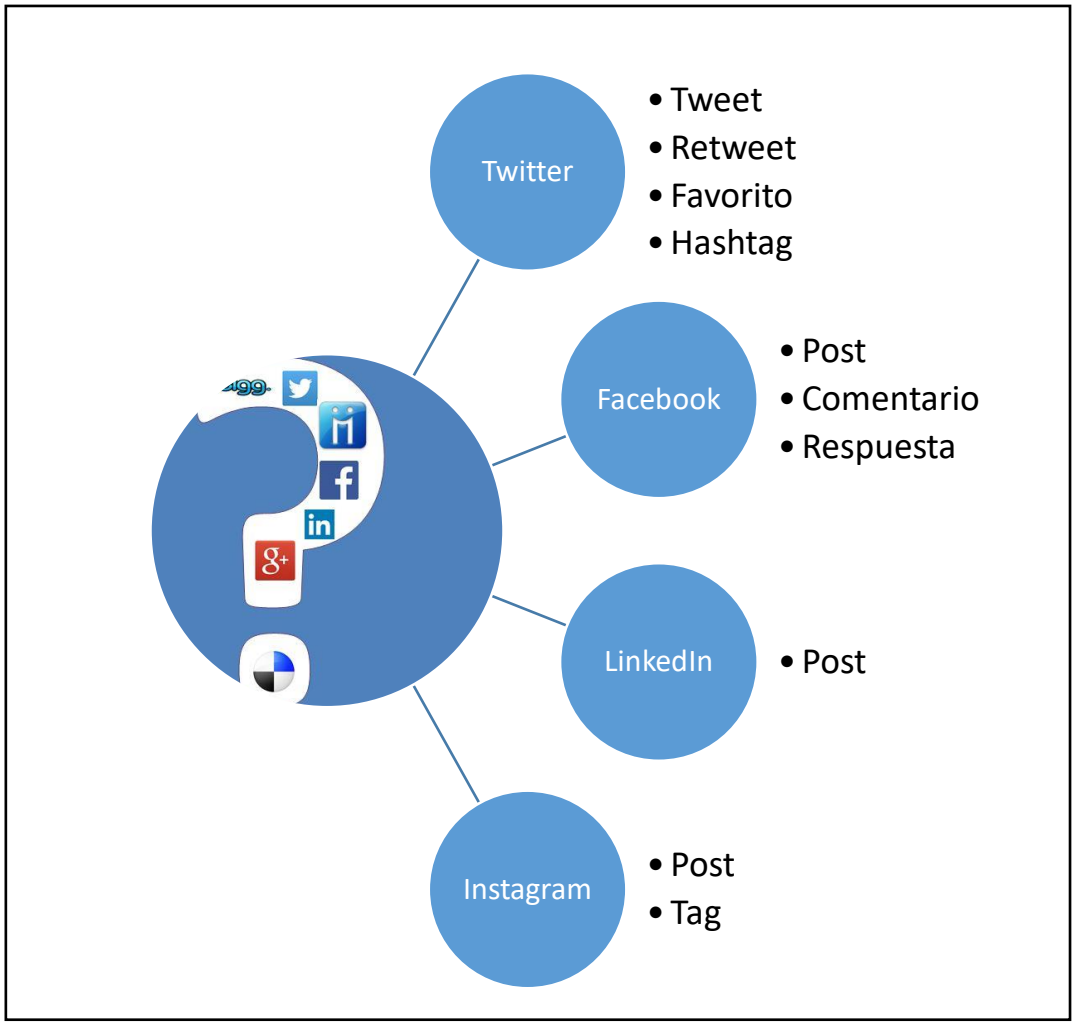


Figure 5. Tipo de entrada de datos

Elaboración Propia

Para este caso se selecciona a Twitter como ámbito de estudio de los algoritmos de clasificación, por ser una red social orientada a opiniones expresadas en micro texto, lo que permitirá minimizar las labores de tratamiento de textos e índices generados.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Twitter

Se procede a realizar el proceso de configuración de entorno de desarrollo API en twitter.

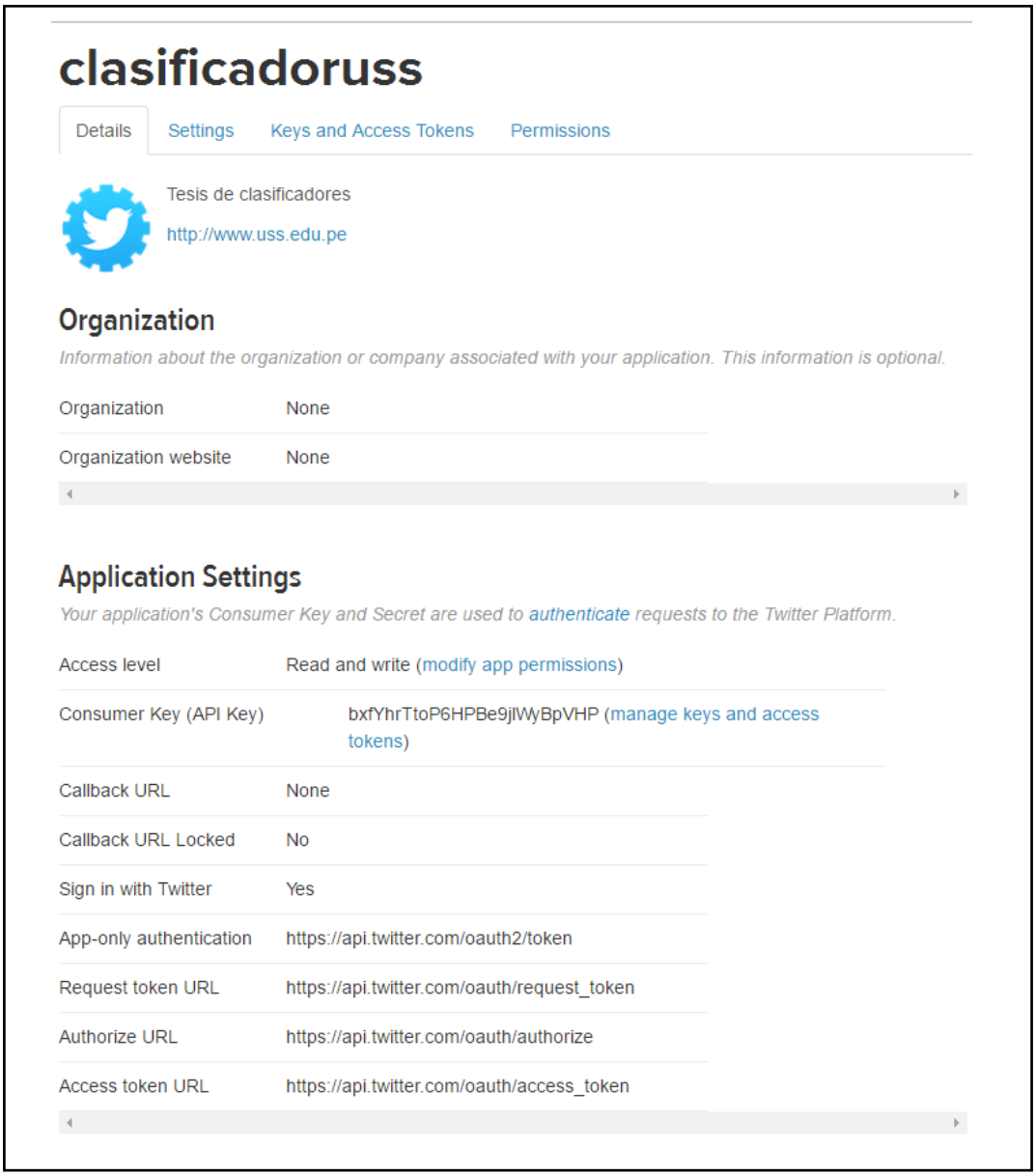


Figure 6. Configuración entorno API Twitter

Captura de Twitter

Una vez que está configurado las credenciales del API de twitter se procede a realizar una prueba de búsqueda, en este caso se utilizara los datos de un determinado Hashtag.

```
> library("twitteR")
> library("base64enc")
>
> download.file(url="https://curl.haxx.se/ca/cacert.pem", destfile="cacert.pem")
probando la URL 'https://curl.haxx.se/ca/cacert.pem'
Content type '' length 250607 bytes (244 KB)
downloaded 244 KB

>
> consumer_key <- '[redacted]'
> consumer_secret <- '[redacted]'
> access_token <- '281220172-byBlQyReDQFgyJdwUxNDQhQKRR8kgBFtGf9Cvca7'
> access_secret <- 'CwKNhMWxHfAkpR0RHp8B3Bj9l7u8mCYOUDRnKV08'
>
> setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_secret)
[1] "Using direct authentication"
```

Figure 07. Verificación de credenciales de API Twitter en R

```
> searchTwitter("#Maléfica", n=10)
[[1]]
[1] "maurilioleite_: Daí existem músicas que você não pode ouvir, porque te lembram momentos fofos ao$

[[2]]
[1] "MMoura73: RT @dan__2000: nem compareci no niver do marcelo pederasta,n fica mal brother,eu sei q$

[[3]]
[1] "gabriwlcrf: RT @CBFudeu: quando alguém pede pra ver o jogo comigo e fica falando mal do meu time$

[[4]]
[1] "_alcoutinho: RT @sacLibriano: As vezes a gente faz umas cagadas irreversíveis na vida, e depois $

[[5]]
[1] "ohfuckyouok: RT @oiestoubebado: vc bebe um poco fica alegre ai no dia seguinte todo mundo NOSSA $

[[6]]
[1] "MilenaStefane2: O povo fica falando mal dele pra mim nao admito mas isso."

[[7]]
[1] "RafaelCunha22: RT @CBFudeu: quando alguém pede pra ver o jogo comigo e fica falando mal do meu t$

[[8]]
[1] "Danny_Francisco: RT @bi_soarres: @Danny_Francisco mal acordo e a nani fica tirando fotinho jansn$
```

Figure 08. Verificación de extracción de datos básicos en Twitter

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

A continuación, el script utilizado para extraer los datos de Twitter.

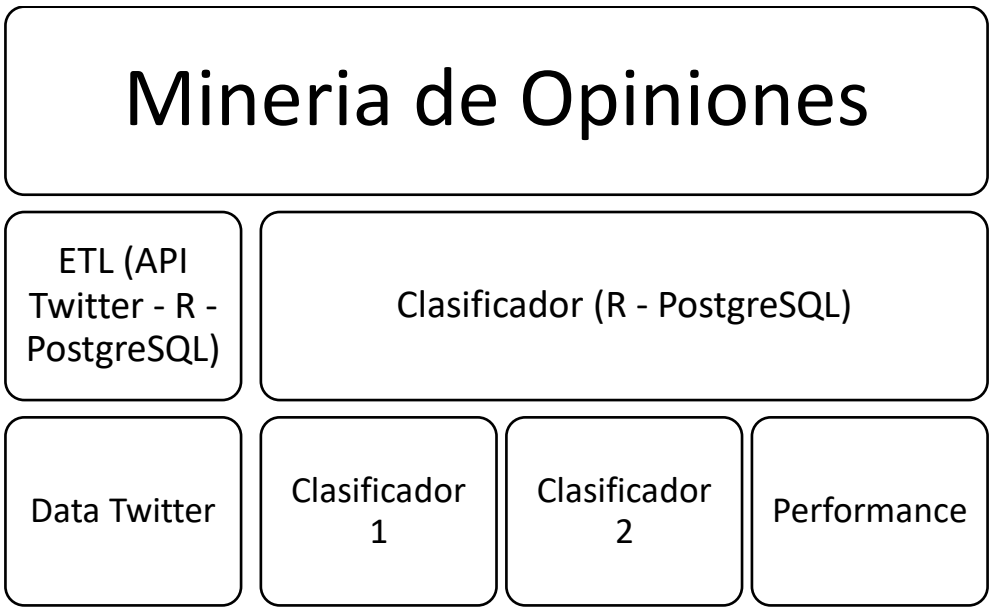
```
#https://apps.twitter.com/app/
#install.packages("twitteR")
#install.packages("wordcloud")
#install.packages("tm")
#install.packages("base64enc")

library("twitteR")
library("wordcloud")
library("tm")
library("base64enc")



#Permisos

download.file(url="https://curl.haxx.se/ca/cacert.pem", destfile="cacert.pem")

#API Twitter permisos

consumer_key <- [redacted]
consumer_secret <- [redacted]
access_token <- '281220172-byBlQyReDQFgyJdwUxNDQhQKRR8kgBFtGf9Cvca7'
access_secret <- 'CwKNhMWxHfAkpR0RHp8B3Bj917u8mCYOUDRnKV08'

setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_secret)
searchTwitter("#Maléfica", n=10)
```

Figure 09. Script del algoritmo para conectarse al API Twitter

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

b) Determinar una estrategia para la extracción y tratamientos de tweets.

