UNIVERSIDAD DEL AZUAY

FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN

ESCUELA DE INGENIERÍA DE SISTEMAS Y TELEMÁTICA

IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERÍA DE TEXTOS EN LOS MENSAJES DEL MEDIO SOCIAL TWITTER.

PARA LOS EVENTOS DEL TRÁFICO VEHICULAR DE LA CIUDAD DE CUENCA.

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SISTEMAS Y TELEMÁTICA

AUTOR: JOSÉ ANDRÉS RAMÍREZ FANTONI

DIRECTOR: ING. MARÍA INÉS ACOSTA URIGÜEN

CUENCA, ECUADOR

2018

Dedicatoria

Este trabajo es dedicado a mis padres, Elena y José, quienes con su infinito amor y apoyo constante me han enseñado a cumplir mis metas y a luchar por mis sueños, quienes día a día con su trabajo esforzado y dedicación han sembrado en mí valores y principios que llevaré conmigo todos los días de mi vida.

A mis hermanos Letty y Mateo, quienes son parte fundamental en mi vida y una de las principales motivaciones para esforzarme y ser mejor cada día, los amo siempre y para siempre.

A mis abuelitos, Plinio y Letty, quienes con su ternura y palabras de apoyo siempre han confiado en mí, y me han motivado a cumplir este logro.

Y como no a todos mis familiares que han estado pendientes de mí y me han transmitido esa energía positiva todo el tiempo, en especial me gustaría mencionar a mis tíos Edgar y Ricardo quienes me ayudaron en más de una ocasión con los recursos necesarios para poder estudiar y terminar la carrera.

A Verónica Flores, mi muy estimadaamiga, quien a lo largo de toda la carrera me ha brindado su ayuda desinteresada, de una u otra manera siempre ha estado para mí, y nunca dejaré de agradecerlepor tanto.

A mi segundo hogar Burger King, que desde el primer día ha sido una aventura que me ha hecho crecer como ser humano, y me ha enseñado el verdadero valor del esfuerzo y la entrega constante.

A un amigo especial que siempre me motivó a continuar el camino y a meterle ganas a la vida con una sonrisa en el rostro, César Mogrovejo, aunque Dios quiso que se nos adelante, estoy seguro que celebra este logro conmigo desde arriba.

Agradecimientos

En primer lugar, debo darle gracias a Dios, por ser tan bueno conmigo, por todas sus bendiciones, por la salud y la vida, por permitirme cumplir este sueño.

A mis padres, por encender esa chispa en mí de esforzarme y seguir adelante hasta alcanzar las metas que me proponga.

A mi directora, Ing. María Inés Acosta, a quien tuve la dicha de conocer en este proyecto, y a quién le estoy inmensamente agradecido por su ayuda, disposición constante y empuje para llevar a cabo y con éxito este trabajo.

A mi director de escuela Ing. Marcos Orellana, quien con sus amplios conocimientos y experticia me encaminó en este proyecto y me facilitó los recursos académicos para lograr la ejecución del proyecto.

A mi gran amiga Belén Arias que me ayudó en el transcurso de todo el proyecto con sus conocimientos y recomendaciones.

Extensivo mi agradecimiento a toda la comunidad educativa Universidad del Azuay y su rector Ing. Francisco Salgado quien tuve la grata experiencia que sea mi asesor metodológico y docente.

Índice de contenidos

Índice de imágenes

Índice de tablas

Resumen

La investigación se centra en los algoritmos para minería de textos aplicados en datos no estructurados acerca de eventos de tránsito vehicular en la ciudad de Cuenca obtenidos del medio social Twitter. Se determinan como los algoritmos a evaluar a SVM, Naïve Bayes y C4.5. Se crea una herramienta denominada Matriz de evaluación de algoritmos para minería de textos, usando como medidas la precisión, la cobertura y el coeficiente de confianza cohen kappa. Se realizan dos tipos de experimentación, análisis de sentimiento usando la aplicación Rapidminer yfrecuencia de palabras usando la representación gráfica mediante nubes de palabras de RStudio. La experimentación permite definir cual/cuales son los mejores algoritmos para minería de textos y su impacto sobre los datos de estudio.

ABSTRACT

The research focused on algorithms for text mining applied in unstructured data obtained from the social media Twitter on traffic events in Cuenca. The algorithms to be evaluated were SVM, Naïve Bayes and C4.5. A tool called algorithm evaluation matrix for text mining was created. This tool used the accuracy, coverage and Cohen kappa coefficient of confidence as measures. Two types of experimentation were carried out, a feeling analysis using the Rapidminer application, and a word frequency analysis using graphical representation through RStudio word clouds. The experimentation allowed to define the best algorithms for text mining and their impact on the studied data.

Translated by

Ing. Paul Arpi

CAPITULO I –Introducción

Introducción

El presente trabajo de investigación profundizará las técnicas de minería de textos y el análisis de sentimiento en los comentarios de la red social Twitter para los eventos del tráfico vehicular de la ciudad. Se abordará el problema de los datos generados por la red social Twitter, catalogados como datos no estructurados, lo que en primer lugar implica un proceso de selección de datos de cuentas y hashtag relacionados al área de estudio, a fin de filtrar la información relevante. En un siguiente paso se procesará los datos con la extracción, almacenamiento y depuración de los elementos texto. Se realizarán pruebas de algoritmos sobre los datos previamente obtenidos y se realizará una selección según los resultados obtenidos por cada uno. Se aplicará los algoritmos probados y seleccionados previamente para el procesamiento a través de modelos de procesos, y se expondrá su impacto en el tratamiento de los datos y posterior obtención de información.

1.1 Objetivos

1.1.1 Objetivo general

Identificar los algoritmos relevantes que determinen la generación de patrones relacionados a un texto emitido por la red social Twitter con respecto al tráfico vehicular en la ciudad de Cuenca.

1.1.2 Objetivos específicos

- Sistematizar información acerca de técnicas y algoritmos de minería de textos, y el análisis de sentimiento en los comentarios del medio social Twitter.
- Establecer cinco algoritmos que se evaluarán en el caso de estudio.
- Diseñar una matriz de evaluación de algoritmos para la minería de textos.
- Encontrar la frecuencia de palabras y la asociación entre ellas a fin de ubicar patrones en colecciones de datos no estructurados con cada algoritmo previamente establecido.
- Registrar los valores obtenidos con cada algoritmo en la matriz de evaluación de algoritmos para la minería de textos.

- Analizar resultados obtenidos de la aplicación de algoritmos de minería de textos a comentarios de la red social Twitter.

1.2 Justificación

Los habitantes de las medianas y grandes ciudades diariamente deben vivir los problemas de la congestión vehicular a causa del continuo crecimiento del parque automotor. Las soluciones que se han planteado hasta el momento has sido fruto de la percepción de las autoridades de tránsito y las continuas quejas de los usuarios. Eventos como: cierre de calles, calles traficadas y no traficadas, colisiones, atropellamientos, derrumbes, eventos organizados, entre otras, son motivo de la generación de comentarios en redes sociales. Es necesario tomar la opinión de los usuarios y de las entidades de control para tratarlas y determinar patrones de comportamiento que retroalimenten a los usuarios en tiempo real, de modo que se tomen las previsiones adecuadas y se agilite el tránsito.

Para desarrollar un sistema como el propuesto es necesario establecer el mejor algoritmo para análisis de los datos que se generan en los medios sociales, en este caso de estudio se usará Twitter que ofrece datos catalogados como no estructurados y que necesitan ser tratados considerando la singularidad de las palabras. Al obtener un algoritmo específico que permita la obtención de los datos necesarios para el caso de estudio se puede continuar con el proceso de selección de herramientas y creación de modelos y técnicas que permitan encontrar los resultados esperados.

1.3 Alcance y limitaciones

En el proyecto se espera que el proceso de indagación, selección y evaluación sirva para determinar el mejor algoritmo de minería de textos aplicable al análisis de datos obtenidos del medio social Twitter.

Como resultado final se obtendrá la Matriz de Evaluación de Algoritmos de minería de textos y el análisis de cuál es el impacto producido de su aplicación en los datos no estructurados obtenidos de Twitter con respecto al tráfico vehicular en la ciudad de Cuenca.

CAPITULO II – Marco Teórico

Introducción

En este capítulo se desarrollará de manera conceptual todas las temáticas a tratar en la investigación, qué abarca la minería de textos dentro del tratamiento de la información, cual ha sido su evolución en el tiempo desde su creación, y en qué áreas de las ciencias se aplica.

Se dará a conocer en que consiste el análisis de sentimiento como metodología para tratar la información obtenida y cómo funciona en relación al caso de estudio.

Para finalizar se abordará el tema principal que consiste en los algoritmos para minería de textos,cuál es su definición, cuáles son sus tipos y clasificaciones, cuáles son los resultados que obtendremos de todo el proceso y como se deberán interpretar.

2.1 Minería de Textos

2.1.1 Definición

La minería de textos es un proceso automático de extracción de conocimiento e información importante partiendo de texto no estructurado.

La minería de textos opera combinando la habilidad humana para comunicar información usando diferentes idiomas, gramática, lenguaje natural y contextualización de las palabras, junto al de las tecnologías de la información que tienen la habilidad de procesar grandes cantidades de información de manera rápida y precisa.

Uno de los desafíos que tiene la minería de textos es interpretar la información que suele carecer de significado debido a la imprecisión del lenguaje natural que tenemos los seres humanos al comunicarnos.

2.1.2 Aplicación de la Minería de Textos

La minería de textos se puede resumir en dos grandes procesos: la conversión de la información a una forma estructurada, y el tratamiento de esta información mediante algoritmos que dan resultados mesurables que se pueden tratar con análisis estadístico.

Para la conversión de la información a estructurada se usan técnicas de almacenamiento en vectores que usan la idea de que un documento o un conjunto de palabras puede ser representado por una secuencia de palabras que tienen un peso dentro del conjunto, este conjunto se conoce como una representación en bolsa de palabras. Este peso se lo puede asignar de manera numérica según la relevancia de la palabra dentro del documento.

"Durante el proceso de minado de texto se aplican varias operaciones sobre los datos, según se necesite alcanzar un objetivo, como, por ejemplo:el análisis lingüístico a través de la aplicación de técnicas de procesamiento de lenguaje natural, filtrado de texto según la media de una palabra en específico, extracción de características, por ejemplo convirtiendo características textuales en características numéricas que un algoritmo de inteligencia artificial pueda procesar, y la selección de características por ejemplo la reducción del número de características para extraer sólo las más relevantes."(D'Andrea, Ducange, Lazzer, & Marcelloni, 2015).

El principal problema que se presenta en la minería de textos es la gran dimensión de información que se tiene que procesar, es por eso que la pre selección de información, y la selección de las características es de gran importancia.

Una vez se tiene la información a procesar se aplican algoritmos de minería de textos e inteligencia artificial como máquinas de vectores de soporte, redes neurales o árboles de decisiones para crear modelos de regresión, nubes de palabras o modelos de clasificación.

Al final los resultados obtenidos son interpretados por medidas estadísticas para verificar el porcentaje de efectividad alcanzado. Si los resultados obtenidos no son lo suficientemente confiables o tienen rangos de resultados muy variables e imprecisos se podría asignar otros valores en los parámetros de los algoritmos y repetir el proceso desde el comienzo.

2.1.3 Evolución en el tiempo

Según(Elder, y otros, 2012), la minería de textos nace de la necesidad de procesar el gran volumen de información textual que aparecía en el mundo, el problema radicaba en que se tenía la información acumulada en vastas

cantidades de texto almacenado en diferentes formas que necesitaban ser ordenadas.

Uno de los primeros ejemplos de clasificación de textos fue en las bibliotecas, el primer catálogo de biblioteca es atribuido a Thomas Hyde en 1674 en la Biblioteca Bodeliana de la Universidad de Oxford.

Luego en 1876 Melvin Dewey crea las tarjetas de índices para el catálogo de fichas de las bibliotecas. En estas fichas se procesaba el texto de tal manera que se extraían las palabras principales para generar los abstract.

Uno de los primeros intentos para resumir un texto más largo y complejo fue en 1958 cuando Luhn trabajando conjuntamente en el Instituto de Ingenieros Eléctricos y La sociedad de físicos de Londres adaptaron una computadora IBM 701 para generar abstracts de documentos. El proceso que realizaba era que contaba las palabras significativas y almacenaba las oraciones en las que se usaba, las palabras que más se repetían se les asignaba un valor de mayor significancia que hacía que las oraciones que las contenían formen el párrafo del abstract.

Durante los siguientes 40 años, los científicos de las bibliotecas se expandieron mucho en los trabajos previamente realizados, Doyle en 1961 basándose en el trabajo de Luhn para sugerir una nueva forma de clasificar la información en una biblioteca en forma de frecuencias de palabras y asociaciones. Afirmó que este sistema, un método altamente sistemático y automatizado para la navegación rápida de la información en una biblioteca era el más eficiente. Y aunque las ventajas para procesar con este método la información era seguro, pasaron varios años intentado aplicarlo con el menor costo posible.

2.1.4 Minería de Textos en Redes Sociales

La propagación del uso de las redes sociales en los últimos años, las vuelven una fuente gigantesca de información para todo tipo de estudio de análisis de datos.

Aunque la información es compartida de manera informal, y no siempre tiene el respaldo de ser veraz, suficiente o completa resulta ser de alta significancia al ser analizada en grandes volúmenes. Es en este punto cuando se combina la minería de textos con las redes sociales, usar las técnicas de minería de textos en datos obtenidos de redes sociales permite extraer información valiosa directamente de

los usuarios, en varios estudios las personas de las cuales se obtiene la información se los llama sensores sociales(Purohit, y otros, 2013), ya que al igual que los dispositivos que miden las temperaturas, presión, humedad, etc. que ofrecen datos para la minería de datos, los usuarios de las distintas redes sociales general datos no estructurados que sirven para análisis como el propuesto en este estudio.

