Clasificación de Oraciones de Artículos Científicos Por Medio de Regresión Logística Multinomial

Luis Orellana Altamirano.

Universidad de Santiago de Chile, Chile
Avenida Ecuador #3659. Estación Central, Santiago de Chile., Chile
luis.orellana.a@usach.cl
fernando.cabrera.ga@usach.cl

Abstracto. El presente trabajo de investigación está centrado principalmente en el etiquetado o clasificación del corpus de documentos científicos. Las etiquetas de los corpus han sido asignadas según el lugar y contenido al cual pertenecen en la estructura general de cada uno de los documentos científicos. Sin embargo, cabe preguntarse, ¿de que manera es posible determinar el tipo de oración de un fragmento específico del texto, tan solo al analizar su contenido? Para tal efecto, se utilizará el método llamado máxima entropía, o también conocido como regresión logística multinomial, el cual permite realizar múltiples clasificaciones de texto.

Palabras Clave: Máxima Entropía, Corpus, Clasificación de Texto, Lime.

1 Introducción

El conocimiento científico es difundido por medio de publicaciones, las cuales están sujetas a evaluación. Por esta razón, la organización y clasificación de estas publicaciones se torna un tema relevante tanto para las editoriales como también para los investigadores ya que permite tener claridad respecto al etiquetado de los documentos basados en algún criterio, logrando que la información se encuentre ordenada con base en una estructura determinada. Para tal efecto, en el presente trabajo, se han utilizado un conjunto de técnicas de minería de texto, con el objetivo de clasificar corpus según sus características.

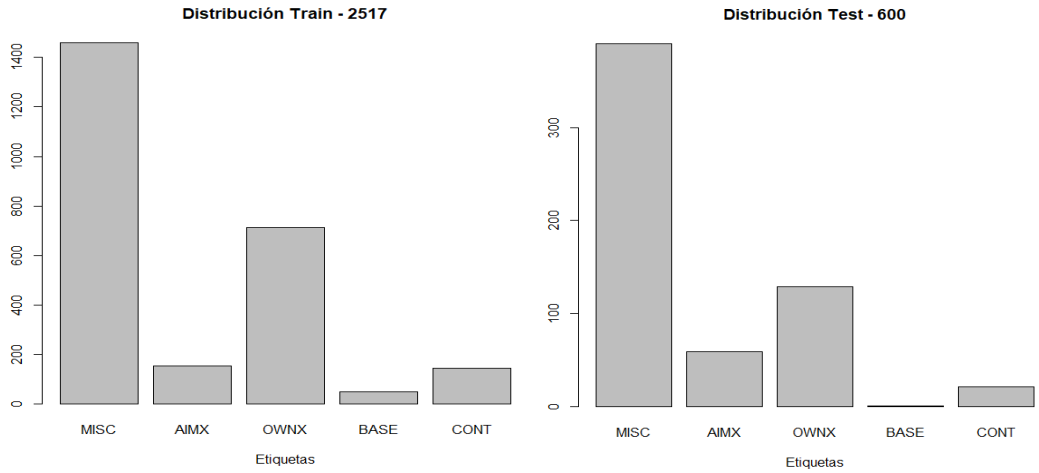
Los datos ocupados para este experimento están contenidos en el dataset publicado por S. Teufel and M. Moens(2002)[1], el cual lleva por nombre “Sentence Corpus”. Este dataset contiene un conjunto de 30 artículos científicos, correspondientes a los tópicos de biología computacional, aprendizaje de máquina, psicología y toma de decisiones. Estos 30 artículos fueron divididos en 3.117 instancias, las cuales están distribuidas en 80,8% de datos de entrenamiento y un 19,2% de datos de validación. Los datos están disponibles en el repositorio de CRAN en R, bajo el nombre de “Lime”[2].

Las etiquetas se pueden clasificar en cinco tipos las cuales son detalladas en la tabla 1. Por otro lado, las distribuciones de dichas etiquetas están descritas en la imagen 1.

Tabla 1. Etiquetas de los Corpus.

