REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2020 PUCV, Chile 53(104) 549-567

Clasificación de Textos Multi-etiquetados con Modelo Bernoulli Multi-variado y Representación Dependiente de la Etiqueta

Multi-label Text Classification with Multi-variate Bernoulli Model and Label Dependent Representation

Rodrigo Alfaro A.
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
CHILE
rodrigo.alfaro@pucv.cl

Héctor Allende O.
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
CHILE
hallende@inf.utfsm.cl

Recibido: 04-X-2018 / **Aceptado:** 24-I-2020
DOI: 10.4067/S0718-09342020000300549

Resumen

La asignación de una o más categorías predefinidas a los textos en lenguaje natural, basados en su contenido, es un componente importante y necesario en muchas tareas al interior de las organizaciones. Esta tarea se realiza comúnmente a través de la clasificación automática de textos, esto es, clasificando documentos dentro de un conjunto de categorías predefinidas por medio de un modelo y método computacional. La representación de los textos para propósitos de clasificación automática ha sido tradicionalmente llevada a cabo usando un modelo de espacio vectorial debido a su simplicidad y buen rendimiento. Por otro lado, la clasificación automática de textos por multi-etiquetados ha sido típicamente abordada utilizando métodos de clasificación de etiqueta simple, lo que implica transformar el problema estudiado para aplicar técnicas binarias o adaptar algoritmos binarios para que funcionen con múltiples etiquetas. En este artículo el objetivo es evaluar un factor de ponderación de las palabras de los textos en el modelo booleano para representación de texto en clasificación multi-etiqueta, usando una combinación de dos enfoques: transformación de problema y adaptación de modelo. Este factor de ponderación y la combinación de enfoques en la clasifiación automática fue puesto a prueba con cuatro diferentes conjuntos de datos textuales utilizados en la literatura especializada y comparado con técnicas alternativas por medio de tres medidas de evaluación. Los resultados presentan mejoras superiores al 10% en el rendimiento de los clasificadores, atribuidas a nuestra propuesta, en todos los casos analizados.

Palabras Clave: Multi-etiqueta, clasificación de textos, representación de textos, transformación del problema, ponderación de términos.

Abstract

The allocation of natural language texts to one or more predefined categories or classes based on their content is an important component and a recent need in many information organization and management tasks. Automatic text classification is the task of categorizing documents to a predefined set of classes by a computational method or model. Text representation for classification purposes has been traditionally approached using a vector space model due to its simplicity and good performance. On the other hand, multi-label automatic text classification has been typically addressed either by transforming the problem under study to apply binary techniques or by adapting binary algorithms to work with multiple labels. In this article, the objective is to evaluate a term-weighting factor in the Boolean model for text representation in multi-label classification, using a mix of two approaches: problem transformation and model adaptation. This term-weighting factor and the combination of approaches in the automatic text classification was tested with four different sets of textual data used in the specialized literature and compared with alternative techniques by means of three measures of evaluation. The results present improvements of more than 10% in the performance of the classifiers, attributed to our proposal, in all the cases analyzed.

Key Words: Multi-label, text classification, text representation, problem transformation, term weighting.

INTRODUCCIÓN

Desde una perspectiva histórica, el análisis de texto se inició en las cercanías del año 1200 cuando el monje dominico Hughe of Saint-Cher creó el primer diccionario de concordacias bíblicas, un índice verbal que enumeraba las palabras bíblicas alfabéticamente, con indicaciones para permitir que el investigador encuentre los pasajes de la Biblia donde aparecen las palabras (Ignaton & Mihalcea, 2017). Junto a lo anterior, también hay evidencia en la inquicisión europea de estudios de conceptos en la Biblia. Asimismo, el primer análisis de texto documentado fue en Suecia en el año 1700 y el primer análisis sistemático cuantitativo fue a principios del siglo XX. Este último mostraba que a fines de 1800 los periódicos de Nueva York habían caído en cobertura científica, religiosa, de deportes, de chismes y escándalos. Porteriormente, hubo estudios similares desarrollados por Wilcox (1900), Fenton (1911) y White (1924).

En el ámbito de la lingüística, desde fines de los años sesenta se desarrolló la denominada lingüística del texto, cuyo fin estaba orientado a reconocer el texto como objeto de estudio. Entre las tareas y aportes más relevantes que ha tenido esta área lingüística ha sido el interés permanente por la tipologización y clasificación de los textos (Ciapuscio, 1994; Loureda, 2003), esto es, discriminar distintas clases de textos, a partir de criterios lingüísticos, textuales, pragmáticos y funcionales (Ciapuscio, 2000). En la actualidad, la cantidad de textos disponibles en la Internet y bases de datos corporativas crece exponencialmente. Es por ello, que existe un creciente interés por ayudar a las personas a encontrar, filtrar y administrar la información disponible en los textos manera eficiente y efectiva (Venegas, 2019).

El análisis de texto es un campo interdisciplinario relativamente nuevo, basado en las ciencias de la computación y la lingüística. Hoy, la disponibilidad de alta capacidad computacional, los nuevos lenguajes de programación y las múltiples fuentes de datos, han fomentado el desarrollo de nuevos métodos desde las humanidades y ciencias sociales, en conjunto con los lingüístas computacionales y ciencias de la computación.

A partir de esta convergencia, surge la clasificación automática de textos como una tecnología que permite asignar etiquetas pre definidas a textos de lenguaje natural, basándose en su contenido. Esta tecnología ha logrado posicionarse como un componente importante de muchas tareas de las organizaciones y su administración. Su aplicación más importante y común hasta la fecha ha sido en la asignación de categorías de tema a un documento para apoyar el filtrado de textos, ruteo de correos electrónico, identificación de lenguaje, evaluación de lecturabilidad (Benjamin, 2012), análisis de polaridad (Lavalle, Montes, Villaseñor, Jiménez & Bárcenas, 2018), perfilamiento de autores (Ortega, López-Monroy, Franco & Montes-y-Gómez, 2018), análisis de opiniones (Chatterjee, Gupta, Chinnakotla, Srikanth, Galley & Agrawal, 2019), entre otros. La clasificación automática de textos juega un papel relevante en una gran variedad de tareas de administración de información las cuales son más flexibles, dinámicas y personalizadas, entre otras, ordenamiento de correos electrónicos u organización jerárquica de carpetas en tiempo real, identificación de tema para el soporte de operaciones, búsquedas o navegación estructurados, encontrar documentos que correspondan a intereses de largo plazo o intereses más dinámicos basados en tareas.

