2024

Proyecto 1: Analítica de Textos

ETAPA 1

SANTIAGO RODRIGUEZ CRUZ

SEBASTIÁN UMAÑA PEINADO

Contenido

1 Entendimiento del negocio y enfoque analítico.

a.
Objetivo: Aislar las características que asocian a un sitio a una opinión positiva o negativa para que el ministerio de turismo y las cadenas hoteleras definan qué características son pertinentes para que las personas perciben hacia un lugar y crear estrategias de mejora para impulsar el turismo en estos lugares.

Criterios de Aceptación:
-Predecir si una reseña es positiva o negativa
-extraer características que definan si una reseña es positiva o negativa
-
El impacto de nuestro proyecto está basado en Colombia y en el turismo desde la perspectiva de COTELCO, este mismo sería muy positivo porque impulsa el crecimiento y la competitividad del sector económico. Además de todo lo anterior, puede llegar a tener un impacto las comunidades locales, propulsar la economía nacional con la atracción de extranjeros y mejorar la imagen del país teniendo en cuenta lo predicho por el modelo.
b.

Ya que los actores interesados en el proyecto requieren extraer, procesar y utilizar las percepciones de un sitio según sus reseñas, que pueden ser positiva o negativa para crear estrategias que fomenten esta industria se usará la técnica de aprendizaje supervisado de clasificación. Se Hará un entendimiento inicial de la estructura general de los datos, Se procesarán las reseñas por medio de técnicas de lenguaje natural y limpieza general de datos y posteriormente se convertirá en un formato entendible por el modelo para crear resultados pertinentes e ilustradores.

Dado que el contexto y corpus de las reseñas es estable, se probará el entrenamiento del modelo vectorizando las reseñas con TF-IDF. Elegimos TF-IDF dado que permite resaltar las palabras de mayor importancia en todo el corpus

del texto ,filtrando las que estén presentes en todos los textos dado que no aportan información útil sin ignorar las que son relevantes en cada documento.

Se valorarán los siguientes 3 algoritmos y se harán pruebas para elegir el mejor según lo que necesite la empresa.

-Logistic Regression

-Random Forest

-SVM

Se eligieron estos algoritmos dado su uso frecuente en tareas similares de análisis de texto y que manejan bien grafos/matrices dispersas, que son las que arroja el TF-IDF.

C.

Estudiante estadística 1: Julio Alberto Gutierrez De Armas 202414355
Estudiante Estadística 2: Karol Biviana clavijo criollo 202412647

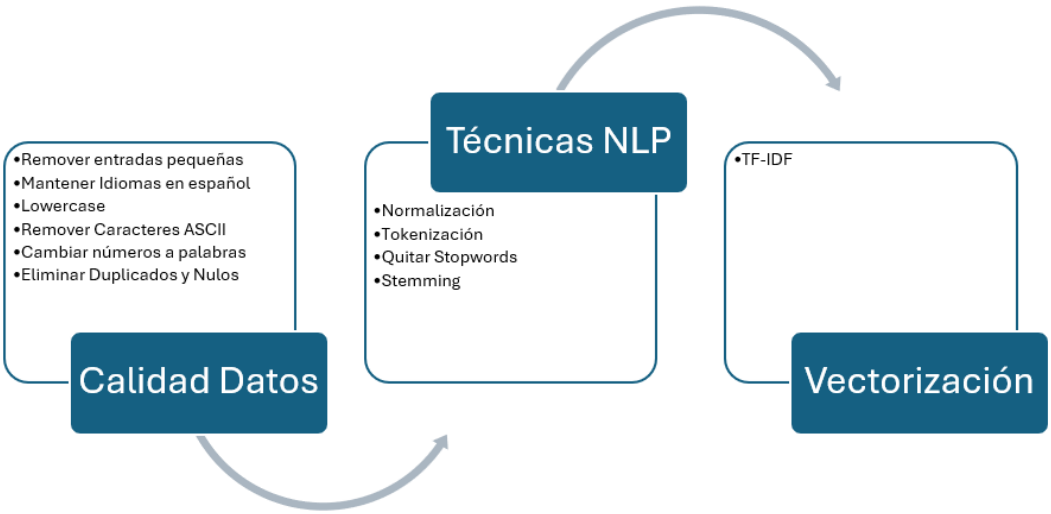
Las reuniones de seguimiento se hacen al inicio de cada reunión antes de tratar el tema pertinente. Esto incluye la primera reunión pues se trabajo de manera individual en semana santa.

