Grupo 12

Entrega 1 proyecto 1 202310

Juan Pablo Hernández, jp.hernandezr1

Juan Pablo Hidalgo, jp.hidalgo

Camilo Otalora, c.otalora

Con asesoría en estadística de

Valery Fonseca, vs.fonseca

Contenido

Entendimiento del negocio y enfoque analítico.

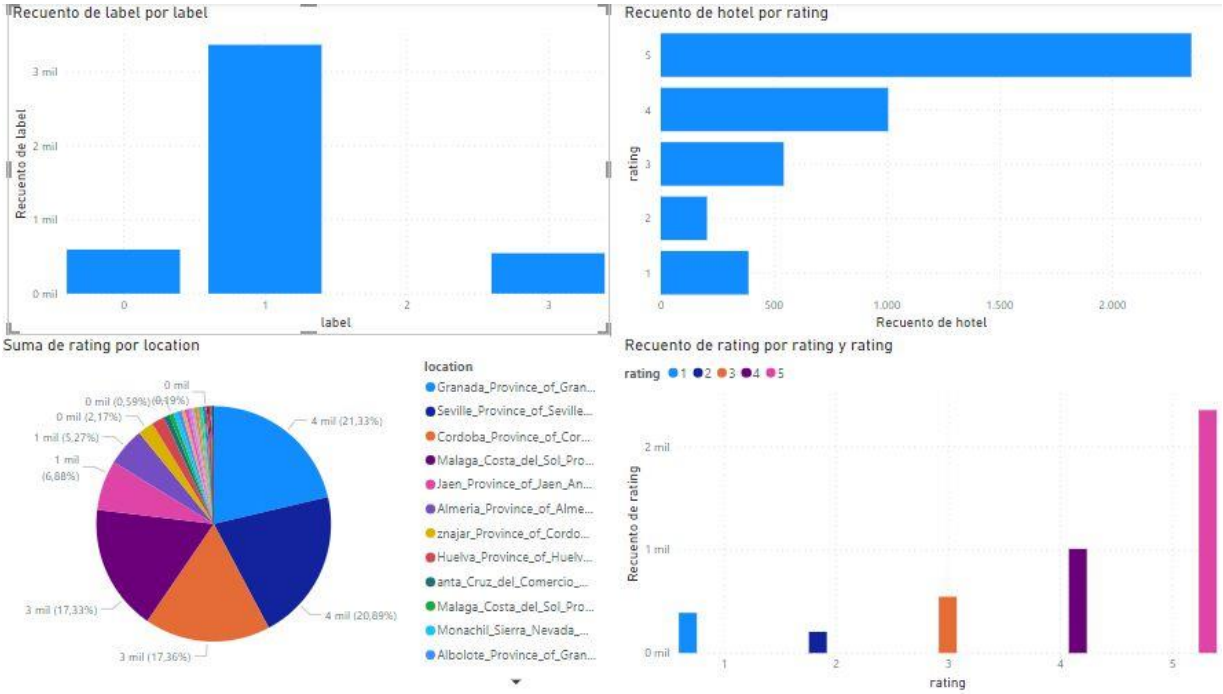
Tabla 1	
Oportunidad/problema Negocio	Análisis de opinión sobre hoteles: Este proyecto comparte comentarios en español sobre hoteles que deben ser analizados para determinar el tipo de revisión que puede obtener (e.g., excelente, perfecto)
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático)	definición del problema: poner labels en un dataset sin estos después de hacer training en uno que si los tiene Recopilación de datos: son suministrados en el zip de HotelsReviews Selección de variables: en este caso se está haciendo directamente sobre analítica de textos, por lo tanto, nos vamos a centrar en 'review_text' y 'label' Modelación: se van a ejecutar diferentes algoritmos de análisis de sentimiento de un texto, incluyendo Naive Bayes classifier, Support vector machines y Maximum entropy clasiffier. Para esto se hace una limpieza y transformación de los datos con el propósito de que funcionen correctamente. Evaluación: se evalua el proyecto con medidas de accuracy y f1_score. Interpretación: se hace una interpretación cualitativa y cuantitativa del proyecto después de ver los resultados de los modelos.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Es una organización de viajes al estilo trip advisor, puede servir para identificar puntos recurrentes que llevan a calificaciones altas y bajas para continuar haciendo y cambiar, respectivamente.
Técnicas y algoritmos a utilizar	• Clasificador Naive Bayes: Este algoritmo funciona calculando la probabilidad de que una oración pertenezca a cada una de las tres categorías (neutral, positiva o negativa) en función de las palabras utilizadas en la oración. Luego, asigna la oración a la categoría con la probabilidad más alta.