Etiqueta	Prefijo	Significado
Aim	AIMX	El objetivo general de la publicación
Own	OWNX	El propio trabajo del autor (métodos, resultados, conclusión)
Contrast	CONT	Contraste, comparación o críticas a anteriores trabajos
Basis	BASE	Trabajos anteriores que proveen la base del presente articulo
Misc	MISC	Cualquier otra oración

Imagen 1. Distribución de Etiquetas.



2.2 Método de Minería de Datos

Con la finalidad de obtener conocimiento por medio de técnicas de minería de texto, se utilizará el método de máxima entropía, el cual es expresado como:

$$p_1(X_1, X_2) = p_1 = E(Y_1) = \frac{\exp(Z_1)}{1 + \exp(Z_1) + \exp(Z_2)}$$

$$p_2(X_1, X_2) = p_2 = E(Y_2) = \frac{\exp(Z_2)}{1 + \exp(Z_1) + \exp(Z_2)}$$

$$p_3(X_1, X_2) = p_3 = 1 - p_1 - p_2 = \frac{1}{1 + \exp(Z_1) + \exp(Z_2)}$$

Donde “P” es la probabilidad de un determinado texto “X” de pertenecer a una etiqueta “Y”, y “Z” es una función multinomial llamada hipótesis que representa un hiperplano de clasificación vectorial.

Este método se encuentra disponible en el repositorio CRAN del software R. Este software es de código abierto, multiplataforma y de programación de alto nivel, el cual trata como caja negra operaciones vectoriales y matriciales. Específicamente, para el análisis de este paper, se utilizaron las librerías *“lime”* y *“maxent”* la cual contiene el método de clasificación. Por otra parte, para evaluar el rendimiento del algoritmo fue utilizada la métrica F-score, la cual es expresada como:

$$Fscore = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}}$$

.Un valor equivalente a 1 en f-score corresponde a una precisión y recall perfecto y el valor 0 es el peor valor que puede alcanzar la métrica[3]. Donde “P” es precisión y “R” es recall. Estos dos valores son calculados a partir de la matriz de confusión

2.3 Preprocesamiento de los textos

Antes de enfrentar el problema y contestar la pregunta de investigación, se utilizará una serie de acciones de procesamiento de los datos, con la finalidad de sacar más provecho al método máxima entropía. Es por ello que fueron utilizados los siguientes recursos disponibles en los paquetes “tm”, “SnowballC” y “wordcloud” de CRAN en R. Los pasos considerados para el pre-procesamiento de texto fueron[4]:

- Remover puntuación
- Remover preposiciones
- Transformación a minúsculas
- Reemplazar palabras por sus orígenes sintácticos
- Remover espacios en blanco
- Remover números

Cabe destacar que este proceso fue realizado acumulativamente, es decir, todos los procesos mencionados contienen a su sucesor.

2.4 Procedimiento de Ajuste del Modelo

Para cada paso en la etapa de pre-procesamiento fue evaluada la métrica de f-score, con el objetivo de medir el comportamiento de la métrica luego de cada filtrado de los datos. De esta manera, se pudo identificar tanto la etiqueta con mayor métrica f-score, como también las etapas del pre-procesamiento más efectivas. Lo anteriormente mencionado, se ve reflejado en la imagen 2.

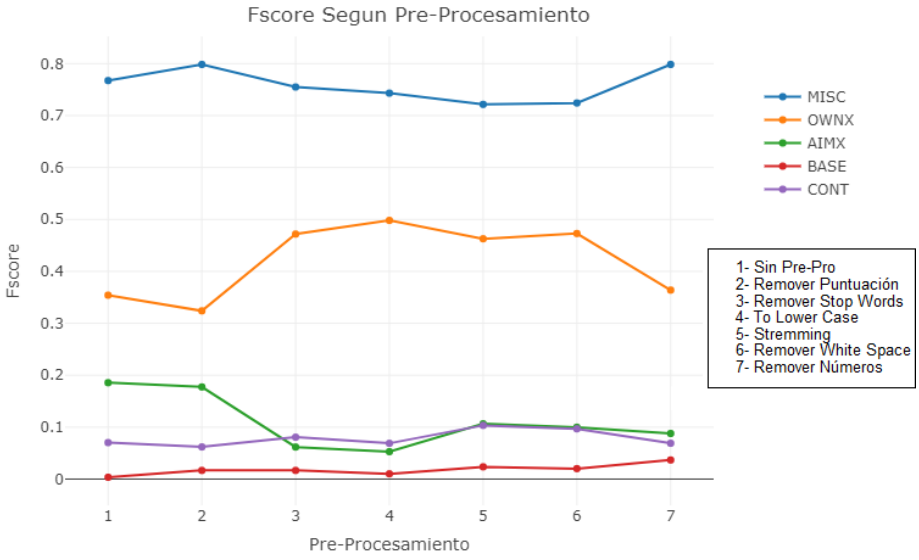
Cabe destacar que cada una de las etapas de pre-procesamiento que fueron nombradas son acumulativas e incluidas una tras otra de manera sistemática.