Fecha Reunión 1: 1/04/2024
Reunión de lanzamiento y planeación
Fecha Reunión 2: 2/04/2024
Reunión de ideación
Fecha Reunión 3: 5/04/2024
Reunión de Finalización

Canal de Comunicación: WhatsApp y Google Meet

2.Entendimiento y preparación de los datos

Se hizo el siguiente proceso antes de probar los modelos



En esta preparación se decidió incluir una columna llamada “Rating”. Esta columna consiste en una simplificación de las reseñas en tres tipos :1,2,3 las cuales corresponden a Negativo, Neutral y Positivo respectivamente, dicha columna se crea según la clase. Reseñas de clase 1 y 2 son tipo 1, clase 3 es tipo 2 y clase 4 y 5 es tipo 3.Esta columna tiene el propósito de ayudar al modelo a entrenarse por medio de condensar características

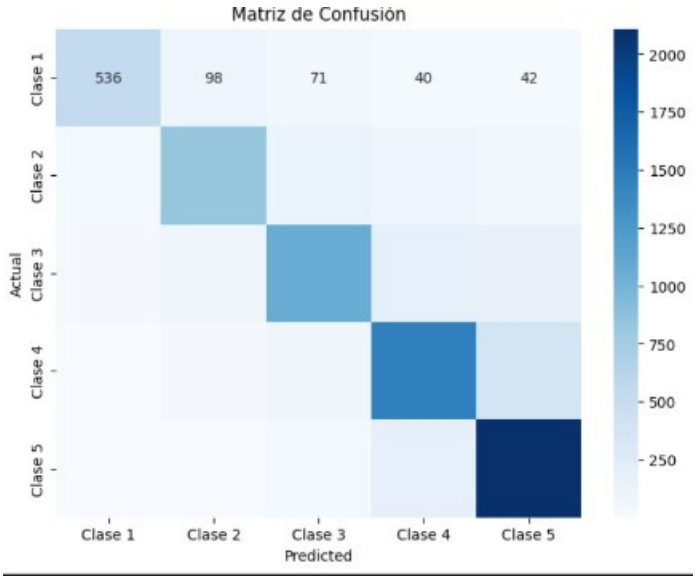
3. Modelado y evaluación.

3.1.Logistic Regression:

La regresión logística es un método estadístico utilizado para tareas de clasificación binaria, donde el objetivo es predecir la probabilidad de que una instancia pertenezca a una de dos clases. Ajusta una función logística (también conocida como función sigmoidea) a los datos de entrenamiento. La función logística asigna cualquier valor de entrada a un valor entre 0 y 1, que puede interpretarse como la probabilidad de que la entrada pertenezca a la clase positiva. La elegimos dado a que las reseñas se pueden clasificar dentro de un espectro de “positivo” o “negativo” el cual es binario ,además de esto funciona bien con datos dispersos que estaría arrojando el TF-IDF y finalmente se puede regularizar de manera fácil para evitar el Overfitting.

Métricas

	precision	recall	f1-score	support
1	0.87	0.68	0.76	787
2	0.79	0.71	0.75	1160
3	0.77	0.69	0.73	1551
4	0.74	0.74	0.74	1962
5	0.76	0.90	0.83	2338
accuracy			0.77	7798
macro avg	0.79	0.75	0.76	7798
weighted avg	0.77	0.77	0.77	7798



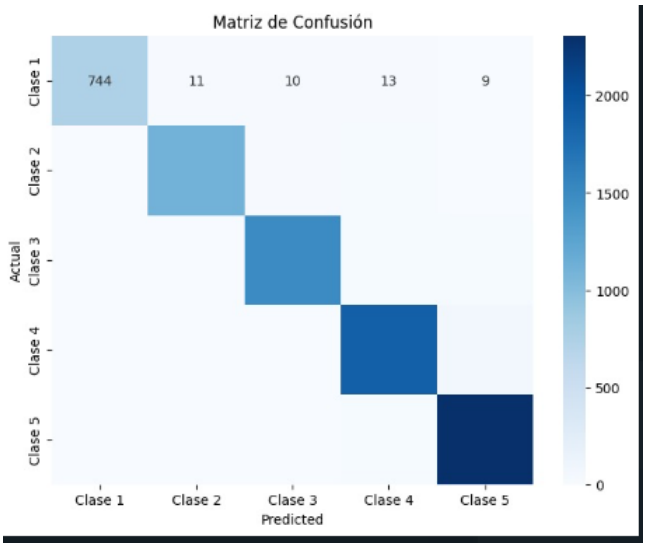
3.2 SVM

Support Vector Machine (SVM) es un algoritmo de aprendizaje automático supervisado que se utiliza para tareas de clasificación y regresión. En el contexto de la clasificación, SVM tiene como objetivo encontrar el hiperplano óptimo que separe las instancias de diferentes clases en el espacio de características.Lo elegimos por que además de ser similar a Logistic Regression , es robusto a overfitting y se adapta a datos dispersos, también se adapta bien a muchas dimensiones, que para esta tarea corresponde al vocabulario del corpus donde cada palabra es una dimensión.