Mapa de Estados



El artefacto es una función que tiene como entrada una lista de textos (comentarios), extraídos de las redes sociales para un determinado hashtag, el comentario viene con una cabecera (ID, COMENTARIO, CLASE).

El artefacto se compone en su interior de un subconjunto de funciones que realizaran un tratamiento sobre la lista de textos ingresados, divididos en etapas (Limpieza, Transformación, Modelo, Validación).

Cada subconjunto a su vez cuenta con métodos propios, por ejemplo, en el caso de Limpieza, se implementan métodos para eliminar palabras vacías (Artículos, Pronombres, etc.) que no tienen relevancia para definir una clase (Positivo, Negativo)

Una vez que el artefacto procesó los métodos y funciones internas, puede imprimir la salida de cada función interna para apreciar el avance y estado de cada una de ellas.

Al finalizar el artefacto imprime la lista de textos (Comentarios) con su clasificación obtenida.

En el caso de procedimiento de entrenamiento, a la colección o lista de textos o comentarios, se le agregara una notación (columna) que define su clase (Positivo, Negativo).

ETL

Ya en los anteriores ítems se ha demostrado la conexión con el API de Twitter para explorar los tweets, sin embargo, el procesamiento de grandes volúmenes de datos no debe realizar en memoria, por cuestiones de optimización como primera etapa se procede a generar una base de datos según el esquema de datos obtenido de Twitter.

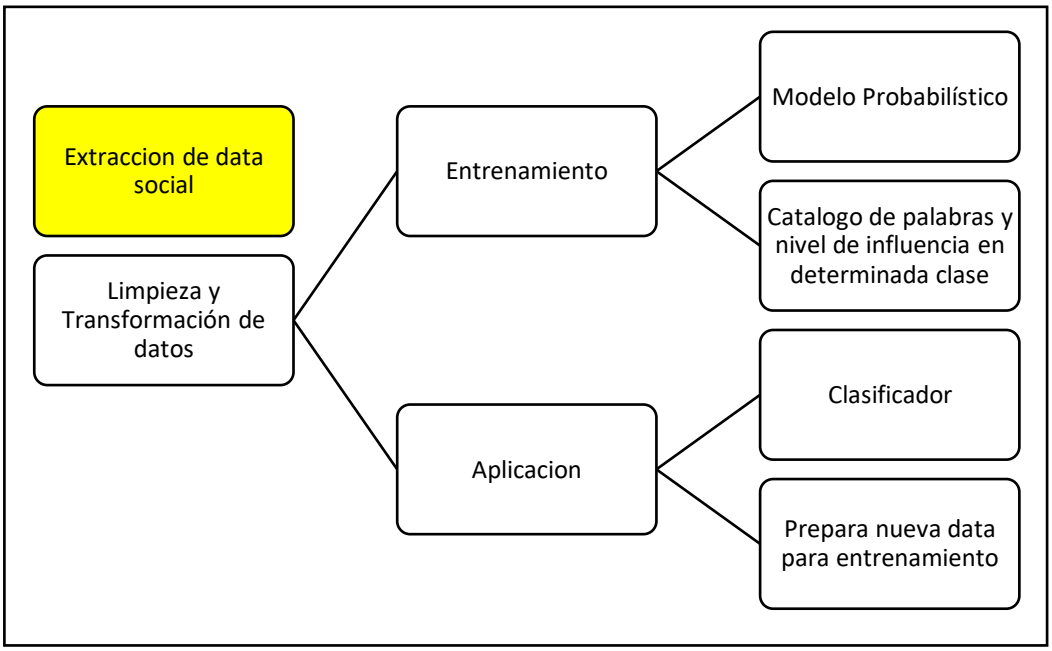


Figure 10. Estrategia Modelo – Fase Extracción de la Data Social

Elaboración Propia

idtopico [PK] integer	nomtopico character varying(200)	idcategoria integer	idredsocial integer	fechaconsulta date	comentarios integer
1	Manchester	1	1	2017-05-24	9
2	Alejandro Toledo	2	1	2017-05-24	28
3	Ingeniero	6	1	2017-06-08	26
4	Estado Islamico	1	1	2017-05-24	55
5	Venezuela	3	1	2017-06-08	6
6	Cristal	4	1	2017-05-24	11
7	Ariana Grande	3	1	2017-05-23	325
8	Movistar	5	1	2017-05-24	181
9	FelizLunes	6	1	2017-05-23	54
10	Emmanuel Macron	5	1	2017-06-08	54

Tabla 01. Formato de muestra de comentarios

La tabla anterior muestra los tópicos consultados, la fecha, el tipo de tópico, la red social y la cantidad (En función a la muestra, ver capítulo 3) de comentarios a extraer.

Una de las particularidades definidas en los lineamientos generales de esta investigación, es que esta extracción está en función al idioma español, filtro que se utilizara con la fecha y el nombre del tópico en la codificación de los algoritmos de extracción.

A partir de este esquema se ha diseñado un pequeño modelo relacional tomando teoría de Inteligencia de Negocios y Modelado de Base de datos para construir el repositorio de comentarios de la red social, este repositorio servirá para las fases de entrenamiento y validación así como la aplicación en sí.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Base de datos

El Repositorio analítico se encontrará en una base de datos PostgreSQL version 9.5

Al aplicar la extracción de datos de la red social se conectará vía API desde la interfaz de R y se extraerán los datos según el siguiente esquema:

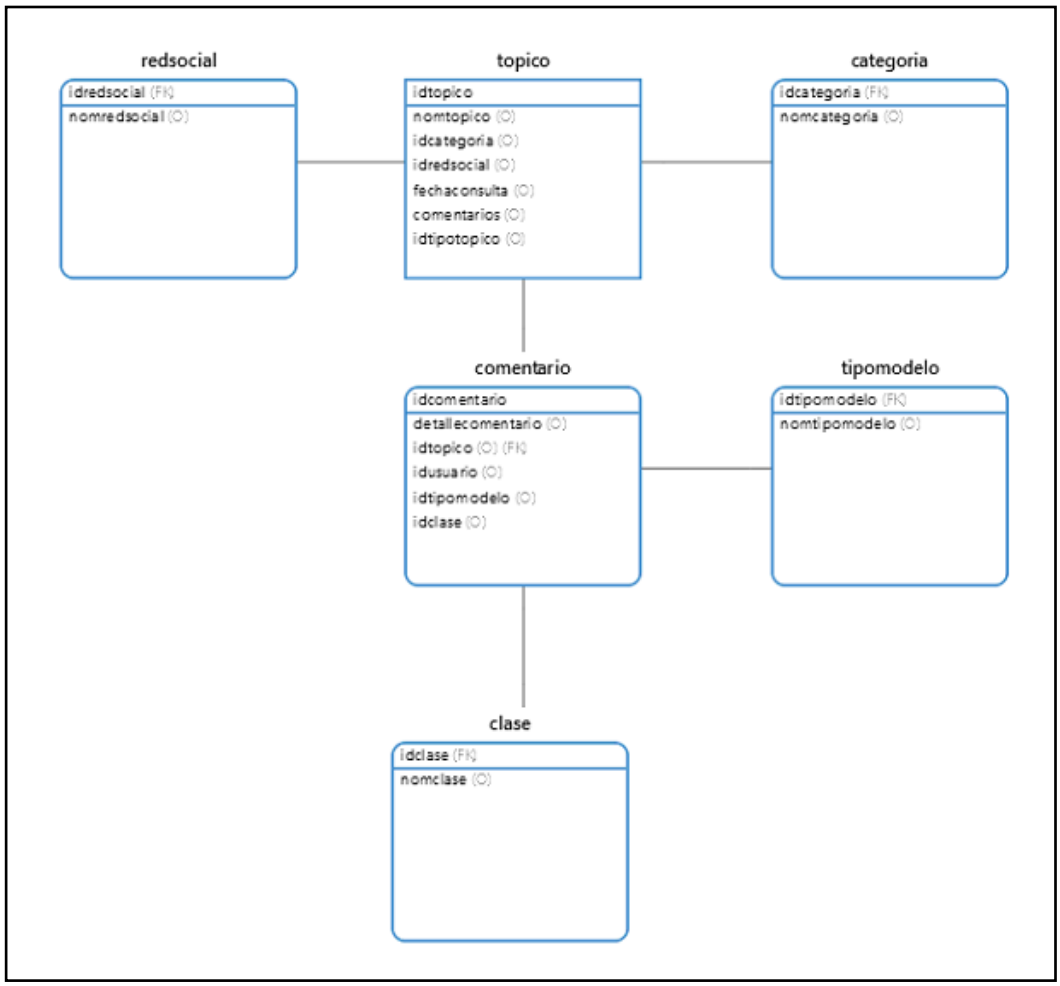


Figure 11. Base de datos de la Data Socia

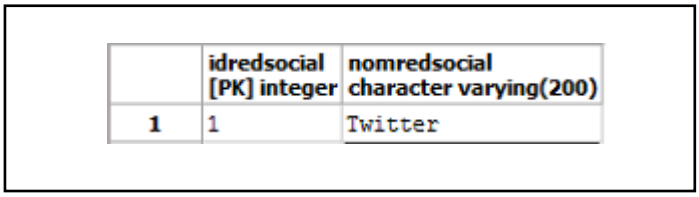
Elaboración Propial

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Donde se obtienen 5 Dimensiones y 1 Tabla Hecho

Dimensión Red Social

Esta dimensión contiene información relevante sobre las redes sociales que utilizara el modelo.



	idredsocial [PK] integer	nomredsocial character varying(200)
1	1	Twitter

Figure 12. Dimensión Red Social

Dimensión Tópico

Esta dimensión contiene información sobre los tópicos a analizar.

idtopico [PK] integer	nomtopico character varying(200)	idcategoria integer	idredsocial integer	fechaconsulta date	comentarios integer
1	Manchester	1	1	2017-05-24	9
2	Alejandro Toledo	2	1	2017-05-24	28
3	Ingeniero	6	1	2017-06-08	26
4	Estado Islamico	1	1	2017-05-24	55
5	Venezuela	3	1	2017-06-08	6
6	Cristal	4	1	2017-05-24	11
7	Ariana Grande	3	1	2017-05-23	325
8	Movistar	5	1	2017-05-24	181
9	FelizLunes	6	1	2017-05-23	54
10	Emmanuel Macron	5	1	2017-06-08	54

Figure 13. Dimensión Tópico

Elaboración Propia

Dimensión Categoría

Esta dimensión contiene información sobre las categorías de los tópicos a analizar.

	idcategoria [PK] integer	nomcategoria character varying(200)
1	1	Internacional
2	2	Politico
3	3	Entretenimiento
4	4	Deporte
5	5	Empresarial
6	6	Variado

Figure 14. Dimensión Categoría

Dimensión TipoModelo

Esta dimensión contiene el tipo de objetivo que tendrá cada elemento para la tabla Hecho, por ejemplo, si el comentario está destinado a ser usado para entrenamiento, validación o aplicación.

	idtipomodelo [PK] integer	nomtipomodelo character varying(200)
1	1	Entrenamiento
2	2	Validacion
3	3	Aplicacion

Figure 15. Dimensión Tipo Modelo

Dimensión Clase

Esta dimensión contiene los tipos de resultado que se espera del modelo (Clasificación Negativo, Positivo, Neutral).

	idclase [PK] integer	nomclase character varying(20)
1	1	Positivo
2	2	Negativo
3	3	Neutro
4	4	Desconocido

Figure 16. Dimensión Clase

Hecho Comentario

En esta entidad se encuentran alojados los comentarios de la red social, para que sean procesados por el modelo.

