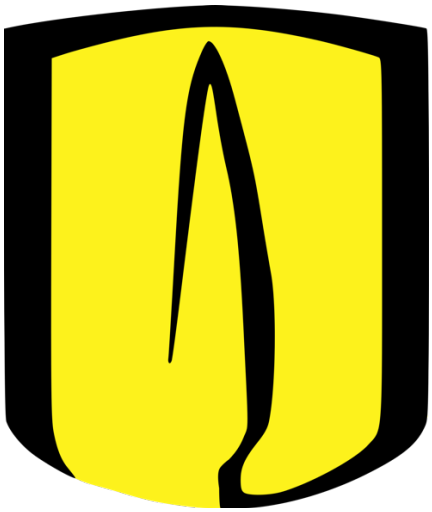
Documento – Proyecto Analítica de textos – Etapa 1



Grupo G30: Tamarindo

Jesús Jiménez – 202020431
Juan Camilo Bonet – 202022466
Thais Tamaio – 202022213

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

Tabla de contenidos

1. Entendimiento del negocio y enfoque analítico

- **Objetivos y criterios de éxito desde el punto de vista del negocio**

  Primero, es pertinente mencionar que se asume que la empresa a la que irá enfocado el proyecto corresponde a una plataforma de streaming de películas en español, en donde los usuarios pueden crear reseñas de las películas que han visto. El objetivo principal es clasificar estas reseñas como positivas o negativas para mejorar la calidad de las recomendaciones de películas.

  Algunos objetivos de negocio adicionales podrían ser:

  - Mejorar la retención de los usuarios: Al proporcionar recomendaciones más precisas y relevantes, se puede aumentar la satisfacción de los usuarios y su compromiso con la plataforma.
  - Incrementar la cantidad de suscripciones: Si los usuarios encuentran recomendaciones útiles y personalizadas, es más probable que decidan suscribirse a la plataforma para acceder a más contenido.
  - Aumentar la satisfacción del cliente: Al ofrecer recomendaciones precisas y personalizadas, se puede mejorar la satisfacción del cliente y disminuir las posibilidades de que cancelen su suscripción.

  Los criterios de éxito para estos objetivos podrían ser:

  - Recall: Es importante tener un alto Recall en este proyecto, ya que se quiere asegurarse de que no se clasifiquen reseñas negativas como positivas. Es decir, se busca minimizar los falsos negativos, es decir, las reseñas que en realidad son negativas, pero son clasificadas como positivas.
  - Precisión: Es importante tener una alta precisión en este proyecto, ya que se quiere asegurarse de que no se clasifiquen reseñas positivas como negativas. Es decir, se busca minimizar los falsos positivos, es decir, las reseñas que en realidad son positivas, pero son clasificadas como negativas.
  - F1: La medida F1 es importante ya que combina tanto la precisión como el Recall, permitiendo obtener una medida general del desempeño del modelo. Por lo tanto, una alta F1 indica un buen equilibrio entre la precisión y el Recall, lo que significa que el modelo está clasificando de manera efectiva tanto las reseñas positivas como las negativas.

  En resumen, el proyecto puede ayudar a la plataforma a mejorar su eficacia en la recomendación de películas y a satisfacer mejor las necesidades y preferencias de sus usuarios.

- **Determinación del enfoque analítico el proyecto**

Un enfoque analítico que podría ayudar a alcanzar los objetivos del negocio será la implementación de un modelo de análisis de sentimientos que permita clasificar las reseñas de las películas como positivas o negativas de manera automatizada. Esto se realizará utilizando técnicas de procesamiento de lenguaje natural y aprendizaje automático para entrenar el modelo con un conjunto de datos de reseñas etiquetadas. Una vez implementado el modelo, se utilizarán diversas métricas de evaluación para medir su precisión y ajustarlo en caso de ser necesario.

Es decir, que se busca implementar un modelo de análisis de sentimientos para mejorar la calidad de las recomendaciones de películas en la plataforma de streaming de películas en español y, por lo tanto, lograr los objetivos de negocio adicionales mencionados anteriormente, como mejorar la retención de los usuarios, incrementar la cantidad de suscripciones y aumentar la satisfacción del cliente.

