Universidad de los Andes

Facultad de Ingeniería
Departamento de Ingeniería de Sistemas y Computación
Inteligencia de Negocios
2024-10

Proyecto 1: Turismo de los Alpes

Grupo 10:
Andrés Arévalo Fajardo - 201923853
María Castro Iregui - 202020850
Mariana Forero Avila - 201922249

6 de abril del 2024
Bogotá D.C.

Tabla de Contenido

Entendimiento de negocio y enfoque analítico

Objetivos y criterios de éxito:
El principal objetivo del negocio es identificar las características que hacen que un sitio turístico sea atractivo o no, por medio de un modelo de clasificación. En específico, que a partir de este se pueda identificar los criterios que hacen a un lugar recomendable, comparar las características de lugares de alta clasificación con los de baja y poder determinar la calificación que tendrá alguno de estos sitios para poder identificar oportunidades de mejora. Esto podría ayudar al ministerio de industria y comercio a fomentar el turismo en Colombia, aumentando así la cantidad de personas que visitan el país, estimulando así la economía local.

Oportunidad/problema Negocio	El problema principal identificado en este contexto es la necesidad de clasificar eficientemente grandes cantidades de información textual recopilada por los actores de turismo en relación con las percepciones de los turistas de diferentes sitios turísticos, para determinar la percepción que se tiene de estos.
Enfoque analítico	Tipo de aprendizaje: Supervisado Técnica de aprendizaje: Clasificación Técnicas: Naive Bayes, Regresión Logistica, Arboles de decisión Algoritmos: Bag of Words BoW, TF-IDF, Random Forest, Naive Bayes multinomial, Regression logística multinomial.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	El mayo beneficiando de esta oportunidad sería el El Ministerio de Comercio, Industria y Turismo de Colombia que podría utilizar el modelo de clasificación para identificar los sitios turísticos que mejor percepción tienen y promoverlos, además de poder identificar los sitios que necesitan mejorar y así fomentar y mejorar el turismo en el país.
Contacto con experto externo al proyecto y detalles de la planeación	Sarita Garzón | s.garzonf2@uniandes.edu.co Diana Rubio | d.rubiog@uniandes.edu.co Canal: correo, Whatsapp, Zoom Reunión para validación: 9 de abril, 6pm

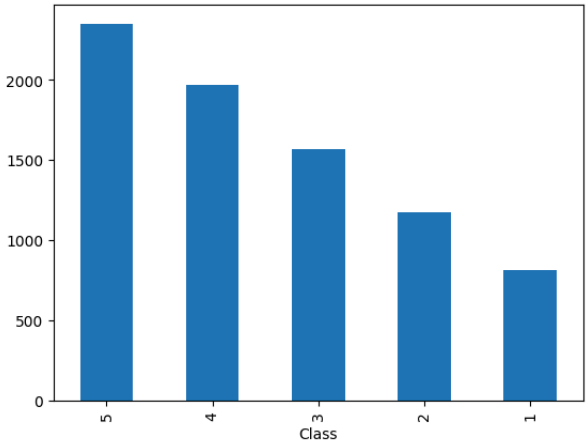
Entendimiento y preparación

Perfilamiento y Análisis de Calidad de Datos

Inicialmente se tienen 7875 registros y dos columnas. Estas dos columnas corresponden a ‘Review’ y ‘Class. La primera hace referencia a la reseña de un sitio turístico por alguna persona y es de tipo ‘object’ al ser una cadena de texto. La segunda simboliza la calificación que otorgó el turista al sitio según el sentimiento que tuvo al visitarlo. Esa columna es un entero entre 1-5.

Se identifica que hay una alta variedad en la longitud de las reseñas, con una cantidad mínima de 6 palabras y una máxima de 2495. En promedio, las reseñas tienen 70-71 palabras.

A partir del análisis de estos datos se pudo determinar que no existían valores nulos, de modo que se cumple con la dimensión de completitud. Todos los datos son válidos ya que todas las reseñas 'Reviews' son de tipo 'string' y todas las calificaciones 'Class' son numéricas y están dentro del rango preestablecido de 1-5. A continuación podemos ver la distribución de las clases. A partir de esta distribución identificamos un desbalance entre las clases, particularmente en la clase del '5' y de '1'. Este desbalance puede generar un bias o sobreajuste en el modelo hacia la clases mayoritarias, como '4' y '5'. Una posible solución para este desbalance es aplicar un algoritmo como SMOTE para balancear los datos, pero en nuestras pruebas observamos una reducción en las métricas al aplicar este algoritmo.



