Revista Signos
ISSN: 0035-0451
revista.signos@ucv.cl
Pontificia Universidad Católica de Valparaíso
Chile

Mendoza, Marcelo; Ortiz, Ivette; Rojas, Víctor
Categorización de texto en bases documentales a partir de modelos computacionales livianos
Revista Signos, vol. 44, núm. 77, diciembre, 2011, pp. 251-274
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile

Revista Signos ISSN 0718-0934
© 2011 PUCV, Chile
DOI: 10.4067/S0718-09342011000300004

44(77) 251-274

Categorización de texto en bases documentales a partir de modelos computacionales livianos*

Text categorization in documentary databases using light computational models

Marcelo Mendoza
marcelo.mendoza@usm.cl
Universidad Técnica Federico Santa María
Chile

Ivette Ortiz
iortiz@iale.cl
IALE Tecnología
Chile

Víctor Rojas
vrojas@iale.cl
IALE Tecnología
Chile

Recibido: 14-IV-2010 / Aceptado: 20-IV-2011

Resumen: En este trabajo se presenta un nuevo categorizador de texto para bases de datos documentales. El categorizador propuesto corresponde a una extensión del categorizador *Naive Bayes* que permite obtener buenos resultados en bases documentales con desbalance en datos de entrenamiento. Resultados experimentales permiten afirmar que el categorizador supera a *Naive Bayes* y se compara favorablemente con otras técnicas más sofisticadas como máquinas de soporte vectorial y regresión logística sin incurrir en costos computacionales significativos en la fase de entrenamiento.

Palabras Clave: Categorización de texto, modelos *Bayesianos*, recuperación de información.

Abstract: We introduce a new text categorization method for documentary databases. The proposed method is an extension of the Naive Bayes text categorization model which allows obtaining good performance results in documentary databases with unbalanced training data. Experimental results allow us to conclude that the categorization method overcomes Naive Bayes and compares favorably with more sophisticated categorization methods such as support vector machines and logistic regression without increasing the use of computational resources in the training phase.

Key Words: Text categorization, Bayesian models, information retrieval.

INTRODUCCIÓN

El explosivo crecimiento de la información disponible en documentos digitales ha forzado el desarrollo de nuevas tecnologías que faciliten la realización de procesos de búsqueda de forma eficiente y efectiva. Es frecuente que para facilitar la búsqueda de información se proceda a la categorización de los documentos en un conjunto acotado de clases. Estas clases permiten representar áreas específicas del conocimiento y son generalmente consolidadas por expertos.

Muchas tareas relacionadas con categorización de documentos han sido tradicionalmente abordadas de forma manual. Por ejemplo, la administración del catálogo de una biblioteca puede ser gestionada por bibliotecólogos u otros expertos que en base a su conocimiento categorizan los documentos. La categorización manual de documentos es costosa y se puede volver poco práctica en escenarios en los cuales los volúmenes de información que deban manejarse crezcan sin mediar control.

Con el surgimiento de la *World Wide Web* (www) a principios de la década de los 90, más y más colecciones documentales han quedado al alcance de los usuarios. En este contexto, el diseño de estrategias que permiten categorizar automáticamente los contenidos de los documentos se ha vuelto una necesidad cada vez más prioritaria. El uso de técnicas provenientes del procesamiento estadístico del lenguaje natural y de la inteligencia artificial ha mostrado alta efectividad.

Las técnicas de categorización automática de texto utilizan algoritmos que son capaces de aprender reglas y/o criterios de clasificación a partir de datos y ejemplos de entrenamiento. La necesidad de categorización manual no es completamente sustituida por las técnicas de categorización automática. De hecho, muchas de estas técnicas requieren de la construcción de datos de entrenamiento, lo cual consiste en la categorización manual por parte de un grupo de expertos de un conjunto reducido de documentos de ejemplo. Usando los datos de entrenamiento se espera que los categorizadores puedan identificar patrones y/o criterios que permitan categorizar nuevos documentos.

La categorización automática de documentos tiene aplicaciones en diversos problemas, como por ejemplo, detección de *e-mail spam,* detección de contenido sexual explícito en sitios/páginas *Web*, mantención de directorios o búsqueda en colecciones documentales enfocadas a temas específicos (búsqueda vertical). Actualmente, la mayoría de los sistemas de recuperación de información contienen múltiples componentes que usan categorización automática de documentos.

Uno de los primeros categorizadores automáticos de documentos fue diseñado a partir del modelo *Naive Bayes* (Maron & Kuhns, 1960). *Naive Bayes* utiliza una representación de los documentos computacionalmente liviana, basada en la aproximación 'bolsa de palabras', denominada en inglés *bag-of-words,* la cual asume independencia estadística en la co-ocurrencia de términos en un mismo documento. Con posterioridad, McCallum y Nigam (1998) estudiaron la factibilidad de otras representaciones basadas en modelos *Bayesianos* multinomiales y de *Bernoulli*. En la actualidad existen diversas extensiones de *Naive Bayes*, las cuales han mostrado obtener buenos resultados de categorización sin introducir costos computacionales altos.

Un segundo conjunto de estrategias de categorización utilizan técnicas que involucran costos computacionales mayores. En este conjunto destacan el algoritmo *Rocchio* (Rocchio, 1971), *k* vecinos más cercanos (Hastie, Tibshirani & Friedman,

2001) y máquinas de soporte vectorial (Vapnik, 1998). Lewis, Yang, Rose y Li (2004) mostraron que en la colección de referencia RCV1[1], los categorizadores basados en máquinas de soporte vectorial superaban en precisión a *Rocchio* y *k* vecinos más cercanos. Posteriormente, Venegas (2007) mostró que en bases documentales de escala media las máquinas de soporte vectorial también obtenían mejores resultados que el categorizador *Naive Bayes*. Sin embargo, muchas veces en sistemas de recuperación de información de escala media, como por ejemplo motores verticales, la introducción de nuevos costos computacionales hace poco práctico el uso de técnicas de este tipo.

En motores de búsqueda de gran escala como Google o en directorios comerciales importantes como Yahoo!, es posible asumir mayores costos computacionales para mejorar los resultados obtenidos. En otros escenarios, donde se dispone de menos recursos computacionales, el factor costo computacional es crítico. Es el caso de la mayoría de los buscadores verticales, como TodoCL[2] o Inquiro[3].

En este trabajo proponemos abordar el problema de categorización de texto en bases documentales de escala media. Para ello introduciremos una nueva variación al modelo *Naive Bayes* que permita mejorar el desempeño en precisión con respecto al método original. Para lograr lo anterior, introduciremos en la representación variables que permiten cuantificar la capacidad descriptiva de cada término con respecto a la colección documental así como su capacidad discriminativa con respecto a la clase en particular. También usaremos una combinación de coeficientes descriptivos/discriminativos. El estado del arte muestra que algunos de los coeficientes cubiertos en este trabajo han sido estudiados previamente con buenos resultados. En este trabajo en particular se estudia por primera vez el desempeño de estos coeficientes de forma conjunta. Otro aspecto novedoso de este trabajo es la introducción del coeficiente de frecuencia inversa desagregado por clase, el cual es estudiado por primera vez en este problema. Mostraremos que la extensión realizada al modelo *Naive Bayes* no introduce nuevos costos computacionales significativos evitando incurrir en el uso de recursos computacionales en fase de entrenamiento y logrando desempeños comparables a los de categorizadores basados en aprendizaje como es el caso de máquinas de soporte vectorial y regresión logística.

El resto del artículo se organiza de la siguiente forma. En la Sección 1 presentamos el apartado antecedentes. En la Sección 2 discutimos el marco conceptual de categorización de texto *Naive Bayes*. En la Sección 3 introducimos las modificaciones propuestas al categorizador *Naive Bayes*. La Sección 4 presenta resultados experimentales. Finalmente, en la Sección 5 discutimos conclusiones y trabajo futuro.

1. Antecedentes

1.1. Categorización de texto basada en modelos Bayesianos

Uno de los primeros categorizadores de texto fue presentado por Maron y Kuhns (1960). El categorizador está basado en el modelo *Naive Bayes*, el cual requiere del supuesto de independencia estadística en la co-ocurrencia de términos en un mismo documento. Cada documento es representado por el conjunto de términos que constituyen su texto. Cada término es representado por un nodo en un grafo dirigido. Un nodo adicional representa a las categorías. Desde cada nodo que representa un término se subtiende un arco hacia el nodo que representa a las categorías. Cada arco es etiquetado por un coeficiente cuyo valor representa la probabilidad de observar dicho término en el documento asumiendo la pertenencia del documento a una determinada categoría. El documento es asignado a la categoría que obtiene la máxima probabilidad de pertenencia.

McCallum y Nigam (1998) probaron dos formulaciones del modelo *Naive Bayes*. Una de ellas, el modelo *Bernoulli* pondera la probabilidad de los términos que ocurren en el texto de un documento con un factor 1 y con 0 los que no ocurren. La segunda formulación, conocida como modelo multinomial, pondera la probabilidad de los términos considerando la frecuencia que estos tienen en cada documento. La precisión de estos modelos sobre distintas colecciones de evaluación fue estudiada con resultados diversos siendo en general superior el desempeño logrado por el modelo multinomial. Entre los factores que explican los resultados obtenidos se encuentra el largo de cada documento, el cual fue estudiado por Bennett (2000), quien atribuyó a este factor la tendencia a obtener estimaciones de las probabilidades *Bayesianas* cercanas a 0 o 1.

Otro factor importante que explica la diversidad de resultados obtenidos usando *Naive Bayes* es la presencia de términos muy frecuentes en los textos, los cuales aminoran el efecto de términos menos frecuentes aun cuando estos tengan una mayor capacidad discriminativa. Rennie, Shih, Teevan y Karger (2003) modificaron el modelo *Naive Bayes* multinomial usando una función de suavizado sobre la frecuencia de los términos, buscado evitar que la presencia de términos muy frecuentes afectara el desempeño del clasificador. Resultados experimentales mostraron que usando esta modificación era posible obtener mejoras sobre *Naive Bayes* multinomial. Posteriormente, Schneider (2005) propuso remover la frecuencia de los términos de la formulación de *Naive Bayes*, lo cual también muestra mejoras sobre *Naive Bayes* multinomial. También aplicaron una estrategia de selección de características, permitiendo construir representaciones de los documentos usando solo aquellos términos que describían de mejor forma a cada clase. Kolcz y Yih (2007) estudiaron una variación sobre el categorizador propuesto por Rennie et al. (2003), obteniendo mejoras en el desempeño. Recientemente, Qiang (2010) propuso usar una nueva función de suavizado sobre la frecuencia de los términos, lo cual también muestra mejoras sobre *Naive Bayes* multinomial.

La aplicabilidad de *Naive Bayes* a la categorización en bases documentales de escala mediana fue estudiada por Venegas (2007). En ese trabajo se presentó el efecto del largo de los documentos en el desempeño de *Naive Bayes*. En esta misma dirección, Wilbur y Kim (2009) concluyeron que en bases documentales de gran escala, y por tanto con grandes vocabularios, el uso de la frecuencia de los términos en cada documento empeoraba el desempeño de los categorizadores *Naive Bayes* multinomiales. Sin embargo, el trabajo argumenta que no existe evidencia que permita mostrar que esto ocurre también en colecciones pequeñas o de escala mediana.

1.2. Categorización de texto basada en técnicas no Bayesianas

Hastie et al. (2001) presentan una completa descripción de técnicas de categorización basadas en *k* vecinos más cercanos. Estas técnicas utilizan un conjunto de ejemplos de entrenamiento representados a través de vectores en un espacio de alta dimensionalidad. Usando una función de distancia, para cada nuevo ejemplo a clasificar se determinan los *k* vecinos más cercanos determinando la categoría a la cual pertenece por mayoría simple. Perkins, Lacker y Theiler (2003) mostraron que en varias colecciones de evaluación, *k* vecinos más cercanos superaba en precisión a *Naive Bayes*. Indyk (2004) ha estudiado el comportamiento de *k* vecinos más cercanos en espacios de alta dimensionalidad, concluyendo que se introducen costos computacionales significativos en las estructuras de datos necesarias para mantener los ejemplos de entrenamiento en memoria. Datar e Indyk (2004) han estudiado como reducir los costos computacionales de mantención de las estructuras de datos que soportan *k* vecinos más cercanos introduciendo estructuras en memoria principal como tablas de *hash*.

Las máquinas de soporte vectorial (Vapnik, 1998) fueron usadas en categorización de texto por Zhang y Oles (2001), quienes estudiaron las ventajas/desventajas de máquinas de soporte vectorial frente a *Naive Bayes*. Fundamentalmente, las máquinas de soporte vectorial buscan determinar el hiperplano de máxima separación entre dos categorías dadas a partir de un conjunto de ejemplos de entrenamiento. Una de las dificultades de esta técnica es que suponen separabilidad lineal entre las categorías. Para abordar esta limitación, el estado del arte propone utilizar funciones de *kernels* que permitan representar a los objetos en espacios de dimensionalidad más alta en la cual la condición de separabilidad lineal pueda relajarse. Otra dificultad radica en que las máquinas de soporte vectorial y sus variantes están parametrizadas por lo que se requiere de un proceso de ajuste de parámetros, lo cual introduce altos costos computacionales. Joachims (2006) ha abordado estas limitaciones estudiando variaciones que permitan entrenar máquinas de soporte vectorial sin incorporar costos computacionales tan altos. Lewis et al. (2004) han mostrado que en la colección de evaluación RCV1 las máquinas de soporte vectorial obtienen mejores resultados que *k* vecinos más cercanos. Resultados experimentales corroboran lo anterior usando evaluaciones sobre la colección de referencia TREC (Voorhees & Harman, 2005) y TDT4 (Ault & Yang, 2002). Recientemente, Mendoza y Becerra (2010) han mostrado que en colecciones documentales de escala mediana el uso de categorizadores basados en regresión logística permite obtener mejores resultados que *Naive Bayes*. Una completa revisión sobre las más conocidas técnicas de categorización de documentos fue presentada por Sebastiani (2002).

2. Categorización de texto *Naive Bayes*

Categorizar documentos corresponde a asignar un valor Booleano a cada par $\langle d,c \rangle \in D \times C$, donde D representa a la colección documental, C representa un grupo de clases, d es un documento que pertenece a D y c es una clase que pertenece a C. El valor Booleano asignado al par $\langle d,c \rangle$ es 1 si se ha tomado la decisión de categorizar a d en c, 0 en caso contrario. A este tipo de categorización se le denomina categorización dura. Existen también categorizadores blandos que asignan un puntaje a cada par $\langle d,c \rangle$, permitiendo que eventualmente un documento pueda ser clasificado en más de una clase. Siguiendo la notación de Sebastiani (2002), la tarea de categorizar texto corresponde a la aproximación de una función objetivo desconocida $\Phi = DxC \rightarrow \{0.1\}$ en el caso de categorización dura y $\Phi : DxC \rightarrow [0,5]$ en el caso de categorización *soft*.

Los clasificadores de texto *Bayesianos* usan un modelo generativo de documentos basado en mezcla paramétrica, donde cada clase corresponde a una de las componentes a mezclar. El modelo tiene la siguiente forma:

$$p(d) = \sum_{j=1}^{|c|} p(cj)\, p(d|cj) \qquad [1]$$

donde $d \in D$ y $cj \in C$. Usando la regla de Bayes podemos obtener la probabilidad de que d genere a c:

$$p(c|d) = \frac{p(c)p(d|c)}{p(d)} \qquad [2]$$

Para clasificar un documento, se selecciona a la clase con máxima probabilidad *a posteriori*, escenario en el cual $p(d)$ se comporta como una constante por lo que el discriminante se reduce al producto $p(c)p(d|c)$. Las probabilidades $p(c)$ corresponden a probabilidades *a priori*, las cuales pueden ser estimadas contando el número de documentos de ejemplo que pertenecen a c.

Las probabilidades $p(d|c)$ pueden ser estimadas usando los términos que componen el contenido de d. Existen varios modelos *Bayesianos* que hacen diferentes supuestos acerca de cómo los documentos son formados a partir de los términos que los componen. Sin embargo, la mayoría de ellos asumen independencia estadística en la co-ocurrencia de los términos en un mismo documento por lo que se denominan modelos *Naive Bayes* (ingenuos).

Los modelos *Naive Bayes* multinomiales asumen que la distribución de términos en un documento es multinomial. Esto es, un documento corresponde a una secuencia de términos y se asume que la posición de cada término es generado independientemente de los otros términos. Dado un vocabulario $V = \{t_1, \ldots, t_n\}$ que permite representar el contenido de la colección D, cada una de las clases $c \in C$ tiene asociado un conjunto de parámetros $\vec{\theta}C = \{\theta_{c,1}, \ldots, \theta_{c,n}\}$, cada uno de los cuales a su vez corresponde a la probabilidad $p(t_i|c)$, que representan cuan probable es observar el término t_i en c. Se cumple que:

$$\sum_{i=1}^{n} p(ti|c) = \forall c \in C$$

Luego, la verosimilitud de un documento d con respecto a una clase c queda dada por:

$$p(c|d) = \frac{Ld!}{\prod_{i}^{n}=1\ Tf_{i,d}!} \prod_{t=1}^{n} p(t_i|c) \qquad [3]$$

donde $Tf_{i,d}$ representa el número de veces que t_i ocurre en d y Ld representa el largo de d, es decir,

$$\sum_{i=1}^{n} Tf_{i,d}!$$

El coeficiente

$$\frac{Ld!}{\prod_{i}^{n}=1\ Tf_{i,d}!}$$

es independiente de c por lo que para un mismo documento se cumple que:

$$p(d|c) \prod_{i=1}^{n} p(ti|c)^{tf_{i,d}} \qquad [4]$$

Finalmente, el discriminante del clasificador *Naive Bayes* multinomial puede expresarse de la siguiente forma:

$$p(c|d) \approx p(c) \prod_{i=1}^{n} p(t_i|c)^{tf_{i,d}} \qquad [5]$$

Dado que muchas de las probabilidades podrían ser muy cercanas a cero (Bennett, 2000), las implementaciones de *Naive Bayes* usan la versión logarítmica suavizada de la probabilidad a *posteriori,* dada por la siguiente expresión:

$$log(p(c|d)) \approx logp(c) + \sum_{i=1}^{n} Tfi,d \; log\,(p(ti\,|c)) \qquad [6]$$

Las probabilidades $p(t_i|c)$ pueden ser estimadas a partir del estimador de máxima verosimilitud:

$$\hat{p}(ti|c) = \frac{Tf_{i,c}}{Tf_c} \qquad [7]$$

donde $Tf_{i,c}$ representa el número de ocurrencias de T_i en C y Tf_c representa el número total de ocurrencias de términos de V en C. Varios trabajos (Rennie et al., 2003; Schneider, 2005; Qiang, 2010) usan una versión suavizada del estimador de máxima verosimilitud, dada por la siguiente expresión:

$$\hat{p}(ti|c) = \frac{1 + Tf_{i,c}}{n + Tf_c} \qquad [8]$$

donde n representa el tamaño del vocabulario.

3. Modificaciones de *Naive Bayes*

Rennie et al. (2003) propusieron varias modificaciones al modelo *Naive Bayes* multinomial de categorización de texto. Lo primero que observaron fue que las distribuciones de frecuencias de términos producidas por el modelo multinomial diferían de las distribuciones empíricas, las cuales seguían leyes de potencia (*power laws*). Debido a esto, los términos que presentaban muchas ocurrencias en el texto atenuaban la importancia de aquellos términos que, aun sin ser muy frecuentes, tenían una alta capacidad discriminativa. Este fenómeno fue denominado *burstiness*. Para atenuar este efecto, propusieron las siguientes transformaciones al factor $Tf_{i,d}$ de la ecuación [6]:

$$Tf'_{i,d} = log(1 + Tf_{i,d}) \qquad [9]$$

$$Tf'_{i,d} = Tf_{i,d}\, log\left(\frac{N}{nt}\right) \qquad [10]$$

$$Tf'_{i,d} = \frac{Tf_{i,d}}{\sqrt{\sum_{t}^{n=1} Tf^2_{i,d}}} \qquad [11]$$

La ecuación [9] representa una versión suavizada del factor $Tf_{i,d}$. La ecuación [10] corresponde al esquema de pesos Tf-Idf usado en recuperación de información por Salton y Buckley (1988), N donde representa el número de documentos de la colección y n_i el número de documentos donde el término ocurre. Finalmente, la ecuación [11] corresponde a la versión normalizada del factor $Tf_{i,d}$ usando la norma L_2. Otra modificación fue estudiada por Schneider (2005), quien propuso eliminar el efecto de los términos más frecuentes, conocido en inglés como *burstiness*, a través de la siguiente modificación:

$$Tf'_{i,d} = Min\{log\,(1 + Tf_{i,d}), 1\} \qquad [12]$$

Una modificación sobre la propuesta de Rennie et al. (2003) fue estudiada por Kolcz y Yih (2007), quienes evaluaron la transformación del factor $Tf_{i,d}$ normalizando con la norma L_1, es decir:

$$Tf'_{i,d} = \frac{Tf_{i,d}}{\sum_{t}^{n=1} Tf_{i,d}} \qquad [13]$$

Finalmente, una última modificación fue estudiada por Qiang (2010), quien propuso la transformación siguiente:

$$Tf'_{i,d} = 1 + log\, Tf_{i,d} \qquad [14]$$

Las transformaciones anteriores provienen del área de recuperación de información y corresponden principalmente a variaciones sobre el conocido modelo vectorial de Salton y Buckley (1988). Sin embargo, estudios posteriores han realizado importantes variaciones sobre este modelo obteniendo mejoras sobre colecciones documentales de evaluación como las producidas por la *Text Retrieval Information Conference* (TREC).

En esta línea, Robertson, Walker, Hancock, Gull y Lau (1992) introdujeron el modelo BM25, el cual en la actualidad es considerado estado del arte para tareas de recuperación de información basado en texto. El modelo BM25 introduce modificaciones tanto sobre el factor Idf como el factor Tf. En este trabajo introduciremos algunas modificaciones al método *Naive Bayes* multinomial consistentes con el modelo BM25.

Una primera modificación introducida en el modelo BM25 fue realizada sobre el factor Idf, el cual es calculado de la siguiente forma:

$$Idf(ti) = log\left(\frac{N - n_i}{n_i}\right) \quad [15]$$

Dado que $nt \in \{1, N\}$, $Idf(ti)$ se indefine cuando $n_i = N$. Para evitar esta situación una versión modificada de Idf propuesta en el modelo BM25 consiste en calcularlo de la siguiente forma:

$$Idf(t_i) = log\left(\frac{N - n_i + 0,5}{n_i + 0,5}\right) \quad [16]$$

Con esta modificación, $Idf(t_i)$ toma valores en

$$\left[log\left(\frac{N - n_i + 0,5}{n_i + 0,5}\right) log\ \left(\frac{N - 0,5}{1.5}\right)\right]$$

Podemos observar que, dependiendo del valor de N, un tramo del recorrido de $Idf(t_i)$ puede tomar valores negativos. Para evitar que esto ocurra, proponemos una versión modificada de $Idf(t_i)$ dada por la siguiente expresión:

$$Idf(ti) = log\left(\frac{2N - n_i + 0,5}{n_i + 0,5}\right) \quad [17]$$

Con esta modificación $Idf(t_i)$, toma valores positivos en

$$\left[log\left(\frac{N+1}{N}\right)\ log\ (2N)\right]$$

El factor Tf del modelo BM25 queda expresado de la siguiente forma:

$$Tf(ti) = \frac{Tf_{i,d}\ (1 + k)}{Tf_{i,d} + k\ ((1-b) + b\ \frac{L_d}{\bar{L}}\ ))} \quad [18]$$

Donde $\bar{L}$ representa el largo promedio de los documentos de la colección. Los valores para los parámetros k y b fueron ajustados empíricamente a la colección de evaluación de TREC, ajuste que corresponde a $k = 2$ y $b = 0,75$.

Haremos dos modificaciones al factor $Tf(t_i)$ que favorezcan su uso en categorización de documentos. Primero, prescindiremos de los parámetros k y b ya que estos no tienen una clara justificación dentro del modelo *Naive Bayes*. La segunda modificación consiste en prescindir del uso de $\bar{L}$ debido a que requiere ser recalculado cada vez que un nuevo documento de entrenamiento es agregado al modelo. Al considerar estas dos simplificaciones, el factor $Tf(ti)$ puede ser expresado de la siguiente forma:

$$Tf(ti) \approx \frac{Tf_{i,d}}{Tf_{i,d} + L_d} \quad [19]$$

Si consideramos que en general se cumple que $L_d \gg Tf_{i,d}$, entonces podemos expresar el factor $Tf(t_i)$de la siguiente forma:

$$Tf(ti) \approx \frac{Tf_{i,d}}{L_d} \quad [20]$$

lo cual corresponde a la versión normalizada de $Tf(t_i)$ según la norma L_1, ya presentada en la ecuación [13]. Considerando el fenómeno de *burstiness*, proponemos una versión modificada de $Tf(t_i)$, introduciendo una función de suavizado, lo cual nos permite obtener la siguiente expresión:

$$Tf(ti) \approx \frac{log(Tf_{i,d})}{L_d} \quad [21]$$

Para evitar que $Tf(t_i)$ se indefina para $Tf_{i,d}=0$, realizamos la siguiente modificación:

$$Tf(t_i) \approx \frac{log(1+Tf_{i,d})}{L_d} \quad [22]$$

lo cual permite que $Tf(t_i)$ solo tome valores positivos.

Usando las versiones derivadas del modelo BM25 para categorización de textos expresadas en las ecuaciones [17] y [22], proponemos la siguiente variación para $Tf'_{i,d}$, dado por el producto $Tf\text{-}Idf$ que queda expresado de la siguiente forma:

$$Tf(t_i) \approx \frac{log(1+Tf_{i,d})}{L_d} log\left(\frac{2N - n_i + 1}{n_i}\right) \quad [23]$$

Finalmente proponemos estudiar otra modificación al discriminante del categorizador *Naive Bayes* multinomial. La estimación de la probabilidad $\hat{p}(t_i|c)$ ya sea por la ecuación [7] o [8] introduce el efecto *burstiness* debido a que usa el coeficiente $Tf_{i,c}$. Para evitar este efecto, proponemos estimar $log\hat{p}(t_i|c)$

usando el coeficiente $Idf(t_i)$ calculado sobre cada clase de C. Sea cf_i la cantidad de clases de C en las cuales t_i ocurre. A partir de nuestra versión modificada de $Idf(t_i)$ expresada en la ecuación [17], proponemos la siguiente estimación para $\log \hat{p}(t_i|c)$:

$$\log \hat{p}(t_i|c) \approx \left(\frac{2|C| - Cf_i + 1}{cf_i}\right) \qquad [24]$$

donde $|C|$ representa la cantidad de clases de C. Dado que Cf_i toma valores en $\{1|C|\}$, el factor $Idf(t_i)$ toma valores en:

$$\left[\log\left(\frac{|C|+1}{|C|}\right) \ \log(2|C|)\right]$$

Normalizando por $\log(2|C|)$, $\hat{p}(t_i|c)$ toma valores en $[0,1]$.

Incorporando las modificaciones expresadas en las ecuaciones [12] y [13] al discriminante del categorizador *Naive Bayes* multinomial de la ecuación [1], ver [25].

Schneider (2005) mostró que otro fenómeno que afectaba el desempeño de los categorizadores *Naive Bayes* multinomiales era el sesgo introducido por las probabilidades *a priori* de las clases que disponían de más ejemplos de entrenamiento. Para ilustrar este fenómeno, comparó el desempeño obtenido por categorizadores *Naive Bayes* multinomiales eliminando las probabilidades *a priori* de las clases, mostrando que permitía obtener mejoras más significativas en la medida que el tamaño de la colección aumentaba. Nosotros también estudiaremos este fenómeno en la sección de resultados experimentales, evaluando el impacto en el desempeño de los categorizadores de texto al eliminar las probabilidades *a priori* de las clases del discriminante definido por la ecuación [25].

4. Evaluación

4.1. Colección documental de análisis

Contamos con un conjunto de documentos consistente en 567 ítems correspondientes a descripciones de sitios, páginas *Web* y otros documentos electrónicos facilitados por "IALE Tecnología"[5] a través de su plataforma de vigilancia Vigiale®[6]. Estos documentos contienen 4175 términos distintos obtenidos luego de un proceso de normalización del texto en el cual hemos extraído los artículos y conjunciones, conocido en inglés como lista de *stopwords*, y símbolos especiales. Sobre este conjunto hemos construido una base documental que considera un índice invertido con 12866 entradas.

La colección documental corresponde a recursos de áreas productivas correspondientes a patentes, informes técnicos, artículos de tecnología y sitios *Web*. Los recursos corresponden a 4 sectores productivos: Agroalimentación, Energía, Materiales y TICs (Tecnologías de la Información y las Comunicaciones). El sector Agroalimentación contiene 128 documentos, Energía contiene 162, Materiales 22 y TICs 248. La categorización considerada es dura, esto es, un documento puede pertenecer solo a una clase y esta ha sido realizada por un grupo de expertos de IALE tecnologías.

Se consideran 21 ejes temáticos, los cuales son subcategorías de los 4 sectores principales. La relación entre ejes temáticos y sectores corresponde a una relación de especialización. Cada eje temático está asociado solo a un sector. La colección documental se distribuye sobre los ejes temáticos siguiendo una categorización dura. En el apartado anexo mostramos más detalles acerca de esta colección.

Para ilustrar las relaciones de dependencia entre los factores usados para el cálculo del discriminante del categorizador propuesto en la ecuación [25], mostramos los gráficos de dispersión para los pares formados por las variables $Tf_{i,d}$, $n1$ y cf_i sobre la colección de IALE. Para cada uno de los términos que conforman el vocabulario de la colección hemos calculado los valores que obtienen las variables anteriores. En el caso de cf_i, el cálculo se ha realizado sobre ejes temáticos.

El Gráfico 1(a) muestra que en el par $\langle n_i, Tf_{i,d} \rangle$ existe una clara relación lineal. El coeficiente de correlación de Pearson entre ambas variables corresponde a 0,9736. El Gráfico 1(b) muestra que en el par

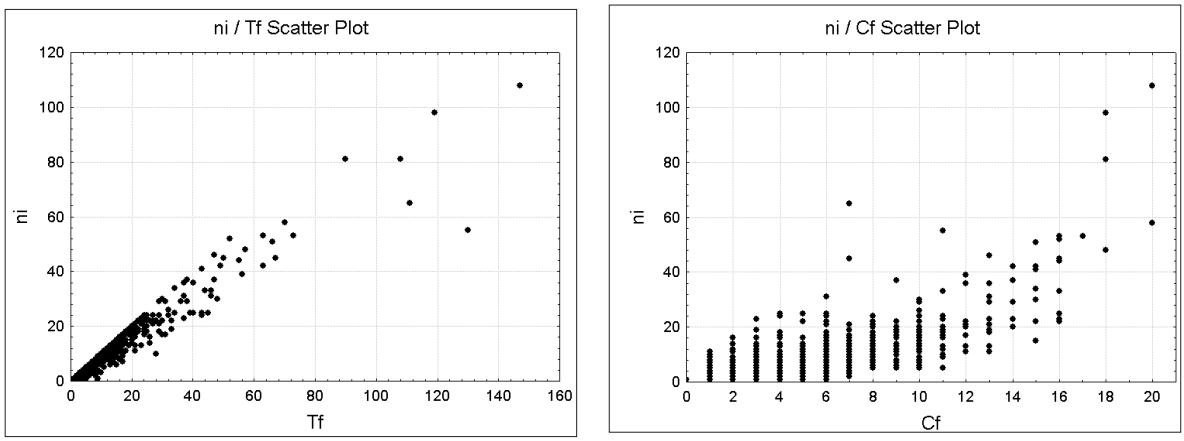
$$\log(p(c|d)) \approx \log p(c) + \sum_{i=1}^{n} \frac{\log(1+Tf_{i,d})}{\log(2|C|)} \ \log\left(\frac{2N - n_i + 1}{n_i}\right) \log\left(\frac{2|C| - Cf_i + 1}{cf_i}\right) \qquad [25]$$

$\langle n_i, Cf_{i,d} \rangle$ la relación lineal es significativa pero menos fuerte que para el par anterior, obteniendo un coeficiente de correlación igual a 0,8015.

El Gráfico 2 muestra el gráfico de dispersión para el par $\langle Tf_{i,d}, Cf_i \rangle$ es el que presenta la correlación más baja de los 3 pares comparados, obteniendo un coeficiente equivalente a 0,7418. La capacidad discriminativa del clasificador dependerá de aquellos términos que aun cuando presenten altos valores de frecuencia local $Tf_{i,d}$ muestren a la vez una alta capacidad discriminativa, es decir, tengan valores bajos asociados a las variables cf_i y n_i.

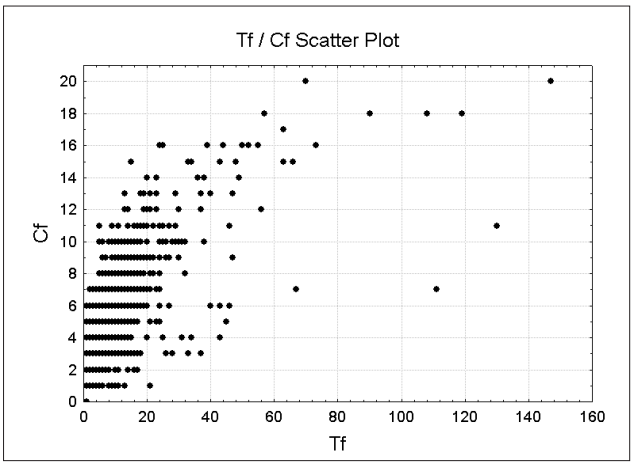
4.2. Aspectos de implementación de los categorizadores Naive Bayes

En esta sección abordaremos aspectos de la implementación de los categorizadores *Naive Bayes*. Para la fase de entrenamiento implementamos un modulo que permite pre-procesar el texto de los documentos a indexar. El módulo permite leer un índice documental y pre-procesar el texto considerando etapas de *tokenización* y procesamiento de caracteres y símbolos especiales. Luego de extraer un conjunto de *stopwords* almacenadas en una lista estática, se dispone de un conjunto



(a) (b)

Gráfico **1**. Gráficos de dispersión para los pares (a)$\langle n_i, Tf_{i,d} \rangle$ y (b) $\langle n_i, Cf_{i,d} \rangle$.



El Gráfico **2.** Gráfico de dispersión para el par $\langle n_i, Cf_{i,d} \rangle$.

de términos para cada documento, los cuales se ingresan a una base de datos relacional. La base de datos relacional registra además la información mutua de cada término con una categoría dada, la cual puede ser usada como criterio de selección de características. Un módulo permite poblar la base de datos automáticamente y mantenerla actualizada cuando nuevos documentos son incorporados a la plataforma. El modelo de la base de datos usada se muestra en la Figura 1.

Como muestra la Figura 1, el modelo permite almacenar documentos categorizados en Sectores y Ejes temáticos, categorías que corresponden a la colección de documentos de vigilancia tecnológica con la cual se realizó el análisis preliminar de la Sección 4.1. Los valores de información mutua son almacenados en tablas que relacionan al vocabulario y las categorías. Las etiquetas de categorización para entrenamiento son almacenadas en las tablas que relacionan a los documentos con las categorías. La tabla índice permite almacenar la frecuencia de cada término del vocabulario en cada documento. Los datos necesarios para calcular las frecuencias inversas de los términos en la colección son almacenados en el vocabulario. La cota superior en espacio en memoria secundaria está determinada por la tabla Índice, la cual relaciona la colección con el vocabulario, por lo que su tamaño corresponde a $\Theta(|V||C|)$, donde V representa al vocabulario de la colección y C representa el conjunto de categorías del problema.

Los categorizadores basados en aprendizaje de modelos incorporan un requerimiento en espacio menor, ya que les basta con almacenar $|C|$ vectores de tamaño $|V|$. Sin embargo, el tiempo de entrenamiento de un categorizador *Naive Bayes* solo considera el llenado de la base de datos. No existen procesos de optimización para sintonización o validación interna lo cual lo compara favorablemente con cualquier método de máquinas de aprendizaje basado en optimización, como es el caso de regresión logística o máquinas de soporte vectorial. En rigor, el tiempo de entrenamiento de un categorizador *Naive Bayes* extendido corresponde a la suma de los tiempos de pre-procesamiento del texto y a la inserción de los documentos en la base de datos. La tabla que tarda más en llenar corresponde a la tabla Índice, la cual requiere $\Theta(|V||C|)$ operaciones de escritura.

El tiempo de procesamiento de un nuevo documento a categorizar considera accesos a las tablas de información mutua y a las tablas de vocabulario y documentos. Cada una de ellas tiene índices,

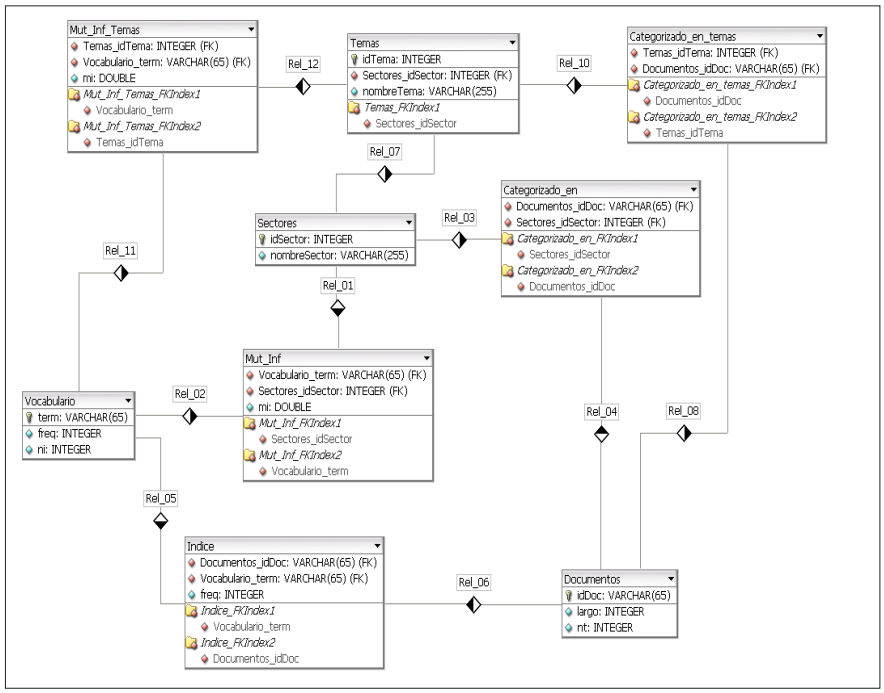


Figura 1. Modelo de la base de datos que permite implementar los categorizadores *Naive Bayes*. La notación usada corresponde al estándar REIN85.

lo cual permite que las búsquedas se reduzcan logarítmicamente sobre el largo del documento. El proceso de validación sobre la colección IALE ha tomado 12 segundos. Cualquier método basado en aprendizaje requerirá de tiempos considerables debido a que agregan etapas de sintonización de parámetros y de optimización de función de evaluación. Lo anterior permite afirmar que los métodos basados en *Naive Bayes* son más livianos computacionalmente en etapa de entrenamiento manteniendo costos similares para tiempos de procesamiento y requiriendo un poco más de espacio para almacenamiento, siendo considerado este último factor como el menos critico de los mencionados.

4.3. Metodología de evaluación

Para evaluar el desempeño de los clasificadores, definiremos en una primera aproximación la metodología de evaluación considerando solo dos categorías. A partir de un conjunto de documentos etiquetados considerados para evaluación, obtenemos una clase nominal y una clase predicha (la determinada por el categorizador). A partir de los cuatro casos posibles que se presentan al comparar la clase nominal con la clase predicha, esto es, verdaderos positivos *(vp)*, falsos positivos *(fp),* falsos negativos *(fn)* y verdaderos negativos *(vn)*, se consideran las siguientes medidas de desempeño:

$$exactitud = \frac{vp + vn}{vp + vn + fp + fn}\ ,\ precision = \frac{vp + vn}{vp + fp}$$

$$tasa\ fp = \frac{fp}{fb + vn}\ ,\ cobertura = \frac{vp}{vp + fn}$$

Además consideraremos la medida F, la que corresponde a la media armónica entre la precisión y la cobertura. Consideramos en todas las evaluaciones como método base (*baseline*) al método *Naive Bayes* multinomial. Las evaluaciones serán realizadas considerando problemas multi-clase, esto es, más de dos clases objetivo en las cuales categorizar. Para poder usar las medidas anteriores, se evalúa el desempeño de cada método considerando una clase objetivo contra el resto, procedimiento que se repite para cada clase considerada. Luego, las medidas obtenidas se promedian sobre todas las clases, obteniendo una medida promedio. El estado del arte muestra que este tipo de evaluación es el más indicado para el caso multi-clase (Sebastiani, 2002).

4.4. Resultados preliminares

En esta sección ilustraremos el desempeño del categorizador propuesto en colecciones documentales de pequeña escala. Como primer experimento, hemos construido un categorizador *Naive Bayes* multinomial (abreviado como NBM) y la versión extendida propuesta (abreviado como NBME) sobre la colección IALE usando dos tercios de la colección para entrenamiento, reservando el tercio restante para evaluación. Para evaluar el impacto de las probabilidades a priori, también hemos considerado versiones de los categorizadores Bayesianos asumiendo uniformidad en la distribución de probabilidades a priori de cada clase. Luego, hemos calculado las medidas de desempeño discutidas en la Sección 4.3. Los promedios ponderados por el número de ejemplos de cada sector (promedio de los promedios por sector, o micro promedio) son mostrados en la Tabla 1.

Como muestra la Tabla 1, las versiones que han supuesto distribuciones uniformes para las probabilidades a priori superan en desempeño por cerca de 4 a 5 puntos a las versiones no uniformes. Este resultado coincide con los resultados obtenidos por Schneider (2005) sobre una colección de escala media conocida como *Linguistics Links Database*. Podemos observar también que el categorizador propuesto supera a *Naive Bayes* multinomial por varios puntos porcentuales. Observemos también que la tasa de falsos positivos de NBME-U está cerca de 7 puntos por debajo de NBM-U, lo cual ilustra las mejoras obtenidas en capacidad discriminativa.

La Tabla 2 muestra los resultados obtenidos por las versiones uniformizadas de los categorizadores *Bayesianos* estudiados, desagregados por sector. Como muestra la Tabla 2, NBME-U obtiene mejores resultados que NBM-U en todos los sectores analizados. Es destacable el desempeño obtenido en el sector 'Materiales', donde NBME-U supera por cerca de 25 puntos porcentuales en medida F a NBM-U, siendo esta precisamente la clase con menos ejemplos.

Posteriormente, evaluamos el desempeño de los categorizadores en ejes temáticos. Para cada uno de los ejes temáticos disponemos de menos ejemplos que en el caso de sectores. Además, la distribución de ejemplos de entrenamiento por eje temático es muy dispar. Los promedios de los resultados obtenidos por ejes temáticos se muestran en la Tabla 3, donde se observa el promedio de los promedios

Tabla 1. Desempeño obtenido por los clasificadores *Bayesianos* usando la colección IALE. El literal U indica uniformidad *a priori.*

	Exactitud	Precisión	Tasa-fp	Cobertura	Medida F
NBM-U	0,8928	0,7660	0,1056	0,9297	0,8397
NBM	0,8534	0,7286	0,1643	0,8854	0,7994
NBME-U	0,9601	0,8850	0,0361	0,9529	0,9167

Tabla 2. Desempeño obtenido por los clasificadores *Bayesianos* usando la colección IALE a nivel de sectores.

NBM-U	**Exactitud**	**Precisión**	**Tasa-fp**	**Cobertura**	**Medida F**
Agroalimentación	0,912	0,782	0,093	0,984	0,871
Energía	0,879	0,764	0,132	0,922	0,836
Materiales	0,803	0,698	0,187	0,894	0,784
TICs	0,901	0,766	0,088	0,911	0,832
NBME-U	**Exactitud**	**Precisión**	**Tasa-fp**	**Cobertura**	**Medida F**
Agroalimentación	0,955	0,811	0,056	1,000	0,896
Energía	0,961	0,887	0,038	0,959	0,922
Materiales	0,974	0,896	0,023	0,963	0,928
TICs	0,962	0,922	0,026	0,925	0,924

Tabla 3. Desempeño obtenido por los clasificadores *Bayesianos* a nivel de temas.

	Exactitud	Precisión	Tasa-fp	Cobertura	Medida F
NBM-U (micro)	0,9201	0,5341	0,0498	0,4858	0,4982
NBM-U (macro)	0,9148	0,5371	0,0481	0,4766	0,4937
NBME-U (micro)	0,9309	0,5419	0,0373	0,4854	0,5007
NBME-U (macro)	0,9346	0,5563	0,0341	0,4831	0,5061

por eje temático (micro promedio) y el promedio global (macro promedio) para cada categorizador.

La Tabla 3 muestra que ambos categorizadores disminuyen su desempeño, lo cual se hace más notorio al evaluar la medida precisión y cobertura. Puede observarse que el categorizador NBME-U logra una menor tasa de falsos positivos y una mejor precisión que el categorizador *baseline*, lo cual indica que presenta mejores desempeños en su capacidad discriminativa. Los detalles del desempeño por eje temático pueden ser observados en el Anexo.

4.5. Validación

En esta sección evaluamos el desempeño del categorizador propuesto usando otras colecciones documentales. Esto nos permite estudiar la validez de los resultados observados en la Sección 4.4. y comparar el desempeño del categorizador propuesto con otros métodos del estado del arte. En un primer experimento comparamos los métodos *Bayesianos* con métodos que consideran fases de entrenamiento más complejas, como es el caso de regresión logística y máquinas de soporte vectorial (que abreviaremos usando la sigla SVM).
En esta primera evaluación, hemos considerado una colección de documentos recolectados desde la plataforma MERLOT[7] (*Multimedia Educational Resource for Learning and Online Teaching*). MERLOT es una plataforma que permite recuperar objetos de aprendizaje, es decir, recursos que prestan utilidad a la confección de material didáctico. Cada uno de los objetos de aprendizaje dispone de un resumen que describe el contenido que el objeto tiene. Para los propósitos de este trabajo, hemos descargado los resúmenes y las palabras clave de 300 objetos de aprendizaje en idénticos números de ejemplos sobre las tres categorías principales del tema Ciencia y Tecnología de la plataforma: 100 objetos

de Computación, 100 objetos de Química, y 100 de Biología. Esta colección tiene un largo de 11191 términos sobre un vocabulario de 4484 términos distintos.

Cada una de las categorías principales de la colección se descompone en varias sub categorías. En particular, la colección de Biología presenta tres sub categorías con igual número de ejemplos. Las otras dos categorías presentan desbalance de ejemplos. La Tabla 4 muestra cómo se distribuyen los ejemplos en cada una de las sub categorías. La evaluación se realiza usando validación cruzada con tres repositorios de igual tamaño (*3-fold cross validation*), reservando dos para entrenamiento y uno para evaluación. Este procedimiento se realiza sobre las tres permutaciones posibles de los repositorios. El uso de validación cruzada nos permite obtener resultados más robustos y generalizables.

La Tabla 5 muestra los resultados obtenidos para exactitud, tasa de falsos positivos y medida F. Los valores obtenidos corresponden a los promedios de las evaluaciones realizadas en cada repositorio *(fold)*.

La Tabla 5 nos muestra que el categorizador *Bayesiano* propuesto obtiene resultados que se comparan favorablemente a los obtenidos por regresión logística y SVM. Al observar los resultados en promedio, NBME-U obtiene los mejores valores en exactitud y medida F siendo solo superado por SVM en tasa de falsos positivos. Los resultados obtenidos por los tres métodos son comparables, estableciéndose una clara diferencia sobre NBM-U de entre 8 y hasta 10 puntos porcentuales en exactitud y medida F, y estando entre 5 a 7 puntos por debajo en tasa de falsos positivos.

La Tabla 6 nos muestra los resultados obtenidos en medida F en cada sub categoría.

Al igual que en la evaluación anterior, hemos usado validación cruzada con tres repositorios. Como podemos observar en la Tabla 6, NBME-U obtiene un desempeño comparable al obtenido por SVM, superando levemente a regresión logística y logrando una diferencia significativa sobre NBM-U por cerca de 4 a 6 puntos en micro y macro promedio, respectivamente. Podemos observar también que la ventaja más significativa de NBME-U sobre regresión logística es obtenida precisamente en aquellas sub categorías donde se dispone de menos documentos. Es el caso, por ejemplo, de "Ingeniería de software" y "Bases de datos", que corresponden a las sub categorías con menos ejemplos de la categoría "Computación" y en donde es esperable que el desempeño de los categorizadores sea menor al de las sub categorías con más ejemplos. Esto ocurre

Tabla 4. Distribución de ejemplos sobre la colección MERLOT.

Sub categoría	Documentos	Categoría
Inteligencia artificial	8	Computación
Simulación por computador	24	Computación
Ingeniería de *software*	12	Computación
Bases de datos	11	Computación
Lenguajes de programación	33	Computación
Multimedia	12	Computación
Química analítica	30	Química
Bioquímica	26	Química
Química inorgánica	20	Química
Química orgánica	24	Química
Citología	33	Biología
Genética	33	Biología
Anatomía	34	Biología

Tabla 5. Evaluación de desempeño sobre la colección MERLOT usando categorías principales.

Exactitud	SVM	Regresión	NBM-U	NBME-U
Computación	0,897	0,877	0,835	0,904
Química	0,894	0,889	0,812	0,911
Biología	0,893	0,888	0,796	0,921
PROMEDIO	0,894	0,884	0,814	**0,912**
Tasa-fp	SVM	Regresión	NBM-U	NBME-U
Computación	0,184	0,198	0,253	0,186
Química	0,102	0,114	0,181	0,116
Biología	0,096	0,084	0,144	0,098
PROMEDIO	**0,1273**	0,132	0,1926	0,1333
Medida F	SVM	Regresión	NBM-U	NBME-U
Computación	0,901	0,878	0,819	0,911
Química	0,852	0,854	0,808	0,846
Biología	0,844	0,832	0,778	0,847
PROMEDIO	0,8656	0,8546	0,8016	**0,868**

Tabla 6. Resultados de medida F sobre trece sub-categorías de la colección MERLOT.

	SVM	Regresión	NBM-U	NBME-U
Inteligencia artificial	0,603	0,567	0,512	0,589
Simulación por computador	0,447	0,418	0,430	0,430
Ingeniería de *software*	0,516	0,449	0,428	0,512
Bases de datos	0,516	0,500	0,307	0,528
Lenguajes de programación	0,548	0,527	0,513	0,534
Multimedia	0,558	0,496	0,413	0,542
Química analítica	0,501	0,491	0,436	0,498
Bioquímica	0,506	0,488	0,470	0,506
Química inorgánica	0,538	0,498	0,474	0,520
Química orgánica	0,512	0,508	0,482	0,482
Citología	0,538	0,529	0,526	0,540
Genética	0,556	0,552	0,532	0,548
Anatomía	0,570	0,568	0,538	0,590
Promedio (macro)	0,531	0,507	0,466	0,524
Promedio (micro)	0,529	0,511	0,481	0,523

con regresión logística y NBM-U, pero no ocurre con NBME-U ni con SVM. Las tres sub categorías pertenecientes a "Biología", que están balanceadas en cuanto a número de ejemplos, muestran buenos resultados para los tres mejores métodos, siendo los mejores resultados obtenidos por NBME-U en dos de las tres comparaciones.

Para ilustrar el desempeño de los categorizadores según la cantidad de ejemplos de entrenamiento de que disponen, mediremos la diferencia en medida F entre cada método evaluado y el *baseline*. Las categorías han sido mostradas en orden decreciente según la cantidad de ejemplos de entrenamiento de que disponen. Los resultados de este experimento son mostrados en el Gráfico 3.

El Gráfico 3 muestra con color gris claro la diferencia entre NBME-U y NBM-U, con color gris oscuro la diferencia entre regresión logística y NBM-U y con

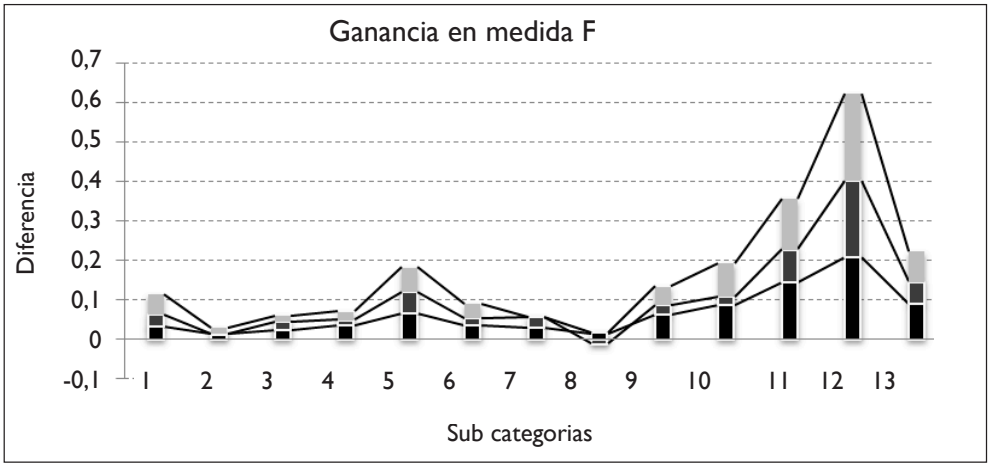
negro la diferencia entre SVM y NBM-U. Como podemos observar, en la medida que el número de ejemplos disminuye, la diferencia a favor de los métodos NBME-U y SVM se hace más significativa. Esto nos indica que mientras NBM-U ve afectado su desempeño cuando dispone de menos ejemplos por categoría, el desempeño de SVM y NBME-U se mantiene, generando una diferencia significativa a favor.

Posteriormente, hemos evaluado el desempeño de NBME-U con otras variaciones de NBM propuestas en el estado del arte. Para ello, hemos considerado dos colecciones documentales adicionales: 20 *Newsgroups*[8] y Web-KB[9], las cuales han sido usadas previamente en los experimentos de Rennie et al. (2003) y Schneider (2005). 20 *Newsgroups* es una colección de aproximadamente 20.000 noticias particionadas sobre 20 temas distintos. Web-KB es una colección de 8282 páginas universitarias clasificadas en 7 categorías. En el caso de 20 *Newsgroups*, solo hemos usado el contenido de las páginas indexadas en la colección, descartando los encabezados, siguiendo el mismo procedimiento usado por Schneider (2005). Lo mismo ha sido realizado sobre la colección Web-KB, usando solo los contenidos y descartando *tags* HTML.

Además de realizar la evaluación del desempeño de NBM-U y NBME-U sobre las colecciones anteriores, hemos evaluado cada una de las extensiones de NBM-U discutidas en la Sección 3. Para ello hemos seguido una metodología de validación cruzada usando 3 repositorios con muestreo estratificado. También hemos evaluado sobre estas colecciones los métodos SVM y regresión logística, previamente estudiados sobre la colección MERLOT. Todos los métodos evaluados consideran uniformidad en distribución de probabilidades *a priori.* La Tabla 7 muestra los promedios sobre los 3 repositorios para las medidas exactitud, tasa de falsos positivos y medida F. Cada una de las extensiones al *baseline* indica la ecuación asociada de acuerdo a las extensiones discutidas en la Sección 3.

La Tabla 7 muestra que el método propuesto se compara favorablemente con los otros métodos discutidos en el estado del arte, superando en todas las mediciones a las variaciones de NBM-U. Entre estas modificaciones, se puede observar que aquellas que consideran Tf sin incorporar el factor Idf obtienen desempeños más bajos. Este fenómeno se explica a partir de las mejoras que muestran en capacidad discriminativa los categorizadores que usan el factor Idf. No se observan diferencias significativas entre las extensiones que solo usan el factor Tf, siendo entre estas ligeramente mejor el desempeño obtenido por Min Tf, la cual elimina el efecto *burstiness*. NBME-U supera a regresión logística (RL) en todas las mediciones y obtiene resultados comparables a los obtenidos por SVM.

Finalmente, mostraremos el efecto del tamaño del vocabulario sobre el desempeño de los categorizadores anteriores. Para ello, usaremos el coeficiente de información mutua como función de criterio para selección de características. El coeficiente de información mutua fue estudiado en



Gáfico 3. Ganancia en medida F para las trece sub-categorías de la colección MERLOT.

Tabla 7. Evaluación de desempeño sobre las colecciones 20 News y Web-KB.

	20 News			Web-KB		
	Exactitud	Tasa-FP	Medida F	Exactitud	Tasa-FP	Medida F
NBM-U	0,812	0,175	0,672	0,833	0,152	0,712
NBME-U	0,868	0,095	0,728	0,894	0,124	0,786
RL	0,858	0,112	0,710	0,882	0,128	0,768
SVM	0,874	0,098	0,722	0,902	0,120	0,794
Tf-soft 1 [4]	0,823	0,164	0,680	0,840	0,136	0,722
Tf-Idf [5]	0,840	0,135	0,694	0,852	0,130	0,742
Tf-L2 [6]	0,820	0,152	0,686	0,836	0,138	0,732
Min Tf [7]	0,838	0,174	0,692	0,846	0,135	0,743
Tf-L1 [8]	0,824	0,137	0,689	0,842	0,140	0,736
Tf-soft 2 [9]	0,820	0,134	0,692	0,846	0,138	0,726

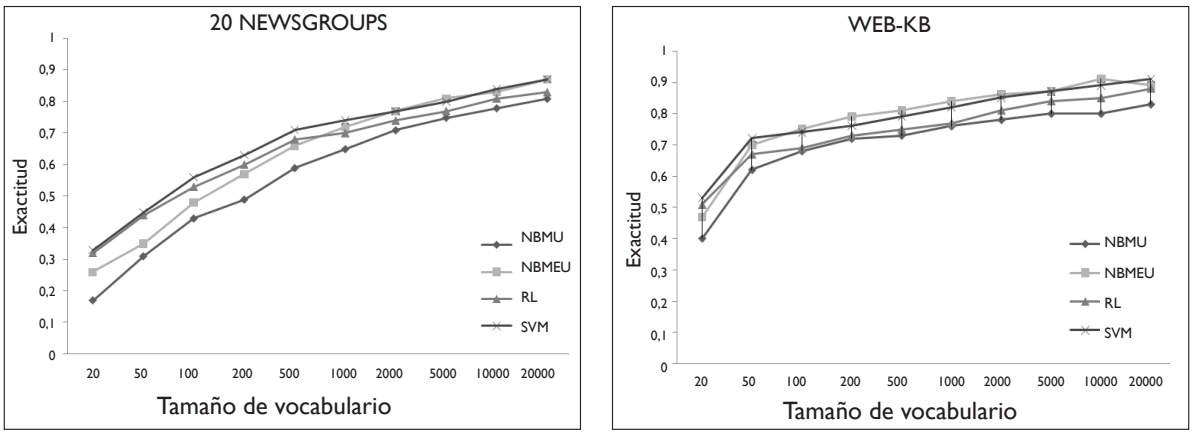
el problema de categorización de texto por Lewis y Ringuette (1994) y permite cuantificar la relevancia de un término para describir una categoría dada. Usando selección de características, hemos evaluado el desempeño de los categorizadores sobre los mejores *n* términos (los *n* mejores términos descriptivos según el coeficiente de información mutua). Se han considerado valores para $n=\{20,50,100,200,500,1000,2000,5000,10000,20000\}$

Los valores obtenidos en exactitud son mostrados en el Gráfico 4.

Como muestra el Gráfico 4, en ambas colecciones de evaluación el desempeño en exactitud mejora en la medida que el tamaño del vocabulario aumenta, resultado que coincide con los encontrados por Schneider (2005) y Kolcz y Yih (2007). En 20 *Newsgroups,* SVM supera levemente en vocabularios pequeños a NBME-U (por 2 a 3 puntos porcentuales), sin embargo, logran resultados similares a partir de un vocabulario de tamaño 1000. En WEB-KB, NBME-U supera levemente a SVM, sin embargo, se equiparan para un vocabulario de tamaño 20.000.

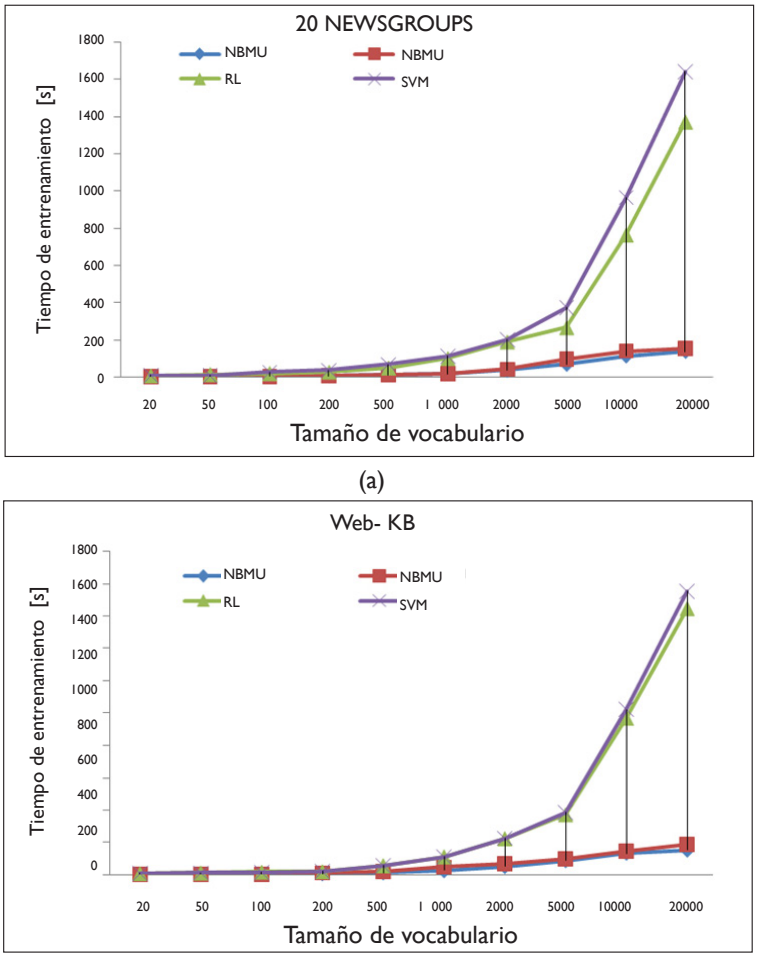
Finalmente, en el Gráfico 3 comparamos los tiempos de entrenamiento requeridos por los categorizadores evaluados en función del tamaño del vocabulario. Para los métodos basados en aprendizaje no hemos considerado la fase de sintonización de parámetros, la cual consiste de un proceso de búsqueda exhaustiva. Cada categorizador basado en aprendizaje ha sido entrenado considerando los valores para sus parámetros que obtienen el mejor desempeño. Los tiempos mostrados en el Gráfico 3 se indican en segundos, medidos sobre un procesador Pentium® M 730 (1.6 GHz, 2 MB L2 cache, 533 MHz FSB).

El Gráfico 3 permite observar que los tiempos de entrenamiento de los categorizadores basados en aprendizaje son superiores a los registrados para categorizadores *Bayesianos*. Esta diferencia se hace más significativa en la medida que el tamaño de la colección de entrenamiento aumenta, siendo para la colección completa la diferencia de un orden de magnitud, en segundos. Sin embargo, la figura 2 nos muestra que el desempeño obtenido por NBME-U es muy similar al obtenido por SVM y RL independientemente del tamaño del vocabulario. Esto nos indica que en colecciones de gran escala el ahorro en costo computacional involucrado en la etapa de entrenamiento de los métodos *Bayesianos* no incide significativamente en el desempeño alcanzado, comparándose favorablemente con los categorizadores basados en aprendizaje.



(a) (b)

Gráfico 4. Exactitud según tamaño del vocabulario sobre (a) 20 *Newsgroups*, (b) WEB-KB.



(a)

(b)

Gráfico 5. Tiempos de entrenamiento en función del tamaño de la colección.

CONCLUSIONES

En este artículo hemos presentado un nuevo categorizador de documentos basado en una extensión al modelo *Naive Bayes*. El categorizador incorpora nuevos factores que permiten modelar adecuadamente la capacidad descriptiva y discriminativa de cada uno de los términos que conforman el vocabulario. Entre estos factores destaca la frecuencia inversa en documentos, usado en su versión original para cuantificar la capacidad discriminativa de cada término con respecto al vocabulario, y usado también en una variación que permite cuantificar la capacidad discriminativa de cada término con respecto al conjunto de categorías a representar. Para evitar el sesgo inherente a conjuntos de datos de entrenamiento con desbalance en el número de ejemplos por categoría, hemos asumido uniformidad en probabilidades a priori.

Los factores incorporados al categorizador *Naive Bayes* extendido no representan nuevos costos computacionales significativos. De hecho, hemos mostrado que en este sentido el categorizador propuesto mantiene las mismas propiedades que el categorizador original. El categorizador *Naive Bayes* extendido es simple y la mayoría de sus factores pueden ser calculados fuera de línea.

Los experimentos han mostrado que el categorizador propuesto obtiene consistentemente mejores resultados que el categorizador *Naive Bayes*. Una de las propiedades observables en los experimentos consiste en que la mejora en rendimiento se produce precisamente en aquellas categorías en las cuales se dispone de menos ejemplos de entrenamiento. También es observable su mejor capacidad discriminativa, aun en categorías con un número significativamente bajo de ejemplos, como ha sido el caso de la evaluación por ejes temáticos.

Hemos mostrado también que el categorizador propuesto se compara favorablemente con categorizadores basados en aprendizaje como regresión logística y máquinas de soporte vectorial. Evaluando el desempeño en colecciones documentales de benchmark como *20 Newsgroups* y WEB-KB hemos mostrado que obtiene resultados similares en tiempos de entrenamiento significativamente menores.

A modo de reflexión de esta investigación podemos señalar que una de sus principales contribuciones consiste en la eventual aplicación de este nuevo categorizador en el ámbito de la lingüística aplicada. Como hemos mostrado en el estado del arte, la administración de grandes volúmenes de texto en bases de datos documentales requiere de la aplicación de técnicas provenientes del área del procesamiento del lenguaje natural, facilitando con esto la identificación automática de conceptos que describan de mejor forma los contenidos de nuevos documentos. El enfoque de esta investigación muestra que esto es posible realizarlo aplicando una aproximación de aprendizaje supervisado, esto es, disponiendo de ejemplos que nos indiquen previamente cuales podrían ser los eventuales contenidos que podríamos observar en cada una de los temas. El categorizador propuesto permite detectar con alta precisión los temas que son tratados en nuevos documentos, facilitando con esto su catalogación.

Los resultados de esta investigación ilustran que la interacción entre la lingüística aplicada y el procesamiento del lenguaje natural produce resultados potentes tanto a nivel de aplicaciones como de avances teóricos. La necesidad de establecer puentes de colaboración permanentes entre ambas comunidades aparece como algo prioritario en el desarrollo de ambas disciplinas, pudiendo enriquecerse la comunidad de lenguaje natural con los avances que la comunidad de lingüística propone acerca de mejores caracterizaciones de los discursos. En el otro sentido, la comunidad de lingüística aplicada puede aplicar las técnicas de procesamiento de lenguaje natural en su ámbito de acción, favoreciendo aspectos como automatización de métodos de caracterización y catalogación, y manejo de grandes volúmenes de información.

Como trabajo futuro evaluaremos el desempeño del categorizador propuesto en colecciones de referencia de gran escala, como es el caso de RCV1 o TREC. En esta dirección, la aplicación de estrategias que permitan aprender la función discriminante usando, por ejemplo, algoritmos genéticos se visualiza como una de las alternativas más atractivas.

REFERENCIAS BIBLIOGRÁFICAS

Ault, T. & Yang, Y. (2002). Information filtering in TREC-9 and TDT-3: A comparative analysis. *Journal of Information Retrieval, 5*(2-3), 159-187.

Bennett, P. (2000). *Assessing the calibration of naive Bayes posterior estimates. Technical Report CMU-CS-00-155*. School of Computer Science: Carnegie-Mellon University.

Datar, M. & Indyk, P. (2004). Locality-sensitive hashing scheme basedon p-stable distributions. En *Annual symposium on computational geometry*. Brooklyn, New York, USA.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer-Verlag.

Indyk, P. (2004). Nearest neighbors in high-dimensional spaces. En J. Goodman & J. O'Rourke (Eds.), *Handbook of discrete and computational geometry* (pp. 877-892). New York: Chapman and Hall/CRC Press.

Joachims, T. (2006). Training linear SVMs in linear time. En *ACM SIGKDD International conference on knowledge discovery and data mining*. Philadelphia, PA, USA.

Kolcz, A. & Yih, W. (2007). Raising the baseline for high-precision text classifiers. En *ACM SIGKDD International conference on knowledge discovery and data mining*. San José, California, USA.

Lewis, D. & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. En *Annual symposium on document analysis and information retrieval*. Las Vegas, NV, USA.

Lewis, D., Yang, Y., Rose, T. & Li. F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research,* 5, 361-397.

Maron, M. & Kuhns, J. (1960). On relevance, probabilistic indexing, and information retrieval. *Journal of the Association for Computing Machinery, 7*(3), 216-244.

McCallum, A. & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. En *International conference on machine learning, Workshop on learning for text categorization*. Madison, Wisconsin, USA.

Mendoza, M. & Becerra, C. (2010). On the design of learning objects classifiers. En *IEEE International conference on advanced learning technologies*. Sousse, Túnez.

Perkins, S., Lacker, K. & Theiler, J. (2003). Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research,* 3, 1333-1356.

Qiang, G. (2010). An effective algorithm for improving the performance on naive Bayes for text classification. En *International conference on computer research and development*. Kuala Lumpur, Malasia.

Rennie, J., Shih, L., Teevan, J. & Karger, D. (2003). Tackling the poor assumptions of naive Bayes text classifiers. En *International conference on machine learning*. Washington, DC, USA.

Robertson, S., Walker, S., Hancock, M., Gull, A. & Lau, M. (1992). Okapi at TREC. En *Text retrieval conference*. Gaithersburg, Maryland, USA.

Rocchio, J. (1971). Relevance feedback in information retrieval. En G. Salton (Ed.), *The SMART Retrieval System–Experiments in automatic document processing* (pp. 313-323). New Jersey: Prentice-Hall.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic retrieval. *Information Processing y Management, 24*(5), 513-523.

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

Schneider, K. (2005) Techniques for improving the performance of naive Bayes for text classification. En *International conference on computational linguistics and intelligent text processing.* Ciudad de México, México.

Vapnik, V. (1998). *Statistical learning theory.* New York: Wiley-Interscience.

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

Voorhees, E. & Harman, D. (2005). *TREC: Experiments and evaluation in information retrieval.* New York: MIT Press.

Wilbur, W. & Kim, W. (2009). The ineffectiveness of within-document term frequency in text classification. *Information Retrieval, 12*(5), 509-525.

Zhang, T. & Oles, F. (2001). Text categorization based on regularized linear classification methods. *Journal of Information Retrieval, 4*(1), 5-31.

NOTAS

1 RCV1 significa *Reuters Corpus Volume 1*. Corresponde a una colección de evaluación de algoritmos para categorización de textos conformada por más de 800.000 resúmenes de noticias de la agencia Reuters, liberados con fines de investigación.

2 TodoCL es un motor de búsqueda vertical chileno restringido a las páginas web y sitios del dominio chileno (.CL). Disponible en: http://www.todocl.cl

3 Inquiro es un motor de búsqueda vertical chileno restringido a las páginas web y sitios de varios dominios de Sudamérica, entre ellos la web mexicana (.MX), argentina (.AR) y chilena (.CL). Disponible en: http://www.inquiro.cl

4 TDT significa Topic Detection and Tracking.

5 IALE Tecnología es una empresa española de desarrollo de tecnologías de la información para sectores productivos. Dispone de una división en Chile. Ver sitio web: http://www.iale.es

6 Vigiale® es una plataforma de vigilancia tecnológica desarrollado por IALE tecnologías.

7 MERLOT es un repositorio de objetos de aprendizaje. Disponible en: http://www.merlot.org

8 *20 Newsgroups* es una colección documental de noticias que permite la evaluación de categorización de texto. Disponible en: http://people.csail.mit.edu/people/jrennie/*20 Newsgroups*/

9 Web-KB es una colección documental de páginas universitarias que permite la evaluación de categorizadores de texto. Disponible en: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

ANEXO

Distribución de la colección documental IALE por temas

Eje temático	Documentos	Sector
FLOSS	29	TICs
Web 2.0	17	TICs
TIC audiovisual	32	TICs
Tecnología RFID	29	TICs
Tecnologías de seguridad	82	TICs
TIC aplicada a la empresa	86	TICs
Agricultura sostenible	86	Agroalimentación
Tecnologías de conservación	7	Agroalimentación
Herramientas TICs de trazabilidad	8	Agroalimentación
Productos funcionales	14	Agroalimentación
Nuevos envases	17	Agroalimentación
Mejora genética de especies cultivadas	19	Agroalimentación
Biotecnología aplicada a la sanidad animal y vegetal	45	Agroalimentación
Cultivos Agro energéticos	3	Energía
Biomasa y Biocombustibles	45	Energía
Arquitectura Bio climatic	52	Energía
Sistemas solares en edificación	71	Energía
Materiales de construcción	15	Materiales
Diseño nuevos materiales	5	Materiales
Bio nanotecnología	9	Materiales
Bio tecnología ambiental	30	Energía

Desempeño del categorizador *Naive Bayes* **multinomial desagregado por eje temático**

Eje temático	Exactitud	Precisión	Tasa-fp	Cobertura	Medida F
FLOSS	0,947	0,716	0,012	0,412	0,523
Web 2.0	0,907	0,501	0,040	0,344	0,408
TIC Audiovisual	0,902	0,513	0,034	0,372	0,431
Tecnología RFID	0,913	0,636	0,028	0,414	0,502
Tecnologías de seguridad	0,914	0,468	0,076	0,640	0,541
TIC aplicada a la empresa	0,935	0,528	0,069	0,558	0,543
Agricultura sostenible	0,943	0,486	0,062	0,534	0,509
Tecnologías de conservación	0,901	0,422	0,068	0,498	0,457
Herramientas TICs de trazabilidad	0,878	0,480	0,089	0,507	0,493
Productos funcionales	0,913	0,509	0,052	0,500	0,504
Nuevos envases	0,918	0,520	0,067	0,532	0,526
Mejora genética de especies cultivadas	0,914	0,533	0,053	0,538	0,535
Biotecnología aplicada a la sanidad	0,920	0,505	0,052	0,560	0,531
Cultivos Agro energéticos	0,977	0,766	0,020	0,422	0,544
Biomasa y Bio combustibles	0,915	0,713	0,028	0,409	0,520
Arquitectura Bio climática	0,894	0,501	0,036	0,348	0,411
Sistemas solares en edificación	0,932	0,564	0,026	0,374	0,450
Materiales de construcción	0,926	0,610	0,022	0,408	0,489
Diseño nuevos materiales	0,901	0,430	0,062	0,608	0,504
Bio nanotecnología	0,869	0,427	0,044	0,520	0,469
Bio tecnología ambiental	0,893	0,449	0,068	0,511	0,478

Desempeño del categorizador Naive Bayes multinomial extendido desagregado por eje temático

Eje temático	Exactitud	Precisión	Tasa-fp	Cobertura	Medida F
FLOSS	0,980	0,764	0,000	0,426	0,547
Web 2.0	0,929	0,528	0,034	0,366	0,432
TIC Audiovisual	0,942	0,573	0,028	0,387	0,462
Tecnología RFID	0,948	0,640	0,020	0,425	0,511
Tecnologías de seguridad	0,914	0,462	0,069	0,622	0,530
TIC aplicada a la empresa	0,919	0,465	0,057	0,542	0,501
Agricultura sostenible	0,922	0,475	0,053	0,520	0,496
Tecnologías de conservación	0,927	0,480	0,054	0,506	0,493
Herramientas TICs de trazabilidad	0,932	0,508	0,044	0,512	0,510
Productos funcionales	0,937	0,516	0,039	0,512	0,514
Nuevos envases	0,935	0,528	0,045	0,541	0,535
Mejora genética de especies cultivadas	0,942	0,561	0,040	0,553	0,557
Biotecnología aplicada a la sanidad	0,934	0,578	0,039	0,571	0,574
Cultivos Agro energéticos	0,986	0,773	0,006	0,433	0,555
Biomasa y Bio combustibles	0,967	0,749	0,008	0,413	0,533
Arquitectura Bio climática	0,910	0,524	0,018	0,360	0,427
Sistemas solares en edificación	0,939	0,559	0,014	0,368	0,444
Materiales de construcción	0,941	0,623	0,015	0,423	0,504
Diseño nuevos materiales	0,901	0,444	0,053	0,612	0,515
Bio nano tecnología	0,913	0,461	0,037	0,534	0,495
Bio tecnología ambiental	0,910	0,473	0,044	0,517	0,494

* Proyecto CORFO N°08IEI-7488 "Implementación de tecnologías de recuperación de información en entornos colaborativos para la generación de conocimiento estratégico".
Proyecto UTFSM-DGIP N°24.11.19 "Propiedades de la información en redes sociales".