Universidad de Lima

Facultad de Ingeniería y Arquitectura

Carrera de Ingeniería de Sistemas

COMPARACIÓN DE MÉTODOS PARA CLASIFICAR COMENTARIOS DE LUGARES TURÍSTICOS POR MEDIO DE ANÁLISIS DE SENTIMIENTO

Tesis para optar el Título Profesional de Ingeniero de Sistemas

Luis Guillermo Herrera Sarmiento

Código 20141885

Asesor

Juan Manuel Gutierrez Cardenas

Lima – Perú

Agosto de 2020

Comparación de métodos para clasificar comentarios de lugares turísticos por medio de análisis de sentimiento

Luis Guillermo Herrera Sarmiento
20141885@aloe.ulima.edu.pe
Universidad de Lima

Resumen: Hoy en día los turistas luego de visitar algún destino plasman sus experiencias como opiniones en diversas fuentes turísticas, redes sociales y/o sitios web turísticos, siendo esta información valiosa para empresas turísticas o relacionadas a ello, para identificar en qué lugar se puede enriquecer la experiencia de la visita (una oportunidad de mejora). Asimismo, promover la atención de los turistas durante la planificación de sus viajes, ya que la información existente puede ser abrumadora. En esta investigación se tomó el sitio web TripAdvisor para adquirir los comentarios acerca los sitios de interés y se realizó la comparación de tres técnicas para la clasificación de estos comentarios: Support Vector Machine (SVM), Naïve Bayes (NB) y Método propuesto basado en SVM y Chi Square como método de selección de características. La técnica híbrida propuesta obtuvo el mejor resultado, seguido de SVM y por último Naïve Bayes cada una con 80.27%, 78.53% y 76.91% de precisión respectivamente. Se concluye que es factible realizar la clasificación automática y obtener los lugares con mayor proporción de reseñas negativas.

Palabras Clave: text mining, sentiment analysis, support vector machine, Naïve Bayes, Online reviews, Text classification, Chi Square, feature selection

Abstract: Nowadays tourists after visiting a destination reflect their experiences and emotions as opinions or reviews in some sources either social networks or on tourist websites such as TripAdvisor, this information is valuable for tourism industry to identify which sites has an opportunity to improve also helps to plan a tourist trips for travelers insomuch as the existing information for the site can be overwhelming. In this paper we consider the website TripAdvisor to obtain comments about touristic places and do the comparison of three techniques for the text classification of these comments: Support Vector Machine (SVM), Naïve Bayes (NB) and Proposed Method based on SVM and Chi Square as a method of features selection. The proposed hybrid technique obtained the best result, followed by SVM and finally Naïve Bayes each with con 80.27%, 78.53% and 76.91%of accuracy respectively.In closing it is feasible to perform the automatic classification and obtain the places with the highest proportion of negative reviews.

Keywords:text mining, sentiment analysis, support vector machine, Naïve Bayes, Online reviews, Text classification, Chi Square, feature selection.

1. INTRODUCCIÓN

Se ha desarrollado una tendencia que va en aumento sobre el número de viajeros independiente o backpackers (Chi, Lo, Chu, Y Lin, 2009). El Ministerio de Comercio Exterior y Turismo del Perú (MINCETUR) indica que el crecimiento del turismo en el Perú en el año 2018 fue de 9.6% a comparación del año anterior y el banco Scotiabank proyectó un incremento del 10% para el año 2019 debido a los diversos eventos internacionales que ocurrieron durante ese año. En el Perú una de las industrias más grandes luego de la pesca y la minería es el turismo, puesto que es una de las principales fuentes de la economía peruana ya que asume la historia, cultura y tradición de la nación en diversos bienes y servicios. Por lo que tomando en consideración la importancia de esta industria, se debe analizar toda información buena o mala acerca de los puntos de interés para encontrar puntos de mejora al servicio ofrecido. He, Lin, Wu, Chen, Ku, Y Chen, (2014) mencionan que los turistas que llegan a su destino no encuentran sitios turísticos de su agrado y estos demuestran su inconformidad en diversos portales web, lo que origina que se dé una mala imagen o reputación de los sitios visitado hacia otros turistas (Li y Yang, 2017) y Kuhamanee et al. (2017) dice que el propósito de su investigación es identificar el nivel de agrado de los turistas a partir de sus comentarios en un sitio web turístico y clasificarlo en negativo o positivo acerca de los lugares visitados esto con ayuda del análisis de sentimiento o clasificación de texto. Chi et al. (2009), mencionan que un viaje para un turista es una actividad personal para conseguir objetivos, por lo que sus opiniones deben ser consideradas primordiales para el cumplimiento de estos y a su vez brindarles un buen servicio. Esta investigación está estructurada en seis partes: Introducción (sección en la que se trata sobre la justificación de la investigación), Trabajos Relacionados (apartado que toca los artículos con temas similares a la investigación), Marco Teórico (sección que mencionada las técnicas y fundamentos usados en la experimentación), Metodología (etapas de la investigación), Experimentación, Resultados, y Conclusiones.

2. ESTADO DEL ARTE

En términos generales existen muchas alternativas para sugerir y recomendar lugares. Esto se ve reflejado en las investigaciones recopiladas sobre métodos para la extracción de datos y clasificación de estos, a través del análisis de sentimiento y muestreo mediante API´s de redes sociales y herramientas de web scraping.

2.1 FUENTES Y TÉCNICAS DE RECOLECCION DE DATOS

Lin et al. (2014) mencionan que las redes sociales presentan una gran oportunidad de obtener opiniones de usuarios, tal como lo refleja en el objetivo de su investigación, que fue encontrar y determinar el sentimiento asociado a los temas de interés de una página de Facebook. Esto gracias a la ayuda de las APIs y técnicas de análisis de sentimiento de trabajos previos y concluyen que estos resultados pueden ser aplicados para mejorar el marketing de la corporación. Nguyen Y Kravets (2016) coinciden con la forma de obtener información de la red social y tienen como objetivo identificar los temas que más agradan al usuario por medio del análisis de sentimiento y determinar la cantidad atribuciones positivas e identificar los temas de interés del usuario.

Becerra (2016) hace mención que los sistemas de clasificación de texto, exploración de datos y el desarrollo de las redes sociales han sido de gran relevancia en los últimos años. El enfoque de su investigación es clasificar opiniones en positivas o negativas de los comentarios de los usuarios o turistas. Por lo que desarrolló una aplicación en Twitter y con apoyo de APIs obtuvo la recolección de tweets. Recopiló estos con la etiqueta #oscars y por medio de técnicas de clusterización identificó el sentimiento del comentario. Ellos concluyen que a pesar de que Twitter tiene limitaciones con ayuda de las APIs es suficiente para comprender mejor el mercado o un acontecimiento. Hodeghatta, (2013) menciona otras técnicas de recolección como in-house tool además de filtro UH-filter para obtener solo lo esencial.

