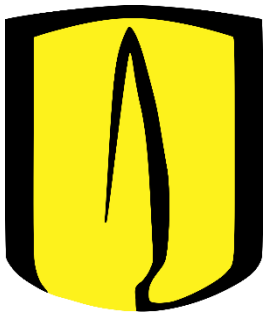
INFORME PROYECTO 01 – NLP



Juan Esteban Rodríguez Ospino
Wilton Esteban Martínez Hernández
Erich Gusseppe Soto Parada

Inteligencia de Negocios
Ingeniería de sistemas y Computación
Universidad de Los Andes
2023

Tabla de contenido

(10%) Entendimiento del negocio y enfoque analítico.

Definición de los objetivos y criterios de éxito desde el punto de vista del negocio. Determinación del enfoque analítico para alcanzar los objetivos del negocio. Descripción de cómo el requerimiento de negocio es resuelto por el enfoque analítico propuesto, para lo cual debe diligenciar la tabla que se presenta a continuación:

Como grupo hemos decidido que la temática que vamos a trabajar va a ser la de comentarios de películas.

Entendimiento del Negocio

Oportunidad/problema Negocio

Identificar el tipo de percepción (positiva) de las personas sobre una serie de películas. Lo anterior, ayudará a poder identificar que tanto una determinada película les gusta a los clientes. Por ejemplo, en servicios de streaming como Netflix se pueden identificar este tipo de comentarios y así determinar el éxito de la película en cuanto al número de clientes que les parece interesante la película (cabe a aclarar que Netflix no posee la funcionalidad de escribir comentarios sobre sus películas). Además, se puede implementar en sitios en diversas plataformas donde las personas pueden poseen de la funcionalidad de dejar comentarios sobre las

películas. Lo anterior, permitirá dar un feedback más preciso a los y por lo tanto las plataformas pueden brindar una calificación más precisa sobre las películas.

Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático**)**

El enfoque analítico para este proyecto implica la construcción de modelos de aprendizaje automático para el análisis de sentimientos en comentarios de películas en español. Se utilizarán técnicas de Procesamiento de Lenguaje Natural (NLP) para procesar los comentarios y extraer características relevantes que puedan ser utilizadas por los modelos. Además, se utilizarán los algoritmos de Bag of Words y TF-IDF para representar documentos de texto.

Organización y rol dentro de ella que se beneficia con la oportunidad definida

La entidad beneficiada es una organización dedica a la producción y reproducción de películas de terceros o propios. Específicamente, en las áreas de marketing y ventas, producción y servicio al cliente. Estas áreas se verán significativamente beneficiadas, esto debido a que en base a las predicciones realizadas por los algoritmos se puede llevar a cabo en campañas publicitarias en base a la percepción de las personas sobre las películas. Además, permite dar una feedback muy valioso para futuras producciones de películas. Por último, el cliente se verá enormemente beneficiado, pues en base a estos comentarios se podrán tomar medidas para refinar la experiencia de usuario.

Técnicas y algoritmos para utilizar

Técnicas

Bag of Words es una técnica que se basa en contar la frecuencia de las palabras en el texto y generar un vector de características. Por otro lado, TF-IDF es una técnica que mide la importancia de una palabra en un documento, mediante la frecuencia de aparición de la palabra en el documento y la frecuencia de aparición de la palabra en la colección de documentos. Ambos algoritmos son ampliamente utilizados en la construcción de modelos de análisis de sentimientos de películas.

Con Bag of Words se pueden representar los comentarios de las películas como vectores de características que contienen la frecuencia de las palabras en los comentarios. Mientras que TF-IDF se utiliza para medir la importancia de las palabras en los comentarios y ajustar la frecuencia de las palabras en función de su importancia. Esto ayuda a identificar las palabras que son más relevantes para la predicción del sentimiento positivo o negativo de un comentario.

Algoritmos

Se utilizarán algoritmos de clasificación supervisada de lenguaje natural para clasificar los comentarios como positivos o negativos en función de las características extraídas: RandomForestClassifier.