Métricas

SVM

	precision	recall	f1-score	support
1	1.00	0.95	0.97	787
2	0.98	0.96	0.97	1160
3	0.98	0.96	0.97	1551
4	0.96	0.96	0.96	1962
5	0.95	0.99	0.97	2337
accuracy			0.97	7797
macro avg	0.97	0.96	0.97	7797
weighted avg	0.97	0.97	0.97	7797



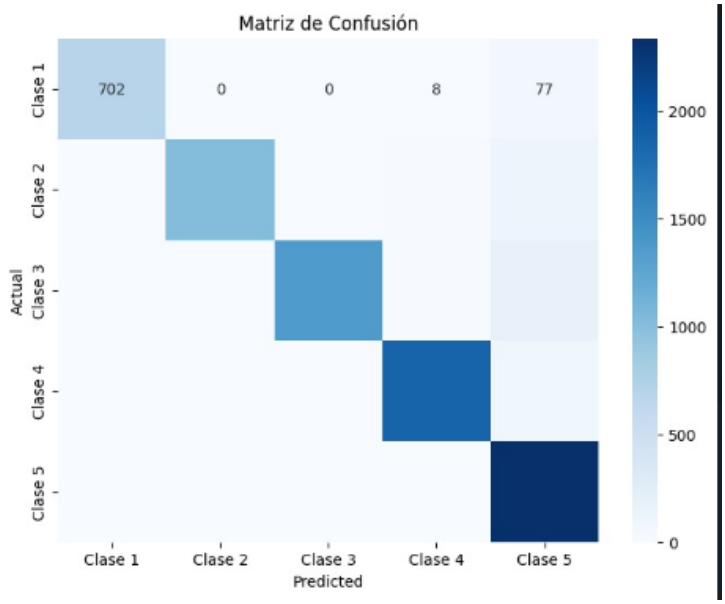
3.3 Random Forest

Random Forest es un método de aprendizaje conjunto que se utiliza tanto para tareas de clasificación como de regresión. Opera construyendo una multitud de árboles de decisión durante el entrenamiento y genera el modo de las clases (para clasificación) o la predicción promedio (para regresión) de los árboles individuales. Lo elegimos por las razones de los otros dos algoritmos pero agregado a esto aprovechamos la capacidad de generar varios arboles como estimadores para una clasificación mas acertada.

Métricas

Random Forest

	precision	recall	f1-score	support
1	1.00	0.89	0.94	787
2	1.00	0.89	0.94	1160
3	1.00	0.88	0.93	1551
4	0.98	0.95	0.97	1962
5	0.83	1.00	0.91	2337
accuracy			0.93	7797
macro avg	0.96	0.92	0.94	7797
weighted avg	0.94	0.93	0.94	7797



4. Resultados

Las métricas de calidad son buenas en todos los algoritmos,todos tienen un recall y precisión encima de 75%, casi perfecto para SVM y Random Forest. Decidimos usar los 3 algoritmos aunque en términos estrictos solo sería necesario usar Logistic Regression con SVM o Randomforest para formar resultados Apropiados.

Entre las palabras mas influyentes encontramos las siguientes a través de las 5 clases.

```
Palabras importantes para el clasificador RandomForestClassifier:
excelent: 0.01383849491186228
mal: 0.009064047545355305
habit: 0.008819072986754219
hotel: 0.0075646026401932995
pesim: 0.00636416359054257
suci: 0.006137814657317129
com: 0.005867246107065754
servici: 0.005481020176657261
recomend: 0.005392617178432088
delici: 0.0053784432633780885
lug: 0.004920784396735126
atencion: 0.004679381771173 02
lleg: 0.004430361942474149
car: 0.0043544545162286855
esper: 0.004247043014309616
increibl: 0.004216649595776959
ciud: 0.004119201740810522
desayun: 0.004057319097844643
restaur: 0.004025447856119509
visit: 0.004013866852601542
```

Encontramos las Siguientes palabras clave en una reseña positiva y negativa.

```
Palabras para malas reseñas:
['excelent' 'ubic' 'limpi' 'delici' 'histori' 'agrad' 'comod' 'ambient'
 'bonit' 'ric' 'ampli' 'camin' 'hermos' 'tom' 'espectacul' 'pued' 'piscin'
 'vist' 'mejor' 'ciud']

Palabras para buenas reseñas:
['pesim' 'suci' 'mal' 'terribl' 'horribl' 'groser' 'rob' 'pag' 'decepcion'
 'desagrad' 'cucarach' 'diner' 'cobr' 'saban' 'asquer' 'recom' 'fatal'
 'iba' 'duch' 'ped']
```

Estrategias propuestas

-Los lugares con percepción negativa es principalmente por la suciedad y la inseguridad. Por el lado de la suciedad, hay presencia de cucarachas y las áreas de higiene como baños están en mal estado.Por el lado de la inseguridad hay una alta tasa de robos por lo que se deben aumentar y mejorar las medidas de seguridad actuales.