Khotimah Y Sarno (2018) para mejorar y proporcionar el mejor servicio de hotel decidieron recopilar los comentarios que sus clientes ofrecen en el portal web Booking.com siendo su fuente de datos lo que garantiza la autenticidad de estos para determinar la satisfacción del cliente o insatisfacción basado en sus comentarios. Además, recopiló estos datos por medio de técnicas de Web Scraping en específico el software WebHarvy que se encarga de recopilar todos los datos necesarios de la página web y con el método PLSA para la clasificación de los comentarios resultando un nivel de 76% de precisión. Al igual que otros autores coincide que el uso de webscraping o crawling data para la recolección de datos (Oliveira,2012), (Prameswari, Surjandari y Laoh,2018), (Li y Yang, 2017)

2.2 ANÁLISIS DE SENTIMIENTO RELACIONADO AL TURISMO

El análisis de sentimiento o también conocido extracción de opinión, análisis de subjetividad o minería de opinión, es un campo que es usado en diversos sectores ya sea en el empresarial, político, turístico entre otros y el uso en cada uno se resume en conocer la opinión o respuesta de parte de la población hacia un producto, lugar, servicio, candidato, entre otros. (Ekawijana y Heryono,2016)

Parikh, Kestar, Dharia, y Gotmare (2018) tiene como objetivo en su investigación la clasificación de los usuarios registrados en el portal web TripAdvisor tomando sus intereses además de otras variables de usuario, fecha, titulo, rating, nombre, lugar; hicieron uso de la técnica K-modes para el desarrollo de clusters sobre lugares de Mumbai indicando que la recomendación se da con gran precisión a pesar de no indicar cuanto es esta. Por otra parte, Li y Yang (2017) para su sistema de minería de datos al igual que los autores anteriores obtuvieron su información de TripAdvisor. Este sistema lo comparo con las técnicas de Support Vector Machine y Regresión Logística, resultando su sistema superior para buzzwords en idioma chino con un 95% de precisión. Gonzalez-Rodriguez, Martinez-Torres, y Toral (2014) monitorean la información que se ofrece en internet sobre sitios turísticos mediante análisis de sentimiento, también haciendo muestreo de datos de un portal web con web scraping y con una técnica basada en una lista de palabras en inglés (AFFINN-11) con lo que concluyen que los usuarios suelen escribir más sobre atracciones y hoteles. Valdivia, Luzón, y Herrera (2017) realiza una comparación de técnicas de aprendizaje supervisado Support Vector Machine, Bi-Grams, y Naive Bayes, resultando los dos primeros superiores mediante la técnica de Validacion cruzada.

Thabtah, Eljinini, Zamzeer, Y Hadi (2009) mencionan en su artículo que desarrollan un sistema de clasificación de texto usando el algoritmo Naïve Bayes basado en el método de selección de características Chi Square. La base de sus comparaciones son la macro F1, la recuperación de macros y las medidas de evaluación de precisión de macros. Sus resultados experimentales comparados con diferentes conjuntos de datos de categorización de texto en árabe proporcionaron evidencia de que la selección de características a menudo aumenta la precisión de la clasificación al eliminar términos poco comunes. Alsaleem (2011) tiene como objetivo en su investigación presentar y comparar los

resultados obtenidos con las colecciones de texto en Periódicos árabe Saudí con las mismas técnicas que los autores previos y tuvo resultado que la técnica de SVM tiene mejor performance que Naïve Bayes sin dar más detalle.

Lin et al. (2014) desarrollaron un sistema que recolecta y clasifica automáticamente la información que existe en internet a través de su caso de estudio de la fan page de Huggies Taiwán. Allí se vio la interacción de la fan page con los usuarios y se clasificó esta mediante el método SVM dando como resultado una precisión de 77% no perfecta, pero considerada buena, por lo que se logró su objetivo que fue darle un valor de sentimiento a cada tema. Akter Y Aziz (2017) indican que haciendo uso de un grupo en una red social; Facebook, como fuente de datos para el análisis de sentimiento mediante dos técnicas, Naive Bayes Machine Learning classifier y Dictionary Based Approach, sin embargo tuvieron problemas con la primera técnica dado que los datos extraídos de una red social son considerados muy ruidosos, es decir, que necesita gran volumen de información para tener la precisión adecuada, mientras que con la segunda técnica se dio un panorama distinto, debido a que esta trabaja mejor con información subjetiva.

Kuhamanee et al. (2017) hicieron un análisis sobre la ciudad de Bangkok con muestreo de twitter con cuatro enfoques (arboles de decisión, support vector machine, naive bayes y redes neuronales) resultando Support Vector Machine y las redes neuronales con mayor precisión superior al 80%

Ye, Zhang, y Law (2009) se basaron en comentarios de internet, dado que consideran que de esta fuente se puede obtener opiniones de lugares específicos, de ese modo hicieron la comparación de técnicas Naive Bayes, SVM y N-gram model en 7 lugares turísticos, resultando los tres con un nivel de precisión superior al 80% , siendo la técnica de SVM la que obtuvo el resultado más alto con un 85% de precisión, tomando en consideración un conjunto de datos con 1191 comentarios y validándolo con la técnica de validación cruzada con un k=17 ,Liu (2012) menciona que usar Dictionary Based Approach, para reconocer el sentimiento de las palabras combinando esta técnica con Wordnet. Esas técnicas fueron usadas por Hu y Liu (2004) con el mismo objetivo resultando un nivel de precisión de 84%.

3. ANTECEDENTES TEÓRICOS

3.1 SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) como técnica de clasificación ha demostrado ser altamente efectivo en la categorización de texto (Joachims,1998). SVM para la clasificación de textos busca un hiperplano representado por un vector $\vec{w}$,que separa los vectores de entrenamiento positivos y negativos de los documentos con un margen máximo para obtener la separación ideal tal como se ve en la Figura 1. (Ye et al., 2009), donde $\vec{d}$ y θ también son vectores

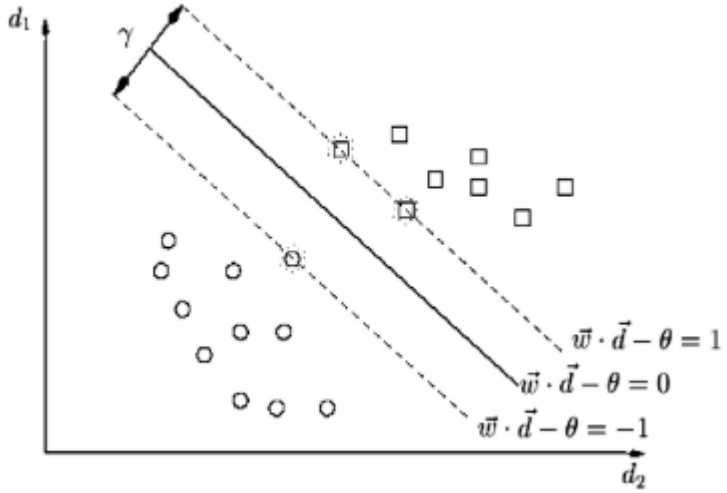

Fig. 1. Máximo margen clasificador SVM
Fuente: Ye et al., (2009)

Donde la mejor separación seria donde el resultado sería 0 y si el resultado da ≥1 se entiende que es una observación de la cluster de datos superior, así como si el resultado fuera ≤-1 es una observación del cluster de datos inferior. Siendo este hiperplano la separación óptima de las características de espacio basado en la teoría de la minimización de riesgo. Primero se comprime el conjunto de datos para dar paso al conjunto de vectores de soporte y se da ganancia de información al usar subconjuntos y las condiciones dadas por los vectores de soporte (Liu, Lv, Liu y Shi, 2010).

