**CLASIFICACIÓN BASE DE DATOS DE CRITICAS PELICULAS**

**Reprocesamiento Data set**

Para poder aplicar las técnicas de clasificación como Arboles, Método de Bayes y K-nearest primero se realizó un trabajo extenso en el procesamiento de los documentos, para tener el datase listo para su respectiva clasificación.

En nuestra base de datos se utiliza un conjunto de datos divididos en críticas negativas y positivas de películas con el fin de que existan atributos válidos para una clasificación se realiza una reprocesamiento con los siguientes pasos:

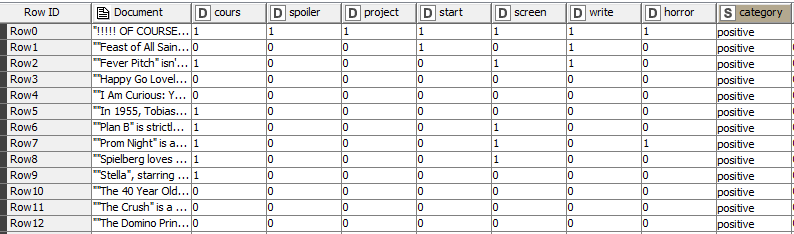
1. Se agrega una columna “category” donde se pone si la crítica pertenece al conjunto positivo o negativo.
2. Eliminación de puntos del documento
3. Eliminar los números del documento.
4. Seleccionar las palabras con más significado para la crítica, en este caso se seleccionan solo los sustantivos y adjetivos en ingles de los conjuntos de datos, y se eliminan los adverbios y conjunciones de el documento.

Por Ejemplo, la palabra “This is an antology horror film” nos quedaría “antology horror film”. Como se puede ver los artículos “is”, “an”.

1. Se aplica un algoritmo de le matización para eliminar inflexiones en las palabras y convertirlas a su raíz, por ejemplo:

|  |  |  |
| --- | --- | --- |
| Palabra | Regla | Resultado |
| Conflated | AT to ATE | Conflate |
| Trouble | BL to BlE | Trouble |
| Sized | IZ to IZE | Size |

1. Se aplica un “Bag Of Words” el cual crea una columna en nuestra tabla de asociación con cada palabra contenidas en un documento.
2. Se realiza cuenta la frecuencia de una palabra en cada documento, esto con el objetivo de asociar que palabras se repiten más en nuestro subconjunto de críticas positivas y negativas.
3. Finalmente se obtiene la tabla siguiente:



La cual nos muestra los documentos los cuales son las críticas que están en la base de datos, palabras las cuales se muestran el número de veces que aparecen en cada critica o documento y finalmente la categoría, la cual puede ser negativa o positiva dependiendo del documento dado por los datos de prueba.

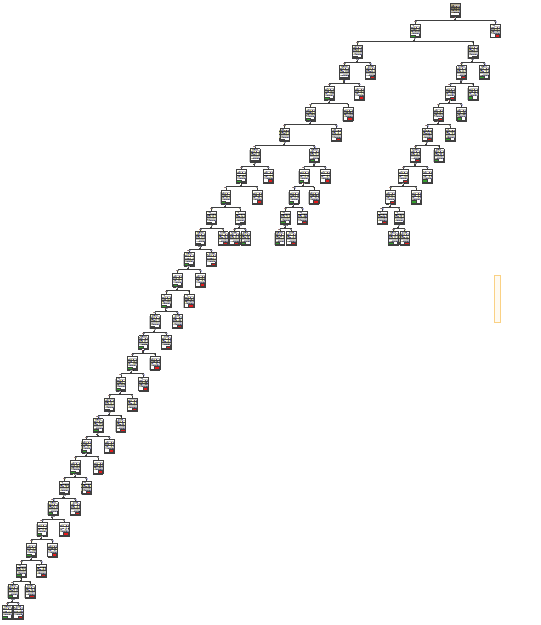
Todo este pre procesamiento se realiza con el objetivo de poder realizar una clasificación por las palabras más frecuentes en un documento, y poder saber con precisión si esta palabra tiene una influencia positiva o negativa en las críticas.

**Clasificación Por Arboles**

Después de procesar nuestro conjunto de pruebas, realizamos una clasificación mediante un árbol binario. Nuestro modelo es construido a partir de las palabras de mayor frecuencia en cada documento y la categoría de las críticas, si son negativas o positivas. El objetivo de nuestro árbol es que clasifique según las palabras si una crítica es positiva o negativa.

Modelo de árbol

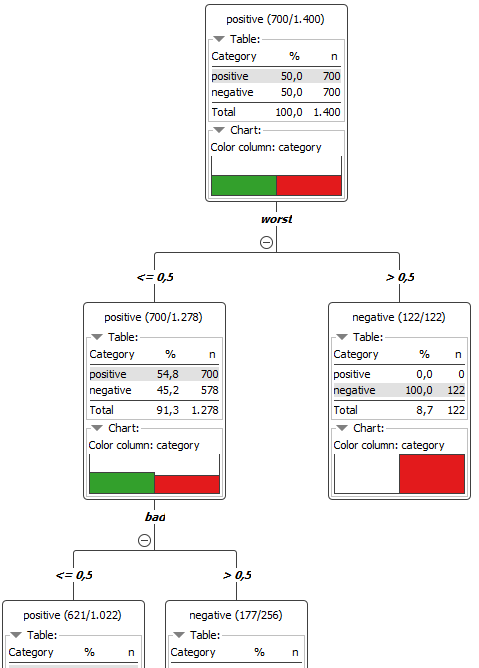
El resultado del modelo de árbol hecho con el algoritmo de poda fue el siguiente:



La imagen anterior nos muestra un árbol binario donde los nodos son las palabras, y dependiendo de la frecuencia de aparición de cada palabra el árbol la clasifica en malo o bueno.

Si bien el árbol resultante es profundo, cabe destacar que existen una gran cantidad de combinaciones de palabras para cada documento.

Sin embargo, como se muestra en la siguiente imagen, vemos que la palabra “worst” que es nuestro nodo raíz y en ingles una palabra negativa, cuando existe una frecuencia de esta palabra mayor a 0.5, de inmediato clasifica la crítica en una categoría Negativa, lo que comprueba que el modelo está funcionando correctamente:



**La matriz de confusión**

Esta matriz nos muestra la cantidad de datos que nuestro modelo de clasificación pudo predecir.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Positivo** | **Negativo** |
| **Positivo** | 258 | 42 |
| **Negativo** | 40 | 220 |

Como se muestra en la tabla anterior fueron 258 verdaderos positivos, 42 falsos positivos 40 verdaderos negativos y 220 falsos negativos. Lo cual muestra un gran acierto a la hora de predecir categorías por parte de nuestro árbol.

**Curva RoC**

Como podemos ver la curva de ROC en cuanto a la hora de clasificar categorías negativas , como podemos observar la curva está por encima de la recta diagonal , con lo cual nos indica un resultado bueno, y con el área bajo la curva de 0.8251 que está muy cerca a 1.

