**Regresión – Informe**

**Relación entre el contenido y la clasificación**

En este estudio se ha hecho un preprocesamiento de los textos de forma igual a la utilizada en la clasificación: para los dos conjuntos de 1.000 *reviews*, se ha eliminado su puntuación, números irrelevantes, palabras demasiado pequeñas (en particular con menos de 3 caracteres) y palabras irrelevantes (*stop words*); después se han convertido todos los caracteres a minúsculos, y las palabras a sus lemas. Enseguida se han eliminado los términos que no aparecían en al menos 20 documentos.

Los documentos resultantes fueron convertidos en vectores, donde la presencia y ausencia de cada palabra se indicaba respectivamente con 1 y 0; a esas columnas se ha añadido una columna con la clasificación del documento, 1 para una *review* positiva y 0para una *review* negativa. Las tablas resultantes fueron unidas y los valores en falta rellenados con 0, por corresponder a términos que no aparecían en ninguno de los documentos de ese conjunto sino solo en los del otro conjunto.

Para predecir la clasificación de una *review* en función de su contenido se ha utilizado un árbol de decisión con potenciación de gradiente o *Gradient Boosting*, con un número de niveles entre 4 y 10 conforme lo recomendado en la descripción del nodo respectivo en Knime. Sus otros parámetros son los estándares de Knime: 100 modelos, y tasa de aprendizaje igual a 0.1.

Para cada número de niveles, se ha obtenido tanto la matriz de confusión con respecto a la predicción de la clasificación, como la correlación linear entre la clasificación prevista y la presencia o ausencia de cada una de las palabras presentes en el conjunto de textos preprocesados.

En el Gráfico 1, se puede ver como de bien el árbol de regresión empleado clasifica cada documento en función de su número de niveles. Existen 1.000 documentos para cada una de las dos clasificaciones, positiva y negativa, y para cada número de niveles se registra el número correcto de clasificaciones tanto positivas como negativas.

En general, para el rango de niveles [4; 10] los documentos son bien clasificados. El rango de clasificaciones correctas es [787; 992] para las negativas y [813; 979] para las negativas, o haciendo la conversión a una tasa de error, respectivamente [0.8; 21.3]% y [2.1; 18,7]%.

Para cada número de niveles en el árbol hay siempre una clasificación que es mejor detectada (no más de 3.5% de tasa de error) y una que se detecta con más dificultad (más de 8.5% de tasa de error), incluso siendo posible diseñar una línea en el gráfico que separe la clasificación mejor de la peor, para cada número de niveles.

La clasificación mejor detectada empieza por ser la positiva, para 4 y 5 niveles en el árbol. Con 6 niveles la situación se invierte y la clasificación negativa pasa a ser la mejor clasificada. Con 7 niveles de nuevo la clasificación positiva es la más acertada, hasta que con 8 o más niveles la clasificación negativa se acierta más.

Las clasificaciones tienden a mejorar con el aumento del número de niveles en el árbol: para 4 niveles se tiene respectivamente 3.3% y 21.3% de tasa de error para la clasificación mejor y peor identificada; para 10 niveles esos valores son respectivamente 0,8% y 8,9%.

Gráfico - Clasificaciones positivas (True Positive) y negativas (True Negative) identificadas correctamente en un universo de 1.000 documentos para cada clasificación, en función del número de niveles del árbol de regresión

Además del enfoque en las clasificaciones, se ha enfocado en la correlación existente entre la clasificación prevista (independientemente de estar correcta o no) y cada uno de los términos existentes en el conjunto de todos los documentos.

Para cada número de niveles del árbol de regresión, entre 4 y 10, se ha registrado en la Tabla 1 el coeficiente más bajo y la palabra asociada; el 1º cuartil; la mediana; el 3º cuartil; y el coeficiente más alto y palabra asociada.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Niveles** | **Mínimo** | **Palabra Mínimo** | **1º Cuartil** | **Mediana** | **3º Cuartil** | **Máximo** | **Palabra Máximo** |
| **4** | -0,44 | bad | -0,06 | 0,0 | 0,07 | 0,15 | love |
| **5** | -0,39 | bad | -0,06 | 0,0 | 0,08 | 0,15 | love |
| **6** | -0,41 | bad | -0,06 | 0,0 | 0,08 | 0,15 | fantast |
| **7** | -0,38 | bad | -0,06 | 0,0 | 0,08 | 0,15 | fantast |
| **8** | -0,38 | bad | -0,06 | 0,0 | 0,08 | 0,15 | fantast |
| **9** | -0,37 | bad | -0,06 | 0,0 | 0,08 | 0,15 | fantast |
| **10** | -0,32 | bad | -0,06 | 0,0 | 0,08 | 0,16 | fantast |

Tabla 1 - Análisis estadística de los coeficientes de correlación entre la predicción de la clasificación y cada uno de los términos existentes en el conjunto completo de documentos

Se puede observar que los resultados cambian muy poco con el número de niveles del árbol de regresión: el coeficiente mínimo es el parámetro que más cambia, variando en un rango de [-0.44; -0.32]. Está siempre asociado a la palabra “bad”, malo; y la palabra “worst”, peor, siempre está asociada al segundo coeficiente más bajo. Por lo tanto se puede concluir que cuando una de esas palabras o sus derivados vienen en el documento es muy posible que la review sea negativa, aunque no sea sorprendente si no es así.

En el extremo opuesto, “love”, amor; y “fantast”, raíz de palabras como “fantastic”, fantástico, son dependiendo del número de niveles las palabras con los coeficientes más altos. Sin embargo su valor absoluto es entre 2 a 3 veces más bajo que el valor absoluto del coeficiente más negativo, lo que permite concluir que la presencia de esos términos en el documento tiene mucho menos influencia en su clasificación prevista.

Es decir, cuando un documento tiene la palabra “bad”, es más probable que esa *review* sea negativa do que ser positiva cuando tiene la palabra “love”. Esto sugiere que la palabra “bad” es más exclusiva de *reviews* negativas, mientas “love” y palabras con raíz “fantast” se utilizan de forma más indiscriminada en las *reviews.*