Este método es el adecuado para un gran conjunto de datos, especialmente la clasificación de texto. (James Kwok , 2000)

Finalmente, se aplica el clasificador SVM con ganancia de información (IG) como un método de selección de características. Asi mismo, en los experimentos, se elige la frecuencia de las palabras para presentar un resultado en lugar de la cantidad de las palabras para la estimación de probabilidad. (Ye et al., 2009)

3.1.2 KERNEL

La función kernel son funciones matemáticas. Estas son las que le permiten convertir lo que es un problema de clasificación no lineal en uno lineal, pero con un espacio dimensional mayor. De acuerdo con la teoría funcional si satisface la condición Mercer. Por ello el uso del kernel apropiado transforma la clasificación no lineal a una lineal en un espacio de alta dimensión. (Liu et al., 2010)

3.1.2.1 CONDICION DE MERCER

Dado que los datos se presentan como productos escalares (xi ·xj) se genera una transformación (Φ) de un universo a otro de alta dimensionalidad (H) donde los datos ya puedan ser procesados y separados de manera lineal.).(Cristianini y Shawe-Taylor, 2000).

La condición de Mercer permite determinar para qué núcleos existe un par {H,Φ} con las propiedades descritas anteriormente y para cuáles no. Es decir, las condiciones que deben cumplirse para que el producto escalar en el espacio de salida se puede escribir a través de un cierto núcleo K(xi ,xi) = Φ(xi)·Φ(xj).(Cristianini y Shawe-Taylor, 2000).

3.1.3 TFIDF

En la representación vectorial de cada palabra en la colección está asociada con una coordenada en un espacio dimensional alto. El valor numérico de cada coordenada a veces se llama el peso de la palabra. Aquí, se utiliza **TF** (términos de frecuencia) × **IdF** (Frecuencia de documentos de frecuencia inversa a término) (Baeza-Yates y Ribeiro-Neto, 2011) como procedimiento de ponderación de arranque (N cantidad de términos) ver ecuación 1. (Tellez et al., 2017)

$$W_{ij} = TF_{iJ} \cdot log\left(N|\,\mathrm{dF}_J\right)$$

Ecuación 1

Donde:

$\boldsymbol{W_{ij}}$ representa el peso asociado al término K_i del documento L_j,

$\boldsymbol{TF_{ij}}$ son los términos de frecuencia

$\mathbf{I\,dF_J} = log\left(N|\,\mathrm{dF}_J\right)$ Frecuencia de documentos de frecuencia inversa a término

3.2 FEATURE SELECTION CHI SQUARE

La selección de características es el proceso de seleccionar los mejores K términos como un subconjunto de los términos que aparecen en el conjunto de entrenamiento y usar solo este subconjunto como características en TC. (Thabtah, Eljinini, Zamzeer, y Hadi, 2009) y lograr dos objetivos principales. Primero, hace que la capacitación aplicada a un clasificador sea más eficiente al disminuir la alta dimensionalidad del vocabulario efectivo. En segundo lugar, la selección de características a menudo aumenta la precisión de la clasificación mediante un término poco común de reducción (Thabtah et al. 2009).

Chi Square evalúa la correlación entre dos variables y determina si son independientes o están correlacionadas (Snedecor y Cochran, 1989). La prueba de independencia, cuando se aplica a una población de sujetos, determina si están correlacionados positivamente o no.

La fórmula estadística de Chi-cuadrado está relacionada con las funciones de selección de características que intentan capturar los mejores K términos para la clase C. (Bahassine, Madani, Al-Sarem, Y Kissi, 2018).

$$x^2 = \frac{N(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)}$$

Ecuación 1.1

Donde:

N = Número de documentos en el corpus.

A = Número de la clase C que contiene el término K.

B = Número de documentos que contiene el término K en otra clase que no sea C.

C = Número de documentos que no contiene el término K.

D = Número de documentos que no contiene el término K en otra clase que no sea C.

3.3 NAIVE BAYES

Naive Bayes asume un modelo estocástico de generación de documentos. Usando la regla de Bayes, el modelo se invierte para predecir la clase más probable para un nuevo documento. (Ye, Zhang, y Law, 2009)

Naive Bayes es un clasificador bayesiano que hace una suposición simplificadora acerca de cómo interactúan las características. (Daniel Jurafsky, 2018)

Han y Kamber (2006) menciona que el clasificador trabaja de la siguiente manera:

1. Escoge atributos específicos de las palabras más importantes para que determinen el sentimiento. Donde lo atributos son definidos en la siguiente ecuación

$$X = (x_1, x_2, \dots, x_n)$$

Ecuación 2

2. Los atributos específicos resultantes que devuelven el resultado que queremos (positivo o negativo) se encuentran definidos en:

$$c_1, c_2, \dots, c_m$$

Ecuación 3

3. El clasificador Naïve Bayes para asumir la condicion de independencia de cada atributo, resulta la ecuación queda de la siguiente manera:

$$P(X|C_i) = \prod_{K=1}^{n} P(x_k|C_i) = P(X_1|C_i)\ x\ P(X_2|C_i)\ x\ \dots\ x\ P(X_n|C_i)$$

Ecuación 3

4. Para validar la predicción resultante del Clasificador *X* en la clase *C,* si solo si:

$$P(X|C_i)P(C_i)\ >\ P(X|C_J)P(C_J)\ \text{para}\ 1\ \leq\ j\ \leq\ m, j\ \neq\ i$$

Ecuación 4

3.4 VALIDACIÓN CRUZADA

La Validación Cruzada o Cross Validation es un método de evaluación conocido en minería de datos, donde los datos de entrenamiento se dividen aleatoriamente en n bloques, cada bloque se retiene una vez y el clasificador se entrena en los n-1 bloques restantes; entonces su tasa de error se evalúa en el bloque de exclusión. Por lo tanto, el procedimiento de aprendizaje se ejecuta N veces en conjuntos de datos de entrenamiento ligeramente diferentes

(Alsaleem, 2011). Por lo que dividiremos el conjunto de datos en test set y training set donde se adoptará una cantidad de dobleces o k, donde k es la cantidad de veces que los datos se dividirán al azar para estimar la precisión del modelo al igual que su variación.

4. METODOLOGÍA

4.1 RECOLECCION DE DATA

Existen muchas fuentes para obtener la opinión de los usuarios acerca de lugares específicos como: redes sociales, sitios web, eWom, o las tradicionales encuestas. En dichas opciones, se pueden obtener reseñas al igual que otras características ya sea: usuario, fecha, título, texto, rating como se muestra en la figura 2. Esto forma un conjunto de datos para posteriormente ser procesados. En este caso, se tomó como fuente de datos al sitio web Tripadvisor (https://www.tripadvisor.com.pe/), dado que como mencionan diversos autores esta es una fuente de comentarios u opiniones de lugares exactos y específicos. Para su desarrollo se usó técnicas de Web scraping sobre cinco lugares turísticos cifra superior a la que consideró Ye et al. (2009) por país, pues son los que poseerían la mayor cantidad de apreciaciones, como el Circuito Mágico de Agua (670 comentarios), Museo Larco (502 comentarios), Huaca Pucllana (642 comentarios), Parque Kennedy (586 comentarios) y Barrio Chino (391 comentarios). Posteriormente se hará uso del software Webharvy para la técnica de Web Scraping al igual que Gössling Y Lane, (2015) lo usaron para recopilar datos del sitio de reservas de hoteles booking.com del mismo modo que Khotimah Y Sarno, (2018) para posteriormente hacer sus clasificadores de texto.