Distribución de las clases.

Se identificaron 71 instancias de reseñas repetidas, en base al texto y calificación que obtuvieron, estos datos duplicados fueron eliminados. Finalmente, la consistencia de los datos no se puede verificar pues no contamos con alguna otra fuente de datos para realizar la verificación.

Procesamiento y preparación

Como se mencionó anteriormente, se encontraron registros duplicados. El primer paso de la preparación fue eliminar dichos datos. También se eliminan las reseñas que no estaban en español, que eran solo 6, pues estos datos se entienden como anómalos dentro del dataset.

Posteriormente, y teniendo en cuenta las técnicas y algoritmos con los cuales se va a resolver la tarea se dividieron los datos en conjuntos de entrenamiento y de prueba. El 70% de los datos quedaron en el conjunto de entrenamiento y el 30% en el de prueba.

El siguiente paso del procesamiento se enfocó en la vectorización y la extracción de características. Usando la librería nltk se establecieron los 'stop_words' a eliminar de los

textos, estas son palabras comunes que generalmente no aportan en el significado del texto. Además se definió una función de tokenización en la cual se realizan los siguientes pasos:

1. Transformar el texto en minúsculas para reducir la dimensionalidad, eliminación del ruido y normalizar los textos.
2. Eliminar caracteres numéricos los cuales introducen ruido en el vocabulario a construir/
3. Eliminar la puntuación.
4. Eliminar caracteres especiales, como emojis y caracteres no ascii. Por su parte los emojis tienen una interpretación subjetiva y decidimos que era mejor eliminarlos de las reseñas que los contenían.
5. Transformar el texto (string) en una tokens, una lista de strings.
6. Se eliminan las palabras del texto que estén en stop words.
7. Usar el SnowballStemmer para eliminar prefijos y sufijos de palabras

Con este tokenizador se usaron dos algoritmos para la vectorización de los textos.

- Bag of Words o CountVectorizer: este algoritmo codifica si una palabra aparece o no en un texto dada su frecuencia (número de instancias).
- Tf-IDF: este algoritmo mide la importancia de una palabra en un texto ajustado con el hecho que algunas palabras aparecen con más frecuencia que otras.

Independiente al tokenizador utilizado, considerando la eliminación de stopwords, emojis y uso de stems, esto resulta en un diccionario de 12465 palabras. En este sentido, debido a que se tienen 7798 registros, lo que evidencia un problema de sobre dimensionalidad entre los features y los registros.

La delimitación de palabras realizada en la vectorización de texto utilizando el enfoque Bag of Words o CountVectorizer se llevó a cabo con el objetivo de reducir la dimensionalidad de los features, ya que un número excesivo de características puede llevar a problemas de sobreajuste y aumentar el tiempo de entrenamiento del modelo. Se aplicó un umbral de frecuencia mínima para eliminar palabras menos relevantes, lo que ayudó a seleccionar solo las palabras más significativas para la representación del texto.

Específicamente, se observó que al eliminar palabras con una frecuencia menor a 50, se redujo drásticamente el tamaño del diccionario a 699 palabras. Al aumentar este umbral a una frecuencia mínima de 10, el número de palabras en el diccionario aumentó a 2123. Finalmente, al considerar palabras con una frecuencia mínima de 1, el diccionario se amplió a 6139 palabras.

Es importante destacar que la elección del umbral de frecuencia mínima puede variar según el algoritmo de predicción, pues reducir las dimensiones del problema también tiene afectaciones en la información disponible para predecir. De este modo, se debe considerar que tan sensibles son los algoritmos frente a la sobredimensión de features.

Modelado y evaluación

Regresión Logística - Mariana Forero

La regresión logística multinomial es un modelo de predicción que utiliza la función logística para modelar la probabilidad de que una observación pertenezca a cada una de las categorías. En este caso de análisis de textos, se toman las palabras de las reseñas como las variables dependientes en la regresión a las cuales el algoritmo le asignará un coeficiente a cada una dependiendo de cuanto contribuyen a que la reseña en la que están presentes quede en determinado puntaje. Para poder procesar las palabras se probó con 2 métodos de vectorización Bag of words y TF-IDF.

