Informe Proyecto 1

Construcción de modelos de analítica de textos:

Análisis de sentimientos de películas

Inteligencia de Negocios

2023-10

Integrantes:
Juan Felipe Patiño 201922857
Juan Ricardo Diaz 201922167

1. Entendimiento del negocio y enfoque analítico.

Oportunidad/problema Negocio	La oportunidad de negocio al alcance es la rápida identificación de la opinión pública sobre una película según los tuits publicados al respecto. De esta manera, un teatro de cine podrá saber como aumentar o disminuir la disponibilidad de funciones acorde al interés y opinión pública desde las primeras funciones.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático)	Para un Modelo de Machine Learning es necesario tener los textos, en Español, a analizar y etiquetados por una persona sobre una percepción positiva o negativa. Estos datos deberían estar dentro de un archivo CSV
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Para los teatros el uso de esta consultoría y creación de un Modelo de Machine Learning permitirá maximizar las ganancias por medio de optimizar el uso de asistencia a los teatros acorde a la opinión pública sobre las películas.
Técnicas y algoritmos a utilizar	Machine Learning Supervisado de Clasificación, utilizando los algoritmos de: KNN, *Decision Tree* y *Random Forest*

2. Entendimiento y preparación de los datos.

a. Perfilamiento

En los datos suministrados contamos con tres columnas, la primera un identificador único de tuit, la segunda el contenido del tuit y la tercera la etiqueta ‘positivo’ o ‘negativo’ sobre dicho tuit. Los datos, según el diccionario, contienen solo tuis del lenguaje español lo cual es importante para posterior análisis. Además, no contamos con más datos de quien realizó el tuit, el horario, los followers del tuitero, ni otros datos, por lo cual únicamente trabajaremos con los datos por medio de un Procesamiento de Lenguaje Natural.

b. Análisis de la calidad de los datos

La calidad de los datos la analizamos con búsqueda de factores como, nulos, repetidos y datos no válidos. En este caso no encontramos esto, ya que todos los tuits eran únicos, no habían identificadores repetidos, la columna de sentimientos siempre era

‘positivo’ o ‘negativo’, y el contenido de los tuits es del tipo esperado. Entonces la calidad es buena respecto a lo definido en el diccionario. Sin embargo, sería ideal tener más información para poder saber el peso de dispersión de la opinión y como un tuit, por medio de interacciones de likes o replies, pueden demostrar que tan representativa es la opinión, y connotación, de un tuit.

c. Tratamiento de los datos

Al contar con un extenso string de texto como datos es necesario vectorizarlo para posteriormente realizar una clasificación. Como opciones de algoritmos de vectorización está Bag of Words y TF-DIF, seleccionamos Bow sobre TF-DIF ya que utilizamos ambos métodos en una primera instancia y comparamos las métricas cuantitativas con un Random Forest. Entonces encontramos un mejor desempeño en Bow el cual fue implementado. Además, para los 3 algoritmos con Bow solo utilizamos palabras con mínimo 2 instancias, pues de esta manera se busca evitar problemas de overfitting. También transformamos la columna de ‘sentimiento’ para que positivo fuera 1 y negativo 0, así contando con una columna binaria que representa lo mismo y es utilizable (a diferencia de la de texto) en todos los modelos a utilizar. Finalmente seleccionamos los modelos de clasificación de: KNN, Decision Tree y Random Forest, con el análisis cuantitativo de Precision, Recall y F1.

3. Modelado y evaluación

El primer modelo utilizado fue K-nearest neighbor, bajo el uso de 3 vecinos. Este se basa en la similitud entre las instancias del conjunto de datos. Los resultados del modelo KNN muestran un desempeño aceptable, aunque con una precisión y recall poco ideales. En especial se demuestra un desempeño significativamente peor en el test a comparación del train, demostrando un problema de overfitting de datos.

La elección de probar este modelo se basa en la evaluación exhaustiva de varios modelos y en la selección de aquel que tenga el menor tiempo de utilización, para asegurar que el modelo pueda ser escalable en un caso de aplicación a una mayor base de datos. Es importante tener en cuenta que cada algoritmo tiene sus propias ventajas y desventajas, entonces la selección del modelo se realizará al final de la prueba de los 3.

El segundo modelo fue Decision Tree, para el cual se utilizaron en pruebas 100 árboles con una profundidad promedio de 152. El modelo se ha evaluado en el conjunto de pruebas utilizando las mismas medidas que el caso previo, donde se resalta una mejor

desempeño que el KNN y resultados superiores al 80% en todas las medidas de interés.

Es decir que los resultados obtenidos muestran que el modelo Decision Tree tiene un desempeño bastante bueno. En especial el valor de F1, que combina precisión y recall, también es alto, lo que indica que el modelo tiene un buen equilibrio entre ambas medidas y no posee problemas de overfitting.

Finalmente, se utilizó “Random Forest” pues después de utilizar los árboles de decisión y tener tan buenos resultados, se nos hizo la mejor opción. Entonces, para hacer un análisis más robusto utilizamos Random Forest que implementa “Bootstrapping” y “Aggregation” lo cuál simplemente significa que utilizamos muchos parámetros y bastantes árboles al mismo tiempo para hacer una predicción más acertada. En este caso, utilizamos 50, 100 y 200 árboles con profundidades de 4, 5 y 6 utilizando los criterios de gini y entropía. Para que el algoritmo fuera efectivo, sólo revisamos alrededor del 50% de las combinaciones totales (lo cuál tomó al algoritmo cerca de 3 minutos).