(25%) Entendimiento y preparación de los datos.

El conjunto de datos presentado posee tres features: Unnamed: 0, reviews_es y sentimiento.

Unnamed: 0: representa el id del review

reviews_es: comentario de la persona sobre una determinada película.

sentimiento: es un valor positivo o negativo, el cual se deduce de lo escrito por la persona que escribe el comentario.

Inicialmente, se eliminó la columna **Unnamed: 0**, si bien no tiene ningún afectó sobre el proceso por comida se eliminó. Luego, se removieron todos los caracteres no alfanuméricos: !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~. Posteriormente, se procedió a eliminar aquellas palabras que no aportan un significado relevante para el análisis de sentimiento, dado que se tienen tanto palabras en español como en inglés se descargaron stop Words de los idiomas. También, con la intención de tener una homogeneidad en las palabras. Es decir, evitar que dos palabras que se escriben de diferentes maneras se interpreten como diferentes, se pasaron todas las palabras a minúsculas.

Por otra parte, se observó que muchas palabras fueron mal escritas por los clientes o simplemente son nombres propios de algunas actores y actrices. Por ejemplo, para el caso de palabras mal escritas, algunas personas escribieron **carnajería** en lugar de **carnicería**, y así con muchos otros casos. Para lo anterior, se descargaron diccionarios de español e inglés para la identificación de estas palabras y proceder a eliminarlas. Además, se identificaron las palabras menos frecuentes y se eliminaron, pues su aporte no es relevante para la determinación del sentimiento del cliente asociado en el escrito. De igual manera, se procedió a eliminar las palabras más frecuentes, ya que según lo observado son palabras que nos relevantes para la determinación del sentimiento. Lista de palabras: **película, si, películas, solo, ser, historia, tan, ver, realmente, vez.** Por último, se procedió a eliminar las tildes de algunas palabras, lo anterior se hizo, pues logro ver que hay una mejora en las métricas de error para el algoritmo Random Forest cuando se remueven las tildes.

(30%) Modelado y evaluación

RandomForestClassifier – Wilton Esteban Martinez

En este caso, se decidió utiliza el algoritmo Random Forest, pues este permite la clasificación en diferentes categorías. Además, porque nos permite identificar patrones y relaciones en grandes conjuntos de datos. Por otra parte, este está diseñado para tratar con datos ruidosos, lo cual es especialmente para nuestro caso, ya que los comentarios tienen muchos valores que afectan la calidad del modelo.

Inicialmente, primero se vectoriza cada una de las palabras usando la técnica de Bag of words. Al mismo tiempo, se utiliza la técnica TF-IDF Term Frequency-Inverse Document Frequency. Posteriormente, se ejecuta el algoritmo de Random Forest para cada una de estas técnicas y se muestra la importancia de cada una de las palabras y sus respectivas métricas de error. Adicional, a lo anterior por cada una de las técnicas anteriormente mencionadas se implementó el algoritmo Random Forest utilizando la técnica de validación cruzada, pues los modelos originales mostraron overfitting.