Tal como lo plantea Venegas (2007), en el ámbito del procesamiento del lenguaje natural (PLN) se integran dos ámbitos disciplinares: la lingüística y la matemática. La primera, a través del estudio del texto y sus componentes. Mientras que la segunda, en conjunto con la estadística, plantean la aplicación de métodos y técnicas de análisis a los textos de lenguaje natural o humano.

Tradicionalmente, la clasificación (o categorización) de textos se ha definido como la asignación de un valor Booleano (verdadero o falso) a cada par $\langle d_j, c_i \rangle \in D \times C$, donde *D* es el dominio de los documentos (*corpus*) y $C = \{c_1, \ldots, c_{|c|}\}$ es el conjunto de etiquetas (clases) predeterminadas. Si un documento está categorizado solo bajo una etiqueta (categorías no sobrepuestas) o bajo múltiples etiquetas a la vez (categorías sobrepuestas), se le llama un 'problema de una etiqueta' o un 'problema multi-etiqueta' respectivamente (Sebastiani, 2002). El caso más estudiado para resolver problemas de clasificación de texto es el de 'una etiqueta' y el enfoque principal es el llamado de Clasificación Binaria (*Binary classification*, BC), donde un documento es clasificado, ya sea a la categoría c_i o a su complemento $\overline{c_i}$. Este enfoque puede ser extendido y utilizado para resolver problemas con más categorías.

En un problema multi-etiqueta existe un número de etiquetas finito $L = \{\lambda_j : j = 1 \dots l\}$, donde λ_j corresponde a la etiqueta *j-ésima,* y al set de documentos etiquetados $D = \{f\{\mathbf{x_i}, \mathbf{Y_i} : i = 1 \dots d\}\}$, donde $\mathbf{x_i}$ representa el vector de características y $\mathbf{Y_i} \subset L$ es el conjunto de etiquetas del texto *i-ésimo.*

La clasificación de multi-etiqueta es un problema importante para la aplicación real, como se puede observar en muchos dominios, como por ejemplo genómica funcional, categorización de textos, minería de música y clasificación de imágenes, por nombrar algunos. En Alfaro y Allende (2011) se presenta una representación de texto modificada para un problema de multi-etiqueta llamado *tf-rfl*. Su rendimiento fue comparado aplicándola en el conjunto de datos conocido como Reuters-21578 y se concluyó que existe una mejora estadísticamente significativa comparado con enfoques alternativos, considerando solo la medida de desempeño *Hamming Loss*.

El propósito de este artículo es evaluar una nueva representación llamada *bin-rfl*, como una extensión de *tf-rfl*, presentada en Alfaro y Allende (2011). Esta nueva representación utiliza representación binaria de términos, en lugar de la frecuencia de los mismos. Se muestra el impacto de la representación considerando diferentes medidas de desempeño para problemas de clasificación de multi-etiqueta. Para ello, se han utilizado dos clasificadores automáticos lineales, los cuales son: una máquina de soporte vectorial lineal (SVM) y redes neuronales artificiales de una capa (ANN). El uso de clasificadores automáticos lineales permite evaluar las mejoras en el rendimiento de los algoritmos, tan solo modificando el espacio de entrada mediante la nueva representación.

La contribución de este artículo es proponer una nueva representación más simple, la cual comprueba nuestra hipótesis, esto es, que una modificación supervisada a la representación del texto en el espacio de entrada, basada en representaciones binarias y una ponderación de términos asociada a los ejemplos conocidos según sus etiquetas, puede mejorar significativamente el rendimiento de los clasificadores. Esta nueva propuesta está probada en cuatro conjuntos de datos textuales multi-etiquetados, los cuales han sido ampliamente usados en la literatura, con diferentes medidas de desempeño.

Este paper está estructurado de la siguiente forma. En la sección 1, introducimos brevemente la clasificación de textos multi-etiqueta. En la sección 2, analizamos diferentes representaciones de texto y presentamos nuestra propuesta. En la sección 3, comparamos el rendimiento de nuestra propuesta con otros algoritmos. La última sección está dedicada a conclusiones finales, discusión de resultados y trabajos futuros.

1. Marco teórico

El interés por la clasificación automática de textos multi-etiqueta ha ido aumentando en los últimos años. Entre las propuestas presentadas por Tsoumakas y Katakis (2007), el enfoque más usado es el denominado transformación del problema. La Tabla 1 presenta un ejemplo de textos multi-etiquetados. En ella se representa que el Texto 1 pertenece a las clases 'Deportes' y 'Política', que el Texto 2 pertenece a las clases 'Ciencias' y 'Política', que el Texto 3 pertenece solo a la clase 'Deportes' y finalmente que el Texto 4 pertenece a las clases 'Religión' y 'Ciencias'.

Tabla 1. Representación de un conjunto de ejemplos multi-etiquetados.

Etiqueta / Textos	Deportes	Religión	Ciencias	Política
Texto 1	x			X
Texto 2			x	X
Texto 3	x			
Texto 4		x	x	

Los métodos para resolver este problema se encuentran agrupados en dos enfoques, los cuales son: transformación del problema y adaptación del modelo (Tsoumakas & Katakis, 2007). El enfoque de transformación del problema es independiente del algoritmo, es decir, transforma la tarea de aprendizaje de multi-etiqueta en una tarea de clasificación de una sola etiqueta. De esta forma, este método puede ser implementado usando algoritmos existentes. El método más común de transformación del problema, llamado Relevancia Binaria (*Binary Relevance*, BR) aprende $|L|$ clasificadores binarios $H_{\lambda_j}: X \to \{\lambda_j, \neg\lambda_j\}$, uno para cada etiqueta diferente λ_j en *L*. A través del uso de Relevancia Binaria se transforma el conjunto de datos original en $|L|$ conjuntos de datos D_{λ_j}. Cada D_{λ_j}etiqueta cada ejemplo de texto en *D* con λ_j si es que λ_j es contenida en el ejemplo o $\neg\lambda_j$ si el texto de ejemplo no contiene la etiqueta. BR brinda la misma solución, tanto para problemas de una sola etiqueta, como para problemas multi-etiqueta usando clasificadores binarios. Para la clasificación de una nueva instancia *x*, este método genera un set de etiquetas como la unión de las etiquetas generadas por el clasificador $|L|$, $H_{BR}(x) = \cup_{\lambda_j \in L}\{\lambda_j\} : H_{\lambda_j}(x) = \lambda_j$. Esta suele ser la transformación más común y es la misma solución utilizada cuando se pretende lidiar con un problema de clasificación multi-clase usando clasificadores binarios (ver Tabla 2).