	• Máquinas de Soporte Vectorial (SVM): Este algoritmo funciona creando un modelo que puede clasificar oraciones en función de las palabras que contienen. El modelo se entrena con un conjunto de datos etiquetados (es decir, oraciones que ya han sido clasificadas como neutrales, positivas o negativas) y aprende a clasificar nuevas oraciones en función de los patrones que identifica en los datos. • Clasificador de Entropía Máxima (MaxEnt): Este algoritmo es similar al Clasificador Naive Bayes, pero utiliza una técnica matemática diferente para calcular las probabilidades de cada categoría.

Entendimiento y preparación de los datos.

En general, los datos proporcionados por el cliente son útiles (title, rating, review_text, location, hotel, label) y pueden ser aprovechados en el proceso de análisis. Sin embargo, es importante tener en cuenta que algunos registros contienen valores nulos en campos que podrían afectar la calidad de los resultados, por lo tanto, en la etapa de preparación, los registros con valores nulos se cambiaran por “na”. Estos registros no se tendrán en cuenta en el análisis y no afectarán la calidad de los datos.

Adicionalmente, en la fase de postprocesamiento de los datos se definieron y ejecutaron varias funciones para estandarizar la información. Se eliminaron los caracteres no ASCII, se convirtió todo a minúsculas, se suprimieron las puntuaciones y se cambiaron los valores numéricos por texto, también se eliminaron las stopwords. Para garantizar la consistencia y la calidad de los datos, lo que a su vez permite obtener resultados más precisos y confiables en el análisis posterior.



Modelado y evaluación

El primer modelo que se utilizo fue Naive Bayes Classifier que se define por asignar probabilísticamente cada frase a una categoría (positiva, negativa y neutra), después de vectorizar cada review. Dado que existen principalmente reviews positivos, se hizo una modificación para darle más peso a las categorías menos representadas, con la modificación de la base de sklearn MultinomialNB. Esto mejoro significativamente el accuracy y f1score del modelo, subiéndolo de 0.206 a 0.86, aunque todavía requiere modificaciones para identificar sentimientos neutros en los reviews. Para obtener el resultado del modelo se separaron los datos en 80% training 20% testing.

El siguiente modelo utilizado fue Support Vector Machines (SVM), que identifica patrones en la frase y su label asignado para ponérselo a nuevas frases. Este logro un menor accuracy que MultinomialNB de 0.84, pero sigue identificando sentimientos positivos una gran mayoría del tiempo.

El ultimo algoritmo usado fue Maximum Entropy (MaxEnt) Classifier, que tiene el mismo propósito de NBC, al asignar probabilísticamente una categoría, pero con un modelo matemático diferente. Tiene un accuracy cercana a MultinomialNB de 0.86 pero con menor f1score de identificar sentimientos neutros.

Resultados

El analizar nuestros modelos, obtuvimos unos resultados interesantes. Basados en la premisa de que queremos conocer que tipos de reseñas dan los huéspedes, y que

queremos clasificar en tres diferentes tipos, los cuales son positiva, neutra y negativa, sabiendo eso nuestros resultados fueron un poco adversos al aplicar el primer modelo, *Naive Bayes Classifier,* pues nos dimos cuenta de que en este modelo en las reseñas positivas se genera un desbalance, es decir muchas de estas que deben ser positivas, en realidad las clasificaba en otro tipo, al cual no pertenecía, por lo que de esa manera se obtuvo una exactitud tan solo de 20%, lo cual nos dice que el modelo no funciona bien en este caso, por lo que decidimos implementar un modelo derivado con mayor exactitud, el modelo usado fue *MultinomialNB,* este modelo ya puede ser capaz de identificar con mayor precisión todos los tipos de reseña, siendo así, pudimos obtener una precisión, recall y F1-score, en su primera iteración del 79%, 82% y 81% respectivamente. Por lo que podemos decir que la implementación de este modelo fue exitosa a nuestro parecer, pues tenemos unos resultados de precisión, recall y F1-score, superiores al 75%, lo que indica que se tiene una muy buena calidad del modelo, que en conclusión es lo que queremos tener de un modelo, como el propuesto, no obstante este modelo genera algunos problemas al identificar reseñas neutras, no obstante no afecta mucho la precisión y exactitud de este modelo.