Fig. 2. Comentario en TripAdvisor
Fuente: TripAdvisor

4.2 LIMPIEZA DE DATOS Y PREPROCESAMIENTO

Los comentarios muchas veces se extraen con caracteres extraños, es decir caracteres que no pueden ser identificados o clasificados ya sean: “”, *,#,%,Y,’,|. Por lo que es necesario eliminar esos elementos. De mismo modo en los comentarios se encuentran palabras que no dan un sentimiento claro (stopwords) por lo que al igual que el caso anterior estas se deben extraer para dar una mejor clasificación a cada elemento de cada opinión dentro del conjunto de datos, es decir eliminar los elementos extraños (tildes, @, etc.).

4.3 CONSTRUCCIÓN DEL CLASIFICADOR

Una vez se haya hecho la limpieza y el preprocesamiento al conjunto de datos, se procede con el etiquetado de cada opinión o comentario del conjunto de datos como positivo o negativo, según su nivel de rating (10-50), siguiendo el modelo que usó Ye et al. (2009) donde 10,20,30 son denominados corpus negativo y los demás como positivo.

Este conjunto de datos ya etiquetado según su puntaje pasa a formar parte del desarrollo del clasificador. El cual deberá otorgar un peso a cada palabra como un contador, y se verá si se considera palabra por palabra o cada par de palabras; fuera de extraer las características del conjunto de datos que se obtuvieron del paso previo, donde fue eliminada la data ruidosa e irrelevante y se determinó un subconjunto de data representativa, de tal manera que se vea reducido todo el nivel de complejidad del proceso con ayuda del estadístico Chi Square.

Es por esto por lo que el input; conjunto de datos o dataset; se dividirá en 80% para formar el training set y 20% el test set, para luego usar un kernel lineal pues es el adecuado e ideal para la clasificación de texto (Yang y Liu ,1999), con Chi Square como selección de características ya que este método a menudo aumenta la precisión de la clasificación al eliminar términos pocos comunes y reducir la alta dimensionalidad de palabras efectivas. Ya que la data que se extrae de un portal web es denominada muy ruidosa, (Thabtah et al,2009), es decir, que es necesario poseer una gran cantidad de registros para obtener resultados adecuados (Akter Y Aziz, 2017).

Las fases o etapas del algoritmo desarrollado son: categorizar, clasificar y comparar con métodos tradicionales como SVM y Multinomial Naive bayes. Diversos autores hacen mención y comparación de tres técnicas siendo dos de estos, los más utilizados e ideales para la clasificación de texto, pues sobresalen de los demás clasificadores (Ekawijana y Heryono 2016), (Khotimah Y Sarno, 2018). El motivo por el cual el método propuesto es SVM con Chi Square como selección de características, es que en las clasificaciones de texto o análisis de sentimiento en la industria turística este método es el que obtiene los mejores resultados en sus métricas a comparación de otras técnicas como: Naive Bayes, redes neuronales, arboles de decisión, entre otros. (Alsaleem,2011), (Kuhamanee et al 2017), (Ye et al. 2009), (Bahassine, et al.2018).

4.4 VALIDACIÓN DE RESULTADOS

Para la validación de los resultados se aplicará la técnica de validación cruzada con una cantidad de 12 dobleces (k= 12) para estimar la precisión del modelo al igual que su variación, dado que diversos autores recomiendan usar un k igual o mayor a 10 (Alsaleem, 2011). Al igual que la comparación de métricas de performance como Precision, Recall, Accuracy, F1 Score como fue usado por Thabtah et al. (2009) y otros autores.

5. EXPERIMENTACIÓN

El proceso de experimentación está representado en la Figura 3.1 y 3.2.

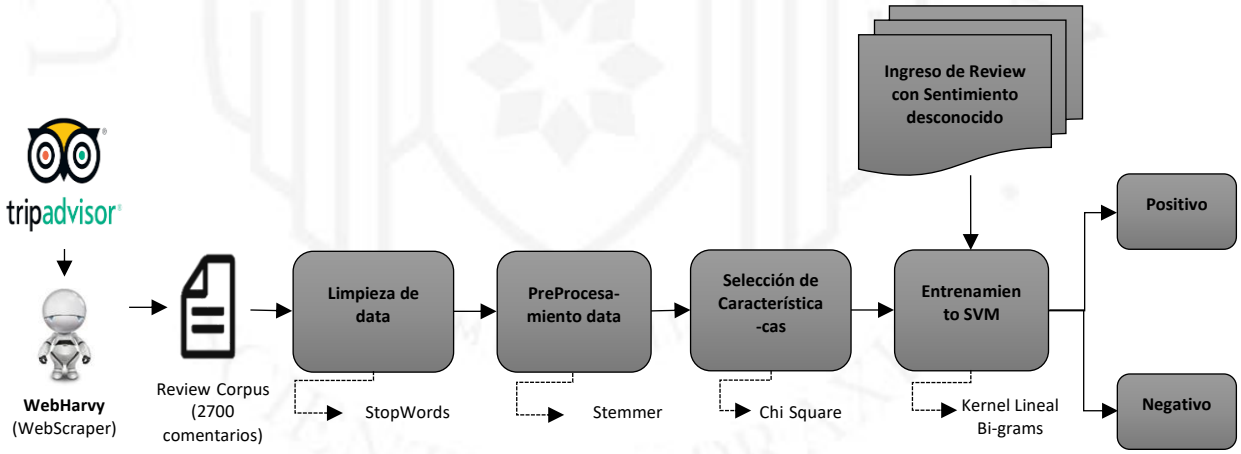

Fig. 3.1. Flujo de Modelo Propuesto
Fuente: Elaboración Propia

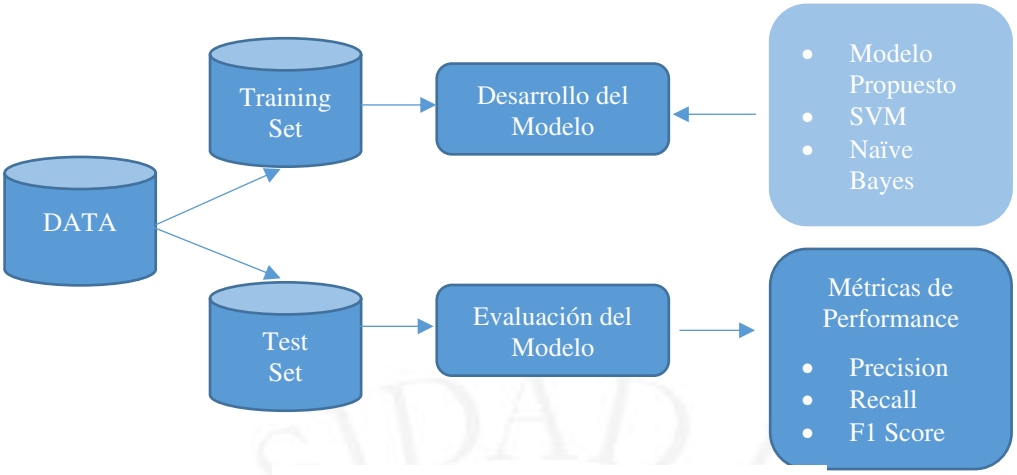

Fig. 3.2. Proceso de Experimentación
Fuente: Elaboración Propia

En la Figura 3.1 se demuestra el proceso de la experimentación para la técnica propuesta desde la obtención de la data del portal TripAvisor, hasta la clasificación de los comentarios en negativos o positivos con apoyo del estadístico chi cuadrado como selección de características, mientras que la figura 3.2 nos muestra que con la misma data se realizará las tres técnicas para la posterior confrontación de los resultados de las métricas de performance.