Poblando Hecho

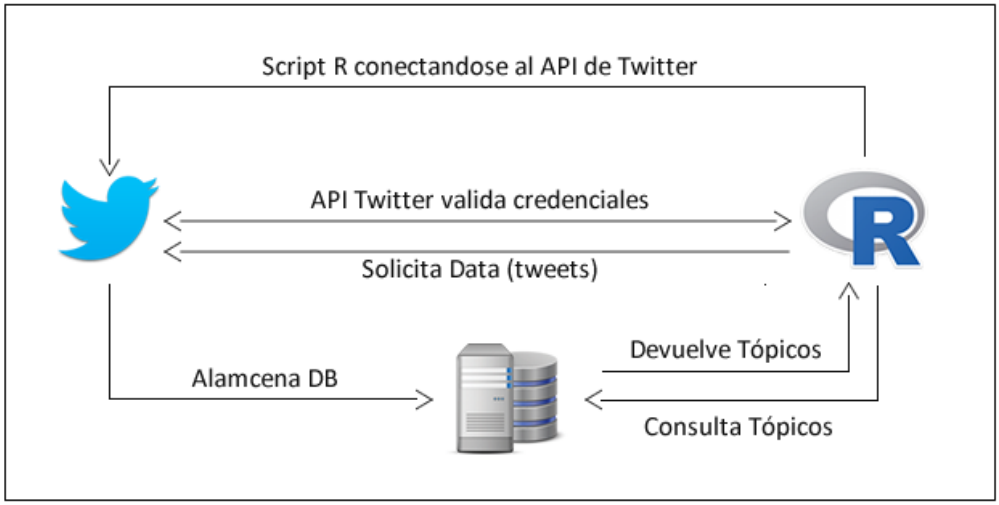


Figure 17. Flujo de procesos para la extracción de la data de Twitter

Elaboración Propia

Se ha construido el siguiente script R para extraer la base de datos con la que trabajara el modelo.

Script R

#install.packages("RPostgreSQL")
#Instancio la librería para conexión a Postgres en R
library("RPostgreSQL")
#Script de cadena de conexión con base de datos, sirve para tener en enlace donde se depositan los datos
con <- dbConnect(PostgreSQL(), user= "postgres", password="postgres", dbname="dm_social",port=5433)

```
#Estructurando consulta de solicitud de topico
qtopic <- "select * from public.topico"
topico<-dbGetQuery(con, qtopic)
```

Lo que devuelve

```
> topico
   idtopico           nomtopico idcategoria idredsocial fechaconsulta comentarios idestado
1         1           Manchester           1           1    2017-05-24          9       NA
2         2    Alejandro Toledo            2           1    2017-05-24         28       NA
3         3 SomosMasPatasCuando            6           1    2017-05-24         26       NA
4         4     Estado Islamico            1           1    2017-05-24         55       NA
5         5        SeQuedaEnEEG            3           1    2017-05-24          6       NA
6         6             Cristal            4           1    2017-05-24         11       NA
7         7       Ariana Grande            3           1    2017-05-23        325       NA
8         8            Movistar            5           1    2017-05-24        181       NA
9         9          FelizLunes            6           1    2017-05-23         54       NA
10       10             A4Latam            5           1    2017-05-24         54       NA
> |
```

Figure 18. Tópicos para extraer comentarios de base de datos muestra

Se complementa el siguiente script para extraer los comentarios y almacenarlos en la base de datos.

Script R

```
library("RPostgreSQL")

#Script de cadena de conexión con base de datos, sirve para tener en enlace donde se depositan los datos
con <- dbConnect(PostgreSQL(), user= "postgres", password="postgres", dbname="dm_social",port=5433)

#Estructurando consulta de solicitud de topico
qtopic <- "select * from public.topico"
topico<-dbGetQuery(con, qtopic)
topico<-as.data.frame(topico)
itopico<-topico[,1]
```

```
can<-length(itopico)

#Instancio la librería para conexión a API Twitter
library("twitteR")
consumer_key <- 'irr0c2aYXbIhJbnlnOfe9Q'
consumer_secret <- 'mho5ScH2PXEZYaomh2FKj8a7fNkAfnxIGvxuhAOPZQ'
access_token <- '281220172-byBlQyReDQFgyJdwUxNDQhQKRR8kgBFtGf9Cvca7'
access_secret <- 'CwKNhMWxHfAkpR0RHp8B3Bj9l7u8mCYOUDRnKV08'
setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_secret)

#Leer cada topico, extraer sus tweets e imprimir en la base de datos

i=0;
z=0;
while (i<can){
        i<-i+1
        nomt<-topico[i,2]
        nomtf <-paste("",nomt," -filter:retweets",sep="")
        canc<-topico[i,6]
        fect<-topico[i,5]
        fectf<-paste("",fect,"",sep="")
        itop<-topico[i,1]

ut <- searchTwitter(nomtf, n=canc, lang="es", since=fectf)tweet<-twListToDF(ut)
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
canc<-tweet[,1]
canc<-length(canc)

        j=0;
        while(j < canc){
          j<-j+1;
          z<-z+1;

                idusuario <- 1
                contenido <- tweet[j,1]
                idtopico  <- itop
                idtipomodelo <- 1
                idclase <- 4


#Estructurando consulta de inserción a base de datos
        queryinsert <- paste("insert into comentario
        (idcomentario,detallecomentario,idtopico,idusuario,idt
        ipomodelo,idclase) values
        (",z,",'",contenido,"',",idtopico,",",idusuario,",",idtipomo
        delo,",",idclase,")",sep="");

        #Ejecutando inserción de FB a BD
        sendfb<-dbGetQuery(con, queryinsert)
        }

}
```

Obteniendo una base de datos de muestra para trabajar la investigación, estas bases de datos con los scripts descritos realizan

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

un sistema de filtros para evitar tweets repetidos u otras irregularidades.

	idcomentario [PK] integer	detallecomentario character varying(300)	idtopico integer	idusuario integer	idtipomodelo integer	idclase integer
567	619	RT @MovistarPeru: @iReykel ¡Movistar se renovó! Pronto vendrán muchas sorpresas para nuestr	8	1	1	4
568	621	@inoesasi ¿Cuál es la nueva imagen de movistar?	8	1	1	4
569	622	El cambio de imagen de Movistar es como cuando tu ex dice que va a cambiar ...POR LAS HUEVA	8	1	1	4
570	623	@iReykel ¡Movistar se renovó! Pronto vendrán muchas sorpresas para nuestros clientes https:	8	1	1	4
571	624	RT @CandeQuiroga3: Soy yo o los datos duran cada vez menos?? La concha de tu madre movistar	8	1	1	4
572	625	El hit de la nueva campaña de Movistar es la torita o me parece?	8	1	1	4
573	626	Vota por el juego del año en los Premios Movistar de FestiGame 2017 https://t.co/tSoR9UWhwt	8	1	1	4
574	628	Video: @NairoQuinCo (@Movistar_Team) a punto de provocar una caída al lanzar bidón en la 17	8	1	1	4
575	630	RT @trafficVALENCIA: via @kikeguevara88: alguien sabe que pasa con instragran y face con	8	1	1	4
576	631	RT @RodrigoC_22: Brother, chévere que seas de otro operador, no gastes batería rajando de M	8	1	1	4
577	632	@AyudaMovistarCL ES INCREÍBLE QUE HAYA HECHO MÁS DE 10 SOLICITUDES PARA MOVISTAR ONE Y TODA	8	1	1	4
578	633	RT @RodrigoC_22: Brother, chévere que seas de otro operador, no gastes batería rajando de M	8	1	1	4
579	634	Y más o menos que le pasa a Movistar?	8	1	1	4
580	635	Vota por el juego del año en los Premios Movistar de FestiGame 2017 https://t.co/YKVcEyPhsn	8	1	1	4
581	636	¿Su empresa esta preparada para contener un ciberataque?, conoce como protegerte https://t.	8	1	1	4
582	638	RT @inoesasi: ¿Por qué todos están hablando de Movistar y un gran cambio? ¿Mejoró su servic	8	1	1	4
583	640	@cbonamus Hola, muy pronto os la comunicaremos. Un saludo, Patricia.	8	1	1	4
584	641	@GringazaPeruana ¡Cambiamos y pronto todas las novedades! Sigue atenta a Movistar https://t	8	1	1	4
585	643	RT @octaranda: #FelizLunes eso y más haría por ti. https://t.co/dfqtz6AFpo	9	1	1	4
586	644	RT @octaranda: #FelizLunes eso y más haría por ti. https://t.co/dfqtz6AFpo	9	1	1	4
587	645	RT @AlistamientFANB: #20May Avance Porcentual de la @REDI_Central en la Segunda Semana del	9	1	1	4
588	647	RT @FsiLauraRuiz19: Todo es del color del cristal con que se mira.. ☎ 5523375695 ó 5549040	9	1	1	4

Figure 19. Tabla hecho poblada (comentarios)

Pre Clasificación

La base de datos que acaba de generarse necesita un proceso manual de enseñanza, en el cual se debe determinar cada comentario según su clase:

Recordando que la dimensión clase soporta:

	idclase [PK] integer	nomclase character varying(20)
1	1	Positivo
2	2	Negativo
3	3	Neutro
4	4	Desconocido

Figure 20. Muestra dimensión clase

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Entonces el trabajo a realizar es clasificar cada comentario, y esta labor debe hacerse manual.

	idcomentario [PK] integer	detallecomentario character varying(300)	idtopico integer	idusuario integer	idtipomodelo integer	idclase integer
482	525	Movistar: Cámbiate a Movistar en plan y llévate un Smar	8	1	1	4
483	526	□ Movistar	8	1	1	4
484	527	RT @ctierz: #MovistarArtsyRBLSFestivalTeatreJove. Entra	8	1	1	4
485	528	RT @CESAR_RV2: 8 Minicoopers dañados por la caída de an	8	1	1	4
486	529	Vota por el juego del año en los Premios Movistar de Fe	8	1	1	4
487	530	Movistar+ comienza a emitir publicidad adaptada a la ed	8	1	1	4
488	531	RT @OscarJim3nez: En @movistar_es les gusta mucho estaf	8	1	1	4
489	532	RT @FloroPeruano: ¿Volverá el internet ilimitado? Te es	8	1	1	4
490	533	RT @Eurosport_ES: Gorka Izagirre de @movistar_Team en e	8	1	1	4
491	534	@soonpleplan Seguí los tips https://t.co/EeyzBdjtoT pa	8	1	1	4
492	535	RT @angellacamila: Movistar cambió y tú para cuando?	8	1	1	4
493	536	RT @angellacamila: Movistar cambió y tú para cuando?	8	1	1	4
494	537	Movistar hoy anda todo indio	8	1	1	4
495	538	@alejanvalverde @Movistar_Team mañana a las 16:00 en @	8	1	1	4

Figure 21. Determinando manualmente la clase de cada comentario.

Una vez finalizado el proceso, se cuenta con una base de datos local para que el modelo pueda procesar, por lo que se pasa a la fase 2 del ETL.

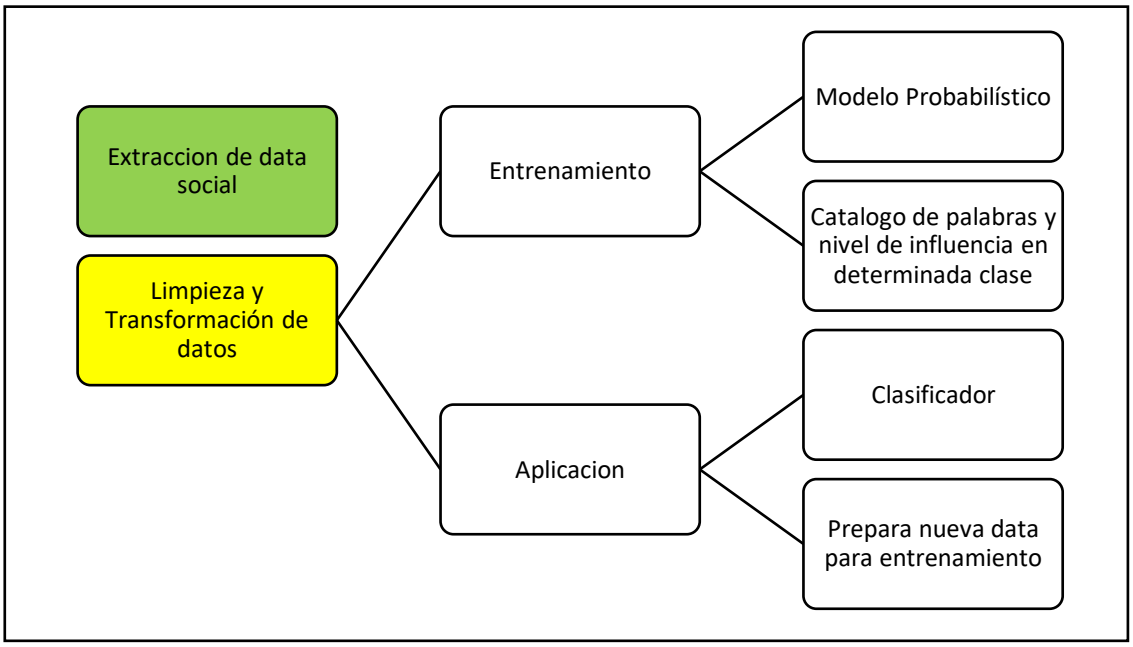


Figure 22. Estrategia Modelo – Fase Limpieza y Transformación de datos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

En la fase de Limpieza y transformación se ha trabajado con ambos entornos, entrenamiento y validación, para no confundir al lector con los scripts, se realizará una simulación sencilla para determinar el trabajo del algoritmo.