Tabla 2. Transformación resultante utilizando Relevancia Binaria.

Etiqueta / Textos	Deportes	No deportes
Texto 1	X	
Texto 2		x
Texto 3	X	
Texto 4		x

Etiqueta / Textos	Religión	No religión
Texto 1	X	
Texto 2	X	
Texto 3		x
Texto 4		x

Etiqueta / Textos	Política	No política
Texto 1	X	
Texto 2	X	
Texto 3		x
Texto 4		x

Etiqueta / Textos	Ciencias	No ciencias
Texto 1		x
Texto 2	X	
Texto 3		x
Texto 4	X	

Otro tipo de transformación del problema se denomina Conjunto Potencia de Etiquetas (*Label Powerset,* LP). En esta transformación cada conjunto de etiquetas se considera una nueva categoría. Luego, si tenemos LP combinaciones, es posible utilizar $|LP|$ clasificadores binarios, uno para cada nueva etiqueta. El conjunto de datos se maneja como uno del tipo etiqueta simple para luego construir un clasificador de etiqueta simple con múltiples clases disjuntas (ver Tabla 3).

Tabla 3. Transformación del problema usando *Label Powerset*.

Etiqueta / Ejemplo	Deportes	Deportes y Política	Ciencias y Política	Ciencias y Religión
Texto 1		x		
Texto 2			X	
Texto 3	x			
Texto 4				x

El segundo método se ocupa de adaptar algunos modelos y algoritmos de aprendizaje específicos para que puedan manejar datos de multi-etiquetados directamente. Estas adaptaciones son logradas gracias a ajustes de los modelos, tales como modificaciones a formulaciones clásicas de estadísticas o teoría de la información. El pre-procesamiento de los documentos para lograr una mejor representación también puede ser considerado dentro de este tipo de transformación.

La clasificación automática de textos multi-etiqueta ha sido abordada también por medio de algoritmos, los cuales capturan directamente las características del problema de multi-etiqueta. Lee y Jiang (2014) proponen un método con base en lógica difusa, donde un texto con multi-etiquetado puede pertenecer a una, o más de una categoría. Los autores exponen que al incorporar técnicas difusas, el método puede sobreponerse a los problemas causados por los altos requerimientos de memoria o bajo rendimiento. Nam, Kim, Mencía, Gurevych y Fürnkranz (2014) se enfoca en resolver las limitaciones que tiene el algoritmo de *Backpropagation Learning* para que

pueda trabajar con datos multi-etiquetados y propone el *Backpropagation Multi-Label Learning* (BP-MLL), en esta propuesta se utiliza un enfoque de red neuronal muy simple para tareas de clasificación de textos multi-etiquetados a gran escala. Por otra parte, Murawaki (2013) propone un Modelo Global Jerárquico para Clasificación de Textos Multi-Etiquetados que busca aprovechar las relaciones de dependencia entre las etiquetas. En línea con la explotación del Aprendizaje Profundo, Liu, Chang, Wu y Yang (2017) plantea un modelo para la clasificación de texto de etiquetas múltiples extremas (XMTC), enfrentando el problema de asignar a cada documento el subconjunto más relevante de etiquetas de clase de una colección de etiquetas extremadamente grande, donde el número de etiquetas podría alcanzar a cientos de miles o millones.

Sin importar el enfoque de soluciones al problema de multi-etiqueta y los algoritmos que lo resuelvan, según Joachims (2002), cualquier tarea de clasificación de texto tiene complejidades debido a que el espacio de características es altamente multidimensional, un uso heterogéneo de los términos y un alto nivel de redundancia. Los problemas multi-etiqueta tienen complejidades adicionales, incluyendo un gran número de etiquetas y el desbalance de las mismas a través del conjunto de documentos.

Aunque las medidas tradicionales de evaluación del desempeño de los clasificadores automáticos, tal como las medidas macro-*F* y *Hamming Loss* son útiles en casos multi-etiqueta, también han surgido nuevas medidas de evaluación con la intención de analizar el desempeño en la asignación del conjunto de etiquetas que le corresponden a cada documento, tal como la exactitud del conjunto de etiquetas, medida denominada *Label-Set Accuracy* (Read, Pfahringer, Holmes & Frank, 2011).

El rendimiento de un sistema de razonamiento automático depende en gran parte de la representación del problema. La misma tarea puede ser fácil o difícil, dependiendo en la forma en la cual está descrita (Fink, 2004). La representación explícita de información relevante tiende a aumentar el desempeño de la máquina de aprendizaje. De esta manera, a través de una representación más compleja podrían obtenerse mejores resultados con algoritmos más simples. En esta línea, hoy se plantean mecanismos que permiten seleccionar y ponderar las características que mejoran el desempeño de un clasificador automático en contextos específicos (Kadhim, 2019). Así también, quienes plantean el análisis de los elementos visuales del texto como una característica adicional (Chatterjee et al., 2019).

En el caso particular de los documentos, la representación del texto tiene un alto impacto en la tarea de clasificación (Keikha, Razavian, Oroumchian & Razi, 2008). El modelo espacio vectorial es uno de los modelos más usados para recuperación de información, principalmente por su simplicidad conceptual y el atractivo de su metáfora subyacente, de usar una proximidad espacial para proximidad semántica

(Manning & Schütze, 1999). En el modelo de espacio vectorial (*Vector Space Model*), los contenidos de un documento son representados por un vector de términos $d = \{w_1, \ldots, w_k\}$, donde *k* es el tamaño del conjunto de términos w_i (o características). Algunos elementos usados en la representación de un texto son los N-gramas, palabras, frases, lógica de términos y declaraciones o cualquier otra unidad léxica, semántica y/o sintáctica que pueda ser utilizada para representar el contenido del texto.