5.1 RECOLECCIÓN DE DATA

Para llevar a cabo la experimentación se obtuvo la data (opiniones o comentarios de lugares turísticos) del sitio web TripAdvisor (Li y Yang, 2017) mediante la técnica de web scraping. Cada comentario viene acompañado de datos como usuario, fecha, titulo, texto, rating, utilidad y ciudad (Parikh et al., 2018). Por lo que se forma un conjunto de datos gracias a esta fuente inmensa de comentarios verídicos que hay sobre lugares específicos y se selecciona las opiniones de usuarios distintos para evitar más de un comentario por usuario. Se tomó en cuenta 5 sitios turísticos de la ciudad de Lima Perú con mayores comentarios en el portal web con un máximo de 5 meses de antigüedad desde febrero del 2019 hasta Junio del 2019, contando con un total 2,700 comentarios; cifra superior a la usada por Ye et al. (2009) y Li y Yang (2017) siendo 1191 y 541 respectivamente de la misma forma que la cantidad de lugares tomados por los primeros autores. Cada recomendación posee un rating otorgado por el usuario que va desde 10 hasta 50 (10-20-30-40-50). Este pre-etiquetado por el usuario se distribuye el conjunto de datos como muestra la figura 4 donde el eje “X” representa cada sitio de interés y su rating (10,20,30,40 y 50), y el eje “Y” la cantidad de opiniones, la figura 5 demuestra la distribución del total del conjunto de datos por lugares y la figura 6 la distribución por rating:

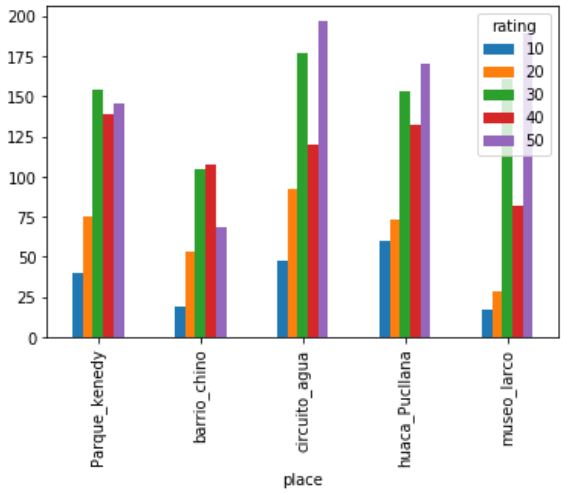

Fig. 4. Distribución de Rating por Lugar
Fuente: Elaboración Propia

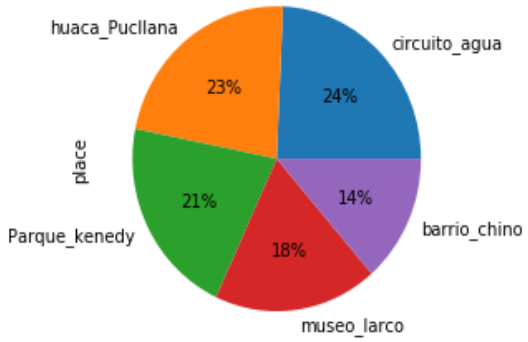

Fig. 5. Distribución de comentarios por Lugar
Fuente: Elaboración Propia

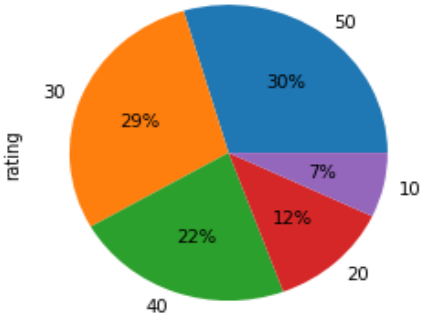

Fig. 6. Distribución del Rating
Fuente: Elaboración Propia

Por lo que se interpreta que a más burbujas o rating se tiene un sentimiento positivo más fuerte en los lugares de interés (Ye et al., 2009). Se observa en la Figura 4 que existe menor cantidad de comentarios clasificados en 10, 20 y 30, y en la figura 6 que la cantidad de comentarios categorizados como 20 y 10 representan casi el 20% del total. Por lo que consideramos como Ye et al. (2009) los ratings 10,20,30 como corpus negativo y 40 y 50 como positivo resultando los siguientes resultados:

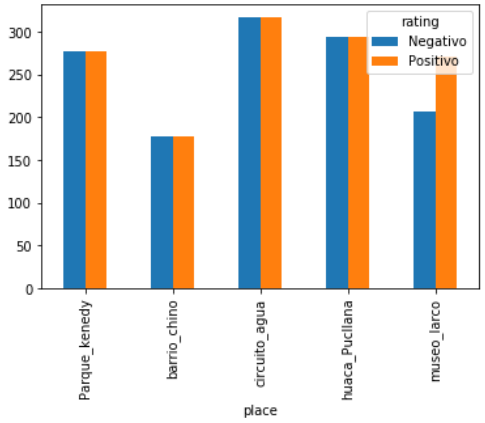

Fig. 7. Distribución de Sentimiento por Lugar
Fuente: Elaboración Propia

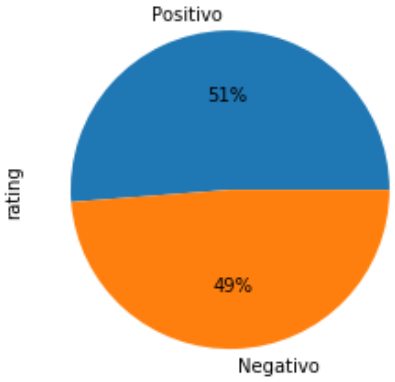

Fig.8. Distribución del Sentimiento
Fuente: Elaboración Propia

Una vez redistribuido el etiquetado del rating a sentimiento positivo y negativo tal como hizo Ye et al. (2009) mencionado en el párrafo anterior, se puede observar una comparación entre 51% positivo contra un 49% negativo en la Figura 8 a diferencia de la figura 6.

5.2 LIMPIEZA DE DATA

Para la limpieza y preprocesamiento de la información se tomó en cuenta el reconocimiento de elementos HTML, stopwords, género de palabras como número de palabra. Por ejemplo, se encontró uno como se muestra en la figura 9

> "***Visitamos el lugar en compañía de mi familia en el mes de abril, El complejo arqueológico es hermoso e interesante, muestra la arquitectura preincaica.

 El guía nos comentó acerca de la forma en que se construyó y cómo ha sobrevivo al paso del tiempo

"***

Fig 9. Comentario elementos HTML
Fuente: Elaboración Propia

Este comentario contiene la opinión del usuario además de elementos HTML que no pueden ser leídos o entendidos al momento de darle pesos a la palabra por lo que se procede a quitar dichos elementos para quedarnos con el mismo comentario de la manera que muestra la Figura 10.

> "***Visitamos el lugar en compañía de mi familia en el mes de abril, El complejo arqueológico es hermoso e interesante, muestra la arquitectura preincaica. El guía nos comentó acerca de la forma en que se construyó y cómo ha sobrevivo al paso del tiempo"***

Fig. 10. Comentario elementos sin HTML
Fuente: Elaboración Propia

En el tema de las Stopwords también se deben suprimir porque al ser solo palabras de nexo no otorgan una información útil o medible al clasificador porque lo que se reemplazan con espacios en blanco.