2.1.4.1 Twitter

Twitter es un medio social que surgió a mediados del año 2006, que tiene una estructura de microblogging en el cual los usuarios publican actualizaciones en un máximo de 140 caracteres (en noviembre de 2017 se aumentó la capacidad a 280 caracteres a todos los usuarios), la plataforma permite que un usuario "siga" a otro es decir que se suscriba a otro usuario para ver sus publicaciones, creándose así la red.

Algunos términos importantes dentro de Twitter:

- Tweet: Conocido también como mensaje, actualización, post o estado de un usuario de Twitter que dentro del límite establecido de caracteres expresa actividades, sentimientos, pensamientos, en general información útil compartida con sus seguidores, que puede ser respondida, creando una conversación.
- Hashtag: O etiqueta que se determina como una palabra precedida por el símbolo # (ejemplo. #tránsito). Esta etiqueta dentro de la plataforma representa un tema de conversación, factor clave para comunicarse con grupos específicos. Un hashtag puede convertirse en un *trending topic* o tema de tendencia si muchos usuarios lo usan en sus tweets. Los hashtags de tendencia se pueden clasificar según la ubicación geográfica de los usuarios.
- Responder (*Reply*): Es una herramienta de la plataforma de Twitter mediante la cual un usuario puede como indica su nombre responder sobre un tweet de otro usuario, generando una publicación en la que automáticamente se menciona al usuario del tweet original, y la respuesta.
- Retwittear (*Retweet*): Esta herramienta permite publicar un tweet de otro usuario dentro del perfil, en este caso se menciona el nombre del usuario del mensaje original.

- Me gusta (*Like*): Herramienta que permite indicar que le agrada el tweet de otro usuario. Se lleva un contador de a cuantos usuarios les gusta la publicación.
- Mensaje Directo: Permite responder mediante un mensaje privado un tweet de otro usuario.
- Mención: Es etiquetar a un usuario dentro de un tweet propio utilizando el símbolo de @ más el nombre del usuario (ejemplo. @joseandresr)
- Url cortas: Los tweets pueden contener enlaces a páginas web externas, y para optimizar la cantidad de caracteres se suelen usar servicios que acortan las url deseadas.

Todos estos términos son importantes al momento de crear el set de datos para la experimentación, ya que se debe realizar tratamiento de los datos antes de procesarlos.

2.2 Datos no estructurados

Una definición apropiada para datos no estructurados es que son aquellos que no pueden ser almacenados en una base de datos tradicional (estructura relacional, jerárquica o de red), que provienen de diversos orígenes como redes sociales, e-mails, documentos de texto, foros, etc. Que tienen una alta tasa de crecimiento en relación a los datos estructurados y que requieren un tratamiento especial.

La secuencia de textos obtenidos de Twitter son por naturaleza no estructurados, pero los sistemas de minería de textos no pueden aplicar los algoritmos para obtener conocimiento en este tipo de datos, por lo tanto dentro del proceso de minería de textos se da alta importancia al hecho de que se debe realizar un pre procesamiento de los datos y crear formatos tratables que resalten las características más importantes en el tratamiento y se eliminen las triviales que retrasan y desvían los resultados. (Santana, Costaguta, & Missio, 2014)

2.3 Análisis de Sentimiento

El análisis de sentimiento de texto también se denomina computación de polaridad emocional; esta nueva área se ha convertido en una frontera floreciente de la minería de textos, por la aplicación de algoritmos para analizar la polaridad emocional contenida en un texto(Li & Wu, 2010). La eficacia de estos métodos es demostrada a través de la obtención de conocimiento a partir de la información expresada por

usuarios de la red social Twitter con respecto a los problemas de tráfico vehicular de la ciudad.

Relación con los medios sociales: Twitter: Con el significativo crecimiento de los mensajes generados por el usuario, Twitter se ha convertido en una red social donde millones de usuarios pueden intercambiar su opinión. El análisis de sentimientos en los datos de Twitter ha proporcionado una manera económica y efectiva de exponer la opinión pública a tiempo, lo cual es crítico para la toma de decisiones en varios dominios.(Tan, y otros, 2013). Sin embargo, solo unos cuantos estudios se enfocan en el campo de la transportación vehicular. En la actualidad se han realizado trabajos que vinculan el tráfico con el análisis de sentimiento por medio de lo que se denomina como el análisis de sentimiento para el tráfico (TSA)(Cao, y otros, 2014).

Las opiniones influyen sobre el comportamiento de las personas, así también, sobre la toma de decisiones, por ejemplo, una compañía al conocer la opinión que tienen sus clientes sobre sus productos puede mejorar la calidad de los mismos con el fin de satisfacer la necesidad de su cliente. Un análisis de sentimiento puede ser aplicado en diferentes ámbitos, por ejemplo, conocer la opinión sobre un político, sobre una campaña de marketing, sobre la popularidad de un cantante, entre otros.

Mediante este método se puede obtener resultados que demuestran el margen de aceptación que se tiene frente a los usuarios en un determinado tema.

El auge de los medios sociales involucra usuarios de todas las edades, poco a poco el rango de edad que estaba muy marcado en un principio, se está rompiendo por lo que es un buen medio para extraer la opinión de los usuarios. (Arias, 2016)

2.4 Metodología CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*) es la guía más usada en proyectos de minería de datos, y es la que se usa en el presente trabajo de investigación.

Cuenta con seis fases que no tienen una secuencia fija, sino adaptable según su necesidad.

Las fases tienen sus tareas generales y específicas que a su vez describen las acciones que deben ejecutarse para situaciones específicas. (Gallardo, 2009)

A continuación, detallo una breve descripción de las fases de la metodología:

1. Compresión del negocio o problema: es la más importante ya que marca el punto de partida del proyecto, posee como tareas determinar los objetivos del negocio, valoración de la situación, determinar los objetivos de la minería de datos, y realizar el plan del proyecto. De esta manera se sabrá desde donde parte, hacia donde se quiere llegar con el proyecto y cuáles serán las reglas a seguir durante su duración.
2. Comprensión de los datos: Comprende la recolección inicial de los datos, la descripción de los datos, la exploración de los datos y la verificación de la calidad de los datos.
3. Preparación de los datos: En esta fase los datos se preparan para adaptarse a las técnicas de minería de datos. Las tareas incluyen seleccionar los datos, limpiar los datos, estructurar los datos, integrar los datos y el formateo de los datos.
4. Modelado: En esta fase se eligen las técnicas de modelado más apropiadas para el proyecto específico, las tareas continúan con la generación del plan de prueba, construir y evaluar el modelo
5. Evaluación: Es esta fase se evalúa el modelo considerando los criterios de éxito del problema. Se evalúan los resultados, se revisa el proceso y se determinan los próximos pasos.
6. Implantación: En esta fase una vez se ha terminado el modelo y se ha validado, se crea un plan de implantación, un plan de monitoreo y mantención, un informe final y la documentación de experiencias en una revisión final del proyecto.

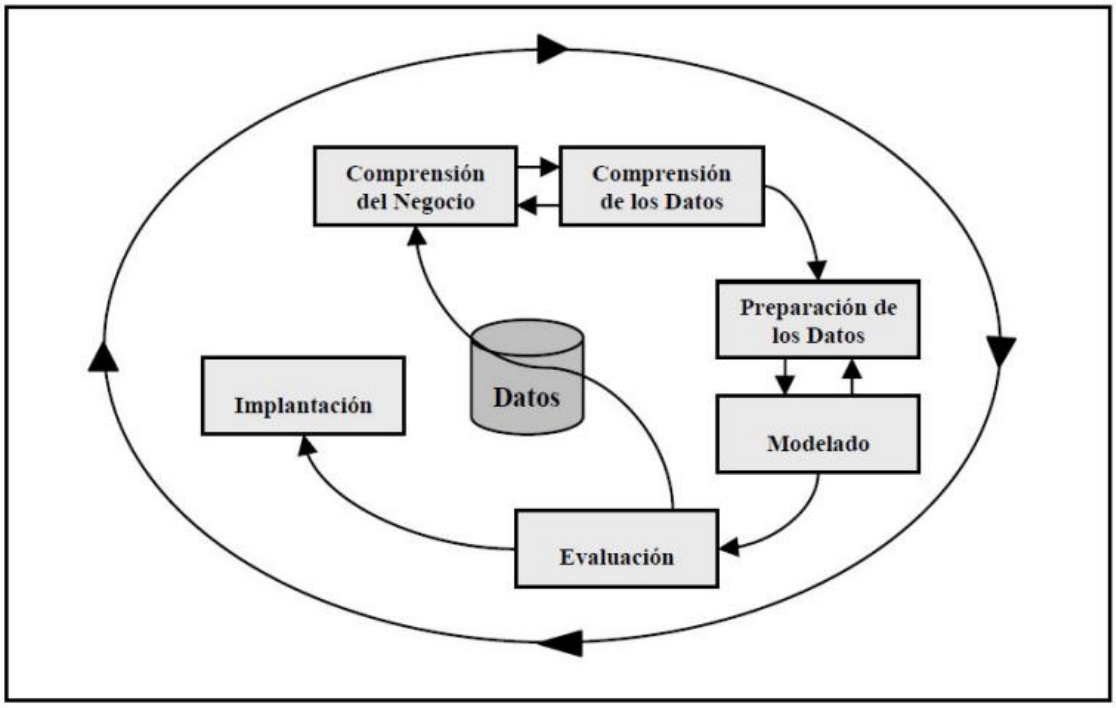


Ilustración 1. Modelo de proceso CRISP–DM ([CRISP-DM, 2000])

Fuente: *(Gallardo, 2009)*

2.5 Algoritmos para minería de textos

La obtención de información mediante la minería de textos requiere la implementación de algoritmos específicos que procesen los datos y generen resultados que nos faciliten tomar decisiones. Particularmente en la temática que se realiza en esta investigación se necesita usar algoritmos para clasificación de textos que nos permitan catalogar la base de datos disponible, y que nos permita ejecutar las tareas planteadas de identificación de frecuencia de palabras, patrones y comportamientos con respecto al tránsito vehicular.

En la revisión sistemática de la literatura realizada se determinaron diferentes algoritmos que son usados en proyectos de minería de textos, entre ellos tenemos: Support Vector Machines, Naïve Bayes, C4.5, k-means, Kohonen SOM, Geo-SOM, y variaciones de los mismos como Semi Naïve Bayes o k-core. También se encontraron algoritmos híbridos que eran usados sólo en ese proyecto de investigación.

Realizando un resumen de los algoritmos encontrados se decidió que los algoritmos que más se repiten dentro de la bibliografía son 3: Support Vector Machines (SVM), Naïve Bayes y C4.5. Los demás son usados una sola vez, y de acuerdo a los objetivos planteados se necesita encontrar el mejor algoritmo, así que consideramos que la

indagación y posterior experimentación se realizará en los tres algoritmos previamente detallados y sus variaciones.

Con los tres algoritmos seleccionados abarcaremos tres grandes campos de algoritmos para la clasificación de textos que son: los lineales, los bayesianos y los árboles de decisión.

2.5.1 SVM

Support Vector Machines es el algoritmo clasificador más usado en cuanto a detección de sentimientos(Montesinos García, 2014). Usa algoritmos lineales que clasifican los datos en dos categorías separadas lo más posible en un hiperplano.

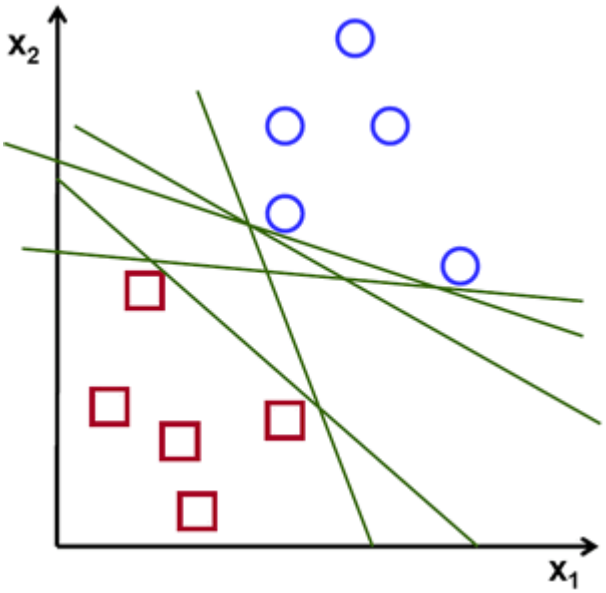


Ilustración 2 - Representación gráfica SVM

Tiene diferentes opciones de configuración, que son organizadas en métodos kernel o núcleos, los cuales establecen los datos dentro de espacios de características adecuadas.

Entre los métodos kernel disponibles en la herramienta de experimentación que se usará tenemos los siguientes según su documentación disponible:

1. **Punto**: se define por:

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \mathbf{x} * \mathbf{y}$$

es decir, es el producto interno de x y y.

2. **Radial**: El núcleo radial se define por:

$$\mathbf{exp}\ (-\mathbf{g}\ ||\ \mathbf{x}-\mathbf{y}\ ||\ \hat{}\ \mathbf{2})$$

donde g es gamma, que determina un papel importante en el rendimiento del kernel, y debe ajustarse según al problema en cuestión.

3. **Polinomio**: El núcleo polinomial se define por:

$$\mathbf{k}\ (\mathbf{x},\mathbf{y})\ =\ (\mathbf{x}\ *\ \mathbf{y}\ +\ \mathbf{1})\ \hat{}\ \mathbf{d}$$

donde d es el grado de polinomio. Los núcleos polinomiales son muy adecuados para problemas donde todos los datos de entrenamiento están normalizados.