Independientemente de las características usadas para representar un texto, a partir de la existencia de estas características se determinará a qué clases pertenece el texto, para ello, la medida más utilizada es el indicador de relevancia $f_{t,d}$, la cual representa cuánto contribuye la característica o término *t* a la semántica del documento *d*. Este tipo de representación, donde el indicador $f_{t,d}$ puede tener valores entre cero y uno ([0,1]) es llamado Modelo Multinomial por McCallum y Nigam (1998), y es diferente al Modelo Bernoulli Multi-variado, donde el indicador es $bin_{t,d}$, el cual es representado por uno cuando el término *t* existe al menos una vez en el documento *d*, es decir, puede tener valor cero o uno ($\{0,1\}$). El factor basado en el Modelo Bernoulli Multi-variado es llamado representación Binaria o modelo Booleano. Muchos problemas, ya sea por su naturaleza o por las medidas que se pueden obtener de ellos, utilizan el modelo de representación basado en el Modelo Bernoulli Multi-variado.

Existen diferentes formas de describir las características de un texto para que los diferentes clasificadores de textos puedan trabajar sobre ellos. Leopold y Kindermann (2002), por ejemplo, combinan transformaciones con diferentes funciones de *kernel* en máquinas de soporte vectorial. Por su parte, de acuerdo a Lan, Tan, Su y Lu (2009) se deben tomar dos decisiones importantes al escoger la representación basada en el modelo de espacio vectorial. 1) ¿qué debería constituir un término? ¿debería ser una raíz de palabra, una palabra, un conjunto de palabras o su significado? 2) ¿cómo deberá ser pesado o ponderado ese término? La ponderación podría ser por medio de una función binaria o de la frecuencia inversa en los documentos (*tf-idf*) desarrollada por Salton y Buckley (1988), usando métricas de selección de características como chi-cuadrado (χ^2), ganancia de información (IG), razón o ratio de ganancia (GR), etc. Los métodos de ponderación de términos mejoran la efectividad de la clasificación de textos a través de una apropiada selección de pesos para los términos. Aunque la clasificación de textos se ha estudiado durante varias décadas, los métodos de ponderación de los términos para la clasificación de textos suelen tomarse del campo de la recuperación de información (IR), incluyendo, por ejemplo, el modelo Booleano, *tf-idf* y sus variantes. En general, para ponderar los términos en el modelo de espacio vectorial, se puede utilizar la frecuencia de términos o frecuencia de documentos que contienen un término.

2. Nuestra propuesta

En esta sección se explica la bien conocida representación *tf-idf* (Salton & Buckley, 1988) y se presenta la representación basada en el modelo multinomial *tf-rfl*. Sobre la base de esta última, proponemos una nueva representación basada en el modelo Bernoulli Multinomial llamada *bin-rfl*. Se plantea con ello la hipótesis de que una modificación supervisada a la representación del texto que considere representaciones binarias, junto con una ponderación de los términos que está basada en los ejemplos conocidos, según sus etiquetas, puede mejorar significativamente el rendimiento de los clasificadores. Para el método de ponderación de términos para problemas de múltiples etiquetas usaremos como variables: a_{t,λ_j}, la cual representa el número de documentos en la categoría λ_j que contiene el término *t* y d_{t,λ_j} que representa el número de documentos en la categoría λ_j que no contiene el término *t*.

Tabla 4. Variables utilizadas para ponderar en un problema multi-etiqueta dado un término t y 4 categorías.

Categoría / Ejemplo	Deportes (1)	Religión (2)	Ciencias (3)	Política (4)
t (contiene el término)	$a_{t,1}$	$a_{t,2}$	$a_{t,3}$	$a_{t,4}$
$\bar{t}$ (no contiene el término)	$d_{t,1}$	$d_{t,2}$	$d_{t,3}$	$d_{t,4}$

2.1. Representación Term frequency-Inverse document frequency (tf-idf)

Según Sebastiani (2002), la representación de textos más utilizada para clasificación de textos es *tf-idf* de Salton y Buckley (1988). Donde, cada componente del vector es calculado según la Ecuación 1:

$$tf - idf_{td} = f_{t,d} \times \log_{10}\left(\frac{N}{N_t}\right), \quad (1)$$

donde $f_{t,d}$ es la frecuencia del término *t* en el documento *d*. Para el problema de dos categorías $N = (a_{t,\lambda_1} + d_{t,\lambda_1} + a_{t,\lambda_2} + d_{t,\lambda_2})$ es el número de documentos, y $N_t = (a_{t,\lambda_1} + a_{t,\lambda_2})$ es el número de documentos que contienen el término *t*.

La principal contribución de esta representación es que pondera con menor importancia los términos que son muy frecuentes en la colección de documentos a través de del factor N/N_t.

2.2. Representación Term frequency-Relevance frequency for a label (tf-rfl)

En la investigación realizada por el Alfaro y Allende (2011) se presentaron resultados preliminares de la representación *Relevance frequency for a label*, *tf-rfl*. Esta representación se describe en la siguiente ecuación, como una nueva representación para problemas multi-etiqueta.

$$tf - rfl_{tdl} = f_{t,d} \times \log_2\left(2 + \frac{a_{t,l}}{\max\left(1, mean(a_{t,\lambda_{j/l}})\right)}\right), \quad (2)$$

donde $f_{t,d}$ es la frecuencia del término *t* en el documento *d*, $a_{t,l}$ es el número de documentos bajo la categoría en evaluación *l* que contienen el término *t*, y $mean(a_{t,\lambda_{j/l}})$ es el promedio del número de documentos que contienen el término *t* entre el conjunto de documentos etiquetados en otra categoría diferente de *l*, es decir $a_{t,\lambda_{j/l}} = \{a_{t,\lambda_1},\ldots,a_{t,\lambda_{l-1}},a_{t,\lambda_{l+1}},\ldots,a_{t,|L|}\}$.