-Los lugares con percepción negativa se ven afectados también por un mal servicio por parte de los empleados a los clientes. Hay que cambiar o re-entrenar al personal actual para mejorar el servicio. Para los lugares de buena calificación, hay que mantener el buen desempeño.

-Para los lugares con buen rating se debe mantener el aseo y mantener la calidad de los servicios ofrecidos como la calidad de la comida,piscinas,etc.

-Se debe capitalizar en las características de la ubicación. Por ejemplo, ofrecer tures a lugares históricos cercanos,ubicar estratégicamente las facultades del hotel con una buena vista(ej: la piscina con vista a la playa).

Justificación Utilidad de datos.

Los datos son relevantes dado que permiten hacer afirmaciones concretas sobre las características de los hoteles en base a las reseñas. Es afín a lo que desea el negocio pues se identifica que palabras influyen en una reseña , las características positivas o negativas de un hotel según esas reseñas y en ese orden de ideas la elaboración de estrategias claras,concisas y enfocadas a un área en particular.

5.Mapa de Actores

Rol dentro de la empresa: Junta Directiva COTELCO

Tipo de Actor: Interno - Asociación

Beneficio: Toma de decisiones estratégicas y dirección del enfoque del proyecto desde la perspectiva de COTELCO.

Riesgo: Posible falta de alineación con los objetivos del proyecto o resistencia al cambio por parte de la junta directiva.

Rol dentro de la empresa: Funcionarios COTELCO

Tipo de Actor: Interno - Asociación

Beneficio: Implementación operativa del proyecto, asegurando el cumplimiento de los objetivos establecidos por la junta directiva.

Riesgo: Posible falta de capacitación o recursos insuficientes para llevar a cabo eficazmente las tareas relacionadas con el proyecto.

Rol dentro de la empresa: Subdivisiones/Capítulos COTELCO

Tipo de Actor: Interno - Asociación

Beneficio: Apoyo localizado y adaptación del proyecto a las necesidades y realidades específicas de cada región o capítulo de COTELCO.

Riesgo: Posible falta de cohesión o divergencia en la implementación del proyecto entre las diferentes subdivisiones o capítulos.

Rol dentro de la empresa: Aliados COTELCO

Tipo de Actor: Externo - Asociaciones, Empresas, etc.

Beneficio: Colaboración y apoyo externo al proyecto, ampliando el alcance y la influencia de COTELCO en la industria turística.

Riesgo: Posible falta de compromiso por parte de los aliados o conflictos de intereses que puedan surgir durante la colaboración.

Rol dentro de la empresa: Hoteles

Tipo de Actor: Externo - Empresas

Beneficio: Participación activa en la oferta turística y potencialmente en la implementación de iniciativas de desarrollo turístico promovidas por COTELCO.

Riesgo: Posible falta de interés o adhesión por parte de algunos hoteles, lo que podría limitar el impacto del proyecto.

Rol dentro de la empresa: Turistas

Tipo de Actor: Usuarios

Beneficio: Consumo de servicios turísticos ofrecidos por los hoteles y otros actores involucrados en el proyecto, contribuyendo así al desarrollo económico del sector.

Riesgo: Posible percepción negativa o falta de interés en las ofertas turísticas promovidas por el proyecto, lo que podría afectar la afluencia de turistas y su impacto económico.

6.Experiencia del trabajo en equipo

Como tal fue mas complejo dado que solo éramos dos en el grupo de BI. Entre Investigación,desarrollo,pruebas,mejoras de las métricas ,documento y reuniones con el grupo de estadística fue aprox 12h cada uno. La investigación fue

interesante, tuvimos la oportunidad de analizar diferentes enfoques como complementar con análisis de VADER,Comparación de modelos,etc. Fue interesante la parte de investigación y desarrollo pues tuvimos oportunidad de cruzar los temas vistos en la materia con la materia SISTEMAS DE RECOMENDACIÓN en la parte de NLP, la representación de palabras usando el modelo vectorial de Salton , técnicas como Word2Vec y Doc2Vec y finalmente como interpretar y manejar las clases .

Santiago Rodriguez Cruz: Líder de Proyecto,Líder de Datos

Sebastián Umaña Peinado:Líder de de Negocio,Líder de Datos

Si tuviéramos que repartir un porcentaje de 100 entre los dos lo haríamos 50 y 50.El aporte de cada uno fue apropiado, nos repartimos según las aptitudes y debilidades de cada uno y resultamos con una solución apropiada para lo que dicta el negocio. El proceso general de desarrollo del proyecto está en el siguiente diagrama