4. **Neural**: el núcleo neural está definido por una red neural de dos capas

$$\mathbf{tan\ h}\ (\mathbf{a}\ \mathbf{x}\ *\ \mathbf{y}\ +\ \mathbf{b})$$

donde a es alfa y b es la constante de intersección. Un valor común para alfa es 1 / N, donde N es la dimensión de los datos. No todas las elecciones de a y b conducen a una función de kernel válida.

5. **Anova**: El núcleo de anova se define por el incremento de la suma de:

$$\mathbf{exp}\ (-\mathbf{g}\ (\mathbf{x}-\mathbf{y}))$$

donde g es gamma y se puede aumentar o disminuir el grado a usar mediante una variable d.

6. **Epachnenikov**: El kernel epachnenikov usa la función:

$$(\mathbf{3}/\mathbf{4})\ (\mathbf{1}-\mathbf{u2})$$

para u entre -1 y 1 y cero para u fuera de rango.

7. **Combinación gaussiana:** este kernel de combinación gaussiana tiene los parámetros ajustables sigma1, sigma2 y sigma3.

8. **Multicuadrático**: El núcleo multicuadrático está definido por la raíz cuadrada de

$$|| x - y || \hat{} 2 + c \hat{} 2.$$

SVM asigna el criterio de separación entre clases que se encuentre 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 entorno a este tendrán una menor probabilidad de ser catalogados correctamente."(Montesinos García, 2014).

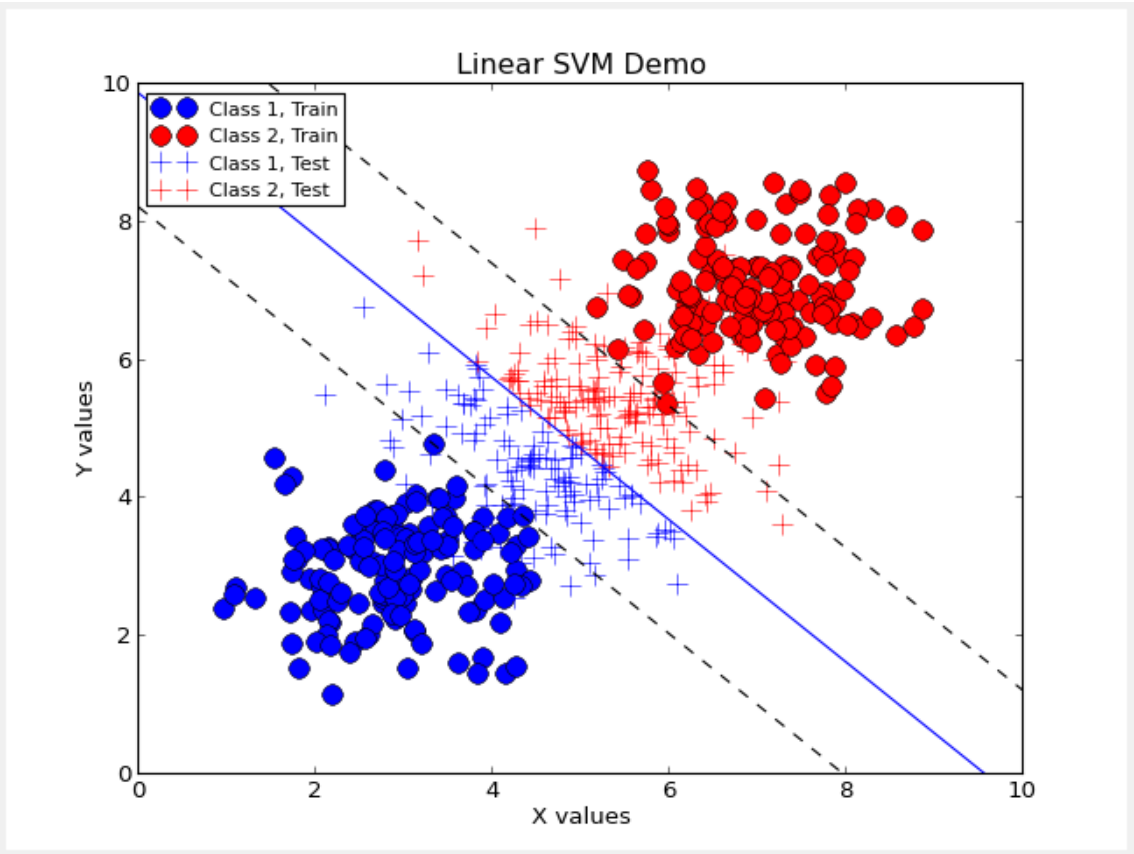


Ilustración 3- Ejemplo SVM

2.5.2 Naïve Bayes

Un clasificador de Naïve Bayes es un clasificador probabilístico simple basado en la aplicación del teorema de Bayes (de las estadísticas bayesianas) con suposiciones de independencia fuertes.

Un clasificador de NaïveBayes supone que la presencia de una característica particular de una clase no está relacionada con la presencia (o ausencia) de ninguna otra característica. Incluso si estas características dependen unas de otras o de la existencia de otras características, un clasificador Naïve Bayes considera que todas estas propiedades contribuyen independientemente a la probabilidad de que se cumpla la respuesta correcta.

"Sea $X\{x_{1,}x_{2,}\ldots\ldots,x_{n,}\}$, un ejemplo del cual se determinará la clase C de pertenencia. Sea H alguna hipótesis. Para problemas de clasificación se desea determinar, la probabilidad de que la hipótesis sostenida da al ejemplo X. El teorema de BAYES proporciona un algoritmo eficiente para calcular esta probabilidad a priori, mediante la siguiente formula:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

La clasificación de un ejemplo de X viene dada por el valor de máxima probabilidad a posterior de la clase C, donde n representa el número de atributos del ejemplo X $\{5\}$."(Suca, Córdova, Condori, Cayra, & Sulla, 2016).

$$C = \operatorname{argmax} P(Hj) \prod_{i=1}^{n} P(\frac{X_i}{H_j})$$

El clasificador Naïve Bayes ofrece una alta exactitud cuándo se aplica a grandes bases de datos. Las ventajas de Naïve Bayes incluye:

- Menor costo computacional.
- Requiere sólo de una pequeña cantidad de datos de entrenamiento.
- Fácil de entender al ser un clasificador basado en probabilidades, y requiere un solo paso de procesamiento si los datos son discretos.
- Comparado con las redes bayesianas el clasificador Naïve Bayesiano es mucho más rápido y si los atributos son independientes puede tener la mayor tasa de acierto.

2.5.3 C4.5

El C4.5 es una máquina de aprendizaje para predicciones con una variable dependiente que puede llegar al objetivo deseado, basándose en los atributos de los datos disponibles. Este método también es conocido como J48 en la herramienta Weka.

Los nodos internos son los atributos, las ramas son los posibles valores y los nodos finales u hojas son la clasificación.

Este método al ser iterativo va colocando los posibles valores de las características, y cuando todas son clasificados y ya no exista ambigüedad entonces se asigna una raíz.

El árbol de decisión J48 tiene una alta precisión en comparación con los clasificadores SVM y Naïve Bayes.

Entre algunas características de este algoritmo tenemos que:

- Permite el manejo de atributos continuos y discretos.
- Para manejar atributos continuos, C4.5 crea un umbral y luego se divide la lista en aquellos cuyo valor de atributo es superior al umbral y los que son menores o iguales a él.
- Manejo de los datos de formación con valores de atributos faltantes
- Manejo de atributos con costos diferentes.
- Para poder podar árboles después de la creación, C4.5 se regresa a través del árbol una vez que ha sido creado e intenta eliminar las ramas que no ayudan, reemplazándolos con los nodos hoja.

Un árbol básicamente consta de una raíz y puede tener de dos a más nodos hijos o intermedios. En la mayoría de árboles se considera un nodo intermedio como un nodo de predicción o regla, la cual nos dirá de acuerdo a la categoría o valor numérico porque rama debe bajar. Esto puede ser visto como nodos circulares para los de predicción y rectangulares para los de decisión en la figura siguiente. Todo esto se usará al momento de predecir nuevos valores.(Suca, Córdova, Condori, Cayra, & Sulla, 2016).

CAPITULO III – Matriz de Evaluación de Algoritmos para Minería de Textos

Introducción

En este capítulo se diseña una matriz en la cual se registrarán los resultados de la experimentación. Se eligen los parámetros a evaluar y se les asigna un valor que repercutirá en la selección del mejor algoritmo.

3.1 Medidas a evaluar

Las medidas que se han elegido para la evaluación se han tomado de acuerdo a estudios similares(Suca, Córdova, Condori, Cayra, & Sulla, 2016), (Ing. Corso, 2009), y se complementarán entre ellas para darnos una representación precisa de los resultados de la experimentación.

- Falsos Positivos:Es el número de predicciones negativas que se reconocieron como positivas.
- Falsos Negativos: Es el número de predicciones positivas que se reconocieron como negativas.
- Verdaderos Positivos: Es el número de predicciones positivas que se clasificaron como positivas.
- Verdaderos Negativos: Es el número de predicciones negativas que se clasifican como negativas.
- Precisión: Porcentaje de predicciones correctas.

$$\text{precisión} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

VP= Verdaderos positivos

FP=Falsos positivos

- Error: Porcentaje de las predicciones incorrectas.

- Cobertura (*Recall*): Permite medir la proporción de verdaderos positivos con respecto a todos los términos reales.

$$\text{cobertura} = \frac{\text{VP}}{\text{VP} + \text{VN}}$$

VP= Verdaderos positivos

VN=Verdaderos negativos

- Coeficiente Cohen Kappa:es un valor estadístico de concordancia que tiene en cuenta la predicción correcta que se produce por casualidad.

Según (Landis & Koch, 1977)la fuerza de concordancia del coeficiente kappa se puede regir según la siguiente tabla:

Fuerza de la concordancia	**Coeficiente kappa**	**Porcentual asignado**
Pobre	0.00	0
Leve	0.01 a 0.20	20
Aceptable	0.21 a 0.40	40
Moderada	0.41 a 0.60	60
Considerable	0.61 a 0.80	80
Casi perfecta	0.81 a 1.00	100

3.2 Diseño de Matriz de Evaluación de Algoritmos

Se han desarrollado modelos de matrices para cada algoritmo y sus parámetros de configuración, en las cuales se determinará la mejor configuración para cada algoritmo. Y posteriormente una matriz global en la cual se resumirán los mejores algoritmos con su configuración correspondiente de cada categoría para elegir el mejor de todo el grupo.

- Matriz de evaluación Support Vector Machines

MATRIZ DE EVALUACIÓN - SUPPORT VECTOR MACHINES									
Parámetro	**Análisis de Sentimiento**		**Coeficientes**				**Resultados**		**Total**
	Positivos	**Negativos**	**Kappa (0.1)**			**Cobertura (%) (0.15)**	**Precisión (%) (0.75)**	**Margen (+/-)**	**Calificación (%)**
			Valor	**Fuerza de concordancia**	**Asignación (%)**				
Punto									
Radial									
Polinomio									
Neural									
Anova									
Epachnenikov									
Combinación gaussiana									
Multicuadrático									

Tabla 1 – Matriz de evaluación Support Vector Machines

MATRIZ DE EVALUACIÓN - NAIVE BAYES									
Parámetro	**Análisis de Sentimiento**		**Coeficientes**				**Resultados**		**Total**
	Positivos	**Negativos**	**Kappa (0.1)**			**Cobertura (%) (0.15)**	**Precisión (%) (0.75)**	**Margen (+/-)**	**Calificación (%)**
			Valor	**Fuerza de concordancia**	**Asignación (%)**				
Con corrección de Laplace									
Sin corrección de Laplace									

- Matriz de evaluación Naïve Bayes

Tabla 2 – Matriz de evaluación Naïve Bayes

- Matriz de evaluación C4.5

MATRIZ DE EVALUACIÓN – ALGORITMOS PARA MINERÍA DE TEXTOS										
Algoritmo	**Parámetro**	**Análisis de Sentimiento**		**Coeficientes**				**Resultados**		**Total**
		Positivos	**Negativos**	**Kappa (0.1)**			**Cobertura (%) (0.15)**	**Precisión (%) (0.75)**	**Margen (+/-)**	**Calificación (%)**
				Valor	**Fuerza de concordancia**	**Asignación (%)**				
(Mejor de SVM)										
(Mejor de NB)										
(Mejor de C4.5)										

Tabla 3 – Matriz de evaluación C4.5

- Matriz de comparación de resultados de todos los algoritmos

Tabla 4 – Matriz de evaluación Algoritmos para minería de textos

3.3 Establecimiento de parámetros y rangos de evaluación

Para crear la calificación final de cada algoritmo, de todos los valores revisados anteriormente se ha considerado tres parámetros principales que son la precisión de la clasificación, y los coeficientes kappa y cobertura. La precisión es el valor que mayor incidencia tiene en la calificación, ya que le corresponde el 75% del total, el coeficiente de kappa el 10% y el coeficiente de cobertura el 15%.

Para elegir el mejor algoritmo se considera el mejor de cada categoría, y de entre estos el que tenga la mejor calificación.

3.4 Conclusiones

En este capitulo se ha definido una matriz para registrar los resultados de la

MATRIZ DE EVALUACIÓN - C4.5									
Parámetro	**Análisis de Sentimiento**		**Coeficientes**				**Resultados**		**Total**
	Positivos	**Negativos**	**Kappa (0.1)**			**Cobertura (%) (0.15)**	**Precisión (%) (0.75)**	**Margen (+/-)**	**Calificación (%)**
			Valor	**Fuerza de concordancia**	**Asignación (%)**				
Default									

experimentación, en toda la bibliografía estudiada no se ha encontrado una herramienta

similar, así que se ha creado un instrumento nuevo que servirá para registrar los resultados de la experimentación.