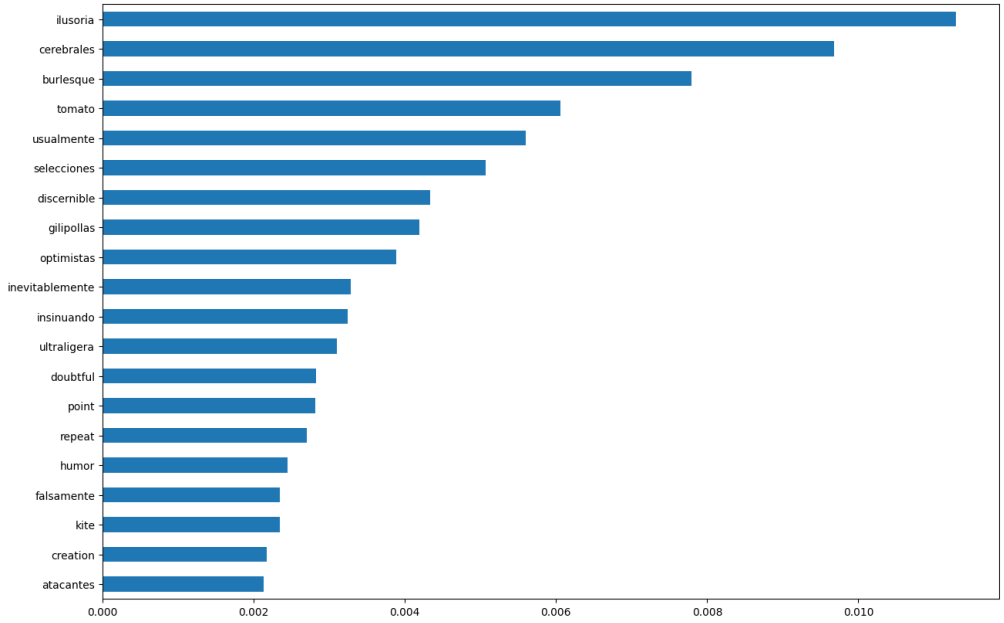


Ilustración 1. Variables importantes - Bag of Words.

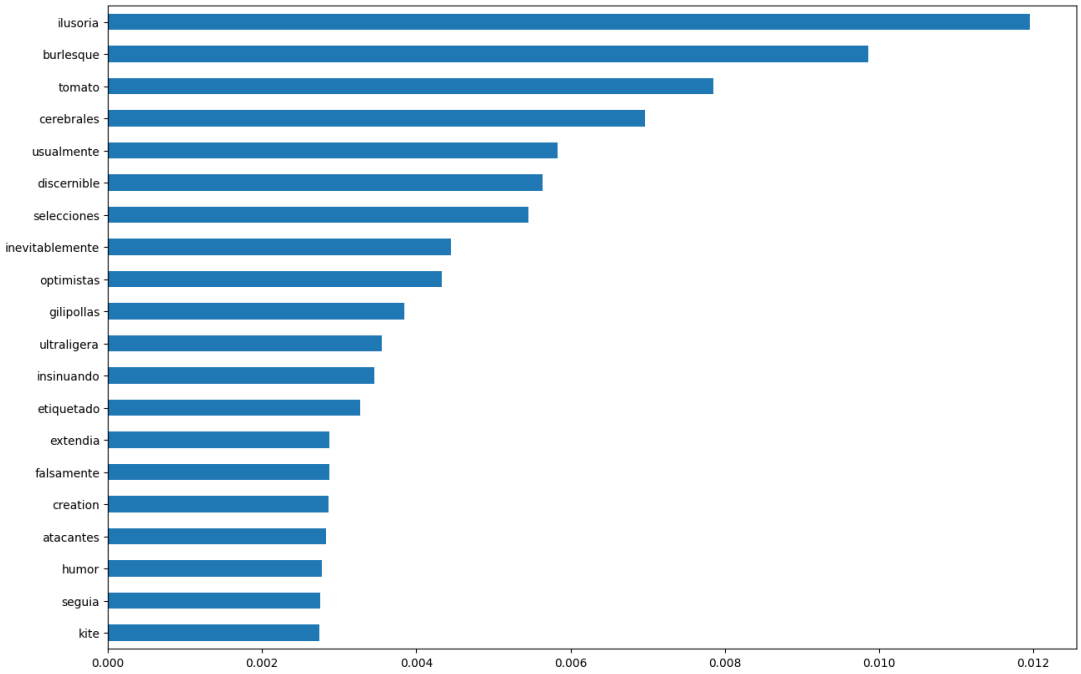


Ilustración 2. Variables importantes - TF-IDF.

	Train	Test
Precision	1.0	0.813906
Recall	1.0	0.796000
F1	1.0	0.804853

Ilustración 3. Métricas error -Bag of Words.

