Universidad de los Andes
Ingeniería de Sistemas y Computación
Inteligencia de negocios

Turismo de los alpes
Etapa 1. Construcción de modelos de analítica de textos

Integrantes (GProy26)

- Henry Santiago Antolinez - 202121785
- Juan David Orduz - 202123170
- Abel Arismendy - 202020625

Grupo de estadística

- Gabriela Coronado
- Alejandra Ramos

1. Entendimiento del negocio su enfoque analítico

Elemento	Descripción
Oportunidad/problema	Construir un modelo analítico que permita realizar la calificación automática de nuevas reseñas de sitios turísticos, con un alto nivel de precisión y sensibilidad.
Negocio	Mejorar la promoción y popularidad de los sitios turísticos en Colombia, aumentando el flujo de turistas locales e

	internacionales, lo que impulsará el sector turístico y la economía del país.
Enfoque analítico	El enfoque analítico se centra en construir modelos de análisis de textos para clasificar reseñas de sitios turísticos según su sentimiento. Se propone utilizar técnicas de procesamiento de lenguaje natural (NLP) y algoritmos de aprendizaje automático como SVC, Logistic Regression y Naive Bayes. Se aplicará la técnica de Bag of Words para representar las reseñas como vectores de frecuencia de palabras, junto con tokenización y lematización para preparar los datos antes del análisis.
Organización	Ministerio de Comercio, Industria y Turismo de Colombia, la Asociación Hotelera y Turística de Colombia – COTELCO, cadenas hoteleras como Hilton, Hoteles Estelar, Holiday Inn, y hoteles pequeños en diferentes municipios de Colombia.
Contacto con experto	Se planificará una reunión con las estudiantes de estadística Alejandra Ramos y Gabriela Coronado para validar el enfoque del proyecto y discutir los impactos esperados. **La reunión se realiza de forma virtual.**

Rol dentro de la empresa	**Tipo de actor**	**Beneficio**	**Riesgo**
Director del Ministerio de Comercio,	Cliente/Financiador	Obtener insights para mejorar la promoción	Dependencia excesiva: Existe el riesgo de que el

Industria y Turismo		turística: El ministerio puede utilizar los resultados del modelo analítico para comprender mejor las preferencias de los turistas y mejorar la promoción de destinos turísticos, lo que puede llevar a un aumento en el turismo y en los ingresos relacionados.	ministerio dependa en exceso de los resultados del modelo, descuidando otras fuentes de información o decisiones estratégicas importantes. Además, si el modelo no es preciso o confiable, podría llevar a decisiones erróneas que afecten negativamente al sector turístico.
Equipo de Desarrollo de Aplicaciones del Ministerio	Proveedor	Desarrollo de herramientas para interactuar con los resultados del modelo: El equipo de desarrollo puede beneficiarse al	Fallas en el desarrollo: Existe el riesgo de que el equipo de desarrollo no cumpla con las expectativas del ministerio o que

		crear aplicaciones o sistemas que permitan al ministerio interactuar fácilmente con los resultados del modelo analítico, lo que facilita su uso y maximiza su impacto.	surjan problemas técnicos durante el desarrollo de las herramientas, lo que podría retrasar su implementación o afectar su funcionalidad.
Investigadores de la Universidad colaboradora	Colaborador	Acceso a conocimientos especializados: El ministerio puede beneficiarse al colaborar con investigadores de una universidad, ya que estos pueden aportar conocimientos especializados en analítica de textos y ayudar a mejorar la calidad del modelo analítico desarrollado.	Divergencia de intereses: Existe el riesgo de que los investigadores tengan objetivos o agendas diferentes a los del ministerio, lo que podría llevar a conflictos de interés o a que los resultados no se alineen completamente con las necesidades del ministerio.

Empresas turísticas y agencias de viajes	Beneficiados/Clientes	Mejora en la calidad de los servicios turísticos: Las empresas turísticas y las agencias de viajes pueden beneficiarse al obtener información precisa sobre las preferencias y opiniones de los turistas, lo que les permite mejorar la calidad de sus servicios y adaptarse mejor a las necesidades del mercado.	Privacidad y confidencialidad: Existe el riesgo de que la recopilación y análisis de datos de reseñas de turistas infrinja la privacidad de los clientes o revele información confidencial, lo que podría afectar la confianza de los clientes en las empresas turísticas o agencias de viajes.