El valor constante 2 en el lado derecho de la fórmula se asigna porque la base de la operación logarítmica es 2. Sin la constante 2, podría tener el efecto de dar valor cero a otros términos.

La principal contribución de esta representación es que pondera con menor importancia los términos que son igualmente frecuentes en las diferentes categorías y pondera con mayor importancia los términos que son más frecuentes en la categoría bajo evaluación.

También es posible utilizar *bin-idf* basado en el modelo de múltiples variables Bernoulli en lugar de *tf-idf* basado en el modelo Multinomial. En este caso, en lugar de utilizar $f_{t,d}$ se utiliza $bin_{t,d}$.

Con el fin de evaluar la mejora del desempeño debido al uso de la ponderación *rfl*, este trabajo presentará una nueva representación basada en la aparición de términos en cada documento, es decir, representación binaria o representación booleana. Esta representación, basada en el Modelo de Bernoulli Multivariado, utiliza menos información que la basada en el Modelo Multinomial, ya que solo se utiliza la información de existencia o no de una palabra en el texto y no su frecuencia de aparición.

2.3. Representación Multi-variate Bernoulli Model - Label Dependent (bin-rfl)

La nueva representación para el problema de etiquetado múltiple, que se propone en este trabajo, llamado *bin-rfl*, se basa en una representación del modelo de Bernoulli

multivariado que se pondera usando el término frecuencia de una etiqueta y se calcula como en la Ecuación número 3:

$$bin - rfl_{tdl} = bin_{t,d} \times \log_2\left(2 + \frac{a_{t,l}}{\max\left(1, mean(a_{t,\lambda_{j/l}})\right)}\right) \quad (3)$$

Donde $bin_{t,d}$ toma el valor 1, si el término *t* está presente en el documento *d* y 0, si el término *t* no está presente en el documento *d*, $a_{t,l}$ es el número de documentos de la categoría bajo evaluación que contiene el término *t*, y $mean(a_{t,\lambda_{j/l}})$ es el número promedio de documentos que contienen el término *t* para cada el conjunto de documentos etiquetados distinto de *l*. Esta nueva representación ayuda a hacer una mejor distinción de los términos, que se refleja en una mejor clasificación de rendimiento, como se verá en la sección resultados.

2.4. Método de clasificación

La propuesta de método de ponderación de términos incluye información de la frecuencia de ocurrencia de un término *t* en cada grupo de documentos etiquetados con otras etiquetas distintas de las que están bajo evaluación. Con *rfl*, se espera que $mean(a_{t,\lambda_{j/l}})$ sea mayor si el término *t* aparece más frecuentemente en documentos con etiqueta $\lambda_j = 1$ que en documentos con otras etiquetas $\lambda_{j/l}$, y tendrá un valor más bajo, si el término *t* es más frecuente en documentos con etiquetas que no sean *l*. De este modo, la ponderación *rfl* resulta ser un mejor discriminador entre categorías.

Nuestra propuesta se basa en las representaciones *bin-rfl* y en los clasificadores binarios, utilizando las transformaciones del problema: Relevancia Binaria y *Label Powerset*. Para ello se transforma el problema de etiquetado múltiple en problemas binarios, y luego para cada documento *d* se construyen las representaciones *bin-rfl* para cada etiqueta y se clasifican usando clasificadores binarios. Se debe destacar que cada documento está representado con un vector diferente cuando está bajo evaluación de cada etiqueta, porque el factor de ponderación depende de la etiqueta bajo evaluación.

3. Marco metodológico

3.1. Conjuntos de datos

Para evaluar la representación, se utilizaron cuatro conjuntos de datos textuales multietiquetados conocidos: REUTERS-21578, OHSUMED, ENRON y MEDICAL. Para REUTERS-21578, que es un conjunto de textos de noticias, se consideró un subconjunto modificado que se propuso en Read et al. (2011) con el fin de poder obtener medidas de desempeño comparativas. El conjunto de datos OHSUMED es una partición de la base de datos MEDLINE, que es una biblioteca de artículos científicos publicados en revistas médicas. La colección OHSUMED también se ha

reducido de 50.216 a 13.929 textos. Este subconjunto contiene las 10 categorías más representativas de las 23 categorías originales. El conjunto de datos de Enron es una colección de textos creados por el proyecto CALO (*Cognitive Assistant that Learns and Organizes*), que contiene 1.702 mensajes de correo electrónico y 52 categorías. Por último, el conjunto de datos Medical fue creado por la *Computational Medicine Center*, 2007 a propósito del *Language Processing Challenge,* 2007, contiene 978 textos clínicos de informes de radiología y considera 45 categorías de códigos médicos. La Tabla 5 presenta las características del conjunto de datos pre-procesados.

Tabla 5. Características del conjunto de datos pre procesados. LabelCard indica el número promedio de etiquetas para cada documento y el Tamaño del Vocabulario considera el volumen de palabras distintas.

Data Set	Número de Etiquetas	Número de Documentos	Tamaño del Vocabulario	LabelCard
Ohsumed	23	13.929	1.002	1,66
Reuters	103	6.000	500	1,46
Enron	52	1.702	1.001	3,38
Medical	6	593	72	1,25

3.2. Medidas de desempeño

Las medidas tradicionales de evaluación como la medida *F* y la *Hamming Loss* son útiles en el caso de conjuntos multi-etiquetados. Sin embargo, como ya hemos planteado, existen nuevas medidas destinadas a evaluar el desempeño en la asignación del conjunto de etiquetas, como lo es la precisión del conjunto de etiquetas (*Label-Set Accuracy*).

Para describir las medidas de desempeño, se utiliza la siguiente notación: considerando el vector $\mathbf{Y_i} \in \{0,1\}^{|L|}: i = 1 \dots d$, entonces cada etiqueta será relevante si $y_{i,j} = 1$, y por su parte, la predicción del clasificador automático será $y'_{i,j} = 1$, donde *d* es el número de documentos y $|L|$ es el número de posibles etiquetas.