Comparándolo con el Árbol de decisión, nos damos cuenta que las diferencias entre los resultados son no muy notables. Mientras que el árbol tiene una precisión 4% mejor, el “Random Forest” cuenta con un recall 4% más alto. La decisión depende realmente de qué necesite el negocio, si su prioridad es minimizar los falsos positivos, entonces deberíamos optar con el Árbol de decisión si lo importante es buscar más casos positivos, entonces deberíamos utilizar el “Random Forest”.

A nivel de la evaluación cuantitativa del modelo deben trabajar con el grupo de expertos en temas de estadística que le fue asignado (ver en el Excel de registro de grupos, los datos del grupo asignado) y mostrar evidencia de las mejoras realizadas gracias a ese trabajo interdisciplinario.

4. Resultados

Descripción de los resultados obtenidos, que permita a la organización comprenderlos, haciendo énfasis en el análisis de las medidas arrojadas por los modelos utilizados y cómo aportan en la consecución de los objetivos del negocio. Incluir posibles estrategias que la organización debe plantear relacionadas con los resultados obtenidos en los modelos y una justificación de por qué esa información es útil para ellos. Este resultado debe estar en el documento y deben generar un video de máximo 7 minutos explicando su proyecto y resultados. El video debe ser publicado en el padlet respectivo (ver

condiciones de entrega). En este punto, deben trabajar con el grupo de expertos en temas de estadística que le fue asignado (ver en el Excel de registro de grupos, los datos del grupo asignado) y mostrar evidencia de las mejoras realizadas gracias a ese trabajo interdisciplinario.

Finalmente, optamos por utilizar el modelo del “Random Forest” ya que, para una empresa cuyo negocio es proyectar películas es preferible tener varias funciones de una película muy “buena” arriesgándose a poner varias funciones de una película que no es tan popular. A perder las boletas de aquellas personas que querían ir a ver una película que la han recomendado mucho pero las 2 funciones en la que la presentaban estaban ya vendidas. Tenemos que tener en cuenta que el costo operacional de una función es muy bajo, una empresa como CineColombia podría tener una sala de cine corriendo películas todas las horas del día y lo único que tendría que pagar sería la luz y la limpieza de la sala. Realmente a las empresas de cine no les importa tener funciones de más, a lo mejor con el 30% de la sala vendida ya se cubre el costo que esto implica. En cambio, si una película muy popular tiene pocas funciones, la empresa pierde mucho dinero, ya que se venderían menos boletas por tener las pocas funciones que existen, llenas.

Por las razones expuestas previamente, nuestro Líder de negocio decidió elegir el modelo de “Random Forest”, porque aunque pueda predecir que cierta película es popular cuando no lo es (un 22% de probabilidad), el modelo predecirá más películas que merecen tener más funciones, algo que le beneficia mucho más a la compañía de cine. En general, ambos modelos tienen muy buenas métricas, (y realmente un 4% de diferencia no es mucho) así que utilizar cualquiera de los dos sería una buena idea, pero si alguna vez ese 4% hace la diferencia, será una ganancia muy importante para la compañía, ya que una función llena es bastante dinero, que no es comparable con los costos de tener una función casi vacía.

5. Trabajo en Equipo

a. Roles

Integrante	**Rol**
Juan Felipe Patiño	Líder de Proyecto, Líder de Analítica
Juan Ricardo Diaz	Líder de Negocio, Líder de Datos

Juan Felipe Patiño, como Líder de Proyecto, estuvo a cargo de coordinar las fechas de reuniones y de trabajo, llenar los formatos de pre-entrega y dividir el trabajo a realizarse de manera equitativa y justa. Además, estuvo pendiente de los avances constantes y notificó al grupo sobre los detalles de la entrega. Finalmente en este rol fue el que subió y entregó todos los entregables pertinentes. Como líder de analítica Juan Felipe fue quien se aseguró de identificar el mejor modelo bajo los supuestos y restricciones de negocio, y también revisar los desarrollos analíticos y su coherencia.

Por otro lado Juan Ricardo, como líder de negocio fue responsable por indagar de la temática y encontrar la estrategia para el negocio. Teniendo en cuenta el uso de un lenguaje enfocado al cliente, y crear el puente de técnico a negocio. Como líder de datos Juan Ricardo gestionó los datos y realizó la identificación, limpieza, y organización pertinente para el desarrollo.

Con los roles de cada uno asignados y desarrollados, cada integrante se planteó utilizar una misma cantidad de tiempo individual y un tiempo grupal de reuniones y de apoyo mutuo. Este tiempo individual de desarrollo fue menor a 5 horas cada uno, dependiendo ya de la velocidad de desarrollo y condiciones del software y su entorno, mientras que de manera grupal es tuvieron múltiples reuniones, alrededor de 3 formales y múltiples informales, de cada una alrededor de 1 hora

Entre los retos principales fue el tiempo disponible, ya que si bien se tuvo el enunciado disponible con antelación el conocimiento necesario para desarrollar fue después y además transcurre durante épocas de poco tiempo libre para fines académicos. Entonces el reto de obtener el tiempo para desarrollar la actividad de manera completa y de buen nivel fue un gran reto. Durante el transcurso de la primera etapa se llevaron a cabo múltiples reuniones, donde vale la pena recalcar una primera de decisión de tema, una segunda de división de roles, dos reuniones posteriores de avances y retroalimentación mutua y una reunión final para finalización de entregables y posterior entrega.

b. División de puntuación

La división de puntos, acorde a la planeación y realización del proyecto se realizará de manera equitativa. Si bien en algunos tramos un integrante había avanzado más que el otro, analizando la situación como una suma del total el trabajo fue equitativo entre ambos. Es por esto que dividiremos los puntos 50-50.