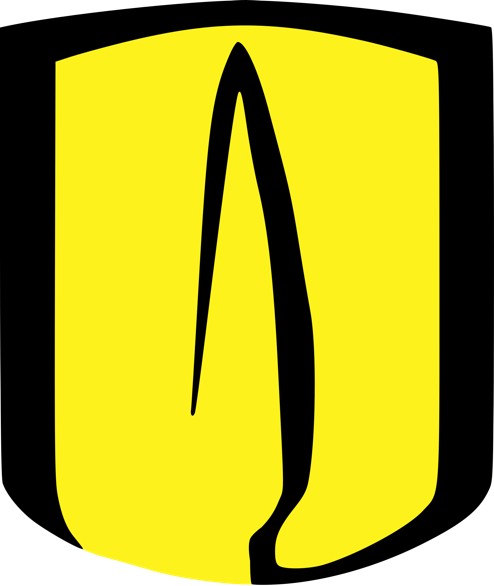
**Documento – Proyecto Analítica de textos – Etapa 1**

****

**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 2](#_Toc131387181)

[ Objetivos y criterios de éxito desde el punto de vista del negocio 2](#_Toc131387182)

[ Determinación del enfoque analítico el proyecto 3](#_Toc131387183)

[ Requerimientos del negocio 4](#_Toc131387184)

[2. Entendimiento y preparación de los datos 4](#_Toc131387185)

[ Perfilamiento y análisis de datos 4](#_Toc131387186)

[ Tratamiento de los datos 4](#_Toc131387187)

[3. Modelado y evaluación 5](#_Toc131387188)

[ BoW – Implementado por Juan Camilo Bonet 5](#_Toc131387189)

[ TF-IDF – Implementado por Jesús Jiménez 5](#_Toc131387190)

[ HashingVectorizer – Implementado por Thais Tamaio 6](#_Toc131387191)

[4. Resultados 6](#_Toc131387192)

[ BoW – Implementado por Juan Camilo Bonet 6](#_Toc131387193)

[o Métricas de desempeño con datos de entrenamiento: 6](#_Toc131387194)

[o Métricas de desempeño con datos de prueba: 7](#_Toc131387195)

[o Validación cruzada: 7](#_Toc131387196)

[ TF-IDF – Implementado por Jesús Jiménez 7](#_Toc131387197)

[o Métricas de desempeño con datos de entrenamiento: 7](#_Toc131387198)

[o Métricas de desempeño con datos de prueba: 7](#_Toc131387199)

[o Validación cruzada: 8](#_Toc131387200)

[ HashingVectorizer – Implementado por Thais Tamaio 8](#_Toc131387201)

[o Métricas de desempeño con datos de entrenamiento: 8](#_Toc131387202)

[o Métricas de desempeño con datos de prueba: 8](#_Toc131387203)

[o Validación cruzada: 9](#_Toc131387204)

[5. Trabajo en equipo 9](#_Toc131387205)

# ***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:

* + Exactitud de las recomendaciones: Medido a través de la precisión en la clasificación de las reseñas como positivas o negativas.
  + Incremento de la retención de usuarios: Medido a través de la tasa de retención de los usuarios y la frecuencia de uso de la plataforma.
  + Incremento de las suscripciones: Medido a través del número de nuevas suscripciones y la tasa de conversión de usuarios no suscritos en suscriptores.
  + Mejora de la satisfacción del cliente: Medido a través de la retroalimentación de los usuarios, como el nivel de satisfacción expresado en encuestas o comentarios.

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 y algoritmos para utilizar** | Clasificación con Random Forest, utilizando técnicas de Bag of Words (BoW), TF-IDF y HashingVectorizer para procesamiento de texto. |

# ***Entendimiento y preparación de los datos***

* **Perfilamiento y análisis de datos**
* **Tratamiento de los datos**

# ***Modelado y evaluación***

* **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.

Este algoritmo fue seleccionado dado a que corresponde a técnica 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.

Este algoritmo fue seleccionado dado a que corresponde a una técnica 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.

Este algoritmo fue seleccionado dado a que corresponde a una técnica 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 este algoritmo es apropiado 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.

# ***Resultados***

* **BoW – Implementado por Juan Camilo Bonet**
  + Métricas de desempeño con datos de entrenamiento

|  |  |  |
| --- | --- | --- |
| **Precisión** | **Recall** | **F1** |
| 1 | 1 | 1 |

En esta parte del código se están mostrando las métricas de evaluación de desempeño del modelo BoW entrenado con los datos de entrenamiento. Las métricas que se están calculando son precisión, Recall y F1. En este caso, los valores que se muestran son 1.0 para todas las métricas, 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

|  |  |  |
| --- | --- | --- |
| **Precisión** | **Recall** | **F1** |
| 1 | 1 | 1 |

En esta parte del código se están mostrando las métricas de evaluación de desempeño del modelo TF-IDF entrenado con los datos de entrenamiento. Las métricas que se están calculando son precisión, Recall y F1. En este caso, los valores que se muestran son 1.0 para todas las métricas, 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

|  |  |  |
| --- | --- | --- |
| **Precisión** | **Recall** | **F1** |
| 1 | 1 | 1 |

En esta parte del código se están mostrando las métricas de evaluación de desempeño del modelo HashingVectorizer entrenado con los datos de entrenamiento. Las métricas que se están calculando son precisión, Recall y F1. En este caso, los valores que se muestran son 1.0 para todas las métricas, 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.

# ***Trabajo en equipo***