Finalmente, se busca encontrar un mejor modelo, esto es importante porque a medida que se prueban diferentes enfoques, se pueden identificar aquellos que brinden los mejores resultados y permitan la mejor toma de decisiones en el negocio.

- **Requerimientos del negocio**

Oportunidad/problema del negocio	La plataforma de streaming de películas en español busca mejorar la calidad de las recomendaciones de películas y aumentar la retención y suscripciones de los usuarios mediante la clasificación automatizada de reseñas como positivas o negativas.
Enfoque analítico	Implementación de un modelo de análisis de sentimientos utilizando técnicas de procesamiento de lenguaje natural y aprendizaje automático para entrenar el modelo con un conjunto de datos de reseñas etiquetadas.
Organización y rol que se beneficia con la oportunidad definida	La plataforma de streaming de películas en español se beneficiaría directamente al mejorar la calidad de las recomendaciones, aumentar la retención y suscripciones de los usuarios, y mejorar la satisfacción del cliente. Los usuarios también se beneficiarían al recibir recomendaciones más precisas y relevantes.

Técnicas para utilizar	Bag of Words (BoW), TF-IDF y HashingVectorizer para procesamiento de texto.
Algoritmos para utilizar	Clasificación con Random Forest, árboles de decisión y Gradient Boosting.

2. Entendimiento y preparación de los datos

- **Perfilamiento y análisis de datos**

  Primero, el DataFrame con los datos suministrados de *MovieReviews* contiene 5000 filas y 3 columnas: id, review_es y sentimiento. El sentimiento de una review puede ser clasificado en dos: positivo o negativo y este corresponde a la sensación que genera un review. A su vez, se tiene que para el dataset entregado el 50% de las reviews son positivas y el otro 50% son negativas, es decir, 2500 reviews.

  Por otra parte, se observa que las reviews en su mayoría, un 96%, son en español, y el 4% restante corresponde a inglés e indonesio. Por último, se observa que no existen valores duplicados ni nulos en todo el dataset.

- **Tratamiento de los datos**

  Se realizaron varias modificaciones al DataFrame original. Primero, se eliminan las reviews en inglés e indonesio, ya que en el análisis sólo nos importan aquellas que son en español. Luego, se modifica la variable *sentimiento* de categórica a numérica, ya que los algoritmos a utilizar sólo trabajan con datos numéricos.

  Por otra parte, se eliminan símbolos de puntuación o comillas de las reviews para evitar que los caracteres de puntuación interfieran en la interpretación de las palabras por el algoritmo. A su vez, también se eliminan caracteres que no sean ASCII, se pasan todos los caracteres a minúsculas y se transforman los números en su representación textual en español. Además, se eliminan stopwords y palabras que no sean en español. El resultado final es una versión limpia y preprocesada del texto que puede ser utilizada para análisis posteriores.

3. Elección de algoritmos

Se escogieron tres técnicas para la evaluación del proyecto y a partir de estas se crearon tres modelos. Estos fueron los algoritmos elegidos:

- **Random Forest**

Random Forest es un algoritmo de aprendizaje supervisado que se utiliza tanto para la clasificación como para la regresión. Este algoritmo se basa en la combinación de múltiples árboles de decisión, donde cada árbol se entrena con una submuestra aleatoria del conjunto de datos de entrenamiento. Durante la predicción, el algoritmo promedia las predicciones de todos los árboles de decisión para obtener la salida final.

- **GradientBoostingClassifier**

  El algoritmo Gradient Boosting es una técnica de aprendizaje automático que se utiliza para mejorar la precisión de un modelo de clasificación. Es un enfoque de conjunto (ensemble) que combina varios modelos de aprendizaje débil en un modelo de aprendizaje fuerte. En cada iteración, el algoritmo construye un nuevo modelo débil que se enfoca en los casos que el modelo anterior no ha clasificado correctamente, y agrega ese modelo al conjunto de modelos débiles ya existentes.

- **Árboles de decisión**

  El algoritmo de árboles de decisión es un método de aprendizaje supervisado utilizado para clasificar datos en categorías o predecir valores numéricos. En este algoritmo, se construye un árbol de decisiones a partir de los datos de entrenamiento, donde cada nodo del árbol representa una característica del conjunto de datos y las ramas representan las posibles respuestas a esa característica. El árbol se construye de forma recursiva, dividiendo el conjunto de datos en subconjuntos más pequeños y homogéneos según ciertas características hasta alcanzar una determinada profundidad o un criterio de parada.