Script R

```
library(RTextTools)
library(e1071)

#protocolo de diseño
#declara tweet de entrenamiento y validación

#declara tweets de entrenamiento histórico positivo

pos_tweets =  rbind(
  c('Yo amo Chiclayo', 'positivo'),
  c('Este panorama es asombroso', 'positivo'),
  c('Yo me siento bien en las mañanas', 'positivo'),
  c('Estoy tan emocionado por el concierto', 'positivo'),
  c('El es mi mejor amigo', 'positivo')
)

#declara tweets de entrenamiento histórico negativo

neg_tweets = rbind(
  c('A mi no me gusta este carro', 'negativo'),
  c('Este panorama es desagradable', 'negativo'),
  c('Yo me siento cansado por las mañanas', 'negativo'),
  c('Yo no estoy tan emocionado por el concierto', 'negativo'),
  c('El es mi enemigo', 'negativo')
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
)

#declara tweets para validación con la que se va a contrastar el aprendizaje del modelo según los clasificadores

test_tweets = rbind(
  c('Me siento alegre esta mañana', 'positivo'),
  c('Amigo asombroso', 'positivo'),
  c('No me agrada este hombre', 'negativo'),
  c('Esta casa no es grande', 'negativo'),
  c('Tu musica es horrible', 'negativo')
)

#set final (LO QUE SE HA CONSOLIDADO POR BASE DE DATOS)

tweets = rbind(pos_tweets, neg_tweets, test_tweets)
```

El código anterior nos devuelve una lista de 15 comentarios, 10 comentarios para entrenar y 5 para validar.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
> tweets
      [,1]                                          [,2]
 [1,] "Yo amo Chiclayo"                             "positivo"
 [2,] "Este panorama es asombroso"                  "positivo"
 [3,] "Yo me siento bien en las mañanas"            "positivo"
 [4,] "Estoy tan emocionado por el concierto"       "positivo"
 [5,] "El es mi mejor amigo"                        "positivo"
 [6,] "A mi no me gusta este carro"                 "negativo"
 [7,] "Este panorama es desagradable"               "negativo"
 [8,] "Yo me siento cansado por las mañanas"        "negativo"
 [9,] "Yo no estoy tan emocionado por el concierto" "negativo"
[10,] "El es mi enemigo"                            "negativo"
[11,] "Me siento alegre esta mañana"                "positivo"
[12,] "Amigo asombroso"                             "positivo"
[13,] "No me agrada este hombre"                    "negativo"
[14,] "Esta casa no es grande"                      "negativo"
[15,] "Tu musica es horrible"                       "negativo"
>
```

Figure 23. Data de entrenamiento y validación para explicación

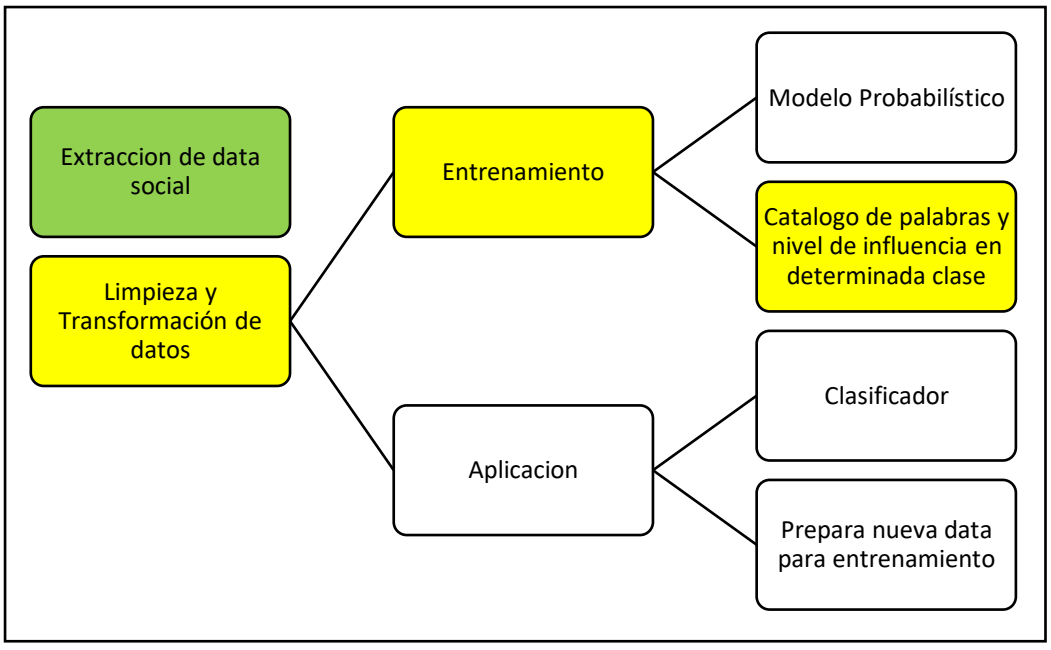


Figure 24. Generación de matriz de términos

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Script R
matrix= create_matrix(tweets[,1], language="spanish", removeStopwords=FALSE, removeNumbers=TRUE, stemWords=FALSE)

```
> matrix
<<DocumentTermMatrix (documents: 15, terms: 31)>>
Non-/sparse entries: 50/415
Sparsity           : 89%
Maximal term length: 12
Weighting          : term frequency (tf)
>
```

En este apartado el modelo ha realizado una serie de funciones y métodos de limpieza de datos explicados en este diagrama:

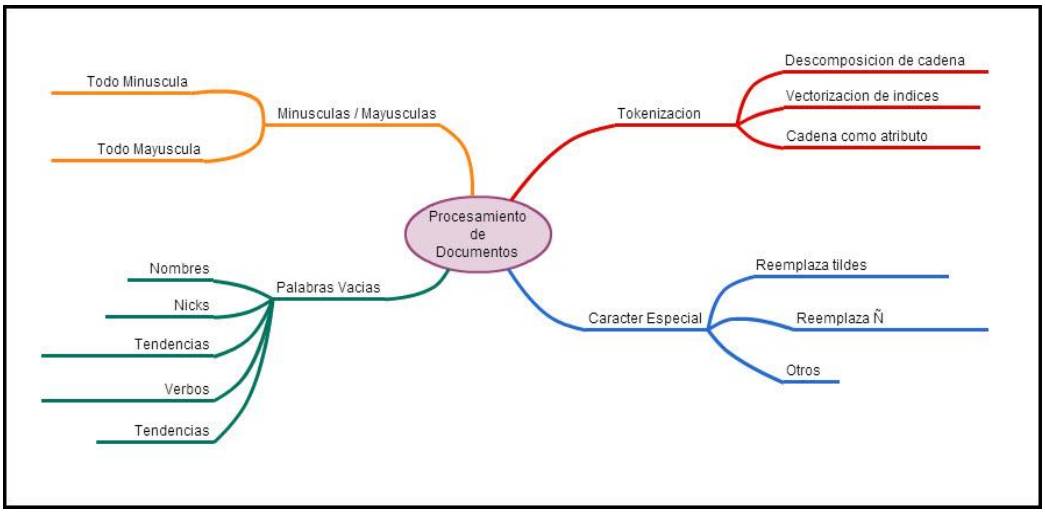


Figure 25. Protocolo inicial de análisis de textos para la investigación

Obteniendo como resultado un conjunto de palabras VALIDAS para que el clasificador genere el aprendizaje.

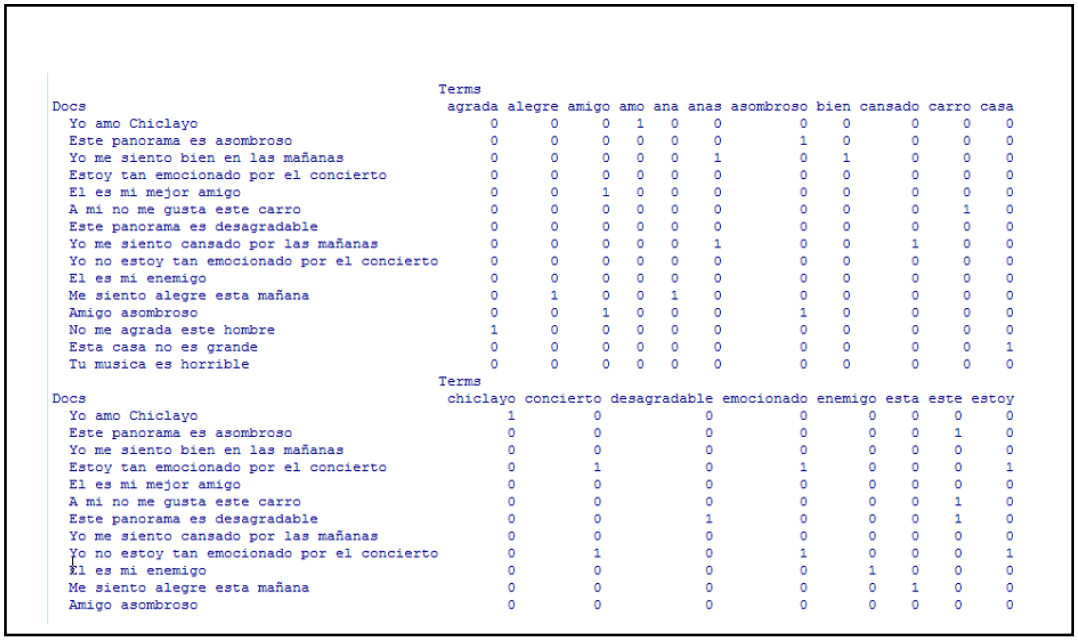
Si se transforma la matriz generada en el código anterior en un formato de datos, se podrá observar el catalogo que utilizará el

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

clasificador:

Script R
mat = as.matrix(matrix)

Obteniendo una matriz donde se consulta cada comentario, y cada palabra se vuelve una columna, para evaluar la ocurrencia de dicha palabra en cada comentario, al final de la matriz se establece la clase a la que pertenece (POSITIVO, NEGATIVO, ETC).



	Terms										
Docs	agrada	alegre	amigo	amo	ana	anas	asombroso	bien	cansado	carro	casa
Yo amo Chiclayo	0	0	0	1	0	0	0	0	0	0	0
Este panorama es asombroso	0	0	0	0	0	0	1	0	0	0	0
Yo me siento bien en las mañanas	0	0	0	0	0	1	0	1	0	0	0
Estoy tan emocionado por el concierto	0	0	0	0	0	0	0	0	0	0	0
El es mi mejor amigo	0	0	1	0	0	0	0	0	0	0	0
A mi no me gusta este carro	0	0	0	0	0	0	0	0	0	1	0
Este panorama es desagradable	0	0	0	0	0	0	0	0	0	0	0
Yo me siento cansado por las mañanas	0	0	0	0	0	1	0	0	1	0	0
Yo no estoy tan emocionado por el concierto	0	0	0	0	0	0	0	0	0	0	0
El es mi enemigo	0	0	0	0	0	0	0	0	0	0	0
Me siento alegre esta mañana	0	1	0	0	1	0	0	0	0	0	0
Amigo asombroso	0	0	1	0	0	0	1	0	0	0	0
No me agrada este hombre	1	0	0	0	0	0	0	0	0	0	0
Esta casa no es grande	0	0	0	0	0	0	0	0	0	0	1
Tu musica es horrible	0	0	0	0	0	0	0	0	0	0	0

	Terms							
Docs	chiclayo	concierto	desagradable	emocionado	enemigo	esta	este	estoy
Yo amo Chiclayo	1	0	0	0	0	0	0	0
Este panorama es asombroso	0	0	0	0	0	0	1	0
Yo me siento bien en las mañanas	0	0	0	0	0	0	0	0
Estoy tan emocionado por el concierto	0	1	0	1	0	0	0	1
El es mi mejor amigo	0	0	0	0	0	0	0	0
A mi no me gusta este carro	0	0	0	0	0	0	1	0
Este panorama es desagradable	0	0	1	0	0	0	1	0
Yo me siento cansado por las mañanas	0	0	0	0	0	0	0	0
Yo no estoy tan emocionado por el concierto	0	1	0	1	0	0	0	1
El es mi enemigo	0	0	0	0	1	0	0	0
Me siento alegre esta mañana	0	0	0	0	0	1	0	0
Amigo asombroso	0	0	0	0	0	0	0	0

Figure 26. Documento de términos, matriz de ocurrencia y catálogo de palabras válidas a clasificar.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

c) Seleccionar los algoritmos de clasificación.

(Hassan , Muhamad, & Muhamad, 2014) En su investigación sobre un estudio de algoritmos de aprendizaje resaltan estos 10 Algoritmos, de los cuales según la cita "Métodos de aprendizaje tales como RANDOM FOREST, ensacado y SVMs logran un excelente rendimiento que habría sido difícil obtener hace sólo 15 hace años"

Además, nos dicen "La calibración mejora drásticamente el rendimiento de árboles potenciados, SVMs, Naive Bayes"

Para esta investigación se someterá a evaluación en el entorno de texto los algoritmos de Naive Bayes, SMV y Árbol de regresión

ALGORITMOS DE APRENDIZAJE PARA CLASIFICACION		
Algoritmo	**Detalle**	**Aplicado a esta investigación**
Regresión Logística		**NO**
Árbol de Decisiones	**Árbol de regresión**	**SI**
Redes Neuronales		**NO**
Redes Bayesianas	**Naive Bayes**	**SI**
Máquina de Soporte Vectorial	**SVM**	**SI**
k-Nearest Neighboard	**kNN**	**NO**
Random Forest		**NO**

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

d) Diseñar y construir un modelo clasificador para el minado de opiniones en Twitter.

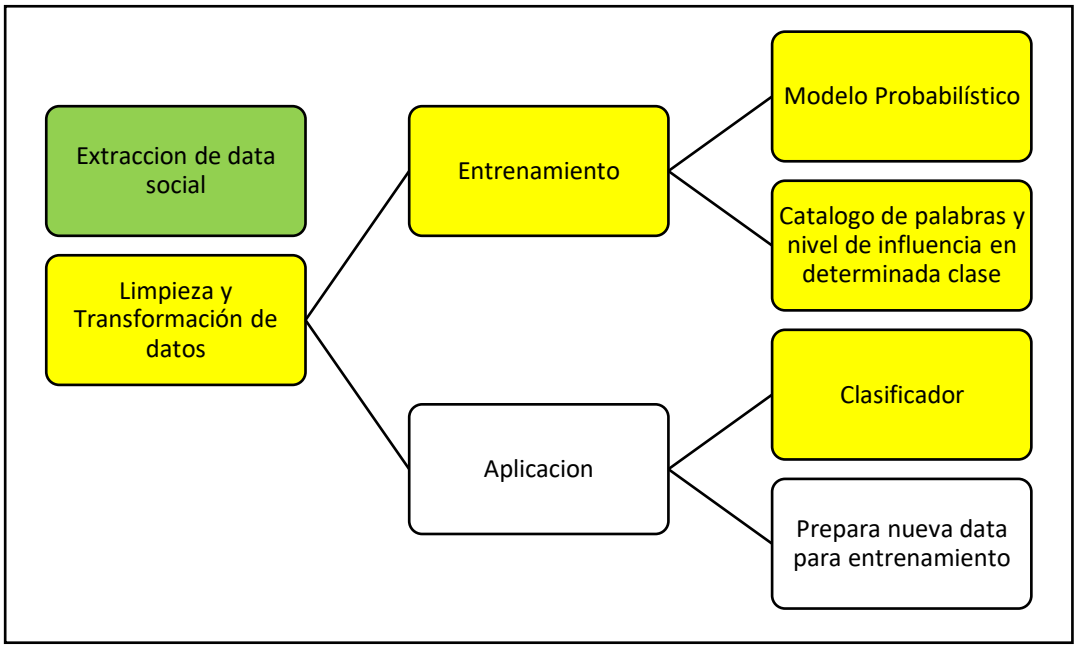


Figure 27. Modelo probabilístico y clasificador

Clasificador Naive Bayes

Fundamento

Basado en el teorema de Bayes, es un procedimiento de probabilidad que evalúa clases o estados

$$P(A \mid B) = P(A) * P(B \mid A) / P(B)$$

En función al comportamiento que ha podido tener estas clases o estados, asumiendo la independencia de los atributos que puedan condicionar dicha clase.

$$P(A \text{ and } B) = P(A) * P(B \mid A)$$

$$P(B \mid A) = P(A \text{ and } B) / P(A)$$

$$P(B \mid A) = P(B) * P(A \mid B) / P(A)$$

Script R
classifier = naiveBayes(mat[1:10,], as.factor(tweets[1:10,2])) class(classifier) summary(classifier) print(classifier)

Obteniendo la siguientes salidas