Basándose en la notación anterior, *Hamming Loss* se define como en la Ecuación 4:

$$Hamming - Loss(\mathbf{Y}, \mathbf{Y}') = \frac{1}{d} \frac{1}{|L|} \sum_{i=1}^{d} \sum_{j=1}^{|L|} |y'_{i,j} \Delta y_{i,j}| \text{ , (4)}$$

Esta medida busca medir la diferencia entre cada etiqueta que los textos realmente tienen, con cada etiqueta que asignó el clasificador automático a dichos textos. Mientras más bajo es el valor obtenido, mejor es el desempeño.

Otra medida multi-etiqueta es la precisión del conjunto de etiquetas (*Label-Set Accuracy*) y es definida como en la Ecuación 5:

$$Label - Set - Accuracy\ (D) = \frac{1}{d} \sum_{i=1}^{d} \frac{\mathbf{Y_i} \cap \mathbf{Y'_i}}{\mathbf{Y_i} \cup \mathbf{Y'_i}} \text{ , (5)}$$

Esta media de desempeño permite medir, para cada texto, la razón entre las etiquetas que el texto tiene y las etiquetas que el clasificador asignó. Mientras más alto es el valor obtenido, mejor es el desempeño.

La medida *F,* comúnmente utilizada en recuperación de información, es muy popular en clasificación de textos multi-etiquetados. La medida *F* es la media armónica entre precisión y exhaustividad (*recall*). La medida *F* (F_1) para cada etiqueta se calcula como se muestra en la Ecuación 6:

$$F_1(\mathrm{Y_i}, \mathrm{Y'_i}) = \frac{2 \times precision \times recall}{precision + recall}, \quad (6)$$

donde la *precisión* es la fracción de las predicciones que son realmente relevantes y *recall* es la fracción de relevancia real con respecto a las predicciones. Mientras más alto es el valor de *F*, mejor es el desempeño.

Para el caso multi-etiqueta es necesario combinar los diferentes F_1 de cada evaluación de la etiqueta. Para eso usamos macro-F_1 que es el promedio del F_1 para cada etiqueta.

3.3. Transformación del problema

En esta investigación se ha seleccionado el enfoque de transformación del problema (Tsoumakas & Katakis, 2007), que consiste en tomar textos multi-etiquetados y cambiar su representación para utilizar técnicas de una clase. Para ello se utilizaron dos técnicas, la primera que se utilizó es la Relevancia Binaria (BR), que se basa en la creación de nuevos conjuntos binarios que representan cada una de las categorías, para las que si el documento pertenece a la categoría se le da un valor de 1, y un valor 0 para las categorías a las que no pertenece. A partir de lo anterior, el conjunto de datos multi-etiquetados original se trasforma en varios conjuntos nuevos con etiqueta simple. La segunda transformación del problema es *Label Powerset* (LP), que es la creación de nuevas etiquetas a partir del conjunto de etiquetas originales del problema, por lo tanto cada documento se asocia a una nueva etiqueta única, el enfoque se basa ahora en un problema de etiqueta simple.

3.4. Clasificadores utilizados

Para evaluar la mejora del rendimiento de la función de ponderación *rfl*, usamos solo clasificadores lineales, la máquina de soporte vectorial (SVM) con kernel lineal y redes neuronales artificiales con una capa (ANN). Para ambos clasificadores usamos implementaciones de SVM y ANN en el *software* WEKA descrito en Hall, Frank, Holmes, Pfahringer, Reutemann y Witten (2009). De las distintas alternativas para

evaluar el desempeño, hemos utilizado la validación cruzada de 3 veces y reportamos el promedio de 3 ciclos de clasificación.

3.5. Resultados

Con el objetivo de comparar el efecto de modificar la representación, realizamos diferentes experimentos de clasificación utilizando cuatro conjuntos de datos diferentes (Reuters, Ohsumed, Enron, Medical), utilizando representaciones *bin-rfl* y *bin-idf*, ambas con transformaciones de Relevancia Binaria y la *Label Powerset*, y dos clasificadores lineales diferentes (SVM y ANN). Las Tablas 6 a 11 muestran los diferentes métodos y su desempeño en términos de diferentes medidas de desempeño descritas previamente.

Tabla 6. Resultados experimentales de diferentes trasformaciones del problema (TP: BR y LP) y Representaciones con SVM en términos de *Label-Set Accuracy*.

TP	Representación	Reuters	Ohsumed	Enron	Medical
BR	*bin-idf*	0,319	0,402	0,381	0,73
BR	*bin-rfl*	0,326	0,414	0,406	0,752
LP	*bin-idf*	0,303	0,381	0,376	0,71
LP	*bin-rfl*	0,317	0,401	0,397	0,705

Tabla 7. Resultados experimentales de diferentes trasformaciones del problema (TP: BR y LP) y Representaciones con ANN en términos de *Label-Set Accuracy*.

TP	Representación	Reuters	Ohsumed	Enron	Medical
BR	*bin-idf*	0,291	0,388	0,375	0,704
BR	*bin-rfl*	0,313	0,395	0,393	0,736
LP	*bin-idf*	0,293	0,373	0,369	0,703
LP	*bin-rfl*	0,305	0,388	0,371	0,694

Acerca de los clasificadores, como se espera en este tipo de problema, SVM supera siempre el modelo de ANN. Relevancia Binaria en general tiene un mejor rendimiento que la *Label Powerset*, a menos que la evaluación sea en términos de *Hamming Loss*, donde algunos conjuntos de datos LP tienen un mejor rendimiento que BR.

Tabla 8. Resultados experimentales de diferentes trasformaciones del problema (TP: BR y LP) y Representaciones con SVM en términos de *Hamming Loss*.

TP	Representation	Reuters	Ohsumed	Enron	Medical
BR	*bin-idf*	0,0663	0,064	0,057	0,0109
BR	*bin-rfl*	0,0657	0,0631	0,0562	0,0102
LP	*bin-idf*	0,0625	0,0601	0,059	0,0115
LP	*bin-rfl*	0,0621	0,0593	0,0577	0,011

Tabla 9. Resultados experimentales de diferentes trasformaciones del problema (TP: BR y LP) y Representaciones con ANN en términos de *Hamming Loss*.