4. Modelado y evaluación

Luego del análisis realizado, se determinó que el algoritmo con el mejor resultado fue RandomForest, por lo que la explicación de las técnicas de modelamiento se basó en este algoritmo:

- **BoW – Implementado por Juan Camilo Bonet**

  BoW es una técnica de procesamiento de lenguaje natural que se utiliza para representar un texto como un conjunto de palabras sin considerar su orden o estructura. En este enfoque, se crea un diccionario de todas las palabras únicas en un conjunto de datos y luego se crea una matriz que representa la frecuencia de cada palabra en cada documento.

  Esta técnica fue seleccionada dado a que es muy útil cuando se desea realizar una clasificación basada en la presencia o ausencia de palabras en los textos, por

lo que es adecuada para el análisis de sentimientos en el que se busca identificar las palabras más comunes en las reseñas positivas y negativas.

Para este modelo, se creó un objeto CountVectorizer que tokeniza el texto, elimina las palabras vacías y transforma el texto en una matriz de conteo de palabras. Luego, se ajustó y transformó la matriz de características con los datos de entrenamiento, se entrena un modelo RandomForestClassifier con esta matriz de características y se utiliza la importancia de las características para visualizar cuáles son las palabras más importantes en la clasificación. Finalmente, se realizó una predicción tanto en los datos de entrenamiento como en los de prueba.

- **TF-IDF – Implementado por Jesús Jiménez**

  TF-IDF es una técnica de procesamiento de lenguaje natural que se utiliza para evaluar la importancia de una palabra en un documento. TF-IDF considera tanto la frecuencia de una palabra en un documento como la frecuencia de la misma palabra en todo el corpus, lo que ayuda a reducir la importancia de palabras comunes.

  Esta técnica fue seleccionada dado a que es muy útil para evaluar la importancia relativa de cada palabra en el texto en función de su frecuencia en el documento y en la colección de documentos. Esta técnica es útil cuando se desea que el modelo sea sensible a las palabras que son más importantes para distinguir entre las clases de interés. En el contexto de este proyecto, es importante que el modelo sea sensible a las palabras que indican si una reseña es positiva o negativa para mejorar la precisión de las recomendaciones.

  Se utilizó la implementación de TF-IDF que consistió en crear un objeto TfidfVectorizer con un tokenizer y lista de stop words específicos, para luego ajustarlo al conjunto de entrenamiento. Posteriormente, se utilizó un modelo de clasificación RandomForestClassifier y se ajustó a los datos de entrenamiento generados por TF-IDF. Se visualizó la importancia de las características y se inspeccionó el número y profundidad de los árboles del modelo. Finalmente, se realizaron predicciones tanto en el conjunto de entrenamiento.

- **HashingVectorizer – Implementado por Thais Tamaio**

  HashingVectorizer es una técnica de procesamiento de lenguaje natural que se utiliza para transformar un conjunto de datos en una matriz de características. En lugar de crear un diccionario de palabras únicas como en BoW, HashingVectorizer utiliza una función de hash para asignar cada palabra a una posición en la matriz. Esto permite una representación de texto más eficiente, ya que no se necesita almacenar el diccionario completo de palabras únicas.

  Esta técnica fue utilizada dado a que utiliza una función hash para convertir cada palabra en un número entero único, y luego transforma los textos en vectores de

frecuencias de estos números enteros. Esta técnica es útil cuando se desea reducir el tamaño del vocabulario y el espacio de almacenamiento requerido para los vectores de características. Es decir que esta técnica es apropiada para el contexto de este proyecto, dado a que se busca reducir el espacio de almacenamiento requerido y acelerar el tiempo de procesamiento, ya que se trabaja con un gran conjunto de datos de reseñas de películas.

En la implementación, se utilizó la técnica HashingVectorizer para vectorizar el texto de las reseñas de películas y convertirlo en una representación numérica que se puede utilizar para entrenar el modelo de clasificación de análisis de sentimientos. Se seleccionó un número de características de 2^16 para la vectorización. Luego, se ajustó el modelo RandomForestClassifier a los datos vectorizados y se utilizó el gráfico de importancia de características para ver las características más importantes del modelo. También se evaluó la profundidad media de los árboles del modelo y se utilizaron las predicciones del modelo para clasificar tanto los datos de entrenamiento como los de prueba.