5.3 DESARROLLO DEL CLASIFICADOR

Una vez preprocesada y limpiada la información dentro de una nueva columna en nuestro conjunto de datos pasará a dividirse en training set y test set. El primero con un 80% del conjunto de datos y el segundo el 20% (Tellez et al., 2017) considerando los campos de conjunto de datos “Rating” y “Cleaned”, siendo este último el texto que contiene la data que resulta del paso anterior donde se obtendrán las palabras para procesar. Lo que de esa manera generará una gran matriz denominada “bag of words” que almacenará la cantidad de palabras de cada una.

El algoritmo propuesto consiste en dos fases categorización y clasificación que a su vez está compuesto en una serie de pasos usando la medición con TF-IDF (término frecuencia de documento de frecuencia inversa), el modelo de Bi-grams y el estadístico Chi Square como selección de características para comparar el training set con las categorías negativo y positivo para encontrar las palabras índices de cada categoría.

Paso 1 (Etapa de clasificación): Eliminar las stop_words, es decir quitar las palabras vacías dado que carecen de significado, tales como artículos, pronombres y/o preposiciones.

Paso 2: Normaliza el resto del training set que consta de sub- pasos como:

- Eliminar símbolos extraños: “”, *,#,%,Y,’,|.
- Eliminar espacios en blanco
- Obtener la raíz de cada palabra

Paso 3 (Transformación): Se calculó el peso de cada palabra usando TFIDF- Term Frequency (tfij) and the Inverse Document Frequency (log (N/dfj)), tomando en cuenta que el modelo propuesto escala el término frecuencia en escala logarítmica y el factor de frecuencia inversa siendo estos parámetros comunes que se usan en la técnica de SVM para procesamiento de textos, adicionando el parámetro de Bi-Grams, es decir tomando en cuenta cada par de palabras dado que es poco común encontrar coincidencias en una sola palabra (Daniel Jurafsky, 2018)

Paso 4 (Selección de características): En este paso las características más distintivas (relevantes y no redundantes) son seleccionadas del texto original (pre procesado) para ser usada como inputs en el siguiente paso. (Bahassine, et al. ,2018).

Para la selección de características se usó el estadístico Chi Square como método de extracción de características del corpus resultante del paso anterior del mismo modo que se usó en la investigación de Bahassine, et al. (2018) donde mencionó que esta técnica arroja mejores resultados. Asimismo, que su empleo logra dos objetivos, el primero es que convierte el clasificador más eficiente al reducir su alta dimensionalidad de palabras efectivas (seleccionadas) ya que mide la correlación entra palabras; y el segundo que incrementa el accuracy de la clasificación por la disminución de términos raros (no identificados). Thabtah et al. (2009).

Se calculó los puntajes del estadístico Chi Square para identificar cuales conforman el top 20 en la figura 11. Donde valor de la prueba Chi Square de una característica (término) permite evaluar su importancia para determinar una clase, es decir mostrar que palabras tienen mayor importancia al momento de definir si él comentario es positivo o negativa.

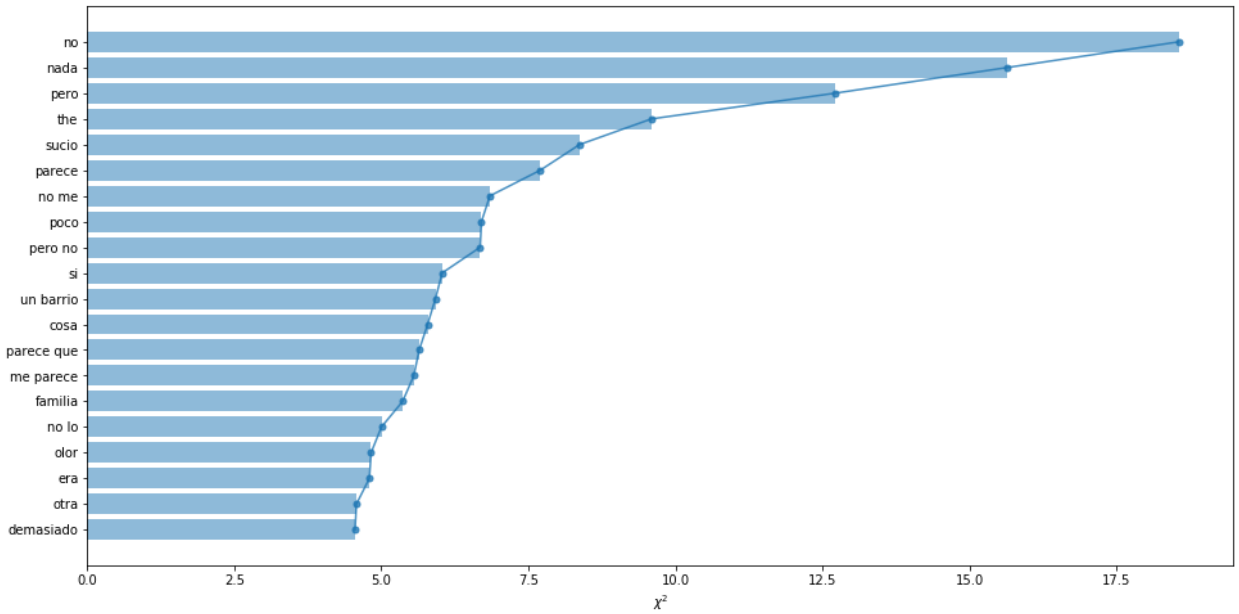

Fig. 11. Puntajes de Chi Square
Fuente: Elaboración Propia

Paso 5 (clasificación): Por medio de la clasificación SVM con un núcleo lineal y con la selección de características mencionadas en los pasos anteriores y el desarrollo de los modelos de análisis de sentimiento Multinomial Naive Bayes y SVM ambos con la misma fuente de información para posteriormente el desarrollo de la comparación de las métricas: accuracy, precisión, recall, F1-score de cada técnica.

Paso 6: Validar resultados de las técnicas desarrolladas con la técnica validación cruzada con un total de 12 dobleces (k=12).

Paso 7: Implementar una interfaz gráfica para la visualización de resultados.

6. RESULTADOS

Los resultados de la experimentación están descritos en dos ámbitos: resultados de las visitas encontradas en el portal web de TripAdvisor y análisis de la comparación de los tres modelos descritos en la investigación.

6.1 RESULTADOS DE VISITAS

En la figura 4 demuestra la proporción de los puntos de visita de los turistas en Lima, siendo el 60% de la muestra extraída parques públicos (Barrio Chino, Parque Kenedy y Parque de la Reserva) y la diferencia de la muestra en lugares arqueológicos (museo Larco y Huaca Pucllana). El sitio de interés con mayor proporción de comentarios es el Parque de la Reserva o circuito mágico, lugar que muestra diversas piletas y juego de luces, así como canciones típicas del Perú. Uno de los objetivos de esta investigación es obtener información sobre el sentimiento del turista sobre cada destino visitado con el fin de mejorar el turismo en Lima Perú, por ejemplo, en las imágenes 4 y 6 se sugiere en que lugares existe más visitas y cuáles no, de manera análoga que lugares poseen mayor proporción de comentarios negativos; para lo cual se detectó que más del 50% de los comentarios en total son positivos con rating de 40 y 50 por lo que se recategorizó en positivo y negativo; los que deben ser cubiertos para promover el turismo como Barrio Chino y el museo Larco.