1.1 Trabajo en equipo

- **Lider de Proyecto:** Abel Arismendy. Estuvo a cargo de la gestión del proyecto. Definió las fechas de reuniones, pre-entregables del grupo y verificó las asignaciones de tareas para que la carga fuera equitativa.
    - Implementación de lematización (4h)
    - Implementación Random Forest (2h)
    - Implementación Naive Bayes Complement (2h)
    - Etiquetado de los datos (1h)

- **Líder de Negocio y Analítica:** Juan David Orduz. Fue responsable de velar por resolver la oportunidad identificada y estar alineado con la estrategia del negocio para el cual se plantea el proyecto. Garantizó que el producto se puede comunicar de forma apropiada y definió la fecha en la que se realizará la reunión con los expertos de estadística. Asimismo, gestionó las tareas de analítica del grupo y se encargó de verificar que los entregables cumplieran con los estándares de análisis y que se tiene el "mejor modelo" según las restricciones existentes.
    - Entendimiento de los datos y tokenización(1h)
    - Implementación Regresión Logística (2h)
    - Análisis de resultados (2h)
- **Líder de Datos:** Santiago Antolinez. Fue el encargado de gestionar los datos que se usaron en el proyecto y de las asignaciones de tareas sobre los datos. Los dejó disponibles para todo el grupo de tal forma que la implementación de los algoritmos fuera tarea sencilla.
    - Limpieza de los datos: Eliminación de stopwords y signos de puntuación y poner las palabras en minúsculas(1h)
    - Implementación SVC (2h)
    - Implementación MLP Classifier (2h)
    - Identificación de palabras relevantes para la clasificación (1h)

El trabajo en equipo ha sido fundamental para el desarrollo de este proyecto. Cada miembro del equipo ha desempeñado un papel crucial y ha dedicado un tiempo considerable para garantizar que cumplimos con nuestros objetivos.

En cuanto a la distribución de los puntos, se repartieron de manera equitativa entre los miembros del equipo. Cada uno contribuyó de manera significativa al proyecto demostrando un alto nivel de compromiso y dedicación. Por lo tanto, proponemos que cada miembro del equipo reciba 33.3 puntos.

Para la próxima entrega del proyecto, podríamos mejorar la planificación de nuestro tiempo. Además de perfeccionar las técnicas del procesamiento de los datos para mejorar aún más la precisión de nuestros modelos.

2. Entendimiento y preparación de los datos

2.1 Entendimiento de los datos

En un estado inicial se contó con 7875 registros equivalentes a reseñas que contaban con texto descriptivo propio de la reseña y su calificación. De la misma manera no hay presencia de valores nulos. Se interpreta entonces lo siguiente sobre los datos (previo a su verificación):

Lo coherente es pensar que las reseñas con calificaciones de 5 deben tener una descripción que se ciña a lo mencionado anteriormente. No tiene sentido entonces, que una mala reseña tenga una calificación tan alta

De modo contrario, se entiende que las reseñas de calificaciones menores como 1 o 2 son de reseñas con comentarios negativos.

El reconocimiento de lo anteriormente mencionado va a ser clave para poder hacer la distinción apropiada de palabras y de la misma manera que la clasificación de las reseñas se haga de forma apropiada.

Ya centrándonos en partes más específicas de las reseñas podemos encontrar que sobre el set completo de reseñas el valor máximo de la longitud de palabras fue de 76 y la más corta de cero. Esto resulta extraño teniendo en cuenta que no hubo ninguna reseña con valores nulos y nos permite en cierta medida comprender qué tipos de errores son necesarios corregir para poder obtener resultados que permitan cumplir los objetivos principales del negocio.

2.2 Preparación de los datos

Para poder implementar modelos que permitan hacer la clasificación correcta y apropiada de las reseñas es necesario preparar los datos de tal manera que se puedan interpretar. Es decir, construir una representación que cada algoritmo de aprendizaje pueda manejar.

Son diferentes los modelos y las técnicas utilizadas para construir de manera correcta representaciones que los algoritmos puedan interpretar. En este caso particular se implementó el modelo de **bag of word** (BOW).

El modelo de "Bag of Words" es una técnica fundamental en el procesamiento de lenguaje natural y su idea principal es representar, en este caso, las reseñas como un conjunto de palabras, ignorando la gramática y el orden de las palabras, y centrándose únicamente en su frecuencia.

Al representar cada reseña como un vector de frecuencia de palabras, se pueden identificar las palabras más frecuentes en las reseñas positivas y negativas, y así determinar la polaridad del sentimiento expresado en cada una.

En términos generales, el modelo de **Bag of Words** resulta útil para analizar y comprender las reseñas ayudando a identificar el sentimiento, extracción de características relevantes, comparar e identificar temas recurrentes

mencionados. De todas maneras, se tiene en cuenta que el BoW tiene limitaciones, como la pérdida de información sobre el orden de las palabras y su significado.

Pasos de preprocesamiento: Los pasos de preprocesamiento son fundamentales para el procesamiento de textos ya que ayudan a limpiar, normalizar y preparar los datos de texto para su correcto análisis.

