Proyecto 1 - Entrega 1

Universidad de Los Andes

Ingeniería de Sistemas y Computación

ISIS-3301 Inteligencia de negocios

Natalia Ortega	201814519
David Leon	201615216
Juan Camilo Mercado	202021541

Proyecto 1 – Entrega 1

Inteligencia de Negocios – ISIS3301

Presentado por: Natalia Ortega, David Leon y Juan Camilo Mercado

0. Tabla de contenido

Proyecto 1 – Entrega 1 ..... *1*

1. Entendimiento del negocio y enfoque analítico **1**

Tabla 1. Elementos del entendimiento del negocio 1

2. Entendimiento y preparación de datos **2**

3. Modelado y evaluación **4**

4. Resultados **5**

5. Mapa de actores relacionado con el producto de datos creado **6**

Tabla 2. Mapa de actores dentro de la organización 6

6. Elementos del equipo **7**

Tabla 3. Tabla de participación 7

Tabla 4. Roles desempeñados por cada integrante 8

Tabla 5. Planeación de reuniones 8

7. Link del video **8**

8. Referencias **9**

1. Entendimiento del negocio y enfoque analítico

Tabla 1. Elementos del entendimiento del negocio

Elemento	Descripción
Oportunidad/problema Negocio	Hacer la clasificación de reseñas que influyen en la popularidad de los destinos turísticos y en la satisfacción de los turistas para mejorar la oferta en Colombia y fomentar esta área.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.	Se usarán técnicas de procesamiento de lenguaje natural, análisis de texto y aprendizaje automático para procesar y analizar reseñas turísticas. Se propone el uso de algoritmos de clasificación de texto como Support Vector Classifier (SVC), 'Naive-Bayes multinomial', o Redes Neuronales 'MLPClassifier', junto con técnicas de vectorización de texto como TF-

	IDF (Term Frequency-Inverse Document Frequency) o Word Embeddings (Word2Vec). Además, se emplearán métricas de evaluación como precisión, recall y F1-score para evaluar el rendimiento de los modelos y se discutirán observando a detalle sus matrices de confusión respecto a los datos de validación y test. A modo de historia de usuario, una posible descripción del requerimiento es: *"Como empresa o sector perteneciente al sector turístico en Colombia, quiero predecir la puntuación entre calificaciones del 1 al 5 de nuevos comentarios para conocer la satisfacción de los clientes e identificar las oportunidades de mejora pertinentes."*
Organización y rol dentro de ella que se beneficia con la oportunidad definida	El Ministerio de Comercio, Industria y Turismo de Colombia es el principal beneficiario, ya que puede usar la información provista para mejorar la oferta turística del país y diversificar la economía. Además, las empresas turísticas, hoteles, agencias de viajes y otros actores del sector también se beneficiarían al adaptar sus servicios y estrategias de marketing según las preferencias y expectativas de los turistas.
Contacto con experto externo al proyecto y detalles de la planeación	Sofia Botía (k.botia@uniandes.edu.co) & Sophia Feghali (s.feghali@uniandes.edu.co) La planeación se pacta con los expertos y se muestra en la tabla 5 de la sección 6 del presente documento.

2. Entendimiento y preparación de datos

Para entender los datos de manera apropiada se exploró el set de datos etiquetado entregado. Primero, se notó que la columna 'Review' contenía reseñas escritas en español que tienen una respectiva clasificación del 1 al 5, donde tener una reseña con clase 1 le da notación negativa, pues estos comentarios expresan disgusto hacia la experiencia correspondiente. Al contrario, los comentarios de clase 5 expresan agradecimiento y satisfacción. Por esto, las clases pueden ser interpretadas también como la 'calificación' de la experiencia, significado que se le dan a las clases a lo largo del informe generado.