Por otro lado, implementamos un modelo, *Support Vector Machines (SVMs)*, el cual es un modelo que nos presenta buenos resultados y logra identificar y clasificar las reseñas de forma correcta, con una exactitud del 83%, y una precisión, recall y F1-score, en su primera iteración de 76%, 72% y 74%, respectivamente, los cuales son muy buenos para la tarea que se pide de clasificación, estos resultados los podemos entender como una buena clasificación de reseñas, y un proceso exitoso del modelo.

Finalmente se implementó un modelo *Maximum Entropy (MaxEnt) Classifier*, el cual nos brindó unos resultados similares al anterior modelo, pero con una exactitud mayor, aunque tenemos los mismos resultados de precisión, recall y F1-score, en su primera iteración, que son 76%, 72% y 74%, respectivamente, podemos decir que este modelo es una mejor implementación para una tarea como esta, pues aun así va a tener una mayor exactitud sobre los datos tratados, por consecuente este modelo nos brinda una mejor respuesta con respecto al anterior.

En conclusión podemos afirmar, que para estos datos de testeo, los tres modelos(teniendo en cuenta que el primer modelo se cuenta el modelo, MultinomialNB, como modelo 1), nos presentan unos resultados muy buenos, sin embargo estos pueden cambiar o variar al momento de usar un pipeline, con otra serie de datos, más grandes o más pequeños, por ese motivo se debe validar estos modelos, para así encontrar el modelo preciso, sin embargo, para esta primera versión del proyecto se recomendaría al interesado, aplicar el tercer modelo, *Maximum Entropy (MaxEnt) Classifier,* esto debido a que si bien el primer modelo tiene mejores datos, nos genera unas fallas pequeñas en el momento de clasificar las reseñas neutras, por ese motivo puede que los datos de precisión y exactitud sean mayores, ya que puede estar segmentando o clasificando estos datos en otro tipo de datos, lo que hace el sistema reconocerlo como un valor verdadero positivo o falso positivo, datos que pueden llegar a afectar las

métricas, por ese motivo se recomienda el tercer modelo como modelo de clasificación de las reseñas, siendo este un modelo con buenas métricas y buen comportamiento, además que nos presenta un comportamiento normal y no tiene ninguna falla notoria en el momento de clasificar.

Trabajo en equipo

Líder de proyecto: Juan Pablo Hernández

Líder de negocio: Camilo Otalora

Líder de datos y analítica: Juan Pablo Hidalgo

Camilo Otalora se dedicó a entender y preparar los datos entregados por el cliente para su utilización en los diferentes modelos.

Juan Pablo Hidalgo se dedicó a plantear los modelos y hacer las modificaciones necesarias a estos para que presentaran un accuracy de 80%+.

Juan Pablo Hernández lidero la conexión del equipo y el entendimiento de los resultados obtenidos por los modelos llevados a cabo.

La intensidad horaria fue acuerdo a 10 horas a la semana entre los 3 integrantes del grupo por 3 semanas.

Las reuniones se llevaron principalmente entre Juan Pablo Hidalgo y Juan Pablo Hernández, con Camilo Otalora participando de manera asíncrona dado que estamos en diferentes secciones de clase.

División de 100 puntos

33: Juan Pablo Hernández

33: Camilo Otalora

33: Juan Pablo Hidalgo

Referencias

Sebastian Raschka. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition*. Packt Publishing.

https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf chapter 13

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer. https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Smith, N. A., & Eisner, J. (2005). Contrastive estimation: Training log-linear models on unlabeled data. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 354-361)

https://aclanthology.org/J96-1002.pdf