



UniversidadeVigo

Marcos Antonio Mourino García
TESIS DOCTORAL
*Clasificación multilingüe de documentos
utilizando machine learning y la Wikipedia*
2017

TESIS DOCTORAL

*Clasificación multilingüe de documentos
utilizando machine learning y la Wikipedia*

Marcos Antonio Mourino García

2017

UniversidadeVigo

EIDO
Escola Internacional
de Doutoramento

Universidade de Vigo

Escola Internacional de Doutoramento

Marcos Antonio Mouriño García

TESIS DOCTORAL

Clasificación multilingüe de documentos utilizando machine learning
y la Wikipedia.

Dirigida por los doctores:

Luis E. Anido Rifón y Roberto Pérez Rodríguez

Año: 2017

A mi familia de la Casilla.

A Tania.

Agradecimientos

A mis padres, hermano, abuelos, tíos y primos, por su interés y apoyo en todo momento.

A mis directores, Luis y Roberto, por sus fundamentales aportes, críticas y sugerencias, así como por su apoyo, seguimiento, dedicación y orientación durante este viaje.

A todos mis compañeros y amigos, los cuales no han dudado un segundo en prestarme su ayuda cuando así lo he necesitado.

A Tania, por su apoyo y comprensión, por estar siempre a mi lado y por no dejarme nunca caminar solo.

Este momento no habría llegado de no ser por todos vosotros.

¡Muchas gracias!

Resumen

La gran cantidad de información textual disponible en la red, junto con el aumento de la demanda por parte de los usuarios, hace necesaria la existencia de sistemas que permitan un acceso a aquella información de interés de una forma eficiente y efectiva, ahorrando así tiempo en su búsqueda y consulta. Entre las técnicas existentes para proporcionar acceso o facilitar la gestión de información, este trabajo se centra en la clasificación de documentos, concretamente, en la clasificación automática de documentos de texto utilizando técnicas de aprendizaje máquina o *machine learning*. La clasificación automática de documentos utilizando técnicas de aprendizaje máquina es muy útil, y cuenta con un elevado número de aplicaciones en diferentes ámbitos.

Los algoritmos en los que se basan los sistemas de clasificación automática requieren que los documentos estén representados de forma que puedan entenderlos y/o relacionarlos, siendo la representación más ampliamente utilizada el modelo bolsa de palabras (*Bag of Words*, BoW). A pesar de ser una de las representaciones más utilizadas, este modelo no es óptimo, puesto que no tiene en cuenta la semántica de las palabras ni las relaciones semánticas entre ellas, siendo de esta manera vulnerable a problemas del lenguaje, como la sinonimia o la polisemia, que afectan al rendimiento de la clasificación. Este inconveniente se ve agravado en las tareas de clasificación multilingüe, las cuales han sido abordadas tradicionalmente utilizando la combinación del modelo bolsa de palabras y las técnicas de traducción automática de documentos. Estas últimas presentan una serie de desventajas que afectan de forma negativa a la calidad de las traducciones, y por ende, a la calidad de la clasificación. Por consiguiente, cuando la representación BoW se combina con la utilización de técnicas de traducción automática, las desventajas de ambas propuestas se suman, lo que conduce a un incremento de la probabilidad de error del clasificador.

Con el objetivo de mitigar los problemas de las propuestas tradicionales para la clasificación monolingüe y multilingüe de documentos, en este trabajo se explora el uso de una representación de los documentos en forma de bolsa de conceptos (*Bag of Concepts*, BoC) de la Wikipedia (WikiBoC), obtenidos a través del anotador semántico de propósito general Wikipedia Miner, para la clasificación monolingüe y multilingüe de documentos de texto en diferentes ámbitos de aplicación. Para demostrar la aplicabilidad y beneficios aportados por la representación WikiBoC de los documentos, se han realizado diversos experimentos de clasificación monolingüe y multilingüe utilizando la propuesta presentada

y las propuestas más relevantes presentes en el estado del arte, y se ha realizado el análisis y la evaluación comparativa del rendimiento de cada una de las propuestas.

La principal contribución de esta tesis es un modelo para la clasificación monolingüe y multilingüe de documentos de texto, pertenecientes a diversos ámbitos de aplicación, que hace uso del conocimiento enciclopédico contenido en la Wikipedia para crear representaciones de los documentos basadas en conceptos, proporcionando mejores resultados que las propuestas presentes en el estado del arte. Esta contribución principal se divide a su vez en una serie de contribuciones menores, como son el análisis y la evaluación comparativa de la aplicación de la representación de los documentos WikiBoC y de las representaciones más relevantes del estado del arte a la clasificación monolingüe y multilingüe de documentos y la creación de conjuntos de datos más allá de los presentes en el estado del arte para obtener una evaluación exhaustiva de la propuesta presentada.

Los resultados obtenidos en este trabajo nos permiten concluir que el uso de la representación WikiBoC en las tareas de clasificación monolingüe y multilingüe es ventajoso, ya que los conceptos extraídos por el anotador semántico Wikipedia Miner proporcionan información muy relevante para el algoritmo de clasificación. Este comportamiento es especialmente significativo cuando los datos disponibles para entrenar el algoritmo de clasificación son escasos, y cuando los documentos involucrados en el problema de clasificación tratan sobre cuestiones o temas biomédicos.

Palabras clave

Clasificación automática de documentos, aprendizaje automático, bolsa de palabras, bolsa de conceptos, Wikipedia, Wikipedia Miner

Abstract

The large amount of textual information available on the Internet, together with the increasing demand by users, makes it necessary the existence of systems that allow users to access to those information that is of their interest in an efficient and effective way, thus saving time searching and querying. Among the existing techniques to provide access or to facilitate the management of information, this research focuses on the classification of documents, particularly, the automatic classification of text documents using machine learning techniques. The automatic classification of documents using machine learning techniques is highly convenient, and it has a large number of applications in several different areas.

Text documents have to be represented in a way that the classification algorithms can understand and relate them, being the most widely used representation the Bag of Words (BoW) paradigm. Despite being one of the most used representations in text classification tasks, this model is not optimal, since it does not take into account the semantics of words and the semantic relations between them, causing the appearance of language problems that affect classification performance. This drawback is aggravated in multilingual classification tasks, which have traditionally been approached by using a combination of the Bag of Words model and automatic translation techniques. The latter present a number of disadvantages that negatively affect the quality of translations, and therefore, the quality of classification. Thus, when the BoW representation is combined with machine translation techniques, the disadvantages of each proposal add up, which leads to an increased error probability.

With the aim of mitigating the problems of the traditional proposals for the monolingual and multilingual classification of documents, this research explores the use of a Wikipedia-based Bag of Concepts (WikiBoC) representation of documents, being these concepts extracted using the general purpose semantic annotator Wikipedia Miner, for the monolingual and multilingual classification of text documents in different application fields. In order to demonstrate the applicability and the benefits provided by the WikiBoC representation of documents, we performed different monolingual and multilingual classification experiments using the approach presented, analysed the results obtained, and compared them with the results obtained using the most relevant state-of-the-art proposals.

The main contribution made by this thesis is a model for monolingual and multilingual classifying text documents, belonging to different fields of application, that leverages the encyclopaedic knowledge contained in Wikipedia to create concept-based representations of documents, obtaining better results than state-of-the-art proposals. This contribution is in turn divided into a series of smaller contributions, such as the analysis and benchmarking of WikiBoC against other state-of-the-art document representations to perform monolingual and multilingual classification of documents, and the creation of datasets beyond those present in the state-of-the-art in order to perform a comprehensive evaluation of the proposal presented.

The results obtained in this research allow us to conclude that the use of the WikiBoC representation in monolingual and multilingual text classification is advantageous, since the concepts extracted through the Wikipedia Miner semantic annotator provide very relevant information to the classification algorithm. This behaviour is especially significant when the data available to train the classification algorithm are scarce, and when the documents involved in the classification problem are about biomedical topics.

Keywords

Automatic document classification, machine learning, Bag of Words, Bag of Concepts, Wikipedia, Wikipedia Miner

Índice de contenidos

1. Introducción	1
1.1. Introducción	1
1.2. Motivación	3
1.3. Hipótesis	4
1.4. Objetivos	4
1.5. Metodología	5
1.6. Publicaciones	8
1.7. Estructura del documento	9
2. Materiales y métodos	11
2.1. Aprendizaje máquina	11
2.2. Clasificación de documentos	13
2.3. Support Vector Machines	15
2.4. Métricas de evaluación	15
2.5. Representación de los documentos	17
2.6. Wikipedia	21
2.7. Wikipedia Miner	21
2.8. Modelos empleados para la representación de los documentos	26
2.9. Conjuntos de datos	33

3. Resultados y discusión	37
3.1. La clasificación automática de documentos y la interoperabilidad entre repositorios de recursos educativos	37
3.2. La representación WikiBoC y la clasificación monolingüe	41
3.3. La representación WikiBoC y la clasificación multilingüe	47
4. Contribuciones, conclusiones y trabajos futuros	57
4.1. Contribuciones	57
4.2. Conclusiones	58
4.3. Trabajos futuros	59
4. Contributions, conclusions and future work	61
4.1. Contributions	61
4.2. Conclusions	62
4.3. Future work	63
Apéndice I: Publicaciones	77
Cross-repository aggregation of educational resources	79
Biomedical literature classification using encyclopedic knowledge: a Wikipedia-based bag-of-concepts approach	101
Wikipedia-based cross-language text classification	125
A Bag of Concepts Approach for Biomedical Document Classification Using Wikipedia Knowledge: Spanish-English Cross-language Case Study	145

Índice de figuras

1.1. Proceso de investigación seguido en la tesis.	7
2.1. Aprendizaje supervisado.	13
2.2. Arquitectura del anotador semántico Wikipedia Miner (extraída del trabajo de Milne and Witten [67]).	22
2.3. Proceso de obtención de la bolsa de conceptos de un documento de texto utilizando Wikipedia Miner.	24
2.4. Ejemplo de ejecución del servicio <i>wikify</i>	25
2.5. Representación de un documento según el modelo BoW.	26
2.6. Representación de un documento según el modelo WikiBoC.	27
2.7. Representación de un documento según el modelo BoW-MT.	28
2.8. Representación de un documento según el modelo ESA.	28
2.9. Representación de un documento según el modelo Bi-LDA.	29
2.10. Representación de un documento según el modelo BWE.	30
2.11. Representación de un documento según el modelo MetaMap.	30
2.12. Representación de un documento según el modelo Hybrid-WikiBoC.	30
2.13. <i>Cross-Language Concept Matching</i>	31
2.14. Conversión de la representación WikiBoC de un documento escrito en castellano a su equivalente en inglés.	32

Índice de figuras

2.15. Representación de un documento según el modelo WikiBoC-CLCM.	33
3.1. Funcionamiento de CROERA.	40
3.2. Arquitectura del clasificador SVM utilizando la representación WikiBoC.	42
3.3. $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus OHSUMED (etiqueta única).	43
3.4. $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus OHSUMED (multietiqueta).	44
3.5. Incremento del rendimiento (en %) de la representación WikiBoC sobre la representación BoW para los corpus OHSUMED y UVigoMED.	44
3.6. $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus UVigoMED (etiqueta única).	45
3.7. $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus UVigoMED (multietiqueta).	46
3.8. Arquitectura del clasificador SVM inglés-castellano utilizando la representación WikiBoC-CLCM.	49
3.9. Arquitectura del clasificador SVM inglés-castellano utilizando la representación Hybrid-WikiBoC.	50
3.10. $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus Wikipedia Corpus.	51
3.11. $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus Wikipedia Human Medicine.	52
3.12. $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus JRC-Acquis.	53
3.13. $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus CL-UVigoMED (etiqueta única).	54
3.14. $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus CL-UVigoMED (multietiqueta).	55

Índice de tablas

2.1.	Número de ocurrencias de cada palabra en cada uno de los documentos.	18
2.2.	Peso o relevancia de cada concepto en cada uno de los documentos.	20
3.1.	Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus OHSUMED (etiqueta única).	43
3.2.	Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus OHSUMED (multietiqueta).	43
3.3.	Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus UVigoMED (etiqueta única).	45
3.4.	Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus UVigoMED (multietiqueta).	46
3.5.	$F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus Wikipedia Corpus.	51
3.6.	$F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus Wikipedia Human Medicine.	52
3.7.	$F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus JRC-Acquis.	53
3.8.	Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus CL-UVigoMED (etiqueta única).	55
3.9.	Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus CL-UVigoMED (multietiqueta).	56

Lista de acrónimos

Bi-LDA	Bilingual Latent Dirichlet Allocation
BoC	Bag of Concepts - bolsa de conceptos
BoW	Bag of Words - bolsa de palabras
BoW-MT	Bag of Words and Machine Translation - bolsa de palabras y traducción automática
BWE	Bilingual Word Embeddings
CL-UVigoMED	Cross-Language UVigoMED
CLCM	Cross-Language Concept Matching
DSRM	Design Science Research Methodology
ESA	Explicit Semantic Analysis
FN	False Negative - falso negativo
FP	False Positive - falso positivo
GFS	Google File System
Hybrid-WikiBoC	WikiBoC and BoW - WikiBoC y BoW
LDA	Latent Dirichlet Allocation
ML	Machine Learning - aprendizaje máquina
MT	Machine Translation - traducción automática
P	Precision - precisión
PLN	Procesado del Lenguaje Natural
R	Recall - retirada

REA	Recurso Educativo Abierto
SVM	Support Vector Machines - máquinas de vectores de soporte
TN	True Negative - verdadero negativo
TP	True Positive - verdadero positivo
UMLS	Unified Medical Language System
UVigoMED	University of Vigo MEDLINE
VSM	Vector Space Model - modelo de espacio vectorial
WE	Word Embeddings
WikiBoC	Wikipedia Bag of Concepts - bolsa de conceptos Wikipedia
WikiBoC-CLCM	Wikipedia Bag of Concepts and the CLCM technique - bolsa de conceptos Wikipedia y la técnica CLCM
WM	Wikipedia Miner
XML	eXtensible Markup Language

Capítulo 1

Introducción

Este capítulo introduce la tesis. Se presenta en primer lugar la clasificación automática de documentos y su relevancia, así como las principales propuestas en el estado del arte para la realización de clasificación monolingüe y multilingüe de documentos de texto. Tras esto, se describen los principales inconvenientes y desventajas de las propuestas tradicionales, que lastran el rendimiento de los sistemas de clasificación. A continuación, se presenta la hipótesis de investigación y los objetivos a alcanzar, así como la metodología utilizada para la realización de esta tesis. El capítulo continúa con la presentación de las cuatro publicaciones derivadas de la realización de esta investigación. Finalmente, este capítulo termina con una descripción del resto de la tesis.

1.1. Introducción

La sociedad de la información y la comunicación conlleva la existencia de una gran cantidad de información textual disponible en la red, la cual es creada de forma continua, a través de diversas fuentes y en diferentes idiomas [31]. La demanda de información por parte de los usuarios aumenta día tras día, lo que hace necesaria la existencia de sistemas que permitan a los usuarios acceder a aquella información de su interés de una manera sencilla, rápida, eficiente y efectiva, ahorrando así tiempo en su búsqueda y consulta [20]. Tres de las principales técnicas utilizadas para la gestión de información de forma automática son la búsqueda y recuperación de información, el *clustering* de documentos y la clasificación de documentos. La búsqueda y recuperación de información [16, 105] es un subcampo de las ciencias de la computación que consiste en la obtención automática de recursos relevantes para satisfacer una necesidad de información acerca de un tema en concreto dentro de una colección de recursos de información. El *clustering* o agrupamiento [1] consiste en agrupar un conjunto de objetos de tal manera que aquellos que pertenecen al mismo grupo (*cluster*) guardan más similitud (en un sentido u otro) entre ellos que

con cualquiera de los objetos pertenecientes al resto de los grupos. La clasificación se define como la acción y efecto de ordenar o dividir un conjunto de elementos en clases a partir de un criterio determinado [93]. Así, podemos definir la clasificación automática de documentos como la asignación algorítmica de documentos a un conjunto de categorías o clases previamente definido [108]. La clasificación automática de documentos de texto es extremadamente útil y cuenta con un elevado número de aplicaciones en diferentes ámbitos. Como ejemplos, podemos citar el análisis de sentimientos [57], el filtrado de correo basura [123], la clasificación de recursos educativos en materias [71], o la clasificación de noticias en sus respectivas secciones en un periódico [62].

La clasificación automática de documentos se puede modelar como un problema de aprendizaje máquina. El aprendizaje máquina, o aprendizaje automático – *machine learning* (ML) – es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a los ordenadores aprender. Concretamente, la clasificación automática de documentos se puede modelar de tres formas diferentes en función del conocimiento existente a priori: como un problema de aprendizaje máquina supervisado [12, 108], no supervisado [44] o semi-supervisado [143]. En esta investigación nos hemos centrado en la clasificación de documentos modelada como un problema de aprendizaje supervisado. En primer lugar, un algoritmo de clasificación se entrena con un cierto número de ejemplos – documentos cuya categoría es conocida – y posteriormente, el algoritmo entrenado se aplica sobre otro conjunto de documentos cuya categoría es desconocida [108].

La representación de los documentos – tanto aquellos utilizados para entrenar el algoritmo como aquellos que van a ser clasificados – y el conjunto de elementos utilizado para entrenar el algoritmo tienen un papel fundamental en el rendimiento de un sistema de clasificación automática de documentos. Por lo tanto, se hace necesario utilizar las mejores secuencias de entrenamiento y representaciones de los documentos posibles con el fin de maximizar el rendimiento del clasificador.

Por un lado, los algoritmos de clasificación requieren que los documentos estén representados de una manera común, de manera que puedan entenderlos y/o relacionarlos. Estas representaciones están basadas en la aplicación de técnicas de Procesado del Lenguaje Natural (PLN), las cuales hacen uso de las características del lenguaje natural contenidas en los documentos, como la frecuencia de ocurrencia de las palabras o la estructura del lenguaje utilizado [109]. A pesar de la diversa cantidad de representaciones existentes, el modelo de espacio vectorial [106] – VSM por sus siglas en inglés – es la representación más utilizada, según la cual cada documento en una colección se representa como un vector en un espacio de alta dimensionalidad, comúnmente utilizando como pesos la frecuencia de ocurrencia de las palabras. Cuando se utilizan las palabras como características o *features*, el modelo se conoce como bolsa de palabras (*Bag of Words*, BoW), siendo una bolsa un conjunto de elementos que pueden ocurrir más de una vez [8]. Según el modelo bolsa de palabras, cada documento se representa por medio del conjunto de palabras que contiene y la frecuencia de ocurrencia de dichas palabras en el documento.

Por otro lado, generalmente, cuanto mayor es el conjunto de elementos de entrenamiento, mejor es el rendimiento del clasificador [61]. De esta forma, la clasificación

algorítmica puede ofrecer un rendimiento bajo cuando no existe una secuencia de documentos lo suficientemente grande para entrenar el clasificador de forma adecuada [42], como a veces sucede en idiomas diferentes al inglés. Por ejemplo, aunque el inglés es el principal idioma utilizado en la diseminación de la investigación biomédica, la relevancia de la investigación realizada y publicada localmente en idiomas diferentes al inglés no debe ser subestimada [40]. Además, la información gestionada por los centros clínicos y de investigación (e.g. expedientes médicos) generalmente se encuentra escrita en el idioma oficial de cada país o región. En escenarios como este la clasificación entre idiomas o multilingüe se vuelve pertinente. La clasificación entre idiomas consiste en el entrenamiento de un algoritmo de clasificación utilizando un conjunto de documentos etiquetados escritos en un idioma L_n – en el cual existe un conjunto de datos lo suficientemente grande para entrenar adecuadamente el algoritmo – para clasificar un conjunto de documentos escritos en un idioma diferente L_m [5].

1.2. Motivación

Tras la revisión del estado del arte realizada como parte de esta investigación, hemos comprobado que el paradigma más empleado para la representación de los documentos en las tareas de clasificación automática es el modelo bolsa de palabras. A pesar de ser la representación más utilizada, el modelo BoW no es óptimo, debido a que solo tiene en cuenta la frecuencia de ocurrencia de las palabras en los documentos, ignorando así importantes relaciones semánticas entre ellas, lo que implica la aparición de una serie de problemas del lenguaje que afectan a la calidad de la clasificación, como la sinonimia, polisemia, ortogonalidad [28, 50, 69, 131], hipónimia, hiperónimia [60, 130, 131], la dispersión de los datos y la diversidad de uso de las palabras [53, 118].

Además, la clasificación algorítmica puede ofrecer un rendimiento bajo cuando no existe una secuencia de documentos lo suficientemente grande para entrenar el clasificador de forma adecuada. Tradicionalmente, este problema ha sido abordado a través de la clasificación entre idiomas o multilingüe, utilizando la combinación de la representación BoW y las técnicas de traducción automática de documentos, bien traduciendo los documentos de entrenamiento al idioma de los documentos a clasificar [110] o viceversa [5]. Este planteamiento presenta una serie de desventajas, que surgen de la combinación de las técnicas que los componen. Por una parte, nos encontramos con las limitaciones inherentes a la representación BoW presentadas previamente. Por otro lado, las técnicas de traducción automática tienen dos grandes desventajas, la ambigüedad léxica y estructural [51, 110], que afectan negativamente a la calidad de las traducciones. La selección de una traducción incorrecta puede distorsionar la precisión de un clasificador debido a la introducción de características erróneas. Por consiguiente, cuando la representación BoW se combina con la utilización de técnicas de traducción automática, las desventajas de ambas propuestas se suman, lo que conduce a un incremento de la probabilidad de error del clasificador.

La literatura recoge diversos intentos que tienen como objetivo solventar los problemas inherentes a la representación BoW de los documentos (principalmente la sinonimia y la

polisemia) a través de la exploración de un nuevo modelo o paradigma: la representación de los documentos como bolsas de conceptos (*Bag of Concepts*, BoC), definiendo concepto como “unidad de significado” [18, 65, 116, 131]. Por definición, los conceptos son no ambiguos, lo que mitiga los problemas introducidos por la sinonimia y la polisemia. Entonces, de acuerdo al modelo BoC, un documento se representa en base al conjunto de conceptos sobre los que trata o lo componen. Dicho con otras palabras, un documento se modela como un vector de pesos de conceptos, siendo asignados estos pesos en función de la relevancia de cada uno de los conceptos en el propio documento. Diversos estudios previos han demostrado que esta representación proporciona buenos resultados en su aplicación a tareas de clasificación automática de documentos [104, 131]. Las principales propuestas presentes en el estado del arte para la creación de representaciones en forma de bolsas de conceptos incluyen aquellas que hacen uso únicamente de la información contenida en los documentos – como Latent Dirichlet Allocation (LDA) [7] o Word Embeddings (WE) [6, 126] – y aquellas que utilizan fuentes de conocimiento externo, como por ejemplo la Wikipedia, conocidas como anotadores semánticos [34, 67]. A pesar de los diversos intentos existentes en la literatura que utilizan el modelo BoC para la realización de clasificación de documentos de texto, los resultados obtenidos no han sido lo suficientemente relevantes como para desbancar al modelo basado en palabras, que sigue siendo el enfoque dominante [15, 45, 56, 88, 104, 107, 140].

1.3. Hipótesis

Uno de los principales resultados del proyecto iTEC, enmarcado en el Séptimo Programa Marco de Investigación y Desarrollo de la Unión Europea y en el cual el autor trabajó como ingeniero de proyecto a la vez que comenzaba su tesis doctoral en el campo de la clasificación automática de documentos, ha sido la creación de una plataforma de búsqueda y recuperación de recursos educativos [96], representados estos como bolsas de conceptos de la Wikipedia – obtenidos a través del anotador semántico Wikipedia Miner [67]. Los excelentes resultados obtenidos de la utilización del anotador semántico Wikipedia Miner para representar los documentos en el anterior sistema de recuperación de información, junto con los problemas de los modelos tradicionales para la representación de los documentos en tareas de clasificación presentados previamente, nos llevan a plantearnos la siguiente hipótesis de investigación:

Hipótesis: La utilización de una representación de los documentos basada en conceptos de la Wikipedia (WikiBoC), obtenidos a través del anotador semántico de propósito general Wikipedia Miner, mejora el rendimiento de las propuestas actuales para la clasificación monolingüe y multilingüe de documentos de texto.

1.4. Objetivos

El objetivo final de esta tesis es demostrar la aplicabilidad y los beneficios aportados por el uso de una representación de los documentos en forma de bolsa de conceptos que hace

uso del conocimiento enciclopédico e información semántica almacenados en la Wikipedia (WikiBoC), a la clasificación de documentos de texto, a través de la experimentación del uso de la representación propuesta en diferentes tareas de clasificación monolingüe y multilingüe de documentos, pertenecientes a diferentes ámbitos de aplicación. Para alcanzar este objetivo final, se identifican los siguientes objetivos intermedios.

1. Revisión de las representaciones de los documentos más relevantes en el estado del arte para la realización de tareas de clasificación de documentos de texto.
2. Familiarización con los métodos tradicionales de clasificación y representación de los documentos.
3. Diseño y desarrollo de un banco de pruebas que permita evaluar la aplicabilidad y mostrar los beneficios aportados por la propuesta presentada frente a las propuestas más relevantes existentes en el estado del arte.
4. Creación de conjuntos de datos, más allá de los estándares presentes en el estado del arte, para obtener una evaluación más exhaustiva de la propuesta presentada.
5. Validación de la aplicabilidad y exposición de los beneficios aportados por la propuesta WikiBoC a la clasificación monolingüe de documentos de texto.
6. Validación de la aplicabilidad y exposición de los beneficios aportados por la propuesta WikiBoC a la clasificación multilingüe de documentos de texto.

1.5. Metodología

El hecho de que tanto el trabajo presentado en esta investigación, como el presentado en cada una de las publicaciones derivadas se traten de trabajos de ingeniería, implica que existen contribuciones en forma de conocimiento y/o artefactos – entendiendo artefacto por todo aquello que no es natural sino construido por el hombre [111] – lo que va más allá de una mera observación de la naturaleza. Las técnicas de clasificación propuestas, su análisis, la evaluación comparativa con otras propuestas presentes en la literatura y la creación, depuración y anotación de los conjuntos de datos suponen un avance del estado del arte.

Una vez diseñados y desarrollados, estos artefactos deben ser evaluados con el objetivo de comprobar su efectividad a la hora de resolver los problemas identificados, de manera que puedan pasar a formar parte de un nuevo conjunto de conocimiento a través de la comunicación de las contribuciones aportadas. De esta forma, se ha optado por seguir la metodología de investigación DSRM (*Design Science Research Methodology*) [95] para la realización del trabajo de investigación presentado en esta tesis. Esta metodología se compone de 6 etapas o fases: identificación del problema y motivación, definición de los objetivos de la solución, diseño y desarrollo de la solución propuesta, demostración, evaluación y comunicación.

1. **Identificación del problema y motivación.** En esta etapa se define el problema de investigación específico y se justifica el valor que aportaría la solución a dicho problema. Dado que la definición del problema será utilizada para desarrollar un artefacto que proporcione una solución al problema, puede ser útil atomizar el problema conceptualmente de manera que la solución pueda capturar su complejidad. La justificación del valor de la solución implica dos objetivos: motivar al investigador y a la audiencia a la que se dirige la solución, y ayudar a entender el razonamiento asociado con la visión del problema por parte del investigador. Esta primera etapa requiere conocimiento acerca del estado del arte del problema así como de la importancia de la solución.
2. **Definición de los objetivos a alcanzar con la solución.** En esta fase se realiza la inferencia de los objetivos de la solución a partir de la definición del problema y el conocimiento de lo que es posible y factible. Estos objetivos pueden ser cuantitativos o cualitativos, y deben ser inferidos de forma racional a partir de los requisitos del problema. Esta etapa requiere conocimiento acerca del estado del arte del problema, así como de las posibles soluciones existentes, si existen, y de su efectividad.
3. **Diseño y desarrollo.** Esta etapa consiste en la creación de los artefactos, entendiendo artefactos como construcciones, modelos, métodos, instanciaciones, o nuevas propiedades de recursos técnicos, sociales y/o informacionales. Esta etapa incluye la definición de la funcionalidad deseada por el artefacto, su arquitectura y la posterior creación del artefacto.
4. **Demostración.** Esta fase consiste en la demostración del uso del artefacto para solucionar una o más instancias del problema. Esto incluye la realización de experimentos, simulaciones, casos de estudio, pruebas, o cualquier otra actividad apropiada. Los recursos requeridos para la realización de esta etapa incluyen conocimiento acerca de cómo utilizar el artefacto diseñado y desarrollado para solucionar el problema.
5. **Evaluación.** En esta etapa se realiza la observación y medida de cuan bien el artefacto proporciona la solución al problema. Se requiere para ello conocimiento de las técnicas de análisis y métricas más relevantes. En función de la naturaleza del problema y del artefacto, la evaluación se puede llevar a cabo de diversas formas, incluyendo la comparación de la funcionalidad del artefacto con los objetivos a alcanzar, medidas de rendimiento objetivas y cuantitativas, encuestas de satisfacción, opiniones de los usuarios, etc. Al final de esta fase, los investigadores pueden decidir si volver al paso 3 con el objetivo de mejorar la efectividad del artefacto, en caso de ser factible, o pasar a la siguiente fase y dejar las posibles mejoras para futuros proyectos.
6. **Comunicación.** Esta última etapa consiste en la comunicación del problema y su importancia, el artefacto, su utilidad e innovación, el rigor de su diseño y su efectividad a la comunidad científica y a cualquier otra audiencia relevante, a través de, por ejemplo, su publicación en revistas y conferencias internacionales.