Por otra parte, estas mediciones son determinadas en base a la bondad del clasificador sobre los datos de prueba, luego de haber entrenado al clasificador por medio de los datos de aprendizaje. Este procedimiento fue aplicado para cada una de las etiquetas disponibles en el dataset, con la finalidad de determinar el tipo de corpus con mayor relevancia.

Para el proceso de selección de parámetros fue utilizada la librería “maxent” para obtener el modelo clasificador, el cual provee un método con el mismo nombre para tal fin. Dicho método utiliza una serie de parámetros[5], de los cuales, la siguiente especificación fue considerada como óptima por la librería “maxent” obteniendo el modelo con la mejor métrica f-score:

- l1_regularizer = 0.0
- l2_regularizer = 0.8
- use_sgd = FALSE
- set_heldout = 0

Imagen 2. F-score Según Pre-Procesamiento.



2.5 Resultados Obtenidos (discusión)

Por medio de la métrica f-score (que considera importantes la precisión y el recall al momento de calcular la puntuación) fueron evaluados los resultados obtenidos por el método a lo largo del proceso de filtrado de datos.

Con base en el análisis en la fase de pre-procesamiento, se puede observar que la mejor métrica, la cual corresponde a la mayor relevancia en el proceso de recupera-

ción de información fue “MISC” (misceláneos). En la tabla 2 se presentan los resultados de la matriz de relevancia. A partir de ellos, es posible mencionar:

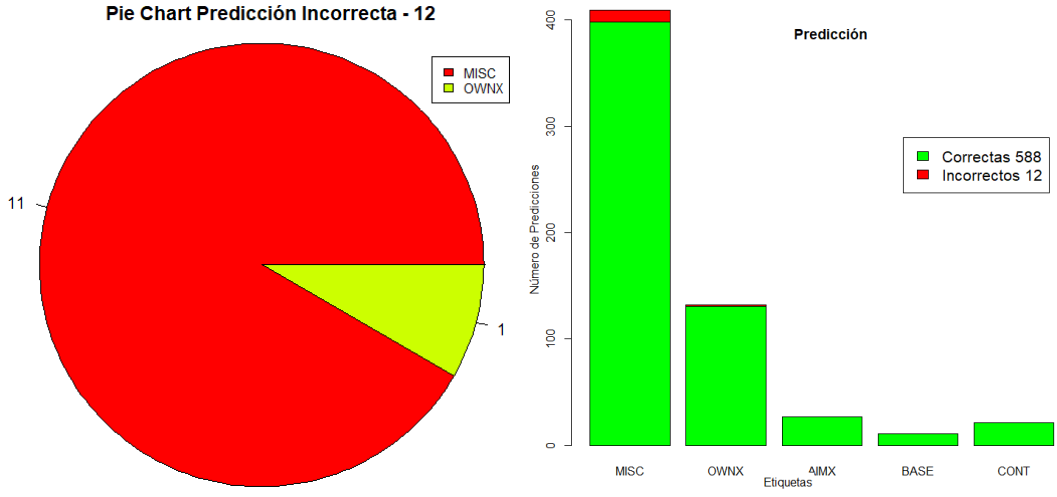
Para la etiqueta MISC (misceláneos), se recuperó un 97% de los corpus relevantes. El valor obtenido para la métricas es: f-score 79.8%, precisión de un 67%, y por último un recall de 0.99%.

Tabla 2. Matriz de Confusión Para la Etiqueta MISC (misceláneos) con Todo el Conjunto de Preprocesamiento.