Igualmente, se observa la tendencia de las palabras de los comentarios, por lo que en el entendimiento se hace análisis de medidas de tendencia central para las palabras de los comentarios, como lo son observar la moda de los comentarios, la longitud de palabras, ver cuanto miden las palabras más larga y más cortas de cada comentario, el máximo y mínimo de palabras entre los comentarios, la cantidad de palabras diferentes, entre otros, para entender con qué tipo de reseñas se trabaja. Con esto se entendió que existe una gran cantidad de palabras repetidas en los comentarios que no aportan significado al comentario y, por lo tanto, al análisis, como lo son los denominados 'stopwords' que suelen ser la moda de cada una de las reseñas.

De la misma forma, al observar las frecuencias halladas de palabras, algunas de las más comunes son signos de puntuación, espacios y otros signos que entorpecen la clasificación por lo que se comienza el proceso de preparación de datos con la tokenización de los comentarios, dividiéndolos en sus palabras o tokens para poder tratarlos de manera aislada.

Luego, se hace la limpieza de las palabras identificadas que entorpecen el entrenaiento, como lo son las stopwords, signos de puntuación y el uso de mayúsculas, por lo que se construye un objeto capaz de llevar a cabo este proceso. En este mismo se hace el análisis lexicográfico, que consiste en extraer los lemas de las palabras (formas significativas sin importar su conjugación) y se hace a su vez el proceso de 'stemming' que consiste en reducir la palabra a su raíz. Se hacen ambos procesos ya que, al preparar el cuaderno se demuestra una mayor efectividad de los clasificadores con el uso de las técnicas 'lemmatize' y 'stemming' que usar solo una de forma aislada, pues se sabe que el uso de las técnicas depende del contexto de las reseñas y del ejercicio a realizar.

En este punto se tiene un vector de raíces sin conjugación de cada una de las palabras significativas, no stopwords del español. Por lo que en este punto, para el entendimiento del algoritmo se juntan las palabras procesadas en reseñas procesadas y se vectorizan para que la máquina pueda tener entendimiento de ellas. Para este proceso se hace una función capaz de probar 3 formas de vectorización de palabras, con CountVectorizer, TfidfVectorizer y Word2vec. Se sabe que no siempre el mejor vectorizador en un modelo va a ser siempre el mismo, por lo que, teniendo en cuenta las ventajas y desventajas de cada uno de los vectorizadores se busca elegir el mejor [1].

En la función se hace el entrenamiento de un modelo pasado como parámetro con los 3 vectorizadores, retorna el modelo entrenado cuando los datos de validación tienen el mejor resultado y el vectorizador usado y también entrenado con el cuál se obtiene el mejor resultado f1 para el algoritmo, esto implica que tiene la mejor métrica de la combinación lineal de recall y precisión, que garantiza que las reseñas clasificadas corresponden en su mayoría a su verdadera clase y que clasifica en la clase correcta a la mayoría de reseñas. Con la este entendimiento y preparación de las reseñas se procede a hacer el modelado por 3 algoritmos de clasificación elegidos.

3. Modelado y evaluación

Para el modelado se eligen 3 clasificadores, el clasificador red neuronal MLPClassifier, el clasificador Support Vector o SVC y el clasificador multinomial de Naive-Bayes multinomialNB.

El primer clasificador corresponde a modelos de aprendizaje automático que pueden aprender patrones y características complejas en los datos, incluyendo el texto en las reseñas para análisis de sentimientos. Al entrenar una red neuronal con datos de texto etiquetados, puede aprender a reconocer patrones y realizar predicciones precisas en nuevos textos lo cual es el objetivo en este contexto [2].

El segundo clasificador, correspondiente a SVC, en su implementación de scikit learn usa un método que es capaz de manejar altas dimensiones, pues, en el análisis de sentimientos, estas altas dimensiones de pueden obtener al vectorizar las reseñas como se muestra en el procedimiento de preparación de datos. SVC puede manejar eficientemente espacios de alta dimensión [2]. Igualmente, este cuenta con alta eficiencia con datos dispersos, puesto que los vectores que representan texto suelen ser muy dispersos (muchos valores son cero). Los algoritmos SVC son eficientes con este tipo de datos. Igual que los otros algoritmos, es capaz de aprender y clasificar nuevas reseñas, como lo es el objetivo del negocio planteado [3].