La figura 1.1 resume el proceso de investigación seguido en esta tesis. En primer lugar, tras la revisión del estado del arte realizada, se ha comprobado que las técnicas

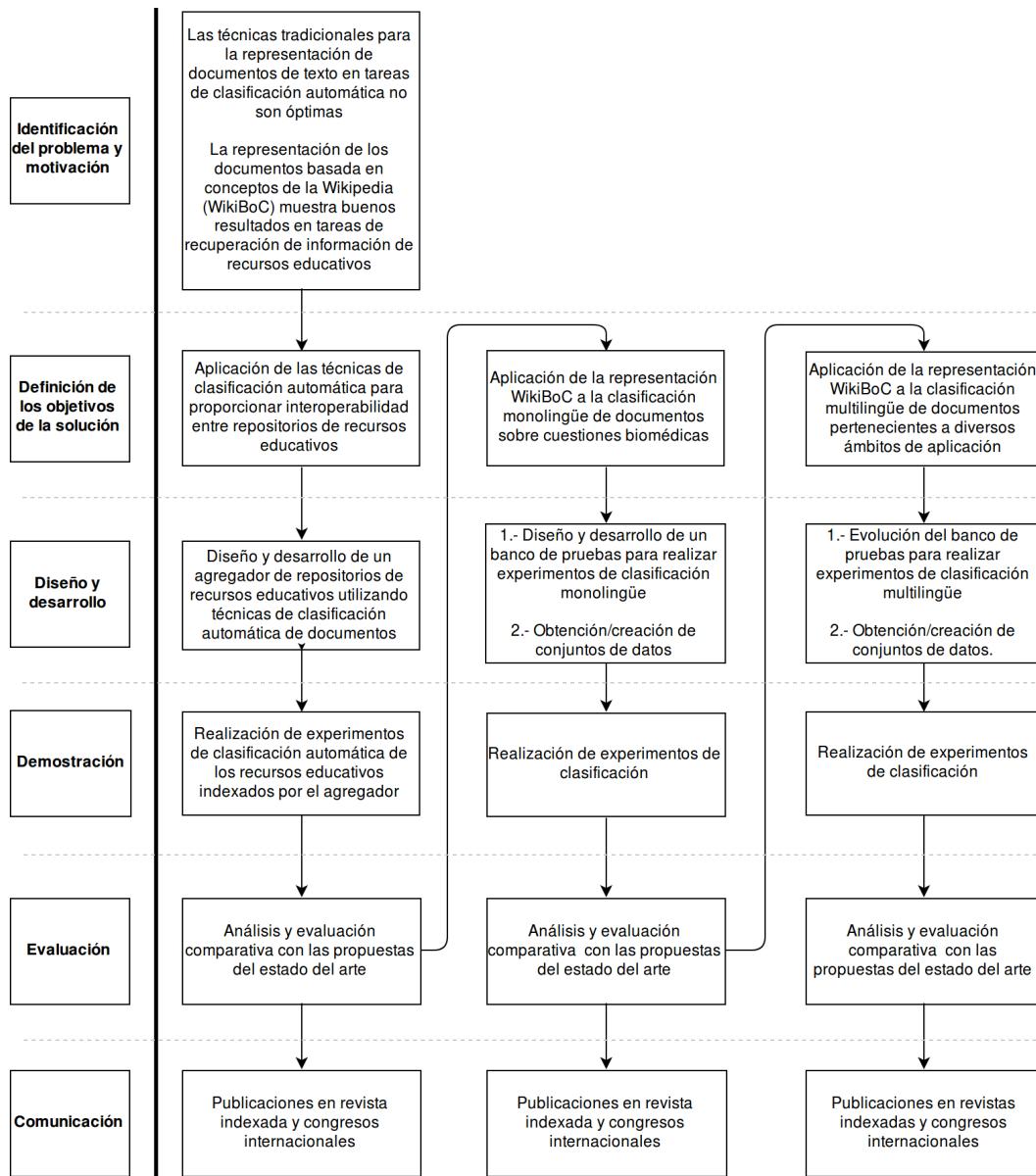


Figura 1.1: Proceso de investigación seguido en la tesis.

tradicionales para la representación de documentos de texto en tareas de clasificación automática no son óptimas y presentan una serie de inconvenientes que lastran su rendimiento. Esto, junto a trabajos previos realizados en el grupo de investigación, los cuales muestran buenos resultados en la aplicación de la representación de los documentos basada en conceptos de la Wikipedia (WikiBoC) a la tarea de recuperación de información de recursos educativos, conforma la etapa de identificación del problema y motivación de la metodología DSRM.

Sin cambiar de ámbito de aplicación, y con el objetivo de familiarizarse con las técnicas de clasificación automática de documentos, se explora en primer lugar la aplicación de dichas técnicas a la clasificación de recursos educativos. El trabajo desarrollado, detallado en el capítulo 3.1, consiste en un agregador de repositorios de recursos educativos que tiene como objetivo proporcionar interoperabilidad entre dichos repositorios, sin hacer uso de las técnicas clásicas para tal efecto basadas en el mapeado entre taxonomías. La evaluación de la propuesta presentada se realizó a través del análisis y evaluación comparativa con las propuestas presentes en el estado del arte.

Esta investigación continúa con la aplicación de la representación WikiBoC a la clasificación de documentos. El trabajo desarrollado, que se encuentra detallado en el capítulo 3.2, consiste en la aplicación de la representación WikiBoC a la clasificación monolingüe de documentos, concretamente en uno de los dominios más relevantes para su aplicación, el dominio biomédico. La evaluación de la propuesta se llevó a cabo a través de la realización de diversos experimentos de clasificación de documentos sobre cuestiones biomédicas, y del análisis y evaluación comparativa de los resultados obtenidos utilizando la propuesta presentada, con aquellos proporcionados por las propuestas más relevantes del estado del arte.

Por último, esta investigación explora el uso de la representación WikiBoC en la clasificación multilingüe de documentos pertenecientes a diversos ámbitos de aplicación, tal y como se detalla en el capítulo 3.3. De nuevo, la evaluación de la propuesta se llevó a cabo a través de la realización de diferentes experimentos de clasificación utilizando diversos conjuntos de datos, y del análisis y evaluación comparativa de los resultados obtenidos utilizando la solución presentada, con aquellos ofrecidos por las propuestas más relevantes presentes en el estado del arte.

1.6. Publicaciones

De la investigación realizada y expuesta en este documento han surgido cuatro publicaciones en las siguientes revistas indexadas en el *Journal Citations Reports: Computers & Education, PeerJ, Information Sciences, y Methods of Information in Medicine*.

- La primera publicación, Mouriño-García et al. [81], sirve como primera toma de contacto previa a las propuestas centrales de este trabajo. Se presenta CROERA,

un agregador de repositorios de recursos educativos (REA) que proporciona acceso a los recursos indexados independientemente de la taxonomía utilizada por cada uno de los repositorios integrados. Esto se consigue a través de la clasificación automática de cada uno de los recursos – representados según el modelo bolsa de palabras – en función de cada una de las taxonomías de los repositorios agregados.

Mouriño-García, M., Pérez-Rodríguez, R., Anido-Rifón, L., Fernández-Iglesias, M. J., & Darriba-Bilbao, V. M. Cross-repository aggregation of educational resources, In *Computers & Education*, Volume 117, 2018, Pages 31-49, ISSN 0360-1315, <https://doi.org/10.1016/j.compedu.2017.09.014>.

- La segunda publicación, Mouriño-García et al. [74], describe el diseño, desarrollo y evaluación de un clasificador monolingüe de literatura biomédica escrita en inglés, utilizando la representación de los documentos en forma de bolsa de conceptos extraídos de la Wikipedia.

Mouriño-García, M. A., Pérez-Rodríguez, R., & Anido-Rifón, L. E. (2015). Biomedical literature classification using encyclopedic knowledge: a Wikipedia-based bag-of-concepts approach. *PeerJ*, 3, e1279. ISSN: 2167-8359

- La tercera publicación, Mouriño-García et al. [80], describe las bases y presenta los resultados de evaluación de un clasificador multilingüe de documentos de texto (inglés – castellano) que aprovecha el conocimiento de la Wikipedia para representar los documentos como bolsas de conceptos.

Mouriño-García, M. A., Pérez-Rodríguez, R., & Anido-Rifón, L. (2017). Wikipedia-based cross-language text classification. *Information Sciences*, 406, 12-28. ISSN: 0020-0255

- La cuarta publicación, Mouriño-García et al. [79], describe las bases y presenta los resultados de evaluación de un clasificador de documentos de texto multilingüe (inglés – castellano) que hace uso del conocimiento enciclopédico contenido en la Wikipedia para mapear documentos de texto biomédicos a vectores de pesos de conceptos Wikipedia.

Mouriño-García, M. A., Pérez-Rodríguez, R., & Anido-Rifón, L. E. (2017). A Bag of Concepts Approach for Biomedical Document Classification Using Wikipedia Knowledge. *Methods of Information in Medicine*, 56. ISSN: (Print): 0026-1270 (Online): 2511-705X

1.7. Estructura del documento

A continuación, se muestran los capítulos en los cuales se ha dividido el documento, así como un breve resumen del contenido de cada uno de ellos.

Capítulo 1. Introducción. Este capítulo realiza una presentación general de la investigación realizada. En primer lugar, se introduce la clasificación automática de documentos, se pone de manifiesto su relevancia en la sociedad actual y se presentan los principales problemas de las representaciones tradicionales de los documentos en las tareas de clasificación automática. A continuación se presenta la hipótesis de investigación. Esta surge de los buenos resultados obtenidos de la utilización del anotador semántico Wikipedia Miner para la creación de un sistema de recuperación de información bajo el paraguas del proyecto iTEC, durante el cual el autor tuvo sus primeros contactos con Wikipedia Miner. El capítulo continúa con la descripción de los objetivos a alcanzar en esta tesis y de la metodología utilizada para su realización. A continuación, se exponen las cuatro publicaciones derivadas de esta investigación. Finalmente, este capítulo termina con la descripción del resto de la tesis.

Capítulo 2. Materiales y métodos. Este capítulo introduce las principales formulaciones matemáticas, técnicas, herramientas y recursos utilizados durante la realización de esta investigación con el objetivo de facilitar al lector la correcta comprensión de este documento, así como de las publicaciones presentadas.

Capítulo 3. Resultados y discusión. Este capítulo presenta y discute los principales resultados obtenidos de la investigación realizada. En primer lugar, y como paso previo a las propuestas centrales presentadas en esta tesis, se exponen y discuten los resultados de la aplicación de las técnicas de clasificación automática de documentos para la creación de un agregador de repositorios de recursos educativos. Tras esto, la investigación se centra en la aplicación de la representación WikiBoC a la clasificación automática de documentos, y se presentan y discuten los resultados de la aplicación de la representación WikiBoC a la clasificación monolingüe y multilingüe de documentos de texto en diversos ámbitos de aplicación.

Capítulo 4. Contribuciones, conclusiones y trabajos futuros. Este capítulo presenta las contribuciones realizadas por esta investigación, las conclusiones obtenidas tras su realización, las cuales permiten verificar la hipótesis de investigación planteada, y una serie de líneas futuras que permiten continuar la investigación presentada en esta tesis.

Apéndice I. Publicaciones. Este apéndice incluye de forma íntegra los artículos publicados como resultado de la investigación detallada en el presente documento.

Capítulo 2

Materiales y métodos

Este capítulo presenta las principales formulaciones matemáticas, técnicas, herramientas y recursos utilizados durante la realización de esta investigación. En primer lugar se introducen las técnicas de aprendizaje máquina, el problema de la clasificación de documentos, el algoritmo *Support Vector Machines*, y las métricas empleadas para evaluar el rendimiento de la clasificación de documentos. A continuación, se describe la representación de los documentos, se introduce brevemente la Wikipedia y se presenta el anotador semántico Wikipedia Miner. Tras esto, se definen los diferentes modelos de representación de los documentos empleados en esta investigación, tanto aquellos existentes en el estado del arte (BoW, BoW-MT, ESA, Bi-LDA, BWE y MetaMap) como los propuestos como parte de esta investigación (WikiBoC, Hybrid-WikiBoC, y WikiBoC-CLCM). Por último, se describe la técnica *Cross-Language Concept Matching* (CLCM) y se detallan los conjuntos de datos empleados para evaluar las propuestas presentadas.

2.1. Aprendizaje máquina

El aprendizaje automático o aprendizaje máquina es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a los ordenadores aprender [108]. En concreto, las técnicas de aprendizaje máquina exploran el estudio y la construcción de algoritmos que puedan aprender de un conjunto de datos y hacer predicciones acerca de ellos. Típicamente se clasifican en tres grandes grupos en función del conocimiento existente a priori: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje semi-supervisado.

2.1.1. Aprendizaje supervisado

El aprendizaje supervisado [12, 108] consiste en inferir una función a partir de una serie de ejemplos etiquetados (secuencia de entrenamiento) para posteriormente predecir una salida para otro conjunto distinto de ejemplos no etiquetados (secuencia de test).

Definición 1. Sea \mathbb{D} un conjunto de documentos. La secuencia de entrenamiento $A \in \mathbb{D}$, se define como

$$A = \{d_1^{L_n}, d_2^{L_n}, \dots, d_{|A|}^{L_n}\} \quad (2.1)$$

siendo $d_i^{L_n}$ un documento etiquetado escrito en un idioma L_n .

Definición 2. La secuencia de test $Z \in \mathbb{D}$, se define como

$$Z = \{d_1^{L_m}, d_2^{L_m}, \dots, d_{|Z|}^{L_m}\} \quad (2.2)$$

siendo $d_i^{L_m}$ un documento escrito en un idioma L_m .

Como veremos posteriormente, si $L_n = L_m$ nos encontramos ante un problema de clasificación monolingüe, es decir, los documentos pertenecientes a la secuencia de entrenamiento se encuentran escritos en el mismo idioma que los documentos pertenecientes a la secuencia de test. Sin embargo, cuando $L_n \neq L_m$ nos encontramos ante un problema de clasificación entre idiomas o multilingüe, lo que significa que los documentos de la secuencia de entrenamiento se encuentran escritos en un idioma diferente a los elementos de la secuencia de test.

Cada elemento perteneciente a la secuencia de entrenamiento se representa generalmente como un par compuesto por un objeto de entrada (típicamente un vector) y una etiqueta de clase o categoría (como sucede en los problemas de clasificación) o un valor numérico (como sucede en los problemas de regresión). Así, un algoritmo de aprendizaje máquina supervisado analiza los datos comprendidos en la secuencia de entrenamiento y produce o infiere una función capaz de predecir el valor de salida correspondiente para cualquier elemento perteneciente a la secuencia de test. La figura 2.1 muestra de forma gráfica un problema de aprendizaje supervisado genérico. Durante la fase de entrenamiento se le proporciona al algoritmo una serie de ejemplos etiquetados lo suficientemente grande para que el algoritmo sea entrenado de forma adecuada. Tras un entrenamiento suficiente, el algoritmo es capaz de predecir la salida correspondiente a un documento de entrada no etiquetado gracias al conocimiento inferido durante la fase de entrenamiento.

2.1.2. Aprendizaje no supervisado

El aprendizaje no supervisado [44] es la tarea de inferir una función que describe estructuras ocultas a partir de un conjunto de datos no etiquetados. La principal

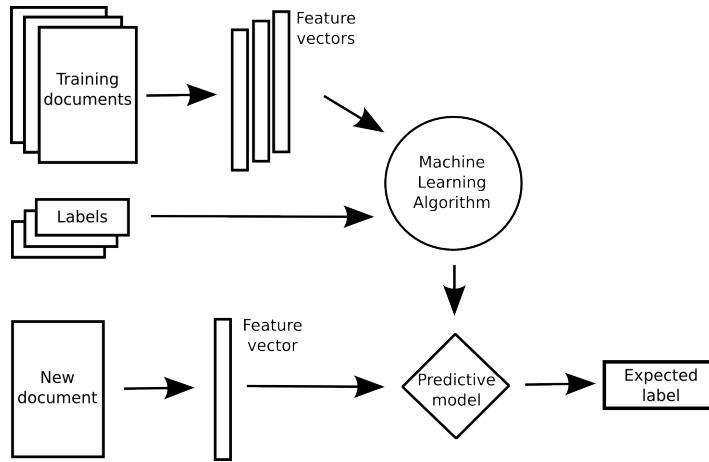


Figura 2.1: Aprendizaje supervisado.

diferencia del aprendizaje no supervisado frente al aprendizaje supervisado es que no existe conocimiento a priori, es decir, no existe una fase de entrenamiento del algoritmo. Generalmente, las técnicas de aprendizaje no supervisado tratan los datos de entrada como un conjunto de variables aleatorias y construyen un modelo de densidad para dicho conjunto de datos.

2.1.3. Aprendizaje semi-supervisado

Las técnicas de aprendizaje semi-supervisado [144] se encuentran entre las dos anteriores. El aprendizaje semi-supervisado se basa en un conjunto de técnicas de aprendizaje supervisado que también hacen uso de datos sin etiquetar durante la fase de entrenamiento. Típicamente, se hace uso de una pequeña cantidad de datos etiquetados junto con una gran cantidad de datos sin etiquetar.

2.2. Clasificación de documentos

El enfoque dominante en la comunidad investigadora para abordar el problema de la clasificación automática de documentos se basa en la aplicación de técnicas de aprendizaje máquina [4, 12, 108]. En particular, la investigación llevada a cabo en esta tesis se ha centrado principalmente en la clasificación automática de documentos modelada como un problema de aprendizaje máquina supervisado. La clasificación automática de documentos se puede definir como la asignación algorítmica de documentos a un conjunto de categorías, etiquetas o clases previamente definido. De la misma forma que Sebastiani [108]:

Definición 3. La clasificación de documentos puede definirse como

$$\check{C} : \mathbb{D} \times \mathbb{C} \rightarrow \{\text{T}, \text{F}\} \quad (2.3)$$

donde \mathbb{D} es el dominio de los documentos, es decir, el conjunto de documentos involucrados en un problema de clasificación concreto, $\mathbb{C} = \{c_1, c_2, \dots, c_{|\mathbb{C}|}\}$ es el conjunto de categorías y T, F denotan verdadero (*true*) y falso (*false*). $\check{C}(d_i, c_j) = \text{T}$ cuando el documento d_i pertenece a la categoría c_j y $\check{C}(d_i, c_j) = \text{F}$ cuando el documento d_i no pertenece a la categoría c_j .

En función de si los documentos involucrados en el proceso de aprendizaje del algoritmo se encuentran asociados a una o a más categorías, podemos dividir los problemas de clasificación automática de documentos en problemas de clasificación de etiqueta (categoría, clase) única y problemas de clasificación multietiqueta (multicategoría, multiclasificación).

2.2.1. Clasificación de etiqueta única y multietiqueta

Los problemas de clasificación de etiqueta única se caracterizan porque cada documento se encuentra asociado a una única etiqueta l de entre un conjunto disjunto de etiquetas L , $|L| > 1$. Sin embargo, en la clasificación multietiqueta, cada documento puede ser asociado a un conjunto de etiquetas $Y \subset L$, $|Y| > 1$ [121]. Los métodos de clasificación multietiqueta, a su vez, se pueden agrupar en dos categorías diferentes: métodos de transformación del problema, y métodos de adaptación del algoritmo [121]. Los métodos de transformación del problema son aquellos que transforman un problema de clasificación multietiqueta en uno o varios problemas de etiqueta única, mientras que los métodos de adaptación del algoritmo extienden o modifican de forma específica algún algoritmo, de manera que puedan abordar problemas de clasificación multietiqueta sin necesidad de realizar ningún tipo de transformación del problema de clasificación original. Para la realización de los problemas de clasificación multietiqueta involucrados en este trabajo de investigación hemos utilizado los métodos de la primera categoría, es decir, se han transformado los problemas multietiqueta en n problemas de etiqueta única. Para realizar esto, hemos seleccionado la estrategia *one-vs-rest* o *one-vs-all* disponible en la librería *scikit-learn* [94], que consiste en el ajuste de un clasificador por cada una de las categorías del problema. Esta estrategia cuenta con una elevada eficiencia e interpretabilidad – ya que cada categoría se representa por medio de un único clasificador – de manera que resulta sencillo obtener más información acerca de una categoría específica simplemente inspeccionando el clasificador correspondiente.

2.2.2. Clasificación monolingüe y multilingüe

Otra posible división de los problemas de clasificación de documentos radica en el idioma de los propios documentos. Por un lado, las tareas de clasificación monolingüe

son aquellas en las que los documentos a clasificar se encuentran escritos en el mismo idioma que los documentos utilizados para entrenar el clasificador. Por otro lado, las tareas de clasificación multilingüe consisten en la clasificación de un conjunto de documentos escritos en un idioma L_m , utilizando un algoritmo de clasificación entrenado con documentos etiquetados escritos en un idioma diferente L_n .

2.3. Support Vector Machines

Las máquinas de vectores de soporte (*Support Vector Machines*, SVM [46]) son un conjunto de algoritmos de aprendizaje máquina ampliamente utilizados en tareas de *clustering*, regresión y clasificación automática de documentos. Este algoritmo ha sido seleccionado para la realización de los experimentos presentados en esta investigación debido a que es uno de los algoritmos más exitosos para la realización de tareas de clasificación automática de documentos [24, 36, 55, 102]. La idea básica consiste en que, dado un conjunto de elementos que pertenecen a un conjunto de categorías, el algoritmo SVM construye un modelo capaz de predecir a qué categoría pertenecerá un nuevo elemento que aparezca en el sistema. *Support Vector Machines* representa cada elemento como un punto en el espacio, separando las clases a espacios lo más amplios posibles, mediante hiperplanos de separación llamados vectores de soporte. De esta forma, cuando los nuevos elementos se ponen en correspondencia con el modelo, se clasificarán en una clase u otra en función de los espacios a los que pertenezcan. Para la implementación de este algoritmo se ha utilizado la clase *sklearn.svm.LinearSVC* de la librería *scikit-learn*. Es necesario destacar que, como el principal objetivo de este trabajo de investigación se centra en la representación de los documentos y no en el propio algoritmo de clasificación, en todos los experimentos realizados se han utilizado los parámetros por defecto proporcionados por la implementación de este algoritmo.

2.4. Métricas de evaluación

La eficiencia de los clasificadores se mide generalmente en términos de las nociones clásicas de recuperación de información precisión (P) y retirada (R), adaptadas al caso de la clasificación de documentos. De acuerdo a Sahlgren and Cöster [104], cuando se utilizan clasificadores para predecir la categoría de un documento, se producen cuatro posibles resultados: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), y *False Negative* (FN). Positivo significa que un documento se clasificó en una categoría concreta, mientras que negativo significa lo contrario. Verdadero significa que la clasificación fue correcta, mientras que falso significa que no lo fue.

2.4.1. Precisión

De acuerdo a Sahlgren and Cöster [104] y Sebastiani [108], la precisión se define como la probabilidad de que si un documento aleatorio $d_i \in D$ es clasificado en una categoría $c \in C$ esta decisión sea correcta. Dados los cuatro posibles resultados anteriores, TP , TN , FP y FN , la precisión se define como

$$P = \frac{TP}{(TP + FP)} \quad (2.4)$$

La librería *scikit-learn* incluye un módulo para el cálculo de las métricas de evaluación presentadas¹. En concreto, para la obtención de la precisión se ha utilizado la clase *sklearn.metrics.precision_score*²

2.4.2. Retirada

De manera análoga a la definición de precisión, la retirada se define como la probabilidad de que si un documento aleatorio $d_i \in D$ debiera haber sido clasificado en la categoría $c \in C$, esa decisión haya sido tomada. Con los cuatro posibles anteriores resultados, TP , TN , FP y FN , la retirada se define como

$$R = \frac{TP}{(TP + FN)} \quad (2.5)$$

Para el cálculo de la retirada se ha utilizado la clase *sklearn.metrics.recall_score*³ de la librería *scikit-learn*.

2.4.3. F-score

A mayores, utilizamos una medida que combina la precisión y la retirada, el $F_\beta - score$, que se define como

$$F_\beta = \frac{(1 + \beta^2) * P * R}{\beta^2 * P + R} \quad (2.6)$$

La medida $F_\beta - score$ permite ponderar el peso de la precisión sobre la retirada y viceversa, a través del parámetro de control β . Si $\beta < 1$, significa que la precisión es más relevante que la retirada, mientras que si $\beta > 1$, la retirada es más relevante que la precisión. Cuando $\beta = 1$ obtenemos la media armónica, el $F_1 - score$ [30, 124]. Para el cálculo de esta métrica se ha utilizado la clase *sklearn.metrics.f1_score*⁴, contenida en la librería *scikit-learn*.

¹<http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics>

²http://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html

³http://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html

⁴http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

En esta investigación hemos reportado los resultados como macro- P , macro- R y macro- F_1 , debido a que son las mejores métricas para mostrar el rendimiento de los clasificadores en conjuntos de datos cuyos documentos no están distribuidos de forma uniforme sobre las categorías disponibles [142]. El cálculo de las métricas macro se basa simplemente en calcular las métricas para cada etiqueta o clase y posteriormente obtener su media no ponderada.

2.5. Representación de los documentos

La representación de los documentos es clave en el diseño de sistemas de clasificación de documentos de texto, debido a que los algoritmos de clasificación requieren que los documentos estén representados de forma que un ordenador o programa software pueda entenderlos. Típicamente, un documento de texto se representa a través de un vector de pesos de características extraídas del propio documento – como la frecuencia de ocurrencia de las palabras o la estructura del lenguaje usado – en un espacio que contiene tantas dimensiones como características diferentes [106, 109].

2.5.1. Bolsa de palabras

Cuanto las palabras y la frecuencia de ocurrencia de estas en los documentos se utilizan como características para crear los vectores anteriormente citados, la representación o modelo se conoce como bolsa de palabras. Entonces, de acuerdo al modelo bolsa de palabras, un documento se representa por medio de un vector de pesos de características, donde cada característica es una palabra, y el peso de cada característica es su frecuencia de ocurrencia en el propio documento. Dicho de otra forma, un documento se caracteriza por el conjunto de palabras que contiene, repetidas tantas veces como estén presentes en el documento. Las siguientes definiciones muestran de forma teórica el modelo bolsa de palabras.

Definición 4. El dominio de las características – en este caso palabras – está compuesto por todas las palabras presentes en el conjunto de documentos involucrados en el problema de clasificación, exceptuando las *stop words* [134] y aplicando previamente el algoritmo de *stemming* de Porter [99]. Se define como

$$\mathbb{W}^{L_j} = \{w_1^{L_j}, w_2^{L_j}, \dots, w_{|\mathbb{W}^{L_j}|}^{L_j}\} \quad (2.7)$$

siendo cada $w_k^{L_j}$ una palabra contenida en el conjunto de documentos escrito en un idioma L_j . Las *stop words* son palabras sin significado, como artículos, pronombres y preposiciones, que no se utilizan en tareas de clasificación de textos, puesto que con elevada probabilidad aparecerán en todos los documentos. El *stemming* es un método para reducir las palabras a su raíz, con el fin de crear características más generales. Estas dos técnicas permiten reducir la dimensionalidad del conjunto de características,

	mercurio	planeta	está	próximo	sol	pequeño	cálido	sistema	solar	cerca
Doc. A	2	2	1	1	1	1	0	0	0	0
Doc. B	2	1	1	0	1	0	1	1	1	1

Tabla 2.1: Número de ocurrencias de cada palabra en cada uno de los documentos.

perjudicial para el rendimiento de las tareas de clasificación automática de textos [35, 135].

Entonces,

Definición 5. Un documento escrito en un idioma L_j , representado según el modelo bolsa de palabras $\overline{BoW_d_i^{L_j}}$, se define como

$$\overline{BoW_d_i^{L_j}} = (ww_{i1}, ww_{i2}, \dots, ww_{i|\mathbb{W}^{L_j}|}) \quad (2.8)$$

donde ww_{ik} es el peso de la característica $w_k^{L_j}$ en el documento $d_i^{L_j}$, o lo que es lo mismo, la frecuencia de ocurrencia de la palabra $w_k^{L_j}$ en el documento $d_i^{L_j}$.

Ejemplo 1. Supongamos que el conjunto de documentos está compuesto únicamente por dos documentos:

- Documento A: “Mercurio es el planeta que está más próximo al Sol. Mercurio es el planeta más pequeño.”
- Documento B: “El planeta más cálido del sistema solar es Mercurio. Mercurio está muy cerca del Sol.”

Tras la eliminación de las *stop words*, los documentos quedan de la siguiente manera:

- Documento A: “Mercurio planeta está próximo Sol Mercurio planeta pequeño”
- Documento B: “planeta cálido sistema solar Mercurio Mercurio está cerca Sol”

Tomando como base los dos documentos anteriores, obtenemos el dominio de las características de este conjunto de documentos, formado por todas las palabras únicas contenidas en el conjunto: “Mercurio”, “planeta”, “está”, “próximo”, “Sol”, “pequeño”, “cálido”, “sistema”, “solar”, y “cerca”:

El proceso de creación de los vectores se puede ver de forma clara a través de la tabla 2.1. En el documento A, las palabras “Mercurio” y “planeta” se repiten dos veces, las palabras “está”, “próximo”, “Sol” y “pequeño” se repiten una vez, mientras que las palabras “cálido”, “sistema”, “solar” y “cerca” no están presentes en el documento. De forma similar se obtiene el vector para el documento B.

De esta forma, siguiendo el modelo bolsa de palabras, los documentos quedan representados en forma de vector de la siguiente manera:

- Documento A: (2, 2, 1, 1, 1, 1, 0, 0, 0, 0)
- Documento B: (2, 1, 1, 0, 1, 0, 1, 1, 1, 1)

2.5.2. Bolsa de conceptos

Con el objetivo de mitigar los problemas asociados a la representación de los documentos como bolsas de palabras, varios autores han propuesto y explorado un modelo diferente: la representación como bolsas de conceptos. De acuerdo a este modelo, un documento se representa en base a los conceptos sobre los que trata, siendo un concepto una “unidad de significado” [65, 116, 131]. De esta forma, un documento se modela como un vector de pesos de conceptos, indicando estos pesos la relevancia de cada uno de los conceptos dentro del documento. De forma análoga a la representación en forma de bolsa de palabras, las siguientes definiciones muestran de forma teórica el modelo bolsa de conceptos.

Definición 6. El dominio de las características – en este caso conceptos – está compuesto por todos los conceptos presentes en el conjunto de documentos involucrados en el problema de clasificación

$$\mathbb{C}^{L_j} = \{c_1^{L_j}, c_2^{L_j}, \dots, c_{|\mathbb{C}^{L_j}|}^{L_j}\} \quad (2.9)$$

siendo cada $c_k^{L_j}$ un concepto presente en el conjunto de documentos escrito en un idioma L_j .

Definición 7. Un documento escrito en un idioma L_j , representado como bolsa de conceptos $\overline{BoC_d_i^{L_j}}$, se define como

$$\overline{BoC_d_i^{L_j}} = (cw_{i1}, cw_{i2}, \dots, cw_{i|\mathbb{C}^{L_j}|}) \quad (2.10)$$

donde cw_{ik} es el peso de la característica $c_k^{L_j}$ en el documento $d_i^{L_j}$, que se define como la relevancia del concepto $c_k^{L_j}$ en el documento $d_i^{L_j}$.