```
Conditional probabilities:
                            agrada
as.factor(tweets[1:10, 2]) [,1] [,2]
                  negativo    0    0
                  positivo    0    0

                            alegre
as.factor(tweets[1:10, 2]) [,1] [,2]
                  negativo    0    0
                  positivo    0    0

                            amigo
as.factor(tweets[1:10, 2]) [,1]      [,2]
                  negativo  0.0 0.0000000
                  positivo  0.2 0.4472136

                            amo
as.factor(tweets[1:10, 2]) [,1]      [,2]
                  negativo  0.0 0.0000000
                  positivo  0.2 0.4472136

                            ana
as.factor(tweets[1:10, 2]) [,1] [,2]
                  negativo    0    0
                  positivo    0    0

                            anas
as.factor(tweets[1:10, 2]) [,1]      [,2]
                  negativo  0.2 0.4472136
                  positivo  0.2 0.4472136
```

Figure 28. Modelo probabilístico

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Por cada palabra del catálogo de palabras se extraen los coeficientes, que representan al peso de cada palabra según su ocurrencia en cada comentario de la base de entrenamiento, y la condición de su clase (Negativo, Positivo, etc).

Para ver a detalle el clasificador Naive bayes de R puede consultarse el script de codigo abierto en: https://github.com/cran/e1071/blob/master/R/naiveBayes.R

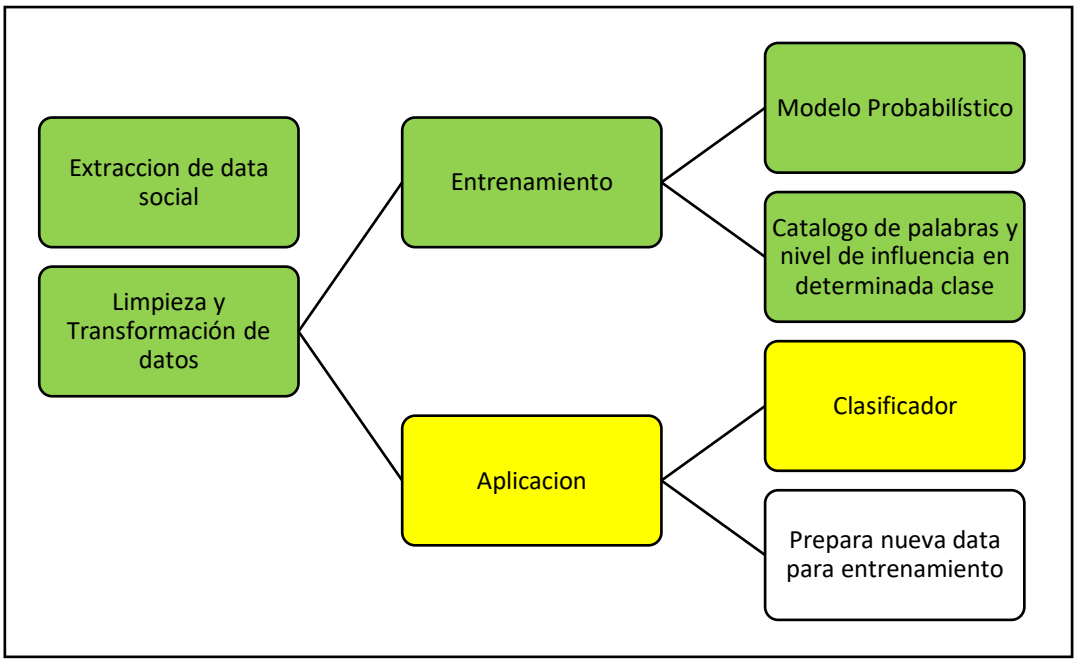


Figure 29. Modelo probabilístico y clasificador

Script R
predicted = predict(classifier, mat[11:15,]); predicted table(tweets[11:15, 2], predicted) recall_accuracy(tweets[11:15, 2], predicted)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
> predicted = predict(classifier, mat[11:15,]); predicted
[1] negativo positivo negativo negativo negativo
Levels: negativo positivo
>
> table(tweets[11:15, 2], predicted)
          predicted
           negativo positivo
  negativo        3        0
  positivo        1        1
> recall_accuracy(tweets[11:15, 2], predicted)
[1] 0.8
>
>
```

Figure 30. Matriz de confusión y confiabilidad del modelo

SVM – Maquina de Soporte Vectorial y Árbol de Regresión - TREE

Fundamentos SVM

Sea D un conjunto de datos de clasificación con n puntos en un espacio d-dimensional D = {(xi, yi)}, con i = 1, 2, ..., n y que haya sólo dos etiquetas de clase tales que yi sea +1 o -1. Un hiperplano h (x) da una función discriminante lineal en dimensiones d y divide el espacio original en dos medios espacios:

$$h(x) = w^T x + b = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b\,,$$

Si el conjunto de datos es linealmente separable, se puede encontrar un hiperplano de separación tal que para todos los puntos con etiqueta -1, h (x) <0 y para todos los puntos etiquetados +1, h (x)> 0. En este caso, h (x) sirve como un clasificador lineal o discriminante lineal que predice la clase para cualquier punto. Por otra parte, el vector de peso w es ortogonal al hiperplano, dando por tanto la dirección que es

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

normal a él, mientras que el sesgo b fija el desplazamiento del hiperplano en el espacio d-dimensional.

$$\delta_i = \frac{y_i h(x_i)}{||w||}$$

Dado un hiperplano de separación h (x) = 0, es posible calcular la distancia entre cada punto xi y el hiperplano por:

$$\delta^* = \min_{x_i} \{ \frac{y_i h(x_i)}{||w||} \}$$

Fundamentos Tree Regression

Un árbol de regresión parte de la premisa de un árbol de decisión, El algoritmo básico para el árbol de decisiones es el algoritmo codicioso que construye árboles de decisión de una manera recursiva de arriba hacia abajo. Usualmente empleamos estrategias codiciosas porque son eficientes y fáciles de implementar, pero usualmente llevan a modelos sub óptimos. También podría utilizarse un enfoque de abajo hacia arriba. El algoritmo se detiene cuando se cumplen las condiciones de parada.

$$H(D) = -\sum_{i=1}^{k} P(c_i|D) \log_2 P(c_i|D)$$

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Script R
container = create_container(matrix, as.numeric(as.factor(tweets[,2])), trainSize=1:10, testSize=11:15,virgin=FALSE)

En el script anterior se declara que, del contenedor para esta demostración, desde el índice 1:10 servirán como datos de entrenamiento, y de 11:15 serán los datos de validación.

```
An object of class "matrix_container"
Slot "training_matrix":
An object of class "matrix.csr"
Slot "ra":
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Slot "ja":
 [1]  4 12  7 18 28  6  8 24 25 30 13 15 19 29 31  3 26 10 18 21 14 18 28  6  9
[26] 24 25 29 30 13 15 19 29 31 16

Slot "ia":
 [1]  1  3  6 11 16 18 21 24 30 35 36

Slot "dimension":
[1] 10 31


Slot "classification_matrix":
An object of class "matrix.csr"
Slot "ra":
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Slot "ja":
 [1]  2  5 17 25 30  3  7  1 18 22 11 17 20 23 27

Slot "ia":
[1]  1  6  8 11 14 16

Slot "dimension":
[1]  5 31


Slot "training_codes":
 [1] 2 2 2 2 2 1 1 1 1 1
Levels: 1 2

Slot "testing_codes":
[1] 2 2 1 1 1
Levels: 1 2

Slot "column_names":
 [1] "agrada"      "alegre"      "amigo"       "amo"         "ana"
 [6] "anas"        "asombroso"   "bien"        "cansado"     "carro"
[11] "casa"        "chiclayo"    "concierto"   "desagradable" "emocionado"
[16] "enemigo"     "esta"        "este"        "estoy"       "grande"
[21] "gusta"       "hombre"      "horrible"    "las"         "maÃ"
[26] "mejor"       "musica"      "panorama"    "por"         "siento"
[31] "tan"
```


Figure 31. Estructura de conocimiento antes de aplicar SVM

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Una vez aplicada la función “create container” se obtienen los datos, las clases de entrenamiento y las palabras que formaran parte del catálogo de palabras.