El tercer clasificador es el clasificador de Naive-Bayes en su forma multinomial. El algoritmo de clasificación de Naive Bayes multinomial es un método de aprendizaje automático que se utiliza comúnmente en el análisis de emociones en texto. Este algoritmo se basa en el teorema de Bayes y asume que las características son independientes entre sí. Para este caso, el algoritmo de Naive Bayes multinomial es útil en este contexto porque puede aprender a partir de un conjunto de datos etiquetados previamente y luego clasificar nuevos textos en función de las características que se extraen de ellos, lo cuál es el objetivo de esta entrega [2].

Para cada uno de los clasificadores se hace el mismo procedimiento. Primero, se halla la mejor combinación de hiperparámetros por medio de GridSearch con validación cruzada, pues, con entrena con datos de modelado y verifica la mejor combinación con datos de validación que genere la puntuación f1 de mayor magnitud que implica lo discutido previamente. Segundo, al elegir esta combinación suponiendo que es la óptima se elige el mejor vectorizador para el modelo con la función construída con anterioridad. Tercero, se hace el entrenamiento con todo el set de datos de entrenamiento (modelamiento + validación) tratado para el entendimeinto de la máquina, y contar con mayores instacias de reseñas para una predicción más exacta y, luego, se evalúa su putantaje f1 generado por el conjunto de datos de test.

Con el puntaje f1 y la matriz de confusión del conjunto de datos de test se hace la validación cuantitativa por medio de las métricas generadas. En cada uno de los casos se observan los porcentajes de precisión y recall, también se recalca el resultado de f1, los cuales se discuten por cada uno de los algoritmos dentro del cuaderno.

Igualmente, al observar la matriz de confusión se obtienen resultados similares en cada uno de los algoritmos, donde casi la mitad de los datos en los tres algoritmos los datos de una clase efectivamente se clasifican como datos de esta, lo cuál es un buen indicador de la forma en cómo clasifica las reseñas. Igualmente, se debe mencionar que la mayoría de comentarios se concentran al rededor de la clase correcta, por ejemplo, en el caso de la Clase 5 (comentarios muy positivos) la mayoría de comentarios se encuentran clasificados entre las clases 4 o 5, lo cuál denota que, aunque o se clasifican como clase 5, se sigue notando la clasificación positiva de los comentarios otorgándoles una clase 4, que significa positivo. Este elemento se considera pertinente ya que, aunque el modelo no tenga una predicción completamente exacta se hacen muy buenas aproximaciones de los contextos de los comentarios.

Por otra parte, se observan que los comentarios muy negativos (Clase 1) se clasifican la mayoría entre clases 1 y 2, algunos en la clase 3, siendo el mayor volumen en la clase, pero se le da una buena notación, ya que los comentarios negativos los clasifica efectivamente como comentarios negativos, pero con diferente fuerza, la cual puede llegar a ser subjetiva. Igualmente, se observa que muy pocos comentarios son clasificados completamente de forma errónea, por ejemplo, los comentarios clasificados como clase 5 (muy positivos) que en verdad eran clase 1 (muy negativos).

Para cada modelo se hace la evaluación de las razones del porqué del modelo de clasificar erroneamente los comentarios con gran nivel de discrepancia de la realidad. En este análisis se encuentran casuas comunes como comentarios mal etiquetados, comentarios mal escritos o con mala ortografía, también comentarios con una conotación pero que destacan aspectos contrarios a la connotación original (por ejemplo colocar 1 estrella a un hotel pero destacar la ateción del restaurante o detalles del destino turístico), incluso comentarios escritos en idiomas diferentes al español (mezcla de lenguajes) que es un aspecto que la máquina no tiene en cuenta. Habiendo generado el análisis por cada uno de los modelos descritos se elige el mejor modelo según resultados y métricas el cual se presenta como el resultado de esta etapa.