Ejemplo 2. Debido a la existencia de diferentes propuestas para obtener representaciones en forma de bolsa de conceptos de los documentos, realizaremos este ejemplo utilizando conceptos genéricos. Supongamos de nuevo que el conjunto de documentos está compuesto únicamente por dos documentos que tratan sobre un conjunto de conceptos.

- Documento A: {c1, c3, c4, c5, c8, c10}
- Documento B: {c1, c2, c4, c6, c7, c9, c10}

Tomando como base los dos documentos anteriores, obtenemos el dominio de las características de este conjunto de documentos, formado por todos los conceptos presentes en el conjunto: “c1”, “c2”, “c3”, “c4”, “c5”, “c6”, “c7”, “c8”, “c9” y “c10”.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
Doc. A	0.2	0	0.3	0.5	0.1	0	0	0.4	0	0.9
Doc. B	0.3	0.6	0	0.5	0	0.5	0.3	0	0.8	0.1

Tabla 2.2: Peso o relevancia de cada concepto en cada uno de los documentos.

El proceso de creación de los vectores se puede ver de forma clara a través de la tabla 2.2. El documento A trata los conceptos “c1”, “c3”, “c4”, “c5”, “c8” y “c10”, mientras que no trata los conceptos “c2”, “c6”, “c7” y “c9”. De forma similar se obtiene el vector para el documento B. Los pesos – seleccionados de forma aleatoria para el ejemplo – indican la relevancia de cada concepto dentro del documento. Por ejemplo, vemos que en el documento A, el concepto “c5” tiene un peso igual a 0.1, lo que implica que tiene muy poca relevancia en el documento. De forma opuesta, el concepto “c10” tiene un peso de 0.9, lo que indica elevada relevancia dentro del documento.

De esta forma, siguiendo el modelo bolsa de conceptos, los documentos quedan representados en forma de vector de la siguiente manera:

- Documento A: (0.2, 0, 0.3, 0.5, 0.1, 0, 0, 0.4, 0, 0.9)
- Documento B: (0.3, 0.6, 0, 0.5, 0, 0.5, 0.3, 0, 0.8, 0.1)

Conocer los conceptos referenciados en un documento de texto no es una tarea trivial. Las propuestas tradicionales para abordar estos problemas se basaban en la generación manual de anotaciones semánticas [101], seleccionadas de entre un conjunto de posibles anotaciones llamado taxonomía. El principal problema de la creación manual de anotaciones es que no es un proceso escalable, debido a su elevado consumo en tiempo y propensión a errores. Para solucionar este problema, varios autores han propuesto diferentes técnicas para la extracción automática de conceptos a partir de documentos de texto, utilizando técnicas de procesado de lenguaje natural y aprendizaje máquina [6, 7, 34, 67]

2.5.3. Híbridas

Varios autores han demostrado que el uso de representaciones híbridas proporciona incrementos de rendimiento en tareas de clasificación de documentos [49, 104, 137]. De esta forma

Definición 8. La representación combinada – híbrida – BoW-BoC de un documento $d_i^{L_j}$ se define como

$$\overline{H_d_i^{L_j}} = \overline{BoW_d_i^{L_j}} + \overline{BoC_d_i^{L_j}} = (ww_{i1}, ww_{i2}, \dots, ww_{i|\mathbb{W}^{L_j}|}, cw_{i1}, cw_{i2}, \dots, cw_{i|\mathbb{C}^{L_j}|}) \quad (2.11)$$

donde ww_{ik} es el peso de la característica $w_k^{L_j}$ en el documento $d_i^{L_j}$, o lo que es lo mismo, la frecuencia de ocurrencia de la palabra $w_k^{L_j}$ en el documento $d_i^{L_j}$, y cw_{ik} es el peso de la característica $c_k^{L_j}$ en el documento $d_i^{L_j}$, que se define como la relevancia del concepto $c_k^{L_j}$ en el documento $d_i^{L_j}$.

2.6. Wikipedia

La Wikipedia es una enciclopedia online, de libre acceso, multilingüe y editada de forma colaborativa. Aunque inicialmente solo se encontraba disponible en inglés, rápidamente se convirtió en multilingüe al comenzar a desarrollarse versiones similares en diferentes idiomas. En la actualidad alberga más de 46 millones de artículos repartidos en 277 ediciones, siendo la versión en inglés la de mayor tamaño, con casi de 5.5 millones de artículos⁵. La información contenida en Wikipedia es de elevado interés para investigadores y desarrolladores, debido a la cantidad de conceptos definidos de forma manual que contiene y a las relaciones semánticas entre los propios conceptos.

Aunque la Wikipedia está abierta al anonimato y es editada de forma colaborativa, podemos asegurar su fiabilidad como fuente de conocimiento externo para la creación de representaciones de documentos como bolsas de conceptos. Giles [38] establece que los artículos científicos presentes en la Wikipedia tienen un nivel de precisión cercano al ofrecido por la Encyclopædia Britannica, ofreciendo un ratio similar de “errores serios”. Wikipedia cuenta con cientos de miles de editores activos que, entre otras tareas, trabajan para mantener su veracidad, bloqueando incluso aquellos artículos más susceptibles de ser saboteados. Viegas et al. [125] han detectado que los actos de vandalismo que se producen, generalmente son reparados tan rápido que la mayor parte de los usuarios ni llegan a detectar sus efectos. Viegas et al. [125] concluyen diciendo que la Wikipedia tiene “capacidades de auto reparación sorprendentemente efectivas”. Por último, la Wikipedia ha sido amplia y exitosamente utilizada como fuente de conocimiento externo en diferentes tareas de recuperación de información y clasificación de documentos [17, 28, 33, 43, 66, 68, 84, 100].

2.7. Wikipedia Miner

Entre la gran cantidad de anotadores semánticos existentes en la literatura, para la implementación de la propuesta presentada en este trabajo hemos optado por la utilización de la herramienta Wikipedia Miner (WM) [67]. Este apartado presentará sus principales características técnicas y de funcionamiento, junto con los motivos que nos han llevado a su utilización.

⁵<https://stats.wikimedia.org/EN/Sitemap.htm>

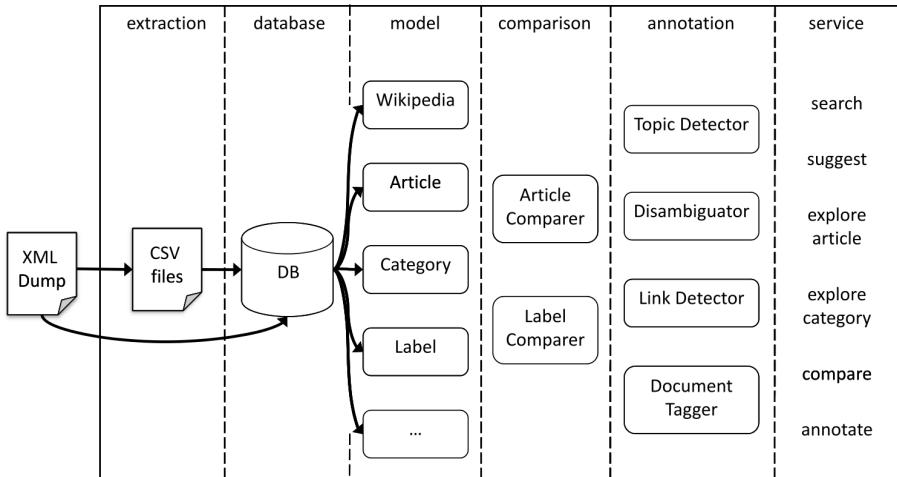


Figura 2.2: Arquitectura del anotador semántico Wikipedia Miner (extraída del trabajo de Milne and Witten [67]).

Wikipedia Miner es una herramienta de código abierto, implementada en Java, que realiza tareas de minería de datos en la Wikipedia, permitiendo así el uso de toda la información semántica que contiene de una manera sencilla, liberando a los usuarios del laborioso esfuerzo intermedio necesario. La figura 2.2 muestra la arquitectura general de la herramienta.

El proceso comienza con un archivo XML de volcado o *dump*, que contiene todo el contenido de una edición de la Wikipedia. Estos *dumps* se pueden obtener directamente desde la Wikimedia Foundation⁶. El proceso de extracción, basado en Hadoop [133], el sistema de archivos GFS de Google [37] y MapReduce [22], se encarga de procesar el *dump* para obtener una serie de archivos en texto plano que recogen la información que contiene. El paquete *model* es el principal punto de acceso de la herramienta Wikipedia Miner, proporcionando un acceso simple al contenido y estructura de la Wikipedia. El paquete *comparison* contiene los algoritmos que obtienen las medidas de relación semántica entre los conceptos contenidos en la Wikipedia. Este paquete contiene dos clases, *ArticleComparer*, que mide la relación semántica entre pares de artículos, y *LabelComparer*, que mide la relación entre pares de términos y frases. El paquete *annotation* proporciona las herramientas para identificar y etiquetar aquellos conceptos de la Wikipedia presentes en un documento de texto. Este paquete es el utilizado en este trabajo para la creación de las representaciones de los documentos en forma de bolsa de conceptos, de manera que mostraremos con un poco más detalle su funcionamiento, el cual está basado en los tres pasos siguientes (cf. figura 2.3):

- El primer paso es la selección del candidato. Dado un documento de texto compuesto por un conjunto de n-gramas – siendo un n-grama una secuencia continua de n

⁶<https://dumps.wikimedia.org/backup-index.html>

palabras – el algoritmo consulta un vocabulario que contiene todos los *anchor texts* de la edición de la Wikipedia utilizada como base de conocimiento, y para cada n-grama del documento verifica si está presente en dicho vocabulario. Para cada coincidencia entre un n-grama y un *anchor text* se obtiene un candidato, siendo los candidatos más relevantes aquellos utilizados con más frecuencia como *anchor texts* en la edición de la Wikipedia utilizada como fuente de conocimiento.

- El segundo paso es la desambiguación. Dado el mismo vocabulario de *anchor texts*, el algoritmo selecciona el destino u objetivo más adecuado para cada candidato obtenido en el paso anterior entre los elementos del vocabulario. Este proceso está basado en aprendizaje máquina, utilizando como elementos de entrenamiento artículos de la Wikipedia, ya que contienen buenos ejemplos de desambiguación realizados de forma manual. La desambiguación se realiza teniendo en cuenta la relación de cada candidato con otros términos no ambiguos del contexto, y en cuan común es la relación entre el *anchor text* y el artículo de la Wikipedia objetivo.
- El tercer paso es la detección de enlace. En este paso se calcula la relevancia de cada concepto extraído del texto. Este paso también se basa en aprendizaje máquina, utilizando de nuevo artículos de la Wikipedia como elementos de entrenamiento, debido a que cada uno de ellos es un buen ejemplo de lo que constituye un enlace relevante o no.

Por último, el paquete *service* proporciona un acceso web al conjunto de funcionalidades ofrecido por la herramienta:

- El servicio *exploreArticle* toma como entrada el título o el identificador de un artículo y devuelve detalles acerca de él, como definiciones textuales, imágenes representativas, etiquetas alternativas y enlaces a otros artículos.
- El servicio *exploreCategory* proporciona una funcionalidad similar a *exploreArticle* para las categorías.
- El servicio *search* toma como entrada una palabra o conjunto de palabras y devuelve una lista con los diferentes conceptos de la Wikipedia a los que puede referirse la entrada.
- El servicio *suggest* toma como entrada una serie de identificadores de artículos y devuelve una lista de artículos relacionados, organizados según las categorías a las que pertenecen.
- El servicio *compare* recibe como entrada una pareja de términos, una pareja de identificadores de artículos o un conjunto de identificadores de artículos y devuelve una medida de cuanto están relacionados entre sí dichos elementos.
- El servicio *wikify* detecta de forma automática los conceptos presentes en un documento de texto proporcionado como entrada, y devuelve una lista con dichos conceptos y su relevancia dentro del propio documento. La figura 2.4 muestra un

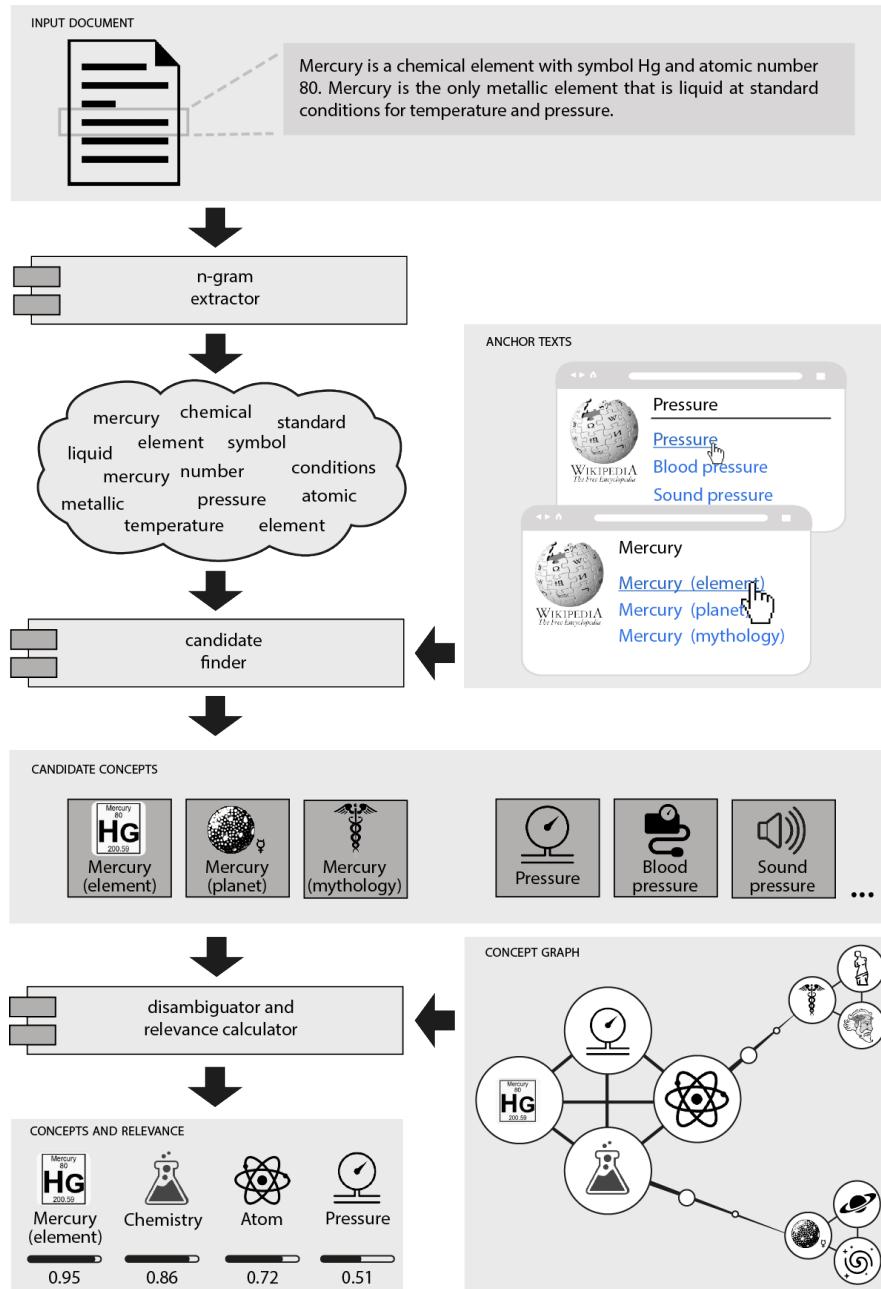


Figura 2.3: Proceso de obtención de la bolsa de conceptos de un documento de texto utilizando Wikipedia Miner.

```

<-<message service="/services/wikify" sourceMode="WIKI"
  documentScore="4.243861377239227">
  -<request>
    <param name="minProbability">0.01</param>
    -<param name="source">
      Mercury is a planet in our solar system. It is the smallest of the eight planets. It is also the closest to the Sun. Mercury goes around the sun the fastest of all the planets. Mercury has no moons.
    </param>
  </request>
  -<wikifiedDocument>
    [[Mercury (planet)|Mercury]] is a [[planet]] in our [[solar system]]. It is the smallest of the eight planets. It is also the closest to the [[Sun]]. Mercury goes around the sun the fastest of all the planets. Mercury has no [[Natural satellite|moons]].
  </wikifiedDocument>
  -<detectedTopics>
    <detectedTopic id="19694" title="Mercury (planet)" weight="0.7783115237231091"/>
    <detectedTopic id="26903" title="Solar System" weight="0.7459290315589117"/>
    <detectedTopic id="26751" title="Sun" weight="0.7290399024009098"/>
    <detectedTopic id="53306" title="Natural satellite" weight="0.7174097983160982"/>
    <detectedTopic id="22915" title="Planet" weight="0.7097966724231624"/>
  </detectedTopics>
</message>

```

Figura 2.4: Ejemplo de ejecución del servicio *wikify*.

ejemplo de la ejecución del servicio *wikify*, al que se le ha proporcionado como entrada un pequeño trozo de texto sobre el planeta Mercurio. El servicio devuelve una lista con los conceptos identificados en el texto (*detectedTopics*) junto con un valor numérico que indica su relevancia dentro del texto proporcionado (*weight*).

Los principales motivos por los que se ha utilizado Wikipedia Miner son los siguientes:

- Utiliza Wikipedia como base de datos de conocimiento, lo que a su vez proporciona las siguientes ventajas
 - Gran cantidad de información disponible.
 - Capacidad multilingüe.
 - Enlaces interlingüísticos o *interwikis*. Wikipedia proporciona enlaces entre una página presente en una edición de la Wikipedia en un idioma y su página equivalente en otro idioma.
 - Relaciones semánticas entre artículos.
- Proporciona desambiguación.
- Asigna pesos a los conceptos extraídos de un documento en función de su relevancia en este.
- Sencillez a la hora de crear anotadores semánticos en diferentes idiomas. Simplemente es necesario proporcionarle a Wikipedia Miner un *dump* de la edición de la Wikipedia en el idioma deseado.
- Es software de código abierto.

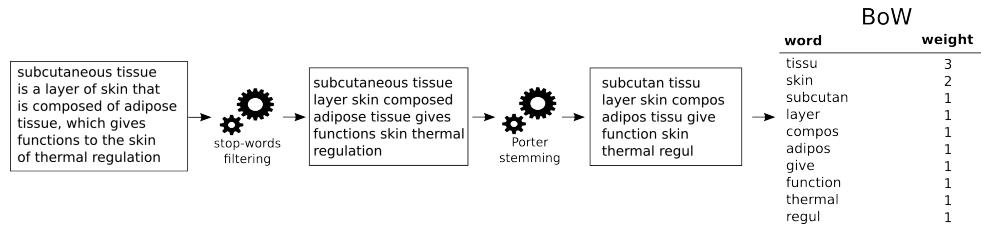


Figura 2.5: Representación de un documento según el modelo BoW.

2.8. Modelos empleados para la representación de los documentos

Esta sección detalla los diferentes modelos de representación de documentos empleados en este trabajo. Se encuentra dividida en dos subsecciones. La primera agrupa los modelos utilizados en las tareas de clasificación monolingüe, mientras que la segunda reúne aquellos empleados en las tareas de clasificación entre idiomas o multilingüe.

2.8.1. Clasificación monolingüe

Se presentan en esta sección los dos modelos empleados para representar los documentos en las tareas de clasificación monolingüe realizadas como parte de este trabajo: BoW y WikiBoC.

2.8.1.1. BoW

Como se ha descrito en la sección 2.5.1, el modelo BoW representa cada documento como un vector de pesos de términos, donde cada término es una palabra, y el peso de cada componente del vector es la frecuencia de ocurrencia de cada palabra en el propio documento. Para crear las representaciones BoW de los documentos (cf. figura 2.5), el primer paso consiste en filtrar las *stop words*, y tras esto se aplica un algoritmo de *stemming* con el objetivo de crear características más genéricas. Para la realización de este proceso utilizamos el algoritmo de Porter [99], el algoritmo de *stemming* más utilizado para textos escritos en inglés [118]. Por último, se obtiene la frecuencia de ocurrencia de cada característica en el documento.

2.8.1.2. WikiBoC

De acuerdo a la sección 2.5.2, el modelo Wikipedia Bag of Concepts (WikiBoC) [80] representa cada documento como un vector de pesos de términos, donde cada

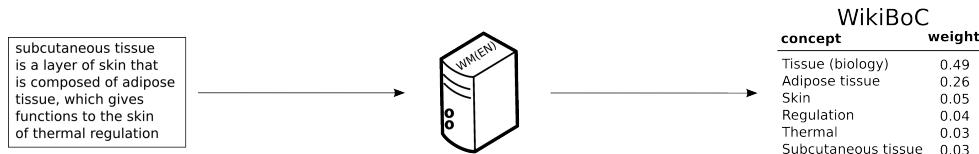


Figura 2.6: Representación de un documento según el modelo WikiBoC.

término es un concepto (artículo, entrada) de la Wikipedia, extraídos por medio del anotador semántico de propósito general Wikipedia Miner (cf. figura 2.6). El peso de cada componente del vector es el peso que Wikipedia Miner asigna a cada concepto en función de la relevancia de este en el propio documento.

2.8.2. Clasificación *cross-lingue* o multilingüe

Esta sección presenta los modelos empleados para representar los documentos en las tareas de clasificación entre idiomas o multilingüe realizadas en este trabajo – BoW-MT, ESA, Bi-LDA, BWE, MetaMap, Hybrid-WikiBoC, y WikiBoC-CLCM – así como la técnica *Cross-Language Concept Matching* (CLCM).

2.8.2.1. BoW-MT

La traducción automática de documentos (en inglés *Machine Translation* (MT)) [51] es un área de la lingüística computacional que investiga la utilización de herramientas software – como Google Translate⁷ o Bing Translator⁸ – para traducir documentos entre idiomas. La combinación de la representación BoW y las técnicas de traducción automática (BoW-MT) (cf. figura 2.7) es una de las propuestas más maduras y utilizadas para abordar tareas de clasificación multilingüe de documentos de texto [27]. A su vez, estas técnicas se dividen en dos categorías generales: la traducción de los documentos completos y la traducción de las características que los representan. La primera propuesta consiste en realizar en primer lugar la traducción de los documentos completos para posteriormente obtener la representación en forma de bolsa de palabras. Existen dos variantes, la traducción de los documentos de la secuencia de entrenamiento al idioma de la secuencia de los elementos a clasificar [102] y la traducción de los documentos a clasificar al idioma de los documentos de la secuencia de entrenamiento [61, 129]. Por el contrario, las técnicas de la segunda propuesta extraen en primer lugar las características (en este caso palabras) de los documentos, para posteriormente proceder a su traducción. Al igual que con las técnicas de traducción de los documentos completos, existen dos variantes: la traducción de las características de los documentos de la secuencia de entrenamiento al idioma de los documentos a clasificar [110] y la traducción de las características de los documentos a clasificar al idioma de los documentos de la secuencia de entrenamiento [5, 91, 110, 132].

⁷<https://translate.google.com>

⁸<https://www.bing.com/translator>

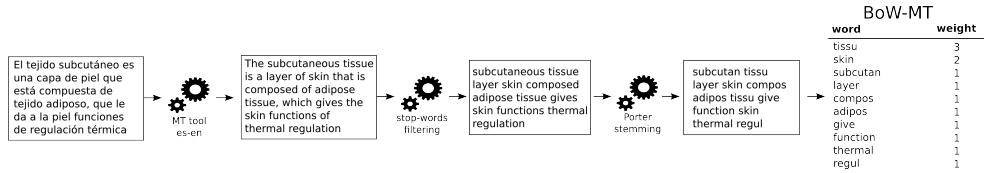


Figura 2.7: Representación de un documento según el modelo BoW-MT.

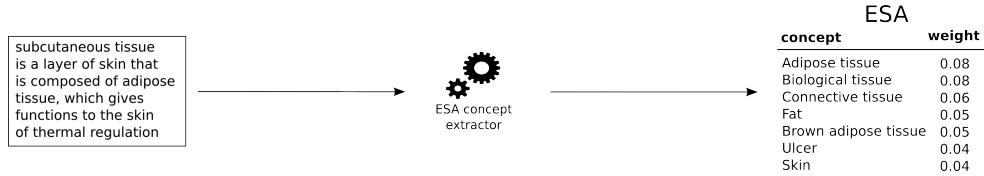


Figura 2.8: Representación de un documento según el modelo ESA.

2.8.2.2. ESA

Explicit Semantic Analysis (ESA) [34] es una representación vectorial que hace uso de fuentes de conocimiento externo (generalmente la Wikipedia) para la generación de las características que componen dicho vector. ESA realiza un análisis semántico explícito del texto, identificando cuestiones que se encuentran presentes de forma explícita en la base de conocimiento empleada. En otras palabras, y suponiendo la Wikipedia como base de conocimiento, ESA indexa los documentos con conceptos de la Wikipedia basándose en un análisis completo del texto, es decir, ESA indexa un texto con artículos de la Wikipedia que se solapan con el propio texto.

De acuerdo a la sección 2.5.2, ESA representa cada documento como un vector de pesos de conceptos, donde cada concepto es un artículo de la Wikipedia (cf. figura 2.8). El peso de cada componente del vector es el peso que ESA asigna a cada concepto en función de cuan relacionado esté con el contenido del documento.

2.8.2.3. Bi-LDA

Latent Dirichlet Allocation (LDA) [7] es un modelo estadístico generativo que permite que conjuntos de observaciones puedan ser descritas por grupos de variables ocultas, no observables o latentes, que describen por qué algunas partes de los datos son similares. Por ejemplo, si las observaciones son palabras presentes en documentos, LDA asume o presupone que cada documento es una mezcla de un pequeño número de temas, y que la creación o “generación” de las palabras del documento es atribuible a las cuestiones o temas sobre los que trata el documento. Siguiendo esta idea, LDA encuentra esas cuestiones de forma automática dentro del texto, es decir, LDA intenta “volver hacia atrás” a partir de las palabras de los documentos, para obtener los temas a partir de los cuales podrían haberse “generado”.

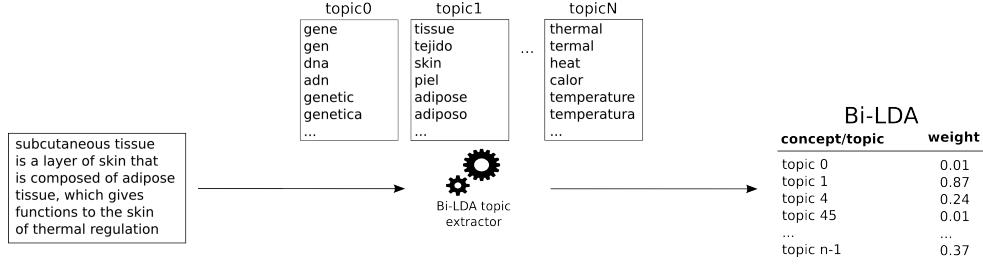


Figura 2.9: Representación de un documento según el modelo Bi-LDA.

El modelo Bilingual LDA (Bi-LDA) [21] es una extensión del modelo LDA, que permite extraer conceptos multilingües a partir de conjuntos de datos paralelos o comparables, para representar los documentos en función de dichos conceptos (cf. figura 2.9). La implementación de la propuesta Bi-LDA utilizada en este trabajo se ha realizado de acuerdo a la metodología propuesta por De Smet et al. [21], Ni et al. [88] y Vulić et al. [126], descrita detalladamente en Mouriño-García et al. [80].

2.8.2.4. BWE

Los *Word Embeddings* [6, 127] han sido introducidos recientemente como formas ricas y coherentes de representar palabras. *Word Embeddings* es el nombre colectivo de un conjunto de técnicas de procesado del lenguaje natural para el modelado del lenguaje y creación de características (*feature learning*), donde las palabras o frases se representan como vectores de números reales. De forma conceptual, esta propuesta implica una integración (*embedding*) de un espacio con una dimensión por palabra a un espacio continuo de vectores de mucha menor dimensionalidad.

Bilingual Word Embeddings (BWE) es una extensión del modelo *Word Embeddings* del escenario monolingüe al multilingüe [58, 120] que permite descubrir *embeddings* para palabras que definen conceptos similares que están muy próximos en el espacio bilingüe compartido. Es decir, las representaciones para la misma palabra en dos idiomas diferentes debieran ser muy similares. Entonces, estos *embeddings* bilingües pueden ser utilizados en diferentes escenarios de aplicación (como recuperación de información o clasificación de documentos entre idiomas) utilizando el espacio de *embeddings* bilingüe inducido.

La implementación de la propuesta BWE se ha realizado de acuerdo a las directivas metodológicas de los trabajos de Vulić and Moens [127] y Upadhyay et al. [122], detalladas en Mouriño-García et al. [80].

2.8.2.5. MetaMap

El modelo MetaMap está basado en la utilización del anotador semántico específico del dominio biomédico homónimo [3], comúnmente utilizado para identificar conceptos

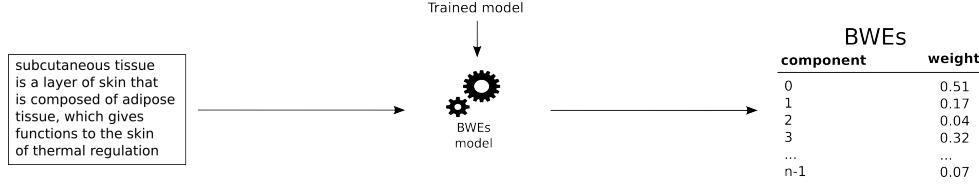


Figura 2.10: Representación de un documento según el modelo BWE.



Figura 2.11: Representación de un documento según el modelo MetaMap.

biomédicos en documentos de texto y mapearlos a entradas del metatesauro UMLS [19]. De acuerdo a la sección 2.5.2 y al trabajo de Carrero et al. [13], el modelo MetaMap representa cada documento como un vector de pesos de conceptos, donde cada concepto es una entrada del metatesauro UMLS (cf. figura 2.11). El peso de cada componente del vector es el peso que el anotador semántico MetaMap asigna a cada uno de los conceptos extraídos del documento.