CAPITULO IV – Experimentación

Introducción

En el capítulo a desarrollarse a continuación se realiza la experimentación que permitirá identificar el mejor algoritmo para minería de textos en el caso de estudio propuesto.

Usando la metodología CRISP-DM como guía se procede a realizar la minería de textos con los algoritmos previamente expuestos, usando las matrices diseñadas usando como herramienta Rapidminer.

4.1 Obtención de los datos

Para la obtención de los datos para la experimentación se usa una cuenta de Twitter en la que hay que iniciar sesión y realizar una secuencia de pasos.

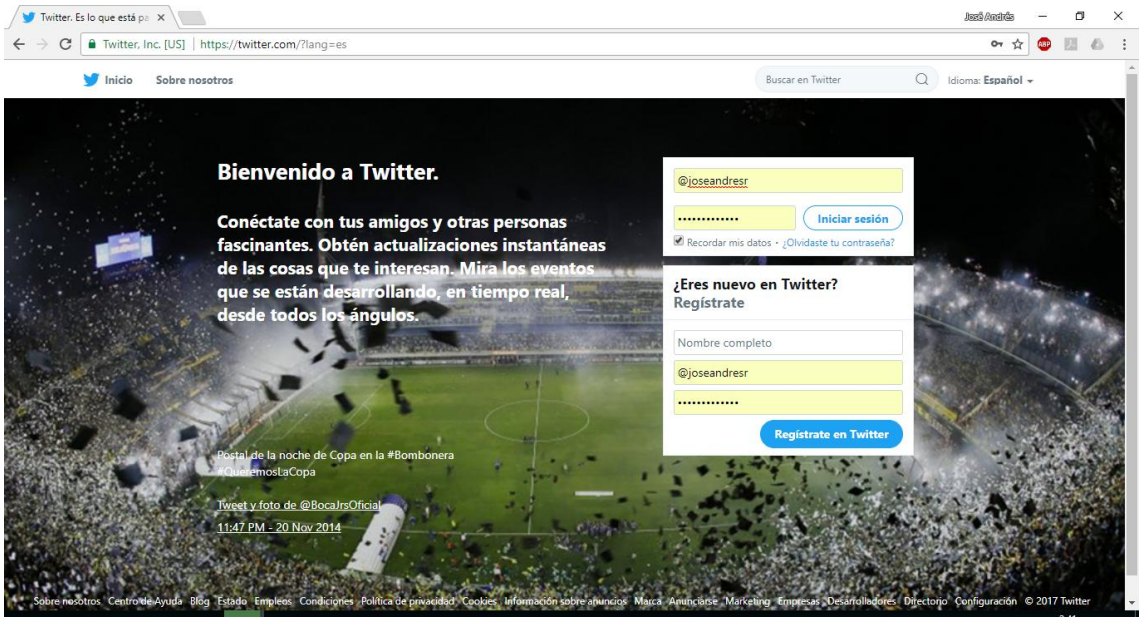


Ilustración 4 - Página de inicio de sesión Twitter.

Tanto para los datos de entrenamiento como para los datos de prueba se sigue un proceso similar que se detalla a continuación.

4.1.1 Datos de entrenamiento

Se extraen los tweets de entrenamiento utilizando la siguiente cadena de búsqueda:

"(circulación OR circulacion OR tránsito OR transito OR flujo OR congestión OR congestion OR tráfico OR trafico) AND vehicular"

esta cadena se construyó buscando los sinónimos de la palabra congestión con el fin de abarcar la mayor cantidad de tweets que se relacionen a la congestión vehicular.

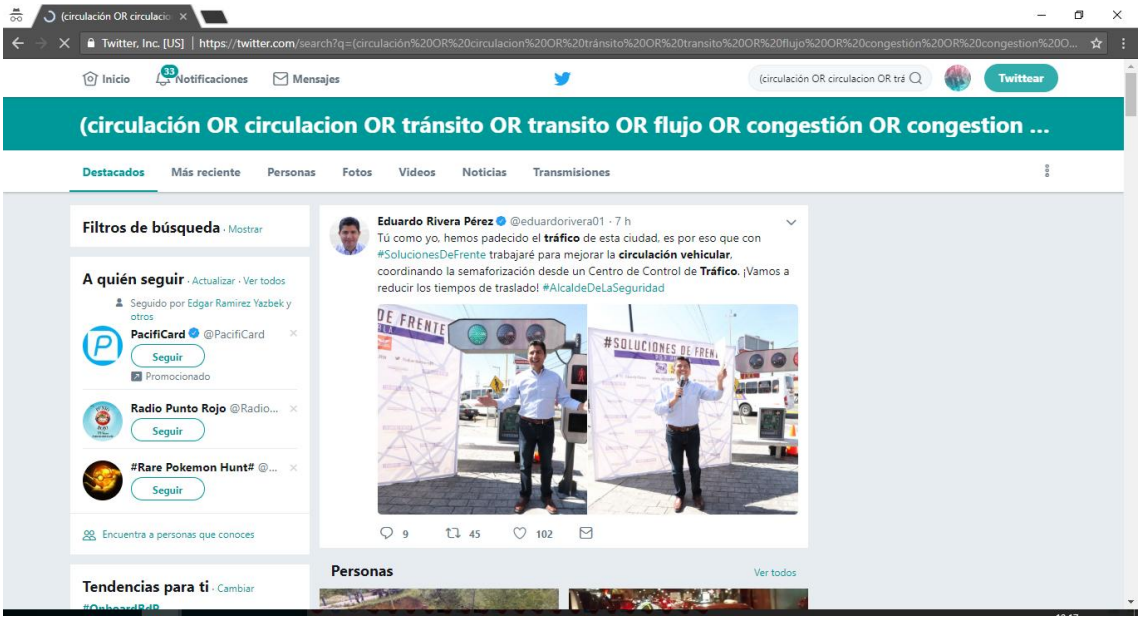


Ilustración 5- Búsqueda de tweets para base de datos de entrenamiento.

Para la obtención de los tweets se utiliza la API de Twitter y el lenguaje estadístico R, y se realiza el siguiente proceso:

1. Se ingresa al sitio web de desarrolladores de Twitter que es: https://apps.twitter.com/.

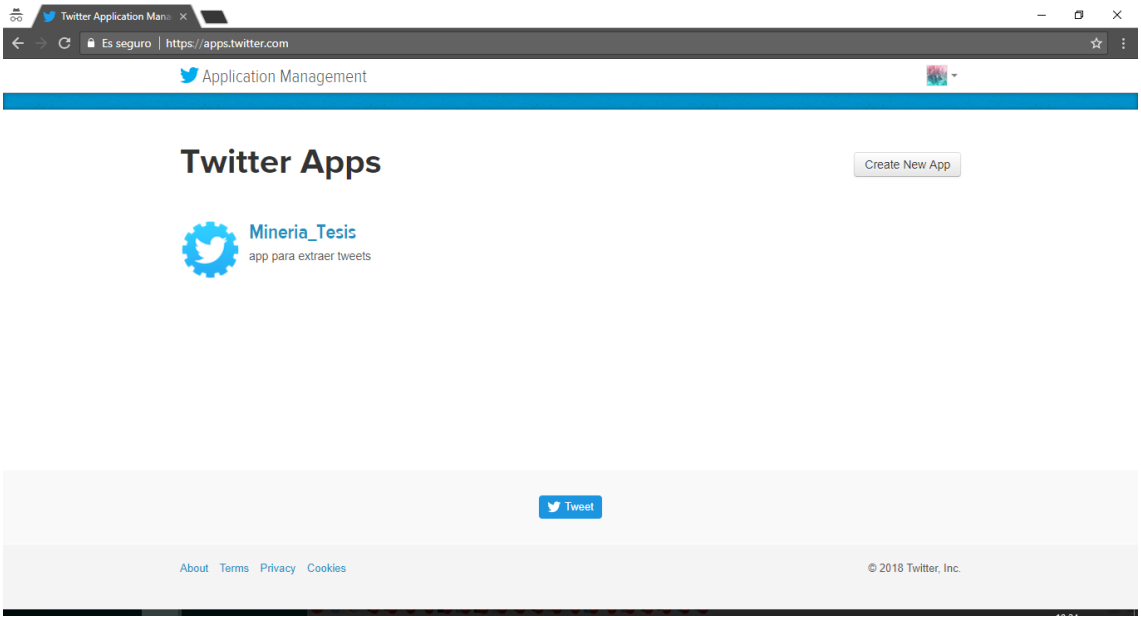


Ilustración 6 Página principal Twitter Apps

2. Se crea una aplicación nueva, y se registran los campos requeridos.

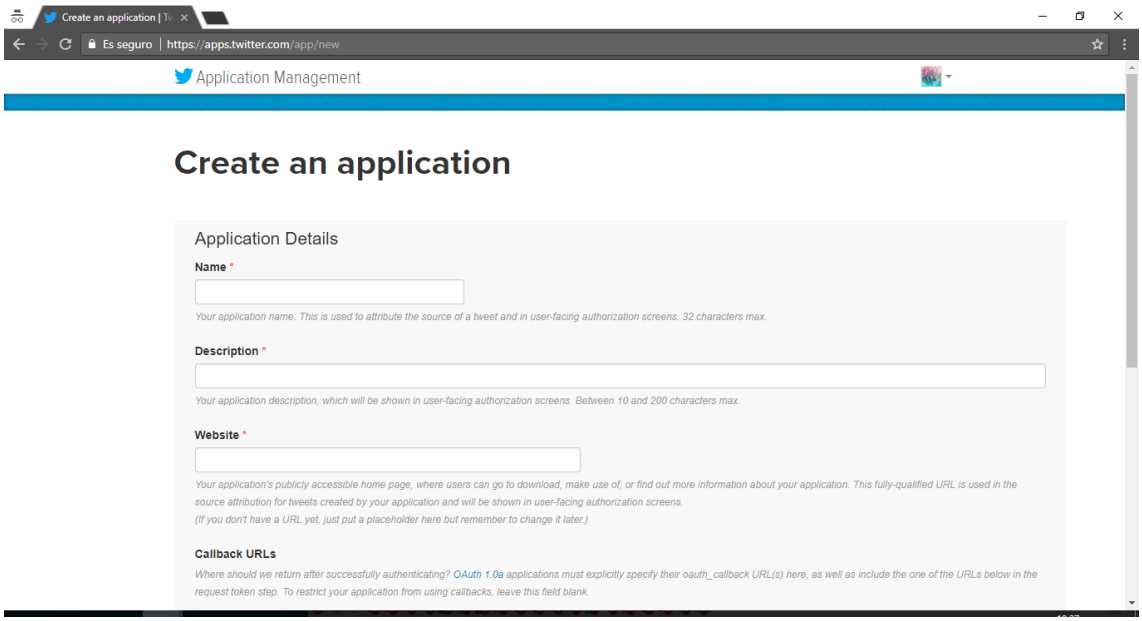


Ilustración 7 Pantalla para crear una nueva aplicación en Twitter Apps

Los campos para la creación de la nueva aplicación son:

Nombre: Se le asigna un nombre a la aplicación, este no debe tener más de 32 caracteres.

- Descipción: Una descripción breve de la aplicación, dispone de entre 10 y 200 caracteres para registrarse.
- Website: Es la dirección web que usará la API, en este caso se registra la que da Twitter como ejemplo para las pruebas: http://127.0.0.1:1410.
- Callback URLs: En este campo se puede registrar la dirección a la que se quiere redirigir luego de la conexión de la API, sin embargo, no es un campo requerido y al no ser necesario en la experimentación se deja en blanco.

Una vez llenos los datos requeridos, se aceptan los términos y condiciones de uso de la API de Twitter y se crea la aplicación.

3. Una vez creada la aplicación tenemos a nuestra disposición las claves para consumir los servicios en R mediante la librería twitteR, estos datos son el *consumer key*, *consumer secret*, *access token* y *access secret.*

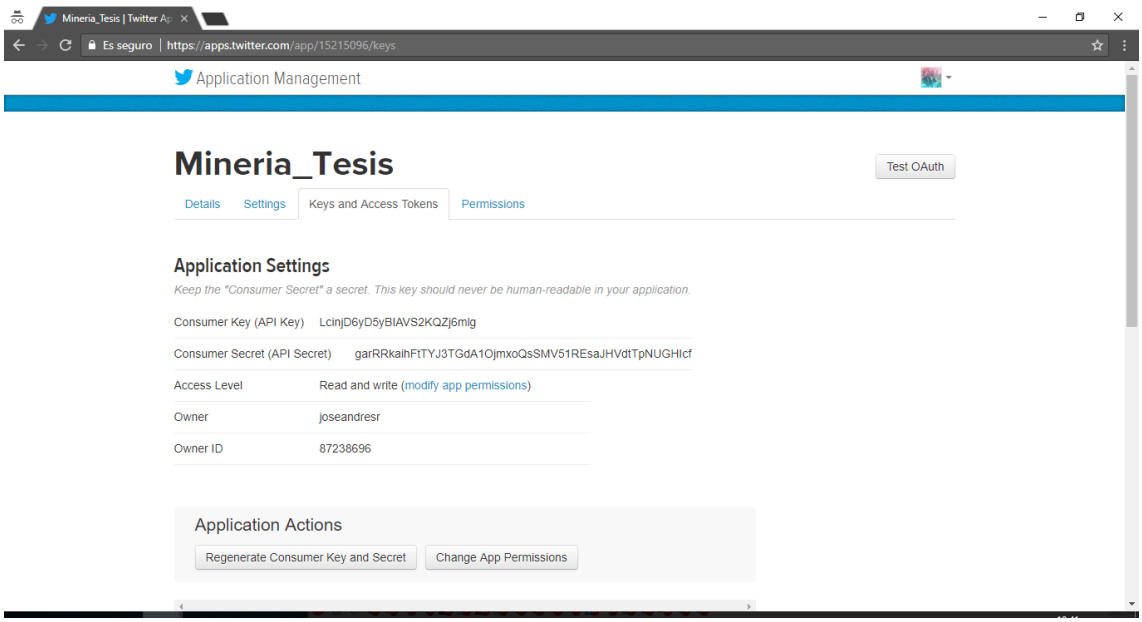


Ilustración 8 Pantalla de configuración de la aplicación de Twitter.

