Medición de la reputación corporativa en los medios de comunicación. Un caso de migración de RapidMiner a R.

Rocio Curti Frau[1], Juan Pablo Sokil[1]

[1] Universidad de Buenos Aires, Facultad de Ciencias Sociales, Buenos Aires, Argentina

Keywords/Palabras Claves: minería de texto, análisis de sentimiento, código abierto, migración de modelos, aprendizaje automático, boosting

La clasificación automática de grandes volúmenes de datos no estructurados y de texto resulta cada vez más importante en áreas de comunicación y prensa, principalmente ante el crecimiento de medios de comunicación online. Este trabajo tiene como objetivo presentar nuestra experiencia en relación a la migración de modelos de categorización de texto y análisis de sentimiento de Rapidminer hacia R.

Los modelos iniciales fueron entrenados con un corpus de notas periodísticas que mencionan a ciertas empresas publicadas en distintos medios de comunicación argentinos entre septiembre 2016 y julio 2017. Cada nota fue clasificada según la valoración (positiva, negativa e informativa) y la temática en torno a la reputación corporativa (calidad de productos y servicios, conducta corporativa, liderazgo, innovación, público interno y sustentabilidad).

En el proceso de migración se trabajó con el mismo conjunto de datos y se buscó replicar el modelo anterior con el fin de validarlo y garantizar una comparabilidad entre ambos desarrollos. El texto de la nota fue convertido en una representación estructurada del mismo mediante el modelo de espacios vectoriales. Primero se llevaron a cabo tareas de tokenización del texto, eliminación de palabras vacías, construcción de bigramas y aplicación de algoritmos de stemming. Luego se elaboró una matriz de frecuencia de términos (tf) y sobre ella una matriz de frecuencia inversa del documento (idf). Por último, se realizó una poda del vocabulario quedando conformada la base de datos de entrenamiento.

Cada una de las categorías se trabajó como variable dicotómica por lo que se entrenaron y optimizaron los parámetros de 10 modelos diferentes, a partir de un ensamble de árboles de decisión basado en boosting utilizando validación. Se alcanzó una curva roc superior al 90% en todas las categorías de análisis. Finalmente se definieron criterios estadísticos para asignar las categorías a cada nota respecto al tema y la valoración.

Los resultados obtenidos no mostraron una mejora significativa con respecto a Rapidminer. Sin embargo, permitieron estandarizar procesos y mejorar la velocidad de procesamiento de los datos. En la actualidad nos encontramos explorando nuevas técnicas, entre las que se destacan la integración de R con Python.

Librerías de R utilizadas: tm, quanteda, stringr, data.table, text2vec, xgboost, Matrix, mlrMBO y DiceKriging.