2.8.2.6. Hybrid-WikiBoC

El modelo Hybrid-WikiBoC [80] combina los modelos WikiBoC y BoW. En este trabajo hemos utilizado una combinación de ambas que consiste en el enriquecimiento de la representación BoW de cada documento con los conceptos extraídos del mismo documento utilizando el anotador semántico Wikipedia Miner (cf. figura 2.12)

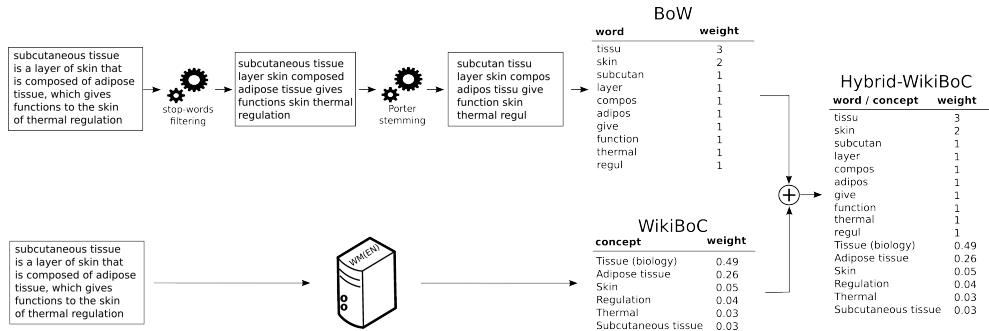
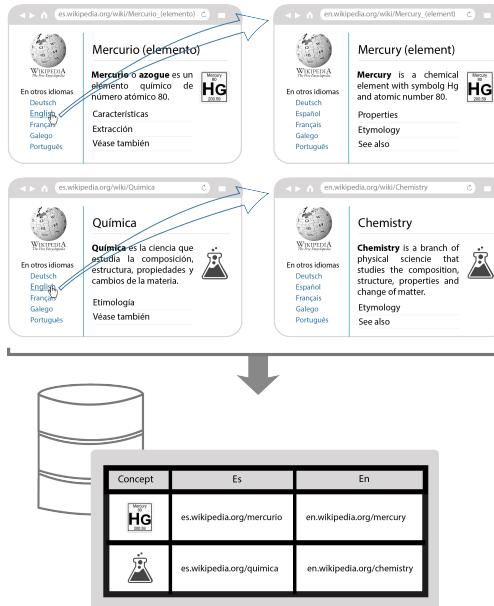


Figura 2.12: Representación de un documento según el modelo Hybrid-WikiBoC.

Figura 2.13: *Cross-Language Concept Matching*.

2.8.2.7. La técnica *Cross-Language Concept Matching*

La técnica *Cross-Language Concept Matching* (CLCM) [80] convierte la representación en forma de bolsa de conceptos Wikipedia de un documento en un idioma L_m a otro idioma diferente L_n . La base de esta técnica subyace en los enlaces interlingüísticos de la Wikipedia, que enlazan un artículo en un idioma a su equivalente en otro idioma. Por ejemplo, partiendo del artículo *Mercurio_(elemento)*, perteneciente a la edición de la Wikipedia en castellano, podemos obtener el correspondiente artículo en inglés (*Mercury_(element)*), en francés (*Mercure_(chimie)*) o en alemán (*Quecksilber*) (cf. figura 2.13). Así, para convertir la representación BoC de un idioma L_m a un idioma L_n únicamente es necesario obtener el concepto (artículo) equivalente en el idioma L_n para cada concepto de la bolsa en el idioma L_m . La figura 2.14 muestra el proceso de conversión de la representación BoC de un documento escrito en castellano a inglés. La tabla representa la función de transformación lineal $\tilde{\mathcal{M}}$, que mapea los conceptos de la Wikipedia en un idioma L_m – en este caso castellano – a otro idioma diferente L_n – en este caso inglés. Entonces, para cada par de conceptos, si existe una correspondencia entre ellos se marcará como *true* (T), y si no existe tal correspondencia, se marcará como *false* (F). La matriz de transformación \mathbf{M} se obtiene de la anterior tabla, siendo “1” el valor de cada componente si existe correspondencia y “0” si no existe. La bolsa de conceptos en castellano se representa mediante el vector d^{L_m} . Finalmente, el vector d^{L_m} se multiplica por la matriz \mathbf{M} para obtener el vector d^{L_n} , que representa la bolsa de conceptos del documento convertida de castellano a inglés.

Es necesario destacar que el espacio de características – conceptos – según el cual se

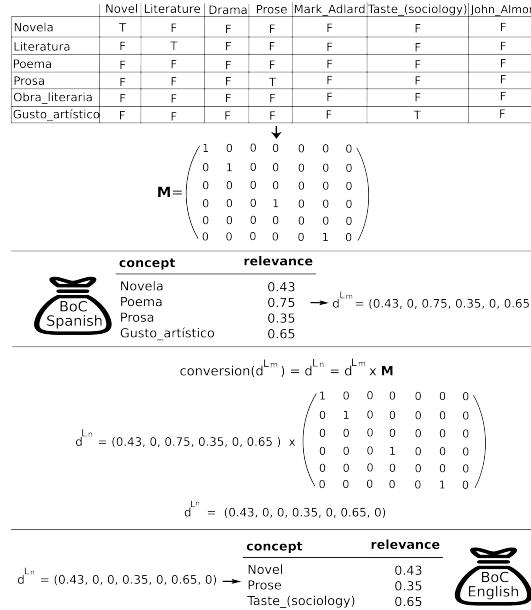


Figura 2.14: Conversión de la representación WikiBoC de un documento escrito en castellano a su equivalente en inglés.

representa un documento de texto está limitado al conjunto de artículos presentes en la edición de la Wikipedia en cada uno de los idiomas. En el ejemplo presente en la figura 2.14, un documento de texto escrito en inglés se representa a través del conjunto de conceptos de la edición en inglés de la Wikipedia. De la misma forma, un documento de texto escrito en castellano se representa por medio del conjunto de conceptos de la edición en castellano de la Wikipedia. Como hemos visto, para convertir la representación WikiBoC de un documento entre idiomas hacemos uso de los enlaces interlingüísticos de Wikipedia. De esta forma, un documento en castellano que se ha convertido a inglés se encuentra representado en el espacio de artículos resultante de la intersección de los artículos de las ediciones de la Wikipedia en castellano e inglés, o lo que es lo mismo, el conjunto de artículos que se encuentran enlazados entre las dos ediciones de la Wikipedia.

Las principales ventajas de esta propuesta son las siguientes: *i*) minimiza la pérdida semántica en el proceso de conversión, debido a la correspondencia entre las distintas versiones de un mismo artículo en diferentes idiomas; y *ii*) no se necesitan corpus paralelos o comparables, ni diccionarios bilingües [112].

A pesar de que la literatura contiene varios trabajos que explotan los enlaces interlingüísticos entre artículos de la Wikipedia, su aplicación está centrada en la recuperación de información [17, 100, 112] y en el cálculo de la relación semántica entre idiomas [43], y no tenemos constancia en su aplicación a la clasificación de documentos.

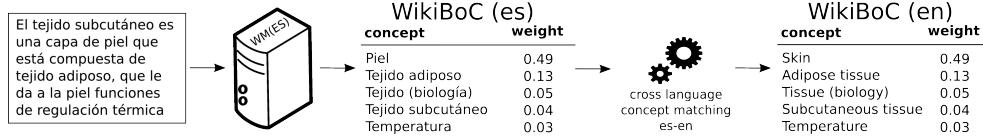


Figura 2.15: Representación de un documento según el modelo WikiBoC-CLCM.

2.8.2.8. WikiBoC-CLCM

El modelo WikiBoC-CLCM [80] se basa en la combinación de la representación WikiBoC y la técnica CLCM. En primer lugar, se obtiene la representación WikiBoC de un documento en un idioma L_m , a través de una instancia del anotador semántico Wikipedia Miner que utiliza como base de conocimiento un *dump* de la Wikipedia en el idioma L_m . Tras esto, se utiliza la técnica *Cross-Language Concept Matching* para convertir la representación WikiBoC del idioma L_m al idioma L_n (cf. figura 2.15).

2.9. Conjuntos de datos

Esta sección describe brevemente los diferentes conjuntos de datos utilizados para la evaluación de las propuestas presentadas en esta investigación. Se describen tanto aquellos conjuntos de datos ya existentes en el estado del arte (OHSUMED y JRC-Acquis) como aquellos creados de forma expresa durante la realización de este trabajo (UVigoMED, CL-UVigoMED, Wikipedia Corpus y Wikipedia Human Medicine Corpus), debido a que existen diferentes aplicaciones o situaciones para las cuales no existe un conjunto de datos estándar para su evaluación [110]. Estos últimos han sido generados a partir de la extracción de información de diversos sitios web, haciendo uso de técnicas de *web scraping* [70].

2.9.1. OHSUMED

La compilación original del corpus OHSUMED [47] está compuesta por 348,566 abstracts biomédicos procedentes de la base de datos MEDLINE, entre los años 1978 y 1991. De los 50,216 documentos del año 1991 que contenían abstracts, Joachims [54] seleccionó los 10,000 primeros como secuencia de entrenamiento y los siguientes 10,000 como secuencia de test, los cuales se encuentran clasificados en una o varias de las 23 subcategorías de la categoría *Diseases* del árbol MeSH. Este subconjunto (al que se refiere de igual manera como OHSUMED) conforma uno de los corpus de referencia para la realización de experimentos de clasificación automática de documentos.

2.9.2. JRC-Acquis

El corpus JRC-Acquis [115] es un conjunto de datos multilingüe compuesto por documentos disponibles en 22 idiomas acerca de temas legales de la Unión Europea, lo que lo hace particularmente adecuado para realizar todo tipo de investigación multilingüe. Para la realización de este trabajo se han utilizado únicamente aquellos documentos escritos en inglés y en castellano, dando lugar a un corpus formado por una secuencia de entrenamiento que cuenta con 20,411 documentos escritos en inglés y una secuencia de test compuesta por 20,507 documentos escritos en castellano. El proceso de creación de este conjunto de documento se encuentra detallado en [80].

2.9.3. UVigoMED

El corpus UVigoMED [74] está compuesto por 92,651 abstracts biomédicos extraídos de MEDLINE escritos en inglés, clasificados en 26 categorías. Siguiendo el procedimiento de creación del corpus OHSUMED, el conjunto de clases está formado por las subcategorías de la categoría *Diseases* del árbol MeSH. Es necesario destacar que este corpus fue creado durante el año 2015, por lo cual se utilizó el árbol MeSH del año 2015, en el cual el grupo *Diseases* está compuesto por 26 categorías en lugar de las 23 que contenía cuando se creó originalmente OHSUMED. El proceso de creación de este conjunto de datos se encuentra detallado en Mouriño-García et al. [76].

2.9.4. CL-UVigoMED

Cross-Language UVigoMED (CL-UVigoMED) [79] es un corpus bilingüe (inglés-castellano) compuesto por 12,832 abstracts biomédicos extraídos de MEDLINE escritos en inglés, y 2,184 escritos en castellano. La construcción de CL-UVigoMED se ha realizado siguiendo la metodología utilizada en la creación de los corpus OHSUMED y UVigoMED, de manera que podemos considerar el corpus CL-UVigoMED como una evolución directa de este último. El proceso de creación se encuentra detallado en Mouriño-García et al. [75].

2.9.5. Wikipedia Corpus

Wikipedia Corpus [80] es un corpus bilingüe (inglés-castellano) compuesto por 3,851 documentos extraídos de la Wikipedia, clasificados en las siguientes categorías: *Culture and the arts*, *Geography and places* y *Mathematics and logic*. El conjunto se divide en una secuencia de entrenamiento formada por 3,019 documentos escritos en inglés y una secuencia de test compuesta por 832 documentos escritos en castellano. El proceso de creación de este conjunto de datos se encuentra detallado en Mouriño-García et al. [77].

2.9.6. Wikipedia Human Medicine Corpus

Wikipedia Human Medicine Corpus [80] es un corpus bilingüe (inglés-castellano) compuesto por 2,612 documentos extraídos de la Wikipedia acerca de cuestiones biomédicas relacionadas con la medicina humana, clasificados en las 22 siguientes categorías: *Alternative medicine, Cardiology, Endocrinology, Forensics, Gastroenterology, Human genetics, Geriatrics, Gerontology, Gynecology, Hematology, Nephrology, Neurology, Obstetrics, Oncology, Ophthalmology, Orthopedical surgical procedures, Pathology, Pediatrics, Psychiatry, Rheumatology, Surgery and Urology*. El conjunto se divide en una secuencia de entrenamiento compuesta por 2,143 documentos escritos en inglés, y una secuencia de test que contiene 469 documentos escritos en castellano. El proceso de creación de este conjunto de datos se encuentra detallado en Mouriño-García et al. [78].

Capítulo 3

Resultados y discusión

Este capítulo presenta y discute los resultados obtenidos de la realización de esta tesis. Con el objetivo de facilitar el seguimiento de la investigación realizada, el capítulo se encuentra estructurado en tres secciones. Cada una de estas secciones contiene a su vez la información necesaria para la contextualización de los resultados presentados. La sección 3.1 sirve como paso previo a las propuestas centrales de esta tesis. En ella se muestran y discuten los resultados obtenidos de la utilización de las técnicas de clasificación automática de documentos en el ámbito educativo, concretamente para proporcionar interoperabilidad entre repositorios de recursos educativos, como alternativa a las técnicas clásicas basadas en el mapeado de taxonomías. Tras esto, la investigación se centra en la aplicación de la representación WikiBoC a la clasificación automática de documentos. En particular, la sección 3.2 presenta los resultados de la aplicación de la representación WikiBoC a la clasificación monolingüe de documentos en uno de los dominios más relevantes para su aplicación, el dominio biomédico. Por último, la sección 3.3 presenta y discute los resultados obtenidos de la aplicación de la representación WikiBoC a la clasificación multilingüe de documentos en diversos ámbitos de aplicación, como el biomédico y el legal.

3.1. La clasificación automática de documentos y la interoperabilidad entre repositorios de recursos educativos

Uno de los principales objetivos de las tareas de gestión de información – y en concreto de la clasificación de documentos – es hacer que los usuarios sean conscientes de su existencia y proporcionar acceso a ella [20, 114].

El estudio presentado en Mouríño-García et al. [81] se trata de una primera toma de contacto previa a las propuestas centrales de este trabajo de investigación. Consiste en la implementación y validación de CROERA, un agregador de repositorios de recursos educativos, basado en la aplicación de técnicas de clasificación automática y la representación tradicional de los documentos en forma de bolsa de palabras, como alternativa a la utilización de las técnicas clásicas de mapeado entre las taxonomías de los diferentes repositorios agregados.

3.1.1. Estado del arte

Debido a la gran cantidad de repositorios de recursos educativos existentes, han surgido diversas alianzas entre repositorios, redes y agregadores para fomentar la compartición y reutilización de material educativo, como ARIADNE [119], MACE [10], MELT [59], GLOBE, ELENA [26] u Open Discovery Space [89]. El principal desafío de la agregación de repositorios es la interoperabilidad, es decir, un acceso integrado y unificado a los recursos, independientemente del esquema de metadatos utilizado [114].

De hecho, la heterogeneidad de las taxonomías de los repositorios existentes [23] ha provocado la aparición de diferentes propuestas para superarla, siendo las más utilizadas las técnicas de mapeado entre taxonomías [25, 86, 87, 117]. El principal inconveniente de estas técnicas es la situación de mapeado “uno a ninguno” (*one-to-one matching*). Así, cuando esta situación sucede, los elementos clasificados bajo las categorías que no han podido ser mapeadas no serán accesibles desde una taxonomía diferente a la original [14, 48].

3.1.2. Enfoque

CROERA proporciona acceso a todos los recursos educativos indexados con independencia de la taxonomía utilizada por cada uno de los repositorios integrados, contribuyendo de manera relevante a facilitar la localización de recursos educativos. La propuesta presentada elimina el problema de la heterogeneidad de taxonomías y permite la realización de búsquedas exploratorias [85, 103] a través de cualquiera de las taxonomías de los repositorios agregados, sin necesidad de realizar ningún tipo de mapeado entre ellas, evitando así los inconvenientes de dichas técnicas.

CROERA funciona de acuerdo a los pasos presentados a continuación (cf. figura 3.1). Es necesario destacar que, aunque el agregador presentado ha sido desarrollado para integrar cualquier número de repositorios, para la versión inicial se han seleccionado OERCommons [90], MERLOT [11] y Open Stax CNX [92], debido a la alta calidad de los metadatos que contienen, así como por estar mantenidos por comunidades de expertos. En primer lugar se obtienen los recursos educativos de los repositorios integrados, a través de la utilización de técnicas de web *scraping* [70]. Tras esto, y para cada uno de los repositorios integrados, se entrena un algoritmo de clasificación SVM con la información

obtenida de los metadatos (concretamente el título, la descripción y las palabras clave) de los recursos educativos obtenidos en el paso anterior. Por último, todos los recursos obtenidos en el primer paso son categorizados utilizando los clasificadores entrenados en el paso dos. Esto permite a los usuarios el acceso al 100 % de los recursos educativos indexados por el agregador, utilizando la taxonomía que se considere más adecuada, ya que todos ellos se encuentran clasificados de acuerdo a cada una de las taxonomías de los repositorios integrados.

3.1.3. Resultados

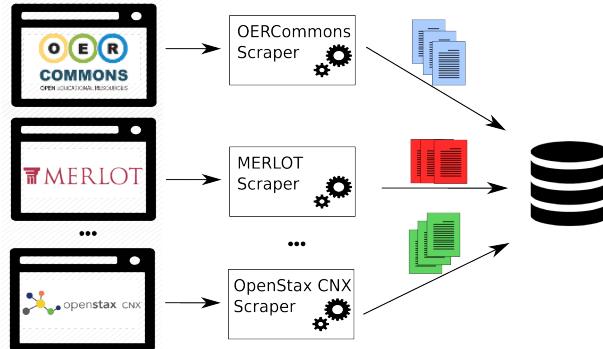
Debido a las diferencias en las taxonomías de los repositorios integrados, la evaluación del rendimiento de la propuesta presentada fue llevada a cabo utilizando dos estrategias complementarias. Por un lado, la calidad de la clasificación de aquellos recursos clasificados originalmente en el conjunto de categorías comunes a los tres repositorios se realizó automáticamente, proporcionando valores medios de precisión, retirada y $F_1 - score$ del 80 %, 74 % y 77 % respectivamente.

Por otro lado, para evaluar la calidad de la clasificación de aquellos recursos no clasificados bajo las categorías comunes, se utilizó un enfoque alternativo basado en el juicio de expertos humanos como referencia o verdad base para calcular el rendimiento de la clasificación realizada por CROERA. En este caso, los valores medios de precisión, retirada y $F_1 - score$ se reducen al 62 %, 52 % y 56.9 % respectivamente. El procedimiento seguido en cada una de las estrategias, así como los resultados detallados, se encuentran en Mouríño-García et al. [81].

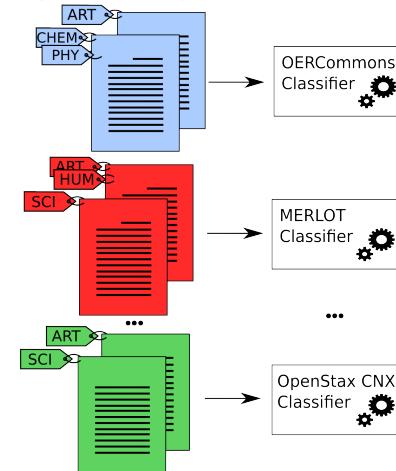
3.1.4. Discusión

La propuesta presentada en CROERA resuelve el principal problema de las técnicas de mapeado entre taxonomías, el mapeado “uno a ninguno”, proporcionando acceso al 100 % de los recursos indexados independientemente de la taxonomía seleccionada, ya que todos ellos se encuentran clasificados de acuerdo a las tres taxonomías. Los resultados presentados en la anterior sección muestran que CROERA proporciona acceso a la totalidad de los recursos con un rendimiento medio del 77 % para aquellos recursos clasificados bajo alguna de las categorías comunes a los tres repositorios, y del 56.9 % para aquellos no clasificados bajo las categorías comunes, mostrando así el potencial de la propuesta presentada como herramienta para la clasificación de recursos educativos. Esto podría aplicarse, por ejemplo, para proporcionar una clasificación preliminar automática que únicamente requeriría correcciones menores por parte de los clasificadores humanos.

Step 1: Resources collection



Step 2: Training



Step 3: Classification

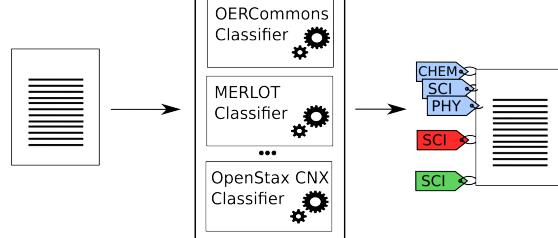


Figura 3.1: Funcionamiento de CROERA.

3.2. La representación WikiBoC y la clasificación monolingüe

Tras una primera exploración de la aplicación de la clasificación automática de documentos a la categorización de recursos educativos representados según el modelo BoW, nos hemos centrado en la clasificación de documentos utilizando la representación basada en conceptos de la Wikipedia (WikiBoC). Entre las múltiples aplicaciones de la clasificación automática de documentos [108], la clasificación de literatura biomédica se erige como un importante dominio de aplicación. El personal médico, científicos e investigadores biomédicos manejan grandes volúmenes de literatura e información biomédica diariamente [39], y la capacidad de revisar de forma eficiente la literatura existente es esencial para el rápido progreso de la ciencia [98].

3.2.1. Estado del arte

La representación más utilizada en tareas de clasificación monolingüe de documentos – y por extensión en tareas de clasificación de literatura biomédica – es el modelo bolsa de palabras [52, 113, 118, 138]. Como hemos visto previamente, este modelo no es óptimo, ya que solo tiene en cuenta la frecuencia de ocurrencia de las palabras en los documentos, ignorando la semántica y las relaciones semánticas entre ellas, lo que propicia la aparición de problemas del lenguaje que afectan a la calidad de la clasificación, como la sinonimia, polisemia, ortogonalidad [28, 50, 69, 131], hipónimia, hiperónimia [60, 130, 131], la dispersión de los datos y la diversidad de uso de las palabras [53, 118]. Con el propósito de solucionar los problemas inherentes al modelo BoW, varios autores, como Gabrilovich and Markovitch [34], Kim et al. [56], Phan et al. [97], Täckström [118], Zhang et al. [139], Zheng et al. [141] han realizado diversos intentos para la creación de representaciones de los documentos como bolsas de conceptos y su aplicación a tareas de clasificación de literatura biomédica monolingüe.

3.2.2. Enfoque

Mouriño-García et al. [74] explora y analiza la aplicación del modelo WikiBoC a la tarea de clasificación de literatura biomédica monolingüe, concretamente en inglés. Para ello, se presenta el diseño, el desarrollo y la evaluación comparativa de la eficiencia de un clasificador que hace uso de la representación de los documentos WikiBoC, así como los beneficios aportados por la aplicación del clasificador propuesto. Para evaluar el rendimiento de la propuesta presentada se realizaron diversos experimentos utilizando los conjuntos de datos OHSUMED y UVigoMED [74, 76] y las representaciones WikiBoC y BoW.

La metodología utilizada para la realización de los experimentos consta de los siguientes pasos, que se pueden seguir de forma gráfica a través de la figura 3.2. En primer lugar, es

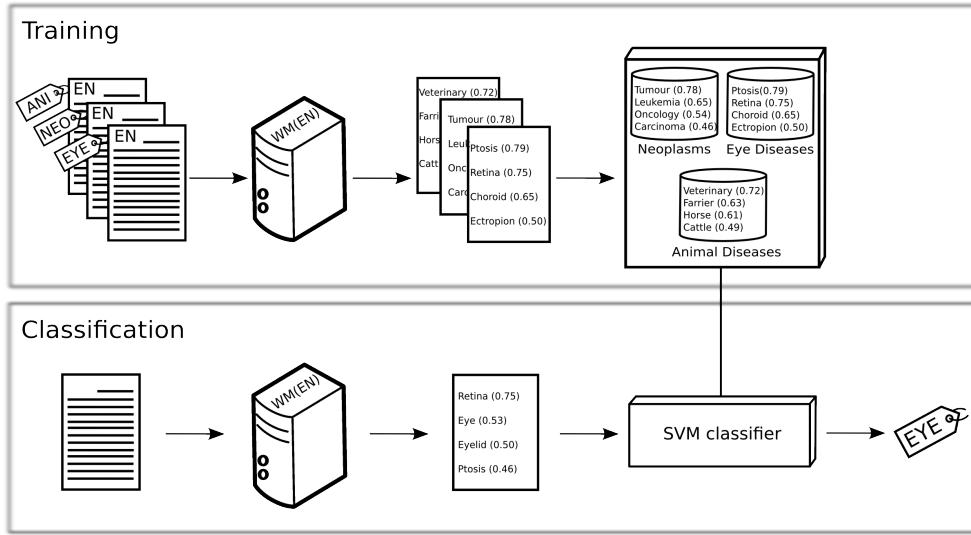


Figura 3.2: Arquitectura del clasificador SVM utilizando la representación WikiBoC.

necesario obtener la representación en forma de bolsa de conceptos de la Wikipedia de los conjuntos de documentos involucrados en los experimentos, tanto aquellos pertenecientes a la secuencia de entrenamiento como aquellos pertenecientes a la secuencia de test. Para ello, se hace uso del anotador semántico utilizado en esta investigación, Wikipedia Miner. Tras esto, las representaciones WikiBoC del conjunto de documentos de entrenamiento son utilizadas para entrenar el algoritmo seleccionado para la realización de esta investigación, el algoritmo SVM. Una vez entrenado el clasificador, las representaciones en forma de bolsa de conceptos de los elementos a clasificar (secuencia de test) son introducidas en el clasificador para que este prediga la categoría o categorías a las que pertenecen. Por último, los anteriores pasos se repiten utilizando la representación de los documentos basada en palabras, y se compara el rendimiento de clasificación ofrecido por ambas propuestas utilizando las métricas de evaluación presentadas en la sección 2.4.

3.2.3. Resultados

La figura 3.3 y la tabla 3.1, y la figura 3.4 y la tabla 3.2 muestran la evolución de los valores de precisión, retirada y $F_1 - score$ proporcionados por el clasificador utilizando las representaciones WikiBoC y BoW, en función de la longitud de la secuencia de entrenamiento, para el corpus OHSUMED, en sus versiones etiqueta única y multietiqueta. Se percibe de forma clara en la figura 3.5 que el rendimiento ofrecido por el clasificador cuando se utiliza la representación de los documentos basada en conceptos es claramente superior al ofrecido cuando se utiliza el modelo basado en palabras, obteniendo incrementos en el rendimiento de hasta un 157 %.

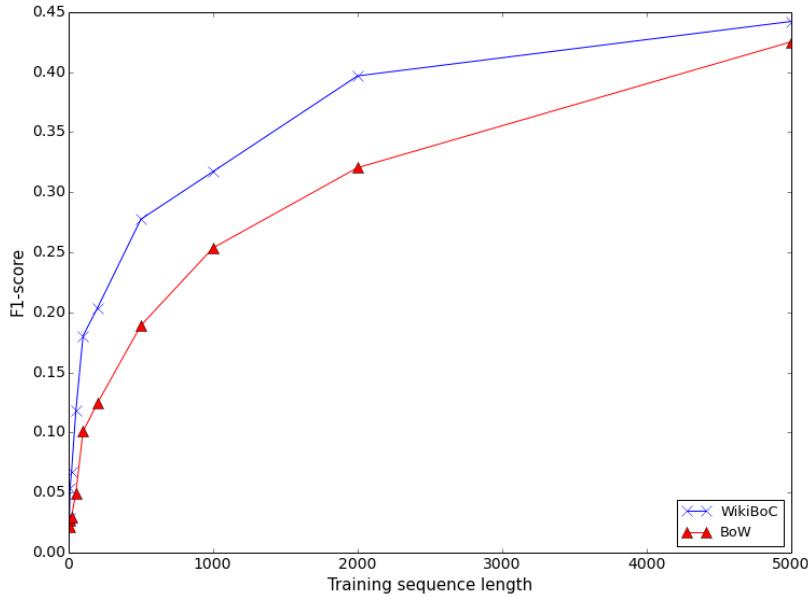


Figura 3.3: $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus OHSUMED (etiqueta única).

	5	10	20	50	100	200	500	1000	2000	5000	
BoW	P	0.058	0.080	0.102	0.160	0.218	0.276	0.345	0.418	0.467	0.528
	R	0.129	0.074	0.106	0.163	0.248	0.307	0.377	0.426	0.471	0.519
	F1	0.027	0.021	0.030	0.050	0.101	0.125	0.189	0.254	0.320	0.425
WikiBoC	P	0.089	0.134	0.213	0.281	0.308	0.332	0.421	0.460	0.512	0.535
	R	0.078	0.151	0.173	0.237	0.309	0.355	0.421	0.470	0.502	0.535
	F1	0.029	0.054	0.067	0.118	0.181	0.204	0.277	0.317	0.397	0.442

Tabla 3.1: Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus OHSUMED (etiqueta única).

	5	10	20	50	100	200	500	1000	2000	5000	
BoW	P	0.180	0.063	0.147	0.300	0.380	0.461	0.507	0.532	0.560	0.571
	R	0.001	0.002	0.031	0.047	0.056	0.082	0.200	0.288	0.385	0.482
	F1	0.001	0.002	0.019	0.028	0.041	0.080	0.172	0.244	0.343	0.424
WikiBoC	P	0.021	0.063	0.300	0.415	0.457	0.526	0.543	0.553	0.556	0.591
	R	0.001	0.001	0.054	0.085	0.121	0.182	0.273	0.340	0.404	0.481
	F1	0.001	0.001	0.038	0.042	0.080	0.156	0.238	0.289	0.364	0.438

Tabla 3.2: Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus OHSUMED (multietiqueta).

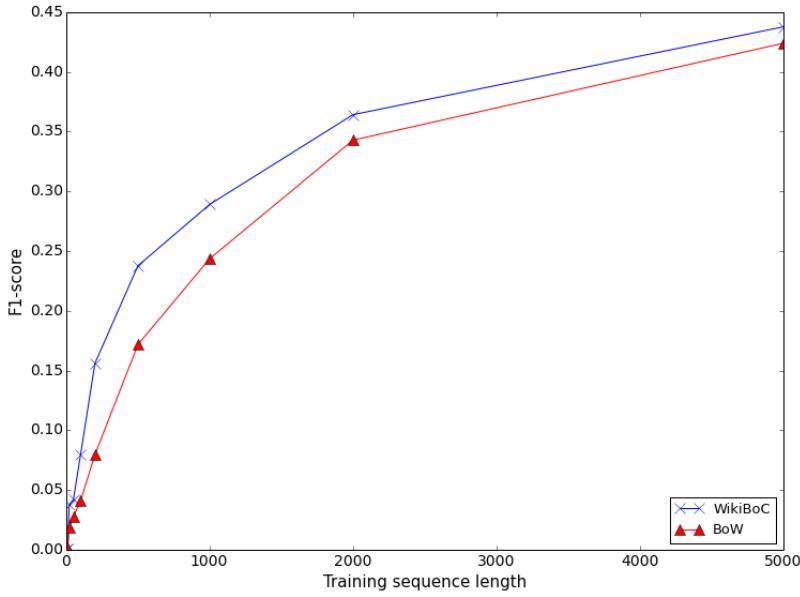


Figura 3.4: F_1 – score en función de la longitud de la secuencia de entrenamiento para el corpus OHSUMED (multietiqueta).