En la siguiente tabla se muestran las métricas obtenidas para cada métodos:

Método Vectorización	Accuracy	Recall	F1
BoW	Train: 97.86% Test: 45.42%	Train: 98% Test: 45%	Train: 98% Test: 45.30%
TF-IDF	Train: 79.96% Test: 48.55%	Train: 80% Test: 49%	Train: 80% Test: 48.57%

A partir de estos resultados se puede apreciar que al utilizar TF-IDF para vectorizar las palabras hubo una mejora en las métricas. Además de que al utilizar bag of words se podría estar teniendo un sobre ajuste del modelo a los datos de train ya que las métricas de las predicciones de los datos de train son muy altas en comparación a las de los datos de test mientras que con TF-IDF esta brecha es menor. A raíz de esto se decidió utilizar TF-IDF como método de vectorización para el modelo de regresión.

A partir de de los coeficientes fue posible determinar las palabras más significativas para cada puntaje:

Clase	15 Palabras más significativas
1	pesim, peor, mal, suci, terribl, horribl, cobr, dij, habi, duch, ningun, diner, rob, groser, pag
2	mal, decepcion, pobr, nadi, habit, descuid, suci, esper, asign, pareci, dolar, ventan, parec, cuart, ped
3	bastant, normal, aunqu, falt, embarg, men, demasi, elev, lent, mayor, general, verd, interes, ten, acondicion
4	buen, bien, comod, fresc, ciud, tunel, sencill, excelent, histori, antigu, limpi, tambi, hermos, agrad, centr
5	excelent, delici, recomend, increibl, encant, atencion, hermos, maravill, perfect, ampli, sup, gran, graci, experient, recom

Las palabras obtenidas para describir cada puntaje tienen sentido ya que describen lo que se esperaría de una reseña con ese puntaje. Por ejemplo para las puntuaciones más bajas las palabras en general son cosas negativas como pésimo, terrible, sucio, mal, peor, por mencionar algunas. Y para las puntuaciones más altas las palabras son positivas como excelente, increible, recomendado, delicioso, entre otras. Y en el 3 que sería una clasificación intermedia palabras ambiguas o imparciales como normal, general, embargo. Estos resultados al ser coherentes demuestran que los resultados de este modelo pueden ser de gran valor para el negocio y sus objetivos.

Random Forest - Andrés Arevalo

El algoritmo Random Forest es una técnica de aprendizaje automático que se basa en árboles de decisión y se emplea tanto para clasificación como para regresión. En el análisis de texto, Random Forest utiliza una técnica de conjunto donde se construyen múltiples árboles de decisión, cada uno entrenado con una muestra aleatoria de características y registros. Los árboles de decisión evalúan las palabras presentes en cada reseña para clasificarla. La clasificación final se decide por votación mayoritaria entre los árboles, lo que refleja la opinión colectiva del conjunto. Para abordar la posible sobre-dimensionalidad, se optó por utilizar Bag of Words (BoW) como algoritmo vectorizador.

Para el desarrollo del algoritmo Random Forest se realizaron varias iteraciones, pero se muestra el resultado sobre cuatro iteraciones. Durante estas se ajustaron los hiperparámetros asociados al modelo. Estos ajustes incluyen la variación en la cantidad de estimadores, es decir, los árboles de decisión en el bosque, así como la máxima profundidad de los árboles. Además, en las iteraciones se llevó a cabo variación el filtro por umbral de frecuencia, mencionado anteriormente, buscando entender que tanto afectaba la sobre-dimensionalidad sobre el modelo. En este orden de ideas se presentan los resultados.

Parámetros por iteración	Accuracy	Recall	F1
Sin filtro por umbral de frecuencia, número de estimadores: 150 y máxima profundidad: none	Train: 100% Test: 44.9%	Train: 100% Test: 45.6%	Train: 100% Test: 42.8%
Filtro por umbral de frecuencia: 50, número de estimadores: 150, máxima Profundidad: none	Train: 100% Test: 43.2%	Train: 100% Test: 44.7%	Train: 100% Test: 42.9%
Filtro por umbral de frecuencia: 50, número de estimadores: 250, máxima Profundidad: 30	Train: 91% Test: 42.1%	Train: 90% Test: 43.9%	Train: 90% Test: 40.7%
Filtro por umbral de frecuencia: 50, número de estimadores: 250, máxima Profundidad: none	Train: 100% Test: 44.4%	Train: 100% Test: 45.5%	Train: 100% Test: 43.8%

Observamos en los resultados que en todos los casos se evidencia una fuerte presencia de overfitting. Cabe mencionar, que realizando un análisis teórico y práctico del modelo, dado el problema de análisis de textos, Random Forest es lento para estimar, costoso para predecir y peor en precisión en comparación a otros algoritmos [1]. A continuación, se presentan las palabras de mayor importancia respecto a las clases. Este análisis se hizo mediante el cálculo

de importancia estándar para un Rando Forest, el cual obtiene el porcentaje de importancia de las palabras sobre el modelo. A partir de esto se obtienen las 28 palabras más importantes, y con un análisis de frecuencia sobre estas se encuentran cuales son las más importantes para predecir entre clases.