6.2 RESULTADOS DE LAS TÉCNICAS DE CLASIFICACIÓN

La tabla 1 y la figura 12 muestra los resultados de las métricas Accuracy, Precision, Recall, F1 score generado por el método propuesto donde se consideró 80% del conjunto de datos como training set y el restante como test set. Luego

de analizar la Figura 12 se encuentra que el clasificador otorga un mejor resultado cuando la cantidad de características es de 5,000.

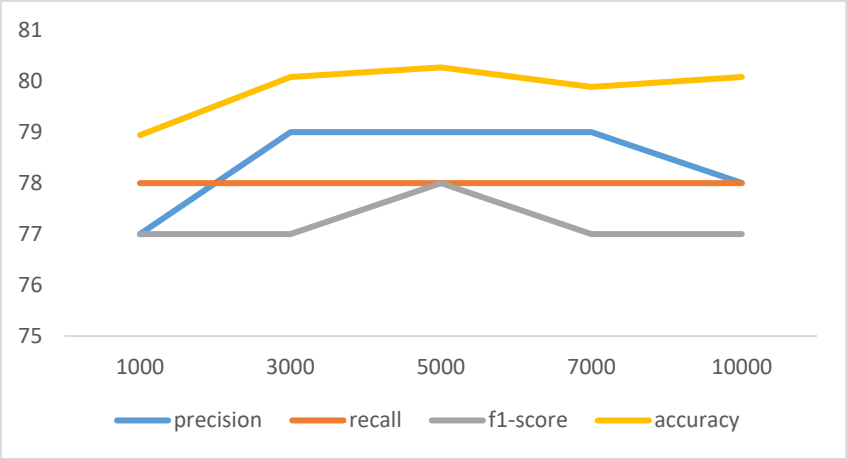

Fig. 12. Variación de Métricas según el número de características del método propuesto
Fuente: Elaboración Propia

Tabla 1. Variación de métricas según el número de características del método propuesto

Características	Precision	Recall	F1-score	Accuracy
1,000	77.0%	78.0%	77.0%	78.9%
3,000	79.0%	78.0%	77.0%	80.1%
5,000	79.0%	78.0%	78.0%	80.3%
7,000	79.0%	78.0%	77.0%	79.9%
10,000	78.9%	78.0%	77.0%	80.1%

Tabla 2 nos muestra la comparación de los niveles de accuracy o exactitud de cada técnica aplicada, donde la que tuvo la mejor performance fue la técnica propuesta (SVM con Chi Square como selector de características) seguida por SVM y Multinomial de NB. Esto debido a la selección de características de la función Chi Square que es la técnica adicional al método propuesto.

Tabla 2. Comparación de Accuracy

Técnica de Clasificación de sentimiento	**SVM**	**NB**	**Método propuesto** (Chi-Square + SVM)
Accuracy	78.53%	76.91%	80.27%

Elaboración propia

Tabla 3 muestra los resultados de las métricas de precisión, recall, F1-score donde sigue resaltando y con mejor performance la técnica propuesta seguida de SVM.

Tabla 3. Comparación de Técnicas

		precision	**recall**	**f1-score**	**support**
Prop (Chi-quare + SVM)	**Negativo**	**0.79**	**0.78**	**0.78**	**238**
	Positivo	0.80	0.79	0.79	284
	avg/total	0.75	0.77	0.75	522
SVM	**Negativo**	**0.74**	**0.75**	**0.74**	**254**
	Positivo	0.76	0.75	0.75	267
	avg/total	0.75	0.75	0.75	521
NB	**Negativo**	**0.74**	**0.77**	**0.75**	**254**
	Positivo	0.79	0.72	0.76	267
	avg/total	0.77	0.77	0.77	521

Elaboración propia

Se obtuvo que la técnica propuesta al momento de la clasificación del sentimiento negativo de un comentario es correcta en un 79% de las veces, además identifica correctamente el 78% el sentimiento de todos los comentarios dando como resultado una exactitud de la clasificación del 78%. La técnica de Support Vector Machine obtuvo una predicción del 74% de las veces y un 75% al momento de identificar el sentimiento de todos los comentarios y un 74% de exactitud en la clasificación. Finalmente, la técnica de Naïve Bayes dio un nivel de predicción correcto del 74% de las veces y 77% al momento de identificar el sentimiento de todos los comentarios y un 75% de exactitud en la clasificación, lo cual lo cataloga como la técnica con el menor nivel de precisión. Se concuerda con Thabtah et al. (2009) que chi cuadrado como selección de características aumenta la exactitud del modelo y del mismo modo con Alsaleem,(2011), Kuhamanee et al (2017), Ye et al. (2009) y Bahassine, et al. (2018). Que la técnica SVM tiene mejores resultados de performance seguido de Naive Bayes.

6.3 VALIDACIÓN DE RESULTADOS

Para la validación de nuestros resultados se decidió aplicar la técnica de validación cruzada con 12 dobleces que es un método de evaluación conocido en temas de minería de datos, donde se recomienda que la cantidad de dobleces o k sea mayor a 10 y 30 (Alpaydin, 2010)

Al finalizar la experimentación, se obtuvo que el método propuesto basado en SVM y Chi Square alcanzó una exactitud de 78.50% (desviación estándar de +/- 3.15), seguido del método de SVM clásico 73.5% (con desviación estándar de +/- 2.04), y por último el modelo de Naïve Bayes con una exactitud 73.5% (con desviación estándar de +/- 2.10)

El experimento se realizó en una laptop con sistema operativo Windows 10, procesador Intel Core i5-4210U con tarjeta de video integrada, RAM de 8GB DDR3L 1600MHz

Para el desarrollo de las pruebas de complejidad se tomó en consideración el tiempo de ejecución con la cantidad de registros resultando de la siguiente manera

Tabla 3. Pruebas de complejidad

Cantidad de Registros	Tiempo de Ejecución
2700	59.82 sec
6000	135.25 sec

Elaboración propia

Además, se tomó en cuenta el tiempo de procesamiento de otro modelo de clasificación con las mismas características que el propuesto, pero en idioma inglés con la misma cantidad de comentarios (2000). Tabla 3. Pruebas de complejidad

Tabla 4. Pruebas de complejidad

Idioma	Tiempo de Ejecución
Español	59.82 sec
Inglés	59.82 sec

Elaboración propia

De acuerdo con los resultados observado en las tablas 3 y 4 se puede determinar que la variable cantidad de registros afectaría el tiempo de ejecución de manera aproximadamente proporcional al aumento de estos mientras que la variable idioma no demuestra ninguna variación

7. CONCLUSIONES

En esta investigación se ha aplicado tres técnicas para la clasificación de texto: SVM con el estadístico Chi Square como selección de características, SVM clásico y Naïve Bayes para la clasificación de comentarios turísticos del portal web TripAdvisor en específico sobre 5 lugares turísticos de la ciudad de Lima, Perú. Sosteniendo que las técnicas aplicadas pueden detectar y clasificar el sentimiento de los comentarios, En términos de exactitud el modelo propuesto muestra la mejor performance alcanzando el 80.27% en clasificación seguido del SVM clásico con 78.53% y por último Naïve Bayes con 76.91% concluyendo que Chi Square como selección de características aumenta la exactitud del modelo. Este trabajo de investigación ha demostrado que es factible poder realizar la clasificación de manera automática e identificar que lugares poseen la mayor ratio de reseñas negativas y que se tome estos resultados para poder realizar acciones sobre dichos lugares dado que la imagen que se muestra en los portales web sobre un sitio de interés puede verse directamente afectada con el turismo de dicho lugar. Se espera que esta investigación sea útil para apoyar al procesamiento de la información que existe en el internet dado que los potenciales turistas prefieren hacer búsquedas sobre los lugares a conocer y basarse en lo que encuentren.