	Train	Test
Precision	0.786842	0.763636
Recall	0.897000	0.840000
F1	0.838318	0.800000

Ilustración 4. Métricas error - Bag of Words con validación cruzada.

	Train	Test
Precision	1.0	0.818548
Recall	1.0	0.812000
F1	1.0	0.815261

Ilustración 5. Métricas error - TF-IDF.

	Train	Test
Precision	0.830373	0.774135
Recall	0.935000	0.850000
F1	0.879586	0.810296

Ilustración 6. Métricas error - TF-IDF con validación cruzada.

En base a las diferentes métricas de error se puede observar que el mejor modelo utilizando el algoritmo Random Forest se obtiene utilizando **Bag of Words con validación cruzada** (ilustración 4)**.** En esta podemos ver que las métricas de entrenamiento respecto a las de prueba no son muy alejadas en comparación con los otros resultados. Es decir, la proporción de registros positivos identificados correctamente es alto (recall), y la proporción de verdaderos positivos que son realmente positivos en el total (precision). De manera general, podemos ver que no hay overfitting y tampoco underfitting.

Decision Tree – Erich

Se decidió utilizar un algoritmo de desition tree dado que este se utiliza para realizar predicciones, después de entrenado, en este caso se vio conveniente implementar un algoritmo de estas características para a partir de las diferentes representatividades por palabra, prediga si el sentimiento es positivo o negativo.

Con respecto a los resultados este modelo se entrenó uno con BoW y el otro con TF-IDF.

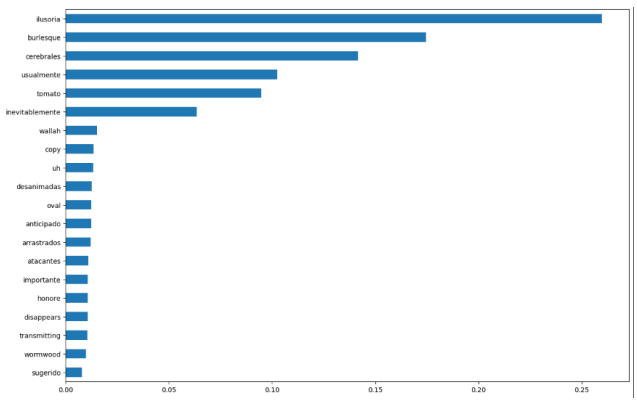
Modelo Con BoW:

Este arrojó unas métricas con los siguientes valores:

```
training
2
Precision: 0.6325985303941216
Recall: 0.947
F1: 0.7585102122547056
test
Precision: 0.6117804551539491
Recall: 0.914
F1: 0.7329591018444266
```

Por ende podemos decir que no hay Overfiting pero por ejemplo el valor de precisión es muy bajo. A pesar de tener un Recall alto y un F1 alto, lo más recomendable seria tener cuidado con las predicciones de variables positivas, dado el valor de la precisión.

Si tenemos este modelo en cuenta nos da una tabla de importancia como la siguiente:



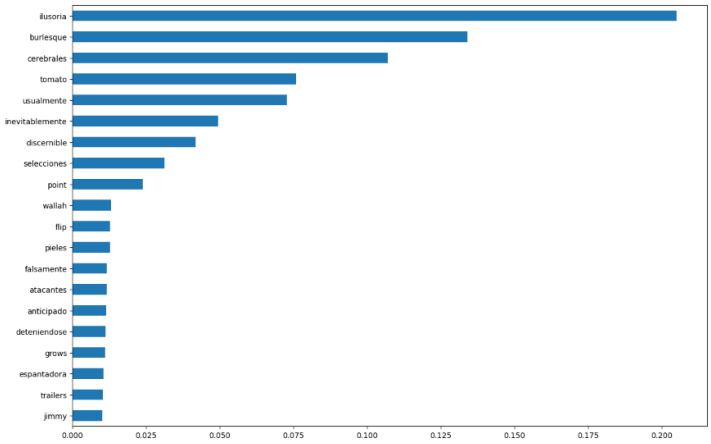
Donde tenemos a ilusoria en un primer puesto de relevancia, y sugerido en un último nivel de relevancia.