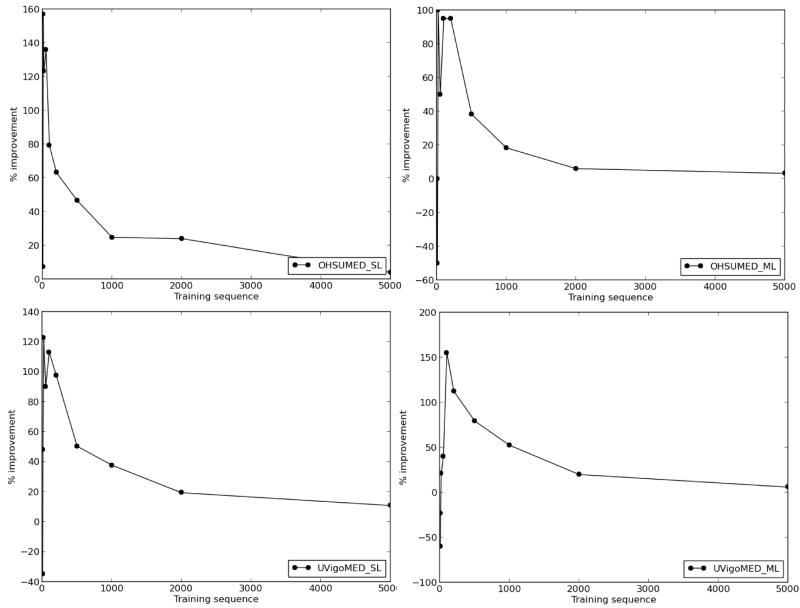


Figura 3.5: Incremento del rendimiento (en %) de la representación WikiBoC sobre la representación BoW para los corpus OHSUMED y UVigoMED.

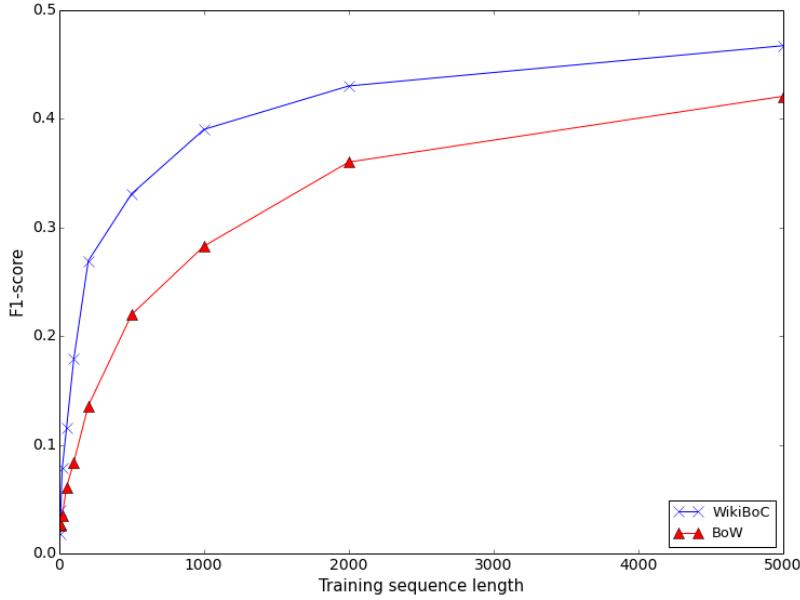


Figura 3.6: $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus UVigoMED (etiqueta única).

De la misma forma, la figura 3.6 y la tabla 3.3, y la figura 3.7 y la tabla 3.4 muestran la evolución, en función de la longitud de la secuencia de entrenamiento, de las métricas precisión, retirada y $F_1 - score$ proporcionadas por el clasificador para el corpus UVigoMED, en sus versiones etiqueta única y multietiqueta. La figura 3.5 muestra como la representación basada en conceptos ofrece mejoras en el rendimiento de hasta un 155 % sobre el modelo BoW.

	5	10	20	50	100	200	500	1000	2000	5000	
BoW	P	0.059	0.122	0.102	0.116	0.179	0.276	0.377	0.460	0.518	0.629
	R	0.060	0.074	0.097	0.150	0.183	0.272	0.397	0.457	0.511	0.631
	F1	0.026	0.027	0.035	0.061	0.084	0.136	0.220	0.283	0.360	0.421
WikiBoC	P	0.095	0.222	0.284	0.259	0.308	0.436	0.500	0.544	0.586	0.594
	R	0.049	0.093	0.148	0.247	0.321	0.432	0.515	0.557	0.590	0.598
	F1	0.017	0.040	0.078	0.116	0.179	0.269	0.331	0.390	0.430	0.467

Tabla 3.3: Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus UVigoMED (etiqueta única).

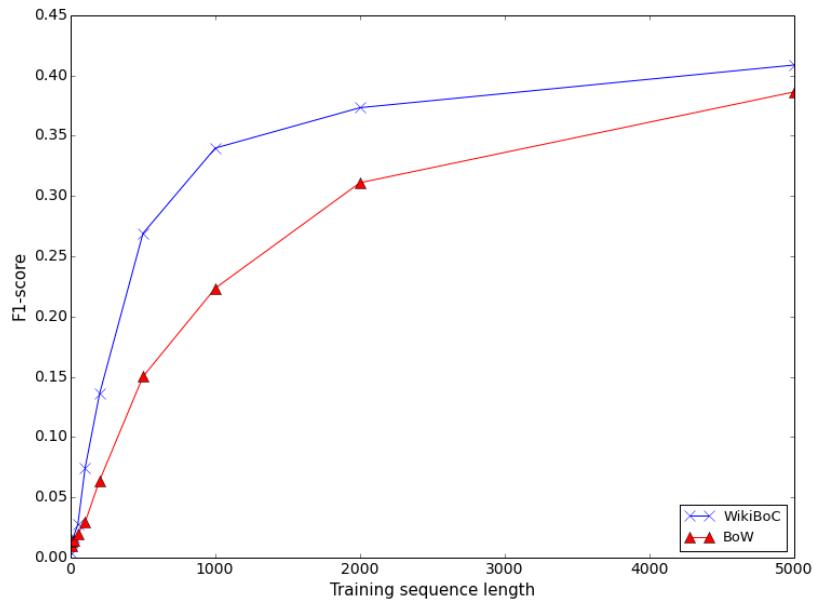


Figura 3.7: $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus UVigoMED (multietiqueta).

	5	10	20	50	100	200	500	1000	2000	5000
BoW	P	0.069	0.033	0.114	0.107	0.160	0.328	0.489	0.544	0.573
	R	0.024	0.031	0.015	0.016	0.029	0.062	0.152	0.229	0.312
	F1	0.010	0.013	0.014	0.020	0.029	0.064	0.150	0.223	0.311
WikiBoC	P	0.001	0.140	0.225	0.186	0.411	0.536	0.589	0.601	0.606
	R	0.021	0.021	0.014	0.023	0.069	0.138	0.282	0.364	0.414
	F1	0.004	0.010	0.017	0.028	0.074	0.136	0.269	0.340	0.409

Tabla 3.4: Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus UVigoMED (multietiqueta).

3.2.4. Discusión

Los resultados presentados en la sección anterior muestran de forma clara que el rendimiento ofrecido por el clasificador SVM cuando se utiliza la representación WikiBoC es superior al ofrecido por el mismo clasificador cuando se utiliza la representación de los documentos basada en palabras. Es necesario destacar, tal y como se aprecia en la figura 3.5, que los mayores incrementos de rendimiento suceden con secuencias de entrenamiento cortas, debido a que los problemas ocasionados por la sinonimia y la polisemia quedan enmascarados cuando existe una gran cantidad de datos para entrenar el algoritmo. Aunque en este trabajo únicamente se ha aplicado la propuesta presentada a la clasificación de documentos pertenecientes al ámbito biomédico, el modelo ha sido aplicado de forma exitosa a dominios diferentes, como puede verse en Mouríño-García et al. [73, 82, 83], en los cuales se ha aplicado la representación WikiBoC a la clasificación de noticias.

3.3. La representación WikiBoC y la clasificación multilingüe

Tras los positivos resultados obtenidos de la aplicación del modelo WikiBoC a la clasificación monolingüe de documentos, esta investigación continúa con la aplicación de dicho modelo a la clasificación multilingüe.

3.3.1. Estado del arte

Tradicionalmente, la clasificación multilingüe de documentos ha sido abordada a través de la combinación del modelo BoW y de las técnicas de traducción automática de documentos (BoW-MT), siendo las principales propuestas *Cross-Lingual Training* y *Multi-View Learning*. Por un lado, los trabajos basados en *Cross-Lingual Training* entran los algoritmos de clasificación con conjuntos de documentos traducidos [61, 128], haciendo uso de herramientas automáticas [27, 51] como *Google Translate* o *Bing Translator*. Una variación de esta propuesta consiste en traducir las características extraídas de los documentos en lugar de los propios documentos [72, 110, 132], bien durante la fase de entrenamiento o bien durante la fase de clasificación. Por otro lado, los estudios basados en *Multi-View Learning* realizan el entrenamiento de los algoritmos de clasificación con los documentos originales y sus traducciones, considerando cada idioma como una visión independiente de los datos [2, 32, 41, 63, 129].

Las principales limitaciones del modelo BoW-MT surgen de la combinación de las técnicas que lo componen. Por una parte, nos encontramos con las limitaciones inherentes al modelo BoW presentadas previamente. Por otro lado, las técnicas de traducción automática de documentos tienen dos grandes desventajas, la ambigüedad léxica y

estructural [51, 110], que afectan de forma negativa a la calidad de las traducciones. De este modo, la selección de una traducción incorrecta puede distorsionar la precisión de un clasificador debido a la introducción de características (*features*) erróneas. Por consiguiente, cuando la representación BoW se combina con la utilización de técnicas de traducción automática, las desventajas de ambas propuestas se suman, lo que conduce a un incremento de la probabilidad de error del clasificador. Con el propósito de solucionar los problemas del modelo BoW-MT, varios autores han explorado la aplicación de representaciones basadas en conceptos a la clasificación multilingüe de documentos, como Ni et al. [88] y De Smet et al. [21], que utilizan Bi-LDA, Maleshkova et al. [64], que utiliza ESA, Vulić and Moens [127] y Upadhyay et al. [122], que hacen uso de *Bilingual Word Embeddings*, o Carrero et al. [13] y Elberrichi et al. [29], que hacen uso del anotador semántico MetaMap.

3.3.2. Enfoque

Los trabajos presentados en Mouríño-García et al. [79, 80] exploran la aplicación de la representación WikiBoC a la construcción de un clasificador multilingüe. Se presenta el diseño y desarrollo de dos propuestas de clasificación inglés-castellano que hacen uso del conocimiento proporcionado por la Wikipedia para representar los documentos. La primera, WikiBoC-CLCM, representa los documentos haciendo uso únicamente de conceptos Wikipedia, y se basa en la combinación de la representación WikiBoC de los documentos y la técnica *Cross-Language Concept Matching* descrita en la sección 2.8.2.7. La segunda, Hybrid-WikiBoC, combina la propuesta WikiBoC-CLCM y el modelo BoW-MT, a través del enriquecimiento de las representaciones en forma de bolsas de palabras con conceptos de la Wikipedia extraídos de los propios documentos.

Para verificar la aplicabilidad y evaluar el rendimiento de las propuestas presentadas, WikiBoC-CLCM y Hybrid-WikiBoC, se han realizado una serie de experimentos de clasificación utilizando diversos conjuntos de datos pertenecientes a diferentes ámbitos: un ámbito genérico, utilizando para ello el conjunto de documentos Wikipedia Corpus [77, 80]; el ámbito biomédico, utilizando para ello los conjunto de datos Wikipedia Human Medicine Corpus [77, 80] y CL-UVigoMED [75, 79]; y el ámbito legal, utilizando el conjunto de datos JRC-Acquis. El rendimiento ofrecido por las dos propuestas presentadas se ha comparado con el ofrecido por las propuestas más relevantes del estado del arte para la clasificación multilingüe de documentos de texto: la representación BoW-MT, el modelo ESA, el modelo Bi-LDA, el modelo BWE y el modelo MetaMap, descritos en la sección 2.8.

La metodología utilizada para la realización de los experimentos utilizando la propuesta WikiBoC-CLCM consta de los siguientes pasos, los cuales se pueden seguir de forma gráfica a través de la figura 3.8. En primer lugar, es necesario obtener la representación WikiBoC de todos los documentos involucrados en el experimento, tanto aquellos pertenecientes a la secuencia de entrenamiento como aquellos pertenecientes a la secuencia de test. Como los documentos de la secuencia de entrenamiento están escritos en inglés, para obtener las representaciones WikiBoC se utilizará una instancia de Wikipedia

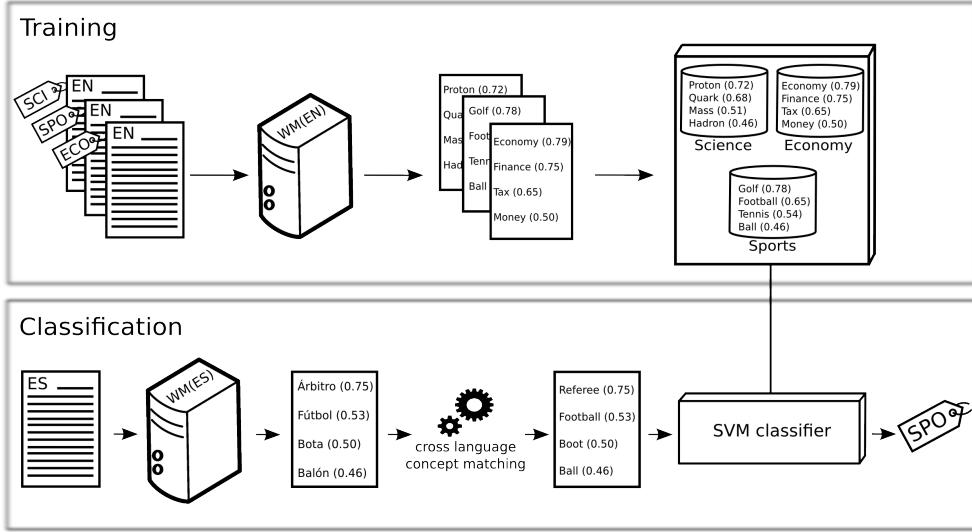


Figura 3.8: Arquitectura del clasificador SVM inglés-castellano utilizando la representación WikiBoC-CLCM.

Miner en inglés, que proporcionará como resultado representaciones de los documentos compuestas por conceptos presentes en la edición en inglés de la Wikipedia. En cuanto a la secuencia de test, formada por documentos escritos en castellano, se utilizará una instancia de Wikipedia Miner en castellano para la obtención de sus representaciones WikiBoC, las cuales estarán compuestas por conceptos presentes en la edición en castellano de la Wikipedia. Tras esto, las representaciones WikiBoC de los documentos de entrenamiento son utilizadas para entrenar el algoritmo de clasificación SVM. Como hemos mencionado previamente, las representaciones de los documentos a clasificar están formadas por conceptos de la edición en castellano de la Wikipedia. Entonces, es necesario convertir dichas representaciones al espacio de conceptos de la edición en inglés de la Wikipedia, de forma que todos los documentos se encuentren representados en el mismo espacio de características, en este caso, conceptos. Para realizar esta conversión se hace uso de la técnica *Cross-Language Concept Matching*. Una vez entrenado el clasificador, las representaciones en forma de conceptos de los documentos a clasificar – ya convertidas a inglés – son introducidas en el clasificador para que este prediga el conjunto de categorías al que pertenece. Por último, los anteriores pasos se repiten utilizando las diferentes propuestas del estado del arte para la realización de clasificación multilingüe de documentos de texto, y se compara el rendimiento de clasificación ofrecido por las diferentes propuestas [79, 80].

La metodología utilizada para la realización de los experimentos utilizando la propuesta Hybrid-WikiBoC es similar a la de la propuesta puramente basada en conceptos (cf. figura 3.9). Durante la fase de entrenamiento, las representaciones WikiBoC y BoW de los documentos en inglés se combinan para entrenar el clasificador SVM. Durante la fase de test, las representaciones WikiBoC de los documentos, ya convertidas a

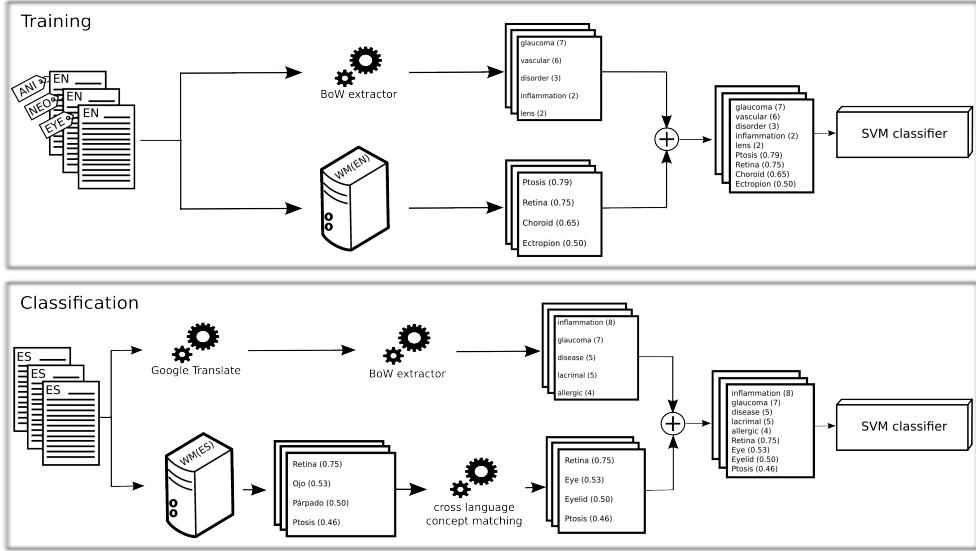


Figura 3.9: Arquitectura del clasificador SVM inglés-castellano utilizando la representación Hybrid-WikiBoC.

inglés, se combinan con las representaciones BoW obtenidas de los documentos en inglés, previamente traducidos del castellano. Estas representaciones combinadas serán introducidas en el clasificador SVM para que prediga a qué categoría o categorías pertenecen.

3.3.3. Resultados

En primer lugar, es necesario destacar que, aunque el enfoque utilizado en los dos trabajos acerca de la clasificación multilingüe de documentos es similar, el análisis comparativo realizado es diferente en cada uno de ellos. Presentaremos los resultados de cada trabajo por separado con el objetivo de facilitar el seguimiento de este apartado.

Mouriño-García et al. [80] presenta los resultados de los experimentos de clasificación realizados sobre los conjuntos de datos Wikipedia Corpus, Wikipedia Human Medicine Corpus y JRC-Acquis, utilizando los modelos BoW-MT, ESA, Bi-LDA, BWE, WikiBoC y Hybrid-WikiBoC. Presentaremos los resultados para cada corpus de forma independiente.

La figura 3.10 y la tabla 3.5 corresponden a los experimentos realizados sobre el conjunto de datos Wikipedia Corpus. Los resultados muestran de forma clara como la propuesta puramente basada en conceptos – WikiBoC-CLCM – ofrece el mejor rendimiento de clasificación cuando la longitud de la secuencia de entrenamiento es corta, mostrando mejoras en el rendimiento de hasta un 11.57% sobre BoW-MT. La propuesta WikiBoC también supera a las propuestas Bi-LDA y BWE en todo el rango

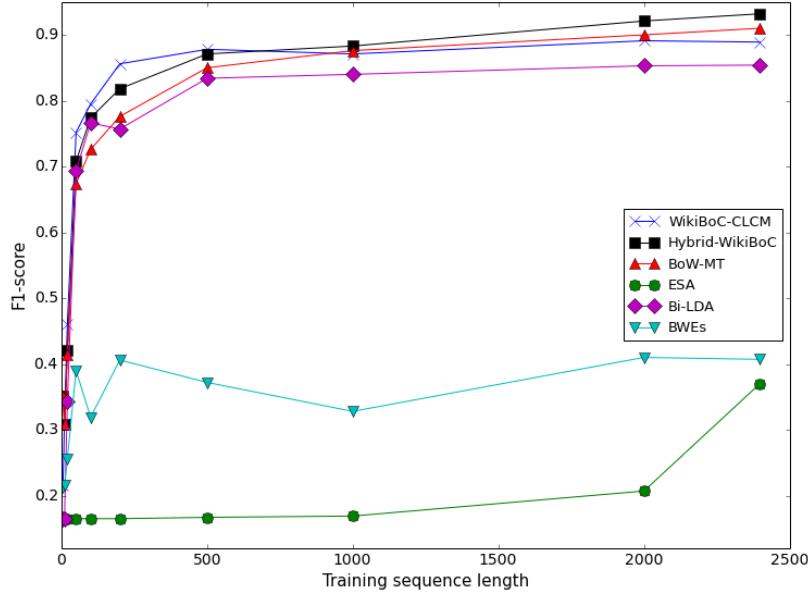


Figura 3.10: $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus Wikipedia Corpus.

de longitudes de la secuencia de entrenamiento, mostrando incrementos en el rendimiento de hasta un 87.87% y 165.55% respectivamente. Por su parte, la propuesta híbrida – Hybrid-WikiBoC – supera a las demás propuestas presentes en el estado del arte en todo el rango de longitudes de la secuencia de entrenamiento, ofreciendo incrementos de un 6.61%, 113.33% y 169.21% sobre BoW-MT, Bi-LDA y BWE respectivamente.

La figura 3.11 y la tabla 3.6 muestran los resultados de los experimentos realizados sobre el conjunto de datos Wikipedia Human Medicine. La propuesta WikiBoC-CLCM ofrece el mejor rendimiento cuando las secuencias de entrenamiento son cortas, superando al modelo BoW-MT en todo el rango de longitudes de la secuencia de entrenamiento excepto para la más grande, y al modelo Bi-LDA en el rango completo, mostrando incrementos en

	5	10	20	50	100	200	500	1000	2000	2398
WikiBoC-CLCM	0.216	0.310	0.461	0.752	0.795	0.856	0.878	0.871	0.891	0.889
BoW-MT	0.351	0.308	0.415	0.674	0.726	0.776	0.850	0.876	0.900	0.910
ESA	0.165	0.165	0.165	0.165	0.165	0.165	0.167	0.169	0.207	0.370
Bi-LDA	0.165	0.165	0.343	0.694	0.766	0.757	0.834	0.840	0.853	0.854
BWE	0.213	0.216	0.256	0.390	0.319	0.406	0.372	0.328	0.410	0.407
Hybrid-WikiBoC	0.352	0.309	0.420	0.708	0.774	0.818	0.871	0.883	0.921	0.932

Tabla 3.5: $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus Wikipedia Corpus.

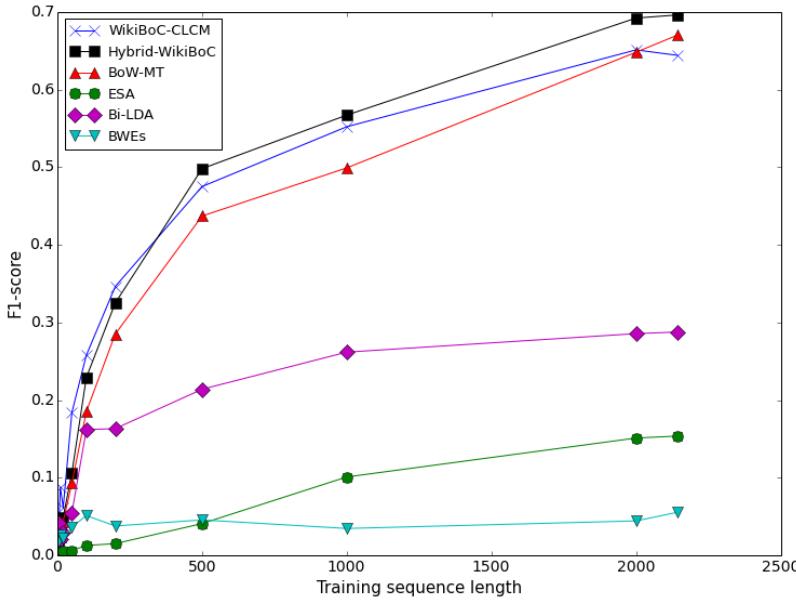


Figura 3.11: $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus Wikipedia Human Medicine.

el rendimiento de un 266.66 % y de un 319.04 % respectivamente. Por su parte, el modelo Hybrid-WikiBoC supera a las demás propuestas en todo el rango de longitudes de la secuencia de entrenamiento, obteniendo mejoras en el rendimiento de hasta un 23.78 % sobre el modelo BoW-MT y de hasta un 141.95 % sobre Bi-LDA. Las propuestas basadas en ESA y BWE son las que peor rendimiento ofrecen.

Por último, la figura 3.12 y la tabla 3.7 detallan los resultados obtenidos para el corpus JRC-Acquis. En este caso, la propuesta WikiBoC-CLCM se ve superada por la representación tradicional BoW-MT. Sin embargo, la propuesta híbrida sí ofrece un rendimiento igual o superior a la propuesta basada en palabras, obteniendo mejoras en el rendimiento de un 0.92 %. En cuanto a los modelos Bi-LDA y BWE, las propuestas

	5	10	20	50	100	200	500	1000	2000	2143
WikiBoC-CLCM	0.037	0.088	0.061	0.183	0.257	0.346	0.475	0.552	0.651	0.644
BoW-MT	0.016	0.024	0.043	0.093	0.185	0.285	0.437	0.499	0.648	0.670
ESA	0.006	0.006	0.006	0.106	0.012	0.015	0.041	0.109	0.151	0.154
Bi-LDA	0.037	0.021	0.029	0.055	0.162	0.163	0.214	0.262	0.286	0.288
BWE	0.018	0.026	0.022	0.036	0.051	0.038	0.046	0.035	0.044	0.055
Hybrid-WikiBoC	0.017	0.030	0.049	0.106	0.229	0.325	0.498	0.567	0.692	0.696

Tabla 3.6: $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus Wikipedia Human Medicine.

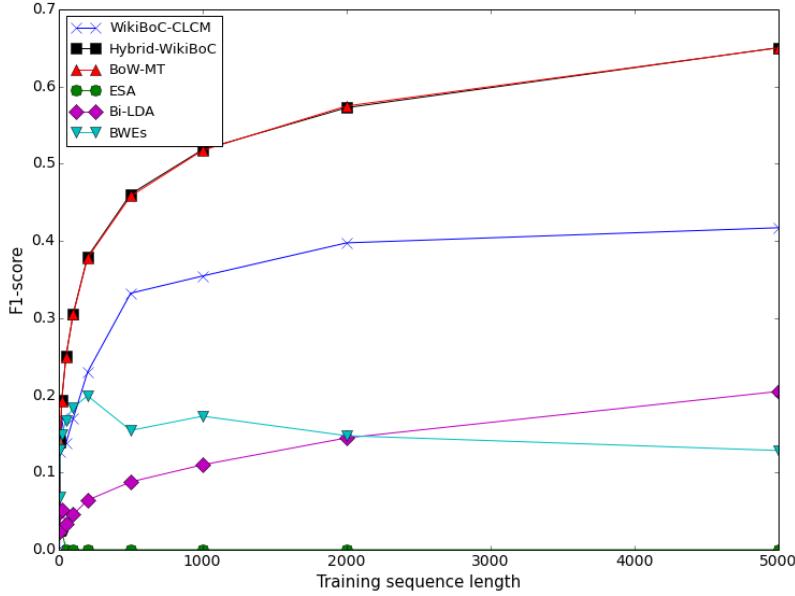


Figura 3.12: F_1 – score en función de la longitud de la secuencia de entrenamiento para el corpus JRC-Acquis.

presentadas ofrecen mejores resultados, con independencia del tamaño de la secuencia de entrenamiento, obteniendo mejoras en el rendimiento de hasta un 637.70 % y 407.89 % respectivamente.

El trabajo presentado en Mouríño-García et al. [79] muestra los resultados de los experimentos de clasificación realizados sobre el corpus CL-UVigoMED, utilizando los modelos BoW-MT, ESA, MetaMap, WikiBoC e Hybrid-WikiBoC. Las figuras 3.13 y 3.14, y las tablas 3.8 y 3.9 muestran que la propuesta WikiBoC supera *i*) al modelo BoW-MT cuando las secuencias de entrenamiento son cortas, proporcionando mejoras en el rendimiento de hasta un 106.26 %; *ii*) al modelo MetaMap en la práctica totalidad del rango de longitudes de la secuencia de entrenamiento, obteniendo incrementos de hasta

	5	10	20	50	100	200	500	1000	2000	5000
WikiBoC-CLCM	0.024	0.127	0.169	0.138	0.170	0.230	0.332	0.354	0.397	0.417
BoW-MT	0.025	0.140	0.192	0.250	0.305	0.378	0.458	0.517	0.575	0.650
ESA	0.026	0.026	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Bi-LDA	0.024	0.048	0.051	0.034	0.046	0.064	0.088	0.110	0.145	0.205
BWE	0.069	0.130	0.149	0.168	0.184	0.199	0.155	0.173	0.148	0.128
Hybrid-WikiBoC	0.025	0.140	0.194	0.250	0.306	0.379	0.461	0.518	0.572	0.650

Tabla 3.7: F_1 – score en función de la longitud de la secuencia de entrenamiento para el corpus JRC-Acquis.