8. TRABAJOS FUTUROS

Esta investigación se centró en 5 lugares turísticos de la ciudad de lima, Perú por lo que se recomienda en trabajos futuros el aumento de destinos además de tomar muestras de distintos periodos de tiempos y ver si se tomaron acciones referentes a los lugares con mayor cantidad de sentimientos negativos. Por otro lado, aprovechando el uso de las redes sociales y para mantener los resultados a tiempo real sería recomendable usarlas para visualizar y analizar si existe alguna variación en los resultados sobre los mismos lugares tomados en cuenta para este trabajo y ver cuanto varia en los resultados. Además, sería conveniente el uso de otros métodos como: el uso de Naive bayes con Chi Square como selección de características o la implementación de algoritmos genéticos para el análisis de sentimiento y ver como se desenvuelve frente a los métodos ya mencionados.

9. REFERENCIAS

Alpaydin, E. (2010). Introduction to Machine Learning Third Edition. *Introduction to Machine Learning*, 350-380. https://doi.org/10.1007/978-1-62703-748-8_7

Alsaleem, S. (2011). Automated Arabic Text Categorization Using SVM and NB. *International Arab Journal of e-Technology*, *2*(2), 124-128.

Bahassine, Said & Madani, Abdellah & Al-Sarem, Mohammed & Kissi, Mohamed. (2018). Feature selection using an improved Chi-square for Arabic text classification. Journal of King Saud University - Computer and Information Sciences. 32. 10.1016/j.jksuci.2018.05.010.

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval the Concepts and Technology Behind Search*.

Becerra, C. (2016). Análisis de sentimiento en Twitter : El bueno , el malo y el > :(. *Universidad ´ Nacional de Cordoba, Argentina,*.

Chi, T. H., Lo, H. H., Chu, Y. H., & Lin, W. C. (2009). A mobile tourism application model based on collective interactive genetic algorithms. *ICCIT 2009 - 4th International Conference on Computer Sciences and Convergence Information Technology*, 244-249. https://doi.org/10.1109/ICCIT.2009.280

Choi, S., Lehto, X. Y., & Morrison, A. M. (2007). Destination image representation on the web: Content analysis of Macau travel related websites. *Tourism Management*, *28*(1), 118-129. https://doi.org/10.1016/j.tourman.2006.03.002

Cristianini, N., & Shawe-Taylor, J. (2000). Support Vector Machines. In An Introduction to Support Vector Machines and Other Kernel-based Learning Methods (pp. 93-124). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511801389.008

Daniel Jurafsky, J. H. M. (2018). Speech and Language Processing (2008), *1*. https://doi.org/10.1162/089120100750105975

Ekawijana, A., & Heryono, H. (2016). Composite Naive Bayes Clasification and semantic method to enhance sentiment accuracy score. 2016 4th International Conference on Cyber and IT Service Management. doi:10.1109/citsm.2016.7577591

Gonzalez-Rodriguez, M. R., Martinez-Torres, M. R., & Toral, S. L. (2014). Monitoring travel-related information on social media through sentiment analysis. *Proceedings - 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, UCC 2014*, 636-641. https://doi.org/10.1109/UCC.2014.102

Gössling, S., & Lane, B. (2015). Rural tourism and the development of Internet-based accommodation booking platforms: a study in the advantages, dangers and implications of innovation. *Journal of Sustainable Tourism*, *23*(8-9), 1386-1403. https://doi.org/10.1080/09669582.2014.909448

Han, J., &Kamber, M. (2006) "Data Mining Concepts and Techniques", University of ILLionis at Urbana Champaign, Elsevier, 2006, pp. 310– 317.

Hodeghatta, U. R. (2013). Sentiment analysis of Hollywood movies on Twitter. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, 1401-1404.

https://doi.org/10.1145/2492517.2500290

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. *Choice Reviews Online*, *50*(08), 50-4466-50-4466. https://doi.org/10.5860/choice.50-4466

Khotimah, D. A. K., & Sarno, R. (2018). Sentiment detection of comment titles in booking.com using probabilistic latent semantic analysis. *2018 6th International Conference on Information and Communication Technology, ICoICT 2018*, *0*(c), 514-519. https://doi.org/10.1109/ICoICT.2018.8528784

Kuhamanee, T., Talmongkol, N., Chaisuriyakul, K., San-Um, W., Pongpisuttinun, N., & Pongyupinpanich, S. (2017). Sentiment analysis of foreign tourists to Bangkok using data mining through online social network. *Proceedings - 2017 IEEE 15th International Conference on Industrial Informatics, INDIN 2017*, 1068-1073. https://doi.org/10.1109/INDIN.2017.8104921

Kwok, James. (2000). Automated Text Categorization Using Support Vector Machine.

Li, J. Bin, & Yang, L. B. (2017). A Rule-Based Chinese Sentiment Mining System with Self-Expanding Dictionary - Taking TripAdvisor as an Example. *Proceedings - 14th IEEE International Conference on E-Business Engineering, ICEBE 2017 - Including 13th Workshop on Service-Oriented Applications, Integration and Collaboration, SOAIC 207*, 238-242. https://doi.org/10.1109/ICEBE.2017.45

Lin, K. C., Wu, S. H., Chen, L. P., Ku, T., & Chen, G. D. (2014). Mining the user clusters on Facebook fan pages based on topic and sentiment analysis. *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration, IEEE IRI 2014*, 627-632. https://doi.org/10.1109/IRI.2014.7051948

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*.

Liu, Zhijie , Lv, Xueqiang , Liu, Kun & Shi, Shuicai. (2010). Study on SVM Compared with the other Text Classification Methods. 10.1109/ETCS.2010.248.

Nguyen, T. T., & Kravets, A. G. (2016). Analysis of the social network facebook comments. *IISA 2016 - 7th International Conference on Information, Intelligence, Systems and Applications*, 1-5. https://doi.org/10.1109/IISA.2016.7785412

Oliveira, Rafael Almeida (2012). Extracting Web Data From Tripadvisor As a Support for tourism indicators development in Minas Gerais. Conference: 14th Global Forum on Tourism Statistics, At Venice - Italy

Parikh, V., Kestar, M., Dharia, D., & Gotmare, P. (2018). A Tourist Place Recommendation and Recognition System. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, (Icicct), 218-222.

Prameswari, P., Surjandari, I., & Laoh, E. (2018). Opinion mining from online reviews in Bali tourist area Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017

Snedecor, W., & Cochran, W. (1989). *Statistical Methods, Eighth Edition* (Eighth Edi).

Tellez, E. S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O. S., & Villaseñor, E. A. (2017). A case study of Spanish text transformations for twitter sentiment analysis. *Expert Systems with Applications*, *81*(July 2018), 457-471. https://doi.org/10.1016/j.eswa.2017.03.071

Thabtah, F., Eljinini, M. A. H., Zamzeer, M., & Hadi, M. (2009). Naïve Bayesian Based on Chi Square to Categorize Arabic Data - Thabtah et al. - 2009.pdf, *10*, 158-163.

Valdivia, A., Luzón, M. V., & Herrera, F. (2017). Sentiment Analysis in TripAdvisor.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, *36*(3 PART 2), 6527-6535. https://doi.org/10.1016/j.eswa.2008.07.035