Clase	Palabras significativas
1	hotel, habit, com, mal, mas, si, habi, servici, lleg, hac, lug, sol, buen, restaur, mejor, suci
2	hotel, habit, mas, com, buen, mal, si, servici, sol, lug, mejor, hac, lleg, pued, habi, bien, restaur,preci, haban, vist, visit, suci
3	hotel, buen, habit, mas, si, com, lug, bien, sol, mejor, servici, hac, pued, restaur, visit, haban, vist, lleg, habi, mal, ciud, atencion, excelent
4	buen, hotel, com, lug, mas, si, habit, bien, pued, visit, servici, restaur, ciud, vist, haban, hac, mejor, excelent, sol, preci, lleg, bastant, habi, atencion
5	buen, com, excelent, lug, hotel, servici, mas, restaur, visit, pued, mejor, si, bien, ciud, vist, atencion, habit, hac, recomend, sol, haban, lleg, preci

Naive Bayes - Maria Castro

El algoritmo de Naive Bayes es un algoritmo de clasificación que se basa en el uso de probabilidades y el teorema de Bayes para predecir la clasificación en base a sus características [2]. Para el caso de análisis de textos, el algoritmo analiza cada reseña como un conjunto de palabras, el cuál es construido con el algoritmo de BoW (Bag of Words) y calcula la probabilidad de que cada palabra se asocie a una de las calificaciones. Después, multiplica estas probabilidades independientes y este resultado lo multiplica por la probabilidad de que sea de cada una de las clasificaciones posibles. Finalmente, selecciona la clasificación con la probabilidad más alta. Es importante notar que este algoritmo supone todas las palabras entre sí, por lo cual es un algoritmo con alto sesgo (bias) y baja varianza.

Para el desarrollo de este algoritmo se realizaron dos iteraciones. En la primera se tiene en cuenta la vectorización del algoritmo Bag of Words. Este modelo presentó mejores métricas en cuanto al conjunto de test *(accuracy: 0.779, precision: 0.801, recall: 0.757, f1: 0.775)* que la segunda iteración pero a partir de la métricas del conjunto de train *(accuracy: 0.454, precision: 0.476, recall: 0.433, f1: 0.442)* se puede evidenciar un nivel significativo de overfitting. El segundo modelo de Naive Bayes adiciona el paso del filtro sobre las palabras de alta frecuencia, dejando solo estas y reduciendo la dimensionalidad de la matriz. Las métricas sobre el conjunto de test mejoran en el accuracy, se reduce el recall, la precisión y el f1-score se mantienen muy parecidos (*accuracy: 0.569, precision: 0.558, recall: 0.553, f1: 0.553)*. Sin embargo, observamos que las métricas del conjunto de train no están tan alejadas como en el caso anterior y aunque son significativamente menores, no presentan tanto overfitting como antes *(accuracy: 0.483, precision: 0.435, recall: 0.438, f1: 0.433).*

Para el negocio es importante entender que factores (palabras) influyen en las calificaciones obtenidas, para darle entendimiento al algoritmo de Naive Bayes, por cada clase se obtienen las palabras de más alta probabilidad. A continuación se muestran las 20 palabras con mayor probabilidad para la primera iteración del algoritmo:

Clase	Palabras significativas
1	hotel, habit, mal, com, mas, si, habi, servici, lleg, hac, noch, sol, lug, buen, teni, personal, mejor, recepcio, restaur, reserv
2	habit, hotel, mas, com, buen, si, mal, servici, sol, lug, hac, mejor, lleg, pued, habi, bien, restaur, personal, teni, esper
3	hotel, buen, habit, mas, com, si, lug, bien, mejor, servici, sol, hac, restaur, pued, visit, haban, noch, asi, desayun, vist,
4	buen, hotel, lug, com, mas, si, habit, bien, pued, visit, servici, restaur, vist, ciud, haban, hac, mejor, excelen, sol, pas
5	buen, com, excelen, lug, servici, hotel, mas, visit, restaur, pued, si, mejor, bien, ciud, vist, habit, persona, atencio, hac, pas

Selección del modelo

En base a los resultados presentados anteriormente de cada uno de los algoritmos y el objetivo del negocio, el modelo seleccionado es Regresión Logística Multinomial dado que se obtuvo la mejor métrica del f1-score de los modelos propuestos. Además desde una perspectiva de negocio las palabras en cada calificación son dicientes y no hay tan alta repetición entre las calificaciones como en los otros modelos.