A continuación se mencionan los pasos en el preprocesamiento y su justificación:

Tokenización: la tokenización se refiere a dividir el texto en palabras individuales. En este caso posteriormente a la separación en palabras individuales, se tomó la decisión de eliminar el “ruido” (información no deseada, irrelevante y disruptiva)

Normalización: En este apartado se nos presentaron dos situaciones a implementar, o **lematización** o **stemming**. Nosotros nos decantamos por usar lematización por las siguientes razones:

1. La lematización utiliza un conocimiento más profundo del idioma y reglas gramaticales para determinar el lema de una palabra. Esto significa que la lematización produce formas base que son palabras reales en el idioma, lo que resulta en una mayor precisión lingüística en comparación con el stemming.

2. La lematización es capaz de manejar palabras irregulares y excepciones gramaticales, ya que tiene en cuenta las reglas lingüísticas específicas del idioma.

Finalmente y ya con los datos preparados se dió paso a la construcción de los vectores de palabras previamente mencionados con el método “Bag of Words”

3. Modelado y evaluación

tipo de aprendizaje: Supervisado tarea de aprendizaje: Clasificación

El objetivo de un algoritmo de clasificación es asignar etiquetas o clases a los datos de entrada basados en ciertas características. De esta manera y en búsqueda de cumplir con nuestro principal objetivo (predecir la clase a la que pertenece cada dato) se aplicaron los siguientes algoritmos de clasificación.

En este apartado veremos cada uno de los modelos utilizados y las métricas respectivas para poder saber cuál fue el algoritmo y el modelo que mejor se adapta al caso

3.1 SVC

El algoritmo de clasificación SVC (Support Vector Classification) busca encontrar el hiperplano óptimo que mejor separa las clases en un espacio de características, maximizando el margen entre clases y minimizando errores de clasificación. Utiliza vectores de soporte para definir dicho hiperplano. Una vez entrenado, clasifica nuevos datos proyectándose en el espacio de características y asignándoles a una clase basándose en su posición con respecto al hiperplano.

Para el anterior algoritmo se calcularon las siguientes métricas: Exactitud: 0.51 Recall: 0.49 Precisión: 0.52 Puntuación F1: 0.50

3.2 MLP Classifier

El MLP classifier es capaz de manejar problemas de clasificación tanto lineales como no lineales y es especialmente útil en conjuntos de datos de alta dimensionalidad o con características complejas. Una vez entrenado, el modelo puede realizar predicciones sobre nuevos datos alimentándose a través de la red neuronal y asignándoles una etiqueta de clase basada en la salida de la capa de salida
Para el anterior algoritmo se calcularon las siguientes métricas: Exactitud: 0.44 Recall: 0.43 Precisión: 0.44 Puntuación F1: 0.44

3.3 Logistic Regression

El algoritmo de clasificación Logistic Regression busca encontrar los coeficientes óptimos que mejor modelan la relación entre las características de entrada y la probabilidad de pertenecer a una clase específica. A través de la función logística, transforma la salida lineal en una probabilidad que varía entre 0 y 1. Durante el entrenamiento, ajusta los coeficientes utilizando técnicas como la maximización de la verosimilitud. Una vez entrenado, clasifica nuevos datos calculando la probabilidad de pertenencia a cada clase y asignándolos a la clase con la mayor probabilidad.

Para el anterior algoritmo se calcularon las siguientes métricas: Exactitud: 0.49 Recall: 0.49 Precisión: 0.48 Puntuación F1: 0.48

3.4 Random Forest

El algoritmo de Random Forest es un método de aprendizaje automático que opera mediante la construcción de múltiples árboles de decisión durante el entrenamiento y combinando sus predicciones para obtener una predicción final. Cada árbol de

decisión se entrena en una muestra aleatoria del conjunto de datos y en un subconjunto aleatorio de las características. Durante la predicción, los resultados de todos los árboles se promedian para obtener una predicción final o se utiliza el voto mayoritario. Este enfoque ayuda a reducir el sobreajuste y a mejorar la generalización del modelo.

Para el anterior algoritmo se calcularon las siguientes métricas: Exactitud: 0,40 Recall: 0.37 Precisión: 0.40Puntuación F1: 0.37