5. Resultados

- **BoW – Implementado por Juan Camilo Bonet**

    - Métricas de desempeño con datos de entrenamiento

      Las métricas de evaluación de desempeño del modelo BoW entrenado con los datos de entrenamiento muestra que la precisión, el Recall y F1 son todas igual a 1. En este caso, lo que sugiere que el modelo tiene un desempeño perfecto en la clasificación de las reseñas como positivas o negativas.

    - Métricas de desempeño con datos de prueba

Precisión	Recall	F1
0.830	0.838	0.834

      En este caso, los valores obtenidos indican que el modelo tiene una precisión del 83.1%, lo que significa que el 83.1% de las predicciones positivas son correctas. El Recall es del 83.75%, lo que indica que el modelo identifica correctamente el 83.75% de los casos positivos. El F1-score es del 83.4%, lo que sugiere que el modelo tiene un buen equilibrio entre la precisión y el Recall. En general, estos resultados son prometedores y sugieren que el modelo es capaz de clasificar correctamente las reseñas como positivas o negativas.

    - Validación cruzada

El valor promedio de la puntuación de validación cruzada es de 0.813, lo que sugiere que el modelo tiene un buen rendimiento en la clasificación de las reseñas.

- **TF-IDF – Implementado por Jesús Jiménez**

  - Métricas de desempeño con datos de entrenamiento

    Las métricas de evaluación de desempeño del modelo TF-IDF entrenado con los datos de entrenamiento muestra que la precisión, el Recall y F1 son todas igual a 1. En este caso, lo que sugiere que el modelo tiene un desempeño perfecto en la clasificación de las reseñas como positivas o negativas.

  - Métricas de desempeño con datos de prueba

Precisión	Recall	F1
0.853	0.800	0.826

    En este caso, los valores obtenidos indican que el modelo tiene una precisión del 85.3%, lo que significa que el 85.3% de las predicciones positivas son correctas. El Recall es del 80%, lo que indica que el modelo identifica correctamente el 80% de los casos positivos. El F1-score es del 82.6%, lo que sugiere que el modelo tiene un buen equilibrio entre la precisión y el Recall. En general, estos resultados son prometedores y sugieren que el modelo es capaz de clasificar correctamente las reseñas como positivas o negativas.

  - Validación cruzada

    El valor promedio de la puntuación de validación cruzada es de 0.805, lo que sugiere que el modelo también tiene un buen rendimiento en la clasificación de las reseñas, aunque ligeramente inferior al del modelo BoW.

- **HashingVectorizer – Implementado por Thais Tamaio**

  - Métricas de desempeño con datos de entrenamiento

    Las métricas de evaluación de desempeño del modelo HashingVectorizer entrenado con los datos de entrenamiento muestra que la precisión, el Recall y F1 son todas igual a 1. En este caso, lo que sugiere que el modelo tiene un desempeño perfecto en la clasificación de las reseñas como positivas o negativas.

  - Métricas de desempeño con datos de prueba

Precisión	Recall	F1
0.856	0.815	0.835

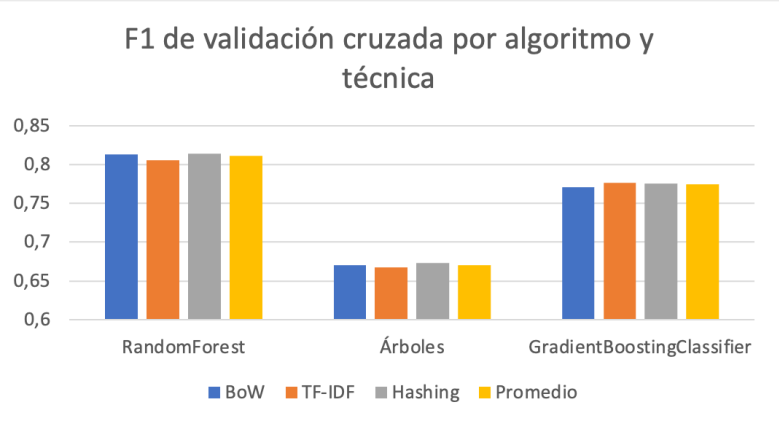
En este caso, la precisión es del 85,55%, lo que significa que de todas las reseñas que el modelo clasificó como positivas, el 85,55% realmente lo son. El Recall es del 81,45%, lo que significa que de todas las reseñas positivas en el conjunto de prueba, el modelo identificó el 81,45%. El F1 es del 83,46%, lo que es una medida combinada de precisión y Recall. En general, estos resultados indican que el modelo tiene un buen desempeño en la clasificación de reseñas de películas como positivas o negativas.