Resultados

En base al modelo de Regresión Logística las palabras que caracterizan cada calificación son las siguientes:

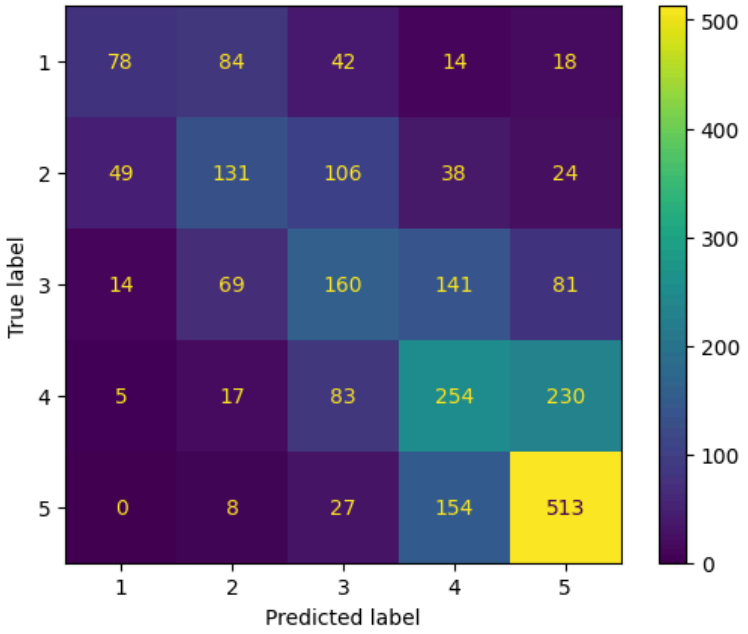
Clase	15 Palabras más significativas
1	pesim, peor, mal, suci, terribl, horribl, cobr, dij, habi, duch, ningun, diner, rob, groser, pag
2	mal, decepcion, pobr, nadi, habit, descuid, suci, esper, asign, pareci, dolar, ventan, parec, cuart, ped
3	bastant, normal, aunqu, falt, embarg, men, demasi, elev, lent, mayor, general, verd, interes, ten, acondicion
4	buen, bien, comod, fresc, ciud, tunel, sencill, excelent, histori, antigu, limpi, tambi, hermos, agrad, centr
5	excelent, delici, recomend, increibl, encant, atencion, hermos, maravill, perfect, ampli,

	sup, gran, graci, experient, recom

Las palabras obtenidas para describir cada calificación tienen sentido ya que describen lo que se esperaría de una reseña con ese puntaje. Por ejemplo para las puntuaciones más bajas las palabras en general son cosas negativas como pésimo, terrible, sucio, mal, peor, por mencionar algunas. Y para las puntuaciones más altas las palabras son positivas como excelente, increible, recomendado, delicioso, entre otras. Y en el 3 que sería una clasificación intermedia palabras ambiguas o imparciales como normal, general, embargo. Estos resultados al ser coherentes demuestran que los resultados de este modelo pueden ser de gran valor para el negocio y sus objetivos.

En cuanto a las métricas obtenidas, se obtuvo un Accuracy del 48.55%, un Recall del 45% y u F1 score del 45.30% los cuales dentro del contexto de un problema multinomial se pueden considerar buenas.

Dados los resultados observamos, consideramos que un posible camino para mejorar las métricas del modelo y que esté más listo para un ambiente de producción sería la reducción de etiquetas en la variable objetivo.



En particular, a partir de la matriz de confusión del modelo de regresión logística podemos identificar que la mayor confusión es entre clasificaciones adyacentes, como el 4 y el 5. Consideramos que para el negocio, puede ser más valioso cuales son las reseñas negativas, sin la especificación del grado de negatividad que dan estas dos clases. Dicho esto, se puede desarrollar una siguiente iteración con la siguiente reducción de clases:

- Negativo (anteriormente clases 1 y 2)
- Neutro (anteriormente clase 3)
- Positivo (anteriormente clases 4 y 5)

Esta información es útil para la organización pues le ayuda a cumplir sus objetivos de identificar las características que determinan que un sitio turístico sean clasificados positiva o negativamente con lo cual va a ser posible predecir la clasificación de un sitio a partir de sus reseñas. Así la organización podrá implementar estrategia que permita a que los diferentes actores puedan adoptar medidas para recomendar los sitios turísticos con mejores calificaciones y mejorar los que tienen una menor clasificación. Otra posible estrategia es el desarrollo de programas para la mejora de sitios con reseñas negativas.