4. El paso final de la obtención de los tweets es extraerlos desde R. Se debe realizar una conexión entre R y Twitter, luego se inicia sesión en una cuenta

de Twitter, y de esta manera comenzar la extracción en tiempo real. Usando el parámetro *search* se ingresa la cadena de búsqueda, y los tweets obtenidos son transformados a textos.

Una vez obtenida la base de textos, para crear la base binomial de polaridad positiva o negativa los tweets se procesaron con un clasificador humano. Para mejorar la precisión de la clasificación, fueron procesados en la extensión *Meaning Cloud* de Rapidminer que clasifica a los tweets según su polaridad.

Así mismo, se utilizó el factor de cohen kappa para la validación de la clasificación.

4.1.2 Datos de prueba

Para los tweets de prueba se utiliza la siguiente cadena de búsqueda:

(circulación OR circulacion OR tránsito OR transito OR flujo OR congestión OR congestion OR tráfico OR trafico) AND vehicular near:CUE

Near es una palabra reservada de Twitter que hace referencia a la localidad en la que fue emitido el tweet, CUEcorresponde a la ciudad de Cuenca en Ecuador. (IATA, 2017)

Dentro de el código en R es necesario escribir la cadena de búsqueda de la siguiente manera:

("(circulación OR circulacion OR tránsito OR transito OR flujo OR congestión OR congestion OR tráfico OR trafico) AND vehicular", n=1000, geocode = '-2.8974039,-79.00448340000003,10mi')

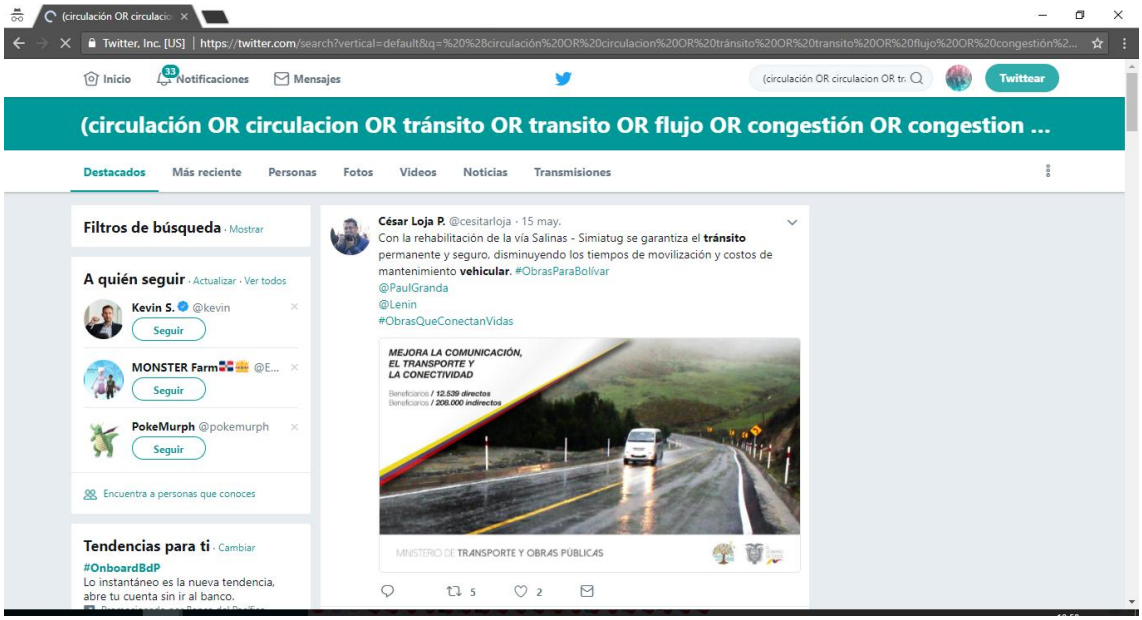


Ilustración 9 - Búsqueda de tweets para base de datos de prueba.

4.2 Preprocesamiento de los datos

Para el preprocesamiento de los datos se han considerado los siguientes tratamientos:

4.2.1 Tokenización

Consiste en la separación del texto de cada tweet en secuencias o tokens. Existen diferentes maneras de especificar cuales serán los puntos de separación, sin embargo, para este tipo de tratamiento se necesita separar usando la opción de caracteres que no son letras, de esta manera el texto se separa en palabras simples que se almacenaran en el vector de palabras.

4.2.2 Transformación de letras

Para trabajar con texto unificado es necesario transformar todas las letras de las palabras a minúsculas o mayúsculas, en la experimentación se convierte todo a minúsculas.

4.2.3 Eliminación de palabras innecesarias

Se aplica un filtro al vector de palabras, para que elimine aquellas que coinciden con un diccionario de palabras conectoras que para el texto en conjunto facilitan su compresión, pero que para el análisis a realizarse no se requieren.

4.3 Procesamiento de los datos

4.3.1 Rapidminer

Para realizar la experimentación se usó el modelo definido para análisis de sentimiento que ofrece Rapidminer.

Es un proceso de 4 pasos en el que se detecta el sentimiento deltexto mediante un modelo de clasificación, éste a su vez es entrenado con una base clasificada manualmente por observadores.

- Paso 1: Importar datos de texto que tengan la evaluación del sentimiento relacionado con él. Se realiza el pre procesamiento para extraer las palabras y crear un vector de palabras (una representación numérica del texto).
- Paso 2: Entrenar el modelo probando los diferentes algoritmos, y ejecutando una validación cruzada para recolectar los datos de rendimiento.
- Paso 3: Crear un nuevo documento a partir de la base de datos de prueba, luego procesar el texto como se realizó en el paso 1.
- Paso 4: El modelo entrenado con la base de entrenamiento se aplica al nuevo documento que se genera de la base de prueba.

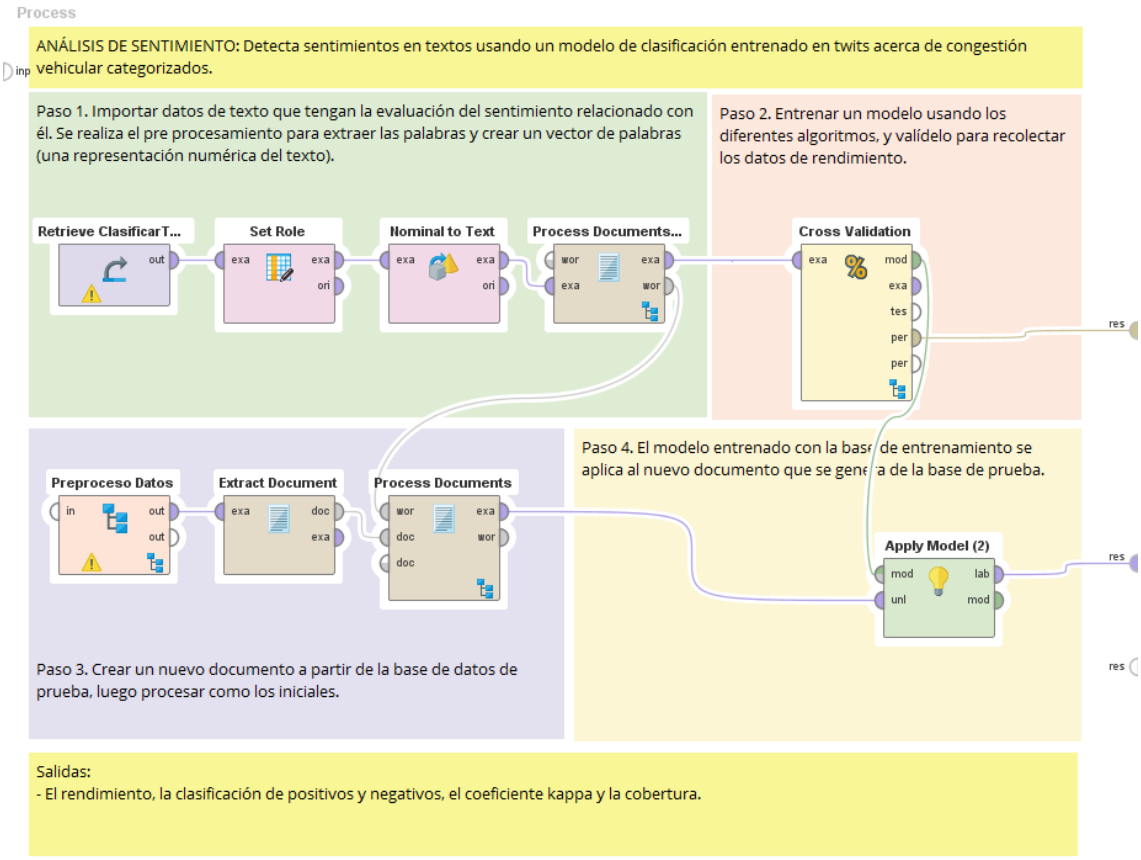


Ilustración 10 Modelo de Análisis de Sentimiento en Rapidminer

4.3.2 Especificaciones técnicas de equipo para experimentación

El equipo en el que se realizó la experimentación tiene las siguientes características técnicas:

- Sistema Operativo: Windows 10
- Procesador: Intel® Core™ i7-4702MQ CPU @ 2.2GHz 2.20GHz
- Memoria RAM: 12.0 GB
- Tipo de sistema: Sistema operativo de 64 bits, procesador x64
- Descripción del equipo: HP ENVY 15-j108la

4.4 Experimentación

4.4.1 Resultados de experimentación análisis de sentimiento por algoritmos

4.4.1.1 SVM

- Punto

  Matriz de confusión:

	Verdaderos negativos	Verdaderos positivos	Presición de la clase
Predicciones negativo	2611	263	90.85%
Predicciones positivo	153	447	74.50%
Class recall	94.46%	62.95%	

- Radial

  Matriz de confusión:

	Verdaderos negativos	Verdaderos positivos	Presición de la clase
Predicciones negativo	2731	542	84.44%
Predicciones positivo	33	168	83.58%
Class recall	98.81%	23.66%	

- Polinomio

  Matriz de confusión:

	Verdaderos negativos	Verdaderos positivos	Presición de la clase
Predicciones negativo	2757	671	80.43%
Predicciones positivo	7	39	84.78%
Class recall	99.75%	5.49%	

- Neural

  Matriz de confusión:

	Verdaderos negativos	Verdaderos positivos	Presición de la clase
Predicciones negativo	2764	710	79.56%
Predicciones positivo	0	0	0%
Class recall	100%	0%	

- Anova

  Matriz de confusión:

	Verdaderos negativos	Verdaderos positivos	Presición de la clase
Predicciones negativo	2764	710	79.56%
Predicciones positivo	0	0	0%
Class recall	100%	0%	

- Epachnenikov

  Matriz de confusión:

	Verdaderos negativos	Verdaderos positivos	Presición de la clase
Predicciones negativo	2731	542	84.44%
Predicciones positivo	33	168	83.58%
Class recall	98.81%	23.66%	

- Combinación gaussiana

  Matriz de confusión:

	Verdaderos	Verdaderos	Presición de la

	negativos	positivos	clase
Predicciones negativo	N/A	N/A	N/A
Predicciones positivo	N/A	N/A	N/A
Class recall	N/A	N/A	

Observación: Con este método kernel la experimentación no terminó en un periodo de procesamiento de 36 horas, por lo tanto, se descartó de la evaluación.

- Multicuadrático

  Matriz de confusión:

	Verdaderos negativos	Verdaderos positivos	Presición de la clase
Predicciones negativo	N/A	N/A	N/A
Predicciones positivo	N/A	N/A	N/A
Class recall	N/A	N/A	

Observación: Con este método kernel la experimentación no terminó en un periodo de procesamiento de 36 horas, por lo tanto, se descartó de la evaluación.