- Validación cruzada

  El valor promedio de la puntuación de validación cruzada es de 0.813, lo que sugiere que el modelo tiene un rendimiento similar al del modelo BoW en la clasificación de las reseñas.

- **Comparación de resultados, conclusiones y recomendaciones**

Para cada una de las técnicas de procesamiento de texto utilizadas (BoW, TF-IDF, y HashingVectorizer), se obtuvieron los mejores resultados con el algoritmo RandomForest. Esto se puede observar en la siguiente gráfica, en el que se compararon los distintos algoritmos con cada técnica de modelamiento presentada y RandomForest siempre arrojó las mejores métricas:



Random forest es un algoritmo de ensamblado de árboles de decisión que combina múltiples árboles de decisión individuales para mejorar la precisión y la generalización del modelo. En comparación con un árbol de decisión único, un bosque aleatorio crea múltiples árboles de decisión a partir de muestras aleatorias de los datos de entrenamiento y características aleatorias de esas muestras, lo que reduce la probabilidad de sobreajuste y mejora la generalización del modelo.

Los resultados de la validación cruzada indican que los tres modelos tienen un rendimiento similar en términos de precisión, sin embargo, el modelo HashingVectorizer obtuvo el valor más alto en Recall y F1, lo que indica que es capaz de identificar correctamente más casos positivos. Por lo tanto, basándonos en las métricas de evaluación y la validación cruzada, se podría seleccionar el modelo HashingVectorizer como la mejor opción para clasificar los datos. El modelo de clasificación con HashingVectorizer tiene un buen desempeño en términos de precisión, Recall y F1, lo que indica un buen equilibrio entre precisión y Recall. Además, al realizar la validación cruzada, se obtuvo un valor promedio de 0.8131, lo que indica que el modelo tiene un buen desempeño en la generalización a nuevos datos.

En general, estas métricas indican que el modelo de clasificación usando RandomForest y HashingVectorizer es una buena opción para resolver este problema en particular y podría ser una herramienta valiosa para un negocio que busca recomendar películas a sus usuarios y mejorar la calidad de su catálogo de películas.

Por ejemplo, si un usuario ha mostrado una tendencia a disfrutar películas con reseñas positivas similares a las de otras películas que haya visto en el pasado, el modelo podría recomendarle otras películas con reseñas positivas similares a las de las películas que ha disfrutado previamente. Por otro lado, si un usuario ha dado críticas negativas a ciertas películas, el modelo podría sugerirle otras películas con reseñas negativas similares a las de las películas que ha mostrado que no le gustan.

Asimismo, el modelo puede ser útil para identificar películas que no son populares o que no son bien recibidas por el público en general. Si un gran número de reseñas para una película en particular son negativas, el modelo podría sugerir que esa película no se recomiende a los usuarios de la plataforma.

Es importante considerar que el modelo se entrena con datos históricos y puede estar sujeto a cambios en los patrones de los usuarios y en los gustos del mercado. Sería necesario monitorear y actualizar regularmente el modelo para asegurarse de que siga siendo relevante y preciso.

Si la empresa tiene acceso a más información sobre los usuarios (como edad, género, ubicación, etc.), se podrían incorporar como características adicionales al modelo para mejorar su precisión.

Finalmente, se debe tener en cuenta que el modelo solo predice si un review es positivo o negativo. Si la empresa desea proporcionar recomendaciones más personalizadas, podría considerar técnicas de filtrado colaborativo o contenido basado en el modelo, que se basan en los datos de usuario y no solo en el contenido del review.