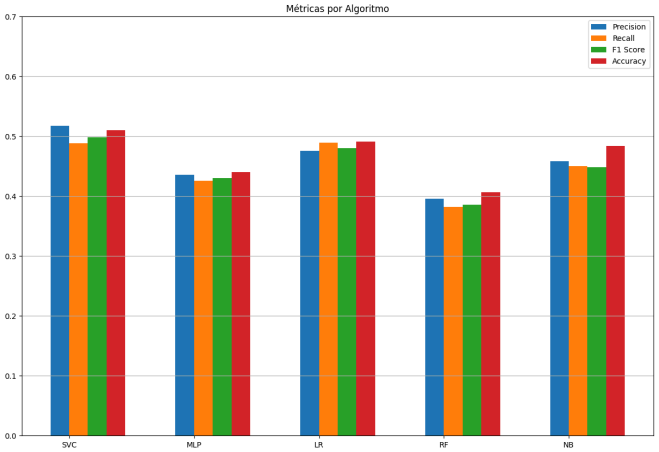
3.5 Naive Bayes

El algoritmo de Naive Bayes es un método de aprendizaje automático que se basa en el teorema de Bayes y asume independencia condicional entre las características. Utiliza la probabilidad condicional para predecir la clase de una instancia dada. Durante el entrenamiento, calcula las probabilidades de las características dadas en cada clase en el conjunto de datos. Durante la predicción, aplica el teorema de Bayes para calcular la probabilidad de que una instancia pertenezca a cada clase, y selecciona la clase con la probabilidad más alta como la predicción final.

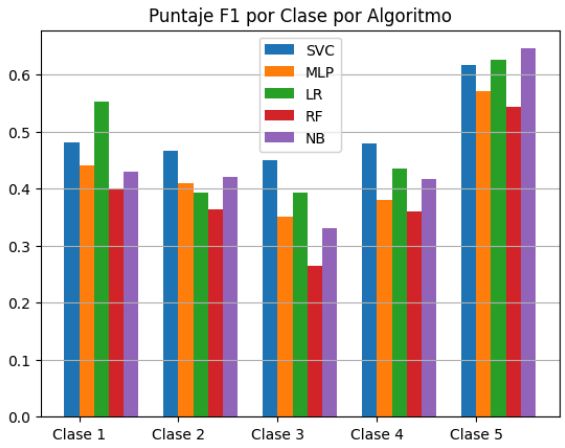
Para el anterior algoritmo se calcularon las siguientes métricas: Exactitud: 0.48 Recall: 0.45 Precisión: 0.46Puntuación F1: 0.48

4. Resultados

Nuestro análisis se centró en el uso de un modelo de clasificación SVC para analizar las reseñas de sitios turísticos. Esto, debido a que fue el modelo que mejores métricas arrojó.El modelo mostró una exactitud del 51%, un recall de 48.87%, una precisión del 51.75% y una puntuación F1 de 49.82%. Aunque estas métricas indican que hay margen de mejora, el modelo proporcionó información valiosa sobre las características que influyen en las calificaciones de las reseñas.



Métricas por algoritmo



F1-score por clase y algoritmo

Analizamos las palabras más frecuentes en las reseñas para cada calificación y encontramos patrones interesantes:

- Las reseñas de 1 estrella a menudo mencionan palabras relacionadas con la limpieza, como ‘mal’, ‘horrible’, ‘sucio’, ‘pésimo’ y ‘cucaracha’. Esto sugiere que mejorar la limpieza podría ser una estrategia efectiva para evitar reseñas de baja calificación.
- Las reseñas de 2 estrellas parecen estar relacionadas con la comida y el precio, con palabras como ‘carne’, ‘agua’, ‘caro’ y ‘malo’ apareciendo con frecuencia. Esto podría indicar que la calidad de la comida y la estructura de precios podrían ser áreas a revisar.
- Las reseñas de 3 estrellas a menudo mencionan palabras como ‘información’, ‘parecer’, ‘normal’, ‘regular’, ‘demasiado’, ‘embargo’ y ‘falta’. Esto sugiere que la cantidad y calidad de la información proporcionada podría ser un factor en las reseñas de calificación media.
- Las reseñas de 4 estrellas a menudo contenían palabras positivas como ‘mucho’, ‘duda’, ‘historia’, ‘encontrar’, ‘cómodo’, ‘excelente’, ‘estupendo’, ‘buen’ y ‘disfrutar’. Esto indica que los visitantes que tuvieron una experiencia positiva a menudo apreciaban el contexto histórico y cultural del lugar.
- Las reseñas de 5 estrellas a menudo mencionan palabras como ‘ambiente’, ‘gran’, ‘súper’, ‘espectacular’, ‘encanto’, ‘hermoso’, ‘increíble’, ‘delicioso’, ‘gracias’ y ‘excelente’. Esto sugiere que los visitantes que dieron altas calificaciones apreciaban el ambiente, la belleza del lugar y la calidad de los servicios o productos ofrecidos.

Estos hallazgos pueden ayudar a la organización a entender mejor las opiniones de los visitantes y a identificar áreas de mejora. Por ejemplo, podrían implementar estrategias para mejorar la limpieza, revisar la calidad de la comida y la estructura de precios, y asegurarse de que proporcionan información clara y precisa a los visitantes. Al hacerlo, podrían aumentar la popularidad de los sitios y fomentar el turismo.