	Relevantes	**No Relevantes**	**Total**
Recuperados	398	190	**588**
No Recuperados	11	1	**12**
Total	**409**	**191**	**600**

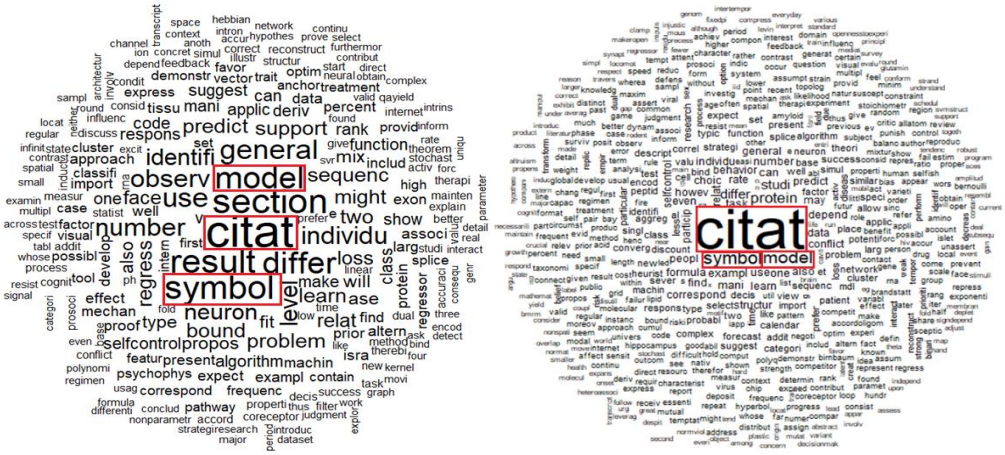
Con respecto a la cantidad clasificaciones erróneas suman un total de 12 documentos. Esto corresponde al 2% de los documentos, de los cuales un corpus fue clasificado como “OWNX” y los restantes 11 fueron clasificados como “MISC”. Esto se ve representado en la imagen 3.

Imagen 3. Error en la Clasificación de Corpus.



Por último, se concluye que el error de clasificación corresponde a prevalencia de algunas palabras que tienen gran frecuencia entre distintas etiquetas, de las cuales, es posible visualizarlas en la imagen 4.

Imagen 4. Frecuencia de Palabras en las Etiquetas “OWNX” y “MISC” Respectivamente.



3 Conclusiones

A partir de los resultados obtenidos en la clasificación podemos corroborar que la etiqueta “MISC” (misceláneos) presenta un error del 2% en la clasificación de los documentos y un f-score de 79,8%. Con respecto a la información del etiquetado inicial del problema, esto correspondería a todas las publicaciones que no sean: objetivos específicos, trabajo experimental de un autor, trabajos basados en publicaciones anteriores, ni comparaciones y/o críticas. Sin embargo, para tener mayor conocimiento respecto a los textos involucrados debería existir un proceso de nueva clasificación de esa etiqueta por ser demasiado general.

Los valores de la métrica f-score para las otras etiquetas presentes en el experimento no fueron evaluados debido a que eran cercanos a 0, lo que es considerado un valor malo para f-score.

Con respecto a la pregunta de investigación, el proceso encargado de realizar el análisis y clasificar un tipo de oración para un fragmento específico fue casi exacto para la etiqueta “MISC”, sin embargo, presenta dificultad en algunas palabras que poseían gran frecuencia entre las distintas etiquetas. Esto sucede debido a que en los textos científicos sin importar la categoría existen términos comunes que se repiten en investigaciones de diferentes índoles.

4 Referencias

1. [1] S. Teufel and M. Moens(2002). Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics, 28(4):409-445.

2. [2]R, C. (s.f.). Documentación de R. Obtenido de https://cran. project.org/web/packages/lime/index.html

3. [3]Zhang,E., Zhang,Y. Encyclopedia of Database Systems, Springer(2009): 1147.

4. [4]Gurusamy,V.,Subbu, K., Preprocessing Techniques for Text Mining. (2014).

5. [5]CRAN. (s.f.). Maxent: Low-memory Multinomial Logistic Regression with Support for Text Classification. Obtenido de https://cran.r-project.org/web/packages/maxent/maxent.pdf.