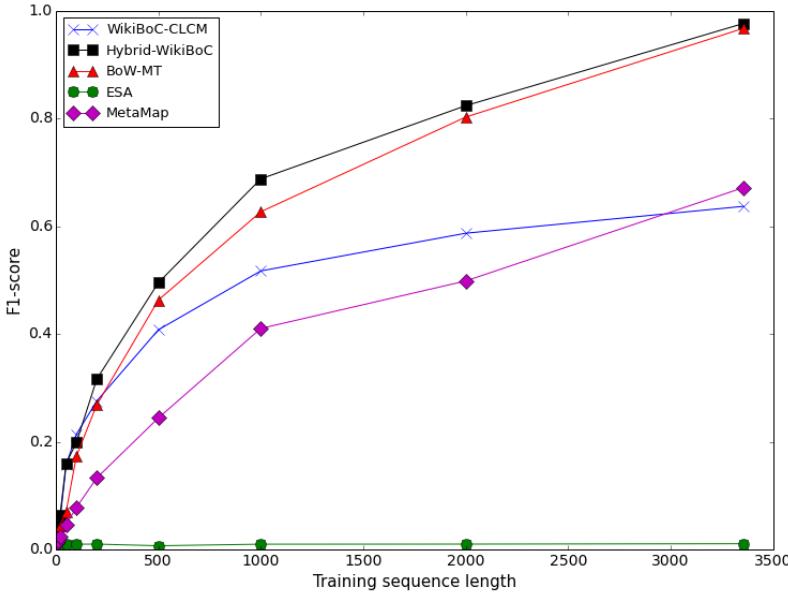


Figura 3.13: $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus CL-UVigoMED (etiqueta única).

un 332.31 %; y *iii)* al modelo ESA, el cual ofrece el peor rendimiento. Por su parte, la propuesta Hybrid-WikiBoC muestra incrementos del rendimiento de hasta un 127,01 % sobre BoW-MT y de hasta un 355.29 % sobre MetaMap. De nuevo, la propuesta basada en ESA es la que peores resultados ofrece.

3.3.4. Discusión

Los resultados de los experimentos realizados sobre los conjuntos de datos Wikipedia Corpus, Wikipedia Human Medicine Corpus y CL-UVigoMED muestran de forma clara como la propuesta basada puramente en conceptos – WikiBoC-CLCM – ofrece el mejor rendimiento de clasificación cuando la longitud de la secuencia de entrenamiento es corta. Cuando se obtienen las representaciones WikiBoC de los documentos se realiza de forma implícita una reducción de la dimensionalidad del conjunto de características, lo cual mitiga la influencia de los problemas de dispersión de los datos y de la alta dimensionalidad del conjunto de características [35, 135], incrementando así el rendimiento [56]. Por su parte, la propuesta híbrida ofrece el mejor rendimiento cuando las secuencias de entrenamiento son elevadas, y supera a las demás propuestas presentes en el estado del arte en todo el rango de longitudes de la secuencia de entrenamiento. Esto significa que los conceptos de la Wikipedia utilizados para enriquecer las representaciones basadas en palabras proporcionan información relevante al clasificador, incrementando así su rendimiento [49, 104]. Es necesario destacar que el incremento relativo de rendimiento

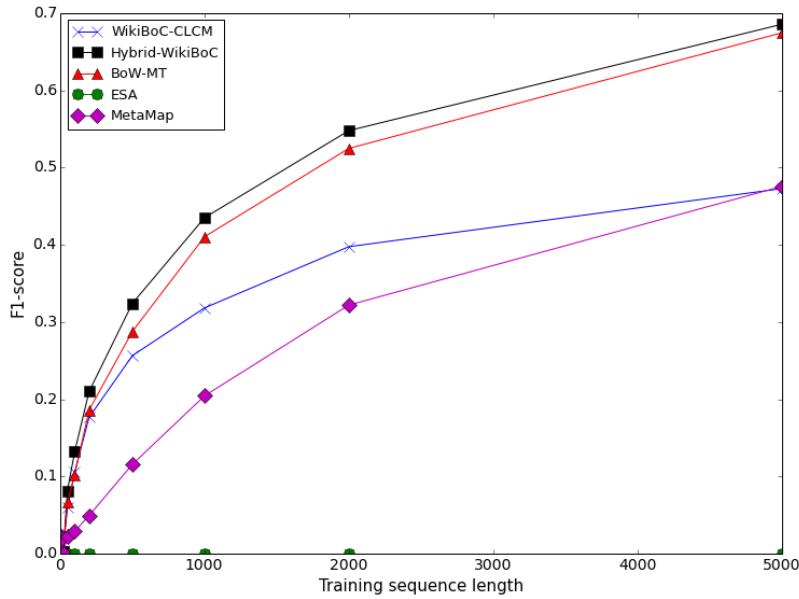


Figura 3.14: $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus CL-UVigoMED (multietiqueta).

		5	10	20	50	100	200	500	1000	2000	3349
BoW-MT	P	0.054	0.055	0.247	0.369	0.477	0.512	0.651	0.760	0.877	0.975
	R	0.100	0.087	0.156	0.186	0.291	0.406	0.570	0.726	0.867	0.974
	F1	0.023	0.025	0.044	0.070	0.173	0.270	0.462	0.627	0.802	0.967
ESA	P	0.008	0.008	0.018	0.018	0.089	0.018	0.008	0.109	0.018	0.098
	R	0.009	0.007	0.135	0.135	0.095	0.135	0.009	0.095	0.135	0.096
	F1	0.007	0.007	0.010	0.010	0.010	0.010	0.007	0.010	0.010	0.011
MetaMap	P	0.105	0.119	0.225	0.267	0.294	0.337	0.494	0.649	0.688	0.876
	R	0.136	0.136	0.136	0.167	0.159	0.242	0.340	0.472	0.575	0.677
	F1	0.011	0.016	0.024	0.046	0.078	0.133	0.244	0.410	0.498	0.671
WikiBoC	P	0.099	0.060	0.171	0.319	0.461	0.490	0.627	0.662	0.715	0.763
	R	0.088	0.128	0.157	0.222	0.297	0.398	0.485	0.586	0.652	0.719
	F1	0.014	0.029	0.053	0.157	0.214	0.276	0.408	0.517	0.587	0.637
Hybrid-WikiBoC	P	0.045	0.100	0.244	0.431	0.431	0.492	0.689	0.780	0.889	0.981
	R	0.093	0.095	0.165	0.265	0.358	0.475	0.655	0.756	0.878	0.981
	F1	0.024	0.043	0.063	0.159	0.199	0.317	0.495	0.688	0.824	0.977

Tabla 3.8: Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus CL-UVigoMED (etiqueta única).

	5	10	20	50	100	200	500	1000	2000	5000	
	P	0.014	0.013	0.105	0.266	0.416	0.464	0.553	0.614	0.670	0.714
BoW-MT	R	0.011	0.009	0.002	0.102	0.164	0.250	0.352	0.473	0.281	0.758
	F1	0.009	0.007	0.002	0.067	0.102	0.186	0.288	0.410	0.524	0.674
ESA	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	R	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	F1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MetaMap	P	0.010	0.010	0.001	0.161	0.235	0.507	0.679	0.675	0.000	0.000
	R	0.063	0.42	0.001	0.064	0.080	0.069	0.130	0.204	0.000	0.000
	F1	0.022	0.020	0.001	0.023	0.0029	0.049	0.116	0.204	0.321	0.475
WikiBoC-CLCM	P	0.028	0.044	0.069	0.533	0.603	0.620	0.670	0.679	0.688	0.702
	R	0.012	0.007	0.003	0.069	0.113	0.192	0.275	0.323	0.400	0.482
	F1	0.014	0.010	0.004	0.060	0.106	0.176	0.256	0.318	0.397	0.473
Hybrid-WikiBoC	P	0.109	0.104	0.286	0.339	0.454	0.516	0.616	0.665	0.704	0.744
	R	0.002	0.002	0.002	0.127	0.183	0.273	0.375	0.482	0.588	0.755
	F1	0.002	0.003	0.003	0.082	0.132	0.210	0.324	0.435	0.548	0.685

Tabla 3.9: Precisión, retirada y $F_1 - score$ en función de la longitud de la secuencia de entrenamiento para el corpus CL-UVigoMED (multietiqueta).

medio de las propuestas presentadas sobre la propuesta basada en palabras es muy superior en los conjuntos de documentos sobre cuestiones biomédicas. Esto, en conjunción con los resultados obtenidos en investigaciones previas [9, 74, 136], demuestra que el uso de conceptos de la Wikipedia es especialmente adecuado para la representación de documentos de contenido biomédico.

En lo referente a las propuestas Bi-LDA y BWE, aunque ofrecen un rendimiento digno en la tarea de clasificación del conjunto de datos Wikipedia Corpus, su rendimiento es muy bajo a la hora de clasificar los conjuntos de datos Wikipedia Human Medicine y JRC-Acquis. Esto es debido a la gran influencia que tiene el corpus comparable que utilizan para extraer los conceptos bilingües, tal y como se puede ver en la sección 6 del trabajo presentado en Mouríño-García et al. [80].

Por último, el hecho de que la propuesta basada en ESA ofrezca los peores resultados en todos los experimentos de clasificación realizados es debido a su bajo rendimiento a la hora de extraer conceptos de documentos completos y a su tendencia a generar *outliers* [28]. Esto afecta de forma muy negativa al rendimiento de clasificación, tal y como se puede ver en la sección 6 de Mouríño-García et al. [80] y en la sección 4 de Mouríño-García et al. [79].

Los resultados obtenidos nos permiten validar la aplicabilidad de la representación de los documentos basada en conceptos de la Wikipedia a la construcción de un clasificador multilingüe de documentos. El uso de la representación WikiBoC es ventajoso, ya que los conceptos extraídos a través del anotador semántico proporcionan información valiosa para el algoritmo de clasificación, incrementando así su rendimiento, especialmente cuando se trata de documentos de contenido biomédico.

Capítulo 4

Contribuciones, conclusiones y trabajos futuros

Este capítulo comienza con la exposición de las contribuciones realizadas por esta investigación. A continuación, se muestran las conclusiones obtenidas tras su realización. Por último, se presenta una serie de propuestas para su extensión futura.

4.1. Contribuciones

La principal contribución realizada por esta tesis es un modelo para la clasificación monolingüe y multilingüe de documentos de texto, pertenecientes a diversos ámbitos de aplicación, que hace uso del conocimiento enciclopédico contenido en la Wikipedia para crear representaciones de los documentos basadas en conceptos, proporcionando mejores resultados que las propuestas presentes en el estado del arte. A su vez, esta contribución general se divide en una serie de contribuciones más concretas:

1. La utilización de técnicas de clasificación automática de documentos para proporcionar interoperabilidad entre repositorios de recursos educativos.
2. El análisis y evaluación comparativa de la aplicación de la representación de los documentos WikiBoC y de la representación basada en palabras a la clasificación monolingüe de literatura biomédica.
3. El corpus de literatura biomédica monolingüe en inglés, UVigoMED.
4. El análisis y evaluación comparativa de la aplicación de la representación de los documentos WikiBoC y de las representaciones más relevantes del estado del arte a la clasificación multilingüe.

5. Los corpus multilingües CL-UVigoMED, Wikipedia Corpus y Wikipedia Human Medicine Corpus.

Los artículos publicados a raíz de la investigación realizada en esta tesis atienden las contribuciones antes presentadas en diferentes escenarios y dominios de aplicación de la siguiente forma. Mouríño-García et al. [81] cubre la primera contribución, descrita en el capítulo 3.1. Mouríño-García et al. [74] cubre las contribuciones 2 y 3, descritas en el capítulo 3.2. Por último, Mouríño-García et al. [79] y Mouríño-García et al. [80] cubren las contribuciones 4 y 5, descritas en el capítulo 3.3.

4.2. Conclusiones

Este trabajo ha presentado la aplicación de la representación de los documentos como bolsas de conceptos de la Wikipedia (WikiBoC), obtenidos dichos conceptos por medio del anotador semántico Wikipedia Miner, a las tareas de clasificación monolingüe y multilingüe de documentos de texto pertenecientes a diversos ámbitos de aplicación.

La obtención de las contribuciones presentadas en la sección anterior se ha alcanzado a través del cumplimiento del conjunto de objetivos planteados al inicio de este trabajo de investigación. Las cuatro publicaciones presentadas como parte de esta tesis nos han permitido familiarizarnos con las técnicas de clasificación automática de documentos de texto y conocer en detalle las principales y más novedosas propuestas presentes en el estado del arte para su representación en tareas de clasificación monolingüe y multilingüe. El banco de pruebas desarrollado y los conjuntos de datos creados durante la realización de esta investigación nos han permitido realizar de una forma exhaustiva, rigurosa y repetible los experimentos de clasificación que constituyen parte fundamental de esta tesis, así como su comparación con las propuestas más relevantes del estado del arte. Por último, la evaluación comparativa de los resultados obtenidos de los experimentos, utilizando las propuestas presentadas, las representaciones más relevantes del estado del arte, y el banco de pruebas y los conjuntos de datos anteriormente citados, nos han permitido demostrar la aplicabilidad y los beneficios aportados por el uso de la representación de los documentos en forma de bolsa de conceptos de la Wikipedia, obtenidos a través de la utilización del anotador semántico Wikipedia Miner, a la clasificación monolingüe y multilingüe de documentos de texto procedentes de diferentes dominios.

El cumplimiento de los objetivos planteados, las contribuciones aportadas y los resultados de los experimentos realizados nos permiten validar la hipótesis de investigación planteada al comienzo de este documento, puesto que, efectivamente, se ha mejorado el rendimiento de las propuestas presentes en el estado del arte para la clasificación monolingüe y multilingüe de documentos de texto, utilizando una representación de los mismos basada en conceptos de la Wikipedia obtenidos a través del anotador semántico de propósito general Wikipedia Miner.

La realización de la investigación presentada nos permite así extraer las siguientes conclusiones:

- El uso de la representación WikiBoC en tareas de clasificación monolingüe y multilingüe de documentos de texto es ventajoso, ya que los conceptos extraídos a través del anotador semántico Wikipedia Miner proporcionan información muy relevante para el algoritmo de clasificación, incrementando así su rendimiento, tal y como se demuestra en los experimentos realizados a lo largo de esta investigación.
- El uso de la representación WikiBoC es especialmente ventajoso cuando se trata de documentos de contenido biomédico. Los resultados obtenidos de los experimentos realizados a lo largo de esta investigación muestran que existe algún tipo de característica distintiva en los documentos biomédicos. Esto va en línea con los resultados obtenidos en investigaciones previas, que establecen que los documentos acerca de cuestiones biomédicas son excelentes candidatos para la aplicación de representaciones basadas en conceptos, ya que albergan características distintivas, como la prevalencia de frases largas, y el hecho de que las frases médicas son más propensas a los problemas de la sinonimia.
- El uso de la representación WikiBoC basada únicamente en conceptos de la Wikipedia es especialmente ventajoso cuando los datos de entrenamiento son escasos. Esto es debido a que los problemas ocasionados por la sinonimia y la polisemia quedan enmascarados cuando existe una gran cantidad de datos para entrenar el algoritmo, lo que no sucede cuando los datos de entrenamiento son escasos.
- El enriquecimiento de las representaciones BoW con conceptos de la Wikipedia extraídos de los propios documentos aumenta el rendimiento clasificadorio a pesar del incremento de la dimensionalidad ocasionado, que generalmente es perjudicial para el rendimiento de las tareas de clasificación. Los resultados de los experimentos realizados evidencian que la información proporcionada por los conceptos de la Wikipedia compensa el posible efecto pernicioso del incremento de la dimensionalidad producido al añadir los conceptos.

La principal limitación de este trabajo viene dada por la posible pérdida de información durante el proceso *Cross-Language Concept Matching*. Esto se debe a que no todos los artículos pertenecientes a una edición de la Wikipedia en un idioma L_m tienen correspondencia en otro idioma L_n , debido a que existen diferencias en el conjunto de artículos que contiene cada edición de la Wikipedia, especialmente entre aquellas ediciones en idiomas mayoritarios y minoritarios. Esto implica que habrá artículos (conceptos, entradas, páginas) en un idioma que no tendrán correspondencia en otros idiomas.

4.3. Trabajos futuros

La investigación realizada deja una serie de líneas abiertas para su extensión futura:

- Aunque la utilización de la representación WikiBoC se ha centrado únicamente en la clasificación monolingüe de documentos en inglés, y en la clasificación

multilingüe castellano-inglés, la propuesta presentada puede ser fácilmente extendida para incorporar documentos escritos en idiomas diferentes a los mencionados, simplemente utilizando *dumps* de la Wikipedia en otros idiomas como base de conocimiento del anotador semántico Wikipedia Miner.

- Sería de interés verificar el rendimiento de la propuesta WikiBoC para clasificar documentos escritos en idiomas minoritarios, los cuales generalmente cuentan con Wikipedias más pequeñas. Una Wikipedia más pequeña significa que el conjunto de conceptos disponible para representar los documentos se reduce, de manera que sería interesante verificar el impacto de la reducción del conjunto de características disponible para representar los documentos.
- En línea con la anterior propuesta de trabajo futuro, sería de interés verificar el rendimiento de la propuesta WikiBoC para realizar tareas de clasificación multilingüe de documentos entre idiomas mayoritarios, los cuales generalmente albergan grandes Wikipedias, e idiomas minoritarios, que como hemos dicho, generalmente cuentan con Wikipedias más pequeñas.
- Otra línea clara de trabajo futuro es la búsqueda de soluciones que mitiguen la posible pérdida de información durante el proceso *Cross-Language Concept Matching*.
- Estudios previos parecen mostrar que la representación WikiBoC proporciona buenos resultados en tareas de recuperación de información monolingüe, lo cual hace prometedora la aplicación de la representación WikiBoC junto con la técnica CLCM para la creación de un sistema de recuperación de información multilingüe.
- Diversos trabajos presentes en el estado del arte muestran que la utilización de una representación basada en conceptos proporciona buenos resultados en tareas de *clustering* de documentos. Sería entonces de interés verificar el rendimiento de la representación WikiBoC y la técnica CLCM para la realización de tareas de *clustering* monolingüe y multilingüe de documentos de texto.

Chapter 4

Contributions, conclusions and future work

This chapter first presents the contributions made by this research. Then, it describes the main conclusions obtained. Finally, it outlines different proposals for future extension.

4.1. Contributions

The main contribution made by this thesis is a model for monolingual and multilingual classifying text documents, belonging to different fields of application, that leverages the encyclopaedic knowledge contained in Wikipedia to create concept-based representations of documents, obtaining better results than state-of-the-art proposals. This contribution is in turn divided into a series of smaller contributions:

1. The use of automatic document classification techniques to provide interoperability between repositories of educational resources.
2. The analysis and benchmarking of WikiBoC against the state-of-the-art Bag of Words document representation to perform monolingual classification of biomedical literature.
3. The monolingual biomedical literature corpus written in English, UVigoMED.
4. The analysis and benchmarking of WikiBoC against the most relevant state-of-the-art document representations to perform multilingual classification of documents.

Conclusions

5. The multilingual corpora CL-UVigoMED, Wikipedia Corpus and Wikipedia Human Medicine Corpus.

The articles published as a result of this thesis address the contributions previously presented in different scenarios and applications domains in the following way. Mouriño-García et al. [81] covers the first contribution, described in Chapter 3.1. Mouriño-García et al. [74] covers contributions 2 and 3, described in Chapter 3.2. Finally, Mouriño-García et al. [79] and Mouriño-García et al. [80] cover contributions 4 and 5, described in Chapter 3.3.

4.2. Conclusions

This work has presented the application of a Wikipedia-based Bag of Concepts document representation (WikiBoC), using to that end the Wikipedia Miner semantic annotator, to the monolingual and multilingual classification of text documents from different fields of application.

Obtaining the contributions presented in the previous section has been achieved through the fulfilment of the set of objectives defined at the beginning of this research work. The four publications presented as part of this thesis have allowed us to get familiar with automatic text document classification techniques, and to know the details of the main and latest proposals on the state-of-the-art to represent documents in monolingual and multilingual classification tasks. The test bench developed and the datasets created in the course of this research have enabled us to conduct classification experiments, which are key part of this thesis, in a comprehensive, rigorous and reproducible way, as well as the benchmarking with the most relevant proposals on the state-of-the-art. Finally, the benchmarking of the classification experiments using the proposal presented, the most relevant document representations on the state-of-the-art, and the test bench and datasets previously mentioned, have allowed us to demonstrate the applicability and to show the benefits of using a Wikipedia-based Bag of Concepts document representation to the monolingual and multilingual classification of text documents belonging to different areas of application.

The compliance with the objectives stated, the contributions presented, and the results of the classification experiments conducted enable us to validate the research hypothesis stated at the beginning of this document, indeed the performance offered by a classifier using the Wikipedia-based Bag of Concepts representation proposed is higher than the performance offered when using state-of-the-art approaches in the task of monolingual and multilingual classifying text documents.

The research conducted allows us to draw the following conclusions:

- The use of the WikiBoC representation in monolingual and multilingual document classification is advantageous, since the concepts extracted through the Wikipedia

Miner semantic annotator add valuable information to the classification algorithm, thus improving its performance, as demonstrated by the classification experiments conducted in the course of this research.

- The use of the WikiBoC is particularly advantageous with documents in the biomedical field. Results obtained from classification experiments show that there is some type of distinctive feature in biomedical documents. This is line with previous research works, which state that biomedical documents are excellent candidates for the Bag of Concepts representation, since they have distinctive characteristics such as the high prevalence of long phrases, and the fact that medical phrases are prone to synonymy.
- The use of the purely concept-based representation WikiBoC is especially advantageous when training data is scarce. This is because, with enough data, the problems of synonymy and polysemy are masked, which is not the case when training data is scarce.
- Enriching the BoW representations with Wikipedia concepts extracted from documents themselves improves classification performance, despite of the dimensionality increase when adding concepts, which is usually harmful to classification performance. The results of the experiments conducted suggest that this added information compensates the possible pernicious effect of the increase of dimensionality produced when adding concepts.

The main drawback of this work is the possible loss of information during the *Cross-Language Concept Matching* process. This is because not all Wikipedia articles belonging to a particular edition of Wikipedia in language L_m have correspondence in a different language L_n , since there exist differences in the number of articles contained in each edition of Wikipedia, especially between those editions in majority and minority languages. This implies that there will be articles (concepts, entries, pages) in a language that will not have correspondence in other languages.

4.3. Future work

This research leaves open lines for future extension:

- Although the use of the WikiBoC representation has focused only on the monolingual classification of documents written in English, and in the Spanish-English multilingual classification, the proposal presented can be easily extended to incorporate documents written in languages other than those mentioned, simply by using dumps of Wikipedia in other languages as knowledge base of the semantic annotator Wikipedia Miner.

Future work

- It would be interesting to verify the performance of the WikiBoC representation to classify documents written in minority languages, which generally keep smaller Wikipedias. A smaller Wikipedia means that the set of concepts available to represent the documents is reduced, so it would be interesting to verify the impact of the reduction of the set of features to represent the documents.
- In line with the previous proposal for future work, it would be interesting to verify the performance of the WikiBoC representation to perform multilingual document classification between majority languages, which generally keep large Wikipedias, and minority languages, which, as we said, generally keep smaller Wikipedias.
- Another clear line for future work is to find solutions that mitigate the possible loss of information during the *Cross-Language Concept Matching* process.
- Previous studies seem to show that the WikiBoC representation provides good results in monolingual information retrieval tasks, which makes promising the application of the WikiBoC representation along with the CLCM technique for the creation of a multilingual information retrieval system.
- Several works in the state-of-the-art show that the use of a concept-based representation provides good results in document clustering tasks. It would be of interest to verify the performance of the WikiBoC representation and the CLCM technique for conducting monolingual and multilingual document clustering tasks.

Bibliografía

- [1] Aggarwal, C. C. and Zhai, C. (2012). A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer.
- [2] Amini, M., Usunier, N., and Goutte, C. (2009). Learning from multiple partially observed views-an application to multilingual text categorization. In *Advances in neural information processing systems*, pages 28–36.
- [3] Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. Suzanne Bakken, American Medical Informatics Association, Washington, USA.
- [4] Behera, R. N., Manan, R., and Dash, S. (2016). Ensemble based hybrid machine learning approach for sentiment classification-a review. *International Journal of Computer Applications*, 146(6).
- [5] Bel, N., Koster, C. H., and Villegas, M. (2003). Cross-lingual text categorization. In *Research and Advanced Technology for Digital Libraries*, pages 126–139. Springer.
- [6] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *journal of machine learning research*, 3(Feb):1137–1155.
- [7] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.
- [8] Blizzard, W. D. et al. (1988). Multiset theory. *Notre Dame Journal of formal logic*, 30(1):36–66.
- [9] Bloehdorn, S. and Hotho, A. (2004). Boosting for text classification with semantic features. In *WebKDD*, pages 149–166. Springer.
- [10] Boeykens, S., Santana Quintero, M., and Neuckermans, H. (2009). Metadata for Architectural Contents in Europe. In *Book of Abstracts*.
- [11] Cafolla, R. and Cafolla, R. (2006). Project MERLOT: Bringing Peer Review to Web-Based Educational Resources. *Journal of Technology and Teacher Education*, 14:313–323.

Bibliografía

- [12] Carbonell, J. G., Michalski, R. S., and Mitchell, T. M. (1983). An overview of machine learning. In *Machine learning*, pages 3–23. Springer Berlin Heidelberg.
- [13] Carrero, F., Cortizo, J. C., and Gomez, J. M. (2008). Testing concept indexing in crosslingual medical text classification. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, pages 512–519. IEEE.
- [14] Chan, L. M. and Zeng, M. L. (2006). Metadata interoperability and standardization - A study of methodology part I: Achieving interoperability at the schema level. *D-Lib Magazine*, 12(6):23–41.
- [15] Chen, M., Jin, X., and Shen, D. (2011). Short text classification improved by learning multi-granularity topics. In *IJCAI*, pages 1776–1781.
- [16] Christopher, D. M., Prabhakar, R., and Hinrich, S. (2008). Introduction to information retrieval. *An Introduction To Information Retrieval*, 151:177.
- [17] Cimiano, P., Schultz, A., Sizov, S., Sorg, P., and Staab, S. (2009). Explicit versus latent concept models for cross-language information retrieval. In *IJCAI*, volume 9, pages 1513–1518.
- [18] Colace, F., De Santo, M., Greco, L., and Napoletano, P. (2015). Improving relevance feedback-based query expansion by the use of a weighted word pairs approach. *Journal of the Association for Information Science and Technology*, 66(11):2223–2234.
- [19] Dai, M., Shah, N. H., Xuan, W., Musen, M. A., Watson, S. J., Athey, B. D., Meng, F., et al. (2008). An efficient solution for mapping free text to ontology terms. *AMIA Summit on Translational Bioinformatics*, 21.
- [20] D'Antoni, S. (2006). *The virtual university: Models and messages, lessons from case studies*. UNESCO-IIEP.
- [21] De Smet, W., Tang, J., and Moens, M.-F. (2011). Knowledge transfer across multilingual corpora via latent topics. *Lecture Notes in Computer Science*, 6634:549–560.
- [22] Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.
- [23] Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N., and Taibi, D. (2012). Linked Education: interlinking educational Resources and the Web of Data. In *Proceedings of the 27th annual ACM symposium on applied computing*, pages 366–371. ACM.
- [24] Ding, S., Yu, J., Qi, B., and Huang, H. (2014). An overview on twin support vector machines. *Artificial Intelligence Review*, pages 1–8.
- [25] Doan, A., Madhavan, J., Domingos, P., and Halevy, A. (2004). Ontology matching: A machine learning approach. In *Handbook on ontologies*, pages 385–403. Springer.

- [26] Dolog, P., Henze, N., Nejdl, W., and Sintek, M. (2004). Personalization in distributed e-learning environments. *Alternate track papers posters of the 13th international conference on World Wide Web WWW Alt 04*, pages 170–179.
- [27] Duh, K., Fujino, A., and Nagata, M. (2011). Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers- Volume 2*, pages 429–433. Association for Computational Linguistics.
- [28] Egozi, O., Markovitch, S., and Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, 29(2):8.
- [29] Elberrichi, Z., Taibi, M., and Belaggoun, A. (2012). Multilingual medical documents classification based on mesh domain ontology. *arXiv preprint arXiv:1206.4883*.
- [30] Fernández, M. (2012). Adquisición y representación del conocimiento mediante procesamiento del lenguaje natural. *A Coruña, Spain*.
- [31] Ferrández, S., Toral, A., Ferrández, Ó., Ferrández, A., and Muñoz, R. (2009). Exploiting wikipedia and eurowordnet to solve cross-lingual question answering. *Information Sciences*, 179(20):3473–3488.
- [32] Fortuna, B. and Shawe-Taylor, J. (2005). The use of machine translation tools for cross-lingual text mining. In *Proceedings of the ICML Workshop on Learning with Multiple Views*.
- [33] Gabrilovich, E. (2007). Feature generation for textual information retrieval using world knowledge. In *ACM SIGIR Forum*, volume 41, pages 123–123. ACM.
- [34] Gabrilovich, E. and Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, pages 443–498.
- [35] Gao, L., Zhou, S., and Guan, J. (2015). Effectively classifying short texts by structured sparse representation with dictionary filtering. *Information Sciences*, 323:130–142.
- [36] Garla, V., Taylor, C., and Brandt, C. (2013). Semi-supervised clinical text classification with laplacian svms: an application to cancer case management. *Journal of biomedical informatics*, 46(5):869–875.
- [37] Ghemawat, S., Gobioff, H., and Leung, S.-T. (2003). The google file system. In *ACM SIGOPS operating systems review*, volume 37, pages 29–43. ACM.
- [38] Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901.
- [39] Gope, H. L., Das, P. K., Islam, M. J., and Seddiqi, M. H. (2014). Medical document classification from ohsuemed dataset. *IJCSN International Journal of Computer Science and Network*, 3(4):215–219.