MATRIZ DE EVALUACIÓN - SUPPORT VECTOR MACHINES									
Parámetro	**Análisis de Sentimiento**		**Coeficientes**				**Resultados**		**Total**
	Positivos	**Negativos**	**Kappa (0.1)**			**Cobertura (%) (0.15)**	**Precisión (%) (0.75)**	**Margen (+/-)**	**Calificación (%)**
			Valor	**Fuerza de concordancia**	**Asignación (%)**				
Punto	447	2611	0.608	Considerable	80	62.92	88.02	1.06	83.45
Radial	168	2731	0.305	Aceptable	40	23.67	83.45	0.82	70.14
Polinomio	39	2757	0.08	Leve	20	5.49	80.48	0.53	63.18
Neural	0	2764	0	Pobre	0	0	79.56	0.16	59.67
Anova	0	2764	0	Pobre	0	0	79.56	0.16	59.67
Epachnenikov	168	2731	0.305	Aceptable	40	23.67	83.45	0.82	70.14
Combinación gaussiana	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.00
Multicuadrático	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.00

Tabla 5 - Matriz de evaluación Support Vector Machines Final

4.4.1.2 Naïve Bayes

- Con corrección de Laplace

  Matriz de confusión:

	Verdaderos negativos	Verdaderos positivos	Presición de la clase
Predicciones negativo	2323	183	92.70%
Predicciones positivo	441	527	54.44%
Class recall	84.04%	74.23%	

- Sin corrección de Laplace

  Matriz de confusión:

	Verdaderos negativos	Verdaderos positivos	Presición de la clase
Predicciones negativo	2323	183	92.70%
Predicciones positivo	441	527	54.44%
Class recall	84.04%	74.23%	

MATRIZ DE EVALUACIÓN - NAIVE BAYES									
Parámetro	**Análisis de Sentimiento**		**Coeficientes**				**Resultados**		**Total**
	Positivos	**Negativos**	**Kappa (0.1)**			**Cobertura (%) (0.15)**	**Precisión (%) (0.75)**	**Margen (+/-)**	**Calificación (%)**
			Valor	**Fuerza de concordancia**	**Asignación (%)**				
Con corrección de Laplace	527	2323	0.514	Moderada	60	74.2	82.04	2.02	78.66
Sin corrección de Laplace	527	2323	0.514	Moderada	60	74.2	82.04	2.02	78.66

Tabla 6 - Matriz de evaluación Naïve Bayes Final

4.4.1.3 C4.5

- Matriz de confusión:

	Verdaderos negativos	Verdaderos positivos	Presición de la clase
Predicciones negativo	2573	255	90.98%
Predicciones positivo	191	455	70.43%
Class recall	93.09%	64.08%	

MATRIZ DE EVALUACIÓN - C4.5									
Parámetro	**Análisis de Sentimiento**		**Coeficientes**				**Resultados**		**Total**
	Positivos	**Negativos**	**Kappa (0.1)**			**Cobertura (%) (0.15)**	**Precisión (%) (0.75)**	**Margen (+/-)**	**Calificación (%)**
			Valor	**Fuerza de concordancia**	**Asignación (%)**				
Default	455	2573	0.591	Moderada	60	64.08	70.74	5.23	68.67

Tabla 7 - Matriz de evaluación C4.5 Final

4.4.2 Resumen de resultados de experimentación de análisis de sentimientos.

MATRIZ DE EVALUACIÓN – ALGORITMOS PARA MINERÍA DE TEXTOS										
Algoritmo	**Parámetro**	**Análisis de Sentimiento**		**Coeficientes**				**Resultados**		**Total**
		Positivos	**Negativos**	**Kappa (0.1)**			**Cobertura (%) (0.15)**	**Precisión (%) (0.75)**	**Margen (+/-)**	**Calificación (%)**
				Valor	**Fuerza de concordancia**	**Asignación (%)**				
SVM	Punto	447	2611	0.608	Considerable	80	62.92	88.02	1.06	83.45
NAIVE BAYES	Con corrección de Laplace	527	2323	0.514	Moderada	60	74.2	82.04	2.02	78.66
C4.5	Default	455	2573	0.591	Moderada	60	64.08	70.74	5.23	68.67

4.4.3 Experimentación frecuencia de palabras

Una vez terminada la experimentación del análisis de sentimiento y la determinación de los mejores algoritmos se procede a la experimentación para obtener la frecuencia de las palabras que más veces se repiten, y la relación que existe entre ellas.

Para realizar la representación gráfica de las palabras y su frecuencia se usa nubes de palabras. Dentro de la utilidad Rapidminer que se usó anteriormente no existe la posibilidad de generar nubes de palabras por lo que se requiere de una utilidad adicional para representar los resultados obtenidos. En este caso se usa el lenguaje R y la utilidad RStudio para realizar la representación.(Arias, 2016)

Luego de extraer la base de textos clasificada tras la experimentación del análisis de sentimiento con cada algoritmo se procedió a graficar nubes de palabras usando la biblioteca *Wordcloud* de R.

La biblioteca *Wordcloud* tiene diferentes parámetros para graficar un set de datos, a continuación, detallo paso a paso el proceso seguido durante el proceso.

- Paso 1: Exportar la base de textos clasificada en Rapidminer usando la función *Export to Excel*.
- Paso 2: Dividir el vector de palabras resultante que cuenta con tres columnas, la primera contiene la palabra, la segunda el número de veces que la palabra fue usada en twits considerados positivos, y en la tercera el número de veces que la palabra fue usada en twits considerados negativos. Se guarda en un archivo de valores separados por coma, la palabra con la frecuencia positiva, y en otro archivo del mismo tipo las palabras con la frecuencia negativa.
- Paso 3: Ordenar las palabras de cada archivo desde la que tiene mayor cantidad de frecuencia, a la menor.
- Paso 4: Dentro de RStudio se importa el set de datos con el que se va a graficar la nube de palabras.
- Paso 5: En RStudio se ejecuta el siguiente script:

```
# Instalar paquetes necesarios
install.packages("tm")  # for text mining
install.packages("SnowballC") # for text stemming
install.packages("wordcloud") # word-cloud generator
install.packages("RColorBrewer") # color palettes
# Cargar paquetes instalados
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")

d <- data.frame("Nombre del set de datos importado en el paso 4")
head(d, 10)
#set.seed(1234)
wordcloud(words = d$word, freq = d$freq, random.order=FALSE,
        colors=brewer.pal(18, "Paired"))
```

- Paso 6: Identificar la nube de palabras generada en la ventana *plots* de RStudio.

Estos 6 pasos se repiten para los resultados de cada uno de los tres algoritmos finales, y para cada clasificación, positivos y negativos.

Las nubes de palabras creadas se presentan a continuación.

4.4.3.1 Nube de palabras de todo el set de datos.

En esta nube de palabas se pueden apreciar las palabras previo a la clasificación, se usa la base de datos de prueba para la representación.

Ilustración 11 - Nube de palabras sin clasificar

4.4.3.2 Nube de palabras del set de datos obtenido con SVM

Las palabras que más usan los usuarios para referirse a eventos de tráfico positivos según la base de datos clasificada con SVM son: vehicular, circulación y flujo. Usan palabras alusivas a detección del evento como visualiza y presenta, más no se encuentras palabras que indiquen lugares específicos con tanta frecuencia.

Ilustración 12- Nube de palabras SVM Positivas

En cuanto a las palabras que más usan para describir eventos de tráfico negativos tenemos: vehicular, flujo, tránsito, alto, congestión, lento, y varias palabras que especifican lugares como túnel, ruta, puente, y otras que indican sentidos o direcciones como oriente y norte.

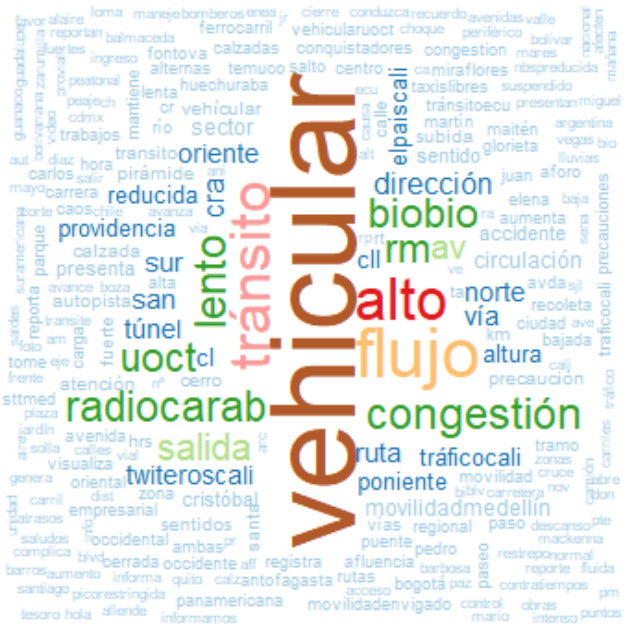


Ilustración 13 - Nube de palabras SVM Negativas

4.4.3.3 Nube de palabras del set de datos obtenido con Naïve Bayes

La base de datos clasificada con Naive Bayes es bastante parecida con respecto a SVM, las palabas con mayor frecuencia para eventos positivos son: vehicular, circulación, flujo y tránsito. Tampoco se encuentran palabras alusivas a direcciones o lugares específicos.

Ilustración 14- Nube de palabras Naïve Bayes Positivas

De igual manera para los eventos negativos existe bastante concordancia en las plabras obtenidas con el algoritmo anterior, las palabras que más se usan según la base clasificada son: vehicular, flujo, alto, tránsito, congestión y lento.

Ilustración 15 - Nube de palabras Naïve Bayes Negativas

4.4.3.4 Nube de palabras del set de datos obtenido con C4.5

En cuanto a la detección de eventos positivos con C4.5 se identifican las mismas palabas de los dos algoritmos , sin embargo, entran varias palabras no alusivas al tema tránsito, que distorsionan aunque de manera mínima, los resultados presentados por el algoritmo ganador.

Ilustración 16 - Nube de palabras C4.5 Positivas

En los eventos de tráfico negativo se presenta la mayor variación de palabras no referentes a tránsito vehicular, aunque estas no se repitan con tanta frecuencia y no sean claramente apreciables en la nube de palabras.

Ilustración 17 - Nube de palabras C4.5 Negativas

4.5 Conclusiones

De acuerdo a la experimentación realizada el mejor algoritmo para minería de textos es SupportVector Machines (SVM) con su configuración Punto.
Con una calificación de 83.45 según la matriz propuesta, se determina que es el mejor algoritmo clasificador para minería de textos. El segundo algoritmo es el de Naïve Bayes con una calificación de 78.66, que difiere del ganador en mayor medida por su grado de fuerza de concordancia determinado por el coeficiente cohen kappa.

En cuanto a la frecuencia de palabras es claramente apreciable que en las clasificaciones positivas figuran las palabras circulación y flujo con una mayor cantidad de repeticiones, es decir las personas al twittear eventos de tráfico positivos se refieren al hecho usando estas palabras.
Sin embargo, de acuerdo a los números, las personas twittean acerca del tráfico más para referirse en forma negativa,en las nubes de palabras se ven las palabras lento, congestión, alto, tránsito en mayor tamaño, representando así el sentimiento negativo de conducir en eventos de tráfico desfavorables.

Conclusiones

En este trabajo de investigación se ha profundizado en conocer y experimentar con los diferentes algoritmos para minería de textos que son mayormente usados en casos de estudio similares, probarlos y entender el impacto que tienen en los datos de estudio, generar información de rendimiento de cada uno y poder decidir cuál es el mejor.

Para identificar los algoritmos que se usaron en la investigación se realizó una revisión sistemática de la literatura con aproximadamente noventa artículos científicos.En los cuales se contabilizó los algoritmos más usados en investigaciones dentro del mismo campo de estudio y afines. Llegando a determinar que los algoritmos eran Support Vector Machines, Naïve Bayes y C4.5 en el 80% de los casos, así que se consideraron únicamente los tres.

Trabajar con datos no estructurados como son las publicaciones de los usuarios en un medio social como Twitter, donde se usa lenguaje natural, y se tienen limitaciones para expresarse es un gran desafío; una de las complicaciones para generar una base de datos robusta, que permita trabajar y entender el modelo para el caso de estudio, fue que en nuestra ciudad no existe la cultura de emitir tweets acerca de eventos de tráfico en un volumen significativo, por lo tanto se tuvo que ampliar el periodo de tiempo para la recolección de tweets, y así ir creando una base consistente.

Otro de los retos fue crear una base de entrenamiento que permita al sistema clasificador aprender a realizarlo y dar valores tan buenos de rendimiento como los obtenidos, se necesitó clasificadores humanos que aporten a este activo importante.

Una vez superada la fase de disponer del suministro de datos para la experimentación, y de conocer a detalle cada uno de los algoritmos, se inició el proceso de creación de una matriz que permita registrar para cada algoritmo, los resultados de precisión en la clasificación, y dos coeficientes de confianza que validen el rendimiento del mismo.De esta manera se estableció un instrumento de evaluación innovador dentro del campo de estudio.

Cuando se realizó la experimentación se usó la mejor configuración que permite la herramienta Rapidminer para cada algoritmo y se eligió como ganador al algoritmo Support Vector Machines con su configuración Punto. Además de la experimentación en análisis de sentimiento, se usó las bases de datos de textos clasificadas en positivas y negativas, para representar gráficamente la frecuencia de las palabras más usadas por los usuarios para describir los eventos de tráfico, y las posibles relaciones entre ellas.

El impacto que tienen los algoritmos de minería de textos sobre datos de medios sociales como Twitter en el campo del tráfico vehicular, y sobre todo poder identificar cuál es el mejor para hacerlo, es muy importante, ya que permite que futuras investigaciones tomen un rumbo mucho más seguro de por donde comenzar, y poder enfocar los recursos en detalles más importantes.

Esta investigación es un paso más que permite disponer de un algoritmo bastante preciso para implementar un sistema mucho más complejo de detección, prevención e incluso predicción de eventos de tráfico que faciliten la gestión de las autoridades y apoyen la toma de decisiones en cuanto a un tema que cada vez se vuelve más crítico en urbes en desarrollo como Cuenca.

Bibliografía

Arias, M. B. (2016). Minería de texto en medios sociales.Caso de estudio del proyecto tranvia de Cuenca. Cuenca, Azuay, Ecuador.

Cao, J., Zeng, K., Wang, H., Cheng, J., Qiao, F., Wen, D., & Gao, Y. (2014). Web-Based Traffic Sentiment Analysis:Methods and Applications. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, 844-853.

D'Andrea, E., Ducange, P., Lazzer, B., & Marcelloni, F. (2015). Real-Time Detection of Traffic From Twitter Stream Analysis. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*.

Elder, J., Miner, G., Nibet, B., Delen, D., Fast, A., & Hill, T. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data.*Elsevier Inc.

Gallardo, J. A. (2009). Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM).

IATA. (2017). *One World - Nations Online*. Obtenido de http://www.nationsonline.org/oneworld/IATA_Codes/IATA_Code_C.htm

Ing. Corso, C. L. (2009). Aplicación de algoritmos de clasificación supervisada usando Weka. Argentina.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 159-174.

Li, N., & Wu, D. (2010). Using text mining and sentiment analysis for online forums hotspot detection. *Decision Support Systems*, 354–368.

Montesinos García, L. (Agosto de 2014). ANÁLISIS DE SENTIMIENTOS Y PREDICCIÓN DE EVENTOS EN TWITTER. Santiago de Chile, Chile.