En este caso se aplican estos dos algoritmos aislados de la red bayesiana, ya que el formato o procedimiento de tratamiento de datos difieren con este último algoritmo.

En el R-Project se encuentra la librería e1071 que contiene algoritmos clasificadores como SVM y Árbol de Regresiones

```
> models = train_models(container, algorithms=c("SVM","TREE"))
> models
$SVM

Call:
svm.default(x = container@training_matrix, y = container@training_codes,
    kernel = kernel, cost = cost, cross = cross, probability = TRUE,
    method = method)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  100
      gamma:  0.03225806

Number of Support Vectors:  10


$TREE
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

1) root 10 13.86 1 ( 0.5 0.5 ) *
```

Figure 32. Aplicando los algoritmos SVM y TREE

Aplicados ambos algoritmos se aplica a los comentarios de validación para realizar la clasificación y determinar la confiabilidad de ambos algoritmos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Script R

```
results = classify_models(container, models)


table(as.numeric(as.factor(tweets[11:15,          2])),
results[,"SVM_LABEL"])
table(as.numeric(as.factor(tweets[11:15,          2])),
results[,"TREE_LABEL"])



recall_accuracy(as.numeric(as.factor(tweets[11:15,          2])),
results[,"TREE_LABEL"])
recall_accuracy(as.numeric(as.factor(tweets[11:15,          2])),
results[,"SVM_LABEL"])
```

Finalmente se obtienen las matrices de confusión para SVM y TREE, se calcula la confiabilidad según las clasificaciones, contrastados contra los datos reales de validación.

```
> table(as.numeric(as.factor(tweets[11:15, 2])), results[,"SVM_LABEL"])

   1 2
  1 2 1
  2 2 0
> table(as.numeric(as.factor(tweets[11:15, 2])), results[,"TREE_LABEL"])

   1
  1 3
  2 2
> recall_accuracy(as.numeric(as.factor(tweets[11:15, 2])), results[,"TREE_LABEL"])
[1] 0.6
> recall_accuracy(as.numeric(as.factor(tweets[11:15, 2])), results[,"SVM_LABEL"])
[1] 0.4
>
```

Figure 33. Resultados de confiabilidad de SVM y TREE

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

e) **Evaluar los algoritmos de clasificación seleccionados.**

Para evaluar los algoritmos de clasificación, se ha utilizado el esquema de indicadores diseñado en el capítulo III de la presente investigación.

Donde:

Indicador	Medida o técnica	Formula	Frecuencia
Exactitud	Matriz de confusión	$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ TP=True Positive (Verdadero Positivo) TN=True Negative (Verdadero Negativo) FP=False Positive (Falso Positivo) FN=False Negative (False Negativo)	Por escenario de prueba
Confiabilidad	Promedio	P=SUMATORIA(EP)/TEP P=PROMEDIO EP=ESCENARIO DE PRUEBA TEP=TOTAL ESCENARIO DE PRUEBA	Una sola vez aplicado al promedio de todos los escenarios
Rendimiento	Promedio	P=SUMATORIA(EP)/TEP P=PROMEDIO EP=ESCENARIO DE PRUEBA TEP=TOTAL ESCENARIO DE PRUEBA	Una sola vez aplicado al promedio de todos los escenarios

Protocolo de pruebas

El protocolo de pruebas se establece para validar los distintos algoritmos a evaluar, en función a iteraciones o escenarios de prueba.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Cada escenario es alimentado por un conjunto determinado de datos y retorna los valores obtenidos de las métricas aplicadas.

Script final para el protocolo de pruebas

Script R

```
library(RTextTools)
library(e1071)
library(RPostgreSQL)
library(tm)
library(TwitteR)

#Script de cadena de conexión con base de datos, sirve para tener en enlace donde se depositan los datos
con <- dbConnect(PostgreSQL(), user= "postgres", password="postgres", dbname="dm_social",port=5433)

#Script para destruir la tablas escenarios de prueba
del<-dbGetQuery(con, "drop table escenario");
del2<-dbGetQuery(con, "drop table rendimiento");

#PROTOCOLO DE DISEÑO
#DECLARA LOS TWEETS DE ENTRENAMIENTO

#DECLARA TWEETS DE ENTRENAMIENTO HISTORICO POSITIVO

pos_tweets<- "select c.idcomentario,c.detallecomentario,cl.nomclase from public.comentario c join public.clase cl on
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
c.idclase=cl.idclase where c.idclase=1 order by
c.idcomentario limit 40 offset 0"
pos_tweets<-dbGetQuery(con, pos_tweets)

#DECLARA TWEETS DE ENTRENAMIENTO HISTORICO
NEGATIVO

neg_tweets<- "select
c.idcomentario,c.detallecomentario,cl.nomclase from
public.comentario c join public.clase cl on
c.idclase=cl.idclase where c.idclase=2 order by
c.idcomentario limit 40 offset 0"
neg_tweets<-dbGetQuery(con, neg_tweets)

#DECLARA DESDE QUE TWEETS VA A RECORRER EL
WHILE PARA CADA ESCENARIO Y CUANTO CRECERA

i=0;
ini=40;
vin=(ini*2)

desfase=10;
iteraciones=5;

vi=(ini*2)+1
vf=(ini*2)+(desfase*2)


f<-matrix(1,0,8,byrow=T)
f<- as.data.frame(f)
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
r1<-matrix(1,0,5,byrow=T)
r1<- as.data.frame(r1)

while(i<iteraciones){

        i=i+1;
        esc<-c(1:desfase)
        idesc <-replace(esc,which(esc>0),i)
        idesc <-as.data.frame(idesc)

        #DECLARA TWEETS PARA VALIDAR LO QUE EL MODELO APRENDIO DEL HISTORICO

        pos_tweetsv<- paste("select c.idcomentario,c.detallecomentario,cl.nomclase from public.comentario c join public.clase cl on c.idclase=cl.idclase where c.idclase=1 order by c.idcomentario limit ",desfase," offset ",ini,"",sep="")
        pos_tweetsv<-dbGetQuery(con, pos_tweetsv)

        neg_tweetsv<- paste("select c.idcomentario,c.detallecomentario,cl.nomclase from public.comentario c join public.clase cl on c.idclase=cl.idclase where c.idclase=2 order by c.idcomentario limit ",desfase," offset ",ini,"",sep="")
        neg_tweetsv<-dbGetQuery(con, neg_tweetsv)

        test_tweets <- rbind(pos_tweetsv,neg_tweetsv)
        v<-test_tweets
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
#SET FINAL DE ENTRENAMIENTO Y VALIDACION
tweets2 <- rbind(pos_tweets, neg_tweets, test_tweets)
colnames(tweets2)<- c("idcomentario","detallecomentario","real")
tweets <- tweets2[2:3]
tweets2<-tweets2[vi:vf,]

ini=ini+10;

# CONSTRUIMOS EL DOCUMENTO CON LOS TERMINOS
matrix= create_matrix(tweets[,1], language="spanish",
removeStopwords=FALSE, removeNumbers=TRUE,
stemWords=FALSE)

# ENTRENAMOS EL MODELO CON UN CLASIFICADOR BAYESIANO
mat = as.matrix(matrix)
classifier = naiveBayes(mat[1:vin,], as.factor(tweets[1:vin,2]),laplace=1)
class(classifier)
summary(classifier)
print(classifier)
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
# PROBAR EL MODELO PARA VALIDACION
predicted = predict(classifier, mat[vi:vf,]); predicted
predicted2 = predict(classifier, mat[vi:vf,],type="raw"); predicted
nb<-as.data.frame(predicted)
colnames(nb)<-"nb_label"
predicted2<-as.data.frame(predicted2)
nb2<-cbind(nb,predicted2)

mc1<-table(tweets[vi:vf, 2], predicted)
recall_accuracy(tweets[vi:vf, 2], predicted)
conf<-recall_accuracy(tweets[vi:vf, 2], predicted)
tima<-system.time(replicate(1,predict(classifier, mat[vi:vf,])))


# FORMATO DE ENTRENAMIENTO PARA MACHINE LEARNING SVM

container = create_container(matrix, as.numeric(as.factor(tweets[,2])),
                             trainSize=1:vin, testSize=vi:vf,virgin=FALSE)

# ENTRENAMOS EL MODELO CON UN CLASIFICADOR SVM
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
                models = train_models(container,
algorithms=c("SVM","TREE"))
                timb<-
system.time(replicate(1,train_models(container,
algorithms=c("SVM"))))
                timc<-
system.time(replicate(1,train_models(container,
algorithms=c("TREE"))))

                # PROBAR EL MODELO PARA VALIDACION

                results = classify_models(container, models)



                mc3<-table(as.numeric(as.factor(tweets[vi:vf,
2])), results[,"SVM_LABEL"])
                mc2<-table(as.numeric(as.factor(tweets[vi:vf,
2])), results[,"TREE_LABEL"])

        recall_accuracy(as.numeric(as.factor(tweets[vi:vf,
2])), results[,"TREE_LABEL"])

        recall_accuracy(as.numeric(as.factor(tweets[vi:vf,
2])), results[,"SVM_LABEL"])
                conf2<-
recall_accuracy(as.numeric(as.factor(tweets[vi:vf,        2])),
results[,"TREE_LABEL"])
                conf3<-
recall_accuracy(as.numeric(as.factor(tweets[vi:vf,        2])),
results[,"SVM_LABEL"])
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
                recall_accuracy(tweets[vi:vf, 2], predicted)

        recall_accuracy(as.numeric(as.factor(tweets[vi:vf,
2])), results[,"TREE_LABEL"])

        recall_accuracy(as.numeric(as.factor(tweets[vi:vf,
2])), results[,"SVM_LABEL"])

                analytics = create_analytics(container, results)
                #summary(analytics)
                #head(analytics@document_summary)
                cas<-analytics@document_summary
                cas<-cas[1:4]
                #analytics@ensemble_summar

                mat2<-cbind(idesc,tweets2,nb,cas)


                mat2$SVM_LABEL
                mat2$SVM_LABEL<-
as.character(mat2$SVM_LABEL)
                mat2$SVM_LABEL[mat2$SVM_LABEL == "1"]
<- "Negativo"
                mat2$SVM_LABEL[mat2$SVM_LABEL == "2"]
<- "Positivo"

                mat2$TREE_LABEL<-
as.character(mat2$TREE_LABEL)
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
                mat2$TREE_LABEL[mat2$TREE_LABEL    ==
"1"] <- "Negativo"
                mat2$TREE_LABEL[mat2$TREE_LABEL    ==
"2"] <- "Positivo"



                colnames(mat2)<-
c("idescenario","idcomentario","detallecomentario","real","nb
_label","svm_label","svm_prob","tree_label","tree_prob")

                mat2$real<- as.character(mat2$real)
                mat2$real[mat2$real == "Positivo"] <- "1"
                mat2$real[mat2$real == "Negativo"] <- "2"
                mat2$nb_label<- as.character(mat2$nb_label)
                mat2$nb_label[mat2$nb_label == "Positivo"] <-
"1"
                mat2$nb_label[mat2$nb_label  ==  "Negativo"]
<- "2"
                mat2$svm_label[mat2$svm_label              ==
"Positivo"] <- "1"
                mat2$svm_label[mat2$svm_label              ==
"Negativo"] <- "2"
                mat2$tree_label[mat2$tree_label == "Positivo"]
<- "1"
                mat2$tree_label[mat2$tree_label            ==
"Negativo"] <- "2"

                mat3<-mat2[-3]
                colnames(f)<-colnames(mat3)
                f <-rbind(mat3,f)
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
print("Reb Bayesiana")
print(conf)
print(mc1)
print("Arbol de Regresion")
print(conf2)
print(mc2)
print("SVM")
print(conf3)
print(mc3)
print(tima)
print(mat2)

r <-rbind(tima,timb,timc)
colnames(r1)<-colnames(r)
r1 <-rbind(r,r1)
}

r1<-r1[1:3]
a<-rownames(r1)
a<-as.data.frame(a)
r1<-cbind(a,r1)
colnames(r1)<-c("idescenario","user","system","elapsed")

dbWriteTable(con, "escenario", f)
dbWriteTable(con, "rendimiento", r1)
```


Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

La salida del código expuesto es la siguiente:

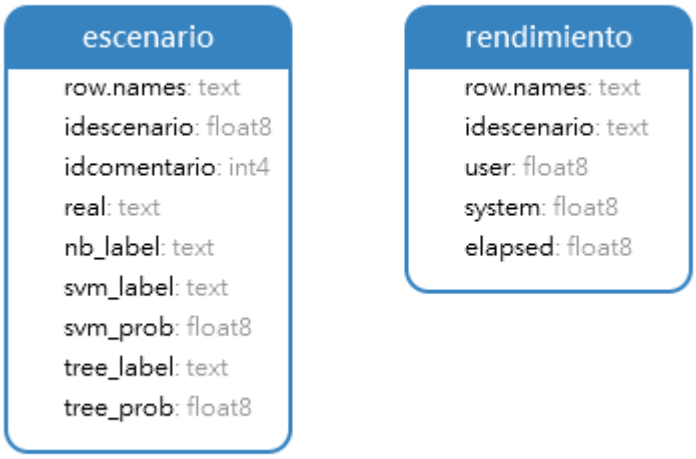
idescenario	idcomentario	real	nb_label	svm_label	svm_prob	tree_label	tree_prob
5	211	1	1	1	0.7956000	1	1.0000000
5	213	1	1	1	0.7956000	1	1.0000000
5	214	1	1	1	0.7956000	1	1.0000000
5	226	1	1	1	0.5827759	1	0.6363636
5	229	1	1	1	0.7260832	1	1.0000000
5	231	1	1	1	0.7956000	1	1.0000000
5	232	1	1	1	0.7956000	1	1.0000000
5	233	1	1	1	0.7956000	1	1.0000000
5	234	1	1	1	0.6586999	1	0.6363636
5	235	1	1	2	0.5750814	1	0.6363636
5	258	2	1	2	0.8123749	2	0.9166667
5	277	2	1	2	0.7664392	2	0.9166667
5	299	2	1	1	0.8994273	1	0.6363636
5	317	2	1	2	0.5968990	2	0.9166667
5	323	2	1	1	0.5515059	2	0.9166667
5	327	2	1	1	0.7956000	1	1.0000000
5	344	2	1	1	0.7445309	2	0.9166667
5	349	2	1	2	0.5669379	2	0.9166667
5	350	2	1	1	0.5074247	2	0.9166667
5	359	2	1	2	0.5895969	2	0.9166667
4	189	1	1	1	0.7429902	2	0.9166667
4	190	1	1	1	0.8276831	1	0.6363636
4	191	1	1	1	0.6689115	1	1.0000000
4	192	1	1	2	0.5808306	2	0.7500000
4	194	1	1	1	0.8400760	2	0.9166667
4	196	1	1	2	0.6482421	1	0.6363636
4	202	1	1	1	0.5246644	2	0.9166667

Figure 34. Resultado de los escenarios de prueba luego de ejecución – entidad escenario

Esta matriz se almacena en la base de datos, se realiza la siguiente consulta para obtener el formato visible al lector, que puede asi comparar los resultados:

```
        select idescenario,idcomentario,c1.nomclase as
        real,c2.nomclase as nb_label,c3.nomclase as
        svm_label,svm_prob,c4.nomclase as tree_label,tree_prob
        from (
select idescenario,idcomentario,cast(real as int) as
```

```
real,cast(nb_label as int) as nb_label,cast(svm_label as int) as svm_label,cast(tree_label as int) as tree_label,svm_prob,tree_prob from escenario) as q join clase c1 on q.real=c1.idclase join clase c2 on nb_label=c2.idclase join clase c3 on svm_label=c3.idclase join clase c4 on tree_label=c4.idclase
```



Obteniendo

idescenario double precision	idcomentario integer	real character var	nb_label character va	svm_label character v	svm_prob double precision	tree_label character va	tree_prob double prec
5	211	Positivo	Positivo	Positivo	0.795599953683685	Positivo	1
5	213	Positivo	Positivo	Positivo	0.795599953683685	Positivo	1
5	214	Positivo	Positivo	Positivo	0.795599953683685	Positivo	1
5	226	Positivo	Positivo	Positivo	0.582775855409765	Positivo	0.63636363
5	229	Positivo	Positivo	Positivo	0.726083154259692	Positivo	1
5	231	Positivo	Positivo	Positivo	0.795599953683685	Positivo	1
5	232	Positivo	Positivo	Positivo	0.795599953683685	Positivo	1
5	233	Positivo	Positivo	Positivo	0.795599953683685	Positivo	1
5	234	Positivo	Positivo	Positivo	0.658699891054614	Positivo	0.63636363
5	235	Positivo	Positivo	Negativo	0.575081424347144	Positivo	0.63636363
5	258	Negativo	Positivo	Negativo	0.812374911379339	Negativo	0.91666666
5	277	Negativo	Positivo	Negativo	0.76643916893479	Negativo	0.91666666
5	299	Negativo	Positivo	Positivo	0.899427263545912	Positivo	0.63636363
5	317	Negativo	Positivo	Negativo	0.596899000603496	Negativo	0.91666666
5	323	Negativo	Positivo	Positivo	0.551505881287387	Negativo	0.91666666
5	327	Negativo	Positivo	Positivo	0.795599953683685	Positivo	1
5	344	Negativo	Positivo	Positivo	0.744530877031998	Negativo	0.91666666

Figure 35. Consolidad final entidad - escenario

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

También en el algoritmo para la evaluación de resultados, se introdujo las funciones para extracción de tiempo, teniendo como resultado:

	idescenario	user	system	elapsed
tima	tima	20.17	0.00	21.27
timb	timb	26.30	0.00	27.20
timc	timc	32.59	0.14	33.78
tima4	tima4	22.21	0.00	22.69
timb4	timb4	26.56	0.00	27.25
timc4	timc4	33.21	0.19	34.60
tima3	tima3	21.91	0.00	22.17
timb3	timb3	26.39	0.00	26.90
timc3	timc3	33.37	0.07	34.58
tima2	tima2	21.17	0.00	21.93
timb2	timb2	26.50	0.00	27.10
timc2	timc2	32.80	0.11	34.03
tima1	tima1	21.60	0.00	21.89
timb1	timb1	26.55	0.00	27.65
timc1	timc1	33.23	0.05	33.89

Figure 36. Resultado para dimensión rendimiento

Lo que se procede a realizar es la validación, aplicando el indicador de exactitud, mediante la técnica de matriz de confusión, para cada escenario:

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

f) Implementar una aplicación para el análisis y visualización de resultados.

Se ha diseñado un aplicativo que sirve como portal para utilizar el modelo diseñado en R.

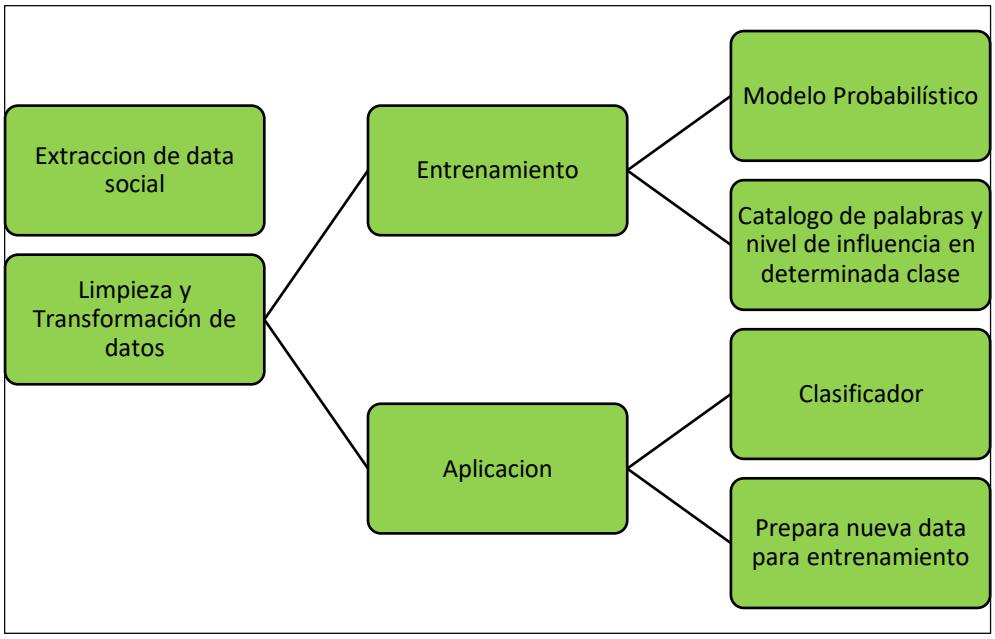


Figure 37. Modelo probabilístico y clasificador concluido en todas sus fases.

El aplicativo se ha construido utilizando herramientas de desarrollo web, como lenguajes de programación PHP versión 5.4, HTML5, CSS3 y JQUERY.

La base de datos es PostgreSQL, y es la misma que se utilizó para el modelo en R.

Aplicando algo de metodología ágil basado en SCRUM, se obtiene la siguiente historia de usuario.

HISTORIA DE USUARIO: ANALIZAR HASHTAG
ACTOR: USUARIO
DESCRIPCION **01.- El usuario ingresa al portal web analítico** **02.- El sistema muestra una interfaz para realizar la búsqueda; solicitando al usuario que ingrese el texto del hashtag.** **03.- El usuario ingresa el texto y presiona el botón buscar.** **04.- El sistema muestra una página al usuario indicando que se está realizando la búsqueda y el análisis del texto ingresado (hashtag).** **5.- EL sistema muestra los resultado en otra página, con una lista de lista de los tweets analizados y su respectiva clasificación.**
OBSERVACION: NINGUNA

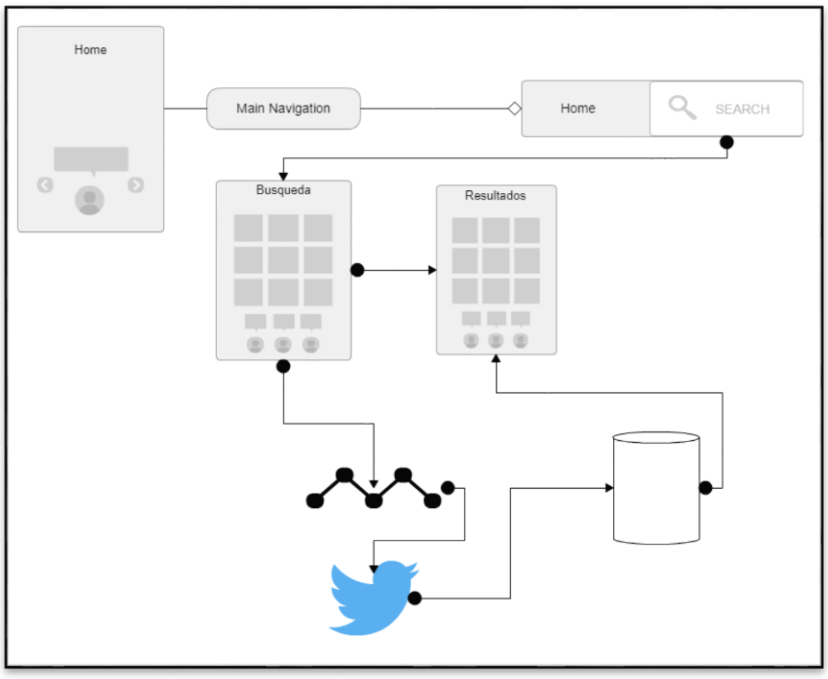


Figure 38. Flujograma de la Aplicación Web

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Dada la naturaleza del applicativo, que es una web pequeña no se ha utilizado arquitectura o patrón especial de software, siendo concebida como una app de arquitectura monolítica.