Bibliografía

- [40] Guindon, G. E., Lavis, J. N., Becerra-Posada, F., Malek-Afzali, H., Shi, G., Yesudian, C. A. K., Hoffman, S. J., et al. (2010). Bridging the gaps between research, policy and practice in low-and middle-income countries: a survey of health care providers. *Canadian Medical Association Journal*, 182(9):E362–E372.
- [41] Guo, Y. and Xiao, M. (2012). Cross language text classification via subspace co-regularized multi-view learning. *arXiv preprint arXiv:1206.6481*.
- [42] Hajmohammadi, M. S., Ibrahim, R., Selamat, A., and Fujita, H. (2015). Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. *Information sciences*, 317:67–77.
- [43] Hassan, S. and Mihalcea, R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1192–1201. Association for Computational Linguistics.
- [44] Hastie, T., Tibshirani, R., and Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer.
- [45] Heap, B., Bain, M., Wobcke, W., Krzywicki, A., and Schmeidl, S. (2017). Word vector enrichment of low frequency words in the bag-of-words model for short text multi-class classification problems. *arXiv preprint arXiv:1709.05778*.
- [46] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.
- [47] Hersch, W., Buckley, C., Leone, T., and Hickam, D. (1994). Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *SIGIR'94*, pages 192–201. Springer.
- [48] Hillmann, D. I. and Westbrooks, E. L. (2004). *Metadata in practice*. American Library Association.
- [49] Huang, A., Milne, D., Frank, E., and Witten, I. H. (2009). Clustering documents using a wikipedia-based concept representation. In *Advances in Knowledge Discovery and Data Mining*, pages 628–636. Springer Berlin Heidelberg.
- [50] Huang, L., Milne, D., Frank, E., and Witten, I. H. (2012). Learning a concept-based document similarity measure. *Journal of the American Society for Information Science and Technology*, 63(8):1593–1608.
- [51] Hutchins, W. J. and Somers, H. L. (1992). *An introduction to machine translation*, volume 362. Academic Press London.
- [52] Jadhav, B. R., Mahajan, M., and GHR CEM, W. (2016). Dual sentiment analysis using adaboost algorithm sentiment analysis. *International Journal of Engineering Science*, 7641.
- [53] Jing, H., Tsao, Y., Chen, K.-Y., Wang, H.-M., et al. (2013). Semantic naïve bayes classifier for document classification. In *IJCNLP*, pages 1117–1123.

- [54] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.
- [55] Khan, A., Baharudin, B., Lee, L. H., and Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20.
- [56] Kim, H., Howland, P., and Park, H. (2005). Dimension reduction in text classification with support vector machines. In *Journal of Machine Learning Research*, pages 37–53.
- [57] Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.
- [58] Klementiev, A., Titov, I., and Bhattacharai, B. (2012). Inducing crosslingual distributed representations of words. *Proceedings of COLING 2012*, pages 1459–1474.
- [59] Kurilovas, E. and Dagiene, V. (2009). Learning Objects and Virtual Learning Environments Technical Evaluation Criteria. *Electronic Journal of e-Learning*, 7(Issue 2):127–136.
- [60] Levelt, W. J. (1993). *Speaking: From intention to articulation*, volume 1. MIT press.
- [61] Ling, X., Xue, G.-R., Dai, W., Jiang, Y., Yang, Q., and Yu, Y. (2008). Can chinese web pages be classified with english data source? In *Proceedings of the 17th international conference on World Wide Web*, pages 969–978. ACM.
- [62] Liparas, D., HaCohen-Kerner, Y., Mountzidou, A., Vrochidis, S., and Kompatsiaris, I. (2014). News articles classification using random forests and weighted multimodal features. In *Information Retrieval Facility Conference*, pages 63–75. Springer.
- [63] Lu, B., Tan, C., Cardie, C., and Tsou, B. K. (2011). Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 320–330. Association for Computational Linguistics.
- [64] Maleshkova, M., Zilka, L., Knoth, P., and Pedrinaci, C. (2011). Cross-lingual web api classification and annotation. *MSW 2011*, page 1.
- [65] Medelyan, O., Witten, I. H., and Milne, D. (2008). Topic indexing with wikipedia. In *Proceedings of the AAAI WikiAI workshop*, volume 1, pages 19–24.
- [66] Mehdi, M., Okoli, C., Mesgari, M., Nielsen, F. Å., and Lanamäki, A. (2017). Excavating the mother lode of human-generated text: A systematic review of research that uses the wikipedia corpus. *Information Processing & Management*, 53(2):505–529.
- [67] Milne, D. and Witten, I. H. (2013). An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194:222–239.
- [68] Milne, D. N., Witten, I. H., and Nichols, D. M. (2007). A knowledge-based search engine powered by wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 445–454. ACM.

Bibliografía

- [69] Ming, Z.-Y. and Chua, T. S. (2015). Resolving polysemy and pseudonymity in entity linking with comprehensive name and context modeling. *Information Sciences*, 307:18–38.
- [70] Mitchell, R. (2015). *Web scraping with Python: collecting data from the modern web*. O'Reilly Media, Inc.
- [71] Moise, G., Vladoiu, M., and Constantinescu, Z. (2014). Maseco: A multi-agent system for evaluation and classification of oers and ocw based on quality criteria. In *E-Learning Paradigms and Applications*, pages 185–227. Springer.
- [72] Montalvo, S., Martínez, R., Casillas, A., and Fresno, V. (2007). Multilingual news clustering: Feature translation vs. identification of cognate named entities. *Pattern Recognition Letters*, 28(16):2305–2311.
- [73] Mouriño-García, M., Pérez-Rodríguez, R., and Anido-Rifón, L. (2015a). Bag-of-concepts document representation for textual news classification. *International Journal of Computational Linguistics and Applications*, 6(1):173–188.
- [74] Mouriño-García, M., Pérez-Rodríguez, R., and Anido-Rifón, L. (2015b). Biomedical literature classification using encyclopedic knowledge: a wikipedia-based bag-of-concepts approach. *PeerJ*, 3:e1279.
- [75] Mouriño-García, M., Pérez-Rodríguez, R., and Anido-Rifón, L. (2016a). Cl-uvigomed corpus. <http://dx.doi.org/10.17632/7ph4hhh429.5>.
- [76] Mouriño-García, M., Pérez-Rodríguez, R., and Anido-Rifón, L. (2016b). Uvigomed corpus. <http://dx.doi.org/10.17632/p3jkppwr29.1>.
- [77] Mouriño-García, M., Pérez-Rodríguez, R., and Anido-Rifón, L. (2016c). Wikipedia corpus. <http://dx.doi.org/10.17632/nsm3ftcjf6.2>.
- [78] Mouriño-García, M., Pérez-Rodríguez, R., and Anido-Rifón, L. (2016d). Wikipedia human medicine corpus. <http://dx.doi.org/10.17632/sp9mcx5594.2>.
- [79] Mouriño-García, M., Pérez-Rodríguez, R., and Anido-Rifón, L. (2017a). A bag of concepts approach for biomedical document classification using wikipedia knowledge. *Methods of Information in Medicine*, 56.
- [80] Mouriño-García, M., Pérez-Rodríguez, R., and Anido-Rifón, L. (2017b). Wikipedia-based cross-language text classification. *Information Sciences*, 406:12–28.
- [81] Mouriño-García, M., Pérez-Rodríguez, R., Anido-Rifón, L., Fernández-Iglesias, M. J., and Darriba-Bilbao, V. M. (2018). Cross-repository aggregation of educational resources. *Computers & Education*, 117(Supplement C):31 – 49.
- [82] Mouriño-García, M., Pérez-Rodríguez, R., Anido-Rifón, L., and Gómez-Carballa, M. (2016e). Bag-of-concepts document representation for bayesian text classification. In *Computer and Information Technology (CIT), 2016 IEEE International Conference on*, pages 281–288. IEEE.

- [83] Mouriño-García, M., Pérez-Rodríguez, R., Vilares-Ferro, M., and Anido-Rifón, L. (2016f). Wikipedia-based hybrid document representation for textual news classification. In *Soft Computing & Machine Intelligence (ISCFMI), 2016 3rd International Conference on*, pages 148–153. IEEE.
- [84] Müller, C. and Gurevych, I. (2009). Using wikipedia and wiktionary in domain-specific information retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 219–226. Springer.
- [85] Neven, F. and Duval, E. (2002). Reusable Learning Objects : a Survey of LOM-Based Repositories. *Evaluation*, 68(4):291–294.
- [86] Nezhadi, A. H., Shadgar, B., and Osareh, A. (2011). Ontology alignment using machine learning techniques. *International Journal of Computer Science & Information Technology*, 3(2):139.
- [87] Ngo, D. and Bellahsene, Z. (2012). Yam++: A multi-strategy based approach for ontology matching task. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 421–425. Springer.
- [88] Ni, X., Sun, J.-T., Hu, J., and Chen, Z. (2011). Cross lingual text classification by mining multilingual topics from wikipedia. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 375–384. ACM.
- [89] Nikolas, A., Sotiriou, S., Zervas, P., and Sampson, D. G. (2014). The open discovery space portal: A socially-powered and open federated infrastructure. In *Digital Systems for Open Access to Formal and Informal Learning*, pages 11–23. Springer.
- [90] OERCommons (2015). OER Commons. *Reference: http://www.oercommons.org/about, Last accessed: November 2017.*
- [91] Olsson, J. S., Oard, D. W., and Hajič, J. (2005). Cross-language text classification. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 645–646. ACM.
- [92] OpenStaxCNX (1999). OpenStax CNX. *https://cnx.org/about, Last accessed: November 2017.*
- [93] Oxford (2016). Oxford dictionaries.
- [94] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.
- [95] Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77.
- [96] Pérez-Rodríguez, R., Anido-Rifón, L., Gómez-Carballa, M., and Mouriño-García, M. (2016). Architecture of a concept-based information retrieval system for educational resources. *Science of Computer Programming*, pages 72–91.

Bibliografía

- [97] Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. (2008). Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In *Proceeding of the 17th international conference on World Wide Web - WWW '08*, pages 91–100.
- [98] Polavarapu, N., Navathe, S. B., Ramnarayanan, R., ul Haque, A., Sahay, S., and Liu, Y. (2005). Investigation into biomedical literature classification using support vector machines. In *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE*, pages 366–374. IEEE.
- [99] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.
- [100] Potthast, M., Stein, B., and Anderka, M. (2008). A wikipedia-based multilingual retrieval model. *Advances in Information Retrieval*, pages 522–530.
- [101] Reeve, L. and Han, H. (2005). Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 1634–1638. ACM.
- [102] Rigutini, L., Maggini, M., and Liu, B. (2005). An em based training algorithm for cross-language text categorization. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 529–535. IEEE Computer Society.
- [103] Roy, D., Sarkar, S., and Ghose, S. (2010). A comparative study of learning object metadata, learning material repositories, metadata annotation & an automatic metadata annotation tool. *Advances in Semantic Computing*, 2:103–126.
- [104] Sahlgren, M. and Cöster, R. (2004). Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th international conference on Computational Linguistics*, page 487. Association for Computational Linguistics.
- [105] Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.
- [106] Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.
- [107] Schmitt, M. and Schuller, B. (2017). openxbow—introducing the passau open-source crossmodal bag-of-words toolkit. *Journal of Machine Learning Research*, 18(96):1–5.
- [108] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.
- [109] Settles, B. (1994). Active learning literature survey. *Machine Learning*, 15(2):201–221.
- [110] Shi, L., Mihalcea, R., and Tian, M. (2010). Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067. Association for Computational Linguistics.

- [111] Simon, H. A. (1969). The sciences of the artificial. *Cambridge, MA*.
- [112] Sorg, P. and Cimiano, P. (2008). Cross-lingual information retrieval with explicit semantic analysis. In *Working Notes for the CLEF 2008 Workshop*.
- [113] Sriram, B., Fuhry, D., Demir, E., Ferhatoğlu, H., and Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM.
- [114] Stefaner, M., Dalla Vecchia, E., Condotta, M., Wolpers, M., Specht, M., Apelt, S., and Duval, E. (2007). MACE-enriching architectural learning objects for experience multiplication. In *European Conference on Technology Enhanced Learning, EC-TEL 2007: Creating New Learning Experiences on a Global Scale*, pages 322–336.
- [115] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.
- [116] Stock, W. G. (2010). Concepts and semantic relations in information science. *Journal of the American Society for Information Science and Technology*, 61(10):1951–1969.
- [117] Straccia, U. and Troncy, R. (2005). omap: Results of the ontology alignment contest. In *Workshop on Integrating Ontologies*, pages 92–96.
- [118] Täckström, O. (2005). *An Evaluation of Bag-of-Concepts Representations in Automatic Text Classification*. PhD thesis, KTH.
- [119] Ternier, S., Verbert, K., Parra, G., Vandepitte, B., Klerkx, J., Duval, E., Ordonez, V., and Ochoa, X. (2009). The Ariadne Infrastructure for Managing and Storing Metadata. *IEEE Internet Computing*, 13(4):18–25.
- [120] Tsai, C.-T. and Roth, D. (2016). Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.
- [121] Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1.
- [122] Upadhyay, S., Faruqui, M., Dyer, C., and Roth, D. (2016). Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425*.
- [123] Uysal, A. K. and Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, 50(1):104–112.
- [124] Van Rijsbergen, C. (1979). Information retrieval.
- [125] Viegas, F. B., Wattenberg, M., and Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 575–582. ACM.

Bibliografía

- [126] Vulić, I., De Smet, W., Tang, J., and Moens, M.-F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, 51(1):111–147.
- [127] Vulić, I. and Moens, M.-F. (2016). Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.
- [128] Wan, C., Pan, R., and Li, J. (2011). Bi-weighting domain adaptation for cross-language text classification. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1535.
- [129] 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: Volume 1-Volume 1*, pages 235–243. Association for Computational Linguistics.
- [130] Wang, P., Hu, J., Zeng, H.-J., Chen, L., and Chen, Z. (2007). Improving text classification by using encyclopedia knowledge. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 332–341. IEEE.
- [131] Wang, P., Hu, J., Zeng, H.-J., and Chen, Z. (2009). Using wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, 19(3):265–281.
- [132] Wei, B. and Pal, C. (2010). Cross lingual adaptation: an experiment on sentiment classifications. In *Proceedings of the ACL 2010 conference short papers*, pages 258–262. Association for Computational Linguistics.
- [133] White, T. (2012). *Hadoop: The definitive guide*. O'Reilly Media, Inc.
- [134] Wilbur, W. J. and Sirotnik, K. (1992). The automatic identification of stop words. *Journal of information science*, 18(1):45–55.
- [135] Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420.
- [136] Yetisgen-Yildiz, M. and Pratt, W. (2005). The effect of feature representation on medline document classification. In *AMIA Annual Symposium Proceedings*, volume 2005, page 849. American Medical Informatics Association.
- [137] Zhang, M.-L. and Zhou, Z.-H. (2005). A k-nearest neighbor based algorithm for multi-label classification. In *Granular Computing, 2005 IEEE International Conference on*, volume 2, pages 718–721. IEEE.
- [138] Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52.
- [139] Zhang, Z. Z. Z., Phan, X.-H. P. X.-H., and Horiguchi, S. (2008). An Efficient Feature Selection Using Hidden Topic in Text Categorization. *22nd International Conference on Advanced Information Networking and Applications - Workshops (aina workshops 2008)*.

- [140] Zhao, R. and Mao, K. (2017). Fuzzy bag-of-words model for document representation. *IEEE Transactions on Fuzzy Systems*.
- [141] Zheng, B., McLean, D. C., and Lu, X. (2006). Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC bioinformatics*, 7:58.
- [142] Zhou, X., Zhang, X., and Hu, X. (2008). Semantic smoothing for bayesian text classification with small training data. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 289–300. SIAM.
- [143] Zhu, X. (2006). Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison, Tech. Rep. 1530*.
- [144] Zhu, X. (2011). Semi-supervised learning. In *Encyclopedia of machine learning*, pages 892–897. Springer.

Apéndice I: Publicaciones

Cross-repository aggregation of educational resources

**Biomedical literature
classification using
encyclopedic knowledge: a
Wikipedia-based
bag-of-concepts approach**

Biomedical literature classification using encyclopedic knowledge: a Wikipedia-based bag-of-concepts approach

Marcos Antonio Mouriño García, Roberto Pérez Rodríguez and Luis E. Anido Rifón

Department of Telematics Engineering, University of Vigo, Vigo, Spain

ABSTRACT

Automatic classification of text documents into a set of categories has a lot of applications. Among those applications, the automatic classification of biomedical literature stands out as an important application for automatic document classification strategies. Biomedical staff and researchers have to deal with a lot of literature in their daily activities, so it would be useful a system that allows for accessing to documents of interest in a simple and effective way; thus, it is necessary that these documents are sorted based on some criteria—that is to say, they have to be classified. Documents to classify are usually represented following the bag-of-words (BoW) paradigm. Features are words in the text—thus suffering from synonymy and polysemy—and their weights are just based on their frequency of occurrence. This paper presents an empirical study of the efficiency of a classifier that leverages encyclopedic background knowledge—concretely Wikipedia—in order to create bag-of-concepts (BoC) representations of documents, understanding concept as “unit of meaning”, and thus tackling synonymy and polysemy. Besides, the weighting of concepts is based on their semantic relevance in the text. For the evaluation of the proposal, empirical experiments have been conducted with one of the commonly used corpora for evaluating classification and retrieval of biomedical information, OHSUMED, and also with a purpose-built corpus of MEDLINE biomedical abstracts, UVigoMED. Results obtained show that the Wikipedia-based bag-of-concepts representation outperforms the classical bag-of-words representation up to 157% in the single-label classification problem and up to 100% in the multi-label problem for OHSUMED corpus, and up to 122% in the single-label classification problem and up to 155% in the multi-label problem for UVigoMED corpus.

Submitted 3 August 2015
Accepted 7 September 2015
Published 29 September 2015

Corresponding author
Marcos Antonio Mouriño García,
marcosmourino@gmail.com
Academic editor
George Perry

Additional Information and
Declarations can be found on
page 19

DOI 10.7717/peerj.1279
© Copyright
2015 Mouriño García et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Bioinformatics, Science and Medical Education, Human-Computer Interaction, Computational Science

Keywords Biomedical literature, Classification, Wikipedia, Encyclopedic knowledge, Document representation, Bag-of-concepts, Bag-of-words, OHSUMED

INTRODUCTION

The ability to automatically classify text documents into a predefined set of categories is extremely convenient. Examples of this include the classification of educational resources

into subjects such as mathematics, science, or history; the classification of books into thematic areas; and the classification of news into sections such as economy, politics, or sports. Among these and other applications, the automatic classification of biomedical literature stands out as an important application to leverage text document automatic classification strategies. Medical staff, scientists, and biomedical researchers handle in their daily work huge amounts of literature and biomedical information, so it is necessary to have a system that allows for accessing to documents of interest in a simple, effective, efficient, and quick way; thus saving time querying or searching for these documents. This implies the necessity for sorting or ranking the documents based on some criterion, i.e., their classification.

Classification is modelled as a supervised learning problem: first, the classifier is trained with a certain number of examples—documents whose category is known—and then, the algorithm is applied to another set of documents whose category is unknown ([Sebastiani, 2002](#)). There is a huge amount of classification algorithms, including *k*-Nearest Neighbor (KNN), Decision Tree (DT), Neural Networks, Bayes, and Support Vector Machines (SVM) ([Yang, 1999](#)).

The functioning of classifiers is based on the application of Natural Language Processing (NLP) techniques to the documents to classify, so that a software agent can recognise which category a given document belongs to, based on some NLP feature contained in it, such as word occurrence frequency or the structure of the language used ([Settles, 2010](#)). Vector Space Model (VSM) ([Salton, Wong & Yang, 1975](#)) is the most often used representation, where each document within a collection is represented as a point in space, commonly using as weights the frequency of occurrence of words. When words are used as features, the model is known as bag-of-words, being a bag—or multiset—a set of elements that can occur more than once ([Blizard, 1988](#)). Then, by using this representation, a document is characterised by a set of words that appear in text, repeated as many times as occurrences in text.

Despite being one of the traditionally used representations in document classification tasks ([Täckström, 2005](#)), the BoW model is suboptimal, because it only accounts for word frequency in the documents, and ignores important semantic relationships between them ([Wang et al., 2008](#)). The main limitations of the BoW representation are redundancy, ambiguity, orthogonality, hyponymy and hypernymy problems, data sparseness and word usage diversity. Redundancy means that synonymous are not unified ([Egozi, Markovitch & Gabrilovich, 2011](#); [Huang & Milne, 2012](#)). For instance, in the BoW model, if a document that contains the word “tumour” was classified into the “cancer” category, this would not provide information to classify a document that contains the phrase “neoplasm”.

- Ambiguity refers to the problem of polysemy—one word can have several meanings ([Täckström, 2005](#); [Egozi, Markovitch & Gabrilovich, 2011](#)). For instance, in the BoW model, if a document that contains the word “tissue” was classified into the “human anatomy” category, it would may cause errors when classifying a document that contains the word “tissue”, but meaning the *Triphosa dubitata* moth.

- Orthogonality problem means that the semantic relatedness between words is not taken into account ([Huang & Milne, 2012](#)). For example, knowing that a document that contains “cardiovascular system” was classified under the label “circulatory system” would not give information on how to classify a document that contains the word “blood”.
- Hyponymy and hypernymy problem means that the hierarchical relations are not leveraged ([Levett, 1993](#); [Wang et al., 2007](#); [Wang et al., 2008](#)). For instance, if a document that contains the word “heart” was classified into “human body” category, this would not provide information to classify a document that contains the word “organ” and vice versa.
- BoW representations often suffer from the problems of data sparseness (zero-probability problem) and word usage diversity. This is because the BoW model only considers frequencies of words occurring in a class, and each document often contains only a small fraction of all words in the lexicon, which degrades the performance of the classifier ([Täckström, 2005](#); [Tsao, Chen & Wang, 2013](#)).

The most relevant works found in the literature focus mainly on solving two of the aforementioned problems: synonymy and the polysemy. To accomplish this, several authors have proposed a concept-based document representation, defining concept as “unit of meaning” ([Medelyan, Witten & Milne, 2008](#); [Wang et al., 2008](#); [Stock, 2010](#)). Several previous works demonstrated that this representation provides good results in classification tasks ([Sahlgren & Cöster, 2004](#); [Wang et al., 2008](#)).

The literature hosts several ways to create this based-of-concepts representation. In Latent Semantic Analysis (LSA) ([Deerwester et al., 1990](#); [Landauer & Dumais, 1997](#)) a concept is a vector that represents the context in which a term occurs; this approach overcomes synonymy but not polysemy. In Latent Dirichlet Allocation (LDA) ([Blei, Ng & Jordan, 2003](#)) each concept consists of a bag-of-words that represents an underlying topic in the text. In Explicit Semantic Analysis (ESA) ([Gabrilovich & Markovitch, 2007](#)) concepts are entries from external knowledge bases such as Wikipedia, WordNet, or Open Directory Project (ODP); these concepts are assigned to documents—annotation process—in accordance with its overlap with each entry in the knowledge base; its main disadvantage is its tendency toward generating outliers ([Egozi, Markovitch & Gabrilovich, 2011](#))—concepts that have a weak relationship to the document to annotate. Semantic annotators—the approach used in our proposal—extract concepts, disambiguate them, link them to domain-specific external sources—such as Unified Medical Language System (UMLS) or Medical Subject Headings (MeSH)—or to general-purpose external sources—such as Wikipedia—and deal with synonymy and polysemy problems.

We think that there is a research gap in the application of BoC representations that leverage encyclopedic knowledge in the building of classifiers of biomedical literature. This article aims at bridging this gap by designing, developing, and evaluating a classifier—single-label and multi-label—of biomedical literature that builds on encyclopedic knowledge and represents documents as bags-of-concepts. In order to evaluate the system,

we conducted several experiments with one of the most commonly used corpora for evaluating classification and retrieval of biomedical information—OHSUMED—as well as with a purpose-built corpus that comprises MEDLINE biomedical abstracts published in 2014—UVigoMED. Results obtained show a superior performance of the classifier when using the BoC representation, and it is an indicative of the potential of the proposed system to automatically classify scientific literature in the biomedical domain.

The remainder of this article is organised as follows: ‘Background’ presents some background knowledge; ‘Materials and Methods’ presents the corpora used, the algorithms, classification strategies, and metrics employed, and the approach proposed; ‘Results’ shows results obtained; ‘Discussion’ discusses the results obtained and presents proposals for future work; finally, ‘Conclusions’ presents the conclusions obtained.

BACKGROUND

In order to create the BoC representation of documents we use a general purpose semantic annotator. The literature contains other proposals for the creation of representations as bags-of-concepts. In this section, we discuss the main proposals for creating representations of documents as bags-of-concepts—Latent Semantic Analysis, Latent Dirichlet Allocation, Explicit Semantic Analysis, domain-specific semantic annotators, general-purpose semantic annotators, and hybrid semantic annotators—and proposals for biomedical literature classification that make use of these representations.

Latent Semantic Analysis

In the theoretical basis of Latent Semantic Analysis model underlies the distributional hypothesis ([Harris, 1968](#); [Sahlgren, 2008](#)): words that appear in similar contexts have similar meanings ([Deerwester et al., 1990](#); [Landauer & Dumais, 1997](#)). In LSA, the meaning of a word is represented as a vector of occurrences of that word in different contexts—being a context a text document. Although LSA combats the synonymy problem, it does not combat polysemy.

The LSA model has been used by several authors for biomedical literature classification tasks. [Kim, Howland & Park \(2005\)](#) explore the dimensionality reduction provided by LSA for classifying a subset of MEDLINE, reporting precision values reaching 90%. [Täckström \(2005\)](#) also makes use of the LSA model for the categorisation of a subset of MEDLINE, obtaining positive results using BoC in categories where BoW fails; despite the fact that results are positive, the author recommends using BoW as the primary representation mechanism and BoC as a punctual complement.

Latent Dirichlet Allocation

Latent Dirichlet Allocation model ([Blei, Ng & Jordan, 2003](#)) presupposes that each document within a collection comprises a small number of topics, each one of them “generating” words. Thus, LDA automatically finds topics in a text, or in other words, LDA attempts “to go back” from the document and find the set of topics that may have generated it. [Zheng, McLean & Lu \(2006\)](#) make use of LDA to identify biological topics—i.e., concepts—from a corpus composed of biomedical articles that belong to MEDLINE;

to that end, first, they use LDA to identify the most relevant concepts, and subsequently, these concepts are mapped to a biomedical vocabulary: Gene Ontology. *Phan, Nguyen & Horiguchi (2008)* get good results in the classification of short texts—OHSUMED abstracts—making use of a BoC document representation whose concepts were extracted using LDA. *Zhang, Phan & Horiguchi (2008)* focus on improving the performance of a classifier, making use of LDA to reduce the dimensionality of the set of features employed; the proposed method is applied to the biomedical corpus OHSUMED, obtaining results that demonstrate that the approach proposed provides better precision values, while reducing the size of the feature space.

Explicit Semantic Analysis

Gabrilovich & Markovitch (2007) propose Explicit Semantic Analysis, a technique that leverages external knowledge sources—as Wikipedia or ODP—to generate features from text documents. Contrary to LSA and LDA, ESA makes textual analysis identifying topics that are explicitly present in background knowledge bases—such as Wikipedia or ODP, among others—instead of latent topics. In other words, ESA analyses a text to index it with Wikipedia concepts. *Gabrilovich & Markovitch (2009)* use ESA to extract features from a text and to classify text documents from MEDLINE in categories. Authors report improvements in classification performance by using ESA to generate features of documents.

Semantic annotators

A semantic annotator is a software agent that is responsible for extracting the concepts that define a document, linking or mapping these concepts to entries from external sources. Semantic annotators usually perform disambiguation, thus combating synonymy and polysemy problems; and, in some cases, they assign a weight to each extracted concept in accordance with its semantic relevance within the document. Depending on the external source employed to link or map the extracted concepts, two kinds of semantic annotators can be distinguished: domain-specific semantic annotators and general-purpose semantic annotators.

Domain-specific semantic annotators

Domain-specific semantic annotators use external sources of a particular domain as knowledge bases to map extracted concepts. In the biomedical domain there are several biomedical ontologies, being the most relevant in the state-of-the-art MeSH (*Lowe & Barnett, 1994; Lipscomb, 2000*) and UMLS (*Bodenreider, 2004*). We can find several domain-specific semantic annotators in the literature. *Elkin et al. (1988)* propose a tool to identify MeSH terms in narrative texts. *Aronson (2001)* describes the MetaMap program, which embeds an algorithm that allows for representing biomedical texts through UMLS concepts. *Jonquet, Shah & Musen (2009)* present Open Biomedical Annotator: first, it extracts terms from text documents making use of Mgrep (*Dai et al., 2008*); second, it maps these terms to biomedical concepts from UMLS and other biomedical ontologies from the National Centre for Biomedical Ontologies (NCBO); and, finally, it annotates

the documents with these concepts. [Kang et al. \(2012\)](#) combine seven domain-specific annotators—ABNER, Lingpipe, MetaMap, OpenNLP Chunker, JNET, Peregrine and StandforNer—to extract medical concepts from clinical texts, providing better results than any of the individual systems alone. Several authors make use of these and other semantic annotators for biomedical classification tasks such as: [Yetisgen-Yildiz & Pratt \(2005\)](#), who use MetaMap to extract concepts from documents and use it to classify biomedical literature; and [Zhou, Zhang & Hu \(2008a\)](#), who use a semantic annotator based on UMLS (MaxMatcher ([Zhou, Zhang & Hu, 2006](#))) for the Bayesian classification of the biomedical literature corpus OHSUMED.

General-purpose semantic annotators

General-purpose semantic annotators use generic knowledge bases—they are not specific of a particular domain—such as Wikipedia, WordNet or FreeBase instead of domain-specific ontologies. [Vivaldi & Rodríguez \(2010\)](#) present a system to extract concepts from biomedical text using Wikipedia as semantic information source, and [Huang & Milne \(2012\)](#) propose the use of a semantic annotator—using Wikipedia and WordNet as knowledge bases—for creating BoC representations from documents and their use in biomedical literature classification tasks.

Hybrid semantic annotators

Hybrid semantic annotators use domain-specific ontologies—such as UMLS or MeSH—and generic knowledge bases—such as WordNet—as background knowledge to extract concepts from a narrative text. Thus, they leverage the advantages of both approaches—the specificity provided by domain-specific ontologies and the generality provided by generic knowledge bases. [Bloehdorn & Hotho \(2004\)](#) use this technique to enrich BoW representations of texts with concepts extracted from the text itself making use of MeSH ontology and the lexical database WordNet. This enriched representation is then used to perform the classification of the biomedical literature corpus OHSUMED, reporting F1-values of 48%.

MATERIALS AND METHODS

Dataset

OHSUMED

In order to evaluate the proposed system, we conducted four experiments with the well-known corpus for information retrieval and classification tasks OHSUMED. To carry out the experiments with the multi-label classifier, we used a subset of OHSUMED composed of 23,166 biomedical abstracts of 1991, classified into one or several of the 23 possible categories ([Joachims, 1998](#)). In order to create train and test sequences, we randomly split the corpus in a training sequence that comprises 18,533 documents and a test sequence composed of the remaining 4,633 documents.