Purohit, H., Hampton, A., Shalin, V., Sheth, A., Flach, J., & Bhatt, S. (2013). What kind of #conversation is Twitter? Mining #psycholinguistic cues for emergency coordination. *Computers in Human Behavior*, 2438–2447.

Santana, P., Costaguta, R., & Missio, D. (2014). Aplicación de Algoritmos de Clasificación de Minería de Textos para el Reconocimiento de Habilidades de E-tutores Colaborativos. *INTELIGENCIA ARTIFICIAL 53*, 57-67.

Suca, C., Córdova, A., Condori, A., Cayra, J., & Sulla, J. (2016). COMPARACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA LA PREDICCIÓN DE CASOS DE OBESIDAD INFANTIL.

Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., . . . He, X. (2013). Interpreting the Public Sentiment Variations on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, 1158 - 1170.

Anexos

Los anexos que se incluyen en el documento son los siguientes:

- Anexo 1: Aprobación de protocolo del Trabajo de Titulación.
- Anexo 2: Aprobación de solicitud de prórroga para entrega de Trabajo de Titulación.
- Anexo 3: Acta de Sustentación de Protocolo/Denuncia del Trabajo de Titulación.
- Anexo 4: Rúbrica para la evaluación del protocolo del Trabajo de Titulación.
- Anexo 5: Solicitud de aprobación del protocolo del Trabajo de Titulación.
- Anexo 6: Diseño del Trabajo de Titulación.

Doctora Jenny Ríos Coello, Secretaria de la Facultad de Ciencias de la Administración de la Universidad del Azuay

C E R T I F I C A:

Que, el Consejo de Facultad en sesión del 29 de mayo de 2017, conoció la petición del estudiante **JOSÉ ANDRÉS RAMÍREZ FANTONI** con código **60856**, que presenta el diseño de su trabajo de titulación denominado: "***IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERÍA DE TEXTOS EN LOS MENSAJES DEL MEDIO SOCIAL TWITTER***", presentado previa a la obtención del título de Ingeniero de Sistemas y Telemática.- El Consejo de Facultad acogió el informe de la Junta Académica de Ingeniería de Sistemas y Telemática y resolvió aprobar el diseño. Designa como **Directora a la ingeniera María Inés Acosta** omo miembros del Tribunal Examinador a los ingenieros Francisco Salgado Arteaga, Ph.D. y Marcos Orellana Cordero.- En esta misma sesión el Consejo de Facultad fija como plazo para la entrega del trabajo de titulación, seis meses contados desde la fecha de su aprobación, esto es hasta el **29 de noviembre de 2017,** debiendo el Director presentar a la Junta Académica, dos informes bimensuales del desarrollo del trabajo de titulación.

Cuenca, mayo 30 de 2017

Dra. Jenny Ríos Coello
Secretaria de la Facultad de
Ciencias de la Administración

Decano de la Facultad de Ciencias de la Administración.- Cuenca, 22 de noviembre de 2017.- Con autorización amplia y suficiente concedida por el Consejo de Facultad en sesión del 25 de febrero de 2016, conoció la petición de los estudiantes **JOSE ANDRES RAMIREZ FANTONI** con código 60856, quien solicita prórroga para la presentación del trabajo de titulación denominado: "**IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERIA DE TEXTOS EN LOS MENSAJES DEL MEDIO SOCIAL TWITTER**", previo a la obtención del título de Ingeniero de Sistemas y Telemática, cuyo plazo de presentación vence el 29 de noviembre de 2017, en apego al Reglamento de Régimen Académico y la normativa Institucional, *resuelve aprobar la solicitud y conceder una prórroga de seis meses, esto es* ***hasta el 29 de mayo de 2018.***

Ing. Oswaldo Merchán Manzano
Decano de la Facultad de
Ciencias de la Administración

rcr.-

Cuenca, 10 de mayo del 2017

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De mi consideración,

Yo, **María Inés Acosta Urigüen** informo que he revisado el protocolo de trabajo de titulación previo a la obtención del título de Ingeniero en Sistemas y Telemática, denominado **"IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERÍA DE TEXTOS EN LOS MENSAJES DEL MEDIO SOCIAL TWITTER"**, realizado por el estudiante **José Andrés Ramírez Fantoni**, con código estudiantil 60856, protocolo que, a mi criterio, cumple con los lineamientos y requerimientos establecidos por la carrera.

Por lo expuesto, me permito sugerir que sea considerado para la revisión y sustentación del mismo,

Sin otro particular, suscribo.

Atentamente

Ing. María Inés Acosta

Oficio Nro. 068-2017-DIST-UDA

Cuenca, 12 de mayo de 2017

Señor Ingeniero
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
Presente.-

De nuestras consideraciones:

La Junta Académica de la Escuela de Ingeniería de Sistemas y Telemática, reunida el día 12 de mayo del 2017, recibió el proyecto de tesis titulado "Impacto de la aplicación de algoritmos de minería de textos en los mensajes del medio social Twitter", presentado por José Andrés Ramírez Fanoni estudiante de la Escuela de Ingeniería de Sistemas y Telemática, y revisado por la Ing. María Inés Acosta, previo a la obtención del título de Ingeniero de Sistemas y Telemática.

Por lo expuesto, y de conformidad con el Reglamento de Graduación de la Facultad, recomendamos como director y responsable de aplicar cualquier modificación al diseño del trabajo de graduación posterior a la Ing. María Inés Acosta y como miembros del Tribunal a Francisco Salgado Ph.D. e Ing. Marcos Orellana.

Atentamente,

Ing. Marcos Orellana Cordero
Cordinador Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay

C O N V O C A T O R I A

Por disposición de la Junta Académica de **Ingeniería de Sistemas y Telemática**, se convoca a los Miembros del Tribunal Examinador, a la sustentación del Protocolo del Trabajo de Titulación: **"IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERÍA DE TEXTOS EN LOS MENSAJES DEL MEDIO SOCIAL TWITTER"**, presentado por el estudiante **José Andrés Ramírez Fantoni**, previa a la obtención del grado de **Ingeniero en Sistemas y Telemática**, para el día **MARTES 16 DE MAYO DE 2017 A LAS 12h20.** La sustentación se realizará en el **laboratorio del IERSE.**

Cuenca, 11 de mayo de 2017

Dra. Jenny Ríos Coello
Secretaria de la Facultad

Ing. María Inés Acosta Uriguen X 103

Ing. Marcos Orellana Cordero

Dr. Francisco Salgado Arteaga

mjmr/

ACTA
SUSTENTACIÓN DE PROTOCOLO/DENUNCIA DEL TRABAJO DE TITULACIÓN

1.1 **Nombre del estudiante: José Andrés Ramírez Fantoni**
1.2 **Director sugerido**: Ing. María Inés Acosta Uriguen|
1.3 **Codirector (opcional)**:
1.4 **Tribunal:** Ing. Marcos Orellana Cordero/ Dr. Francisco Salgado Arteaga
1.5 **Título propuesto: "IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERÍA DE TEXTOS EN LOS MENSAJES DEL MEDIO SOCIAL TWITTER"**
1.6 **Resolución**:

1.6.1 Aceptado sin modificaciones _____

1.6.2 Aceptado con las siguientes modificaciones:

_____

_____

1.6.3 Responsable de dar seguimiento a las modificaciones: Ing. María Inés Acosta Uriguen|

1.6.4 No aceptado
- Justificación:

_____
_____
_____

Tribunal

Ing. María Inés Acosta Uriguen — Ing. Marcos Orellana Cordero — Dr. Francisco Salgado Arteaga

.....
Sr. José Andrés Ramírez Fantoni

Dra. Jenny Ríos Coello
Secretario de Facultad

Fecha de sustentación: día **MARTES 16 DE MAYO DE 2017 A LAS 12h20**

1

14.¿Concuerdan con el objetivo general?	✓			
15.¿Son comprobables cualitativa o cuantitativamente?	✓			
Metodología				
16.¿Se encuentran disponibles los datos y materiales mencionados?	✓			
17.¿Las actividades se presentan siguiendo una secuencia lógica?	✓			
18.¿Las actividades permitirán la consecución de los objetivos específicos planteados?	✓			
19.¿Los datos, materiales y actividades mencionadas son adecuados para resolver el problema formulado?	✓			
Resultados esperados				
20.¿Son relevantes para resolver o contribuir con el problema formulado?	✓			
21.¿Concuerdan con los objetivos específicos?	✓			
22.¿Se detalla la forma de presentación de los resultados?	✓			
23.¿Los resultados esperados son consecuencia, en todos los casos, de las actividades mencionadas?	✓			
Supuestos y riesgos				
24.¿Se mencionan los supuestos y riesgos más relevantes?	✓			
25.¿Es conveniente llevar a cabo el trabajo dado los supuestos y riesgos mencionados?	✓			
Presupuesto				
26.¿El presupuesto es razonable?	✓			
27.¿Se consideran los rubros más relevantes?	✓			
Cronograma				
28.¿Los plazos para las actividades son realistas?	✓			
Referencias				
29.¿Se siguen las recomendaciones de normas internacionales para citar?	✓			
Expresión escrita				
30.¿La redacción es clara y fácilmente comprensible?	✓			
31.¿El texto se encuentra libre de faltas ortográficas?	✓			

(*) Breve justificación, explicación o recomendación.

- Opcional cuando cumple totalmente,
- Obligatorio cuando cumple parcialmente y NO cumple.

.....

.....

.....

Ing. María Inés Acosta Uriguen Ing. Marcos Orellana Cordero Dr. Francisco Salgado Arteaga

UNIVERSIDAD DEL AZUAY

Escuela Sistemas y Telemática	Oficio Estudiante: Solicitud aprobación de Protocolo de Trabajo de Titulación	IST-RE-EST-02 Versión 01 04/04/2017 Página 1 de 1

Lugar de Almacenamiento	Retención	Disposición Final
F: Archivo Secretaría de la Facultad	5 años	Almacenar en archivo pasivo de la Facultad

Cuenca, 10 de mayo del 2017

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De mi consideración,

Estimado Señor Decano, yo **José Andrés Ramírez Fantoni** con C.I. **010535303-1**, código estudiantil 60856; estudiante de la Carrera de Sistemas y Telemática, solicito muy comedidamente a usted y por su intermedio al Consejo de Facultad, la aprobación del protocolo de trabajo de titulación con el tema **"IMPACTO DE LA APLICACIÓN DE ALGORITMOS DE MINERÍA DE TEXTOS EN LOS MENSAJES DEL MEDIO SOCIAL TWITTER"** previo a la obtención del título de Ingeniero en Sistemas y Telemática para lo cual adjunto la documentación respectiva.

Por la favorable acogida que brinde a la presente, anticipo mi agradecimiento.

Atentamente:

José Ramírez Fantoni
Estudiante de la Carrera de Sistemas y Telemática

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500
Nº 0794837

UNIVERSIDAD DEL AZUAY

DOCTORA JENNY RIOS COELLO, SECRETARIA DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACION DE LA UNIVERSIDAD DEL AZUAY

C E R T I F I C A:

Que, el señor **RAMIREZ FANTONI JOSE ANDRES** con código **60856,** alumno de la escuela de **INGENIERIA DE SISTEMAS Y TELEMATICA,** tiene aprobado más del 80% de los créditos de su malla de estudios.

Que, al señor **RAMIREZ FANTONI JOSE ANDRES** le falta aprobar las siguientes asignaturas para finalizar sus estudios:

~~PRODUCCIÓN II~~

CALIDAD DE SOFTWARE

INGENIERÍA DE SOFTWARE II

METODOLOGIA DE LA INVESTIGACION

~~MÉTODOS NUMÉRICOS~~

~~PASANTIAS~~

Cuenca, 29 de marzo de 2017

Derecho No. **001-010-000116614**
mjmr.-

UNIVERSIDAD DEL AZUAY
FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN

Edición autorizada de 10.000 ejemplares
Del 788.501 al 798.500
Nº 0791505

UNIVERSIDAD DEL AZUAY

Universidad del Azuay

Facultad de Ciencias de la Administración

Escuela de Ingeniería de Sistemas y Telemática

Diseño de Trabajo de Graduación

1. **Datos Generales**

1.1. Nombre del estudiante: José Andrés Ramírez Fantoni

1.1.1. Código: 60856

1.1.2. Contacto: 4082545–0995238008

jose.andres.ramirez.fantoni@gmail.com

1.2. Director sugerido: María Inés Acosta Urigüen, Magíster en Administración de Empresas, Ingeniera de Sistemas.

1.2.1. Contacto: macosta@uazuay.edu.ec

1.3. Co-director sugerido:

1.4. Asesor metodológico: Ing. Francisco Salgado Arteaga

1.5. Tribunal Designado:

1.6. Aprobación: Junta Académica

Consejo de Facultad

1.7. Línea de investigación de la carrera: Sistemas de Bases de Datos

1.7.1. Código UNESCO: 1203 Informática de computadores

1203.17 Informática

1.7.2. Tipo de trabajo: Investigación Aplicada

1.8. Área de estudio: Innovación Tecnológica

1.9. Título propuesto: Impacto de la aplicación de algoritmos de minería de textos en los mensajes del medio social Twitter

1.10. Subtitulo propuesto: Para los eventos del tráfico vehicular de la ciudad de Cuenca

2. Contenido

2.1. Motivación de la investigación:

Los habitantes de las medianas y grandes ciudades diariamente deben vivir los problemas de la congestión vehicular a causa del continuo crecimiento del parque automotor. Las soluciones que se han planteado hasta el momento has sido fruto de la percepción de las autoridades de tránsito y las continuas quejas de los usuarios. Eventos como: cierre de calles, calles traficadas y no traficadas, colisiones, atropellamientos, derrumbes, eventos organizados, entre otras, son motivo de la generación de comentarios en redes sociales. Es necesario tomar la opinión de los usuarios y de las entidades de control para tratarlas y determinar patrones de comportamiento que retroalimenten a los usuarios en tiempo real, de modo que se tomen las previsiones adecuadas y se agilite el tránsito.