La realización de cada modelo se hizo por un miembro del equipo, el clasificador red neuronal MLPClassifier por Juan camilo Mercado , el clasificador Support Vector o SVC por David Leon y el clasificador multinomial de Naive-Bayes multinomialNB por Natalia Ortega, lo que se muestra en la tabla de participación.

4. Resultados

Como resultado de esta primera etapa, siendo elegido por las métricas descritas en la sección anterior, se presenta el modelo entrenado correspondiente al clasificador SVC. El algoritmo de clasificación elegido es adecuado en el contexto de análisis de sentimientos por medio de texto debido a que, como se discutió con anterioridad, este es capaz de manejar correctamente los vectores generados en la preparación de datos, lo que es una gran ventaja frente a los demás algoritmos, igualmente, lo muestra obteniendo las mejores

métricas y los valores dentro de rangos aceptados más acertados, asunto a discutir a continuación para desarrollarlo a mayor detalle [3].

Continuando con la discusión, se debe tener en cuenta que la métrica por la que se elige solo tiene en cuenta valores precisos y con buen recall, pero se deberían tener en cuenta valores aproximados, por ejemplo aceptar un comentario clasificado como clase 1 o 2 como comentario muy negativo y tenerlo en cuenta para la clase 1 y 2. Por esto, cabe mencionar que otro criterio de elección tenido en cuenta para la alección del modelo presentado como resultado es la cantidad de reseñas clasificadas correctamente y la cantidad de sus clases aledañas (i+1 y i-1, con i como el número de clase), siendo tenido en cuenta para la clase 1, los clasificados como clase 1 o 2, para la clase 2, los clasificados como 1, 2 y 3, para la clase 3, los clasificados como 2, 3 y 4, y así sucesivamente.

Por último, se realiza el etiquetado de los datos con el modelo elegido y se hace una validación cualitativa observando ejemplos de los datos clasificados por cada una de las clases, observando que efectivamente los comentarios de connotación negativa se clasifican como comentarios de clase 1 o 2, y los de connotación positiva como 4 o 5. Igualmente, como se recomienda a lo largo del cuarderno jupyter entregado, debe haber un especialista que esté atento a comentarios atípicos con los posibles errores mencionados en la discusión del modelado, pues estos son aquellos que el clasificador no logra clasificar de forma correcta o aproximadamente correcta.

De igual manera, cabe resaltar que lo observado en las métricas para cada modelo demuestra que a mayor número de datos de entrenamiento mejor es el recall de la clase que tiene una significancia alta en este contexto. Por esto, se recomienda entrenar el modelo con la mayor cantidad de datos posible para la clase para que así pueda predecir de una mejor forma, pues, estos asuntos se discuten y demuestran a lo largo del cuaderno jupyter anexo en la carpeta de la entrega.

5. Mapa de actores relacionado con el producto de datos creado

Tabla 2. Mapa de actores dentro de la organización

Rol dentro del Ministerio	Tipo de actor	Beneficio	Riesgo
Departamento de Turismo	Cliente	Mejora en la toma de decisiones relacionadas con la promoción turística, la asignación de recursos y el desarrollo de políticas públicas para el sector.	Desconfianza en el modelo si no logra proporcionar resultados útiles y precisos, lo que podría afectar la credibilidad del Ministerio (en general).
Dirección de Presupuesto y Finanzas	Financiador	Mejoraría la eficiencia del gasto público al identificar áreas de inversión	Riesgo de inversión financiera si el modelo no cumple con las