PT	Representation	Reuters	Ohsumed	Enron	Medical
BR	*bin-idf*	0,0669	0,0672	0,0593	0,0113
BR	*bin-rfl*	0,0662	0,0644	0,0583	0,0111
LP	*bin-idf*	0,0633	0,0623	0,0616	0,0121
LP	*bin-rfl*	0,0637	0,0603	0,058	0,0167

Acerca de la representación, en casi todos los casos, la representación *bin-rfl* presenta mejoras a *bin-idf*. Como se muestra en las Tablas 6 y 7, se obtiene una mejora promedio superior al 4% (con SVM) y al 5% (con ANN) en términos de *Label-Set-Accuracy*. Del mismo modo, se obtienen mejoras del 2% en términos de *Hamming-Loss*, como se muestra en las Tablas 8 y 9. Finalmente, como se puede observar en las Tablas 10 y 11, la mejora del rendimiento en términos de F_1 del 10% (con SVM) y 11% (con ANN) promediado usando la representación *bin-rfl* en lugar de *bin-idf* y la transformación Relevancia Binaria.

Tabla 10. Resultados experimentales de diferentes trasformaciones del problema (TP: BR y LP) y Representaciones con SVM en términos de *macro-*F_1.

PT	Representation	Reuters	Ohsumed	Enron	Medical
BR	*bin-idf*	0,213	0,379	0,197	0,354
BR	*bin-rfl*	0,238	0,426	0,227	0,372
LP	*bin-idf*	0,219	0,361	0,18	0,339
LP	*bin-rfl*	0,219	0,42	0,221	0,324

Tabla 11. Resultados experimentales de diferentes trasformaciones del problema (TP: BR y LP) y Representaciones con ANN en términos de *macro-*F_1.

PT	Representation	Reuters	Ohsumed	Enron	Medical
BR	*bin-idf*	0,204	0,35	0,184	0,342
BR	*bin-rfl*	**0,229**	**0,411**	**0,218**	**0,357**
LP	*bin-idf*	0,184	0,349	0,173	0,33
LP	*bin-rfl*	0,21	0,402	0,209	0,316

Para finalizar el análisis de resultados experimentales, en Gráfico 1 se muestra gráficamente como en casi todos los casos, la representación *bin-rfl* presenta mejoras significativas en relación a *bin-idf*. Este porcentaje se calcula como la razón entre la diferencia de la métrica con la nueva representación y la antigua representación. A partir de ella, se puede observar que las mejoras, en muchos casos son superiores al 10%, en términos de *macro-*F_1.

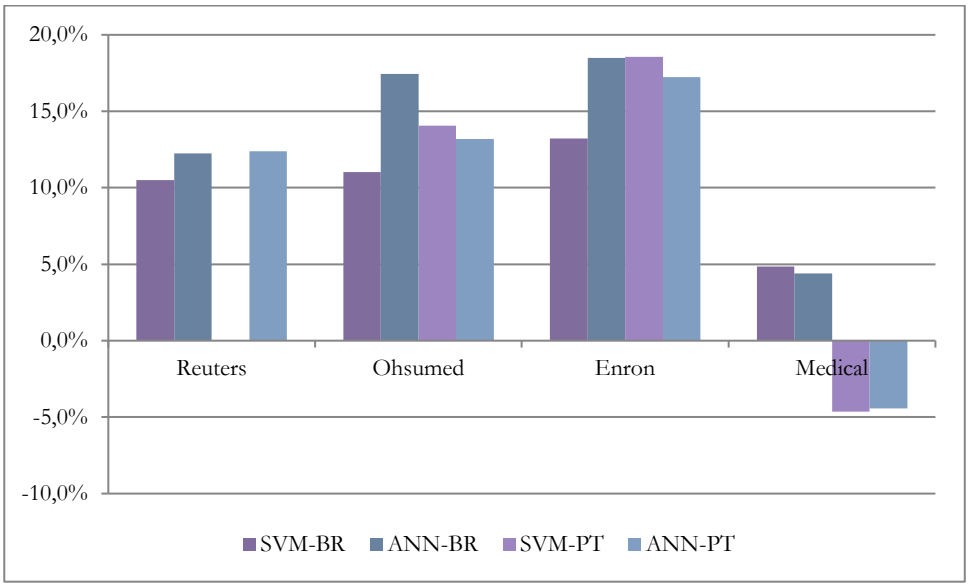

Gráfico 1. Porcentaje de mejora del desempeño en términos de *macro-F_1*.

Para evaluar los resultados como en Alfaro y Allende (2011), se implementó una prueba basada en *two-tailed paired t-test* al nivel de significación del 5%. De acuerdo con estos resultados, la transformación del problema de Relevancia Binaria, con SVM y *bin-rfl* es mejor que la Relevancia Binaria con SVM y *bin-idf* en todas las medidas ($p = 0{,}0296$ para la Precisión del Conjunto de Etiquetas, $p = 0{,}0014$ para *Hamming Loss* y $p = 0{,}0138$ para F_1). El valor de p mostrado entre paréntesis proporciona una cuantificación adicional del nivel de significación.

CONCLUSIONES

La clasificación automática con varias etiquetas es un tema importante en la recuperación de la información y el aprendizaje automático. La representación y clasificación de textos se han tratado tradicionalmente usando *tf-idf* debido a su simplicidad y buen desempeño.

Los cambios en la representación de entrada a los clasificadores automáticos pueden emplear conocimientos sobre el problema, una etiqueta en particular o la categoría a la que pertenece el documento. La representación *bin-rfl* puede ser desarrollada para resolver un problema particular directamente, sin complejas transformaciones de problemas. En este trabajo, hemos presentado el *bin-rfl* como una nueva representación de texto para el enfoque de clasificación multi-etiqueta. Esta representación permite discriminar los términos que mejor describen una categoría, en contraste con otras categorías, aprovechando así las características del dominio de documentos que conforman el corpus. Esta propuesta se evaluó utilizando dos diferentes clasificadores lineales en cuatro conjuntos de textos diferentes, que

corresponden a artículos científicos médicos, documentos periodísticos, informes de diagnósticos médicos y mensajes de correo electrónico. Se realiza una comparación con *bin-idf* y se utilizan dos transformaciones del problema multi-etiquetado (Relevancia Binaria y *Label Powerset*). El desempeño de esta representación muestra una mejora en todos los casos, utilizando Relevancia Binaria y SVM. Solo en la medida *Hamming Loss*, fue mejor usar *Label Powerset* y SVM. Creemos que la contribución del factor de ponderación *rfl* para la representación multi-etiqueta se debe a una mejor resolución del problema considerado, ya que es capaz de hacer una mejor identificación de los términos en los documentos, lo que se refleja en un mejor rendimiento de los modelos de clasificación automática. En futuros estudios, planeamos utilizar la representación *bin-rfl* para la tarea de selección de características o identificación de atributos más significativos para discriminar. También vamos a utilizar otras representaciones, por ejemplo *Part of Speech*, N-gramas o basadas en otras distribuciones de probabilidad para construir una representación dependiente de etiqueta. Finalmente, usaremos la representación para realizar análisis de opiniones, clasificación de correo electrónico y otras aplicaciones de reconocimiento de patrones.

REFERENCIAS BIBLIOGRÁFICAS

Alfaro, R. & Allende, H. (2011). Text representation in multi-label classification: Two new input representations. En A. Dobnikar, U. Lotrič & B. Šter (Eds.), Adaptive and Natural Computing Algorithms. ICANNGA 2011. *Lecture Notes in Computer Science* (pp- 61-70). Springer: Berlin, Heidelberg.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*(1), 63-88.

Chatterjee, A., Gupta, U., Chinnakotla, M. K., Srikanth, R., Galley, M. & Agrawal, P. (2019). Understanding emotions in text using deep learning and Big Data. *Computers in Human Behavior*, 93, 309-317.

Ciapuscio, G. (1994). *Tipos textuales*. Universidad de Buenos Aires: Buenos Aires.

Ciapuscio, G. (2000). Hacia una tipología del discurso especializado. *Discurso y Sociedad*, *2*(2), 39-71.

Fenton, F. (1911). The influence of newspapers presentations upon the growth of crime and other anti-social activity. *American Journal of Sociology*, *16*(3), 342-371.

Fink, E. (2004). Automatic evaluation and selection of problem-solving methods: Theory and experiments. *Journal of Experimental and Theoretical Artificial Intelligence, 16*(2), 73-105.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). *The weka data mining software: an update. SIGKDD Explor. Newsl*, *11*(1), 10-18.

Ignaton, G. & Mihalcea, R. (2017). *Text mining: A guidebook for the social sciences*. Los Angeles: SAGE Publications.

Joachims, J. (2002). *Learning to classify text using support vector machines: Methods, theory, and algorithms*. Dordrecht: Kluwer Academic.

Kadhim A. I. (2019). Term weighting for feature extraction on Twitter: A comparison between BM25 and TF-IDF. En actas de la *International Conference on Advanced Science and Engineering (ICOASE)* (pp. 124-128). Kurdistán: University of Zakho and Duhok Polytechnic University.

Keikha, M., Razavian, N., Oroumchian, F. & Razi, H. S. (2008). (Eds). Document representation and quality of text: An analysis. En *Survey of Text Mining II: Clustering, Classification, and Retrieval* (pp. 135-168). Londres: Springer-Verlag.

Lan, M., Tan, C. L., Su, J. & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 31, 721-735.

Lavalle, J., Montes, M., Villaseñor, L., Jiménez, H. & Bárcenas, E. (2018). Equivalences between polarity algorithms. *Studia Logica*, *106*(2), 371-395.

Lee, S. & Jiang, J. (2014). Multilabel text categorization based on fuzzy relevance clustering. *Fuzzy Systems IEEE Transactions*, *22*(6), 1457,1471.

Leopold, E. & Kindermann, J. (2002). Text categorization with support vector machines. How to represent texts in input space? *Machine Learning, 46*(1-3) 423-444.

Liu, J., Chang, W-Ch., Wu, Y. & Yang, Y. (2017). Deep learning for extreme multi-label text classification. En actas del *40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '17) (pp. 115-124). Nueva York, NY USA: ACM.

Loureda, O. (2003). *Introducción a la tipología textual.* Madrid: Arco Libros.

Manning, C. & Schütze, H. (1999). *Foundations of statistical natural language Processing*. Boston: The MIT Press.

McCallum, A. & Nigam, K. (1998). A comparison of event models for naive bayes text classification. En actas *AAAI-98 workshop on learning for text categorization, Madison, WI* (41-48). Menlo Park, CA: AAI Press.

Murawaki, Y. (2013). Global model for hierarchical multi-label text classification. *International Joint Conference on Natural Language Processing*, 46-54.

Nam, J., Kim, J., Mencía, E., Gurevych, I. & Fürnkranz, J. (2014). Large-scale multi-label text classification -Revisiting Neural Networks. *Machine Learning and Knowledge Discovery in Databases, ECML PKDD*, 437-452.

Ortega, R. M., López-Monroy, A. P., Franco, A. & Montes-y-Gómez, M. (2018). Emphasizing personal information for author profiling: New approaches for term selection and weighting. *Knowledge-Based Systems*, 145, 169-181.

Read, J., Pfahringer, B., Holmes, G. & Frank E. (2011). Classifier chains for multi label classification. *Machine Learning,* 85, 333-359.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management: An International Journal, 24*(5), 13-523.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Survey*, *34*(1), 1-47.

Tsoumakas, G. & Katakis, I. (2007). Multi label classification: An overview. *International Journal of Data Warehouse and Mining, 3*(3), 1-13.

Venegas, R. (2007). Clasificación de textos académicos en función de su contenido léxico-semántico. *Revista Signos. Estudios de Lingüística*, *40*(63), 239-271.

Venegas, R. (2019). Clasificación automatizada de macromovidas discursivas en el género tesis: Escritura académica y aprendizaje de máquinas. En C. Zapata & B. Manríquez (Coords.), *Tecnologías del lenguaje Humano: Aplicaciones desde la lingüística computacional y de corpus*. Medellín: Editorial de la Universidad de Medellín.

White, P. W. (1924). Quarter century survey of press content shows demand for facts. *Editor and Publisher,* 57.

Wilcox, D. F. (1900). The American newspaper: A study in social psychology. *The ANNALS of the American Academy of Political and Social Science*, *16*(1), 56-92.