Modelo Con TF-IDF:

Con respecto a el modelo entrenado con TF-IDF, tenemos las siguientes métricas:

	Train	Test
Precision	0.662138	0.818548
Recall	0.926000	0.812000
F1	0.772149	0.736402

Por ende, el modelo, puede tener un posible underfiting, pero si lo tenemos en cuenta nos arroja la siguiente grafica.



Podemos decir que el modelo es muy parecido al primero solo que tiene pequeñas variaciones en cuanto a las relevancias, donde aparecen y desaparecen palabras y algunas cambian de puesto.

SVM – Juan Esteban Rodríguez

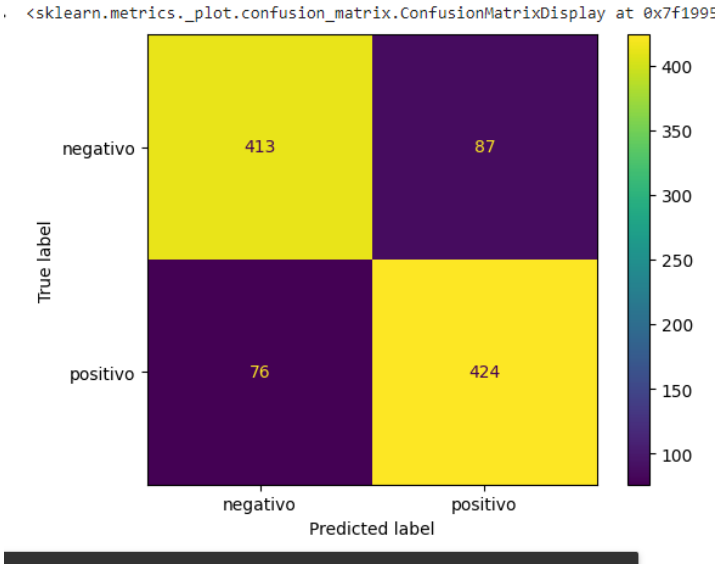
Este algoritmo corresponde a un modelo de clasificación llamado Support Vector Machine (SVM) que se utiliza para clasificar datos en dos o más categorías. En este caso específico, se está utilizando un SVM lineal y se está ajustando al conjunto de datos X_tfidf, que es una representación vectorial de los documentos del conjunto de datos de entrenamiento, y las etiquetas correspondientes y_train.

El algoritmo utiliza una técnica llamada TF-IDF (Term Frequency-Inverse Document Frequency) para crear una matriz de características numéricas que representa el contenido de los documentos. TF-IDF es una medida estadística que indica la importancia relativa de una palabra en un documento y en un conjunto de documentos. La técnica es útil para asignar un peso a las palabras que aparecen con frecuencia en un documento, pero no en todos los documentos, lo que las hace más discriminantes.

Una de las ventajas de este enfoque es que considera la frecuencia de las palabras en todo el conjunto de datos, lo que puede ser útil para identificar palabras que son relevantes en el conjunto de datos en su totalidad y no solo en un subconjunto específico. Además, al utilizar un SVM lineal, se puede entrenar el modelo de manera eficiente incluso en conjuntos de datos grandes.

Una desventaja potencial es que este enfoque puede ser susceptible al ruido en los datos, lo que puede llevar a un rendimiento subóptimo del modelo. Además, en comparación con la técnica de bag of words, el uso de TF-IDF puede ser más difícil de interpretar y visualizar.