		más rentables y estratégicas basadas en los análisis.	expectativas de rentabilidad.
Empresas de Tecnología y Consultoría Especializadas	Proveedor	Oportunidad de desarrollar y ofrecer soluciones tecnológicas y consultoría especializada en análisis de datos y modelos predictivos para el sector turístico.	Posible crítica si las soluciones tecnológicas no cumplen con las expectativas del Ministerio y no proporcionan los resultados esperados.
Ciudadanos y Empresas del Sector Turístico	Beneficiado	Acceso a información y análisis que les permite tomar decisiones más informadas para aprovechar la oferta turística en Colombia.	Riesgo de confusión o decepción si los resultados del modelo no son comprensibles o no reflejan adecuadamente la realidad del sector.

6. Elementos del equipo

Tabla 3. Tabla de participación

Actividad	Camilo Mercado	Natalia Ortega	David Leon
Planeación de reuniones		**X**	
Gestión del repositorio	**X**		
Documentación del negocio			**X**
Documentación del cuaderno		**X**	
Investigación inicial de técnicas de análisis de emociones	**X**		
Ejecución de limpieza de datos	**X**		
Ejecución de entendimiento de datos			**X**
Modelado del algoritmo MLPC	**X**		
Modelado del algoritmo SVC			**X**
Modelado del algoritmo NB multinomial		**X**	
Análisis de resultados			**X**
Documentación de actores		**X**	
Aporte total	**33,34%**	**33,33%**	**33,33%**

Tabla 4. Roles desempeñados por cada integrante

Rol desempeñado	Camilo Mercado	Natalia Ortega	David Leon
Líder del proyecto	**X**		
Líder de negocio			**X**
Líder de datos		**X**	
Líder de analítica	**X**		

Tabla 5. Planeación de reuniones

Reunión	Fecha	Descripción
Reunión de lanzamiento y planeación	18/03/24	Para el lanzamiento del proyecto de inteligencia de negocios se establecen objetivos, alcance, roles, requisitos del cliente y plan de acción. Se asignan roles, se delinean hitos clave, se identifican riesgos y se establecen estrategias de mitigación.
Reunión de ideación	19/03/24	Se generan ideas creativas y soluciones innovadoras para el análisis de emociones. Se fomenta la participación abierta y la colaboración entre los asistentes. Se establecen los siguientes pasos para desarrollar y evaluar las ideas.
Reunión de seguimiento 1	26/03/24	Se hace el seguimiento del progreso del proyecto, revisar el estado actual y discutir los hitos alcanzados desde el lanzamiento. Se identifican posibles desafíos y se establecen medidas correctivas si es necesario. Se asignan responsabilidades para las próximas etapas y se fija la fecha para la próxima reunión de seguimiento.
Reunión de seguimiento 2	01/04/24	Se continua el seguimiento del progreso del proyecto, el avance desde la última reunión y evaluamos el cumplimiento de los hitos establecidos. Se discuten los problemas emergentes y se implementan medidas correctivas.
Reunión de finalización	05/04/24	Se evalúa la finalización exitosa del proyecto y revisamos los logros alcanzados en relación con los objetivos iniciales. Se analizan lecciones aprendidas y se destacan los puntos fuertes y áreas de mejora. Se discuten los próximos pasos, incluyendo la entrega del resultado.

7. Link del video

El link del video en el Padlet es el siguiente: https://uniandes.padlet.org/mavillam/exposici-n-proyecto-anal-tica-de-texto-de-bi-202410-dl1kygd2nk4nzomo/wish/2945025811

8. Referencias

[1] J. Vinagre Triguero, «CLASIFICACIÓN DE COMENTARIOS,» Universitat de Barcelona, Barcelona, 2023.

[2] M. Campos Mocholí, «Clasificación de textos,» Universitat Politécnica de Valéncia, Valencia, 2021.

[3] J. Amat Rodrigo, «Ciencia de datos,» Abril 2017. [En línea]. Available: https://cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines. [Último acceso: 2 Abril 2024].