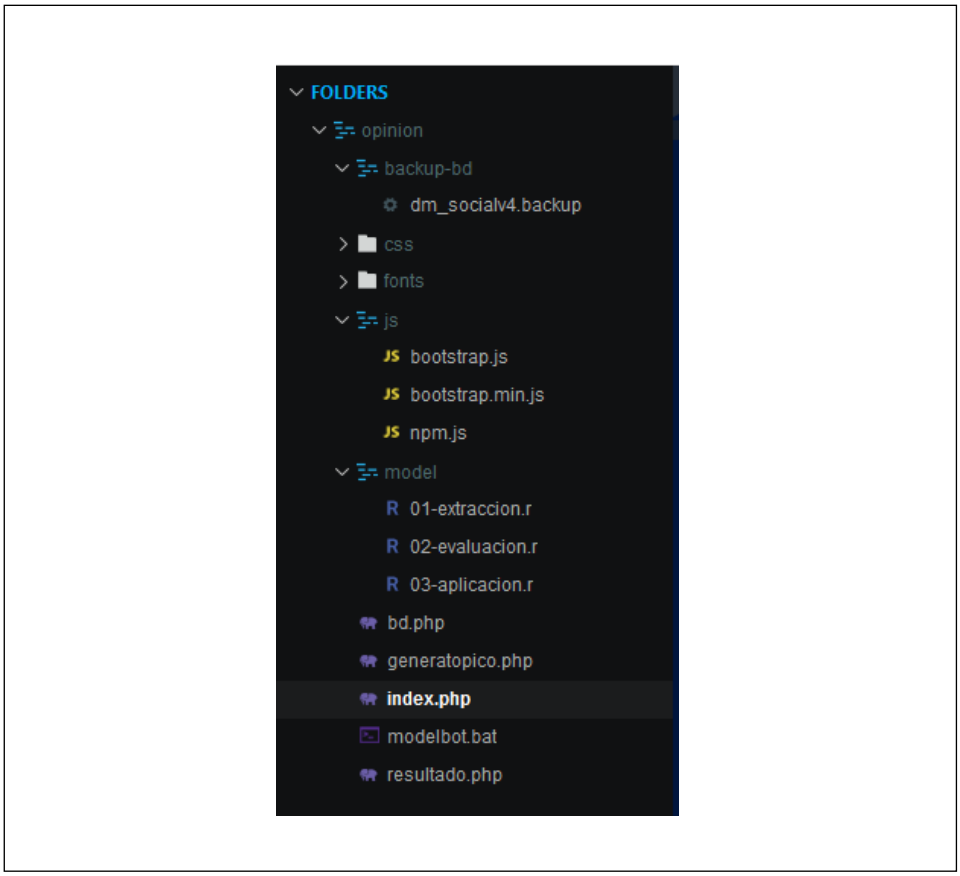


Figure 39. Árbol de directorio del aplicativo web

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

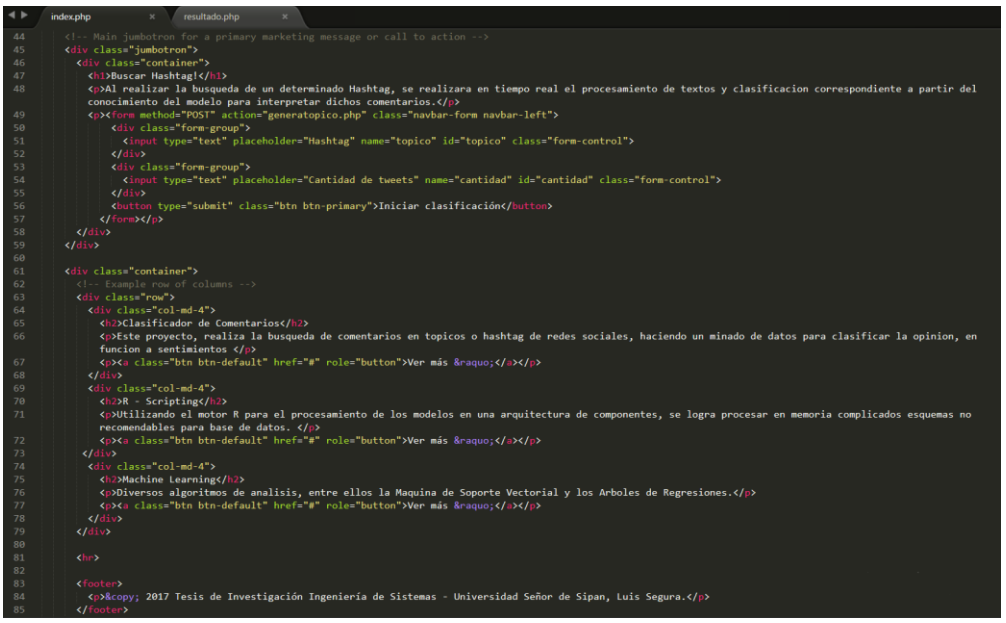


Figure 40. Código de la página index del aplicativo web

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO VI

CONCLUSIONES Y RECOMENDACIONES

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

VI. CAPITULO VI: CONCLUSIONES Y RECOMENDACIONES

6.1. Conclusiones

A. El ámbito de aplicación de los algoritmos clasificadores de textos, se encuentra dado por el universo de datos que genera los comentarios en redes sociales, específicamente la red social Twitter utilizada en esta investigación, aplicando el análisis por tópicos de interés con diversas categorías, ejecutando un muestreo estratificado obteniendo unos 664 tweets aproximadamente.

B. Para la extracción y tratamiento de tweets se ha optado por una estrategia de respaldo mediante la generación de una base de datos que guarda los tweets consultados al API de la red social, divididos en un esquema de tópicos, el cual contiene una cantidad definida de tweets considerando la fecha de consulta al API y guardado en la base de datos.

C. Para seleccionar los algoritmos de clasificación se ha optado por elegir algoritmos que cumplan con una función binomial (POSITIVO, NEGATIVO), así mismo se ha considerado como factor las últimas investigaciones, determinando que los

algoritmos a utilizar serian la RED BAYESIANA, MAQUINA DE SOPORTE VECTORIAL (SVM) y ARBOL DE REGRESION (TREE).

D. Se ha construido un modelo clasificador a partir de un esquema de entrenamiento, validación y aplicación, donde se origina una base de datos de conocimiento, utilizando el lenguaje R, se ha escrito el algoritmo, usando librerías de clasificadores, para estructurar el modelo.

E. En el indicador exactitud, sometido a 05 escenarios de prueba, se obtienen diversos resultados, esto implica, que la naturaleza de los textos tratados, genera una distorsión o ruido que los algoritmos deben tratar. En la mayoría de los casos se observa una red bayesiana con baja exactitud, mientras que SVM y TREE mantienen valores cercanos.

Esto se puede corroborar en el indicador de confiabilidad, que analiza los datos de las 05 iteraciones o escenarios de prueba, obteniendo en promedio que la red bayesiana consigue un 51 % de confiabilidad, mientras que SVM un 70 % no muy alejado de un árbol de regresión con 74 %.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Desde el punto de vista del rendimiento, se obtiene que el árbol, genera mayor tiempo y consumo de recursos que la red bayesiana y el SVM.

6.2. Recomendaciones

El lenguaje de programación utilizado en este estudio fue **R Proyect;** sin embargo, existen otros lenguajes como **Python,** que al igual que **R,** poseen librerías y algoritmos de clasificación para el tratamiento de datos; por lo que, se recomienda utilizar este (u otro) lenguaje de programación para comparar los datos con los resultados de esta investigación.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

REFERENCIAS BIBLIOGRAFICAS

Bing, L. (2012). *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers.

Cuadrado, J. C. (2011). *Un modelo linguistico-semantico basado en emociones para la clasificacion de textos segun su polaridad e intensidad.* Madrid.

Dubiau, L. (2014). *Procesamiento de Lenguaje Natural en Sistemas de Análisis de Sentimientos* (Tesis de Grado). Universidad de Buenos Aires, Argentina.

Echevarria, P. F. (2009). *Text mining aplicado a la clasificacion y distribucion automatica de correo electronico y deteccion de correos SPAM.*

Estévez, S. (2015). *Minería de Opinión en Twitter: una aproximación desde el aprendizaje supervisado* (Tesis de Grado). Universidad de La Habana, Cuba.

Hassan , S., Muhamad, R., & Muhamad, S. (2014). *Comparing SVM and Naïve Bayes Classifiers for Text Categorization with Wikitology as knowledge enrichment.*

Khan A., Atique, M. & Thakare, V. (2015). *Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis.* Recuperado de: http://www.ijecscse.org/papers/ATCON2015/DTM-01.pdf

Kotwal, A., Fulari, P., Jadhav, D. & Kad, R. (2016). *Improvement in Sentiment Analysis of Twitter Data Using Hadoop.* Recuperado de: http://www.imperialjournals.com/index.php/IJIR/article/view/1143/1097

Martín, C. (2016). *Análisis de Sentimientos en Twitter: El bueno, el malo y el >:(* (Tesis de Grado). Universidad Nacional de Coordoba, Argentina.

Merlo, R., Contreras, D. & Puente C. (2010). *Análisis de opiniones en Internet a partir de la red social Twitter.* Recuperado de: http://www.revista-anales.es/web/n_5/pdf/seccion_9.pdf

Montesinos, L. (2014). *Análisis de sentimientos y predicción de eventos en Twitter.* (Tesis de Grado). Universidad de Chile, Chile.

Pak, A. & Paroubek P. (2014). *Twitter as a Corpus for Sentiment Analysis and Opinion Mining.* Recuperado de: https://pdfs.semanticscholar.org/ad8a/7f620a57478ff70045f97abc7aec9687ccbd.pdf

Piehadrita, J. M. (2013). *Desarrollo e implementacion de un sistema de procesamiento de voces para el análisis de tres estados emocionales.* Bogota D.C.

Quantico Trends (2016). *Investiga a tus clientes y competencia en las redes sociales.* Recuperado de: http://www.quanticotrends.com/blog/quanticotrends/como-esta-compuesto-hoy-el-universo-de-twitter-en-peru/

Rodriguez, O. (2011). *Twitter: aplicaciones profesionales y de empresa.* Anaya Multimedia. Recuperado de: https://books.google.com.pe/books?id=nlgtygAACAAJ&hl=es

Saif, H., Heb, Y., Fernandez, M. & Alani, H. (2016). *Contextual semantics for sentiment analysis of Twitter.* Recuperado de: http://www.sciencedirect.com/science/article/pii/S0306457315000242

Tapia, A. & Ruiz, E. (2013). *Diseño de un modelo de clasificación de opiniones subjetivas utilizando minería de textos, aplicado en análisis de redes sociales.* (Tesis de Grado). Universidad Señor de Sipán, Perú.

Ullate, A. (2014). *Análisis de las Universidades de Madrid en Twitter Utilizando Herramientas de Data Discovery* (Tesis de Grado). Universidad Francisco de Vitoria, Madrid.

Villena, J. (2015, Oct 13). *Introducción al Análisis de Sentimientos (Minería de Opiniones).* Recuperado de https://www.meaningcloud.com/es/blog/introduccion-al-analisis-de-sentimientos-mineria-de-opinion

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

ANEXOS Nº 1

Para poder conectarse con el API de Twitter, es necesario darse de alta como usuario en la misma; posteriormente se ingresa al **Application Management,** plataforma que generará los **keys** y **Access Tokens** necesarios para poder conectarse desde R Studio y obtener los tweets para poblar la base de datos. Disponible en: **https://apps.twitter.com/app/new.**

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe