Clasificación de textos utilizando información inherente al conjunto a clasificar

por

Adriana Gabriela Ramírez de la Rosa

Tesis sometida como requisito parcial para obtener el grado de

Maestro en Ciencias en el Área de Ciencias Computacionales en el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Supervisada por:

Dr. Manuel Montes y Gómez, INAOE

Clasificación de textos utilizando información inherente al conjunto a clasificar

Tesis de Maestría

Por:

Adriana Gabriela Ramírez de la Rosa

Asesor:

Dr. Manuel Montes y Gómez

Instituto Nacional de Astrofísica Óptica y Electrónica
Coordinación de Ciencias Computacionales

Tonantzintla, Puebla. Noviembre 2010

Para los que me acompañaron: motivándome, enseñándome, divirtiéndome, abrazándome...
para los que se van conmigo...

y para la luna... donde siempre anduve.

Resumen

El constante crecimiento de la cantidad de documentos digitales disponibles en la Web han motivado el desarrollo de mecanismos automáticos que faciliten su acceso, organización y análisis. En un esfuerzo por organizar esa información de tal forma que se pueda acceder eficientemente a ella, se han desarrollado los métodos de clasificación de textos. La clasificación de textos es una tarea que consiste en la asignación de documentos dentro de un conjunto de categorías o clases predefinidas. A lo largo de los años se han propuesto diferentes algoritmos y métodos para clasificar textos; particularmente enfoques de aprendizaje automático. Dentro de este enfoque, para llevar a cabo la construcción de clasificadores es necesario contar con un número de documentos de ejemplo. Desafortunadamente, en muchos escenarios de clasificación los documentos de ejemplos son escasos o peor aún, no existen; además, generarlos es una tarea demasiado costosa. Con el fin de atacar el problema de insuficiencia e inexistencia de documentos etiquetados para entrenar clasificadores con buen desempeño, en esta tesis se propone un método alternativo para clasificar textos que se basa en un enfoque de clasificación consensuada, esto es, clasificar un documento considerando información tanto de él mismo como información presente en el conjunto de documentos a clasificar (no etiquetado). En particular, se considera la clasificación de los documentos más similares al documento a clasificar con el objetivo de dar soporte al proceso de asignación de clase. El método fue evaluado en tres escenarios de clasificación con características particulares: i) cuando existen pocos documentos etiquetados, ii) en un enfoque de clasificación multi-lenguaje, utilizando documentos etiquetados de un idioma distinto a los documentos que se desean clasificar, y iii) en un enfoque de clasificación multi-dominio, en el cual se utilizan documentos etiquetados de un dominio similar al conjunto de documentos a clasificar. Los resultados

experimentales demostraron que el método propuesto en esta tesis es una alternativa de clasificación de documento que además, es flexible a diferentes escenarios de clasificación; en especial, cuando el conjunto de documentos a clasificar es pequeño o cuando se hace uso de documentos etiquetados en otros lenguajes.

Abstract

The continued growth in the number of digital documents available on the Web has motivated the development of automatic mechanisms that facilitate access, organization and analysis. In an effort to organize that information so that you can access it efficiently, have developed methods for text classification. Text classification task is the assignment of documents within a set of predefined categories or classes. Over the years have proposed different algorithms and methods for classifying texts, particularly machine learning approaches. Within this approach, to carry out the construction of classifiers is necessary to have a set of sample documents. Unfortunately, in many classification scenarios, the sample documents are very few or worse, do not exist; in addition, to generate them is a very expensive task. In order to tackle the problem of insufficient and lack of labeled documents, to train classifiers with good performance, in this thesis is proposed an alternative text classification method based on a consensus classification approach, that means, classifying a document using information himself as both information of set of documents to classify (unlabeled documents), in particular, is considered the classification of closer documents to it, the goal is give support to the assignment class process. The proposed method was evaluated on three classification scenarios with particular characteristics: i) when there are few labeled documents, ii) in a multi-lingual classification approach, using labeled documents in a different language of documents to classify, and iii) in a multi-domain classification approach, where labeled documents of similar domain to the set of documents to classify are used. The experimental results showed that the method proposed in this thesis is an alternative of text classification that also is flexible to scenarios of classification with insufficiency and lack of labeled documents.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

El crecimiento de la Web en los últimos años se ha dado de forma exponencial, la cantidad de información que contiene, así como el número de usuarios han aumentado drásticamente en la última década, alrededor del 400 % (Research (2010)); con esta explosión de información y usuarios surge la necesidad de encontrar lo que se busca de una manera eficiente. En otras palabras, con el rápido crecimiento de este importante recurso global, la organización, recuperación y análisis de la información juega un papel muy importante para encontrar efectivamente lo que se desea.

Los esfuerzos para dar solución a esta necesidad se han encaminado en diversos sentidos. En particular, se han desarrollado métodos enfocados en la *recuperación de la información*, tarea que involucra la representación, almacenamiento, organización y acceso a la información de tal forma que al representarla y organizarla se proporcione a los usuarios un acceso fácil a la información específica en la que están interesados (Baeza-Yates y Ribeiro-Neto (1999)). Debido a lo anterior, la organización juega un papel fundamental para solucionar el problema planteado.

Para organizar la información se han desarrollado métodos, por un lado, centrados en el agrupamiento de documentos; esto es, organizar documentos dentro de grupos donde sus miembros compartan cierta similitud. Por otro lado, métodos que se basan en la clasificación de documentos, dado un conjunto de documentos y un conjunto de clases, decidir para cada documento la clase de éste tomando en cuenta su contenido (Sebastiani (2002)). Esta tesis se centra en la clasificación de documentos.

1.1. Motivación

La organización de la información en la Web requiere de métodos que sean capaces de clasificar grandes cantidades de documentos, por ello se han desarrollado métodos automáticos de clasificación; para lo cual es indispensable contar con documentos *ejemplo* para cada una de las clases o categorías en las que se desean organizar los nuevos.

El desarrollo de estos métodos de clasificación ha evolucionado. Por ejemplo, al inicio su construcción era 100 % manual; consistía en que personas especializadas en un dominio particular se concentraran en leer cada *documento ejemplo* de cada categoría para generar reglas de clasificación del tipo 'Si en el documento d aparece el término t_A y el término t_B y el término t_C, ..., entonces el documento pertenece a la clase X'. Con estas reglas se buscaba determinar la *categoría* de cada nuevo documento. Existen muchas desventajas de este procedimiento, de las más importantes, el alto costo en tiempo y esfuerzo, y la necesidad de contar con expertos en diversos dominios.

A causa de las desventajas anteriores, se empezaron a desarrollar métodos de clasificación automáticos; esto es, que cada *documento ejemplo* de cada clase es representado de tal forma que un programa de computadora pueda obtener una especie de 'reglas de clasificación' (antes generadas de forma manual) a través de algún método de aprendizaje. Este procedimiento es conocido como clasificación supervisada y es uno de los enfoques más utilizados hasta el momento.

De lo anterior se intuye que un componente importante en la construcción de clasificadores dentro de un enfoque de clasificación supervisada es el conjunto de *documentos ejemplo*. Un documento ejemplo es aquél del cual se conoce la clase a la que pertenece en un dominio específico. Es decir, si el problema a resolver fuera la clasificación de documentos de tres clases de noticias, por ejemplo `política`, `deportes` y `cultura`, es indispensable (en el enfoque de clasificación supervisado) tener un conjunto de documentos que *a priori* se sabe son de `política` o de `deportes` o de `cultura`. A este tipo de documentos ejemplo se les conoce como *documentos etiquetados* y al conjunto de todos estos documentos etiquetados se conoce como *conjunto de entrenamiento*, es este conjunto el que se utiliza para entrenar un modelo de clasificación.

Como es de esperarse, en general, entre mayor es el número de ejemplos (conjunto de entrenamiento) mejor será el aprendizaje del modelo de clasificación (Wong y Fu

(2000), Ling et al. (2008)). Sin embargo, no siempre es posible conseguir documentos etiquetados de todos los posibles temas que existen en la Web, por lo que se vislumbra un problema de insuficiencia e inexistencia de documentos etiquetados.

1.2. Planteamiento del problema

Lo anterior ha motivado el desarrollo de métodos que por un lado hacen frente al problema de la insuficiencia de documentos etiquetados; mientras es complicado y costoso conseguir estos documentos, es relativamente sencillo obtener documentos sin etiquetar; en este sentido se han generado métodos dentro de enfoques semi-supervisados (que usan una gran cantidad de documentos no etiquetados junto con documentos etiquetados) o métodos que utilizan técnicas de agrupamiento o *clustering*. Por otro lado, se han desarrollado métodos que hacen frente al problema de inexistencia de documentos etiquetados; a través de la construcción de clasificadores a partir de conjuntos de entrenamiento compuestos por: a) ejemplos etiquetados de dominios similares y b) ejemplos etiquetados del mismo dominio pero en otros idiomas.

Para resumir los enfoques de solución mencionados anteriormente es posible decir que en todos los casos se hace uso de recursos existentes. El primer enfoque de solución consiste en utilizar los insuficientes documentos etiquetados del dominio particular de clasificación y documentos no etiquetados (que son fáciles de conseguir). A los métodos que caen en este enfoque se denominarán 'métodos en condiciones de insuficiencia de documentos etiquetados'. En particular, los métodos semi-supervisados y los métodos que utilizan algoritmos de *clustering* caen dentro de este enfoque. Dentro de los primeros existen métodos que utilizan diferentes técnicas, por ejemplo, *self-trainig*, *co-training*, basados en grafos y un enfoque denominado clasificación colectiva; todos estos métodos hacen uso de documentos etiquetados junto con documentos sin etiquetar. Los segundos, utilizan algoritmos de *clustering* para obtener cierta información del conjunto de documentos no etiquetados, información que puede ser desde incorporar nuevas características a la representación de cada documento, hasta hacer una pre-clasificación considerando los *clusters* formados por el algoritmo.

El segundo enfoque de solución consiste en utilizar documentos etiquetados de un dominio similar al dominio de clasificación. A estos métodos se les denominarán 'métodos de clasificación multi-dominios'. Los métodos dentro de este enfoque tratan

de adaptar un clasificador para clasificar documentos en un dominio similar. Este problema también es conocido como *transfer learning*.

El tercer y último enfoque de solución consiste en utilizar documentos etiquetados del mismo dominio de clasificación pero en otros idiomas, a estos métodos se les denominarán 'métodos de clasificación multi-lenguajes'. El objetivo es generar clasificadores utilizando conjuntos de entrenamiento 'suficientes' del mismo dominio de clasificación pero en otros lenguajes. Esto es muy útil para realizar clasificación en lenguajes poco comunes donde no existen conjuntos de documentos etiquetados, y donde se aprovechan los conjuntos abundantes de otros lenguajes.

Evidentemente, al utilizar un conjunto de entrenamiento que no corresponde al dominio particular de clasificación, el desempeño del clasificador resultante será bajo, por lo que los métodos actuales mencionados anteriormente se centran en mejorar el clasificador resultante para que de esta forma se mejore la clasificación. En esta tesis se propone una solución alternativa, la cual buscará mejorar la clasificación. Los objetivos generales y específicos de este trabajo de investigación se presentan a continuación.

1.3. Objetivos

1.3.1. Objetivo general

Proponer un esquema de clasificación automática de textos que considere información extraída del conjunto de entrenamiento y también considere información contenida en el conjunto de documentos a clasificar (*conjunto objetivo*) para mejorar la clasificación dentro de un enfoque supervisado en situaciones de insuficiencia o inexistencia de conjuntos de documentos etiquetados.

1.3.2. Objetivos específicos

- Utilizar un modelo de clasificación consensuada para incorporar información tanto del conjunto de entrenamiento como del conjunto objetivo haciendo uso de similitudes de documentos.

- Establecer una función que maneje la influencia de la información obtenida del conjunto a clasificar, que favorezca la información de documentos similares al

documento a clasificar.

- Evaluar el esquema propuesto en tres escenario de clasificación: insuficiencia de documentos etiquetados, en un escenario multi-lenguaje y en un escenario multi-dominio.

1.4. Solución propuesta

La solución propuesta está inspirada primero, en que es posible obtener información de un conjunto de documentos no etiquetados, como lo muestran los métodos semi-supervisados (Nigam et al. (1999)), y también en la 'suposición de cluster' (a veces llamada suposición de consistencia): es probable que puntos (en este caso documentos) cercanos y puntos dentro de la misma estructura (en este caso, pequeños grupos de documentos similares) pertenezcan a la misma clase (Driessens et al. (2006), Zhou et al. (2004), Ning y Karypis (2009), Jensen, Neville, y Gallagher (2004)).

Por lo tanto la solución propuesta usa información de similitudes de documentos en el conjunto a clasificar para mejorar la clasificación de instancias individuales utilizando un clasificador aprendido con los datos etiquetados disponibles. En particular, el método propuesto consiste de dos etapas, en la primera se construye un clasificador utilizando el conjunto de documentos etiquetados disponible. En la segunda etapa se agrega información de la clasificación de los documentos más parecidos al que se desea clasificar, utilizando la información generada en la primera etapa. Se busca que este conjunto de documentos más parecidos aporte evidencia para clasificar el documento particular en la clase correcta. Al final, el proceso de asignación de clase considera tanto la clasificación del documento a clasificar como la información de un conjunto de documentos similares a él.

1.5. Estructura de la tesis

En el Capítulo 2 se exponen conceptos y definiciones necesarios para la fácil comprensión de la investigación realizada en esta tesis. Qué es clasificación de textos, qué representación de los documentos se usa, cómo se construye un clasificador y, cómo se evalúa su desempeño una vez construido, son ejemplos de los conceptos y definiciones a encontrar en este capítulo.

Una revisión detallada de los trabajos relacionados con la investigación presentada en esta tesis se muestra en el Capítulo 3, entre ellos los trabajos más relevantes en los siguientes tres escenarios: insuficiencia de documentos etiquetados, clasificación multi-lenguaje y clasificación multi-dominio.

En el Capítulo 4 se presenta el esquema de clasificación propuesto. Para demostrar experimentalmente la utilidad del método propuesto, en el Capítulo 5 se presenta la evaluación hecha en tres distintos escenearios. Finalmente, en el Capítulo 6 se resume el trabajo realizado, se mencionan las conclusiones obtenidas y se plantea el trabajo futuro.

Capítulo 2

Clasificación de textos

En este capítulo se introducen los conceptos y definiciones básicas útiles para la fácil comprensión del resto de esta tesis. El capítulo se centra en la Clasificación de Textos. En primer lugar se describe brevemente el aprendizaje automático, después se describen los procesos involucrados en la clasificación de textos, desde la extracción de características hasta los métodos de aprendizaje utilizados y las medidas de evaluación usadas en esta tarea.

2.1. Aprendizaje automático

Se aprende automáticamente a realizar una tarea T, si después de obtener una experiencia E y una medida de desempeño P, el desempeño en la tarea T medida por P, mejora con la experiencia E (Mitchell (1997)). En general, es necesario indicar la clase de tarea a realizar, el objetivo a cumplir, y de dónde debe obtener la experiencia.

Los algoritmos de aprendizaje automático más utilizados son los algoritmos de clasificación supervisados, éstos consisten en asignar a un objeto (persona, documento, fenómeno físico, etc) diversas categorías o clases previamente especificadas. La clasificación supervisada parte de un conjunto de N casos del problema descritos por un vector de características (se darán más detalles de la construcción de tal vector en las secciones siguientes), de los cuales se conoce el valor c_i de la clase, a esta base de datos se conoce como *conjunto de entrenamiento*. En la Tabla 2.1 se muestra la forma general del conjunto de entrenamiento con N casos, los valores asociados (t_{ij}) a cada uno de los s términos (T) y la clase a la que pertenecen (C). Utilizando el conjunto de entrenamiento, los algoritmos de aprendizaje construyen un modelo o regla general que se utilizará para clasificar objetos nuevos (de los cuales no se conoce la clase)

(Araujo (2006); Debole y Sebastiani (2003)).

Número de caso	T_1	T_2	...	T_s	C
1	t_{11}	t_{21}	...	t_{s1}	c_{11}
2	t_{12}	t_{22}	...	t_{s2}	c_{22}
...	...	...	...	...	...
N	t_{1n}	t_{2n}	...	t_{sn}	c_{Mn}

Tabla 2.1: Forma general del conjunto de entrenamiento y el conjunto de evaluación. Cada caso es representado utilizando un conjunto de s términos y el valor de la clase a la que pertenecen.

La clasificación es uno de los principales dominios en donde se aplica aprendizaje automático. La tarea consiste en organizar un conjunto de objetos en un conjunto específico de categorías o clases. Específicamente, la clasificación de textos es la actividad de construir clasificadores de textos de forma automática, es decir, programas capaces de determinar la etiqueta de clase de un documento específico utilizando un conjunto de etiquetas de clases predefinidas $C = \{c_1, ..., c_{|C|}\}$.

Para construir un clasificador automático de textos es necesaria la existencia de un *corpus inicial* $D = \{d_1, ..., d_{|D|}\}$ de documentos etiquetados con C. Mediante un proceso inductivo se construye un clasificador que aprende las características de cada clase del *conjunto de entrenamiento* $D_t = \{d_1, ..., d_{|D_t|}\}$, es decir, se *aprende* una función $F : D_t \rightarrow C$. Una vez construido el clasificador, su desempeño es probado utilizando un *conjunto de evaluación* $D_e = D - D_t$ y la función F aprendida.

La Figura 2.1 muestra un esquema del proceso de clasificación de textos. La parte central del proceso es la construcción del clasificador que a su vez consta de dos partes: Extracción de características y el algoritmo de aprendizaje (Liu et al. (2007); Aas y Eikvil (1999)). A continuación se presentan a detalle en qué consisten cada una de estas partes. En la última sección se describe cómo se realiza la evaluación de los clasificadores construidos.

2.2. Extracción de características

El primer paso en la clasificación de textos es representar cada documento tal que, pueda ser utilizado por un algoritmo de aprendizaje y el método de clasificación (Sebastiani (2005)). Al transformar un documento a la representación deseada se realizan todas o algunas de las siguientes acciones (Aas y Eikvil (1999)):

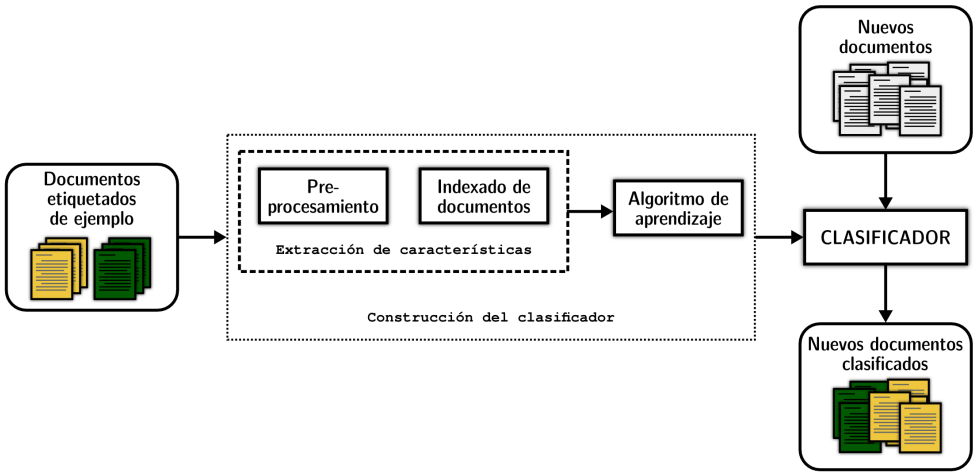


Figura 2.1: Esquema del proceso de clasificación de textos. Dado un conjunto de documentos se extrae un conjunto de características o atributos con los cuales se representará cada documento. Después a través de algún algoritmo de aprendizaje se generará un clasificador el cual determinará la clase de un nuevo documento de acuerdo a lo aprendido

- Eliminar etiquetas HTML u otras,
- Eliminar *palabras vacías*, que son palabras que aparecen frecuentemente en el documento pero que no aportan información del mismo, por ejemplo: pronombres, preposiciones, conjunciones, etc.
- Lematizar las palabras; es decir, eliminar sufijos de las palabras para conservar sólo la palabra raíz.

El proceso anterior es referido como **pre-procesamiento**, en esta tesis sólo se hace uso de la eliminación de palabras vacías. La siguiente parte del proceso corresponde al indexado de documentos y reducción de la dimensionalidad.

2.2.1. Indexado de documentos

En clasificación de textos, el método más utilizado para representar cada documento es a través de un vector de términos con pesos. Comúnmente se tiene una colección de documentos representada por una matriz $\mathbf{A}$, donde cada renglón corresponde a la representación vectorial de un documento $d_j = \{w_{1j}, ..., w_{|\tau|j}\}$, donde w_{ij} es el peso de la palabra (o término) i en el documento j. Dado que una palabra no aparece en todos los documentos de la colección, la matriz $\mathbf{A}$ es dispersa. El número

de columnas en la matriz $|\tau|$ corresponde al número de términos diferentes t_i en la colección, a este conjunto de términos se le conoce como *diccionario* y frecuentemente el número de términos en el diccionario suele ser de cientos o miles.

Dependiendo de la naturaleza de la colección a clasificar, w_{ij} puede calcularse utilizando diferentes esquemas, todos ellos se basan en las siguientes dos observaciones:

- entre más veces ocurra una palabra en un documento, más relevante es la palabra a la categoría del documento y
- entre más veces ocurra una palabra en todos los documentos de la colección, menor será su capacidad de discriminación entre categorías.

A continuación se mencionan los esquemas de pesado más utilizados en clasificación de textos.

- **Booleano**. Es el enfoque más simple, consiste en asignar un peso de 1 al término en el documento si este término aparece en él, y 0 en caso contrario:

$$w_{ij} = \begin{cases} 1 & \text{si el término } t_i \text{ aparece en el documento } j \\ 0 & \text{en caso contrario} \end{cases} \tag{2.2.1}$$

- **Frecuencia del término**. En este enfoque el peso es el número de repeticiones (o frecuencia) de cada término t_i en el documento j. Así un término será más importante entre más aparezca en el documento.

$$w_{ij} = f_{ij} \tag{2.2.2}$$

- **Frecuencia inversa**. En los enfoques anteriores no se considera la frecuencia del término a través de todos los documentos en la colección. Por lo que en este enfoque (también denotado como $tf \cdot idf$) se determina el peso de un término i en el documento j en proporción directa al número de veces que el término aparece en el documento, e inversamente proporcional al número de documentos en el conjunto D_t que contienen el término i. En particular, el peso está dado por la ecuación siguiente,

$$w_{ij} = tf_{ij} \times idf_i \tag{2.2.3}$$

donde tf_{ij} es la frecuencia del término t_i en el documento d_j y idf_i es la frecuencia inversa del documento definida como $\log(\frac{N}{df_i})$; $N = |D_t|$ y df_i es el número de documentos que contienen el término t_i. El factor idf_i es utilizado para eliminar el impacto de términos frecuentes que existen en la mayoría de los documentos (Lertnattee y Theeramunkong (2003); Han y Karypis (2000)).

Debido a que la evaluación del método propuesto en esta tesis se hace sobre escenarios de clasificación en los cuales el conjunto de entrenamiento y el conjunto de documentos a clasificar presentan diferencias significativas al provenir de dominios de clasificación distintos; se busca que los esquemas de pesado no sean muy dependientes del conjunto de entrenamiento usado. Por lo anterior, en esta tesis sólo se utilizan dos de los esquemas mencionados anteriormente, *booleano* y *frecuencia del término*. En el capítulo 4 se detalla el uso de tales esquemas en el método propuesto.

2.2.2. Reducción de la dimensionalidad

Como se ha mencionado anteriormente, el *diccionario* de una colección de documentos normalmente contiene un gran número de términos. Manejar toda esa cantidad de términos es un problema para los algoritmos de aprendizaje. Para resolver este problema se han desarrollado esquemas que ayudan a seleccionar un subconjunto de términos del diccionario, dicho subconjunto debe describir tan bien, o incluso mejor a cada documentos que el conjunto de términos original (Wu y Flach (2002)). A la elección del subconjunto de términos se le conoce como reducción de dimensionalidad que consiste, como se ha mencionado, en eliminar términos del *diccionario* con el objetivo de mejorar la efectividad de la clasificación y reducir la complejidad computacional (Aas y Eikvil (1999)).

Frecuencia de documento (DF), ganancia de información (IG), información mutua (MI) y Estadística χ^2 son los esquemas más utilizados (Yang y Pedersen (1997); Aas y Eikvil (1999)); mismos que se describen brevemente a continuación.

- **Frecuencia de documento (DF)**. La idea básica de este esquema es que los términos que aparecen en pocos documentos no aportan información para predecir la clase, al contrario, estos términos pueden ser considerados ruido y por lo tanto, influir negativamente el desempeño de la clasificación. La frecuencia de documento de un término es el número de documentos en el cual cada

término ocurre. Reducir mediante este método es sencillo, basta con calcular la frecuencia de documento para cada término del diccionario del conjunto de entrenamiento y quitar aquellos que tengan una DF menor a un umbral determinado.

- **Ganancia de información (IG)**. La idea básica es descubrir que tan bien un término puede separar el conjunto de datos. Dado un conjunto de entrenamiento, el procedimiento para obtener el subconjunto de términos es calcular la ganancia de información de cada término único utilizando la Ecuación 2.2.4, luego quitar del espacio de términos aquellos que su IG sea menor que algún umbral predeterminado.

$$IG(t) = -\sum_{i=1}^{|C|} P(c_i)\log P(c_i) + P(t)\sum_{i=0}^{|C|} P(c_i|t)\log P(c_i|t) + P(\bar{t})\sum_{i=1}^{|C|} P(c_i|\bar{t})\log P(c_i|\bar{t}) \tag{2.2.4}$$

En la Ecuación 2.2.4, $P(c_i)$ se estima como la fracción de documentos totales que pertenecen a la clase c_i, $P(t)$ la fracción de documentos en los cuales el término t aparece. Así como $P(c_i|t)$ se calcula como la razón de documentos de la clase c_i que contienen el término t, mientras que $P(c_i|\bar{t})$ es la fracción de documentos de la clase c_i en los que el término t no aparece.

- **Información mutua (MI)**. Es un criterio comúnmente utilizado en modelos de lenguaje estadísticos. Para cada término único se calcula su información mutua con relación a cada una de las clases del conjunto de entrenamiento utilizando la ecuación siguiente:

$$I(t,c) = \log \frac{P(t \wedge c)}{P(t) \times P(c)} \tag{2.2.5}$$

donde $P(t \wedge c)$ se estima como la fracción de documentos de la clase c que contienen el término t.

- **Estadística χ^2 (CHI-cuadrada)**. Es un método de selección de atributos basado en la prueba estadística χ^2, calculada sobre cada atributo con respecto a la clase para obtener el nivel de correlación entre ellas.

En esta tesis se utiliza el esquema de reducción de dimensionalidad DF (frecuencia de documento) por la siguientes dos razones: i) el estudio realizado en Yang y Pedersen (1997) demuestra que este esquema es tan efectivo como IG y χ^2, además de ser el de menor costo; ii) considerando la naturaleza del problema que este trabajo hace frente (insuficiencia e inexistencia de conjuntos de entrenamiento en el dominio particular de la clasificación) no es necesario utilizar enfoques de reducción más complejos que además tienen suposiciones de distribuciones iguales en el conjunto de entrenamiento y el conjunto de evaluación.

2.3. Algoritmos de aprendizaje

Una vez que la matriz del conjunto de entrenamiento está creada, se debe entrenar un clasificador. Para realizar este proceso, existen métodos de clasificación estadísticos y técnicas de aprendizaje automático que han sido aplicadas en esta tarea en años recientes.

Cuatro de los métodos más utilizados se presentan a continuación; para describirlos es necesario tomar en cuenta la siguiente notación: $\mathbf{d} = \{d_1, ..., d_M\}$ el vector del documento a ser clasificado, $c_1, ..., c_K$ las posibles clases, N el número de vectores de documentos $\mathbf{d}_1, ..., \mathbf{d}_N$ del conjunto de entrenamiento con clases $y_1, ..., y_N$ y, N_j el número de documentos de entrenamiento con clase c_j.

2.3.1. K-Vecinos más cercanos

De los métodos de clasificación supervisados utilizados por su simplicidad y bajo costo computacional son los llamados *basados en instancias*, que utilizan criterios de vecindad. La idea fundamental detrás de los métodos basados en instancias, en especial del método llamado K-Vecinos más cercanos (o kNN), es que las muestras pertenecientes a una misma clase, probablemente, se encontrarán cercanas en un espacio de representación común (Araujo (2006)).

La asignación de clase de un documento $\mathbf{d}$ se hace considerando los k documentos, en el conjunto de entrenamiento, más cercanos a $\mathbf{d}$; la clase de la mayoría de esos k documentos es la clase asignada al documento $\mathbf{d}$.

Un vecino más cercano es un documento que se define en términos de la distancia Euclidiana como sigue:

$$distancia(\mathbf{d_i}, \mathbf{d_j}) = \sqrt{\sum_{r=1}^{M}(t_r^i - t_r^j)^2} \tag{2.3.1}$$

donde $\mathbf{d_i}$ y $\mathbf{d_j}$ son los vectores de los documentos a comparar, M es el total de atributos en el diccionario, t_r^i es el r-avo atributo del documento $\mathbf{d_i}$ y t_r^j es el r-avo atributo del documento $\mathbf{d_j}$.

2.3.2. Clasificador bayesiano simple

No solo en la clasificación de textos los clasificadores bayesianos o probabilísticos son ampliamente utilizados debido a que presentan ciertas ventajas (Araujo (2006)):

- Generalmente, son fáciles de construir y entender.
- Las inducciones de estos clasificadores son extremadamente rápidas, requiriendo sólo un paso para hacerlo.
- Es muy robusto considerando atributos irrelevantes.
- Toma evidencia de muchos atributos para realizar la predicción final.

El enfoque de clasificación bayesiana permite asignar el valor de la etiqueta de clase más probable (c_i), dado un conjunto de atributos que describen a un nuevo caso o instancia ($\mathbf{d}$) de la siguiente forma:

$$clase_{\mathbf{d}} = argmax_{c_r} P(c_r|\mathbf{d}) \tag{2.3.2}$$

En clasificación de textos, el clasificador bayesiano simple (*Naïve Bayes*) estima la probabilidad de los documentos que pertenecen a cada clase con base en el peso de cada unos de los términos del vector del documento. Formalmente, un documento $\mathbf{d}$ pertenece a las clase c_j tal que, usando la regla de Bayes, la probabilidad $P(c_j|\mathbf{d})$ puede expresarse como (Rigutini, Maggini, y Liu (2005)):

$$P(c_j|\mathbf{d}) = \frac{P(c_j) \times P(\mathbf{d}|c_j)}{P(\mathbf{d})} \tag{2.3.3}$$

Debido a que $P(\mathbf{d})$ es un factor común en el modelo para cada clase, puede eliminarse. Además, la suposición (más importante aunque no cierta en la práctica) de

que la presencia de cada palabra en el documento es un evento independiente a la presencia o ausencia de las demás, permite definir lo siguiente:

$$P(c_j|\mathbf{d}) = P(c_j)\prod_{i=1}^{M} P(d_i|c_j) \tag{2.3.4}$$

La probabilidad *a priori* de la clase $P(c_j)$ puede estimarse de la distribución de los documentos en el conjunto de entrenamiento que pertenecen a la clase c_j:

$$P'(C = c_j) = \frac{N_j}{N} \tag{2.3.5}$$

Finalmente, para estimar la probabilidad $P(d_i|c_j)$ se utiliza la siguiente ecuación:

$$P'(d_i|c_j) = \frac{1 + d_{ij}}{M + N_j} \tag{2.3.6}$$

donde d_{ij} es el número de veces que la palabra i aparece en los documentos de la clase c_j del conjunto de entrenamiento.

2.3.3. Máquinas de vectores de soporte

Esta técnica tiene raíces en la teoría de aprendizaje estadístico, en los trabajos realizados por Vapnik (Cortes y Vapnik (1995)). La idea es que los documentos representados como vectores sean mapeados a un espacio de muy alta dimensionalidad. En este espacio, el problema es encontrar un plano de decisión que separe los ejemplos de entrenamiento positivos de los negativos con un margen máximo entre estas dos clases. La Figura 2.2 ilustra la idea de puntos de datos linealmente separables. Al plano de decisión en este espacio linealmente separable se le llama *hiperplano*. Las líneas puntadas paralelas a la línea sólida muestra qué tanto el hiperplano puede moverse sin que ésto lleve a una mala clasificación de los datos. *Margen* es la distancia entre estas lineas paralelas y los puntos (ejemplos) más cercanos al hyperplano son llamados *vectores de soporte*.

El hecho de que el hiperplano no tenga ningún error al separar los ejemplos de entrenamiento, no garantiza que con nuevos documentos suceda lo mismo, sin embargo entre mayor sea el *margen* menor será el riesgo de que un documento nuevo sea clasificado de manera errónea (Özgür et al. (2004); Álvarez Romero (2009)).

Formalmente, las máquinas de vectores de soporte clasifican un vector $\mathbf{d}$ dentro

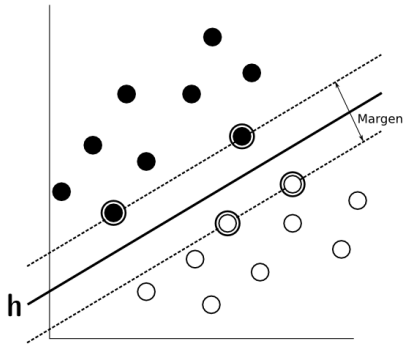


Figura 2.2: SVM encuentra el hiperplano h que separa los ejemplos de entrenamiento positivos de los negativos con un margen máximo. Los vectores de soporte están indicados sobre las líneas punteadas

de dos valores posibles: -1 y 1 (Aas y Eikvil (1999)) usando:

$$s = \mathbf{w}^T \phi(\mathbf{d}) + b = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{d}, \mathbf{d}_i) + b \tag{2.3.7}$$

y

$$y = \begin{cases} 1 & \text{si } s > s_0 \\ \text{-1} & \text{en caso contrario} \end{cases}$$

donde $\{y_i\}_{i=1}^{N}$ es la clase correspondiente a cada documento i y $y_i \in \{-1, 1\}$; $\phi(\mathbf{d})$ indica la representación del documento $\mathbf{d}$ en un espacio diferente al original. Esta representación se logra a partir de un kernel, los más comunes son los polinomios de grado d como el definido por la función:

$$K(\mathbf{d}, \mathbf{d}_i) = (\mathbf{d}^T \mathbf{d}_i + 1)^d \tag{2.3.8}$$

El entrenamiento de la máquina de vectores de soporte consiste en determinar el vector $\mathbf{w}$ que maximice la distancia entre los objetos de la clases diferentes dentro del conjunto de entrenamiento, donde $\mathbf{w}$ puede ser escrita como una combinación lineal de valores α_i, y_i y $\phi(\mathbf{d})$.

A pesar de que este enfoque nació en la clasificación binaria, la idea puede ser generalizada fácilmente para colecciones con más de dos clases de documentos. La idea es dividir el problema multi-clase y convertirlo en varios problemas binarios. Generalmente es usado en *uno contra uno* o *uno contra todos* (Weston y Watkins (1999)). Para *uno contra uno*, si se tienen q clases, se construye $q(q-1)/2$ clasificadores

usando los documentos de cada combinación de dos clases distintas. Para determinar la clase del documento nuevo se usa una estrategia de voto. En *uno contra todos*, se construyen q clasificadores, uno para cada clase, usando los ejemplos de una clase y mezclando todas las demás. En este segundo caso, el clasificador produce una función que da un valor relativamente mayor a una de las dos clases, al documento nuevo se le asigna la clase que obtuvo el valor más alto. En *uno contra uno* se construyen más clasificadores, pero cada clasificador tiene menos ejemplos de entrenamiento.

2.3.4. Clasificador basado en prototipos

La hipótesis detrás de los clasificadores basados en prototipos es asignar un documento a la clase donde la similitud entre el documento y el prototipo de esa clase sea la mayor, en relación a la similitud con los prototipos de todas las demás clases (Tan (2008)).

Este modelo puede resumirse como sigue: en la fase de entrenamiento, por cada clase se construye una instancia representativa, también llamada prototipo. Luego en la fase de prueba, cada documento no etiquetado es comparado contra todos los prototipos y es asignado a la clase del prototipo más similar; en particular, el proceso de asignación de clase es dado por:

$$clase(d) = argmax_i(sim(\mathbf{d}, \mathbf{P}_i)) \tag{2.3.9}$$

Dado que tanto el documento $\mathbf{d}$ y el prototipo $\mathbf{P}_i$ son vectores, la función de similitud $sim(\mathbf{d}, \mathbf{P}_i)$ utilizada es conocida como función de similitud coseno y se calcula como sigue:

$$sim(\mathbf{d}, \mathbf{P}_i) = \frac{\mathbf{d} \cdot \mathbf{P}_i}{||\mathbf{d}|| \times ||\mathbf{P}_i||} \tag{2.3.10}$$

Hasta el momento sólo se sabe que es necesario obtener un prototipo por clase utilizando los documentos etiquetados en cada una de ellas. Ahora se presentan algunas de las formulaciones más utilizadas para construir estos prototipos (Cardoso-Cachopo y Oliveira (2007)). Sea $\mathbf{D}_j$ el conjunto de documentos en el conjunto de entrenamiento etiquetados con la clase c_j y, N_j el número de documentos de la clase j:

- *Prototipo promedio.* Es el enfoque más simple, la idea se basa en la medida estadística conocida como media aritmética. Para cada clase, se crea un proto-

tipo promediando los vectores pertenecientes a esa clase, mediante la siguiente ecuación.

$$\mathbf{P}_j = \frac{1}{N_j} \cdot \sum_{\mathbf{d} \in \mathbf{D}_j} \mathbf{d} \tag{2.3.11}$$

- *Prototipo suma normalizada.* Una vía para representar a todos los vectores de los documentos de una clase en uno solo es, intuitivamente, sumándolos. Esa es la idea de este enfoque, aquí el prototipo de cada clase se calcula sumando los vectores de cada clase y luego normalizando el vector resultante para que su longitud sea unitaria y no se vea afectado por el número de documentos en cada clase, la formulación es la siguiente:

$$\mathbf{P}_j = \frac{1}{||\sum_{\mathbf{d} \in \mathbf{D}_j} \mathbf{d}||} \cdot \sum_{\mathbf{d} \in \mathbf{D}_j} \mathbf{d} \tag{2.3.12}$$

- *Fórmula de Roccio.* Utilizando esta fórmula para construir el prototipo de una clase particular se consideran no solo los documentos que pertenecen a esa clase si no también los que no pertenecen a ella. Para cada clase, el prototipo está definido como:

$$\mathbf{P}_j = \beta \cdot \frac{1}{N_j} \cdot \sum_{\mathbf{d} \in \mathbf{D}_j} \mathbf{d} - \gamma \cdot \frac{1}{N_{k \neq j}} \cdot \sum_{\mathbf{d} \notin \mathbf{D}_j} \mathbf{d} \tag{2.3.13}$$

  donde el valor de β se fija mayor al de γ para darle más peso a la suma de los ejemplos positivos. Los valores usuales de estas variables son $\beta = 16$ y $\gamma = 4$ (Joachims (1997)).

De los algoritmos de aprendizaje presentados, en esta tesis se utiliza el enfoque basado en prototipos. Se ha demostrado que el desempeño de estos métodos es mejor que otros algoritmos tales como el clasificador bayesiano simple, K-Vecinos más cercanos y C4.5 en tareas de clasificación de textos (Han y Karypis (2000)). Aunque no se encontró evidencia de ser mejor que las máquinas de vectores de soporte, son más eficientes que ellas, en particular, en la fase de clasificación donde la memoria y el tiempo gastados son proporcionales al número de clases en la colección. Del clasificador basado en prototipos se utiliza la Ecuación 2.3.12, prototipo suma normalizada, para construir cada prototipo pues es una de las técnicas ampliamente usada en esta tarea (Lertnattee y Theeramunkong (2004); Cardoso-Cachopo y Oliveira (2007)).

2.4. Medidas de evaluación

Una vez generado el clasificador es necesario medir su capacidad de predicción sobre las nuevas instancias, para ello es necesario contar con un conjunto de ejemplos (diferente al conjunto con el cual se generó el clasificador), llamado *conjunto de evaluación*, el cual consiste de un conjunto de casos tal como se describe en la Tabla 2.1.

Cuando no se cuenta con un conjunto de evaluación independiente del conjunto de entrenamiento se realiza una validación cruzada (*k-fold cross validation*), donde se divide el conjunto de entrenamiento en k subconjuntos de forma aleatoria, luego se entrena un clasificador utilizando $k-1$ subconjuntos y el restante se utiliza para evaluar los resultados. Este procedimiento se realiza k veces, al final se obtiene un promedio que resume los resultados obtenidos.

Existen dos métodos para obtener una evaluación general del desempeño del clasificador, *macro promedio* y *micro-promedio*. Para ambos métodos es necesario obtener cuatro valores por cada clase:

- a_j - el número de documentos asignados correctamente a la clase j.
- b_j - el número de documentos asignados incorrectamente a la clase j.
- c_j - el número de documentos rechazados incorrectamente de la clase j.
- d_j - el número de documentos rechazados correctamente de la clase j.

A partir de estos cuatro valores se definen las medidas de desempeño recuerdo, precisión y medida-F. Utilizando *macro promedio*, estas medidas son calculadas como se muestra en las Ecuaciones 2.4.1 a 2.4.3, donde K es el número de clases diferentes.

$$\text{macro recuerdo} = \frac{1}{K}\sum_{j=1}^{K}\frac{a_j}{a_j + c_j} \tag{2.4.1}$$

$$\text{macro precisión} = \frac{1}{K}\sum_{j=1}^{K}\frac{a_j}{a_j + b_j} \tag{2.4.2}$$

$$\text{macro medida-F} = \frac{2 * \text{macro recuerdo} * \text{macro precisión}}{\text{macro recuerdo} + \text{macro precisión}} \tag{2.4.3}$$

Para calcular el *micro promedio* se considera la colección completa, las Ecuaciones 2.4.4 a 2.4.6 muestran cómo se obtienen utilizando los valores a_j, b_j, c_j y d_j descritos anteriormente.

$$\text{micro recuerdo} = \frac{\sum_{j=1}^{K} a}{\sum_{j=1}^{K}(a_j + c_j)} \tag{2.4.4}$$

$$\text{micro precisión} = \frac{\sum_{j=1}^{K} a}{\sum_{j=1}^{K}(a_j + b_j)} \tag{2.4.5}$$

$$\text{micro medida-F} = \frac{2 * \text{micro recuerdo} * \text{micro precisión}}{\text{micro recuerdo} + \text{micro precisión}} \tag{2.4.6}$$

En esta tesis se utilizaron macro promedios pues algunas colecciones con las que se evaluó el método propuesto contenían clases desbalanceadas, es decir, el número de documentos por cada clase varía considerablemente.

2.5. Resumen

En este capítulo se definió la clasificación de textos como una tarea donde se aplican métodos y algoritmos de aprendizaje automático. Se mencionan los procesos principales implicados en el desarrollo de métodos de clasificación de textos y se presentaron más de una técnica o método que pueden ser usados en cada proceso. A continuación se especifica cuál de todas las opciones descritas a lo largo del capítulo serán utilizadas en esta tesis.

Dentro de la etapa de extraer las características de los documentos, se eliminan las palabras vacías en el preprocesamiento y se utiliza un enfoque de pesado booleano por la naturaleza del problema que se quiere resolver. Para reducir el número de términos que describen a cada documento, se utiliza la frecuencia de documento. Los cuatro algoritmos de aprendizaje mencionados, son utilizados como métodos de referencia en la evaluación del método propuesto. Sin embargo, el método que se explicará a detalle en el Capítulo 4 toma como algoritmo base el basado en prototipos. Para evaluar el método de clasificación se utilizaron las medidas de macro promedios pues algunas de las colecciones tienen clases de diferentes tamaños.

Capítulo 3

Trabajo relacionado

La clasificación automática de textos resuelta mediante el enfoque tradicional supervisado, requiere un gran número de documentos de ejemplo (etiquetados) para aprender un clasificador con buen desempeño al clasificar documentos no vistos anteriormente. Los esfuerzos para resolver este problema se han encaminado en tres sentidos: i) métodos que hacen uso de documentos no etiquetados, ii) métodos que utilizan documentos etiquetados de dominios similares y, iii) métodos que utilizan documentos etiquetados del mismo dominio pero en otros lenguajes. En el presente capítulo se describe en qué consisten cada uno de estos escenarios además de presentar ejemplos de métodos desarrollados dentro de cada uno de ellos.

Para finalizar el capítulo se discuten ventajas y desventajas que presentan los tres enfoques y se ubica el método propuesto en este trabajo de investigación dentro de este contexto.

3.1. Métodos en condiciones de insuficiencia de documentos etiquetados

Si el problema es la insuficiencia de datos etiquetados, es necesario generar métodos utilizando los pocos datos disponible y tratar de entrenar buenos clasificadores, tan buenos como lo sería uno aprendido con suficientes datos etiquetados. Estos métodos parten de la observación intuitiva de que a pesar de que no existen suficientes documentos etiquetados, es relativamente sencillo obtener documentos no etiquetados. Estos conjuntos de documentos no etiquetados deben proporcionar algún tipo de información que sea útil para mejorar el clasificador resultante.

Los métodos dentro de este enfoque, a su vez, pueden organizarse dentro de dos

grupos:

- Semi-supervisados. Este enfoque es uno de los más populares dentro de este escenario. Considera, principalmente, el uso de un gran número de datos no etiquetados junto con una pequeña cantidad de datos etiquetado, con el fin de construir mejores clasificadores.

  Existen muchas técnicas dentro de este enfoque, de las más utilizadas en clasificación de textos están: *self-trainig*, aquí inicialmente se entrena un clasificador usando una pequeña cantidad de datos etiquetados; después, este clasificador es usado para clasificar los datos no etiquetados, de éstos, se eligen los documentos más confiables en conjunción con la etiqueta de clase predicha por el clasificador y es agregado al conjunto de entrenamiento; el clasificador es re-entrenado y el proceso es repetido. Otra técnica es conocida como *co-training* aquí se parte de las suposiciones que los atributos pueden ser dividido dentro en dos grupos, cada subconjunto es suficiente para entrenar un buen clasificador, y los dos conjuntos son condicionalmente independientes dada la clase; primero se entrenan dos clasificadores con el conjunto de datos etiquetados utilizando los dos subconjuntos de atributos, luego se clasifican los documentos sin etiquetas y se usan los más confiables de cada clasificador para incorporarlos como entrenamiento y re-entrenar el otro (Zhu (2005)).

  Existen una gran cantidad de trabajos que desarrollan nuevos métodos utilizando el enfoque semi-supervisado, donde las técnicas anteriores son las más aplicadas. Las soluciones planteadas en su mayoría tratan de resolver el problema de insuficiencia de datos etiquetados (Guzmán-Cabrera et al. (2009); Raskutti, Ferrá, y Kowalczyk (2002); Cardoso-Cachopo y Oliveira (2007); Cong et al. (2004); Nigam et al. (1999)), pero algunos usan este enfoque para tratar de solucionar problemas de inexistencia de documentos etiquetados en el dominio particular de clasificación (Aue y Gamon (2005); Raskutti, Ferrá, y Kowalczyk (2002); Wan (2009)).

  Una de las desventajas mas sobresalientes de estos métodos es que el clasificador tiene que ser re-entrenado en cada iteración, lo que es costoso computacionalmente.

- *Clustering* para clasificar. Este enfoque se basa en uno de los algoritmos populares del aprendizaje no supervisado. La idea es obtener información del conjunto

de documentos a clasificar a través de algoritmos de clustering. La información obtenida es un conjunto de atributos que pueden ser usados como sigue:

- como el conjunto de atributos que se usarán para representar a cada documento (documentos etiquetados y los que se quieren clasificar). Una vez que todos los documentos estén representados en ese espacio de atributos, se utilizan métodos de clasificación supervisados. Derivaux, Forestier, y Wemmert (2008) y Kyriakopoulou y Kalamboukis (2007), presentan un método en el cual agrupan el conjunto total de documentos (etiquetados y no etiquetados) con diferentes algoritmos de *clustering*, luego a cada grupo le asignan un identificador, este identificador será un nuevo atributo. De esta forma generan un conjunto de atributos que consiste de identificadores de grupos generados a partir de diferentes algoritmos de agrupamiento. Luego utilizan un algoritmo de clasificación supervisada.
- como atributos para formar un segundo conjunto de atributos independiente, para utilizarlo en un enfoque de *co-training*. Raskutti, Ferrá, y Kowalczyk (2002), crean un conjunto de atributos a partir de los identificadores de grupos generados por algún algoritmo de *clustering*, además de las similitudes de cada documento a los centroides de cada grupo generado. Así crean una vista alternativa e independiente de los documentos a clasificar. Por otro lado, Fang, Parthasarathy, y Schwartz (2001), utiliza algoritmos de *clustering* para obtener un conjunto de palabras clave, que definen a cada grupo, con el objetivo de agregarlas al vector de términos que representan a cada documento.

3.2. Métodos de clasificación multi-dominio

Al no contar con documentos etiquetados para construir un clasificador con buen desempeño, una opción común de los enfoques propuestos es utilizar documentos etiquetados de dominios similares; de alguna forma, transferir el conocimiento de un dominio a otro. En este proceso es posible notar que las diferencias entre dominios, causan un bajo desempeño en clasificadores generados bajo los enfoques tradicionales de aprendizaje automático. El reto es disminuir esas diferencias para obtener los resultados esperados en la tarea de clasificación de textos. En particular, se desea

que dado un conjunto de documentos etiquetados en un dominio fuente (D_f), un conjunto de documentos no etiquetados en un dominio objetivo (D_o), y un conjunto de clases C, estimar la hipótesis $h : D_o \rightarrow C$. Este enfoque es conocido también como *transfer learning* o *domain adaptation* una formulación de este problema, centrada en clasificación de textos, es presentada en Ben-David et al. (2010).

Una constante en los métodos dentro de este escenario es la idea de etiquetar una pequeña cantidad de documentos del dominio objetivo utilizando un clasificador generado a partir de los documentos etiquetados del dominio fuente, con el objetivo de usarlos de dos maneras distintas:

- re-entrenar un modelo utilizando sólo los documentos etiquetados del dominio objetivo Tan et al. (2007),
- utilizarlos para seleccionar documentos útiles en el conjunto de entrenamiento del dominio fuente que puedan ayudar a la clasificación de los documentos en el dominio objetivo. Usan los documentos etiquetados del dominio fuente y del dominio objetivo para generar un nuevo clasificador que sirva para etiquetar los documentos restantes del dominio objetivo (Zhen y Li (2008))

Otra clase de métodos dentro de este escenario corresponde a los que han desarrollado una versión de los métodos de clasificación tradicionales para adaptarlos a este escenario, por ejemplo Wenyuan Dai et al. en (Dai et al. (2007)) proponen un clasificador bayesiano para transferir un clasificador de un dominio a otro. Dado que D_f y D_o tienen distribuciones distintas, el método que proponen primero estima un modelo inicial bajo la distribución de D_f; luego, se aplica el algoritmo Expectation-Maximization para ajustar el modelo a un modelo óptimo local bajo la distribución de los datos no etiquetados (D_o), de esta manera se transfiere el modelo de predicción de una distribución a otra.

3.3. Métodos de clasificación multi-lenguaje

Clasificación entre lenguajes, también llamado CLTC (*Crosslingual Text Classification*) consiste en explotar documentos en un idioma origen para clasificar documentos en diferentes idiomas (lenguajes objetivos) (Gliozzo y Strapparava (2005)). Formalmente, sea X_f un conjunto de documentos etiquetados en un lenguaje origen (L_f), X_o el conjunto de documentos en un lenguaje objetivo (L_o) sin etiquetas, y

un conjunto de etiquetas de clase C; dados un conjunto de documentos etiquetados de entrenamiento disponibles sólo en un lenguaje, CLTC desea estimar una hipótesis $h : X_o \rightarrow C$ para clasificar documentos escritos en otro lenguaje (Ling et al. (2008)).

Existen diferentes esquemas de clasificación multi-lenguaje. Uno de los más utilizados se centra en representar a todos los documentos (X_f y X_o) en un sólo lenguaje; con el fin de lograr esto se han desarrollado dos esquemas: por un lado traducir de un lenguaje a otro ya sea de L_f a L_o o de L_o a L_f) (Rigutini, Maggini, y Liu (2005); Escobar-Acevedo, Montes-Y-Gómez, y Villaseñor-Pineda (2009)); por otro lado, se pueden traducir en ambos sentidos para generar dos conjuntos de atributos y luego utilizarlos bajo el enfoque *co-training* (Wan (2009)).

Otras clases de métodos tratan de extraer información del conjunto de documentos del lenguaje objetivo. Por ejemplo, Ling et al. (2008), ellos crean un modelo para CLTC basado en el enfoque llamado *information bottleneck* (Tishby, Pereira, y Bialek (1999)), utilizan este enfoque para extraer las partes comunes de documentos en dos diferentes lenguajes. Por otro lado, Leonardo Rigutini et al., proponen otro método basado en el algoritmo *Expectation-Maximization* en el cual mediante un proceso iterativo etiqueta cada documento del conjunto de evaluación, construye un nuevo clasificador utilizando esos documentos ahora etiquetados y vuelve a clasificar el mismo conjunto. El proceso termina cuando las etiquetas del conjunto de prueba no cambian de una iteración a otra.

3.4. Discusión

Los trabajos presentados en las secciones anteriores tienen dos características principales. Por un lado, utilizan recursos existentes para construir clasificadores de textos en distintas áreas, se fundamentan en incorporar información de otros dominios o lenguajes para adaptar un clasificador a las necesidades específicas de cierta aplicación. Cada uno de los métodos mencionados sólo funcionan en la tarea específica para la cual fueron diseñados. Por otro lado, la incorporación de la información, en la mayoría de los trabajos presentados, involucra un proceso iterativo en el cual re-entrenan los modelos de clasificación en cada iteración, esto sucede porque es una manera efectiva de incorporar información de documentos no etiquetados de dominios similares o no.

A diferencia de los métodos descritos en este capítulo, el método propuesto no pretende mejorar el clasificador resultante, en su lugar se desea mejorar la clasificación

de los documentos generada a partir de un clasificador de bajo desempeño (como es el caso al construir clasificadores utilizando como conjunto de entrenamiento los disponibles en los tres escenarios de clasificación mencionados). Además, puede ser utilizado en los tres escenarios, no es dependiente de la tarea específica que se trata de resolver. El método propuesto en esta tesis es simple y no requiere de iteraciones para construir el clasificador resultante.

Igual que los métodos descritos, el método propuesto incorpora información de los documentos de ambos conjuntos (entrenamiento disponible, y evaluación). Pero esta incorporación se realiza utilizando la clasificación individual de las instancias más similares a cada documento a clasificar. En los siguientes capítulos se presentará detalladamente el método propuesto así como la evaluación realizada en los tres escenarios de clasificación.

Capítulo 4

Método propuesto

En este capítulo se describe el método de clasificación de textos propuesto. Su característica principal consiste en determinar la clase de cada documento consensuando información procedente de dos fuente: i) del conjunto de entrenamiento, es decir, los prototipos construidos con los documentos etiquetados y ii) del conjunto de documentos a clasificar, considerando documentos similares al que se desea clasificar.

4.1. Vista general del método propuesto

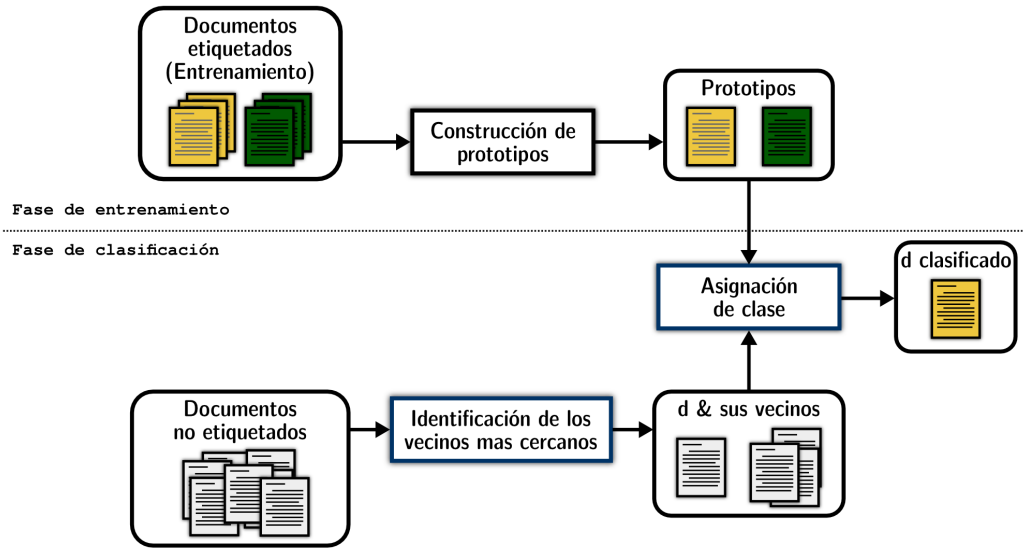


Figura 4.1: Esquema general del método de clasificación de textos propuesto. Consiste de dos fases: Entrenamiento, en la cual se construyen los prototipos por cada clase; y Clasificación, donde se identifican los vecinos más cercanos de cada documentos en el conjunto a clasificar y se hace la asignación de clase utilizando esta información obtenida

En la Figura 4.1 se muestra el esquema general del método propuesto. Este consiste de dos fases principales: **entrenamiento** y **clasificación**.

En la **fase de entrenamiento** se construye un clasificador tal cual se haría en un enfoque de clasificación supervisado. En particular, se utiliza un clasificador basado en prototipos para este fin, debido a su simplicidad y la naturalidad con que puede extenderse dentro de un enfoque de clasificación consensuada, donde sea posible utilizar información del conjunto a clasificar, particularmente, similitudes entre documentos. Para dar una idea general de cómo funcionan los clasificadores basados en prototipos se muestra la Figura 4.2, en la cual se representan dos clases: `verde` y `amarilla`, el conjunto de instancias no etiquetadas están contenidos en el rectángulo y la instancia a clasificar está representada por un estrella ($\star$). El prototipo de cada clase está representado por un cuadrado ($\square$) en el centro de los círculos que representan las clases.

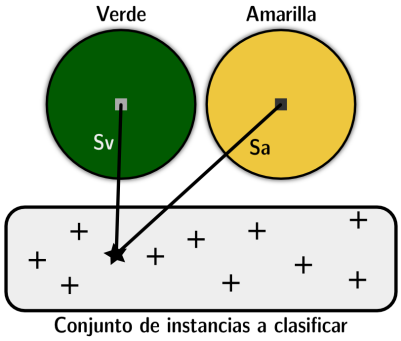


Figura 4.2: Representación gráfica de un clasificador basado en prototipos. Se representa que la instancia $\star$ del conjunto de instancias a clasificar debe ser etiquetada como `verde` ya que está mas cercana al prototipo de esta clase que al prototipo de la clase `amarilla`

La asignación de la clase de una nueva instancia ($\star$) está determinada por la similitud entre esa instancia y cada prototipo. En el caso de la Figura 4.2, la clase de la instancia $\star$ será `verde` (la distancia que los separa es la menor que la distancia al prototipo de la clase `amarilla`, en otras palabras, la similitud es mayor).

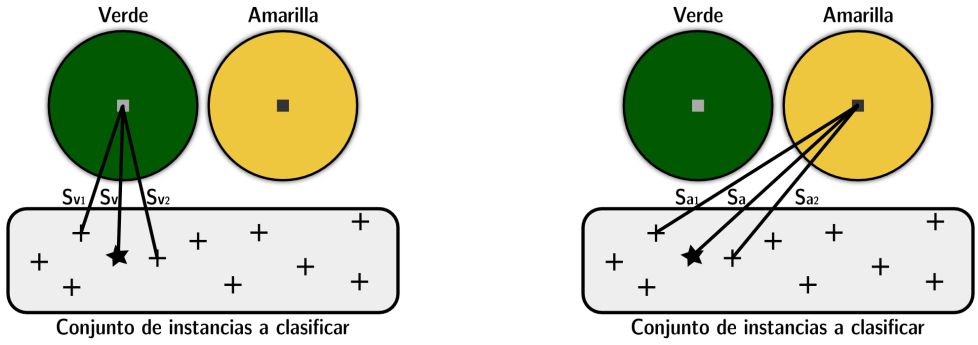
La segunda parte del método es la **fase de clasificación**. En esta fase es donde se hace la mayor aportación al estado del arte. Partiendo de la observación en corpus (en clasificación de textos) y de trabajos previos, en esta fase se consideran las siguientes premisas:

- documentos similares, de acuerdo con alguna medida de similitud, deben pertenecer a la misma clase, y

- el conjunto de documentos a clasificar posee algún tipo de organización intrínseca que es útil para la clasificación de cada instancia en ese conjunto.

Para incluir información del conjunto de documentos no etiquetados, es decir del conjunto de documentos a clasificar, cada documento de este conjunto será asociado a un subconjunto de documentos similares. De este pequeño subconjunto (el documento a clasificar y los documentos más similares a él) se tomará información para la clasificación de un documento particular.

En la Figura 4.3 se muestra de forma gráfica cómo se utilizará la información obtenida del conjunto no etiquetado. En la figura, dentro del conjunto de instancias a clasificar, sólo se consideran dos vecinos de la instancia estrella ($\star$). El método toma en cuenta las similitudes de las tres instancias (la instancia a clasificar y sus dos vecinos) contra cada uno de los prototipos. En el inciso (a) se muestran las similitudes consideradas utilizando el prototipo de la clase `verde`, y en el inciso (b) las similitudes al prototipo de la clase `amarilla`. La instancia estrella es asignada, en este caso, a la clase `verde` al ser la combinación de las similitudes S_v, S_{v1} y S_{v2} mayor que la combinación de las similitudes S_a, S_{a1}, S_{a2}.



(a) Similitudes al prototipo de la clase **verde**

(b) Similitudes al prototipo de la clase **amarilla**

Figura 4.3: Representación del uso de información del conjunto a clasificar por el método propuesto. Además de medir la similitud de la instancia a clasificar a los prototipos de las dos clases representadas, se mide también la similitud de los vecinos más cercanos de esta instancia a ambos prototipos

En resumen, en esta fase se identifican los documentos más similares en el conjunto de documento no etiquetado: se encuentran los *vecinos más cercanos* de cada documento. También se decide la etiqueta de cada uno de esos documentos considerando información tanto de él como de sus vecinos.

En las secciones siguientes se dará una descripción más detallada de cada unos de los procesos involucrados en el método de clasificación propuesto.

4.2. Construcción de prototipos

Debido a la simplicidad, rapidez y efectividad de los métodos basados en prototipos, se consideró a este enfoque como método de clasificación base en el método propuesto.

La primera fase del método corresponde a la construcción de los prototipos. Dado un conjunto de entrenamiento organizado en N clases, se construyen n prototipos, uno por cada clase, utilizando la ecuación del prototipo *suma normalizada* (Ecuación 2.3.12) que se presentó en la Sección 2.3.4 y que se muestra nuevamente a continuación:

$$\mathbf{P}_j = \frac{1}{||\sum_{\mathbf{d}\in\mathbf{D}_j} \mathbf{d}||} \cdot \sum_{\mathbf{d}\in\mathbf{D}_j} \mathbf{d}$$

En resumen, las entradas y salidas de este proceso son:

- **Entrada:** Un conjunto de vectores de documentos etiquetados y organizados en n clases. Estos vectores son construidos a partir de los documentos de cada categoría utilizando los esquemas de indexado descritas en la Sección 2.2.1. En particular, se utilizó un enfoque de pesado *booleano*, pues no es necesario considerar mas información sobre los términos del conjunto de entrenamiento. El método está diseñado para escenarios de clasificación donde el conjunto de entrenamiento y el conjunto de evaluación no pertenecen al mismo dominio de clasificación (recordar que para el entrenamiento se pretende utilizar conjuntos de documentos etiquetados disponibles).

- **Salida:** n prototipos, cada uno representando a todos los documentos de cada una de las n clases.

4.3. Identificación de los vecinos más cercanos

Este proceso se enfoca en la identificación de los k vecinos más cercanos para cada documento $\mathbf{d}$ del conjunto de documentos a clasificar, D_T. Para llevar a cabo esta tarea, primero se calcula la similitud entre cada par de documentos en D_T, usando la ecuación de la similitud coseno (Ecuación 2.3.10); luego, basados en esos valores de similitud se seleccionan los k documentos más similares al documento $\mathbf{d}$ como vecinos

más cercanos. La siguiente ecuación define el conjunto de los k vecinos más cercanos para cada documento $\mathbf{d}_i \in D_T$.

$$N_k^{\mathbf{d}_i} = argmax_{S_j \in \mathbb{S}_k} \left[\sum_{\mathbf{d} \in S_j} sim(\mathbf{d}, \mathbf{d}_i) \right] \tag{4.3.1}$$

donde $\mathbb{S}_k$ y sim están definidos como:

$$\mathbb{S}_k = \{S | S \subseteq D_T \wedge |S| = k\} \tag{4.3.2}$$

$$sim(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{||\mathbf{d}_i|| \times ||\mathbf{d}_j||} \tag{4.3.3}$$

Este es un proceso fundamental en el diseño del método propuesto, de aquí se obtiene información del conjunto de documentos a clasificar, también llamado conjunto objetivo. Detalles de las entradas y salidas de este proceso se listan abajo.

- **Entrada:** Un conjunto de vectores de documentos no etiquetados (los que se van a clasificar). Un aspecto importante de este proceso es la forma de representación de los documentos del conjunto objetivo; para capturar toda la información posible de este conjunto se utiliza el esquema de pesado de frecuencia del término (detalles de este esquema se han proporcionado en la Sección 2.2.1); por lo tanto, el espacio de términos de la representación de estos documentos depende sólo del conjunto D_T.

- **Salida:** El mismo conjunto de entrada organizado en paquetes, cada paquete consiste de un documento $\mathbf{d_i}$ en D_T y el conjunto de vecinos más cercanos a él $N_k^{d_i}$. La cantidad de paquetes es igual al número de documentos en el conjunto D_T.

4.4. Asignación de clases

En la clasificación basada en prototipos, la clase de un documento $\mathbf{d}$ es tradicionalmente determinada por la Ecuación 4.4.1, donde $\mathbf{P_i}$ corresponde al vector prototipo de la clase i. La idea de esta clasificación es simple, la clase del documento $\mathbf{d}$ será la clase representada por el prototipo más cercano a dicho documento.

$$clase(\mathbf{d}) = argmax_i \left(sim(\mathbf{d}, \mathbf{P}_i) \right) \tag{4.4.1}$$

La propuesta presentada en esta tesis, extiende esta estrategia de asignación de clase considerando no sólo la información del propio documento (la similitud entre el documento y el prototipo de cada clase), sino también la información de la clase asignada a otros documentos similares del mismo conjunto objetivo. Esta nueva forma de asignación de clase se basa en la idea de que documentos cercanos pertenecen a la misma clase (como se ha mencionado previamente); si esto ocurre, es evidente que el prototipo de la clase de **d** y sus vecinos más cercanos, producirá la mayor similitud.

En particular, dado un documento del conjunto objetivo ($\mathbf{d} \in D_T$) en conjunción con sus k vecinos más cercanos (N_k^d), se asigna una clase a **d** usando la ecuación 4.4.2.

$$clase(\mathbf{d}) = argmax_i \left(\lambda \; sim(\mathbf{d}, \mathbf{P}_i) + (1 - \lambda) \frac{1}{k} \sum_{n_j \in N_k^d} [inf(\mathbf{d}, \mathbf{n}_j) \times sim(\mathbf{n}_j, \mathbf{P}_i)] \right) \tag{4.4.2}$$

donde,

- $sim(\mathbf{v}_i, \mathbf{v}_j)$ es la función de similitud coseno definida en la Ecuación 4.3.3, donde $\mathbf{v_i}$ y $\mathbf{v_j}$ son vectores.
- N_k^d es la k-vecindad de **d**, obtenida a partir de la Ecuación 4.3.1.
- λ es la constante usada para determinar la importancia relativa tanto de la información del propio documento **d** como la información de sus vecinos. Valores pequeños de λ indican una mayor contribución de los vecinos, y viceversa. Este parámetro puede variar dependiendo del problema específico de clasificación; es de nuestro interés analizar el comportamiento del método con diferentes valores de esta constante.
- $inf(\mathbf{d}, \mathbf{n})$ es una función de influencia usada para pesar la contribución de cada vecino **n** a la clasificación de **d**. El propósito de esta función es dar mas relevancia a vecinos más cercanos. En particular, se definió esta influencia directamente proporcional a la similitud entre cada vecino y el documento, esta similitud es calculada utilizando la fórmula coseno (ver Ecuación 4.3.3). Se utiliza un mismo número de vecinos para todos los documentos; sin embargo, se

deja que la función por su definición, se encargue de considerar sólo los documentos muy cercanos al documento a clasificar. Por ejemplo,si en el conjunto de vecinos existen algunos muy alejados del documento a clasificar, es posible decir que su influencia es también muy pequeña, por lo tanto se define su influencia directamente proporcional a la distancia que los separe.

Para comparar cada documento del conjunto objetivo con los prototipos de cada clase se necesita que todos los vectores estén en el mismo espacio de términos. Estos detalles se mencionan a continuación.

- **Entrada:**
  - Un conjunto formado por el documento a clasificar **d**, y sus vecinos N_k^d, representados bajo el mismo espacio de términos que los prototipos creados en la fase de entrenamiento. Es decir, tanto **d** como N_k^d son representados ahora utilizando el esquema de pesado *booleano* y el conjunto de términos obtenido del conjunto de entrenamiento, con la única finalidad de poder realizar las comparaciones necesarias.
  - Los prototipos de las n clases calculados en la fase de entrenamiento.
- **Salida:** El documento **d** etiquetado, es decir, clasificado dentro de las n clases del conjunto de entrenamiento.

4.5. Resumen

El método descrito tiene como clasificador base un enfoque basado en prototipos el cual es simple y eficaz en la clasificación de textos. Consta de dos fases, en la de entrenamiento se construyen los prototipos que representan a cada clase, utilizando un conjunto de documentos etiquetados. Posteriormente se realiza la clasificación de un conjunto de documentos utilizando una combinación lineal de las similitudes de cada documento y un conjunto de documentos vecinos (obtenidos del conjunto de documentos no etiquetado) a cada uno de los prototipos obtenidos en la fase anterior.

4.5.1. Principales características del método propuesto

La integración de la información del conjunto objetivo se realiza de forma sencilla, sin procesos complicados o costosos, mediante una combinación lineal de similitudes.

Mientras otros métodos descritos en el trabajo relacionado concentran sus esfuerzos en mejorar el proceso de aprendizaje en la fase de entrenamiento, el método propuesto no construye un clasificador cada vez que se traslade el método a otro problema de clasificación, sus esfuerzos están centrados en incluir información relevante del conjunto de documentos objetivo.

Estas características hacen que la principal ventaja del método propuesto sea su **flexibilidad**. Por lo tanto, puede ser llevado a diferentes escenarios, como los tres utilizados en la evaluación del método: 1) insuficiencia de documentos en el conjunto de entrenamiento, 2) clasificación multi-lenguaje y, 3) clasificación multi-dominio; como se explicará en el siguiente capítulo.

Capítulo 5
Evaluación

En problemas reales de clasificación de textos el objetivo es clasificar un conjunto de documentos en un dominio específico; sea X_T el conjunto de documentos a clasificar de un dominio D_T escritos en un lenguaje L_T, puede suceder las siguientes tres situaciones:

1.- No existen suficientes documentos etiquetados en el dominio D_T para generar clasificadores con buen desempeño, por lo tanto se desea evaluar el desempeño del método cuando el tamaño del conjunto de entrenamiento es muy pequeño.

2.- Existen suficientes documentos etiquetados en el dominio D_T pero estos documentos están en un lenguaje distinto L_F, en este caso se desea probar el método en un esquema de clasificación multi-lenguaje en el cual se entrena con documentos de un lenguaje (L_F) y se clasifican documentos en un lenguaje distinto (L_T).

3.- No existen conjuntos de documentos etiquetados en el dominio D_T sin embargo existen suficientes documentos etiquetados en un dominio similar D_F. La evaluación consistirá en medir el desempeño en un esquema multi-dominio.

Con el objetivo de evaluar la flexibilidad y robustez del método se diseñaron y ejecutaron experimentos en las tres situaciones anteriores.

5.1. Configuraciones globales

El método es evaluado utilizando macro-promedios, en particular se usó la medida-F, medida de desempeño descrita en la Sección 2.4.

Para generar resultados de referencia (o *baseline*) que sirvan para comparar los obtenidos por el método propuesto, se clasificaron los documentos del conjunto X_T con cuatro de los clasificadores más utilizados en clasificación de textos, **Máquinas de Vectores de Soporte** (SVM), **Clasificador bayesiano simple** (NB), **k-Vecinos más cercanos** (kNN) y un **Clasificador basado en Prototipos** (CBP), descritos en la Sección 2.3. La implementación usada de SVM, NB y kNN es la del software WEKA (Witten y Frank (2005)), con la configuración por omisión. Para la implementación del CBP, se utiliza un prototipo de suma normalizada (Ecuación 2.3.12) y para asignar la clase se utiliza la Ecuación 2.3.9.

La asignación de clase del método propuesto involucra la utilización de dos valores que pueden ser establecidos dependiendo del problema que se trate. Particularmente, se está interesado en analizar el impacto del método con diferentes valores de las variables k y λ de la Ecuación 4.4.2. Por un lado, el número de vecinos que deben considerarse para apoyar la clasificación de cada documento (el valor de la variable k) y por otro lado, se desea analizar la constante λ que es utilizada para establecer la proporción en que se utilizará la información proveniente del documento mismo a clasificar y la información proveniente de los vecinos más cercanos de dicho documento.

Los valores utilizados para cada parámetro son los mismos a través de los tres escenarios:

- El número de vecinos k fue variado desde $k = 1$ hasta $k = 30$.

- El valor de λ fue cambiado en el rango de valores entre $[0, 1]$ en incrementos de 0.1.

Para determinar que los resultados no fueron obtenidos por azar, se realizó la prueba estadística z-test con una confianza del 95 % [1].

En las siguientes secciones se detalla la realización de la evaluación para cada uno de los tres escenarios, incluyendo el corpus, los resultados de referencia, los resultados obtenidos por el método propuesto y las conclusiones obtenidas.

[1] La prueba z-test utilizada investiga si existe diferencia entre un valor dado y un valor observado (Kanji (2006)).

5.2. Escenario 1: Ejemplos insuficientes en el conjunto de entrenamiento

El primer escenario de clasificación en que se evaluó el desempeño del método propuesto es cuando no existen suficientes documentos etiquetados en el dominio D_T (dominio del conjunto de documentos a clasificar) para generar clasificadores con buen desempeño, por lo tanto se desea usar el conjunto de pocos datos etiquetados.

A continuación se describe el corpus utilizado, la configuración utilizada y los resultados obtenidos en este caso.

5.2.1. Corpus: Reuters-21578

Esta colección ha sido ampliamente utilizada durante la última década en investigaciones de recuperación de información, aprendizaje automático, clasificación de textos y otras investigaciones basadas en corpus. Consta de un conjunto de 21578 documentos de noticias del año 1987 de la agencia de noticias Reuters. El conjunto contiene 135 clases temáticas. Existe una versión de esta colección donde se ha dividido en dos subconjuntos, entrenamiento y evaluación, a la que se ha llamado *Distribución 1.0 de Reuters-21578* (Lewis (1991)).

Esta colección contiene clases con distintas cantidades de documentos, haciendo de ella una colección *desbalanceada*. A partir de ella se han creado subconjuntos, con la finalidad de poder utilizarla por diferentes tipos de métodos.

En esta tesis se utiliza la partición llamada R8, previamente utilizada por Cardoso-Cachopo (Cardoso-Cachopo y Oliveira (2007)) y Álvarez-Romero (Álvarez Romero (2009)), entre otros. Consiste de las 8 clases más grandes de Reuters-21578 y los documentos sólo pertenecen a una de ellas. Información sobre este conjunto se presenta en la Tabla 5.1.

Con el objetivo de evaluar el método en situaciones de insuficiencia de documentos en el conjunto de entrenamiento, se generaron tres colecciones pequeñas a partir de la colección R8. Se eligieron, de forma aleatoria, 41, 20 y 10 documentos etiquetados del conjunto de entrenamiento por cada clase [2]; de esta manera se crearon tres colecciones: R8-reducida-41, R8-reducida-20 y R8-reducida-10.

[2]La selección aleatoria de estos subconjuntos de documentos etiquetados de la colección R8 se realizó una sola vez.

Clase	Doctos. en el conjunto de entrenamiento	Doctos. en el conjunto de evaluación	Total
acq	1596	696	2292
crude	253	121	374
earn	2840	1083	3923
grain	41	10	51
interest	190	81	271
money-fx	206	87	293
ship	108	36	144
trade	251	75	326
Total	5485	2189	7674

Tabla 5.1: Colección R8: información sobre el número de documentos por clase divididos en entrenamiento y evaluación. La clase con más documentos consta de 3923, mientras que la de menos documentos contiene 51, claramente, la colección R8 es desbalanceada

La Tabla 5.2 muestra algunos datos sobre estas nuevas colecciones, como el número de documentos en el conjunto de entrenamiento y el número de términos en el diccionario para cada una de ellas. El número de documentos en el conjunto de evaluación es el mismo para todas las colecciones utilizadas y corresponden al conjunto de prueba de la colección R8, 2189 documentos (ver Tabla 5.1).

Colección	Doctos. en el conjunto de entrenamiento	Número de términos en el diccionario
R8	5485	3711
R8-reducida-41	328	2887
R8-reducida-20	160	1807
R8-reducida-10	80	1116

Tabla 5.2: Colecciones R8-reducidas: información de las tres colecciones derivadas de la colección R8 utilizadas para evaluar el método propuesto en condiciones de insuficiencia de documentos etiquetados, se muestra el número de términos en el diccionario de cada colección

5.2.2. Resultados de referencia

Los resultados de comparación utilizando cuatro métodos de clasificación (kNN, NB, SVM y CBP) se muestran en la Tabla 5.3. Es posible notar que los mejores resultados son obtenidos al utilizar el Clasificador basado en Prototipos, en especial con las colecciones reducidas. Entre menor es el conjunto de entrenamiento, el clasi-

ficador basado en prototipos obtiene una mejora significativa en relación a los otros tres clasificadores.

Colección	kNN	NB	SVM	CBP
R8	-	0.828	**0.886**	0.876
R8-reducida-41	0.520	0.747	0.812	**0.836**
R8-reducida-20	0.530	0.689	0.760	**0.803**
R8-reducida-10	0.393	0.634	0.646	**0.767**

Tabla 5.3: Colección R8-reducidas: resultados de referencia utilizando la medida-F obtenida al clasificar el conjunto de evaluación con los métodos de clasificación: k-Vecinos más cercanos (kNN), Bayesiano Simple (NB), Máquina de Vectores de Soporte (SVM) y un Clasificador basado en Prototipos (CBP) bajo el enfoque de clasificación supervisada

5.2.3. Resultados experimentales y discusión

Después de realizar la serie de experimentos con los valores de λ y k mencionados anteriormente, la Tabla 5.4 muestra los valores de la medida-F alcanzados por el método propuesto en las tres colecciones reducidas, con los valores de λ que alcanzaron mejores resultados, $\lambda = 0.4$, $\lambda = 0.3$ y $\lambda = 0.2$ (las tablas de los resultados para todos los valores de λ se pueden ver en el Apéndice A.1). Los resultados en 'negrillas' indican que el método es significativamente mejor que los resultados obtenidos con CBP, clasificador utilizado como *baseline*. Los resultados marcados con $\star\star$ señalan el mejor resultado por cada colección. Para una mejor vista de los resultados alcanzados se presentan las gráficas de la Figura 5.1.

Los resultados obtenidos muestran que el método propuesto mejoran la clasificación de las tres colecciones, especialmente en la que tiene el conjunto de entrenamiento más pequeño. Por ejemplo, para la colección *R8-reducida-10* el mejor resultado mejora en 10.5 % al resultado del *baseline* de esa colección (ver Tabla 5.5). Nótese también que el método no es sensible a los valores de k y λ, alcanzando, de manera general, el mejor resultado con $k < 10$ y $\lambda = 0.2$.

Con el objetivo de resumir la información de los resultados obtenidos en este escenario, la Tabla 5.5 presenta el mejor *baseline* (para este escenario corresponde al clasificador basado en prototipos -CBP-), el mejor resultado alcanzado por el método propuesto con su correspondiente configuración de k y λ, así como el resultado de la clasificación del método utilizando una configuración fija para los parámetros k y λ.

De los resultados anteriores se concluye lo siguiente:

	R8-reducido-41 (*baseline*=0.836)			R8-reducido-20 (*baseline*=0.803)			R8-reducido-10 (*baseline*=0.765)		
k	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$
1	0.843	0.841	**0.816**	**0.827**	0.817	0.797	**0.811**	**0.794**	0.761
2	**0.869**	**0.865**	0.847	**0.823**	**0.83**	0.815	**0.813**	**0.82**	**0.814**
3	**0.872**	**0.872**	**0.859**	**0.84**	**0.844**	**0.841**	**0.821**	**0.83**	**0.823**
4	**0.867**	**0.877**⋆⋆	**0.868**	**0.843**	**0.841**	**0.85**	**0.83**	**0.833**	**0.834**
5	**0.862**	**0.871**	**0.861**	**0.86**⋆⋆	**0.84**	**0.84**	**0.829**	**0.834**	**0.842**
6	**0.869**	**0.873**	**0.865**	**0.854**	**0.855**	**0.832**	**0.829**	**0.833**	**0.839**
7	**0.866**	**0.863**	**0.859**	**0.858**	**0.853**	**0.844**	**0.835**	**0.835**	**0.834**
8	**0.871**	**0.861**	**0.859**	**0.856**	**0.858**	**0.838**	**0.836**	**0.84**	**0.845**⋆⋆
9	**0.875**	**0.862**	**0.867**	**0.856**	**0.848**	**0.849**	**0.832**	**0.836**	**0.835**
10	**0.863**	**0.86**	0.85	**0.849**	**0.842**	**0.838**	**0.828**	**0.829**	**0.822**
11	**0.861**	**0.857**	0.85	**0.847**	**0.842**	**0.837**	**0.827**	**0.835**	**0.82**
12	0.848	**0.854**	0.842	**0.842**	**0.831**	**0.831**	**0.828**	**0.828**	**0.821**
13	**0.853**	**0.855**	0.844	**0.843**	**0.83**	**0.823**	**0.818**	**0.828**	**0.82**
14	**0.853**	**0.856**	0.83	**0.844**	**0.836**	**0.809**	**0.816**	**0.825**	**0.822**
15	0.848	**0.855**	0.829	**0.833**	**0.827**	0.808	**0.817**	**0.815**	**0.82**
16	0.85	**0.854**	0.83	**0.833**	**0.827**	0.809	**0.817**	**0.812**	**0.805**
17	**0.853**	**0.854**	0.825	**0.832**	**0.835**	0.81	**0.81**	**0.813**	**0.805**
18	0.85	**0.854**	0.826	**0.827**	**0.838**	0.805	**0.808**	**0.813**	**0.805**
19	0.849	**0.854**	0.826	**0.828**	**0.836**	0.808	**0.808**	**0.811**	**0.801**
20	0.847	**0.853**	0.826	**0.829**	**0.834**	0.812	**0.81**	**0.808**	**0.802**
21	0.846	**0.852**	0.819	**0.829**	**0.834**	0.82	**0.812**	**0.809**	**0.796**
22	0.849	**0.852**	0.818	**0.829**	**0.833**	0.809	**0.812**	**0.808**	**0.799**
23	**0.857**	**0.852**	0.818	**0.829**	**0.832**	**0.828**	**0.815**	**0.808**	**0.801**
24	**0.856**	0.85	0.819	0.818	**0.832**	**0.828**	**0.812**	**0.808**	**0.797**
25	**0.857**	**0.852**	0.823	0.819	**0.832**	**0.821**	**0.814**	**0.808**	**0.793**
26	**0.854**	**0.852**	0.824	0.819	**0.821**	0.819	**0.812**	**0.806**	**0.794**
27	**0.854**	0.848	0.834	0.817	**0.827**	**0.821**	**0.81**	**0.807**	**0.796**
28	**0.854**	0.845	0.838	0.816	0.818	0.82	**0.81**	**0.806**	**0.791**
29	**0.852**	0.845	0.836	0.815	0.818	0.82	**0.809**	**0.808**	**0.791**
30	**0.854**	0.846	0.834	0.814	0.818	**0.822**	**0.808**	**0.817**	**0.791**
Promedio	0.857 ±0.009	0.856 ±0.009	0.838 ±0.017	0.834 ±0.014	0.834 ±0.011	0.823 ±0.014	0.818 ±0.009	0.819 ±0.012	0.811 ±0.019

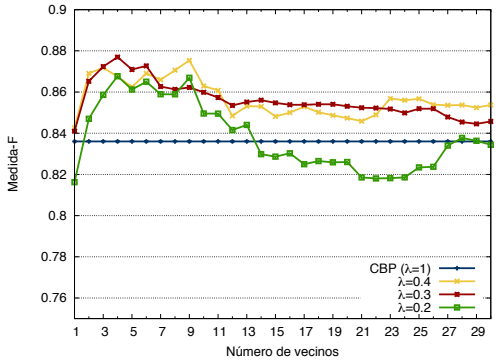
Tabla 5.4: Colección R8-reducidas: resultados de la medida-F del método propuesto sobre tres colecciones de datos con conjuntos de entrenamiento pequeños. Los mejores resultados por colección están indicados por ⋆⋆. El *baseline* presentado corresponde al clasificador basado en prototipos (CBP)

Colección	*Baseline* (CBP)	Mejor resultado $[k, \lambda]$		Configuración $[k = 4, \lambda = 0.2]$
R8-reducida-41	0.836	**0.877** +4.9 %	[4, 0.3]	0.868
R8-reducida-20	0.803	**0.860** +7.1 %	[5, 0.4]	0.850
R8-reducida-10	0.765	**0.845** +10.5 %	[8, 0.2]	0.834

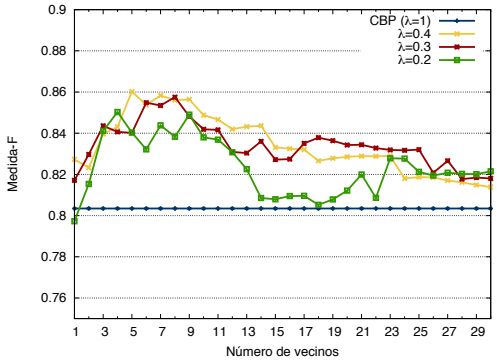
Tabla 5.5: Colección R8-reducida: resumen de los mejores resultados obtenidos en el escenario de insuficiencia de documentos etiquetados. La tercera columna muestra el mejor resultado obtenido, la configuración con la cual se obtuvo y el porcentaje de mejora en relación al *baseline*

- Se requiere un número pequeño de vecinos para alcanzar el mejor resultado en las tres colecciones. Como se esperaba, entre menor sea el número de ejemplos en el conjunto de entrenamiento, es mayor el número de vecinos (k) necesarios para alcanzar el valor de desempeño máximo.

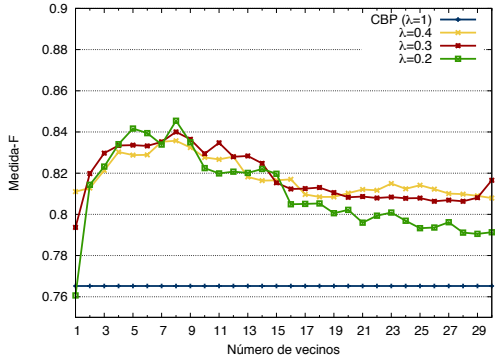
- Entre menor sea el número de ejemplos en el conjunto de entrenamiento, mayor es la mejora alcanzada por el método propuesto en comparación con el *baseline*. De aquí que se puede concluir que *el método es especialmente apropiado cuando el conjunto de entrenamiento es muy pequeño*, en otras palabras, cuando los



(a) R8-reducida-41: CBP=0.836



(b) R8-reducida-20: CBP=0.803



(c) R8-reducida-10: CBP=0.765

Figura 5.1: Colecciones R8-reducidas: gráficas de los resultados obtenidos en las tres colecciones de datos con conjuntos de entrenamiento pequeños, para tres valores de λ (0.4, 0.3 y 0.2) y para diferentes valores de k (k=1, ..., 30)

enfoques de clasificación actuales tienden a generar modelos de clasificación con bajo desempeño.

- En la mayoría de los casos, el mejor resultado es alcanzando al utilizar un $\lambda < 0.4$, lo cual indica que la información de los vecinos es útil, incluso más significativa que la información del propio documento a clasificar.

5.3. Escenario 2: Clasificación multi-lenguaje, utilización de ejemplos en otros idiomas

El segundo escenario de clasificación en que se evaluó el desempeño del método propuesto es cuando existen suficientes documentos etiquetados en el dominio D_T pero escritos en diferente lenguaje al de los documentos a clasificar; por lo tanto, se desea utilizar los documentos etiquetados disponibles en otro lenguaje. El corpus utilizado, resultados y análisis se presentan a continuación.

5.3.1. Corpus: RCV1

En 2000, Reuters Ltd. puso a disposición de la comunidad investigadora una colección más grande de noticas de la agencia Reuters. Este nuevo corpus es conocido como *Reuters Corpus, Volumen 1* o RCV1, éste contiene 810 mil noticias en inglés. La versión 2 de este corpus (llamada RCV2) contiene alrededor de 487 mil noticas en 13 diferentes idiomas, entre ellos, francés, inglés, y español (Lewis et al. (2004)).

Para esta tesis se utilizó un subconjunto de noticias en tres idiomas del grupo GCAT (*Government/Social*), de ella se seleccionaron los documentos que tuvieran sólo una etiqueta de clase. Las clases elegidas fueron: GCRM (*Crimes*), GDIS (*Disasters*), GPOL (*Politics*) y GSPO (*Sports*). Con lo anterior se formaron 3 colecciones, una por cada idioma. Las que se llamarán C-Español, C-Francés, y C-Inglés. Cada colección cuenta con 320 documentos, 80 por cada clase.

En las Tablas 5.6 y 5.7 se muestran algunos detalles de las clases de cada colección, específicamente se muestra información sobre el diccionario de las colecciones (C-Español, C-Francés y C-Inglés). En la primera se muestra el número de términos únicos de cada clase en cada una de las 3 colecciones, en el último renglón de esta tabla se muestra el total de términos únicos de cada colección. Nótese que el número de términos únicos por colección es menor que la suma de los términos únicos por clase. En la segunda tabla se muestra la cantidad de términos en el diccionario de cada colección al ser traducidos a los otros dos idiomas.

De la Tabla 5.7 es posible observar que mientras la colección C-Francés tiene un mayor número de términos únicos en el diccionario, al ser traducido en otro idioma (**español** e **inglés**) el número de términos se ve disminuido. Por otro lado, la colección C-Inglés tiene el menor número de términos en el diccionario, pero al ser traducido a

Clase	C-Español	C-Francés	C-Inglés
GCRM	6,523	7,369	5,206
GDIS	2,909	4,805	4,077
GPOL	7,285	8,003	6,217
GSPO	5,773	8,699	5,639
Términos únicos por colección	15,227	19,081	13,901

Tabla 5.6: Colecciones multi-lenguaje: información sobre el número de términos únicos por clase a través de cada una de las tres colección (C-Español, C-Francés y C-Inglés), además de los términos únicos por colección.

	Lenguaje		
Colección	**Español**	**Francés**	**Inglés**
C-Español	**15,227**	13,848	10,580
C-Francés	18,146	**19,081**	13,845
C-Inglés	16,899	18,237	**13,901**

Tabla 5.7: Colecciones multi-lenguaje: información sobre el tamaño de cada colección en tres lenguajes (español, francés e inglés). Se resaltan las cantidades que corresponden al idioma original de la colección

los otros lenguajes este número aumenta. Cabe mencionar que estas colecciones han sido utilizadas en otras investigaciones tales como las realizadas por Escobar-Acevedo et al. en (Escobar-Acevedo, Montes-Y-Gómez, y Villaseñor-Pineda (2009)).

5.3.2. Configuraciones particulares

Para implementar el método propuesto en este escenario, puesto que es necesario comparar documentos del conjunto objetivo en un lenguaje L_T con prototipos del conjunto de entrenamiento en un lenguaje L_F, se ha incorporado un proceso al modelo general del método propuesto de la Figura 4.1 tal como se muestra en la Figura 5.2.

El nuevo proceso, **traducción de documentos**, consiste en traducir cada documento del conjunto D_T al lenguaje del conjunto de D_S. Se ha decido traducir en ese sentido debido a que en un ambiente de clasificación real resulta menos costoso traducir una única ocasión el conjunto de documentos de entrenamiento que traducir los documentos a clasificar cada vez que nuevos conjuntos de documentos no etiquetados lleguen. Una vez traducidos, se continúa con el procedimiento descrito en el Capítulo 4.

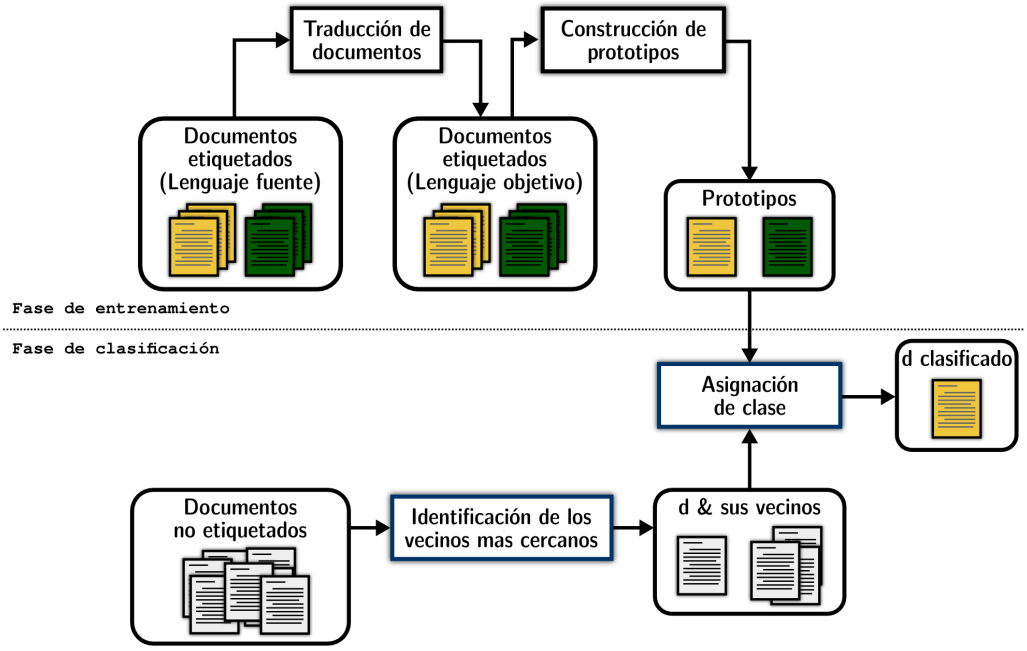


Figura 5.2: Esquema del método de clasificación de textos propuesto para utilizarlo en un enfoque de clasificación multi-lenguaje. Se agrega el proceso de traducción de documentos antes de la construcción de prototipos para representar a todos los documentos en el lenguaje de los documentos no etiquetados, lenguaje objetivo

5.3.3. Diseño de los experimentos

El objetivo de esta serie de experimentos es medir el desempeño del método en un enfoque *multi-lenguaje*. En este contexto se diseñaron 6 experimentos, mostrados en la Tabla 5.8, utilizando todas las combinaciones posibles con las tres colecciones.

Lenguaje del conjunto de entrenamiento (D_T)	**Lenguaje del conjunto objetivo (D_F)**	**Experimento** (D_T–D_F)
Francés	Español	$F_E - E$
Inglés	Español	$I_E - E$
Español	Francés	$E_F - F$
Inglés	Francés	$I_F - F$
Español	Inglés	$E_I - I$
Francés	Inglés	$F_I - I$

Tabla 5.8: Escenario 2: serie de experimentos realizados en el enfoque de clasificación multi-lenguaje. La notación X_Y indica que los documentos escritos originalmente en el idioma X fueron traducidos al idioma Y. E, F y I hacen referencia al idioma español, francés e inglés, respectivamente

5.3.4. Resultados de referencia

Se realizaron los 6 experimentos mostrados en la Tabla 5.8 utilizando los métodos base (kNN, NB, SVM y CBP), los resultados de la medida-F obtenida se muestran en la Tabla 5.9. Nótese que ninguno de los cuatro métodos de referencia es el mejor para todos los experimentos Como lo muestran, no es posible determinar el mejor método para los seis experimentos.

Experimento (D_T–D_F)	**kNN**	**NB**	**SVM**	**CBP**
$F_E - E$	0.292	**0.882**	0.658	0.879
$I_E - E$	0.100	0.791	0.625	**0.814**
$E_F - F$	0.100	**0.802**	0.723	0.790
$I_F - F$	0.107	0.753	**0.764**	0.616
$E_I - I$	0.100	**0.891**	0.486	0.851
$F_I - I$	0.268	0.931	0.616	**0.956**

Tabla 5.9: Colecciones multi-lenguaje: resultados de referencia utilizando la medida-F obtenida al clasificar el conjunto de evaluación con los métodos de clasificación: k-Vecinos más cercanos (kNN), Bayesiano Simple (NB), Máquina de Vectores de Soporte (SVM) y un Clasificador basado en Prototipos (CBP) bajo el enfoque de clasificación supervisada

5.3.5. Resultados experimentales y discusión

Los resultados de la medida-F obtenidos con los tres mejores valores de λ para clasificar las colecciones C-Español, C-Francés y C-Inglés se muestran en las Tablas 5.10, 5.11 y 5.12, respectivamente (los resultados para todos los valores de λ se muestran en el Apéndice A.2). En estas tablas, los valores en 'negrilla' corresponden a los resultados que son significativamente mejores al clasificador basado en prototipos (CBP) tomado como base en el método propuesto y que es utilizado como *baseline* en estas tablas.

Una mejor vista de los resultados obtenidos en los seis experimentos se muestra en las gráficas de la Figura 5.3. Además, con el objetivo de resumir la información mostrada en las tablas y gráficas anteriores se presenta la Tabla 5.13 que muestra en la segunda columna, dos *baselines*: el que corresponde a un método de clasificación tradicional basado en prototipos (CBP) y el mejor *baseline* considerando los cuatro métodos de clasificación presentados en la Tabla 5.9. La tercera columna muestra el mejor resultado obtenido con el método propuesto y la configuración con la cual

	Francés-Español (*baseline*=0.879)			Inglés-Español (*baseline*=0.814)		
k	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.893	0.897	0.894	0.824	0.822	0.818
2	0.906	0.912	0.906	0.852	0.857⋆⋆	0.851
3	0.905	0.906	0.909	0.846	0.854	0.837
4	0.909	0.912	0.909	0.839	0.82	0.821
5	0.913	0.912	0.909	0.832	0.818	0.804
6	0.903	0.906	0.909	0.843	0.83	0.833
7	0.912	0.906	0.91	0.845	0.832	0.818
8	0.915	0.906	0.913	0.854	0.83	0.824
9	**0.919**	0.916	0.906	0.849	0.841	0.83
10	**0.919**	**0.919**	0.91	0.85	0.837	0.821
11	**0.922**⋆⋆	0.91	0.904	0.85	0.837	0.817
12	**0.919**	0.909	0.904	0.853	0.837	0.814
13	**0.919**	0.912	0.907	0.851	0.834	0.817
14	**0.919**	0.916	0.904	0.857	0.835	0.812
15	**0.919**	**0.919**	0.904	0.854	0.838	0.812
16	**0.919**	**0.919**	0.901	0.854	0.835	0.816
17	**0.919**	**0.919**	0.898	0.85	0.835	0.815
18	**0.919**	**0.919**	0.895	0.848	0.838	0.815
19	**0.922**	**0.922**	0.893	0.848	0.835	0.807
20	**0.919**	**0.922**	0.896	0.844	0.841	0.811
21	**0.922**	**0.922**	0.893	0.837	0.842	0.808
22	**0.922**	**0.922**	0.896	0.84	0.835	0.804
23	**0.922**	**0.922**	0.893	0.84	0.831	0.796
24	**0.922**	**0.919**	0.893	0.84	0.828	0.793
25	**0.922**	**0.922**	0.893	0.837	0.828	0.798
26	**0.922**	0.916	0.89	0.83	0.824	0.801
27	**0.922**	0.916	0.893	0.834	0.82	0.798
28	**0.922**	0.916	0.896	0.834	0.824	0.801
29	**0.922**	0.916	0.893	0.834	0.822	0.797
30	**0.922**	0.916	0.893	0.826	0.821	0.793
Promedio	0.917 ±0.007	0.915 ±0.006	0.900 ±0.007	0.843 ±0.009	0.833 ±0.009	0.813 ±0.014

Tabla 5.10: Colecciones multi-lenguaje: resultados de la medida-F del método propuesto sobre la colección C-Español, utilizando como lenguaje de entrenamiento, francés e inglés. Los resultados por experimento están indicados por ⋆⋆. El *baseline* es CBP

es alcanzado dicho resultado. La última columna contiene el resultado del método propuesto bajo una configuración fija para los seis experimentos.

Un análisis de los resultados obtenidos se lista a continuación:

- El método propuesto demuestra tener buen desempeño en todos los experimentos. De la Tabla 5.13 es posible notar que la mejora en la clasificación del método propuesto en comparación con el enfoque tradicional de clasificación basado en prototipos (es decir, cuando $\lambda = 1$) es significativamente superior en 5 de los 6 experimentos e incluso es significativamente mejor que el mejor *baseline* en tres de los seis casos.

  El problema con el sexto caso, clasificar documentos en inglés utilizando documentos escritos en francés, $I_F - F$ (último renglón de la Tabla 5.13), puede deberse a que de los seis experimentos, es el único en el que los resultados del *baseline* CBP supera el 95 % en el desempeño del clasificador. Es decir, el *baseline*, sin usar la información del conjunto a clasificar obtiene buen desempeño, provocando que el margen en el cual la información de los vecinos puede mejorar la clasificación, sea mínima.

	Español-Francés (*baseline*=0.790)			Inglés-Francés (*baseline*=0.616)		
k	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.793	0.788	0.753	0.589	0.577	0.559
2	0.805	0.812	0.795	0.611	0.61	0.614
3	0.826	0.814	0.81	0.628	0.623	0.617
4	0.817	0.831⋆⋆	0.815	0.629	0.642	0.637
5	0.817	0.829	0.8	0.625	0.636	0.629
6	0.814	0.812	0.787	0.623	0.641	0.638
7	0.802	0.807	0.784	0.645	0.655	0.661
8	0.799	0.81	0.783	0.647	0.66	**0.682⋆⋆**
9	0.792	0.815	0.793	0.641	0.655	0.668
10	0.802	0.816	0.806	0.638	0.666	**0.68**
11	0.81	0.82	0.806	0.641	0.661	**0.672**
12	0.803	0.817	0.797	0.633	0.664	0.67
13	0.799	0.81	0.803	0.632	0.656	0.667
14	0.808	0.82	0.817	0.634	0.656	**0.675**
15	0.805	0.823	0.81	0.631	0.656	**0.677**
16	0.804	0.823	0.806	0.627	0.656	**0.68**
17	0.805	0.823	0.803	0.63	0.648	**0.674**
18	0.807	0.813	0.802	0.63	0.648	**0.677**
19	0.798	0.81	0.802	0.617	0.648	**0.674**
20	0.795	0.804	0.797	0.617	0.641	0.667
21	0.795	0.807	0.789	0.614	0.636	0.662
22	0.791	0.804	0.782	0.608	0.63	0.644
23	0.788	0.787	0.777	0.616	0.628	0.644
24	0.788	0.787	0.774	0.616	0.625	0.636
25	0.788	0.786	0.77	0.613	0.625	0.625
26	0.788	0.787	0.766	0.619	0.618	0.617
27	0.788	0.786	0.772	0.613	0.613	0.617
28	0.788	0.786	0.769	0.608	0.613	0.623
29	0.784	0.784	0.761	0.608	0.61	0.609
30	0.784	0.784	0.759	0.608	0.61	0.611
Promedio	0.799 ±0.011	0.807 ±0.015	0.790 ±0.018	0.623 ±0.013	0.637 ±0.0213	0.647 ±0.030

Tabla 5.11: Colecciones multi-lenguaje: resultados de la medida-F del método propuesto sobre la colección C-Francés, utilizando como lenguaje de entrenamiento, español e inglés. Los mejores resultados por experimento están indicados por ⋆⋆. El *baseline* es CBP

- Utilizar la información de los vecinos en este escenario de clasificación es muy importante, los mejores resultados se produjeron con valores de λ cercanos a cero dándole más importancia al factor que incluye a los vecinos en la Fórmula 4.4.2; así mismo, en dos casos, los mejores resultados se alcanzaron con $\lambda = 0.0$ (para los experimentos $I_F - F$ y $E_I - I$), lo que indica que no se tomó en cuenta la información del propio documento a clasificar en la asignación de su clase.

A fin de entender con mayor detalle el desempeño del método en este escenario, en la Figura 5.4 se muestran las matrices de similitud entre el conjunto de documentos de cada idioma. Una matriz de similitud se obtiene al calcular la similitud entre cada par de documentos de un conjunto, entre mayor sea la similitud el punto que la represente deberá ser mas oscuro; los documentos representados en la matriz están organizados por clases. En estas matrices es posible ver la cohesión interna de cada clase ayudando a tener un panorama de la complejidad de la clasificación de un conjunto de documentos.

De las matrices de similitud de la Figura 5.4, la colección que tiene clases más cohesionadas es C-Inglés, mientras que las colecciones C-Español y C-Francés tienen

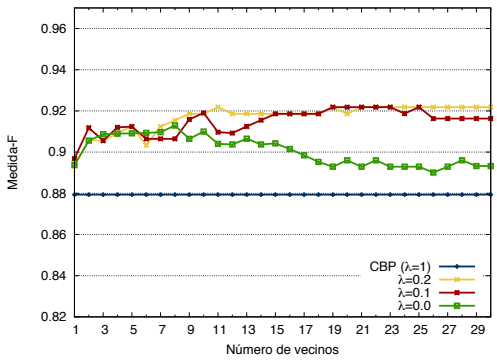
	Español-Inglés (*baseline*=0.851)			Francés-Inglés (*baseline*=0.956)		
k	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.883	0.861	0.851	0.953	0.944	0.925
2	**0.899**	0.883	0.877	0.963	0.95	0.938
3	**0.905**	**0.895**	0.89	0.963	0.959	0.935
4	**0.924**	**0.915**	**0.915**	0.956	0.956	0.935
5	**0.92**	**0.92**	**0.917**	0.963	0.966	0.947
6	**0.927**	**0.92**	**0.921**	0.966	0.969	0.95
7	**0.927**	**0.93**	**0.931**	0.963	0.966	0.95
8	**0.924**	**0.93**	**0.928**	0.969	0.969	0.95
9	**0.924**	**0.934**	**0.934**	0.966	0.969	0.95
10	**0.927**	**0.94**	**0.94**	0.969⋆⋆	0.969	0.96
11	**0.93**	**0.94**	**0.944**	0.966	0.966	0.954
12	**0.927**	**0.94**	**0.934**	0.966	0.966	0.954
13	**0.927**	**0.937**	**0.943**	0.966	0.963	0.951
14	**0.93**	**0.937**	**0.944**	0.963	0.96	0.951
15	**0.927**	**0.937**	**0.94**	0.966	0.963	0.954
16	**0.93**	**0.94**	**0.947**	0.966	0.963	0.954
17	**0.927**	**0.937**	**0.95**⋆⋆	0.963	0.963	0.951
18	**0.93**	**0.934**	**0.947**	0.966	0.966	0.96
19	**0.93**	**0.934**	**0.947**	0.966	0.966	0.957
20	**0.927**	**0.934**	**0.944**	0.966	0.966	0.954
21	**0.924**	**0.937**	**0.941**	0.966	0.966	0.957
22	**0.927**	**0.937**	**0.944**	0.966	0.966	0.957
23	**0.924**	**0.94**	**0.941**	0.969	0.966	0.957
24	**0.924**	**0.94**	**0.941**	0.966	0.966	0.954
25	**0.924**	**0.934**	**0.937**	0.966	0.966	0.957
26	**0.924**	**0.934**	**0.937**	0.969	0.966	0.957
27	**0.924**	**0.937**	**0.937**	0.969	0.966	0.963
28	**0.924**	**0.939**	**0.94**	0.966	0.966	0.96
29	**0.924**	**0.937**	**0.937**	0.969	0.963	0.96
30	**0.924**	**0.937**	**0.937**	0.969	0.966	0.96
Promedio	**0.923** ±0.010	**0.929** ±0.018	**0.931** ±0.022	0.965 ±0.004	0.964 ±0.005	0.952 ±0.009

Tabla 5.12: Colecciones multi-lenguaje: resultados de la medida-F del método propuesto sobre la colección C-Inglés, utilizando como lenguaje de entrenamiento, español e francés. Los mejores resultados por experimento están indicados por ⋆⋆. El *baseline* es CBP

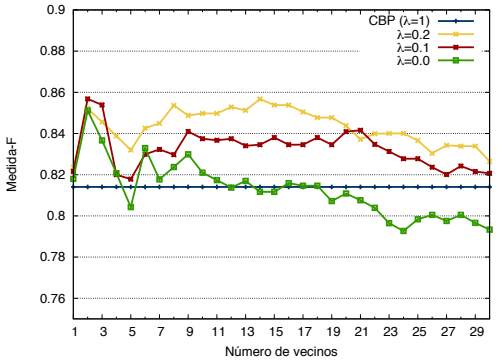
Experimentos	***Baselines***		**Mejor resultado**			**Configuración**
	CBP	Mejor	$[k,\ \lambda]$			$[k = 11,\ \lambda = 0.1]$
$F_E - E$	0.879	0.882	0.922 ⋆†	+4.9 %	[11, 0.2]	0.910
$I_E - E$	0.814	0.814	0.857 ⋆†	+5.3 %	[2, 0.1]	0.837
$E_F - F$	0.790	0.802	0.831 ⋆	+5.2 %	[4, 0.1]	0.820
$I_F - F$	0.616	0.764	0.682 ⋆	+10.7 %	[8, 0.0]	0.661
$E_I - I$	0.851	0.891	0.950 ⋆†	+11.6 %	[17, 0.0]	0.940
$F_I - I$	0.956	0.956	0.969	+1.4 %	[10, 0.2]	0.966

Tabla 5.13: Escenario 2: resumen de los mejores resultados obtenidos al realizar los seis experimentos en un enfoque multi-lenguaje. La tercera columna muestra el mejor resultado obtenido, la configuración con la cual se obtuvo y el porcentaje de mejora en relación al *baseline* CBP. ⋆ indica que el resultado es significativamente mejor que el *baseline* CBP, mientras que † muestra una mejora significativa en relación al mejor *baseline*

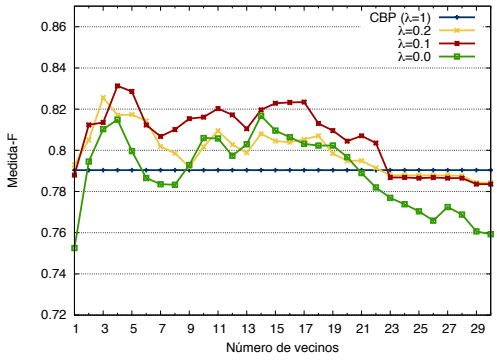
una pobre cohesión entre los documentos con la misma etiqueta de clase. Considerando que el método propuesto se basa en las similitudes entre documentos, al existir clases poco cohesionadas (los documentos de estas clases no se parecen mucho entre si) el desempeño del método propuesto se verá afectado. Con base en la observación de las matrices de similitud y los resultados de la clasificación de esos mismos documentos,



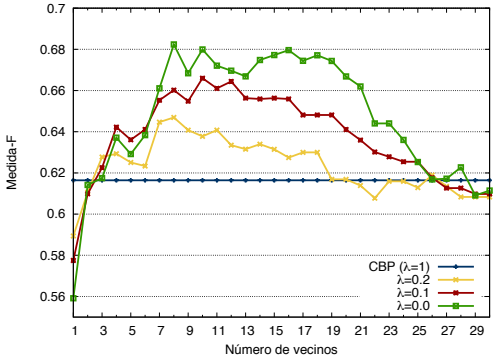
(a) Experimento $F_E - E$: CBP=0.879



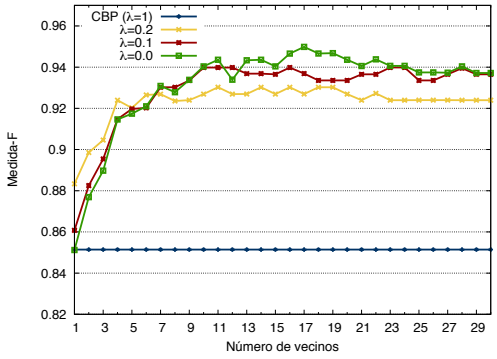
(b) Experimento $I_E - E$: CBP=0.814



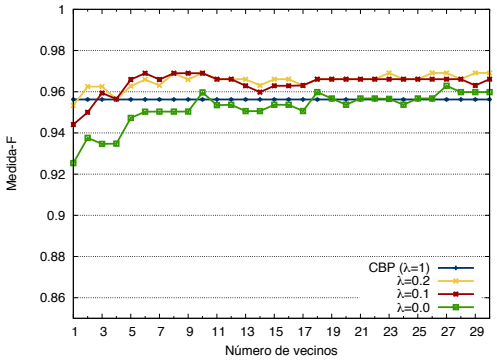
(c) Experimento $E_F - F$: CBP=0.790



(d) Experimento $I_F - F$: CBP=0.616

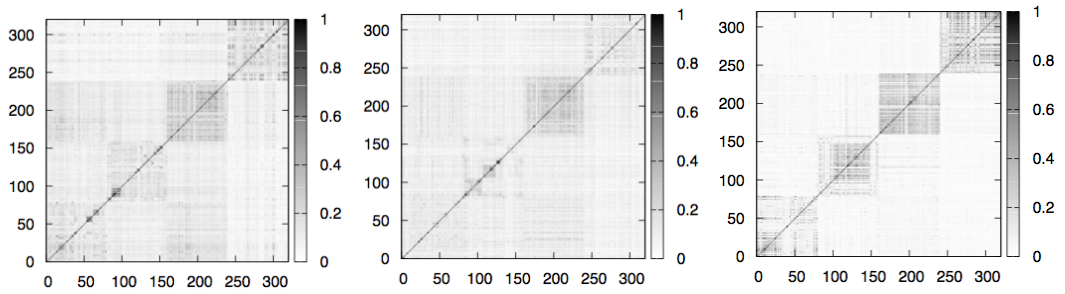


(e) Experimento $E_I - I$: CBP=0.851



(f) Experimento $F_I - I$: CBP=0.956

Figura 5.3: Escenario 2: gráficas de los resultados obtenidos en los experimentos multi-lenguaje para tres valores de λ (0.2, 0.1 y 0.0) y para diferentes valores de k (k=1, ..., 30)



(a) Español (b) Francés (c) Inglés

Figura 5.4: Escenario 2: matrices de similitud de los conjuntos de documentos C-Español, C-Francés y C-Inglés. Los documentos con la misma etiqueta de clase se encuentran consecutivos en las matrices. Los puntos en las gráficas indican la similitud entre cada par de documentos

es razonable concluir que entre mayor cohesión exista en las clases de una colección, mejor será su clasificación con el método propuesto.

5.4. Escenario 3: Clasificación multi-dominio, utilización de ejemplos de dominios similares

En el tercer escenario de clasificación no existen conjuntos de documentos etiquetados en el dominio específico en donde se desean clasificar documentos, en su lugar existen suficientes documentos etiquetados de *dominios similares*, con los cuales se desea entrenar un clasificador.

Dado un conjunto de documentos a clasificar X_T que pertenecen a un dominio específico D_T, y un conjunto de documentos etiquetados X_F que pertenece a un dominio similar D_F, se usa X_F para entrenar un modelo que clasifique los documentos en X_T. En esta tesis, dos dominios son similares si los documentos de ambos pueden clasificarse dentro de las mismas clases.

En las secciones siguientes se describirá el corpus utilizado, detalles de implementación, resultados obtenidos y análisis de los mismos.

5.4.1. Corpus: *Multi-Domain Sentiment Dataset* V2.0

Este corpus consiste de comentarios sobre 25 tipos de productos (o dominios) tomados de Amazon.com. Los comentarios pueden ser: a) opiniones positivas sobre el producto, b) opiniones negativas sobre el producto y c) opiniones sin clasificación

(sólo para algunos productos) (Blitzer, Dredze, y Pereira (2007)).

En esta tesis se utilizó un subconjunto de este corpus. Sólo se consideraron comentarios positivos y negativos de 4 tipos de productos: *Dvd* (D), *Electronics* (E), *Kitchen & Housewares* (K), y *Books* (B). Cada dominio (tipo de producto) contiene 1000 documentos de comentarios positivos y 1000 documentos de comentarios negativos. Esta base de datos fue construida por Blitzer et al. y ha sido utilizada en Blitzer, Dredze, y Pereira (2007) y en Ben-David et al. (2010).

En la Tabla 5.14 se muestra el número de términos en el diccionario de cada uno de los dominios que serán utilizados para evaluar, en este escenario, el método de clasificación propuesto.

Dominio	**Términos en el diccionario**
Books	6337
Dvd	6067
Electronics	4319
Kitchen	3765

Tabla 5.14: Colecciones multi-dominios: número de términos en el diccionario de cada uno de los 4 dominios usados en la clasificación multi-dominio

5.4.2. Diseño de los experimentos

Para evaluar el desempeño del método propuesto en este escenario se han diseñado una serie de experimentos utilizando los cuatro dominios disponibles. Los doce experimentos consisten en utilizar todos los documentos de un dominio como entrenamiento y todos los documentos de otro dominio como documentos a clasificar. En la Tabla 5.15 se muestran los 12 experimentos a llevar a cabo usando la combinación de los cuatro dominios.

5.4.3. Pre-procesamiento

Como se ha descrito, el corpus utilizado para evaluar el método en este escenario consiste de documentos de opiniones, esta característica hace que la clasificación no consista, como en los escenarios anteriores, en determinar la clase temática de los documentos; en este caso la tarea de clasificación consiste en determinar si una opinión

Dominio usado como entrenamiento (X_T)	**Dominio usado como objetivo (X_F)**	**Experimento (X_T–X_F)**
Dvd	Books	$D - B$
Electronics	Books	$E - B$
Kitchen	Books	$K - B$
Books	Dvd	$B - D$
Electronics	Dvd	$E - D$
Kitchen	Dvd	$K - D$
Books	Electronics	$B - E$
Dvd	Electronics	$D - E$
Kitchen	Electronics	$K - E$
Books	Kitchen	$B - K$
Dvd	Kitchen	$D - K$
Electronics	Kitchen	$E - K$

Tabla 5.15: Escenario 3: serie de experimentos realizados en el enfoque de clasificación multi-lenguaje. La notación $Y - Z$ indica que se usarán los documentos del dominio Y como entrenamiento y los del dominio Z como conjunto a clasificar. B, D, E, K corresponden a los dominios *Book*, *Dvd*, *Electronics* y *Kitchen & Housewares*, respectivamente

es positiva o negativa donde el tema que se aborda en cada uno de esos documentos sea el mismo en las dos clases. Por ejemplo suponga que se tienen dos documentos, ambos comentando un libro de la trilogía de *La Fundación* de Issac Asimov, el primero hace una crítica negativa del libro, mientras que el segundo expresa comentarios positivos. En la clasificación temática, ambos documentos deberían pertenecer a la misma clase, digamos, **ciencia ficción**; sin embargo, en este tipo de clasificación deberán pertenecer a clases distintas, el primero a la clase de **opiniones negativas** y el segundo a la clase de **opiniones positivas**.

Por lo anterior, es necesario hacer algunos ajustes en el pre-procesamiento de los documentos con el fin de tratar de aminorar el impacto de este otro tipo de clasificación, particularmente, se pretende eliminar los atributos que puedan considerarse como *temáticos*. Si τ denota un conjunto de términos, para construir los prototipos de cada clase, en lugar de considerar sólo los atributos del dominio utilizando como entrenamiento τ_{D_T}, se utilizaron los atributos resultado de $\tau_{D_T} \cap \tau_{D_F}$, donde D_F es el conjunto de entrenamiento del dominio a clasificar, mientras que D_T el conjunto de documentos etiquetados. Con esta intersección se pretende eliminar atributos temáticos, de modo que sólo se conserven los atributos comunes en ambos dominios.

En este mismo sentido, se han utilizado el conjunto de atributos extraídos por

Blitzer et al. en (Blitzer, Dredze, y Pereira (2007)) en lugar de utilizar las palabras extraídas de cada dominio tal como se ha hecho en los escenarios anteriores.

5.4.4. Resultados de referencia

Para tener resultados de referencia que ayuden a medir el desempeño del método propuesto, se utilizaron los cuatro métodos de clasificación descritos en 2.3. Los resultados se muestran en la Tabla 5.16, en ella se nota que la diferencia entre el desempeño de los cuatro clasificadores es mínima y no representa una mejora significativa de unos sobre los otros.

Experimento (X_T–X_F)	**kNN**	**NB**	**SVM**	**CBP**
$D-B$	0.517	**0.748**	0.730	0.742
$E-B$	0.434	**0.676**	0.662	0.672
$K-B$	0.571	0.689	**0.695**	0.683
$B-D$	0.579	0.712	**0.775**	0.732
$E-D$	0.455	**0.685**	0.677	0.673
$K-D$	0.537	**0.702**	0.701	0.684
$B-E$	0.537	0.685	**0.710**	0.669
$D-E$	0.504	**0.721**	0.700	0.697
$K-E$	0.554	0.775	**0.795**	0.780
$B-K$	0.559	0.685	**0.737**	0.729
$D-K$	0.483	**0.753**	0.731	0.743
$E-K$	0.441	0.802	**0.811**	0.804

Tabla 5.16: Colecciones multi-dominio: resultados de referencia utilizando la medida-F obtenida al clasificar el conjunto de evaluación con los métodos de clasificación: k-Vecinos más cercanos (kNN), Bayesiano Simple (NB), Máquina de Vectores de Soporte (SVM) y un Clasificador basado en Prototipos (CBP) bajo el enfoque de clasificación supervisada. $X_T - X_F$ indica que se usarán los documentos del conjunto X_T como entrenamiento y los documentos del conjunto X_F como conjunto de evaluación. B, D, E, K corresponden a los dominios *Book*, *Dvd*, *Electronics* y *Kitchen & Housewares*, respectivamente

5.4.5. Resultados experimentales y discusión

Las Tablas 5.18, 5.19, 5.20 y 5.21 muestran los valores de la medida-F alcanzados por el método propuesto sobre los 12 experimentos multi-dominio, utilizando los mejores valores de λ (que varían dependiendo del dominio de los documentos a clasificar). Los resultados completos, con todos los valores de λ se muestran en el Apéndice A.3.

Para resumir y tener una mejor vista de los resultados alcanzados se muestran las gráficas de estos resultados en las Figuras 5.6 y 5.7, y la Tabla 5.17. En la tabla se presentan los mejores resultados obtenidos en los 12 experimentos; en la segunda columna se muestran dos *baselines*: el que corresponde a un método de clasificación tradicional basado en prototipos (CBP) y el mejor *baseline* considerando los cuatro métodos de clasificación presentados en la Tabla 5.16. La tercera columna muestra el mejor resultado obtenido con el método propuesto y la configuración con la cual es alcanzado dicho resultado.

Experimentos	***Baselines***		**Mejor resultado**		
	CBP	Mejor		$[k, \lambda]$	
$D-B$	0.742	0.748	0.745	+0.4 %	[8, 0.6]
$E-B$	0.672	0.676	0.690 $\star$	+2.7 %	[19, 0.2]
$K-B$	0.683	0.695	0.686	+0.4 %	[1, 0.6]
$B-D$	0.732	0.775	0.734	+0.3 %	[5, 0.4]
$E-D$	0.673	0.685	0.683 $\star$	+1.5 %	[3, 0.4]
$K-D$	0.684	0.702	0.687	+0.4 %	[6, 0.6]
$B-E$	0.669	0.710	0.677	+1.2 %	[1, 0.6]
$D-E$	0.687	0.721	0.713 $\star$	+2.3 %	[4, 0.2]
$K-E$	0.780	0.795	0.791	+1.4 %	[8, 0.2]
$B-K$	0.729	0.737	0.747 $\star$	+2.5 %	[1, 0.5]
$D-K$	0.743	0.753	0.759 $\star$	+2.2 %	[3, 0.3]
$E-K$	0.804	0.811	0.825 $\star$	+2.6 %	[11, 0.2]

Tabla 5.17: Escenario 3: resumen de los mejores resultados obtenidos al realizar los seis experimentos en un enfoque multi-dominio. La tercera columna muestra el mejor resultado obtenido, la configuración con la cual se obtuvo y el porcentaje de mejora en relación al *baseline* CBP. $\star$ indica que el resultado es mejor que el mejor *baseline*

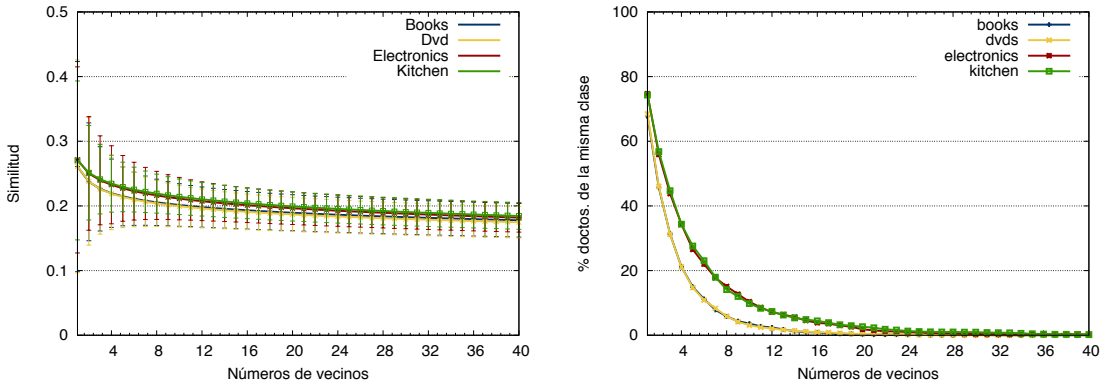
De todo lo anterior es posible concluir que dado que el método propuesto no obtuvo resultados significativamente mayores a los del *baseline* (CBP) los vecinos no aportan información que favorezca la correcta clasificación de cada documento. Con el fin de analizar con más detalle cada colección se presentan las gráficas de la Figura 5.5.

En la Figura 5.5(a) se muestra una gráfica en la cual se pretende ver qué tan cerca están los vecinos, considerados en cada experimento, del documento que se intenta clasificar. Cada punto de dicha gráfica es calculado usando la Ecuación 5.4.1.

$$\frac{\sum_{j}^{|X_F|} \frac{\sum_{i}^{k} sim(d_j, n_i)}{k}}{|X_F|} \tag{5.4.1}$$

donde $d_j \in X_T$ y $n_i \in N_k^{d_j}$.

La gráfica de la Figura 5.5(b) muestra el porcentaje, en promedio, de vecinos que pertenecen a la misma clase de cada documento de las cuatro colecciones, donde cada línea en la gráfica representa a un dominio. Cabe mencionar que para realizar esta gráfica se consideró la etiqueta de clase de cada uno de los documentos en el conjunto a clasificar, sin embargo, en el proceso de clasificación esta información nunca es vista.



(a) Similitudes entre documentos de cada dominio a sus vecinos más cercanos

(b) Porcentaje vecinos que pertenecen a la misma clase del documento a clasificar

Figura 5.5: Escenario 3: gráficas de similitud entre vecinos considerados por el método y porcentaje de vecinos con la misma clase que el documentos a clasificar

Algunos puntos a considerar de las observaciones anteriores se listan a continuación:

- Los cuatro dominios son muy parecidos en cuanto a la similitud de cada documento a sus vecinos más cercanos y el porcentaje de vecinos que comparten la misma etiqueta de clase que el documento a clasificar.
- La similitud de los vecinos con el documento a clasificar, en promedio, es muy pequeña (< 0.3, donde el máximo valor de similitud es 1). Esto es indicador de dos cosas, por un lado, que los atributos utilizados para la representación de estos documentos en particular no sean los adecuados y por otro lado, que la función de similitud utilizada no tome ventaja de los atributos que se utilizaron en la representación. Esto puede ser la causa de que la información generada por los vecinos no tenga tanta influencia en la decisión de clasificación.

- En general, después de 8 vecinos, la similitud es constante para los cuatro dominios y es aproximadamente de 0.2. Esto causa que considerar 10 o 20 vecinos, la decisión no clasificación no se verá afectada.

- Al ver el comportamiento de la gráfica de la Figura 5.5(b) es claro que, considerar la información de los vecinos, aún si los promedios de similitud fueran más altos, no ayudaría a la clasificación pues el porcentaje de vecinos con la misma etiqueta de clase que el documento a clasificar (recordar que esta es una de las hipótesis iniciales) decrece exponencialmente al considerar un mayor número de vecinos. Note que si sólo se consideraran dos vecinos, uno pertenecería a la clase del documento a clasificar y el otro no.

	D - B (*baseline*=0.742)			**E - B (*baseline*=0.672)**			**K - B (*baseline*=0.683)**		
k	$\lambda = 0.6$	$\lambda = 0.4$	$\lambda = 0.2$	$\lambda = 0.6$	$\lambda = 0.4$	$\lambda = 0.2$	$\lambda = 0.6$	$\lambda = 0.4$	$\lambda = 0.2$
1	0.741	0.741	0.721	0.67	0.67	0.673	0.686⋆⋆	0.681	0.668
2	0.738	0.739	0.729	0.675	0.675	0.678	0.683	0.682	0.678
3	0.739	0.738	0.728	0.674	0.678	0.68	0.686	0.686	0.683
4	0.741	0.742	0.733	0.675	0.678	0.686	0.68	0.68	0.68
5	0.743	0.741	0.732	0.676	0.675	0.681	0.683	0.683	0.685
6	0.744	0.74	0.734	0.676	0.677	0.68	0.68	0.683	0.678
7	0.742	0.741	0.733	0.677	0.678	0.682	0.681	0.684	0.68
8	0.745⋆⋆	0.741	0.733	0.674	0.676	0.68	0.681	0.688	0.683
9	0.743	0.743	0.736	0.675	0.679	0.685	0.682	0.684	0.686
10	0.744	0.743	0.733	0.675	0.679	0.682	0.682	0.684	0.685
11	0.742	0.738	0.733	0.675	0.675	0.68	0.683	0.683	0.681
12	0.744	0.74	0.734	0.675	0.676	0.683	0.684	0.681	0.68
13	0.743	0.739	0.733	0.677	0.677	0.683	0.682	0.684	0.682
14	0.742	0.739	0.734	0.678	0.678	0.685	0.681	0.682	0.681
15	0.742	0.739	0.735	0.677	0.678	0.684	0.681	0.686	0.679
16	0.742	0.74	0.734	0.676	0.679	0.687	0.683	0.686	0.681
17	0.741	0.738	0.731	0.675	0.676	0.686	0.681	0.686	0.68
18	0.741	0.739	0.735	0.676	0.674	0.683	0.682	0.686	0.679
19	0.741	0.739	0.733	0.676	0.676	0.69⋆⋆	0.682	0.686	0.675
20	0.742	0.738	0.734	0.677	0.675	0.687	0.683	0.684	0.673
21	0.743	0.736	0.732	0.677	0.675	0.688	0.682	0.685	0.673
22	0.742	0.738	0.733	0.677	0.675	0.69	0.683	0.685	0.674
23	0.743	0.736	0.732	0.677	0.675	0.686	0.683	0.685	0.676
24	0.742	0.737	0.736	0.677	0.675	0.683	0.683	0.684	0.675
25	0.743	0.736	0.733	0.675	0.676	0.68	0.683	0.683	0.679
26	0.742	0.736	0.734	0.675	0.675	0.682	0.685	0.685	0.677
27	0.742	0.737	0.734	0.674	0.674	0.68	0.684	0.684	0.679
28	0.741	0.737	0.732	0.675	0.674	0.681	0.684	0.683	0.681
29	0.742	0.739	0.735	0.675	0.676	0.681	0.683	0.685	0.684
30	0.741	0.738	0.734	0.676	0.676	0.68	0.683	0.685	0.68
Promedio	0.742 ± 0.001	0.739 ± 0.002	0.733 ± 0.003	0.676 ± 0.002	0.676 ± 0.002	0.683 ± 0.004	0.683 ± 0.001	0.684 ± 0.002	0.679 ± 0.004

Tabla 5.18: Colecciones multi-dominio: resultados de la medida-F del método propuesto sobre el conjunto de documentos del dominio *books* (B), utilizando como entrenamiento los dominios: dvd (D), electronics (E) y kitchen (K). Los mejores resultados están indicados por ⋆⋆. El *baseline* es CBP

	B - D (*baseline*=0.732)			E - D (*baseline*=0.673)			K - D (*baseline*=0.684)		
k	$\lambda = 0.6$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.6$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.6$	$\lambda = 0.4$	$\lambda = 0.3$
1	0.733	0.729	0.721	0.677	0.673	0.666	0.681	0.681	0.682
2	0.732	0.73	0.718	0.674	0.678	0.67	0.679	0.681	0.681
3	0.731	0.73	0.719	0.673	0.682⋆⋆	0.681	0.681	0.682	0.684
4	0.732	0.726	0.724	0.675	0.681	0.678	0.684	0.683	0.681
5	0.731	0.734⋆⋆	0.723	0.677	0.679	0.679	0.687	0.685	0.682
6	0.731	0.732	0.726	0.674	0.681	0.681	0.687⋆⋆	0.682	0.682
7	0.729	0.733	0.728	0.674	0.681	0.68	0.686	0.683	0.681
8	0.728	0.728	0.723	0.675	0.678	0.68	0.682	0.683	0.68
9	0.733	0.729	0.723	0.676	0.678	0.681	0.682	0.684	0.684
10	0.73	0.73	0.724	0.674	0.675	0.679	0.683	0.684	0.684
11	0.729	0.73	0.723	0.673	0.675	0.679	0.683	0.683	0.682
12	0.729	0.729	0.726	0.674	0.676	0.675	0.684	0.684	0.68
13	0.727	0.727	0.727	0.674	0.675	0.673	0.682	0.684	0.682
14	0.727	0.729	0.728	0.672	0.673	0.67	0.683	0.683	0.681
15	0.726	0.728	0.728	0.673	0.672	0.67	0.682	0.683	0.683
16	0.727	0.727	0.727	0.672	0.671	0.669	0.682	0.684	0.682
17	0.728	0.727	0.729	0.672	0.671	0.673	0.682	0.683	0.68
18	0.727	0.725	0.727	0.672	0.671	0.675	0.681	0.684	0.682
19	0.728	0.726	0.727	0.672	0.674	0.676	0.681	0.681	0.682
20	0.729	0.725	0.727	0.672	0.672	0.675	0.68	0.681	0.682
21	0.728	0.726	0.725	0.673	0.674	0.673	0.68	0.681	0.682
22	0.728	0.726	0.724	0.673	0.672	0.675	0.68	0.681	0.681
23	0.729	0.725	0.725	0.673	0.674	0.674	0.68	0.681	0.681
24	0.727	0.725	0.725	0.672	0.674	0.674	0.68	0.681	0.679
25	0.728	0.724	0.724	0.672	0.673	0.676	0.681	0.681	0.681
26	0.729	0.725	0.725	0.672	0.672	0.674	0.681	0.681	0.681
27	0.729	0.726	0.725	0.672	0.672	0.673	0.681	0.68	0.681
28	0.728	0.727	0.726	0.673	0.672	0.673	0.681	0.679	0.68
29	0.729	0.727	0.727	0.672	0.672	0.671	0.681	0.68	0.681
30	0.728	0.727	0.727	0.673	0.673	0.672	0.681	0.68	0.68
Promedio	0.729 ± 0.002	0.728 ± 0.003	0.725 ± 0.003	0.673 ± 0.001	0.675 ± 0.003	0.675 ± 0.004	0.682 ± 0.002	0.682 ± 0.002	0.681 ± 0.001

Tabla 5.19: Colecciones multi-dominio: resultados de la medida-F del método propuesto sobre el conjunto de documentos del dominio *dvd* (D), utilizando como entrenamiento los dominios: books (B), electronics (E) y kitchen (K). Los mejores resultados están indicados por ⋆⋆. El *baseline* es CBP

	B - E (*baseline*=0.669)			D - E (*baseline*=0.687)			K - E (*baseline*=0.780)		
k	$\lambda = 0.6$	$\lambda = 0.4$	$\lambda = 0.2$	$\lambda = 0.6$	$\lambda = 0.4$	$\lambda = 0.2$	$\lambda = 0.6$	$\lambda = 0.4$	$\lambda = 0.2$
1	0.677⋆⋆	0.672	0.662	0.708	0.697	0.694	0.785	0.786	0.766
2	0.671	0.664	0.648	0.703	0.702	0.694	0.783	0.783	0.766
3	0.666	0.672	0.666	0.703	0.709	0.7	0.783	0.779	0.781
4	0.666	0.665	0.662	0.706	0.709	0.713⋆⋆	0.78	0.788	0.786
5	0.668	0.668	0.662	0.704	0.707	0.706	0.782	0.786	0.788
6	0.669	0.667	0.66	0.705	0.708	0.708	0.78	0.786	0.788
7	0.667	0.666	0.659	0.706	0.707	0.704	0.782	0.783	0.789
8	0.669	0.663	0.656	0.705	0.707	0.701	0.782	0.785	0.792⋆⋆
9	0.667	0.667	0.657	0.706	0.706	0.701	0.783	0.782	0.783
10	0.667	0.665	0.66	0.706	0.706	0.694	0.781	0.782	0.786
11	0.667	0.665	0.661	0.704	0.706	0.696	0.78	0.782	0.784
12	0.667	0.667	0.662	0.705	0.708	0.694	0.779	0.781	0.777
13	0.667	0.666	0.66	0.704	0.708	0.691	0.782	0.78	0.784
14	0.666	0.665	0.656	0.703	0.707	0.69	0.78	0.781	0.78
15	0.663	0.662	0.657	0.703	0.707	0.693	0.779	0.781	0.78
16	0.664	0.662	0.657	0.7	0.705	0.694	0.78	0.78	0.779
17	0.663	0.663	0.654	0.7	0.705	0.695	0.778	0.779	0.777
18	0.663	0.662	0.657	0.702	0.706	0.698	0.777	0.779	0.78
19	0.661	0.661	0.66	0.701	0.706	0.698	0.775	0.777	0.779
20	0.662	0.661	0.658	0.702	0.706	0.698	0.778	0.78	0.781
21	0.663	0.662	0.656	0.702	0.705	0.696	0.777	0.777	0.782
22	0.664	0.661	0.657	0.702	0.705	0.695	0.777	0.778	0.78
23	0.662	0.661	0.657	0.702	0.702	0.694	0.778	0.778	0.778
24	0.664	0.661	0.656	0.702	0.703	0.691	0.778	0.776	0.779
25	0.663	0.661	0.656	0.702	0.703	0.693	0.776	0.777	0.779
26	0.663	0.661	0.656	0.701	0.701	0.691	0.777	0.779	0.78
27	0.664	0.662	0.655	0.701	0.702	0.69	0.778	0.779	0.779
28	0.664	0.661	0.655	0.7	0.702	0.691	0.778	0.78	0.779
29	0.664	0.662	0.655	0.701	0.701	0.692	0.778	0.781	0.781
30	0.665	0.661	0.656	0.701	0.702	0.693	0.778	0.779	0.781
Promedio	0.665 ± 0.003	0.664 ± 0.003	0.658 ± 0.003	0.703 ± 0.002	0.705 ± 0.003	0.696 ± 0.006	0.779 ± 0.002	0.781 ± 0.003	0.781 ± 0.005

Tabla 5.20: Colecciones multi-dominio: resultados de la medida-F del método propuesto sobre el conjunto de documentos del dominio *electronics* (E), utilizando como entrenamiento los dominios: books (B), dvd (D) y kitchen (K). Los mejores resultados están indicados por ⋆⋆. El *baseline* es CBP

5.5. Resumen de la evaluación

En este capítulo se presentó la evaluación experimental del método de clasificación propuesto. Se utilizaron tres escenarios de clasificación diferentes para este fin. La elección de estos escenarios fue motivada por el problema de insuficiencia e inexistencia de documentos etiquetados para generar clasificadores de textos, así en cada uno de los escenarios utilizados en la evaluación se hace uso de los recursos disponibles, ya sean pocos documentos etiquetados, documentos en idiomas diferentes o de dominios similares.

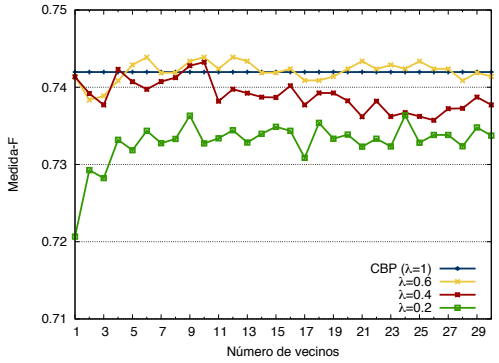
En general, el método funciona bien en los tres escenarios, sin embargo, funciona mejor en el primero donde se probó el método utilizando un conjuntos de entrenamiento muy pequeños en relación al número de documentos a clasificar.

En el segundo escenario, utilizando un enfoque multi-lenguaje, el método mejora los *baselines* pero la mejora depende de las colecciones que se usen de entrenamiento y evaluación; es decir, de los 6 experimentos realizados no en todos las mejoras fueron significativas.

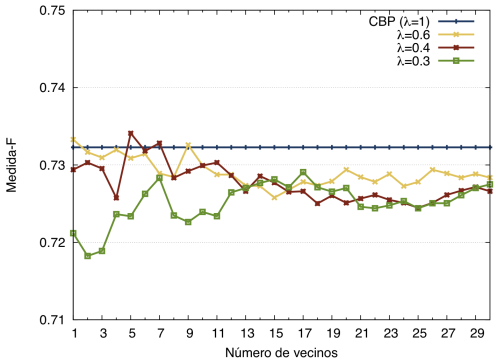
Por último, al utilizar documentos etiquetados de dominios similares al conjunto de documentos a clasificar, en el tercer escenario, no se produjeron mejoras significativas en ninguno de los 12 experimentos realizados; sin embargo, se realizó un análisis para determinar la causa. El análisis consistió en ver el promedio de similitud de los vecinos, considerados por el método, al documento que se quiere clasificar. Se calculó el número de los vecinos de un documento que pertenecen a la misma clase que él. Con los datos anteriores se determinó que la similitud de los vecinos con el documento es muy pequeña y que el porcentaje de vecinos que comparten la clase con el documento a clasificar decae exponencialmente al considerar más vecinos.

	B - K (*baseline*=0.729)			**D - K (*baseline*=0.743)**			**E - K (*baseline*=0.804)**		
k	$\lambda = 0.5$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.5$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.5$	$\lambda = 0.3$	$\lambda = 0.2$
1	0.747⋆⋆	0.734	0.724	0.753	0.747	0.734	0.816	0.806	0.786
2	0.74	0.746	0.742	0.742	0.746	0.748	0.814	0.811	0.803
3	0.738	0.742	0.743	0.747	0.759⋆⋆	0.754	0.813	0.815	0.805
4	0.734	0.737	0.736	0.749	0.748	0.758	0.82	0.825	0.817
5	0.736	0.74	0.736	0.747	0.749	0.754	0.82	0.823	0.822
6	0.736	0.739	0.734	0.749	0.747	0.752	0.815	0.819	0.82
7	0.735	0.742	0.739	0.748	0.75	0.752	0.814	0.822	0.818
8	0.736	0.739	0.737	0.747	0.751	0.754	0.816	0.821	0.819
9	0.734	0.738	0.736	0.746	0.749	0.753	0.813	0.822	0.82
10	0.733	0.737	0.733	0.743	0.746	0.75	0.812	0.819	0.823
11	0.731	0.735	0.729	0.746	0.743	0.749	0.814	0.819	0.825⋆⋆
12	0.731	0.733	0.726	0.743	0.743	0.75	0.813	0.821	0.824
13	0.732	0.735	0.729	0.743	0.743	0.747	0.811	0.821	0.818
14	0.731	0.731	0.732	0.743	0.743	0.749	0.811	0.816	0.817
15	0.732	0.73	0.731	0.744	0.741	0.746	0.811	0.817	0.818
16	0.732	0.731	0.731	0.746	0.742	0.748	0.811	0.814	0.817
17	0.733	0.731	0.73	0.744	0.741	0.747	0.81	0.815	0.818
18	0.732	0.729	0.731	0.744	0.742	0.747	0.81	0.816	0.816
19	0.732	0.729	0.728	0.745	0.743	0.749	0.81	0.816	0.816
20	0.733	0.733	0.729	0.744	0.743	0.746	0.812	0.815	0.818
21	0.732	0.733	0.732	0.745	0.743	0.746	0.811	0.814	0.817
22	0.73	0.733	0.732	0.744	0.744	0.746	0.811	0.815	0.817
23	0.729	0.731	0.729	0.745	0.744	0.745	0.811	0.813	0.814
24	0.73	0.73	0.728	0.744	0.744	0.745	0.811	0.814	0.814
25	0.729	0.729	0.726	0.744	0.744	0.746	0.81	0.814	0.813
26	0.731	0.73	0.726	0.743	0.745	0.745	0.809	0.813	0.815
27	0.73	0.729	0.724	0.744	0.746	0.744	0.808	0.813	0.816
28	0.732	0.728	0.724	0.744	0.745	0.742	0.808	0.813	0.812
29	0.731	0.729	0.72	0.744	0.746	0.742	0.808	0.813	0.815
30	0.729	0.728	0.722	0.743	0.744	0.742	0.808	0.814	0.815
Promedio	0.733 ± 0.004	0.734 ± 0.005	0.731 ± 0.006	0.745 ± 0.002	0.745 ± 0.004	0.748 ± 0.005	0.812 ± 0.003	0.816 ± 0.004	0.816 ± 0.007

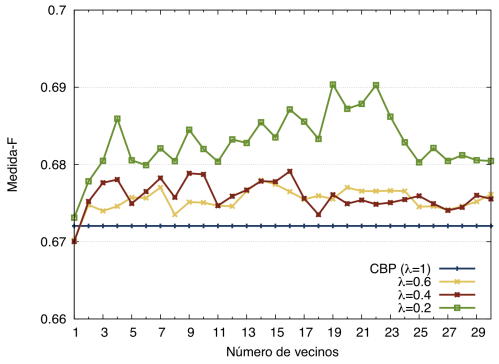
Tabla 5.21: Colecciones multi-dominio: resultados de la medida-F del método propuesto sobre el conjunto de documentos del dominio *kitchen* (K), utilizando como entrenamiento los dominios: books (B), dvd (D) y electronics (E). Los mejores resultados están indicados por ⋆⋆. El *baseline* es CBP



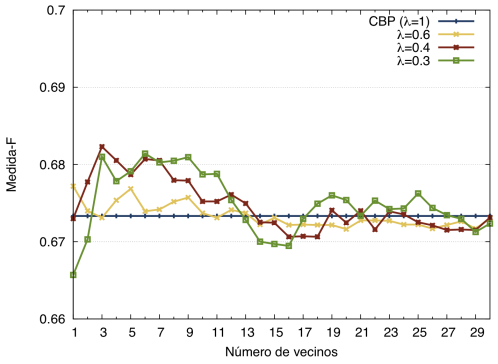
(a) Experimento $D - B$: CBP=0.742



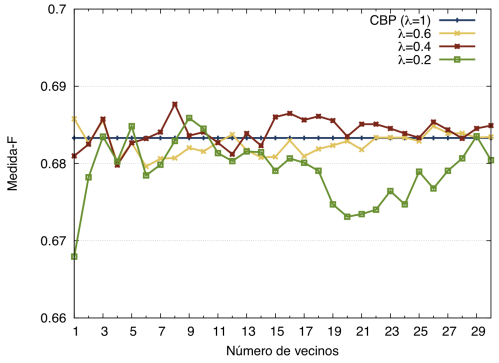
(b) Experimento $B - D$: CBP=0.732



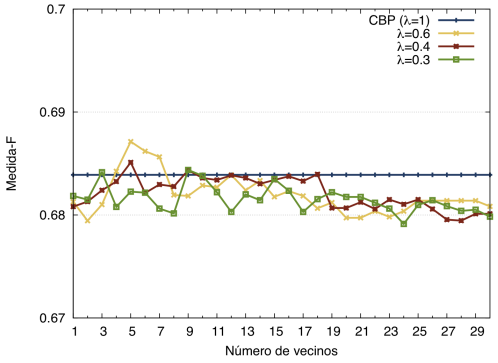
(c) Experimento $E - B$: CBP=0.672



(d) Experimento $E - D$: CBP=0.673

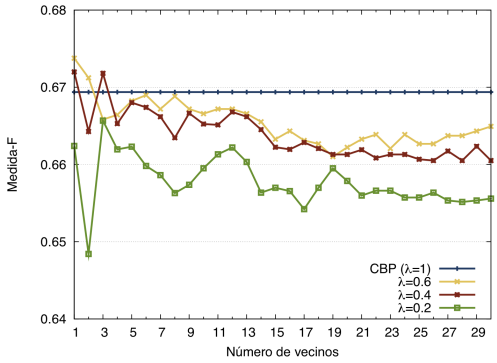


(e) Experimento $K - B$: CBP=0.683

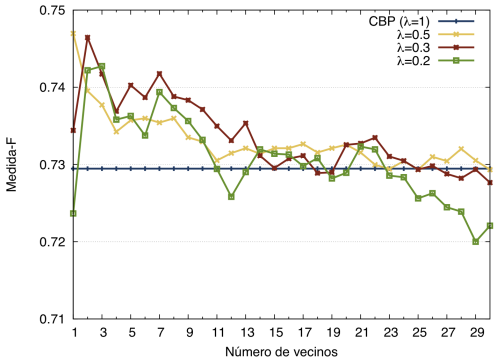


(f) Experimento $K - D$: CBP=0.684

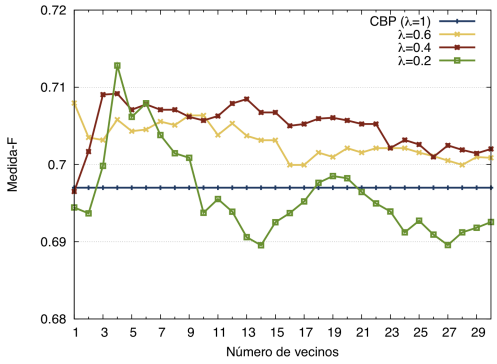
Figura 5.6: Escenario 3: gráficas de los resultados obtenidos en los experimentos multi-dominio para clasificar los dominios BOOK (usando $\lambda = 0.2$, $\lambda = 0.4$ y $\lambda = 0.6$) y DVD (usando $\lambda = 0.3$, $\lambda = 0.4$ y $\lambda = 0.6$)



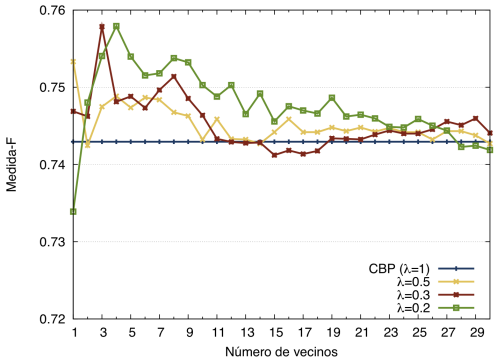
(a) Experimento $B - E$: CBP=0.669



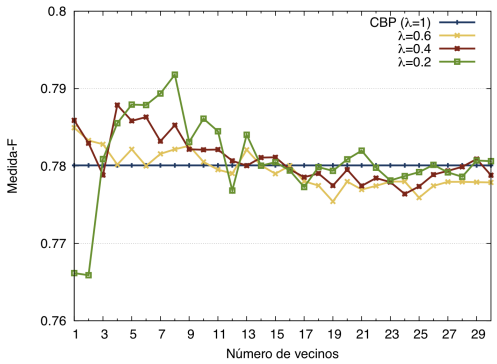
(b) Experimento $B - K$: CBP=0.729



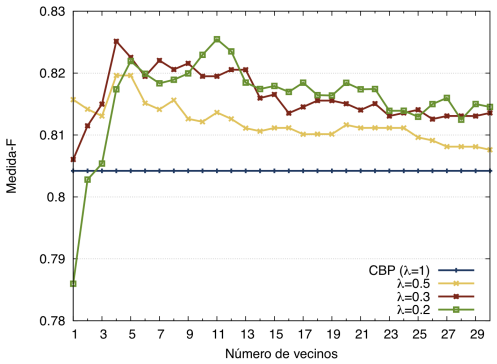
(c) Experimento $D - E$: CBP=0.687



(d) Experimento $D - K$: CBP=0.743



(e) Experimento $K - E$: CBP=0.780



(f) Experimento $E - K$: CBP=0.804

Figura 5.7: Escenario 3: gráficas de los resultados obtenidos en los experimentos multi-dominio para clasificar los dominios ELECTRONICS (usando $\lambda = 0.2$, $\lambda = 0.4$ y $\lambda = 0.6$) y KITCHEN (usando $\lambda = 0.2$, $\lambda = 0.3$ y $\lambda = 0.5$)

Capítulo 6

Conclusiones y trabajo futuro

En un enfoque de clasificación supervisada, es necesario contar con un conjunto de entrenamiento para entrenar clasificadores, en general, entre mayor sea el conjunto de entrenamiento mejor será el desempeño del clasificador construido a partir de él. Desafortunadamente no siempre existen suficientes documentos etiquetados en todos los dominios de clasificación, por lo que se han planteado soluciones que consideran: a) que existen pocos ejemplos etiquetados de un dominio particular, b) que existen documentos etiquetados suficientes en el mismo dominio pero en otro lenguaje, y c) que no existen documentos etiquetados en el dominio particular pero sí en dominios similares. Con el fin de presentar una solución alternativa a este problema se desarrollo la investigación presentada en esta tesis.

En este capítulo se presenta un resumen de la investigación realizada, se listan las conclusiones y las aportaciones principales de la investigación; además, se plantean ideas para trabajo futuro.

6.1. Conclusiones

En esta tesis se desarrolló un esquema de clasificación automática de textos utilizando información inherente en el conjunto de documentos a clasificar. El objetivo es mejorar la clasificación dentro de los siguientes tres escenarios: 1) ejemplos insuficientes en el conjunto de entrenamiento, 2) clasificación multi-lenguaje: utilización de ejemplos etiquetados en otros lenguajes, y 3) clasificación multi-dominio: utilización de ejemplos etiquetados de dominios similares.

Para determinar la clase de un documento se consideran dos partes, la información obtenida de clasificar ese documento y la información de la clasificación de sus docu-

mentos más similares. Esto basado en la suposición de *cluster*: documentos similares pueden pertenecer a la misma clase; el objetivo es dar soporte al proceso de decisión cuando al considerar sólo la primera parte, la clasificación tiene bajo desempeño.

El método se evaluó en los tres escenarios mencionados anteriormente, de los resultados obtenidos es posible concluir lo siguiente:

- En dos de los tres escenarios (escenarios 1 y 2), el método propuesto muestra una mejora significativa sobre cuatro de los modelos de clasificación más utilizados en esta tarea: K-Vecinos más Cercanos, Clasificador Bayesiano Simple, Máquinas de Vectores de Soporte, y Clasificador basado en Prototipos.

- En el tercer escenario de clasificación multi-dominio no se logró superar a los métodos de referencia debido a que la colección utilizada en este escenario consiste de opiniones, y la representación de documentos usada, no ayuda a discriminar entre la clase de opiniones positivas y negativas; esto causa que las similitudes de los documentos, incluso al primer vecino más cercano, no rebasara el 30 %. Por consiguiente la información que se obtiene de los vecinos no ayuda a mejorar la clasificación de un documento particular.

- La asignación de clase del método propuesto puede configurarse utilizando dos parámetros: λ y el número de documentos vecinos a considerar (k). Durante el desarrollo de la evaluación no fue posible determinar un valor fijo para estos dos parámetros; sin embargo, el valor para k que obtuvo mejores resultados a través de los tres escenarios siempre fue menor a 20, incluso en las colecciones con mas de 2000 documentos. En cuanto al valor de λ, se obtuvieron los mejores resultados al utilizar $\lambda < 0.6$ lo que indica que en todos los casos resultó útil considerar la información de los vecinos, incluso más que la información obtenida del propio documento.

- De manera general, el método que se ha propuesto es flexible a diferentes escenarios de clasificación; en particular, a los que se han mencionado previamente. En la mayoría de los experimentos realizados, el resultado de la clasificación se vio afectado positivamente al usar información de la clasificación de los vecinos más cercanos.

- Finalmente, con base en los resultados obtenidos en la evaluación experimental realizada en esta tesis, es recomendable aplicar el método propuesto a conjuntos

de documentos que cumplan lo siguiente: i) que exista una gran similitud entre los documentos a clasificar, entre mayor sea la similitud, se espera un mejor desempeño del método, y ii) que la colección se divida en clases temáticas, es decir, el discurso de los documentos de una misma categoría sean a cerca de un mismo asunto/objeto/materia.

La aportación de este trabajo radica en el desarrollo de un método de clasificación de textos útil en situación de carencia de conjuntos de documentos etiquetados. Hace uso de recursos existentes, mientras este conjunto de documentos existentes compartan las mismas clases que el conjunto de documentos a clasificar. El método desarrollado es simple (se basa en similitudes) y flexible (puede usarse en diferentes problemas de clasificación, al menos en escenarios de insuficiencia de documentos etiquetados, multi-lenguaje y multi-dominio). Se aprovecha la similitud entre documentos similares para obtener mejor porcentaje de clasificación. Cabe recalcar que del conjunto a clasificar sólo se utiliza información sobre similitudes entre documentos de este conjunto.

6.2. Trabajo futuro

Con la finalidad de que se pueda continuar investigando aspectos relacionados a esta tesis se plantean las siguientes ideas como trabajo futuro.

Un escenario difícil de clasificación es la clasificación multi-dominio, transferir conocimiento de un dominio a otro es una tarea que ha sido estudiada en distintas áreas y se ha trasladado a la clasificación de textos, tratando de dar solución a la carencia de conjuntos de entrenamiento. Sin embargo, en la evaluación del método realizada en este escenario no se han logrado mejoras significativas. Posibles soluciones a esto son: a) utilizar atributos que ayuden a discriminar mejor entre las clases, se puede hacer uso de listas de términos subjetivos que ayuden a determinar cuando una opinión es positiva o negativa; b) utilizar enfoques de clasificación que ya han dado buenos resultados en este tipo de clasificación y diseñar una forma de incorporar la información de los documentos similares que sea congruente con ese enfoque.

Aunque el objetivo de la tesis no era definir valores fijos para los parámetros λ y k es deseable hacerlo, por lo tanto se plantea la idea de hacer un análisis profundo del funcionamiento de estos parámetros en función del problema a resolver, es decir,

obtener algún tipo de métrica del conjunto de documentos a clasificar que pueda estar relacionado con los valores de estos parámetros.

Por último se propone utilizar el método desarrollado en esta tesis dentro de un enfoque semi-supervisado, para mejorar la elección de los documentos a ser incorporados en la construcción del modelo. Este enfoque podría reducir el número de iteraciones y hace menos costoso y más flexible el clasificador resultante.

Apéndices

Apéndice

Resultados completos

A continuación se presentan los resultados obtenidos en la evaluación del método en tres escenarios de clasificación.

A.1. Escenario 1: Ejemplos insuficientes en el conjunto de entrenamiento

En las Tablas A.1, A.2 y A.3 se muestran los resultados obtenidos con el método propuesto en las colecciones R8-reducido-41, R8-reducido-20 y R8-reducido-10, respectivamente.

	R8-reducido-41 (*baseline*=0.836)									
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.841	0.844	0.847	0.836	0.844	0.843	0.841	0.816	0.811	0.77
2	0.839	0.844	0.848	0.86	0.865	0.869	0.865	0.847	0.828	0.803
3	0.84	0.842	0.847	0.857	0.867	0.872	0.872	0.859	0.843	0.82
4	0.838	0.845	0.851	0.855	0.864	0.867	0.877	0.868	0.833	0.814
5	0.837	0.845	0.854	0.853	0.859	0.862	0.871	0.861	0.844	0.813
6	0.841	0.842	0.855	0.862	0.863	0.869	0.873	0.865	0.854	0.803
7	0.84	0.841	0.855	0.863	0.873	0.866	0.863	0.859	0.822	0.789
8	0.84	0.841	0.854	0.862	0.871	0.871	0.861	0.859	0.827	0.79
9	0.84	0.843	0.854	0.859	0.867	0.875	0.862	0.867	0.831	0.791
10	0.84	0.844	0.854	0.859	0.862	0.863	0.86	0.85	0.827	0.787
11	0.84	0.846	0.854	0.861	0.854	0.861	0.857	0.85	0.819	0.786
12	0.84	0.845	0.853	0.858	0.851	0.848	0.854	0.842	0.823	0.786
13	0.84	0.845	0.852	0.856	0.851	0.853	0.855	0.844	0.824	0.786
14	0.84	0.845	0.851	0.856	0.85	0.853	0.856	0.83	0.813	0.782
15	0.84	0.846	0.85	0.856	0.85	0.848	0.855	0.829	0.823	0.777
16	0.84	0.845	0.85	0.847	0.843	0.85	0.854	0.83	0.801	0.777
17	0.84	0.845	0.851	0.849	0.849	0.853	0.854	0.825	0.805	0.778
18	0.84	0.845	0.851	0.849	0.85	0.85	0.854	0.826	0.798	0.779
19	0.84	0.846	0.853	0.851	0.852	0.849	0.854	0.826	0.801	0.766
20	0.84	0.846	0.853	0.849	0.852	0.847	0.853	0.826	0.8	0.767
21	0.84	0.846	0.853	0.85	0.852	0.846	0.852	0.819	0.798	0.766
22	0.84	0.844	0.851	0.849	0.853	0.849	0.852	0.818	0.797	0.764
23	0.84	0.844	0.851	0.846	0.852	0.857	0.852	0.818	0.8	0.758
24	0.84	0.844	0.851	0.849	0.849	0.856	0.85	0.819	0.8	0.759
25	0.84	0.844	0.852	0.848	0.849	0.857	0.852	0.823	0.796	0.752
26	0.841	0.844	0.85	0.849	0.846	0.854	0.852	0.824	0.799	0.754
27	0.841	0.844	0.85	0.849	0.846	0.854	0.848	0.834	0.795	0.751
28	0.841	0.844	0.85	0.847	0.846	0.854	0.845	0.838	0.795	0.748
29	0.841	0.844	0.845	0.846	0.844	0.852	0.845	0.836	0.794	0.75
30	0.841	0.844	0.845	0.846	0.844	0.854	0.846	0.834	0.79	0.735
Promedio	0.84 ± 0.001	0.844 ± 0.001	0.851 ± 0.003	0.853 ± 0.006	0.854 ± 0.009	0.857 ± 0.009	0.856 ± 0.009	0.838 ± 0.017	0.813 ± 0.018	0.777 ± 0.021

Tabla A.1: Escenario 1: resultados de la medida-F para 10 valores de λ sobre la colección R8-reducido-41

	R8-reducido-20 (*baseline*=0.803)									
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.814	0.824	0.815	0.816	0.814	0.827	0.817	0.797	0.782	0.717
2	0.809	0.815	0.819	0.828	0.82	0.823	0.83	0.815	0.805	0.786
3	0.809	0.815	0.831	0.833	0.84	0.84	0.844	0.841	0.825	0.802
4	0.809	0.816	0.831	0.829	0.839	0.843	0.841	0.85	0.831	0.796
5	0.809	0.816	0.835	0.832	0.832	0.86	0.84	0.84	0.826	0.803
6	0.809	0.813	0.826	0.832	0.832	0.854	0.855	0.832	0.834	0.797
7	0.809	0.813	0.829	0.832	0.839	0.858	0.853	0.844	0.814	0.792
8	0.809	0.814	0.828	0.829	0.845	0.856	0.858	0.838	0.809	0.78
9	0.809	0.814	0.826	0.83	0.848	0.856	0.848	0.849	0.824	0.776
10	0.811	0.813	0.824	0.826	0.841	0.849	0.842	0.838	0.815	0.775
11	0.81	0.812	0.825	0.827	0.842	0.847	0.842	0.837	0.81	0.778
12	0.81	0.812	0.821	0.825	0.84	0.842	0.831	0.831	0.805	0.779
13	0.809	0.812	0.821	0.83	0.84	0.843	0.83	0.823	0.792	0.775
14	0.81	0.813	0.821	0.822	0.838	0.844	0.836	0.809	0.792	0.784
15	0.809	0.813	0.821	0.822	0.824	0.833	0.827	0.808	0.81	0.782
16	0.809	0.813	0.821	0.821	0.822	0.833	0.827	0.809	0.788	0.774
17	0.809	0.812	0.817	0.822	0.823	0.832	0.835	0.81	0.784	0.774
18	0.809	0.812	0.817	0.822	0.822	0.827	0.838	0.805	0.785	0.762
19	0.809	0.812	0.817	0.824	0.817	0.828	0.836	0.808	0.787	0.766
20	0.809	0.815	0.817	0.82	0.816	0.829	0.834	0.812	0.788	0.762
21	0.809	0.816	0.817	0.82	0.816	0.829	0.834	0.82	0.786	0.761
22	0.809	0.816	0.817	0.82	0.815	0.829	0.833	0.809	0.788	0.763
23	0.809	0.816	0.817	0.826	0.816	0.829	0.832	0.828	0.789	0.749
24	0.809	0.816	0.817	0.83	0.806	0.818	0.832	0.828	0.787	0.757
25	0.809	0.816	0.817	0.829	0.805	0.819	0.832	0.821	0.788	0.75
26	0.809	0.817	0.817	0.82	0.807	0.819	0.821	0.819	0.787	0.752
27	0.809	0.815	0.817	0.82	0.807	0.817	0.827	0.821	0.788	0.751
28	0.809	0.815	0.816	0.821	0.811	0.816	0.818	0.82	0.786	0.753
29	0.809	0.815	0.818	0.821	0.811	0.815	0.818	0.82	0.786	0.734
30	0.809	0.816	0.818	0.821	0.81	0.814	0.818	0.822	0.782	0.735
Promedio	0.809 ± 0.001	0.815 ± 0.002	0.821 ± 0.005	0.825 ± 0.005	0.825 ± 0.014	0.834 ± 0.014	0.834 ± 0.011	0.823 ± 0.014	0.799 ± 0.017	0.769 ± 0.021

Tabla A.2: Escenario 1: resultados de la medida-F para 10 valores de λ sobre la colección R8-reducido-20

	R8-reducido-10 (*baseline*=0.765)									
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.775	0.78	0.781	0.79	0.807	0.811	0.794	0.761	0.723	0.685
2	0.78	0.784	0.796	0.805	0.813	0.813	0.82	0.814	0.8	0.756
3	0.78	0.785	0.791	0.798	0.805	0.821	0.83	0.823	0.81	0.795
4	0.782	0.787	0.795	0.805	0.814	0.83	0.833	0.834	0.823	0.813
5	0.781	0.789	0.793	0.807	0.819	0.829	0.834	0.842	0.844	0.806
6	0.78	0.784	0.789	0.799	0.811	0.829	0.833	0.839	0.828	0.787
7	0.783	0.789	0.792	0.798	0.813	0.835	0.835	0.834	0.823	0.789
8	0.782	0.789	0.794	0.796	0.813	0.836	0.84	0.845	0.802	0.78
9	0.782	0.788	0.792	0.797	0.814	0.832	0.836	0.835	0.822	0.781
10	0.782	0.788	0.792	0.798	0.814	0.828	0.829	0.822	0.819	0.783
11	0.781	0.788	0.792	0.795	0.81	0.827	0.835	0.82	0.815	0.785
12	0.78	0.785	0.79	0.796	0.81	0.828	0.828	0.821	0.81	0.772
13	0.781	0.785	0.791	0.795	0.81	0.818	0.828	0.82	0.8	0.775
14	0.78	0.786	0.792	0.794	0.809	0.816	0.825	0.822	0.801	0.774
15	0.78	0.784	0.791	0.794	0.806	0.817	0.815	0.82	0.798	0.771
16	0.78	0.783	0.792	0.794	0.805	0.817	0.812	0.805	0.801	0.768
17	0.779	0.783	0.791	0.794	0.806	0.81	0.813	0.805	0.795	0.767
18	0.775	0.784	0.792	0.795	0.806	0.808	0.813	0.805	0.794	0.76
19	0.774	0.784	0.791	0.795	0.806	0.808	0.811	0.801	0.791	0.747
20	0.774	0.784	0.791	0.795	0.8	0.81	0.808	0.802	0.785	0.745
21	0.773	0.784	0.791	0.794	0.799	0.812	0.809	0.796	0.781	0.743
22	0.773	0.786	0.79	0.794	0.801	0.812	0.808	0.799	0.783	0.751
23	0.773	0.786	0.79	0.794	0.798	0.815	0.808	0.801	0.783	0.75
24	0.773	0.786	0.791	0.794	0.799	0.812	0.808	0.797	0.781	0.752
25	0.773	0.786	0.791	0.794	0.798	0.814	0.808	0.793	0.781	0.75
26	0.773	0.786	0.791	0.794	0.798	0.812	0.806	0.794	0.78	0.747
27	0.773	0.781	0.791	0.794	0.798	0.81	0.807	0.796	0.779	0.746
28	0.773	0.781	0.794	0.794	0.798	0.81	0.806	0.791	0.78	0.73
29	0.773	0.781	0.791	0.794	0.797	0.809	0.808	0.791	0.779	0.73
30	0.773	0.781	0.791	0.794	0.797	0.808	0.817	0.791	0.776	0.733
Promedio	0.777 ± 0.004	0.785 ± 0.003	0.791 ± 0.002	0.796 ± 0.004	0.806 ± 0.007	0.818 ± 0.009	0.819 ± 0.012	0.811 ± 0.019	0.796 ± 0.023	0.762 ± 0.026

Tabla A.3: Escenario 1: resultados de la medida-F para 10 valores de λ sobre la colección R8-reducido-10

A.2. Escenario 2: Clasificación multi-lenguaje

Los resultados de las evaluaciones completas para el escenarios multi-lenguaje se presentan en las Tablas A.4, A.5, A.6, A.7, A.8 y A.9.

					Francés - Español (*baseline*=0.879)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.879	0.883	0.879	0.886	0.886	0.902	0.903	0.893	0.897	0.894
2	0.879	0.883	0.886	0.889	0.893	0.899	0.906	0.906	0.912	0.906
3	0.879	0.883	0.89	0.893	0.89	0.896	0.899	0.905	0.906	0.909
4	0.879	0.886	0.89	0.887	0.893	0.899	0.903	0.909	0.912	0.909
5	0.879	0.886	0.89	0.887	0.893	0.899	0.906	0.913	0.912	0.909
6	0.879	0.88	0.89	0.887	0.893	0.902	0.902	0.903	0.906	0.909
7	0.879	0.88	0.89	0.887	0.893	0.902	0.906	0.912	0.906	0.91
8	0.879	0.879	0.883	0.887	0.89	0.902	0.912	0.915	0.906	0.913
9	0.879	0.879	0.886	0.89	0.89	0.899	0.915	0.919	0.916	0.906
10	0.879	0.879	0.886	0.886	0.884	0.899	0.915	0.919	0.919	0.91
11	0.879	0.879	0.883	0.886	0.884	0.899	0.912	0.922	0.91	0.904
12	0.879	0.879	0.883	0.886	0.884	0.899	0.912	0.919	0.909	0.904
13	0.879	0.879	0.883	0.886	0.887	0.899	0.912	0.919	0.912	0.907
14	0.879	0.879	0.883	0.886	0.887	0.903	0.915	0.919	0.916	0.904
15	0.879	0.879	0.883	0.886	0.889	0.897	0.915	0.919	0.919	0.904
16	0.879	0.879	0.883	0.886	0.893	0.897	0.915	0.919	0.919	0.901
17	0.879	0.879	0.883	0.886	0.893	0.897	0.915	0.919	0.919	0.898
18	0.879	0.879	0.883	0.886	0.893	0.9	0.912	0.919	0.919	0.895
19	0.879	0.879	0.883	0.886	0.893	0.9	0.912	0.922	0.922	0.893
20	0.879	0.879	0.883	0.886	0.893	0.896	0.912	0.919	0.922	0.896
21	0.879	0.879	0.883	0.886	0.893	0.896	0.919	0.922	0.922	0.893
22	0.879	0.879	0.88	0.886	0.893	0.893	0.919	0.922	0.922	0.896
23	0.879	0.879	0.88	0.886	0.893	0.893	0.912	0.922	0.922	0.893
24	0.879	0.879	0.88	0.886	0.893	0.893	0.916	0.922	0.919	0.893
25	0.879	0.879	0.88	0.886	0.893	0.893	0.913	0.922	0.922	0.893
26	0.879	0.879	0.88	0.886	0.893	0.893	0.913	0.922	0.916	0.89
27	0.879	0.879	0.88	0.883	0.893	0.893	0.91	0.922	0.916	0.893
28	0.879	0.879	0.88	0.886	0.893	0.893	0.916	0.922	0.916	0.896
29	0.879	0.879	0.88	0.883	0.893	0.893	0.91	0.922	0.916	0.893
30	0.879	0.879	0.883	0.886	0.893	0.896	0.903	0.922	0.916	0.893
Promedio	0.879 ± 0	0.88 ± 0.002	0.884 ± 0.003	0.886 ± 0.002	0.891 ± 0.003	0.897 ± 0.003	0.911 ± 0.005	0.917 ± 0.007	0.915 ± 0.006	0.9 ± 0.007

Tabla A.4: Escenario 2: resultados de la medida-F para 10 valores de λ para Francés-Español

					Inglés - Español (*baseline*=0.814)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.814	0.811	0.814	0.817	0.82	0.817	0.833	0.824	0.822	0.818
2	0.818	0.814	0.814	0.814	0.82	0.823	0.839	0.852	0.857	0.851
3	0.818	0.814	0.817	0.826	0.819	0.822	0.838	0.846	0.854	0.837
4	0.815	0.818	0.824	0.827	0.827	0.823	0.841	0.839	0.82	0.821
5	0.815	0.818	0.824	0.817	0.827	0.826	0.836	0.832	0.818	0.804
6	0.815	0.818	0.821	0.825	0.827	0.836	0.847	0.843	0.83	0.833
7	0.815	0.818	0.821	0.825	0.83	0.833	0.836	0.845	0.832	0.818
8	0.815	0.815	0.821	0.825	0.831	0.837	0.851	0.854	0.83	0.824
9	0.815	0.815	0.821	0.828	0.834	0.841	0.847	0.849	0.841	0.83
10	0.815	0.815	0.821	0.828	0.831	0.837	0.844	0.85	0.837	0.821
11	0.815	0.815	0.825	0.828	0.834	0.833	0.844	0.85	0.837	0.817
12	0.815	0.819	0.825	0.828	0.831	0.837	0.845	0.853	0.837	0.814
13	0.815	0.819	0.825	0.83	0.834	0.837	0.845	0.851	0.834	0.817
14	0.815	0.819	0.825	0.83	0.831	0.842	0.849	0.857	0.835	0.812
15	0.815	0.819	0.825	0.827	0.831	0.842	0.849	0.854	0.838	0.812
16	0.815	0.819	0.825	0.83	0.831	0.842	0.849	0.854	0.835	0.816
17	0.818	0.819	0.825	0.83	0.831	0.838	0.849	0.85	0.835	0.815
18	0.818	0.819	0.825	0.83	0.834	0.838	0.849	0.848	0.838	0.815
19	0.818	0.819	0.825	0.827	0.834	0.842	0.849	0.848	0.835	0.807
20	0.818	0.819	0.821	0.827	0.831	0.842	0.846	0.844	0.841	0.811
21	0.818	0.819	0.821	0.827	0.831	0.842	0.846	0.837	0.842	0.808
22	0.818	0.819	0.821	0.827	0.83	0.842	0.846	0.84	0.835	0.804
23	0.818	0.819	0.818	0.827	0.827	0.842	0.846	0.84	0.831	0.796
24	0.818	0.816	0.818	0.827	0.827	0.842	0.846	0.84	0.828	0.793
25	0.818	0.816	0.822	0.827	0.827	0.842	0.846	0.837	0.828	0.798
26	0.818	0.816	0.822	0.827	0.827	0.842	0.846	0.83	0.824	0.801
27	0.814	0.816	0.822	0.827	0.83	0.842	0.846	0.834	0.82	0.798
28	0.814	0.816	0.822	0.821	0.83	0.842	0.842	0.834	0.824	0.801
29	0.814	0.816	0.822	0.821	0.827	0.839	0.842	0.834	0.822	0.797
30	0.814	0.812	0.818	0.817	0.823	0.838	0.839	0.826	0.821	0.793
Promedio	0.816 ± 0.002	0.817 ± 0.002	0.822 ± 0.003	0.826 ± 0.004	0.829 ± 0.004	0.837 ± 0.007	0.844 ± 0.005	0.843 ± 0.009	0.833 ± 0.009	0.813 ± 0.014

Tabla A.5: Escenario 2: resultados de la medida-F para 10 valores de λ para Inglés-Español

					Español - Francés ($baseline$=0.790)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.79	0.793	0.797	0.797	0.808	0.8	0.797	0.793	0.788	0.753
2	0.79	0.793	0.793	0.797	0.803	0.809	0.808	0.805	0.812	0.795
3	0.793	0.793	0.797	0.797	0.803	0.811	0.817	0.826	0.814	0.81
4	0.793	0.79	0.796	0.796	0.803	0.805	0.808	0.817	0.831	0.815
5	0.79	0.79	0.79	0.796	0.803	0.805	0.811	0.817	0.829	0.8
6	0.79	0.793	0.79	0.793	0.803	0.802	0.795	0.814	0.812	0.787
7	0.79	0.79	0.79	0.796	0.797	0.798	0.799	0.802	0.807	0.784
8	0.79	0.793	0.793	0.79	0.791	0.792	0.805	0.799	0.81	0.783
9	0.79	0.793	0.796	0.793	0.792	0.802	0.798	0.792	0.815	0.793
10	0.79	0.793	0.797	0.793	0.787	0.799	0.801	0.802	0.816	0.806
11	0.787	0.793	0.797	0.793	0.787	0.792	0.799	0.81	0.82	0.806
12	0.787	0.793	0.797	0.793	0.793	0.799	0.799	0.803	0.817	0.797
13	0.787	0.793	0.796	0.793	0.792	0.796	0.801	0.799	0.81	0.803
14	0.787	0.793	0.796	0.793	0.795	0.799	0.807	0.808	0.82	0.817
15	0.787	0.793	0.797	0.793	0.795	0.795	0.801	0.805	0.823	0.81
16	0.787	0.793	0.797	0.793	0.796	0.793	0.798	0.804	0.823	0.806
17	0.787	0.793	0.793	0.793	0.796	0.793	0.8	0.805	0.823	0.803
18	0.787	0.793	0.793	0.793	0.796	0.792	0.8	0.807	0.813	0.802
19	0.787	0.793	0.793	0.79	0.796	0.789	0.797	0.798	0.81	0.802
20	0.787	0.793	0.793	0.79	0.796	0.792	0.797	0.795	0.804	0.797
21	0.787	0.793	0.793	0.797	0.796	0.795	0.797	0.795	0.807	0.789
22	0.787	0.793	0.793	0.793	0.796	0.792	0.794	0.791	0.804	0.782
23	0.787	0.793	0.793	0.797	0.796	0.792	0.794	0.788	0.787	0.777
24	0.787	0.793	0.793	0.797	0.793	0.792	0.794	0.788	0.787	0.774
25	0.787	0.793	0.793	0.797	0.793	0.792	0.794	0.788	0.786	0.77
26	0.787	0.793	0.793	0.793	0.793	0.792	0.794	0.788	0.787	0.766
27	0.787	0.793	0.793	0.793	0.793	0.795	0.794	0.788	0.786	0.772
28	0.787	0.793	0.793	0.797	0.793	0.795	0.794	0.788	0.786	0.769
29	0.787	0.793	0.793	0.793	0.793	0.795	0.794	0.784	0.784	0.761
30	0.787	0.793	0.793	0.793	0.793	0.795	0.791	0.784	0.784	0.759
Promedio	0.788 ± 0.002	0.793 ± 0.001	0.794 ± 0.002	0.794 ± 0.002	0.796 ± 0.005	0.797 ± 0.006	0.799 ± 0.006	0.799 ± 0.011	0.807 ± 0.015	0.79 ± 0.018

Tabla A.6: Escenario 2: resultados de la medida-F para 10 valores de λ para Español-Francés

					Inglés - Francés ($baseline$=0.616)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.613	0.621	0.624	0.625	0.617	0.604	0.597	0.589	0.577	0.559
2	0.619	0.615	0.609	0.612	0.618	0.619	0.608	0.611	0.61	0.614
3	0.622	0.621	0.61	0.607	0.608	0.606	0.618	0.628	0.623	0.617
4	0.616	0.624	0.61	0.611	0.611	0.602	0.611	0.629	0.642	0.637
5	0.622	0.622	0.611	0.614	0.609	0.61	0.615	0.625	0.636	0.629
6	0.616	0.62	0.612	0.609	0.602	0.604	0.622	0.623	0.641	0.638
7	0.616	0.625	0.614	0.617	0.611	0.608	0.614	0.645	0.655	0.661
8	0.622	0.625	0.617	0.617	0.611	0.602	0.616	0.647	0.66	0.682
9	0.622	0.625	0.62	0.617	0.615	0.611	0.631	0.641	0.655	0.668
10	0.625	0.625	0.625	0.62	0.618	0.608	0.626	0.638	0.666	0.68
11	0.625	0.625	0.625	0.624	0.618	0.609	0.623	0.641	0.661	0.672
12	0.622	0.625	0.628	0.62	0.618	0.609	0.623	0.633	0.664	0.67
13	0.622	0.625	0.625	0.62	0.615	0.609	0.626	0.632	0.656	0.667
14	0.622	0.625	0.625	0.624	0.618	0.609	0.622	0.634	0.656	0.675
15	0.622	0.625	0.625	0.62	0.622	0.609	0.617	0.631	0.656	0.677
16	0.616	0.625	0.625	0.62	0.622	0.609	0.615	0.627	0.656	0.68
17	0.622	0.625	0.625	0.62	0.618	0.602	0.615	0.63	0.648	0.674
18	0.616	0.625	0.625	0.62	0.618	0.609	0.617	0.63	0.648	0.677
19	0.616	0.625	0.625	0.62	0.622	0.609	0.617	0.617	0.648	0.674
20	0.622	0.625	0.625	0.62	0.622	0.609	0.62	0.617	0.641	0.667
21	0.616	0.625	0.625	0.624	0.622	0.612	0.618	0.614	0.636	0.662
22	0.616	0.625	0.625	0.624	0.625	0.612	0.618	0.608	0.63	0.644
23	0.616	0.625	0.625	0.624	0.622	0.609	0.615	0.616	0.628	0.644
24	0.616	0.625	0.625	0.62	0.619	0.609	0.615	0.616	0.625	0.636
25	0.616	0.625	0.625	0.624	0.619	0.615	0.615	0.613	0.625	0.625
26	0.616	0.62	0.625	0.624	0.619	0.609	0.615	0.619	0.618	0.617
27	0.616	0.625	0.625	0.624	0.619	0.609	0.615	0.613	0.613	0.617
28	0.616	0.62	0.625	0.624	0.619	0.611	0.615	0.608	0.613	0.623
29	0.616	0.62	0.62	0.623	0.613	0.606	0.615	0.608	0.61	0.609
30	0.616	0.62	0.62	0.623	0.613	0.605	0.612	0.608	0.61	0.611
Promedio	0.619 ± 0.003	0.623 ± 0.003	0.622 ± 0.006	0.62 ± 0.005	0.617 ± 0.005	0.608 ± 0.004	0.617 ± 0.006	0.623 ± 0.013	0.637 ± 0.021	0.647 ± 0.03

Tabla A.7: Escenario 2: resultados de la medida-F para 10 valores de λ para Inglés-Francés

					Español - Inglés (*baseline*=0.851)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.861	0.864	0.867	0.874	0.873	0.873	0.876	0.883	0.861	0.851
2	0.864	0.868	0.874	0.883	0.895	0.897	0.894	0.899	0.883	0.877
3	0.864	0.864	0.88	0.882	0.898	0.894	0.894	0.905	0.895	0.89
4	0.861	0.864	0.869	0.875	0.895	0.908	0.917	0.924	0.915	0.915
5	0.861	0.87	0.866	0.879	0.901	0.91	0.917	0.92	0.92	0.917
6	0.861	0.87	0.879	0.882	0.898	0.91	0.92	0.927	0.92	0.921
7	0.861	0.87	0.876	0.885	0.904	0.907	0.921	0.927	0.93	0.931
8	0.861	0.864	0.876	0.882	0.901	0.914	0.921	0.924	0.93	0.928
9	0.861	0.867	0.876	0.882	0.9	0.914	0.921	0.924	0.934	0.934
10	0.858	0.861	0.873	0.879	0.897	0.907	0.917	0.927	0.94	0.94
11	0.861	0.861	0.873	0.888	0.897	0.91	0.92	0.93	0.94	0.944
12	0.861	0.861	0.869	0.885	0.897	0.904	0.917	0.927	0.94	0.934
13	0.861	0.861	0.869	0.885	0.894	0.9	0.907	0.927	0.937	0.943
14	0.858	0.861	0.869	0.882	0.894	0.897	0.91	0.93	0.937	0.944
15	0.858	0.861	0.869	0.885	0.89	0.901	0.91	0.927	0.937	0.94
16	0.858	0.861	0.869	0.878	0.89	0.901	0.917	0.93	0.94	0.947
17	0.858	0.861	0.869	0.878	0.89	0.901	0.917	0.927	0.937	0.95
18	0.858	0.861	0.869	0.882	0.89	0.897	0.914	0.93	0.934	0.947
19	0.858	0.861	0.869	0.881	0.89	0.901	0.914	0.93	0.934	0.947
20	0.858	0.861	0.869	0.882	0.89	0.901	0.91	0.927	0.934	0.944
21	0.858	0.861	0.869	0.884	0.89	0.901	0.91	0.924	0.937	0.941
22	0.858	0.861	0.866	0.882	0.89	0.901	0.91	0.927	0.937	0.944
23	0.858	0.861	0.864	0.879	0.89	0.901	0.914	0.924	0.94	0.941
24	0.858	0.861	0.864	0.879	0.887	0.901	0.914	0.924	0.94	0.941
25	0.854	0.861	0.861	0.869	0.887	0.901	0.91	0.924	0.934	0.937
26	0.854	0.861	0.861	0.872	0.884	0.904	0.907	0.924	0.934	0.937
27	0.854	0.861	0.861	0.872	0.884	0.904	0.91	0.924	0.937	0.937
28	0.854	0.861	0.861	0.872	0.884	0.9	0.907	0.924	0.939	0.94
29	0.854	0.861	0.861	0.876	0.882	0.9	0.907	0.924	0.937	0.937
30	0.854	0.861	0.861	0.872	0.882	0.9	0.907	0.924	0.937	0.937
Promedio	0.859 ± 0.003	0.863 ± 0.003	0.869 ± 0.005	0.88 ± 0.005	0.891 ± 0.007	0.902 ± 0.007	0.911 ± 0.009	0.923 ± 0.01	0.929 ± 0.018	0.931 ± 0.022

Tabla A.8: Escenario 2: resultados de la medida-F para 10 valores de λ para Español-Inglés

					Francés - Inglés (*baseline*=0.956)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.963	0.963	0.959	0.956	0.956	0.959	0.966	0.953	0.944	0.925
2	0.959	0.959	0.963	0.959	0.959	0.959	0.963	0.963	0.95	0.938
3	0.959	0.963	0.963	0.963	0.959	0.959	0.963	0.963	0.959	0.935
4	0.959	0.966	0.966	0.963	0.959	0.956	0.956	0.956	0.956	0.935
5	0.963	0.966	0.966	0.966	0.966	0.963	0.966	0.963	0.966	0.947
6	0.963	0.963	0.966	0.966	0.963	0.963	0.966	0.966	0.969	0.95
7	0.963	0.963	0.966	0.966	0.966	0.963	0.963	0.963	0.966	0.95
8	0.963	0.963	0.966	0.966	0.966	0.963	0.966	0.969	0.969	0.95
9	0.963	0.963	0.966	0.966	0.966	0.963	0.966	0.966	0.969	0.95
10	0.963	0.963	0.963	0.963	0.963	0.963	0.966	0.969	0.969	0.96
11	0.963	0.963	0.963	0.963	0.963	0.96	0.963	0.966	0.966	0.954
12	0.963	0.963	0.963	0.963	0.963	0.96	0.963	0.966	0.966	0.954
13	0.963	0.963	0.963	0.963	0.963	0.96	0.963	0.966	0.963	0.951
14	0.963	0.963	0.963	0.963	0.963	0.963	0.963	0.963	0.96	0.951
15	0.963	0.963	0.963	0.963	0.963	0.963	0.963	0.966	0.963	0.954
16	0.963	0.963	0.963	0.963	0.963	0.963	0.963	0.966	0.963	0.954
17	0.963	0.963	0.963	0.963	0.966	0.966	0.963	0.963	0.963	0.951
18	0.963	0.963	0.966	0.966	0.966	0.966	0.966	0.966	0.966	0.96
19	0.963	0.963	0.966	0.966	0.966	0.963	0.963	0.966	0.966	0.957
20	0.963	0.963	0.966	0.966	0.963	0.963	0.963	0.966	0.966	0.954
21	0.963	0.963	0.966	0.966	0.963	0.963	0.963	0.966	0.966	0.957
22	0.963	0.963	0.966	0.966	0.966	0.963	0.963	0.966	0.966	0.957
23	0.963	0.963	0.966	0.966	0.963	0.963	0.963	0.969	0.966	0.957
24	0.963	0.963	0.966	0.966	0.966	0.963	0.963	0.966	0.966	0.954
25	0.963	0.963	0.966	0.966	0.966	0.963	0.963	0.966	0.966	0.957
26	0.963	0.963	0.966	0.966	0.966	0.963	0.963	0.969	0.966	0.957
27	0.963	0.963	0.966	0.966	0.966	0.966	0.963	0.969	0.966	0.963
28	0.963	0.963	0.963	0.966	0.966	0.963	0.963	0.966	0.966	0.96
29	0.963	0.963	0.963	0.966	0.966	0.966	0.963	0.969	0.963	0.96
30	0.963	0.963	0.966	0.966	0.966	0.966	0.963	0.969	0.966	0.96
Promedio	0.963 ± 0.001	0.963 ± 0.001	0.965 ± 0.002	0.964 ± 0.002	0.964 ± 0.003	0.963 ± 0.002	0.963 ± 0.002	0.965 ± 0.004	0.964 ± 0.005	0.952 ± 0.009

Tabla A.9: Escenario 2: resultados de la medida-F para 10 valores de λ para Francés-Inglés

A.3. Escenario 3: Clasificación multi-dominio

Ahora se presentan doce tablas que corresponden a los resultados de los doce experimentos realizados en el escenario de clasificación multi-dominio, las doce tablas son A.10 a la tabla A.21

					D - B (*baseline*=0.742)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.74	0.742	0.74	0.741	0.74	0.741	0.737	0.721	0.703	0.649
2	0.742	0.742	0.74	0.738	0.738	0.739	0.735	0.729	0.713	0.669
3	0.742	0.741	0.74	0.739	0.738	0.738	0.741	0.728	0.71	0.663
4	0.741	0.742	0.742	0.741	0.739	0.742	0.743	0.733	0.708	0.662
5	0.741	0.743	0.739	0.743	0.736	0.741	0.736	0.732	0.713	0.669
6	0.741	0.743	0.742	0.744	0.74	0.74	0.735	0.734	0.721	0.682
7	0.742	0.745	0.744	0.742	0.743	0.741	0.74	0.733	0.726	0.679
8	0.742	0.745	0.745	0.742	0.743	0.741	0.741	0.733	0.722	0.674
9	0.743	0.744	0.744	0.743	0.741	0.743	0.741	0.736	0.72	0.672
10	0.742	0.746	0.745	0.744	0.743	0.743	0.742	0.733	0.724	0.671
11	0.742	0.745	0.744	0.742	0.742	0.738	0.74	0.733	0.719	0.669
12	0.742	0.745	0.744	0.744	0.741	0.74	0.739	0.734	0.716	0.672
13	0.741	0.744	0.747	0.743	0.741	0.739	0.74	0.733	0.715	0.672
14	0.741	0.743	0.746	0.742	0.74	0.739	0.739	0.734	0.716	0.67
15	0.741	0.744	0.743	0.742	0.741	0.739	0.74	0.735	0.712	0.664
16	0.741	0.744	0.743	0.742	0.741	0.74	0.737	0.734	0.713	0.664
17	0.741	0.744	0.742	0.741	0.739	0.738	0.736	0.731	0.713	0.668
18	0.741	0.742	0.743	0.741	0.74	0.739	0.736	0.735	0.717	0.667
19	0.742	0.742	0.743	0.741	0.74	0.739	0.737	0.733	0.719	0.675
20	0.742	0.742	0.743	0.742	0.739	0.738	0.735	0.734	0.72	0.666
21	0.742	0.742	0.742	0.743	0.743	0.736	0.736	0.732	0.721	0.667
22	0.742	0.742	0.744	0.742	0.74	0.738	0.734	0.733	0.719	0.665
23	0.741	0.742	0.743	0.743	0.739	0.736	0.733	0.732	0.722	0.671
24	0.741	0.743	0.744	0.742	0.74	0.737	0.733	0.736	0.721	0.673
25	0.741	0.744	0.742	0.743	0.74	0.736	0.733	0.733	0.72	0.672
26	0.741	0.743	0.741	0.742	0.739	0.736	0.734	0.734	0.723	0.67
27	0.741	0.742	0.743	0.742	0.739	0.737	0.735	0.734	0.722	0.668
28	0.741	0.742	0.743	0.741	0.74	0.737	0.736	0.732	0.72	0.672
29	0.741	0.742	0.743	0.742	0.741	0.739	0.737	0.735	0.723	0.673
30	0.74	0.742	0.742	0.741	0.74	0.738	0.736	0.734	0.72	0.669
Promedio	0.741 ± 0.001	0.743 ± 0.001	0.743 ± 0.002	0.742 ± 0.001	0.74 ± 0.002	0.739 ± 0.002	0.737 ± 0.003	0.733 ± 0.003	0.718 ± 0.005	0.669 ± 0.006

Tabla A.10: Escenario 3: resultados de la medida-F para 10 valores de λ para Dvd-Book

					E - B (*baseline*=0.672)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.673	0.674	0.672	0.67	0.668	0.67	0.672	0.673	0.651	0.608
2	0.672	0.673	0.673	0.675	0.669	0.675	0.681	0.678	0.667	0.615
3	0.673	0.674	0.671	0.674	0.674	0.678	0.679	0.68	0.676	0.636
4	0.674	0.673	0.674	0.675	0.676	0.678	0.684	0.686	0.671	0.631
5	0.674	0.672	0.675	0.676	0.674	0.675	0.681	0.681	0.673	0.64
6	0.673	0.673	0.674	0.676	0.676	0.677	0.682	0.68	0.668	0.639
7	0.673	0.674	0.674	0.677	0.673	0.678	0.679	0.682	0.667	0.631
8	0.673	0.674	0.672	0.674	0.674	0.676	0.678	0.68	0.669	0.635
9	0.673	0.673	0.675	0.675	0.676	0.679	0.679	0.685	0.676	0.636
10	0.673	0.674	0.672	0.675	0.676	0.679	0.679	0.682	0.67	0.637
11	0.673	0.673	0.672	0.675	0.677	0.675	0.68	0.68	0.672	0.638
12	0.673	0.672	0.672	0.675	0.675	0.676	0.678	0.683	0.673	0.639
13	0.673	0.673	0.672	0.677	0.676	0.677	0.679	0.683	0.67	0.635
14	0.672	0.673	0.673	0.678	0.677	0.678	0.678	0.685	0.673	0.637
15	0.672	0.674	0.674	0.677	0.678	0.678	0.68	0.684	0.672	0.632
16	0.672	0.674	0.674	0.676	0.677	0.679	0.681	0.687	0.677	0.632
17	0.672	0.673	0.674	0.675	0.676	0.676	0.68	0.686	0.679	0.632
18	0.672	0.673	0.674	0.676	0.675	0.674	0.68	0.683	0.684	0.632
19	0.672	0.673	0.674	0.676	0.676	0.676	0.68	0.69	0.681	0.629
20	0.672	0.673	0.674	0.677	0.676	0.675	0.68	0.687	0.685	0.631
21	0.672	0.673	0.675	0.677	0.675	0.675	0.678	0.688	0.68	0.632
22	0.672	0.673	0.674	0.677	0.678	0.675	0.678	0.69	0.68	0.624
23	0.672	0.673	0.674	0.677	0.678	0.675	0.678	0.686	0.674	0.63
24	0.673	0.672	0.674	0.677	0.679	0.675	0.676	0.683	0.679	0.631
25	0.673	0.673	0.673	0.675	0.678	0.676	0.677	0.68	0.679	0.631
26	0.673	0.672	0.672	0.675	0.677	0.675	0.676	0.682	0.677	0.625
27	0.673	0.672	0.673	0.674	0.677	0.674	0.675	0.68	0.677	0.622
28	0.673	0.672	0.673	0.675	0.677	0.674	0.674	0.681	0.674	0.624
29	0.673	0.672	0.673	0.675	0.677	0.676	0.674	0.681	0.676	0.624
30	0.673	0.672	0.674	0.676	0.676	0.676	0.676	0.68	0.676	0.625
Promedio	0.673 ± 0.001	0.673 ± 0.001	0.673 ± 0.001	0.676 ± 0.002	0.676 ± 0.002	0.676 ± 0.002	0.678 ± 0.003	0.683 ± 0.004	0.674 ± 0.006	0.63 ± 0.007

Tabla A.11: Escenario 3: resultados de la medida-F para 10 valores de λ para Electronics-Book

					K - B ($baseline$=0.683)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.685	0.683	0.684	0.686	0.684	0.681	0.674	0.668	0.65	0.613
2	0.681	0.681	0.682	0.683	0.684	0.682	0.685	0.678	0.658	0.613
3	0.684	0.683	0.682	0.686	0.687	0.686	0.688	0.683	0.663	0.621
4	0.682	0.681	0.678	0.68	0.682	0.68	0.679	0.68	0.662	0.611
5	0.683	0.684	0.682	0.683	0.681	0.683	0.683	0.685	0.671	0.616
6	0.683	0.683	0.681	0.68	0.683	0.683	0.678	0.678	0.663	0.623
7	0.683	0.683	0.68	0.681	0.682	0.684	0.681	0.68	0.67	0.615
8	0.683	0.683	0.68	0.681	0.684	0.688	0.684	0.683	0.673	0.625
9	0.684	0.684	0.68	0.682	0.684	0.684	0.68	0.686	0.676	0.629
10	0.684	0.684	0.683	0.682	0.682	0.684	0.682	0.685	0.68	0.624
11	0.684	0.683	0.683	0.683	0.684	0.683	0.682	0.681	0.679	0.625
12	0.684	0.683	0.683	0.684	0.683	0.681	0.681	0.68	0.673	0.624
13	0.683	0.683	0.683	0.682	0.681	0.684	0.681	0.682	0.675	0.626
14	0.683	0.683	0.683	0.681	0.684	0.682	0.683	0.681	0.676	0.62
15	0.683	0.684	0.683	0.681	0.682	0.686	0.683	0.679	0.678	0.618
16	0.683	0.683	0.683	0.683	0.683	0.686	0.681	0.681	0.674	0.618
17	0.683	0.683	0.683	0.681	0.682	0.686	0.682	0.68	0.676	0.614
18	0.683	0.684	0.684	0.682	0.68	0.686	0.679	0.679	0.669	0.615
19	0.683	0.684	0.683	0.682	0.683	0.686	0.682	0.675	0.671	0.609
20	0.684	0.685	0.683	0.683	0.684	0.684	0.68	0.673	0.675	0.608
21	0.684	0.684	0.683	0.682	0.684	0.685	0.679	0.673	0.674	0.607
22	0.683	0.685	0.683	0.683	0.684	0.685	0.681	0.674	0.675	0.608
23	0.684	0.685	0.682	0.683	0.682	0.685	0.681	0.676	0.676	0.607
24	0.684	0.685	0.683	0.683	0.683	0.684	0.681	0.675	0.672	0.604
25	0.684	0.683	0.683	0.683	0.684	0.683	0.681	0.679	0.673	0.603
26	0.684	0.684	0.683	0.685	0.682	0.685	0.68	0.677	0.673	0.607
27	0.684	0.684	0.683	0.684	0.683	0.684	0.68	0.679	0.673	0.6
28	0.684	0.684	0.683	0.684	0.682	0.683	0.681	0.681	0.674	0.604
29	0.684	0.684	0.684	0.683	0.683	0.685	0.681	0.684	0.673	0.606
30	0.684	0.684	0.683	0.683	0.684	0.685	0.682	0.68	0.672	0.6
Promedio	0.683 ± 0.001	0.684 ± 0.001	0.682 ± 0.001	0.683 ± 0.001	0.683 ± 0.001	0.684 ± 0.002	0.681 ± 0.002	0.679 ± 0.004	0.672 ± 0.006	0.614 ± 0.008

Tabla A.12: Escenario 3: resultados de la medida-F para 10 valores de λ para Kitchen-Book

					B - D ($baseline$=0.732)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.731	0.734	0.73	0.733	0.729	0.729	0.721	0.704	0.669	0.637
2	0.733	0.735	0.732	0.732	0.729	0.73	0.718	0.713	0.689	0.653
3	0.734	0.733	0.731	0.731	0.732	0.73	0.719	0.716	0.699	0.651
4	0.732	0.732	0.73	0.732	0.731	0.726	0.724	0.717	0.698	0.655
5	0.732	0.732	0.729	0.731	0.732	0.734	0.723	0.716	0.692	0.653
6	0.733	0.732	0.729	0.731	0.732	0.732	0.726	0.72	0.699	0.65
7	0.731	0.732	0.729	0.729	0.731	0.733	0.728	0.718	0.702	0.644
8	0.732	0.733	0.73	0.728	0.729	0.728	0.723	0.717	0.696	0.64
9	0.731	0.732	0.731	0.733	0.727	0.729	0.723	0.717	0.701	0.643
10	0.731	0.732	0.732	0.73	0.728	0.73	0.724	0.722	0.703	0.638
11	0.732	0.732	0.731	0.729	0.727	0.73	0.723	0.718	0.698	0.628
12	0.732	0.731	0.73	0.729	0.728	0.729	0.726	0.72	0.696	0.625
13	0.732	0.731	0.731	0.727	0.726	0.727	0.727	0.719	0.698	0.626
14	0.732	0.731	0.732	0.727	0.726	0.729	0.728	0.719	0.694	0.628
15	0.732	0.732	0.731	0.726	0.728	0.728	0.728	0.716	0.698	0.619
16	0.732	0.732	0.733	0.727	0.726	0.727	0.727	0.718	0.697	0.617
17	0.732	0.732	0.732	0.728	0.726	0.727	0.729	0.716	0.693	0.605
18	0.732	0.732	0.732	0.727	0.726	0.725	0.727	0.718	0.692	0.613
19	0.732	0.732	0.732	0.728	0.727	0.726	0.727	0.716	0.688	0.612
20	0.732	0.732	0.732	0.729	0.728	0.725	0.727	0.718	0.689	0.608
21	0.732	0.732	0.732	0.728	0.726	0.726	0.725	0.718	0.683	0.602
22	0.732	0.731	0.732	0.728	0.726	0.726	0.724	0.718	0.684	0.6
23	0.732	0.731	0.731	0.729	0.727	0.725	0.725	0.717	0.685	0.597
24	0.732	0.732	0.732	0.727	0.727	0.725	0.725	0.717	0.687	0.598
25	0.732	0.731	0.731	0.728	0.727	0.724	0.724	0.716	0.688	0.597
26	0.732	0.731	0.733	0.729	0.727	0.725	0.725	0.716	0.69	0.598
27	0.732	0.731	0.732	0.729	0.729	0.726	0.725	0.716	0.691	0.596
28	0.732	0.731	0.732	0.728	0.728	0.727	0.726	0.714	0.692	0.591
29	0.732	0.73	0.731	0.729	0.727	0.727	0.727	0.717	0.69	0.591
30	0.732	0.731	0.731	0.728	0.729	0.727	0.727	0.714	0.688	0.591
Promedio	0.732 ± 0.001	0.732 ± 0.001	0.731 ± 0.001	0.729 ± 0.002	0.728 ± 0.002	0.728 ± 0.003	0.725 ± 0.003	0.717 ± 0.003	0.692 ± 0.007	0.62 ± 0.022

Tabla A.13: Escenario 3: resultados de la medida-F para 10 valores de λ para Book-Dvd

					E - D (*baseline*=0.672)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.673	0.671	0.675	0.677	0.673	0.673	0.666	0.656	0.64	0.609
2	0.674	0.675	0.676	0.674	0.679	0.678	0.67	0.66	0.657	0.615
3	0.675	0.675	0.676	0.673	0.678	0.682	0.681	0.677	0.659	0.637
4	0.676	0.675	0.674	0.675	0.678	0.681	0.678	0.676	0.66	0.641
5	0.675	0.676	0.675	0.677	0.678	0.679	0.679	0.678	0.672	0.643
6	0.675	0.674	0.673	0.674	0.676	0.681	0.681	0.676	0.67	0.645
7	0.674	0.676	0.674	0.674	0.679	0.681	0.68	0.676	0.67	0.651
8	0.674	0.674	0.673	0.675	0.675	0.678	0.68	0.677	0.67	0.648
9	0.674	0.675	0.673	0.676	0.676	0.678	0.681	0.68	0.672	0.649
10	0.674	0.676	0.673	0.674	0.674	0.675	0.679	0.678	0.67	0.645
11	0.674	0.675	0.675	0.673	0.674	0.675	0.679	0.68	0.669	0.645
12	0.674	0.675	0.676	0.674	0.673	0.676	0.675	0.676	0.673	0.642
13	0.674	0.676	0.673	0.674	0.672	0.675	0.673	0.678	0.673	0.639
14	0.674	0.675	0.673	0.672	0.67	0.673	0.67	0.677	0.67	0.644
15	0.674	0.675	0.672	0.673	0.671	0.672	0.67	0.675	0.672	0.64
16	0.674	0.675	0.671	0.672	0.67	0.671	0.669	0.676	0.674	0.636
17	0.674	0.675	0.672	0.672	0.671	0.671	0.673	0.676	0.675	0.637
18	0.674	0.675	0.673	0.672	0.673	0.671	0.675	0.677	0.674	0.629
19	0.674	0.674	0.673	0.672	0.674	0.674	0.676	0.68	0.672	0.627
20	0.674	0.674	0.672	0.672	0.672	0.672	0.675	0.675	0.668	0.625
21	0.674	0.674	0.674	0.673	0.672	0.674	0.673	0.677	0.671	0.63
22	0.674	0.674	0.673	0.673	0.674	0.672	0.675	0.677	0.672	0.624
23	0.674	0.674	0.674	0.673	0.673	0.674	0.674	0.674	0.674	0.625
24	0.674	0.674	0.673	0.672	0.673	0.674	0.674	0.674	0.674	0.627
25	0.674	0.674	0.673	0.672	0.672	0.673	0.676	0.674	0.669	0.623
26	0.674	0.674	0.673	0.672	0.671	0.672	0.674	0.672	0.663	0.621
27	0.674	0.674	0.673	0.672	0.672	0.672	0.673	0.672	0.666	0.62
28	0.674	0.675	0.673	0.673	0.673	0.672	0.673	0.671	0.664	0.621
29	0.674	0.675	0.672	0.672	0.673	0.672	0.671	0.674	0.666	0.62
30	0.674	0.675	0.673	0.673	0.673	0.673	0.672	0.674	0.665	0.624
Promedio	0.674 ± 0.001	0.675 ± 0.001	0.673 ± 0.001	0.673 ± 0.001	0.674 ± 0.003	0.675 ± 0.003	0.675 ± 0.004	0.675 ± 0.005	0.668 ± 0.007	0.633 ± 0.011

Tabla A.14: Escenario 3: resultados de la medida-F para 10 valores de λ para Electronics-Dvd

					K - D (*baseline*=0.684)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.683	0.683	0.681	0.681	0.681	0.681	0.682	0.672	0.647	0.607
2	0.683	0.686	0.683	0.679	0.682	0.681	0.681	0.672	0.661	0.626
3	0.682	0.683	0.682	0.681	0.681	0.682	0.684	0.68	0.673	0.643
4	0.682	0.682	0.685	0.684	0.681	0.683	0.681	0.678	0.676	0.634
5	0.683	0.681	0.684	0.687	0.682	0.685	0.682	0.685	0.68	0.647
6	0.683	0.682	0.687	0.686	0.685	0.682	0.682	0.682	0.686	0.644
7	0.683	0.682	0.687	0.686	0.686	0.683	0.681	0.682	0.68	0.649
8	0.683	0.682	0.683	0.682	0.685	0.683	0.68	0.682	0.679	0.642
9	0.683	0.683	0.683	0.682	0.683	0.684	0.684	0.681	0.679	0.645
10	0.684	0.684	0.683	0.683	0.683	0.684	0.684	0.681	0.675	0.636
11	0.685	0.684	0.683	0.683	0.682	0.683	0.682	0.682	0.677	0.637
12	0.685	0.684	0.683	0.684	0.681	0.684	0.68	0.683	0.676	0.642
13	0.684	0.685	0.685	0.682	0.682	0.684	0.682	0.679	0.674	0.636
14	0.684	0.684	0.685	0.683	0.681	0.683	0.681	0.678	0.672	0.631
15	0.684	0.684	0.685	0.682	0.681	0.683	0.683	0.678	0.671	0.631
16	0.684	0.684	0.683	0.682	0.681	0.684	0.682	0.676	0.671	0.628
17	0.684	0.684	0.683	0.682	0.681	0.683	0.68	0.676	0.666	0.616
18	0.685	0.684	0.683	0.681	0.68	0.684	0.682	0.675	0.669	0.616
19	0.684	0.684	0.682	0.681	0.681	0.681	0.682	0.681	0.666	0.615
20	0.684	0.683	0.681	0.68	0.679	0.681	0.682	0.68	0.666	0.615
21	0.685	0.684	0.681	0.68	0.679	0.681	0.682	0.677	0.665	0.614
22	0.685	0.684	0.682	0.68	0.679	0.681	0.681	0.677	0.662	0.608
23	0.685	0.684	0.682	0.68	0.682	0.681	0.681	0.678	0.665	0.61
24	0.685	0.684	0.682	0.68	0.681	0.681	0.679	0.678	0.662	0.605
25	0.685	0.684	0.682	0.681	0.681	0.681	0.681	0.682	0.662	0.602
26	0.685	0.684	0.682	0.681	0.68	0.681	0.681	0.683	0.663	0.602
27	0.685	0.684	0.682	0.681	0.681	0.68	0.681	0.681	0.666	0.601
28	0.685	0.684	0.682	0.681	0.679	0.679	0.68	0.682	0.663	0.599
29	0.685	0.684	0.683	0.681	0.68	0.68	0.681	0.68	0.663	0.595
30	0.685	0.684	0.683	0.681	0.681	0.68	0.68	0.681	0.665	0.593
Promedio	0.684 ± 0.001	0.684 ± 0.001	0.683 ± 0.002	0.682 ± 0.002	0.681 ± 0.002	0.682 ± 0.002	0.681 ± 0.001	0.679 ± 0.003	0.669 ± 0.008	0.622 ± 0.018

Tabla A.15: Escenario 3: resultados de la medida-F para 10 valores de λ para Kitchen-Dvd

					B - E ($baseline$=0.669)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.673	0.673	0.676	0.674	0.676	0.672	0.666	0.662	0.646	0.631
2	0.67	0.669	0.67	0.671	0.67	0.664	0.66	0.648	0.642	0.622
3	0.673	0.67	0.671	0.666	0.67	0.672	0.664	0.666	0.656	0.631
4	0.672	0.672	0.669	0.666	0.666	0.665	0.666	0.662	0.656	0.637
5	0.671	0.672	0.671	0.668	0.667	0.668	0.667	0.662	0.657	0.625
6	0.671	0.671	0.669	0.669	0.67	0.667	0.666	0.66	0.655	0.631
7	0.67	0.672	0.67	0.667	0.669	0.666	0.667	0.659	0.649	0.632
8	0.67	0.67	0.668	0.669	0.669	0.663	0.667	0.656	0.65	0.624
9	0.669	0.67	0.667	0.667	0.667	0.667	0.663	0.657	0.648	0.625
10	0.669	0.668	0.668	0.667	0.667	0.665	0.667	0.66	0.648	0.62
11	0.669	0.669	0.668	0.667	0.665	0.665	0.664	0.661	0.643	0.611
12	0.669	0.67	0.667	0.667	0.668	0.667	0.667	0.662	0.641	0.608
13	0.669	0.67	0.667	0.667	0.666	0.666	0.667	0.66	0.645	0.61
14	0.669	0.67	0.667	0.666	0.666	0.665	0.665	0.656	0.643	0.602
15	0.669	0.669	0.667	0.663	0.664	0.662	0.664	0.657	0.64	0.602
16	0.669	0.669	0.667	0.664	0.664	0.662	0.663	0.657	0.642	0.597
17	0.669	0.67	0.667	0.663	0.665	0.663	0.662	0.654	0.64	0.599
18	0.669	0.67	0.667	0.663	0.665	0.662	0.662	0.657	0.643	0.599
19	0.669	0.67	0.667	0.661	0.662	0.661	0.661	0.66	0.637	0.601
20	0.669	0.67	0.666	0.662	0.662	0.661	0.661	0.658	0.635	0.597
21	0.669	0.67	0.666	0.663	0.663	0.662	0.662	0.656	0.633	0.59
22	0.669	0.67	0.666	0.664	0.664	0.661	0.661	0.657	0.632	0.585
23	0.669	0.67	0.665	0.662	0.663	0.661	0.661	0.657	0.636	0.584
24	0.669	0.669	0.665	0.664	0.663	0.661	0.66	0.656	0.632	0.58
25	0.669	0.668	0.665	0.663	0.664	0.661	0.659	0.656	0.633	0.579
26	0.669	0.668	0.665	0.663	0.664	0.661	0.659	0.656	0.628	0.578
27	0.67	0.668	0.665	0.664	0.665	0.662	0.661	0.655	0.632	0.579
28	0.67	0.669	0.666	0.664	0.666	0.661	0.658	0.655	0.628	0.577
29	0.67	0.669	0.666	0.664	0.664	0.662	0.66	0.655	0.63	0.577
30	0.67	0.669	0.666	0.665	0.666	0.661	0.658	0.656	0.63	0.577
Promedio	0.67 ± 0.001	0.67 ± 0.001	0.667 ± 0.002	0.665 ± 0.003	0.666 ± 0.003	0.664 ± 0.003	0.663 ± 0.003	0.658 ± 0.003	0.641 ± 0.009	0.604 ± 0.02

Tabla A.16: Escenario 3: resultados de la medida-F para 10 valores de λ para Book-Electronics

					D - E ($baseline$=0.687)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.7	0.702	0.707	0.708	0.701	0.697	0.697	0.694	0.674	0.647
2	0.699	0.699	0.701	0.703	0.705	0.702	0.694	0.694	0.683	0.647
3	0.699	0.7	0.701	0.703	0.706	0.709	0.709	0.7	0.689	0.654
4	0.7	0.701	0.703	0.706	0.705	0.709	0.708	0.713	0.699	0.665
5	0.7	0.7	0.702	0.704	0.709	0.707	0.706	0.706	0.7	0.66
6	0.7	0.702	0.704	0.705	0.705	0.708	0.707	0.708	0.706	0.66
7	0.7	0.701	0.702	0.706	0.704	0.707	0.705	0.704	0.698	0.664
8	0.7	0.701	0.704	0.705	0.707	0.707	0.708	0.701	0.693	0.657
9	0.7	0.7	0.704	0.706	0.709	0.706	0.705	0.701	0.69	0.65
10	0.7	0.7	0.704	0.706	0.709	0.706	0.702	0.694	0.684	0.644
11	0.699	0.701	0.703	0.704	0.709	0.706	0.701	0.696	0.687	0.643
12	0.7	0.7	0.705	0.705	0.707	0.708	0.701	0.694	0.689	0.641
13	0.699	0.7	0.704	0.704	0.707	0.708	0.703	0.691	0.685	0.638
14	0.699	0.701	0.704	0.703	0.707	0.707	0.7	0.69	0.684	0.63
15	0.699	0.701	0.702	0.703	0.705	0.707	0.702	0.693	0.684	0.631
16	0.699	0.701	0.702	0.7	0.702	0.705	0.704	0.694	0.678	0.631
17	0.699	0.7	0.701	0.7	0.703	0.705	0.703	0.695	0.682	0.629
18	0.699	0.7	0.702	0.702	0.702	0.706	0.704	0.698	0.681	0.629
19	0.699	0.7	0.701	0.701	0.702	0.706	0.703	0.698	0.682	0.622
20	0.7	0.7	0.702	0.702	0.701	0.706	0.703	0.698	0.675	0.625
21	0.7	0.7	0.702	0.702	0.701	0.705	0.702	0.696	0.676	0.622
22	0.7	0.7	0.701	0.702	0.701	0.705	0.7	0.695	0.671	0.623
23	0.7	0.7	0.702	0.702	0.702	0.702	0.699	0.694	0.668	0.613
24	0.7	0.699	0.703	0.702	0.703	0.703	0.698	0.691	0.668	0.616
25	0.7	0.699	0.701	0.702	0.701	0.703	0.701	0.693	0.669	0.609
26	0.7	0.699	0.701	0.701	0.701	0.701	0.702	0.691	0.668	0.614
27	0.7	0.7	0.701	0.701	0.699	0.702	0.703	0.69	0.672	0.609
28	0.7	0.7	0.702	0.7	0.699	0.702	0.703	0.691	0.668	0.606
29	0.7	0.7	0.701	0.701	0.7	0.701	0.701	0.692	0.671	0.61
30	0.7	0.7	0.702	0.701	0.701	0.702	0.699	0.693	0.671	0.606
Promedio	0.7 ± 0	0.7 ± 0.001	0.702 ± 0.001	0.703 ± 0.002	0.704 ± 0.003	0.705 ± 0.003	0.702 ± 0.003	0.696 ± 0.006	0.681 ± 0.011	0.633 ± 0.019

Tabla A.17: Escenario 3: resultados de la medida-F para 10 valores de λ para Dvd-Electronics

					K - E (*baseline*=0.780)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.78	0.781	0.784	0.785	0.786	0.786	0.775	0.766	0.75	0.713
2	0.78	0.78	0.782	0.783	0.784	0.783	0.775	0.766	0.761	0.721
3	0.78	0.781	0.78	0.783	0.784	0.779	0.786	0.781	0.765	0.741
4	0.78	0.777	0.78	0.78	0.785	0.788	0.788	0.786	0.774	0.75
5	0.781	0.778	0.779	0.782	0.784	0.786	0.787	0.788	0.775	0.746
6	0.781	0.778	0.779	0.78	0.785	0.786	0.79	0.788	0.778	0.751
7	0.781	0.78	0.782	0.782	0.785	0.783	0.787	0.789	0.784	0.76
8	0.781	0.781	0.781	0.782	0.785	0.785	0.788	0.792	0.784	0.759
9	0.781	0.782	0.782	0.783	0.783	0.782	0.781	0.783	0.779	0.761
10	0.781	0.781	0.781	0.781	0.784	0.782	0.785	0.786	0.781	0.762
11	0.782	0.781	0.781	0.78	0.782	0.782	0.786	0.784	0.779	0.762
12	0.781	0.78	0.778	0.779	0.781	0.781	0.779	0.777	0.776	0.757
13	0.781	0.781	0.778	0.782	0.783	0.78	0.781	0.784	0.778	0.762
14	0.782	0.779	0.778	0.78	0.779	0.781	0.782	0.78	0.782	0.761
15	0.781	0.78	0.779	0.779	0.782	0.781	0.78	0.78	0.778	0.772
16	0.781	0.779	0.779	0.78	0.777	0.78	0.782	0.779	0.781	0.772
17	0.781	0.78	0.778	0.778	0.778	0.779	0.777	0.777	0.784	0.768
18	0.781	0.778	0.778	0.777	0.778	0.779	0.779	0.78	0.779	0.769
19	0.781	0.778	0.777	0.775	0.777	0.777	0.777	0.779	0.78	0.767
20	0.781	0.779	0.779	0.778	0.777	0.78	0.781	0.781	0.783	0.764
21	0.781	0.779	0.779	0.777	0.776	0.777	0.778	0.782	0.78	0.769
22	0.781	0.779	0.779	0.777	0.776	0.778	0.777	0.78	0.777	0.77
23	0.781	0.779	0.778	0.778	0.776	0.778	0.778	0.778	0.777	0.773
24	0.781	0.779	0.779	0.778	0.775	0.776	0.778	0.779	0.778	0.77
25	0.781	0.78	0.778	0.776	0.774	0.777	0.779	0.779	0.78	0.771
26	0.781	0.779	0.779	0.777	0.776	0.779	0.778	0.78	0.776	0.769
27	0.781	0.78	0.779	0.778	0.777	0.779	0.779	0.779	0.779	0.773
28	0.781	0.78	0.779	0.778	0.778	0.78	0.78	0.779	0.783	0.772
29	0.781	0.781	0.779	0.778	0.778	0.781	0.78	0.781	0.782	0.778
30	0.781	0.781	0.779	0.778	0.778	0.779	0.781	0.781	0.782	0.774
Promedio	0.781 ± 0	0.78 ± 0.001	0.779 ± 0.002	0.779 ± 0.002	0.78 ± 0.004	0.781 ± 0.003	0.781 ± 0.004	0.781 ± 0.005	0.778 ± 0.007	0.761 ± 0.015

Tabla A.18: Escenario 3: resultados de la medida-F para 10 valores de λ para Kitchen-Electronics

					B - K (*baseline*=0.729)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.733	0.737	0.738	0.745	0.747	0.744	0.734	0.724	0.704	0.671
2	0.731	0.733	0.734	0.736	0.74	0.742	0.746	0.742	0.727	0.686
3	0.731	0.732	0.734	0.731	0.738	0.744	0.742	0.743	0.732	0.701
4	0.73	0.731	0.732	0.732	0.734	0.736	0.737	0.736	0.731	0.709
5	0.73	0.731	0.731	0.733	0.736	0.739	0.74	0.736	0.737	0.71
6	0.73	0.73	0.731	0.733	0.736	0.738	0.739	0.734	0.733	0.709
7	0.729	0.732	0.732	0.735	0.735	0.735	0.742	0.739	0.734	0.707
8	0.73	0.731	0.731	0.733	0.736	0.735	0.739	0.737	0.728	0.712
9	0.73	0.73	0.731	0.732	0.734	0.736	0.738	0.736	0.732	0.703
10	0.729	0.729	0.731	0.732	0.733	0.732	0.737	0.733	0.728	0.706
11	0.729	0.73	0.732	0.733	0.731	0.733	0.735	0.729	0.723	0.698
12	0.729	0.73	0.731	0.732	0.731	0.733	0.733	0.726	0.72	0.695
13	0.729	0.729	0.734	0.732	0.732	0.732	0.735	0.729	0.726	0.706
14	0.729	0.73	0.73	0.731	0.731	0.731	0.731	0.732	0.726	0.7
15	0.729	0.729	0.731	0.732	0.732	0.728	0.73	0.731	0.726	0.699
16	0.729	0.729	0.73	0.732	0.732	0.729	0.731	0.731	0.723	0.699
17	0.729	0.729	0.73	0.732	0.733	0.73	0.731	0.73	0.723	0.697
18	0.729	0.729	0.731	0.731	0.732	0.731	0.729	0.731	0.717	0.698
19	0.729	0.729	0.73	0.732	0.732	0.729	0.729	0.728	0.716	0.692
20	0.729	0.729	0.73	0.732	0.733	0.73	0.733	0.729	0.718	0.691
21	0.729	0.729	0.73	0.73	0.732	0.729	0.733	0.732	0.718	0.697
22	0.729	0.729	0.73	0.732	0.73	0.734	0.733	0.732	0.714	0.694
23	0.729	0.729	0.731	0.731	0.729	0.731	0.731	0.729	0.714	0.689
24	0.729	0.729	0.729	0.731	0.73	0.731	0.73	0.728	0.714	0.687
25	0.729	0.729	0.729	0.73	0.729	0.731	0.729	0.726	0.716	0.689
26	0.729	0.729	0.729	0.731	0.731	0.731	0.73	0.726	0.716	0.688
27	0.729	0.729	0.728	0.731	0.73	0.73	0.729	0.724	0.717	0.686
28	0.729	0.729	0.728	0.73	0.732	0.73	0.728	0.724	0.718	0.678
29	0.729	0.73	0.728	0.73	0.731	0.728	0.729	0.72	0.716	0.683
30	0.729	0.73	0.728	0.731	0.729	0.728	0.728	0.722	0.716	0.681
Promedio	0.729 ± 0.001	0.73 ± 0.002	0.731 ± 0.002	0.732 ± 0.003	0.733 ± 0.004	0.733 ± 0.005	0.734 ± 0.005	0.731 ± 0.006	0.722 ± 0.008	0.695 ± 0.01

Tabla A.19: Escenario 3: resultados de la medida-F para 10 valores de λ para Book-Kitchen

					D - K ($baseline$=0.743)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.745	0.744	0.745	0.749	0.753	0.75	0.747	0.734	0.723	0.688
2	0.744	0.746	0.743	0.746	0.742	0.745	0.746	0.748	0.741	0.7
3	0.746	0.745	0.747	0.747	0.747	0.755	0.758	0.754	0.754	0.722
4	0.746	0.746	0.747	0.746	0.749	0.752	0.748	0.758	0.756	0.727
5	0.744	0.745	0.745	0.747	0.747	0.749	0.749	0.754	0.758	0.744
6	0.745	0.746	0.747	0.748	0.749	0.749	0.747	0.752	0.757	0.741
7	0.746	0.746	0.745	0.748	0.748	0.75	0.75	0.752	0.755	0.747
8	0.745	0.747	0.747	0.748	0.747	0.749	0.751	0.754	0.757	0.746
9	0.744	0.747	0.747	0.745	0.746	0.747	0.749	0.753	0.755	0.743
10	0.745	0.745	0.746	0.746	0.743	0.744	0.746	0.75	0.753	0.74
11	0.744	0.747	0.745	0.746	0.746	0.747	0.743	0.749	0.75	0.737
12	0.743	0.746	0.744	0.743	0.743	0.745	0.743	0.75	0.75	0.737
13	0.744	0.746	0.745	0.744	0.743	0.743	0.743	0.747	0.749	0.739
14	0.744	0.745	0.746	0.745	0.743	0.744	0.743	0.749	0.746	0.739
15	0.744	0.744	0.745	0.744	0.744	0.745	0.741	0.746	0.744	0.746
16	0.743	0.744	0.745	0.745	0.746	0.746	0.742	0.748	0.746	0.739
17	0.744	0.744	0.744	0.744	0.744	0.745	0.741	0.747	0.746	0.736
18	0.744	0.744	0.744	0.744	0.744	0.746	0.742	0.747	0.748	0.734
19	0.744	0.744	0.744	0.743	0.745	0.743	0.743	0.749	0.747	0.726
20	0.743	0.744	0.744	0.743	0.744	0.744	0.743	0.746	0.746	0.729
21	0.743	0.744	0.743	0.743	0.745	0.744	0.743	0.746	0.748	0.73
22	0.743	0.744	0.743	0.743	0.744	0.743	0.744	0.746	0.746	0.728
23	0.743	0.744	0.743	0.745	0.745	0.743	0.744	0.745	0.741	0.728
24	0.743	0.744	0.744	0.744	0.744	0.744	0.744	0.745	0.742	0.73
25	0.743	0.744	0.743	0.744	0.744	0.742	0.744	0.746	0.742	0.732
26	0.743	0.744	0.743	0.745	0.743	0.745	0.745	0.745	0.743	0.73
27	0.743	0.744	0.743	0.744	0.744	0.742	0.746	0.744	0.746	0.73
28	0.743	0.745	0.744	0.744	0.744	0.743	0.745	0.742	0.743	0.732
29	0.743	0.745	0.744	0.744	0.744	0.745	0.746	0.742	0.742	0.726
30	0.743	0.745	0.743	0.743	0.743	0.744	0.744	0.742	0.741	0.726
Promedio	0.744 ± 0.001	0.745 ± 0.001	0.745 ± 0.001	0.745 ± 0.002	0.745 ± 0.002	0.746 ± 0.003	0.745 ± 0.004	0.748 ± 0.005	0.747 ± 0.007	0.732 ± 0.012

Tabla A.20: Escenario 3: resultados de la medida-F para 10 valores de λ para Dvd-Kitchen

					E - K ($baseline$=0.804)					
k	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 0.7$	$\lambda = 0.6$	$\lambda = 0.5$	$\lambda = 0.4$	$\lambda = 0.3$	$\lambda = 0.2$	$\lambda = 0.1$	$\lambda = 0.0$
1	0.809	0.809	0.812	0.812	0.816	0.814	0.806	0.786	0.757	0.709
2	0.806	0.809	0.812	0.813	0.814	0.816	0.811	0.803	0.78	0.739
3	0.806	0.809	0.815	0.812	0.813	0.816	0.815	0.805	0.789	0.753
4	0.808	0.808	0.811	0.816	0.82	0.821	0.825	0.817	0.805	0.77
5	0.809	0.809	0.812	0.817	0.82	0.822	0.823	0.822	0.811	0.784
6	0.807	0.808	0.811	0.813	0.815	0.818	0.819	0.82	0.81	0.786
7	0.807	0.808	0.81	0.812	0.814	0.819	0.822	0.818	0.811	0.782
8	0.807	0.807	0.811	0.813	0.816	0.82	0.821	0.819	0.814	0.787
9	0.806	0.808	0.809	0.812	0.813	0.816	0.822	0.82	0.813	0.788
10	0.806	0.808	0.809	0.811	0.812	0.817	0.819	0.823	0.816	0.791
11	0.806	0.809	0.811	0.811	0.814	0.817	0.819	0.825	0.819	0.787
12	0.806	0.809	0.809	0.811	0.813	0.815	0.821	0.824	0.818	0.796
13	0.806	0.808	0.808	0.81	0.811	0.815	0.821	0.818	0.819	0.795
14	0.806	0.807	0.808	0.811	0.811	0.814	0.816	0.817	0.823	0.797
15	0.805	0.807	0.808	0.81	0.811	0.813	0.817	0.818	0.819	0.797
16	0.806	0.807	0.809	0.81	0.811	0.813	0.814	0.817	0.818	0.799
17	0.805	0.807	0.809	0.812	0.81	0.812	0.815	0.818	0.817	0.803
18	0.806	0.807	0.809	0.81	0.81	0.812	0.816	0.816	0.815	0.8
19	0.805	0.806	0.808	0.81	0.81	0.813	0.816	0.816	0.816	0.8
20	0.805	0.807	0.807	0.81	0.812	0.812	0.815	0.818	0.819	0.803
21	0.806	0.807	0.807	0.809	0.811	0.814	0.814	0.817	0.819	0.802
22	0.805	0.806	0.807	0.808	0.811	0.813	0.815	0.817	0.82	0.802
23	0.805	0.806	0.807	0.808	0.811	0.812	0.813	0.814	0.815	0.804
24	0.805	0.806	0.806	0.808	0.811	0.812	0.814	0.814	0.814	0.807
25	0.805	0.806	0.806	0.807	0.81	0.812	0.814	0.813	0.814	0.802
26	0.805	0.806	0.806	0.806	0.809	0.812	0.813	0.815	0.815	0.803
27	0.804	0.806	0.806	0.806	0.808	0.81	0.813	0.816	0.811	0.803
28	0.804	0.806	0.806	0.807	0.808	0.811	0.813	0.812	0.814	0.802
29	0.804	0.806	0.806	0.806	0.808	0.811	0.813	0.815	0.816	0.805
30	0.804	0.805	0.807	0.807	0.808	0.81	0.814	0.815	0.816	0.804
Promedio	0.806 ± 0.001	0.807 ± 0.001	0.809 ± 0.002	0.81 ± 0.003	0.812 ± 0.003	0.814 ± 0.003	0.816 ± 0.004	0.816 ± 0.007	0.811 ± 0.013	0.79 ± 0.022

Tabla A.21: Escenario 3: resultados de la medida-F para 10 valores de λ para Electronics-Kitchen

Apéndice

Artículos publicados

Los artículos derivados de esta tesis se lista a continuación.

- *Enhancing Text Classification by Information Embedded in the Test Set*. Gabriela Ramírez-de-la-Rosa, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda. 11th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing-2010. Romania, March, 2010. Lecture Notes in Computer Science 6008, Springer, 2010.

- *Using Information from the Target Language to Improve Crosslingual Text Classification*. Gabriela Ramírez-de-la-Rosa, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, David Pinto-Avendaño, Thamar Solorio. 7th International Conference on Natural Language Processing IceTAL-2010, Reykjavik, Iceland, August 2010. Lecture Notes in Computer Science, volume 6233, 2010.

Referencias

Aas, K., y Eikvil, L. 1999. Text categorisation: A survey.

Araujo, B. S. 2006. *Aprendizaje Automático: conceptos básicos y avanzados*. Pearson Prentice Hall.

Aue, A., y Gamon, M. 2005. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of RANLP*.

Baeza-Yates, R. A., y Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; y Vaughan, J. 2010. A theory of learning from different domains. *Machine Learning* 79:151–175. 10.1007/s10994-009-5152-4.

Blitzer, J.; Dredze, M.; y Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*.

Cardoso-Cachopo, A., y Oliveira, A. L. 2007. Semi-supervised single-label text categorization using centroid-based classifiers. In *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing*, 844–851. ACM.

Cong, G.; Lee, W. S.; Wu, H.; y Liu, B. 2004. Semi-supervised text classification using partitioned em. In *11 th Int. Conference on Database Systems for Advanced Applications (DASFAA)*, 229–239.

Cortes, C., y Vapnik, V. 1995. Support vector networks. *Machine Learning* 20(3):273–297.

Dai, W.; Xue, G.-R.; Yang, Q.; y Yu, Y. 2007. Transferring naive bayes classifiers for text classification. In *AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence*, 540–545. AAAI Press.

Debole, F., y Sebastiani, F. 2003. Supervised term weighting for automated text categorization. In *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, 784–788. ACM Press.

Derivaux, S.; Forestier, G.; y Wemmert, C. 2008. Improving supervised learning with multiple clusterings. In *Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications in conjunction with ECAI*, 57–60.

Driessens, K.; Reutemann, P.; Pfahringer, B.; y Leschi, C. 2006. Using weighted nearest neighbor to benefit from unlabeled data. In *PAKDD*, 60–69.

Escobar-Acevedo, A.; Montes-Y-Gómez, M.; y Villaseñor-Pineda, L. 2009. Using nearest neighbor information to improve cross-language text classification. In *MICAI*, 157–164.

Fang, Y. C.; Parthasarathy, S.; y Schwartz, F. 2001. Using clustering to boost text classification. In *Workshop on Text Mining (TextDM'2001)*.

Gliozzo, A., y Strapparava, C. 2005. Cross language text categorization by acquiring multilingual domain models from comparable corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, 9–16.

Guzmán-Cabrera, R.; Montes-Y-Gómez, M.; Rosso, P.; y Villaseñor-Pineda, L. 2009. Using the Web as corpus for self-training text categorization. *Information Retrieval* 12(3):400–415.

Han, E.-H., y Karypis, G. 2000. Centroid-based document classification: Analysis and experimental results. In *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, 424–431. London, UK: Springer-Verlag.

Jensen, D.; Neville, J.; y Gallagher, B. 2004. Why collective inference improves relational classification. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 593–598. ACM.

Joachims, T. 1997. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, 143–151. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Kanji, G. K. 2006. *100 Statistical tests*. SAGE Publications Ltd.

Kyriakopoulou, A., y Kalamboukis, T. 2007. Using clustering to enhance text classification. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 805–806. New York, USA: ACM.

Lertnattee, V., y Theeramunkong, T. 2003. Term-length normalization for centroid-based text categorization. In *Knowledge-Based Intelligent Information and Engineering Systems*, 850–856.

Lertnattee, V., y Theeramunkong, T. 2004. Effect of term distributions on centroid-based text categorization. *Inf. Sci. Inf. Comput. Sci.* 158(1):89–115.

Lewis, D. D.; Yang, Y.; Rose, T. G.; y Li, F. 2004. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* 5:361–397.

Lewis, D. D. 1991. Evaluating text categorization. In *In Proceedings of Speech and Natural Language Workshop*, 312–318. Morgan Kaufmann.

Ling, X.; Xue, G.-R.; Dai, W.; Jiang, Y.; Yang, Q.; y Yu, Y. 2008. Can Chinese web pages be classified with English data source? In *WWW '08: Proceeding of the 17th International Conference on World Wide Web*, 969–978. New York, NY, USA: ACM.

Liu, Y.; Loh, H. T.; Youcef-Toumi, K.; y Tor, S. B. 2007. Handling of imbalanced data in text classification: Category-based term weights. In *Natural Language Processing and Text Mining*, 171–192. Springer London.

Mitchell, T. M. 1997. *Machine Learning*. McGraw-Hill.

Nigam, K.; Mccallum, A.; Thrun, S.; y Mitchell, T. 1999. Text classification from labeled and unlabeled documents using em. In *Machine Learning*, 103–134.

Ning, X., y Karypis, G. 2009. The set classification problem and solution methods. In *SDM*, 847–858.

Raskutti, B.; Ferrá, H.; y Kowalczyk, A. 2002. Combining clustering and co-training to enhance text classification using unlabelled data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 620–625. ACM.

Research, M. M. 2010. Internet world stats. http://www.internetworldstats.com/stats.htm.

Rigutini, L.; Maggini, M.; y Liu, B. 2005. An EM based training algorithm for cross-language text categorization. In *WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, 529–535. Washington, DC, USA: IEEE Computer Society.

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

Sebastiani, F. 2005. Text categorization. In Zanasi, A., ed., *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*. Southampton, UK: WIT Press. 109–129.

Tan, S.; Wu, G.; Tang, H.; y Cheng, X. 2007. A novel scheme for domain-transfer problem in the context of sentiment analysis. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 979–982. New York, NY, USA: ACM.

Tan, S. 2008. An improved centroid classifier for text categorization. *Expert Systems with Applications* 35(1-2):279–285.

Tishby, N.; Pereira, F. C.; y Bialek, W. 1999. The information bottleneck method. 368–377.

Wan, X. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 235–243. Suntec, Singapore: Association for Computational Linguistics.

Weston, J., y Watkins, C. 1999. Support vector machines for multi-class pattern recognition. In *Proceedings of the 6th European Symposium On Artificial Neural Networks*.

Witten, I., y Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. San Fransisco, CA, USA: Morgan Kaufmann, 2nd edition.

Wong, W.-C., y Fu, A. W.-C. 2000. Incremental document clustering for web page classification.

Wu, S., y Flach, P. A. 2002. Feature selection with labelled and unlabelled data. In *Proceedings of ECML/PKDD'02 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, 156–167.

Yang, Y., y Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, 412–420. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Zhen, Y., y Li, C. 2008. Cross-domain knowledge transfer using semi-supervised classification. In *AI '08: Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence*, 362–371. Berlin, Heidelberg: Springer-Verlag.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; y Schölkopf, B. 2004. Learning with local and global consistency. In Thrun, S.; Saul, L.; y Schölkopf, B., eds., *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

Zhu, X. 2005. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.

Álvarez Romero, J. d. D. 2009. Clasificación automática de textos usando reducción de clases basada en prototipos. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).

Özgür, A.; Alpaydın, P. E.; Prof, A.; y Güngör, T. 2004. Supervised and unsupervised machine learning techniques for text document categorization. Technical report.

Agradecimiento especial para: mis asesores Manuel Montes y Luis Villaseñor
mis revisores los doctores: Enrique Sucar, Eduardo Morales y Francisco Martínez
y el soporte económico otorgado por el CONACyT a través de la beca 239516.