Para desarrollar un sistema como el propuesto es necesario establecer el mejor algoritmo para análisis de los datos que se generan en los medios sociales, en este caso de estudio se usará Twitter que ofrece datos catalogados como no estructurados y que necesitan ser tratados considerando la singularidad de las palabras. Al obtener un algoritmo específico que permita la obtención de los datos necesarios para el caso de estudio se puede continuar con el proceso de selección de herramientas y creación de modelos y técnicas que permitan encontrar los resultados esperados.

2.2. Problemática:

Las redes sociales han crecido geométricamente en los últimos años, por lo que este medio se ha constituido en un verdadero canal de comunicación activa. Una de sus principales características ha sido la portabilidad de sus aplicaciones a dispositivos como los smartphones y las tablets; fáciles de usar y con una interacción en tiempo real. Frecuentemente, las personas utilizan una red social para emitir opiniones sobre eventos que suceden en su contexto o simplemente lo hacen sobre un tópico determinado. Un tema de importancia en la vida de las personas del área urbana es el transporte; los habitantes de las medianas y

grandes ciudades diariamente deben vivir los problemas de la congestión vehicular.

Para tratar los datos acerca del tráfico vehicular que se obtienen de los medios sociales mediante minería de textos es necesario evaluar y determinar el algoritmo que ofrece mejores resultados. Se necesita correlacionar la colección de datos que se obtienen para el estudio con varios algoritmos y clasificarlos según la información de salida que se obtenga. Una vez obtenida la clasificación se podrá determinar el mejor algoritmo y su impacto en el ámbito del tráfico vehicular.

2.3. Pregunta de investigación:

¿Cuáles son los algoritmos para minería de texto que se aplican al problema de tráfico vehicular?
¿Cómo están organizados los datos generados por el medio social Twitter?
¿Cuáles son los patrones de texto en los comentarios emitidos por los usuarios de la red social en cuanto al tráfico vehicular?
¿Cuál es (son) el (los) mejor(es) algoritmo(s) para minería de texto que se aplica al problema de tráfico vehicular?

2.4. Resumen:

El presente trabajo de investigación profundizará las técnicas de minería de textos y el análisis de sentimiento en los comentarios de la red social Twitter para los eventos del tráfico vehicular de la ciudad. Se abordará el problema de los datos generados por la red social Twitter, catalogados como datos no estructurados, lo que en primer lugar implica un proceso de selección de datos de cuentas y hashtag relacionados al área de estudio, a fin de filtrar la información relevante. En un siguiente paso se procesará los datos con la extracción, almacenamiento y depuración de los elementos texto. Se realizarán pruebas de algoritmos sobre los datos previamente obtenidos y se realizará una

selección según los resultados obtenidos por cada uno. Se aplicará los algoritmos probados y seleccionados previamente para el procesamiento a través de modelos de procesos, y se expondrá su impacto en el tratamiento de los datos y posterior obtención de información.

2.5. Indagación Exploratoria:

Las redes sociales se han convertido en un fenómeno social que replantearon fenómenos, procesos, etc. de la vida diaria en sus formas básicas de comunicación, interacción y producción de saber, mediado por tecnologías (Fainholc, 2015).

Con el significativo crecimiento de los mensajes generados por el usuario, Twitter se ha convertido en una red social donde millones de usuarios pueden intercambiar su opinión. El análisis de sentimientos en los datos de Twitter ha proporcionado una manera económica y efectiva de exponer la opinión pública a tiempo, lo cual es crítico para la toma de decisiones en varios dominios. (Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., ... & He, X., 2014). Sin embargo, solo unos cuantos estudios se enfocan en el campo de la transportación vehicular. En la actualidad se han realizado trabajos que vinculan el tráfico con el análisis de sentimiento por medio de lo que se denomina como el análisis de sentimiento para el tráfico (TSA), esta herramienta se ha transformado como una nueva adopción (Cao et al., 2014).

El análisis de sentimiento de texto también se denomina computación de polaridad emocional; esta nueva área se ha convertido en una frontera floreciente de la minería de textos, por la aplicación de algoritmos para analizar la polaridad emocional contenida en un texto (Li & Wu, 2010). La eficacia de estos métodos es demostrada a través de la obtención de conocimiento a partir de la información expresada por usuarios de la red social Twitter con respecto a los problemas de tráfico vehicular de la ciudad.

La minería de textos son un conjunto de métodos que, mediante los datos emitidos por los usuarios, obtiene información y encuentra los tópicos más hablados, y permite categorizar a los usuarios según la información aportada. En la minería de textos cada opinión es tratada como un documento particular. La minería de textos según Arias es aquella que ofrece con sistemas automáticos una solución, sustituyendo o completando el trabajo de las personas, sin que importe la cantidad de texto; encuentra información desconocida hasta el inicio del proceso, que puede determinar patrones que de otro modo serían complejos de descubrir. Así también, la minería de textos consiste en extraer la información útil de documentos no estructurados e identificar patrones interesantes de conocimiento. La minería de textos es una extensión de la minería de datos; muchas de las técnicas utilizadas en la minería de datos pueden ser aplicadas a la minería de textos. En resumen, la minería de textos se encarga de examinar una colección de datos no estructurados, cuyo fin es la identificación de patrones que generen información y posteriormente conocimiento, sin importar que esta colección sea de gran tamaño. Cuando la minería de textos se centra en fuentes que provienen de redes sociales, se suele hablar también de la minería de medios sociales, que corresponde al proceso de representar, analizar y extraer patrones significativos desde los datos de medios sociales, resultantes de la interacción social.

Según Estévez-Velarde es importante analizar las diferentes opciones de algoritmos que existen para el minado de texto, específicamente para el minado de opinión o análisis de sentimiento ya que no todas las tecnologías existentes son las ideales para obtener resultados claros.

2.6. Objetivo General

Identificar los algoritmos relevantes que determinen la generación de patrones relacionados a un texto emitido por la red social Twitter con respecto al tráfico vehicular en la ciudad de Cuenca.

2.7. Objetivos Específicos

- Sistematizar información acerca de técnicas y algoritmos de minería de textos, y el análisis de sentimiento en los comentarios del medio social Twitter.
- Establecer cinco algoritmos que se evaluarán en el caso de estudio.
- Diseñar una matriz de evaluación de algoritmos para la minería de textos.
- Encontrar la frecuencia de palabras y la asociación entre ellas a fin de ubicar patrones en colecciones de datos no estructurados con cada algoritmo previamente establecido.
- Registrar los valores obtenidos con cada algoritmo en la matriz de evaluación de algoritmos para la minería de textos.
- Analizar resultados obtenidos de la aplicación de algoritmos de minería de textos a comentarios de la red social Twitter.

2.8. Metodología:

Este proyecto tendrá una metodología orientada a la investigación bibliográfica para determinar el estado de avance en el que se encuentra actualmente la minería de texto y sus algoritmos y cómo se ha venido desarrollando a través de los años.

2.9. Alcances y resultados esperados:

En el proyecto se espera que el proceso de indagación, selección y evaluación sirva para determinar el mejor algoritmo de minería de textos aplicable al análisis de datos obtenidos del medio social Twitter.

Como resultado final se obtendrá la Matriz de Evaluación de Algoritmos de minería de textos y el análisis de cuál es el impacto producido de su aplicación en los datos no estructurados obtenidos de Twitter con respecto al tráfico vehicular en la ciudad de Cuenca.

2.10. Supuestos y riesgos:

Supuestos	Riesgos	Probabilidad	Soluciones
La información utilizada como fundamento para la investigación es de calidad.	La información utilizada no cumple con el nivel de calidad esperado.	Media	Realizar una revisión sistemática de la bibliografía para garantizar que la información es pertinente y de calidad.
Los datos no estructurados obtenidos del medio social Twitter son suficientes y permiten la correcta aplicación de los algoritmos a evaluar.	No hay suficientes datos relacionados al tráfico vehicular en la ciudad de Cuenca en el medio social Twitter, y no se pueden evaluar los algoritmos.	Media	Realizar una búsqueda detallada de las cuentas, hashtag, y palabras claves que usan los usuarios de Twitter para reportar novedades de tráfico vehicular, para garantizar que exista la suficiente información.
Se cuenta con suficientes algoritmos disponibles y documentados para realizar la evaluación.	No se cuenta con los algoritmos suficientes para realizar la evaluación.	Media	Durante el proceso de indagación exploratoria se seleccionan los algoritmos suficientes para realizar la evaluación.
No existen investigaciones similares con respecto a evaluación de algoritmos para minería de textos aplicados a datos sobre tráfico vehicular	Investigaciones similares ya realizadas, sobre comparación de algoritmos para tratar datos sobre tráfico vehicular.	Baja	Buscar un enfoque diferente de la investigación; realizar la evaluación de algoritmos para otra temática.

2.11. Presupuesto:

Rubro – Denominación	Costo USD	Justificación
Conexión a Internet	$180.00	Para acceder a la información bibliográfica disponible en la red, y gestionar revisiones con los interesados en el proyecto.
Suministros y Papelería	$100.00	Para entregar documentos impresos durante y al final del desarrollo del trabajo de titulación.
Equipos	$0	No se considera la compra de equipos ya que son de propiedad del desarrollador del proyecto.
Gastos Varios	$150.00	Se consideran como varios los gastos de transporte o imprevistos que se presenten durante la ejecución del proyecto.
Total	$430.00	

2.12. Financiamiento:

El financiamiento para la realización de este proyecto de investigación será propio, y servirá para sustentar todos los gastos involucrados.

2.13. Esquema tentativo:

Portada

Dedicatoria

Agradecimientos

Abstract

Índice de imágenes

Índice de tablas

Índice de contenidos

UNIVERSIDAD DEL AZUAY

2.14. Cronograma:

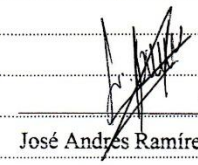
Objetivo Específico	Actividad	Resultado Esperado	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
			15	22	29	5	12	19	26	3	10	17	24	31	7	14	21	28	4	11	18	25	2
			MAYO			JUNIO				JULIO					AGOSTO				SEPTIEMBRE				OCTUBRE
Sistematizar información acerca de técnicas y algoritmos de minería de textos, y el análisis de sentimiento en los comentarios del medio social Twitter.	Revisión sistemática de la bibliografía acerca de minería de textos, algoritmos y análisis de sentimiento.	Obtener toda la información necesaria para el desarrollo de la investigación, que esta sea de calidad y pertinente a la temática.																					
Establecer los algoritmos que se evaluarán en el caso de estudio.	Obtener la mayoría de algoritmos para minería de textos existentes y aplicables a la investigación.	Seleccionar 5 algoritmos para la evaluación y análisis de impacto.																					
	Mediante asesoramiento técnico seleccionar 5 algoritmos con mayor documentación existente y que tengan aplicabilidad en la temática.																						
Diseñar una matriz de evaluación de algoritmos para la minería de textos.	Crear una matriz que sintetice los atributos principales y medibles de los algoritmos a evaluar.	Determinar la herramienta que permitirá evaluar y comparar los algoritmos de la investigación.																					

Encontrar la frecuencia de palabras y la asociación entre ellas a fin de ubicar patrones en colecciones de datos no estructurados con cada algoritmo previamente establecido.	Realizar las pruebas de los algoritmos con intervención de los técnicos del proyecto.	Probar los algoritmos seleccionados y obtener los resultados.	
Registrar los valores obtenidos con cada algoritmo en la matriz de evaluación de algoritmos para la minería de textos.	Llenar la matriz de evaluación de algoritmos con los datos obtenidos de las pruebas.	Disponer de los datos obtenidos de las pruebas en la matriz de evaluación de algoritmos para su interpretación.	
Analizar resultados obtenidos de la aplicación de algoritmos de minería de textos a comentarios de la red social Twitter.	Analizar los datos obtenidos según los parámetros de la matriz de evaluación de algoritmos	Concluir con el análisis de los datos obtenidos y su relación con el proyecto de congestión de tráfico vehicular en la ciudad de Cuenca.	
	Elegir el mejor algoritmo según la aplicabilidad en el proyecto		
Redactar documentación del proyecto.	Escribir el trabajo de graduación.	Terminar redacción del trabajo de graduación.	

2.15. Referencias:

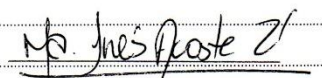
- Fainholc, B. (2015). Un análisis contemporáneo del Twitter. *Revista de Educación a Distancia*, (26).
- Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., ... & He, X. (2014). Interpreting the public sentiment variations on twitter. ieee transactions on knowledge and data engineering, 26(5), 1158-1170.
- Cao, J., Zeng, K., Wang, H., Cheng, J., Qiao, F., Wen, D., & Gao, Y. (2014). Web-Based Traffic Sentiment Analysis: Methods and Applications. Intelligent Transportation Systems, IEEE Transactions on, 15(2), 844–853. https://doi.org/10.1109/TITS.2013.2291241
- Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems, 48(2), 354–368. https://doi.org/10.1016/j.dss.2009.09.003
- Zhañay, A., & Belén, M. (2016). *Minería de texto en medios sociales. Caso de estudio del proyecto Tranvía de Cuenca* (Bachelor's thesis, Universidad del Azuay).
- Estévez-Velarde, S., & Cruz, Y. A. (2013). Evaluación de algoritmos de clasificación supervisada para el minado de opinión en Twitter.

2.16. Firma de responsabilidad:

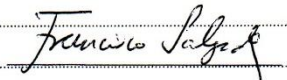


José Andrés Ramírez Fantoni

2.17. Firmas de responsabilidad:



MSc. María Inés Acosta



PhD. Francisco Salgado

2.18. Fecha de entrega

10 de mayo de 2017