6. *Trabajo en equipo*

- **Planeación y división de trabajo.**

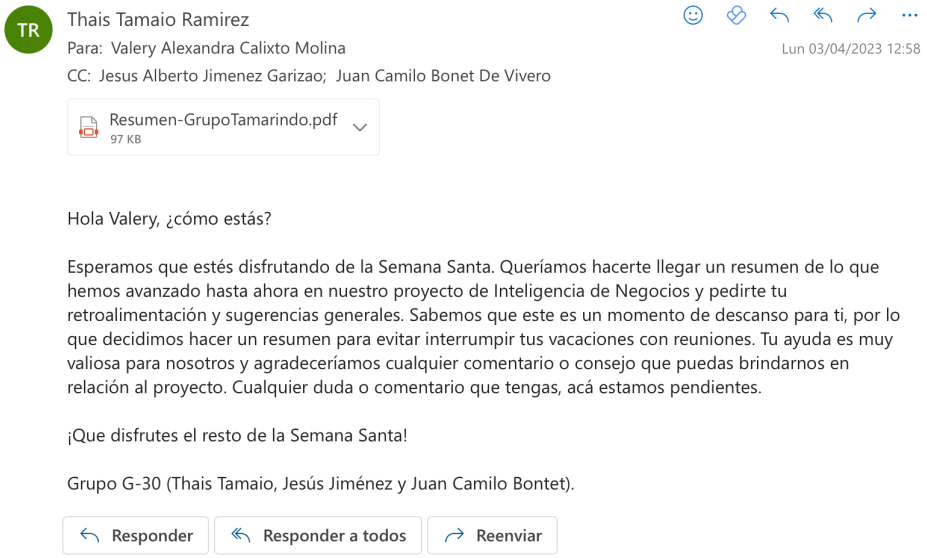
  Durante la primera reunión decidimos como se iba a dividir el trabajo y cuál sería el cronograma. Decidimos que cada integrante estaba encargado de implementar una técnica de procesamiento de texto dentro de cada algoritmo y además tenían un trabajo extra. Thais estaba encargada del perfilamiento, preparación y tratamiento de los datos, Jesús estaba encargado de las conclusiones y recomendaciones y Juan Camilo estaba encargado de documentar el enfoque analítico y la división de trabajo. Además, decidimos que íbamos a tener dos reuniones adicionales a la inicial. La primera de estas seria cuando Thais estaba lista con la preparación de los datos, y en esta reunión íbamos a decidir los algoritmos a implementar. La última reunión seria para finalizar el documento y grabar el video.

  Durante la última reunión, hablamos sobre como fue el trabajo en equipo y como se repartirían los 100 puntos entre los integrantes. Consideramos que todos aportamos una cantidad igual al proyecto y acordamos en la siguiente división de puntos:

  Thais: 33.3 puntos Jesús: 33.3 puntos Juan Camilo: 33.3 puntos

- **Retroalimentación de estudiante de estadística**

  Con respecto a la estudiante de estadística, le mandamos un correo para recibir su retroalimentación sobre el trabajo hecho, pero hasta el momento no hemos recibido una respuesta. Esperamos poder implementar sus recomendaciones para la próxima etapa del proyecto.


TR

Thais Tamaio Ramirez

Para: Valery Alexandra Calixto Molina

CC: Jesus Alberto Jimenez Garizao; Juan Camilo Bonet De Vivero

Lun 03/04/2023 12:58

Resumen-GrupoTamarindo.pdf
97 KB

Hola Valery, ¿cómo estás?

Esperamos que estés disfrutando de la Semana Santa. Queríamos hacerte llegar un resumen de lo que hemos avanzado hasta ahora en nuestro proyecto de Inteligencia de Negocios y pedirte tu retroalimentación y sugerencias generales. Sabemos que este es un momento de descanso para ti, por lo que decidimos hacer un resumen para evitar interrumpir tus vacaciones con reuniones. Tu ayuda es muy valiosa para nosotros y agradeceríamos cualquier comentario o consejo que puedas brindarnos en relación al proyecto. Cualquier duda o comentario que tengas, acá estamos pendientes.

¡Que disfrutes el resto de la Semana Santa!

Grupo G-30 (Thais Tamaio, Jesús Jiménez y Juan Camilo Bontet).

Responder Responder a todos Reenviar