Mapa de actores

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Ministerio de Comercio	Usuario - Cliente	Contribuir al desarrollo y promoción del turismo en Colombia, lo que puede generar beneficios económicos y sociales a largo plazo.	Dependencia excesiva de los resultados del análisis, lo que podría llevar a decisiones erróneas si no se interpreta adecuadamente la información proporcionada.
Cadenas hoteleras y asociaciones turísticas	Financiador	Obtener información detallada y análisis sobre las características de los sitios turísticos, lo que puede ayudar a mejorar la calidad y atractivo de sus destinos turísticos.	Inversión de recursos financieros en un proyecto que podría no cumplir con las expectativas o no generar los resultados esperados.
Universidad de los Andes(entidad colaboradora)	Proveedor	Oportunidad de ofrecer servicios y soluciones especializadas para el análisis de datos turísticos, lo que puede generar ingresos adicionales y mejorar su reputación en el mercado.	Responsabilidad en la calidad y ética de los modelos de clasificación, ya que pueden afectar decisiones importantes en políticas públicas del turismo.
Turistas locales y extranjeros	Beneficiado	Acceso a información más precisa y detallada sobre los sitios turísticos, lo que les permite tomar decisiones informadas y mejorar su experiencia de viaje.	Vulneración de la privacidad y seguridad de los datos personales si la información se maneja de forma inadecuada.

Trabajo en Equipo

Roles

- Andrés Arevalo: Líder de datos, Líder de analítica, desarrollo del algoritmo Random Forest
- Maria Castro: Lider de proyecto, Líder de analítica, desarrollo del algoritmo Naive Bayes
- Mariana Forero: Líder de Negocio, Líder de analítica,, desarrollo del algoritmo Regresión Logística

El entendimiento y preparación de los datos y el documento fue realizado por todo el equipo.

Tiempos

- Reunión de lanzamiento y planeación del trabajo: 2 horas
- Entendimiento del negocio: 1 hora
- Entendimiento y preparación de los datos: 3 horas
- Desarrollo del primero modelo: 4 horas
- Desarrollo del segundo modelo: 4 horas
- Desarrollo del tercer modelo: 4 horas
- Reunión de análisis de resultados y seguimiento: 2 horas
- Creación del documento: 3 horas
- Creación del vídeo: 2 horas

Además de las reuniones, se hizo un seguimiento continuo por medio de canales de texto, por los cuales se informaron de avances, dudas y obstáculos enfrentados.

Retos y Soluciones

Durante la fase de tokenización enfrentamos bastantes retos en cuanto a cómo poder limpiar los datos de la mejor manera y sin perder información valiosa. Por un lado, encontramos que varias reseñas contaban con emojis e inicialmente no teníamos muy claro qué hacer con ellos puesto que aunque sí tienen un valor semántico, este no es necesariamente objetivo y podría introducir ruido en los datos. Todo esto fue discutido dentro del grupo para llegar a un consenso y al final eliminamos todos los caracteres que no fueran ascii de los datos. Por otro lado, después de la vectorización de los datos identificamos un problema de dimensionalidad en los datos, pues nuestro vocabulario era del orden de miles de palabras, ie. decenas de miles de columnas mientras que solo teníamos alrededor de 5000 datos para entrenar el modelo. A partir de este reto, determinamos que había muchas palabras anómalas que aparecen muy poco en el modelo, muchas de las cuales eran errores en la digitación de las reseñas. Con esto fue que planeamos la eliminación de palabras con frecuencia mencionada en el punto de procesamiento y preparación lo cuál redujo a aproximadamente XXX el vocabulario.

Repartición Puntos

Andrés Arévalo: 33.33, Maria Castro: 33.33, Mariana Forero: 33.33

Referencias

[1] Scikit Learn, “Classification of text documents using sparse features” url: https://scikit-learn.org/stable/auto_examples/text/plot_document_classification_20newsgroups.html (accessed Apr. 5, 2024).

[2] S. Ray, “Naive Bayes classifier explained: Applications and practice problems of naive Bayes classifier,” Analytics Vidhya, https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ (accessed Apr. 3, 2024).