Matriz de Confusión



Los resultados muestran que el modelo tiene un mejor rendimiento en el conjunto de entrenamiento que en el conjunto de prueba. En el conjunto de entrenamiento, la precisión es muy alta (0.987562), lo que indica que el modelo puede identificar correctamente el 98,8% de los casos positivos en ese conjunto de datos. Sin embargo, en el conjunto de prueba, la precisión es más baja (0.829746), lo que sugiere que el modelo no es tan bueno para identificar correctamente los casos positivos.

En cuanto al recall, es alto en ambos conjuntos de datos, lo que indica que el modelo puede encontrar la gran mayoría de los casos positivos en los datos. Pero el recall en el conjunto de entrenamiento es un poco más alto que en el conjunto de prueba, lo que sugiere que el modelo puede estar sobre ajustando los datos de entrenamiento.

	Train	Test
Precision	0.987562	0.829746
Recall	0.992500	0.848000
F1	0.990025	0.838773

(8%) Trabajo en equipo

Roles del equipo

Líder de proyecto y Líder de analítica: Wilton Esteban Martinez Hernandez

Tiempo: 12 horas

Algoritmo: RandomForestClassifier

Retos del proyecto: algunas palabras mal escritas por los clientes plantearon un desafío, pues se trató de buscar alguna herramienta que nos ayudará a corregirlas lo cual nos llevó un par de horas poder dar con una solución sensata de acuerdo con el negocio.

Puntos para mejorar: Iniciar el proyecto con anticipación, esto con el fin de evitar retrasos.

Líder de negocio: Juan Esteban Rodríguez Ospino

Tiempo en horas: 8

Algoritmo: SVM

Retos del proyecto: x**.** Los datos y las tendencias en el lenguaje natural están en constante cambio, por lo que es importante mantener el modelo actualizado para garantizar su efectividad a largo plazo

Otro reto es mantener el modelo actualizado,

Puntos para mejorar:

Líder de datos: Erich Giusseppe Soto Parada

Tiempo en horas: 7h

Algoritmo: Arboles de Decisión

Retos del proyecto: Para poder hacer que el algoritmo diera de forma correcta los resultados esperados, se tuvo que esperar varias horas y con diferentes combinaciones para arrojar los modelos aquí expuestos.

Puntos para mejorar: Se podría tal vez corrido el modelo con más combinaciones (para lo que se requeriría unas 6 horas más de entrenamientos) y tener los datos de mejor calidad y con una limpieza de datos mejor, pero de resto está bien.

(15%) Resultados:

No hay un modelo único que sea el mejor. Concluimos que todos los modelos son complementarios a pesar de sus problemas respectivos. Sería necesario revisar a profundidad los problemas de sobreajuste y subajuste en cada uno de ellos. Cada modelo obtuvo resultados decentes, aunque hay sospechas de sobreajuste y subajuste en algunos casos.

Creemos que tenemos una aproximación inicial acertada para utilizar los diferentes modelos en la predicción de análisis de sentimiento basado en comentarios. Esto

puede ayudar a la empresa a proporcionar contenido relevante basado en las palabras clave utilizadas por los modelos para realizar las predicciones.

Por ejemplo, si una de las palabras más importantes es “humor”, se podría proporcionar más contenido de ese estilo. Esto también se aplica a palabras relacionadas, como “gilipollas”, que en cierto contexto puede hacer referencia de forma coloquial al humor de un personaje en concreto. Si otra palabra importante es “atacante”, se podría recomendar más contenido de acción.

Cabe aclarar que aún es necesario realizar un análisis más profundo en cuanto al sentido y relevancia de las palabras en un contexto positivo o negativo. Sin embargo, este es un buen comienzo.

No hay un modelo único que sea el mejor. Concluimos que todos los modelos son complementarios a pesar de sus problemas respectivos. Sería necesario revisar a profundidad los problemas de sobreajuste y subajuste en cada uno de ellos. Cada modelo obtuvo resultados decentes, aunque hay sospechas de sobreajuste y subajuste en algunos casos.

Referencias

Raschka, S. (2014). Ensemble Classifiers. Retrieved from https://sebastianraschka.com/Articles/2014_ensemble_classifier.html