To perform the single-label experiments, we removed from the aforementioned corpus those documents belonging to more than one category, resulting in a corpus formed by 9,034 documents classified in only one of the 23 categories; and, then, we randomised it

again to split it in a training sequence composed of 7,227 documents and a test sequence that comprises 1,807 documents.

UVigoMED

In order to corroborate the results obtained when conducting the experiments over OHSUMED corpus, we expressly created another corpus to conduct the same experiments as in OHSUMED. We named it UVigoMED.¹ In this section, we briefly describe the corpus and the process of collecting documents. First, we selected the classification scheme, consisting of the MeSH general terms of “diseases” group—the same as in OHSUMED. It is worth noting that, to create the UVigoMED corpus, we used the 2015 MeSH tree structure, where the diseases group contains 26 categories instead of the 23 that contained the MeSH tree structure when OHSUMED was created. To build the corpus we performed the following steps (see Fig. 1):

- We downloaded from MEDLINE all the descriptions of the articles (HTML webpages) of year 2014 classified under each one of the 26 categories.
- We extracted from each article description: the title, the abstract, and the categories it belongs to.
- We stored in our database the title, abstract and categories for each article description that was downloaded.

As a result, we obtained a corpus that comprises 92,661 biomedical articles classified in one or several categories of the 26 that were available. Finally, in order to create the training and test sequences, we randomly selected 18,532 documents as the test sequence, remaining 74,129 for the training sequence.

To carry out the single-label experiments, we created a subset of the aforementioned corpus comprising those documents belonging to just one category—by removing those that belonged to more than one category—resulting in a corpus composed of 54,853 documents classified in one of the 26 categories, and split randomly in a training sequence that comprises 43,882 documents and a test sequence composed by 10,971 items.

Multi-label classification methods

There are two main approaches to the multi-label classification problem: problem transformation methods and algorithm adaptation methods. Problem transformation methods are those that transform the multi-label problem in several single-label problems, whereas algorithm adaptation methods consist in performing adaptations of specific algorithms to address multi-label problems directly without performing any transformation.

In our proposal, we opted for using the methods of the first category, i.e., transforming the multi-label problem in N single-label binary problems, one for each category. To perform this, we made use of *Scikit-learn*, a module for Python that provides a set of the most relevant machine learning algorithms in the state-of-the-art (*Pedregosa et al., 2012*). In particular, we made use of the *one-vs-rest* or *one-vs-all* strategy, that automatically implements a classifier for each category. This strategy also allows for using different classification algorithms, including SVM, which is the one what we chose.

SVM algorithm

Support Vector Machines are a set of supervised machine learning algorithms used in clustering, regression, and classification tasks, among others. We selected the SVM algorithm because it is one of the most relevant algorithms in the state-of-the-art—together with Naïve Bayes, *k*-Nearest Neighbor, Decision Trees, or Neural Networks,—it is one of the most successful machine learning algorithms to perform automatic text classification tasks ([Rigutini, Maggini & Liu, 2005](#)) and it offers higher performance than other relevant algorithms of the state-of-the art such as KNN or Naïve Bayes ([Yang, 1999](#)). Although a more detailed definition can be found in [Hearst et al. \(1998\)](#), the basic idea is that, given a set of items belonging to a set of categories, SVM builds a model that can predict which category the new items that appear in the system belong to. SVM represents each item as a point in space, separating the categories as much as possible. Then, when a new item appears in the model, it will be placed in one category or another, depending on their proximity to each one. This algorithm corresponds to the class `sklearn.svm.LinearSVC` of the *Scikit-learn* library.

Evaluation metrics

The single-label and multi-label classification problems make use of different evaluation metrics. Hereafter, we cite the main metrics that the literature shows to evaluate each of the problems.

Single-label classification problem

When predicting the category to which a document belongs, there are four possible outcomes: true positive (TP), true negative (TN), false positive (FP) and false negative (FN), where *positive* means that a document was classified in a certain category, *negative* means the opposite, *true* means that the classification was correct and *false* means that the classification was incorrect ([Sahlgren & Cöster, 2004](#)).

In the same way as [Sebastiani \(2002\)](#) and [Sahlgren & Cöster \(2004\)](#), we define:

$$P = \text{Precision} = \frac{TP}{(TP + FP)} \quad (1)$$

$$R = \text{Recall} = \frac{TP}{(TP + FN)}. \quad (2)$$

We also use a measure that combines precision and recall, F1-score, defined as:

$$F_1 = \frac{2 * P * R}{P + R}. \quad (3)$$

In our work we report the results as macro-F1, because it is the best metric to reflect the classification performance in corpora where data are not evenly distributed over different categories ([Zhou, Zhang & Hu, 2008a](#)).

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed Animal Diseases [Mesh] Search Advanced

Abstract Send to:

Gene. 2015 Sep 1;568(2):117-23. doi: 10.1016/j.gene.2015.05.023. Epub 2015 May 13.

A mutation in the NLRC5 promoter limits NF- κ B signaling after Salmonella Enteritidis infection in the spleen of young chickens.

Chang G¹, Liu X², Ma T¹, Xu L¹, Wang H¹, Li Z¹, Guo X¹, Xu Q¹, Chen G³.

Author information

Abstract

To date, the functions of the NLRC5 in chickens remain undefined. In the current study, chicken NLRC5 was cloned and an A1017G mutation was detected in its promoter region. The relative expression levels of the NLRC5 and key NF- κ B pathway genes, IKK α , IKK β , NF- κ B, IL-6, IL-1 β and IFN- γ , in the spleens of wild and mutant type birds, AA and GG, were determined using FQ-PCR at 7 day post-infection (DPN) with Salmonella Enteritidis. Additionally, the bacterial burden in the caecum and various immune response parameters were measured to evaluate immune responses. All of the examined immune response parameters were significantly different between the AA chickens and the GG chickens. Specifically, the mRNA expression levels of IKK α , NF- κ B, IL-6, IL-1 β and IFN- γ were higher in AA chickens than those in GG chickens, while the mRNA expression levels of NLRC5 were lower in AA chickens than those in GG chickens ($P<0.05$). Moreover, the mRNA expression levels of TLR4 and MyD88 were not affected in either group. Collectively, considering former NLRC5 functional study in vitro, the wild genotype birds presented with better resistance to Salmonella Enteritidis through the actions of the NLRC5 and subsequent inhibition of the NF- κ B pathway in chickens.

Copyright © 2015 Elsevier B.V. All rights reserved.

KEYWORDS: Chicken; Immune response; NLRC5; Salmonella

PMID: 25979675 [PubMed - indexed for MEDLINE]



id	title	abstract	category
1	Clinical and demographic findings of patients with...	Rheumatoid arthritis (RA) and ankylosing spondylitis (AS) ...	Immune System Diseases
2	Luteolin provides neuroprotection in models ...	Luteolin has recently been proven to exert neuroprotect...	Wounds and Injuries
3	A mutation in the NLRC5 promoter limits NF- κ B ...	To date, the functions of the NLRC5 in chickens remain ...	Animal Diseases
...

Figure 1 UVigoMED corpus creation.

Multi-label classification problem

[Schapire & Singer \(2000\)](#) consider in their work the *Hamming Loss*, defined according to [Tsoumakas & Katakis \(2007\)](#) as

$$HL = \text{Hamming Loss}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \Delta Z_i|}{|L|} \quad (4)$$

being D the multi-label corpus, that comprises $|D|$ multi-label elements $(x_i, Y_i), i = 1 \dots |D|$, $Y_i \subseteq L$, L is the set of labels, composed by $|L|$ labels, H is a multi-label classifier, $Z = H(x_i)$ is the set of labels predicted by H for x_i , and Δ represents the symmetric difference between two sets, corresponding to the XOR operation in the Boolean algebra.

The following metrics—*Accuracy*, *Precision*, and *Recall*—are used by [Godbole & Sarawagi \(2004\)](#) and defined again by [Tsoumakas & Katakis \(2007\)](#) as:

$$A = \text{Accuracy}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \quad (5)$$

$$P = \text{Precision}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Z_i|} \quad (6)$$

$$R = \text{Recall}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i|}. \quad (7)$$

We also use the F1-score, defined in the previous section.

Approach

The approach presented consists in the classification—single-label and multi-label—of the two corpora of biomedical literature defined in ‘Dataset’ using a Wikipedia-based bag-of-concepts representation of documents, and the comparison of the performance with the performance of the classifier when using the traditional BoW representation of documents. We used the SVM algorithm (‘SVM algorithm’), and for the multi-label problem, we also made use of the strategy presented in ‘Multi-label classification methods’. With the aim of conducting all the experiments under the same conditions, we selected randomly for both corpora—single-label and multi-label versions—training sequences composed of 5,000 elements and test sequences that comprise 1,000 elements.

First, it was necessary to obtain the BoW and BoC representations of each document in the corpora. [Figure 2](#) shows the differences between the creation of the traditional BoW representation and the BoC representation. In order to create the BoW representation of a document, the first step is to filter the stop words. Stop words are words such as “the”, “if”, and “or” that are of no use for text classification, since they probably occur in almost all documents. The next step is stemming, the removing of common inflexional affixes, in order to perform some form of morphological normalization to create more general features. To that end we use the *Porter stemmer* ([Porter, 1980](#)), which is the most common

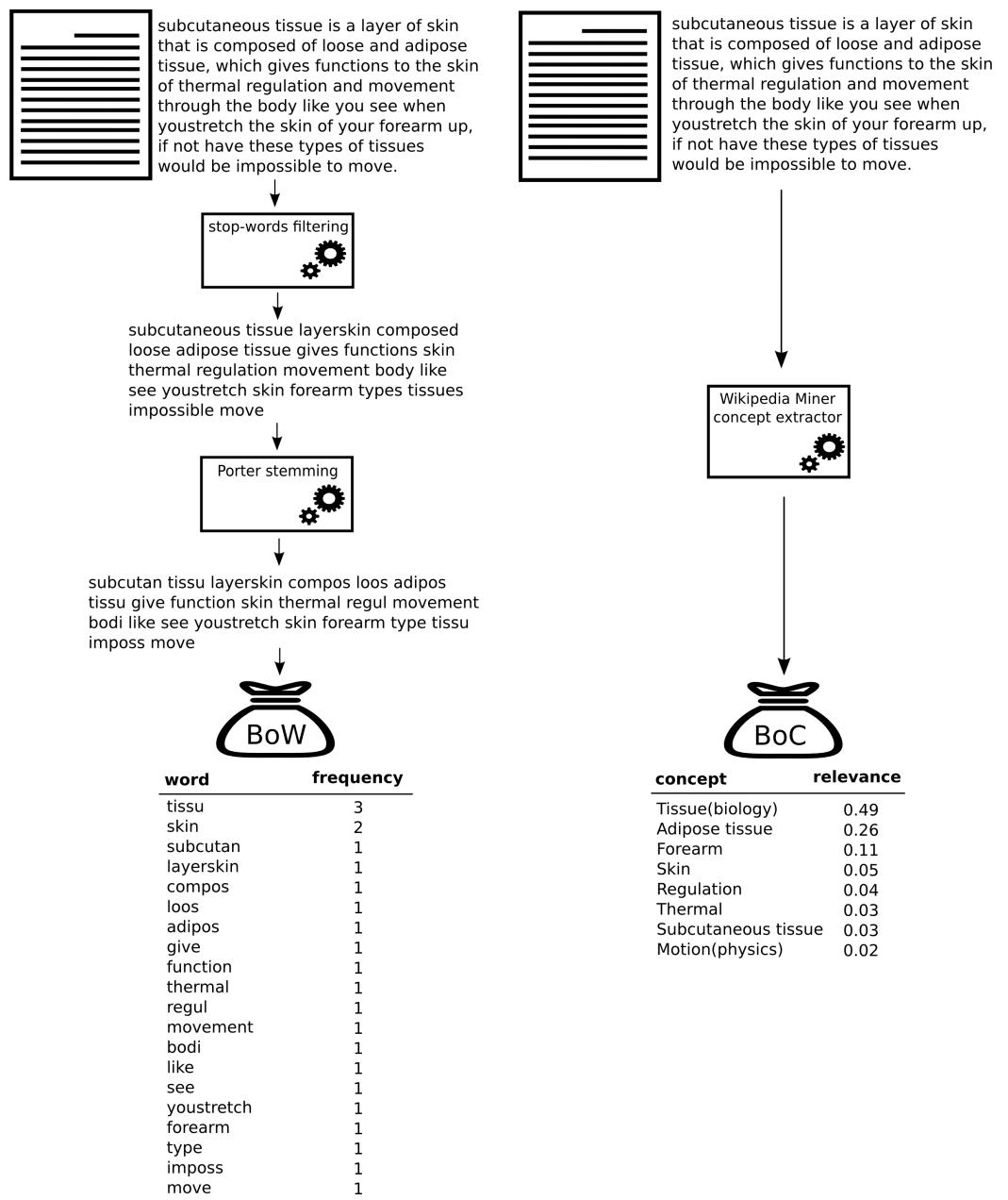


Figure 2 Bag-of-words and bag-of-concepts creation process.

stemming algorithm to work with English text ([Täckström, 2005](#)). Finally, we calculate the frequency of occurrence of stemmed words.

To create the BoC, we opted for using a semantic annotator, in particular, a general-purpose semantic annotator that uses NLP techniques, machine learning, and Wikipedia as a knowledge base: *Wikipedia Miner* ([Milne & Witten, 2013](#)). The implementation of its algorithm is based on three steps:

- The first step is *candidate selection*. It consists in, given a text document composed of a set of *n*-grams—being an *n*-gram continuous sequence of n words—the algorithm

queries a vocabulary that comprises all the *anchor texts* in Wikipedia and verifies whether any of the *n-grams* are present in the vocabulary. Thus, for each matching *n-gram-anchor text* a candidate is obtained, being the most relevant candidates those that are most frequently used as *anchor texts* in Wikipedia.

- The second step is *disambiguation*. Given the same vocabulary of *anchor texts*, the algorithm selects the most suitable target for each candidate. The process is performed making use of machine learning techniques and using Wikipedia articles as the training sequence, since they contain good examples of manually performed disambiguation. Disambiguation is accomplished having into account the relationship of each candidate with other non-ambiguous terms in its context, and also the commonness of the candidate.
- The third step is *link detection*, wherein the relevance of concepts extracted from the text is calculated. To that end, the algorithm uses again machine learning techniques and Wikipedia articles as the training sequence, since each of them is a good example of what constitutes a relevant link and what does not. [Figure 3](#) shows graphically the whole process to obtain a bag-of-concepts—being each one of them a Wikipedia article—from a text document.

Having obtained the BoC representation for each of the documents we proceeded to classify the two corpora—both single-label and multi-label versions—making use of the strategies and algorithms defined in ‘Multi-label classification methods’ and ‘SVM algorithm’.

RESULTS

OHSUMED

[Figure 4](#) and [Table 1](#) show the evolution of the F1-score for BoW and BoC, varying the length of the training sequence for the single-labelled OHSUMED corpus; and [Fig. 5](#) and [Table 2](#), show the F1-score for BoW and BoC, varying the length of the training sequence in the multi-labelled OHSUMED corpus. We can perceive that the performance offered by the classifier using the BoC representation is clearly superior to the one offered by the traditional BoW representation for both experiments—single-label and multi-label. As we can see in [Fig. 8](#), the BoC representation reaches improvements up to 157% for the single-label problem and up to 100% for the multi-label problem.

UVigoMED

[Figure 6](#) and [Table 3](#) show the evolution of the F1-score for BoW and BoC when varying the length of the training sequence for the single-label version of the UVigoMED corpus. We can see that the performance offered by the classifier when using the BoC representation is much higher than that offered when using the BoW one, reaching improvements up to 122%, as shown in [Fig. 8](#). The experiments conducted with the multi-label corpus provide the results shown in [Fig. 7](#) and [Table 4](#), where we can see again that the BoC representation outperforms BoW, reaching increases up to 155%.

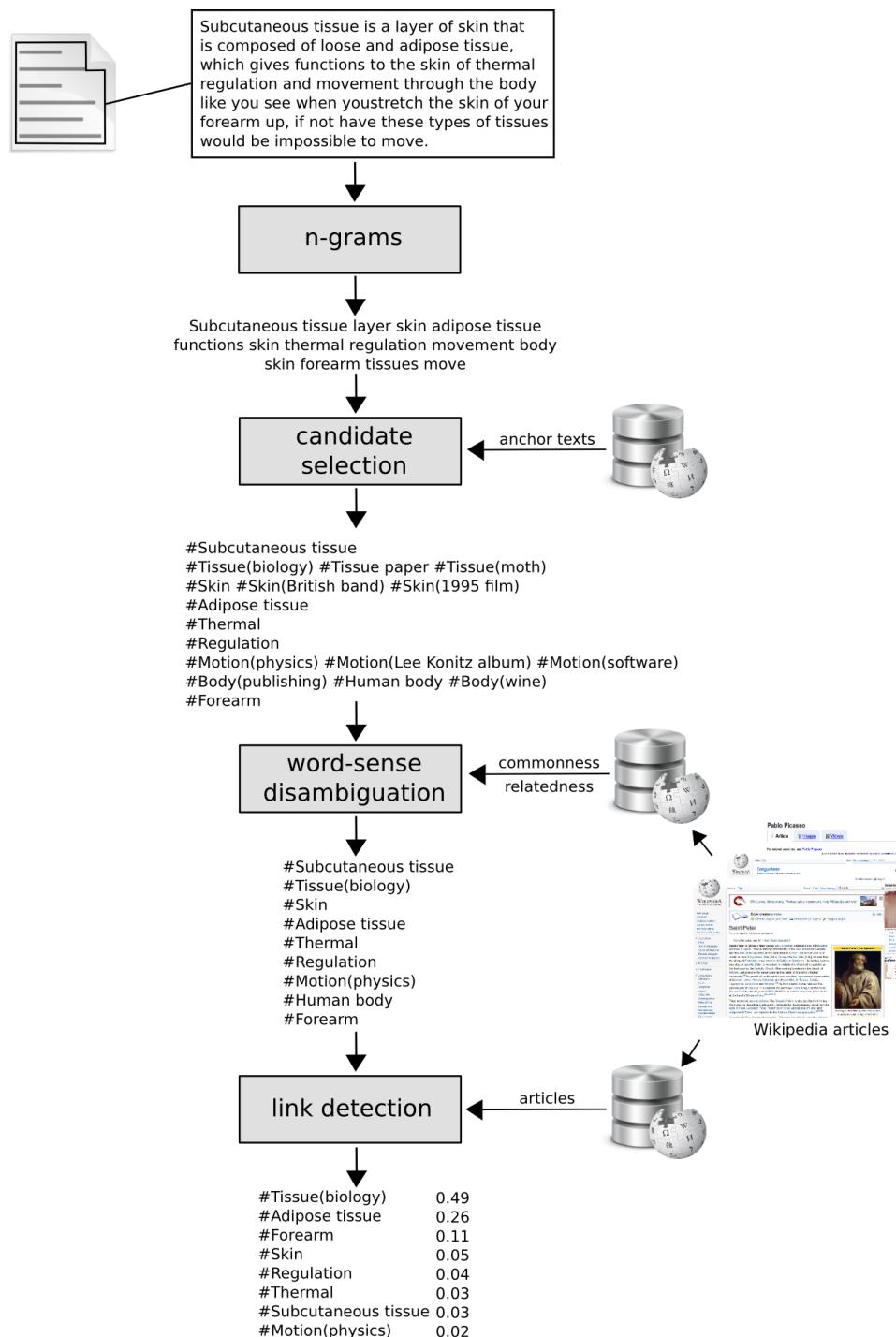


Figure 3 Bag-of-concepts obtainment process of a document using Wikipedia Miner.

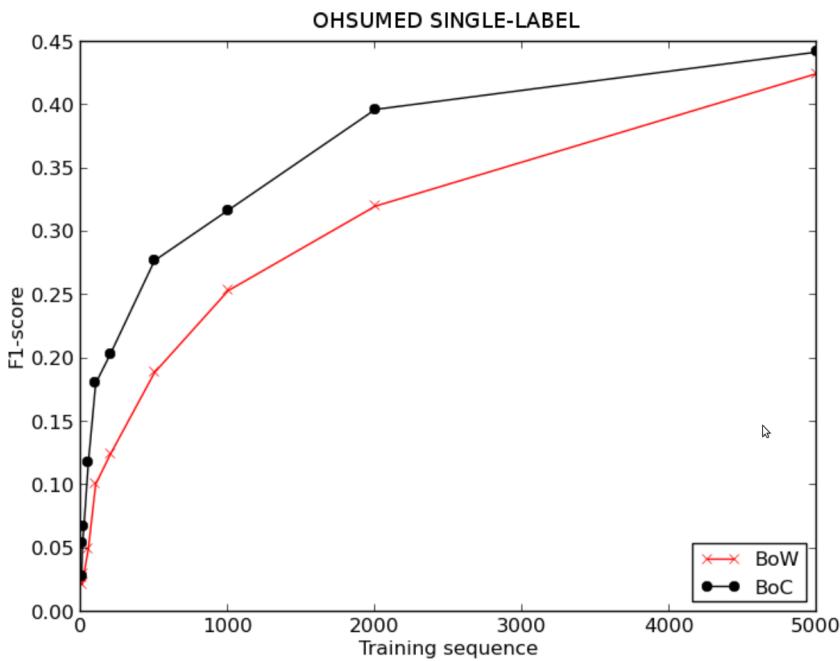


Figure 4 F1 score for BoW and BoC varying the length of the training sequence in single-labelled OHSUMED corpus.

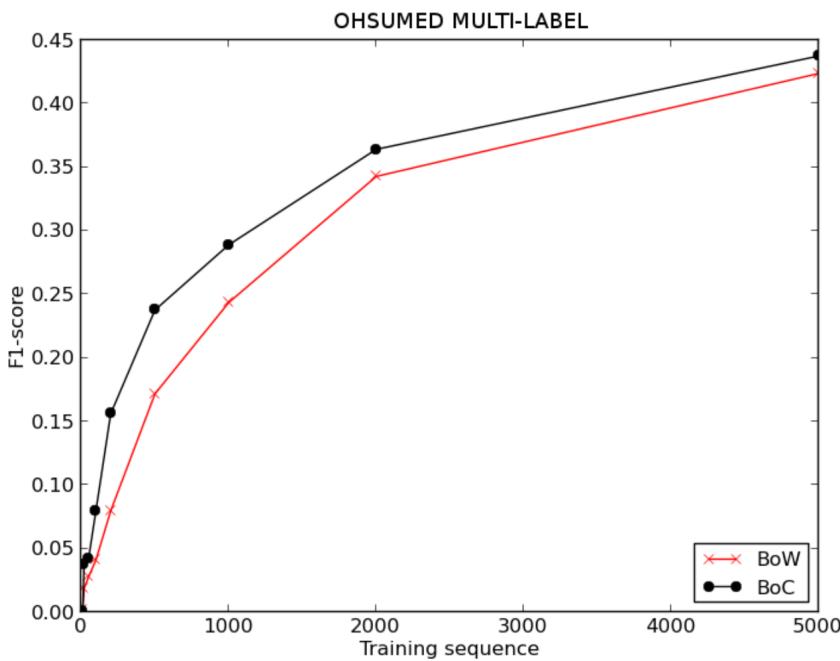


Figure 5 F1-score for BoW and BoC varying the length of the training sequence in multi-labelled OHSUMED corpus.

Table 1 F1-score for BoW and BoC varying the length of the training sequence in single-labelled OHSUMED corpus.

		5	10	20	50	100	200	500	1,000	2,000	5,000
BoW	P	0.058	0.080	0.102	0.160	0.218	0.276	0.345	0.418	0.467	0.528
	R	0.129	0.074	0.106	0.163	0.248	0.307	0.377	0.426	0.471	0.519
	F1	0.027	0.021	0.030	0.050	0.101	0.125	0.189	0.254	0.320	0.425
BoC	P	0.089	0.134	0.213	0.281	0.308	0.332	0.421	0.460	0.512	0.535
	R	0.078	0.151	0.173	0.237	0.309	0.355	0.421	0.470	0.502	0.535
	F1	0.029	0.054	0.067	0.118	0.181	0.204	0.277	0.317	0.397	0.442

Table 2 Hamming loss, precision, accuracy, recall and F1-score for BoW and BoC varying the length of the training sequence in multi-labelled OHSUMED corpus.

		5	10	20	50	100	200	500	1,000	2,000	5,000
BoW	HL	0.061	0.061	0.063	0.063	0.062	0.062	0.059	0.058	0.056	0.054
	P	0.180	0.063	0.147	0.300	0.380	0.461	0.507	0.532	0.560	0.571
	A	0.000	0.000	0.016	0.026	0.030	0.049	0.121	0.156	0.198	0.192
	R	0.001	0.002	0.031	0.047	0.056	0.082	0.200	0.288	0.385	0.482
	F1	0.001	0.002	0.019	0.028	0.041	0.080	0.172	0.244	0.343	0.424
BoC	HL	0.061	0.061	0.060	0.060	0.059	0.058	0.057	0.057	0.056	0.051
	P	0.021	0.063	0.300	0.415	0.457	0.526	0.543	0.553	0.556	0.591
	A	0.000	0.000	.0033	0.060	0.077	0.111	0.148	0.171	0.184	0.202
	R	0.001	0.001	0.054	0.085	0.121	0.182	0.273	0.340	0.404	0.481
	F1	0.001	0.001	0.038	0.042	0.080	0.156	0.238	0.289	0.364	0.438

Table 3 F1-score for BoW and BoC varying the length of the training sequence in single-labelled UVigoMED corpus.

		5	10	20	50	100	200	500	1,000	2,000	5,000
BoW	P	0.059	0.122	0.102	0.116	0.179	0.276	0.377	0.460	0.518	0.629
	R	0.060	0.074	0.097	0.150	0.183	0.272	0.397	0.457	0.511	0.631
	F1	0.026	0.027	0.035	0.061	0.084	0.136	0.220	0.283	0.360	0.421
BoC	P	0.095	0.222	0.284	0.259	0.308	0.436	0.500	0.544	0.586	0.594
	R	0.049	0.093	0.148	0.247	0.321	0.432	0.515	0.557	0.590	0.598
	F1	0.017	0.040	0.078	0.116	0.179	0.269	0.331	0.390	0.430	0.467

DISCUSSION

The results presented in the previous section clearly show the increase in performance of a SVM classifier for categorising biomedical literature when using a Wikipedia-based bag-of-concepts document representation instead of the classical representation based-on-words. It is worth noting that, as can be seen in Fig. 8, the highest increases occur when training sequences are short, because, with enough data, the problems of synonymy and polysemy are masked, and surface overlap performs well.

The increase in classifiers' performance yields important benefits for users—fundamentally medical staff, researchers and students—since a suitable and correct

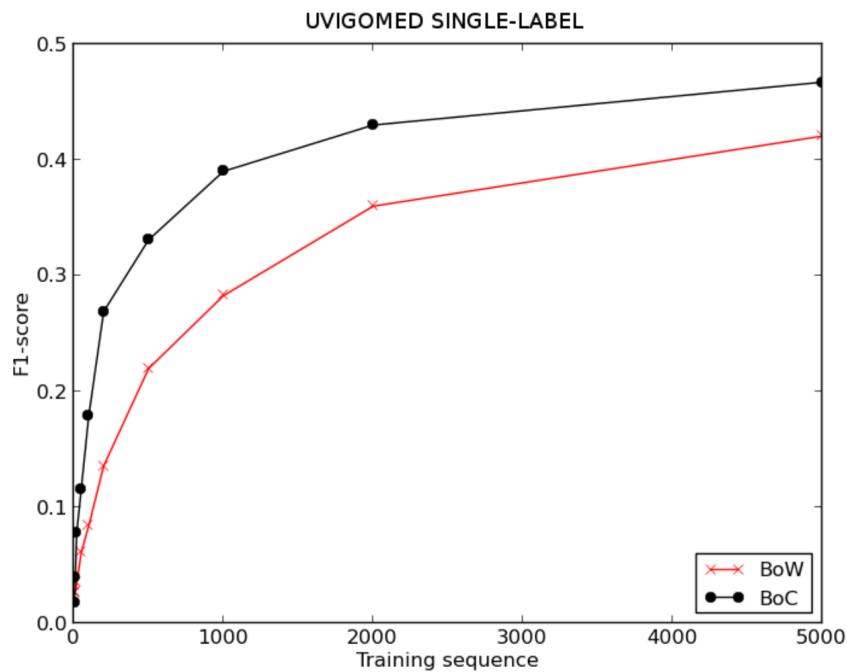


Figure 6 F1 score for BoW and BoC varying the length of the training sequence in single-labelled UVigoMED corpus.

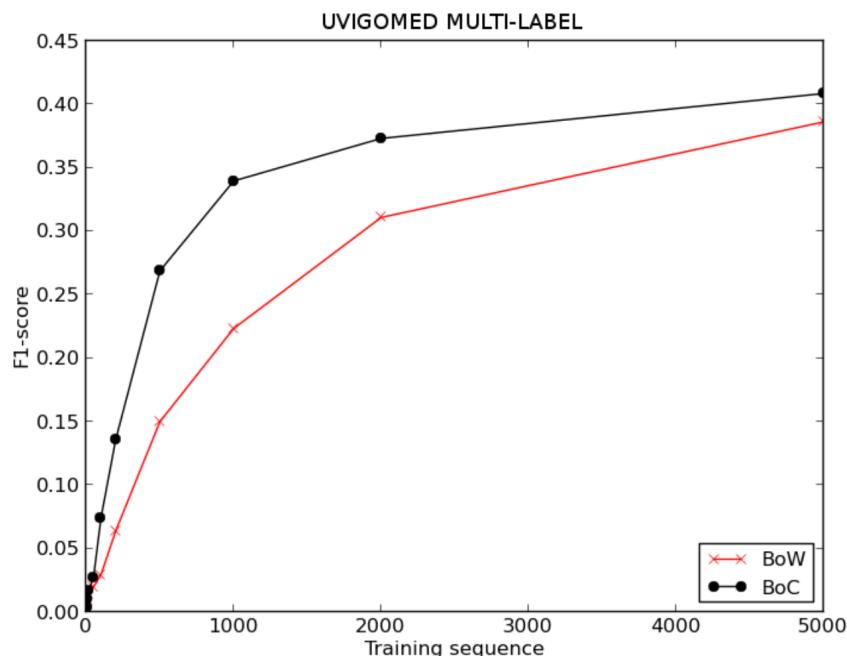


Figure 7 F1-score for BoW and BoC varying the length of the training sequence in multi-labelled UVigoMED corpus.

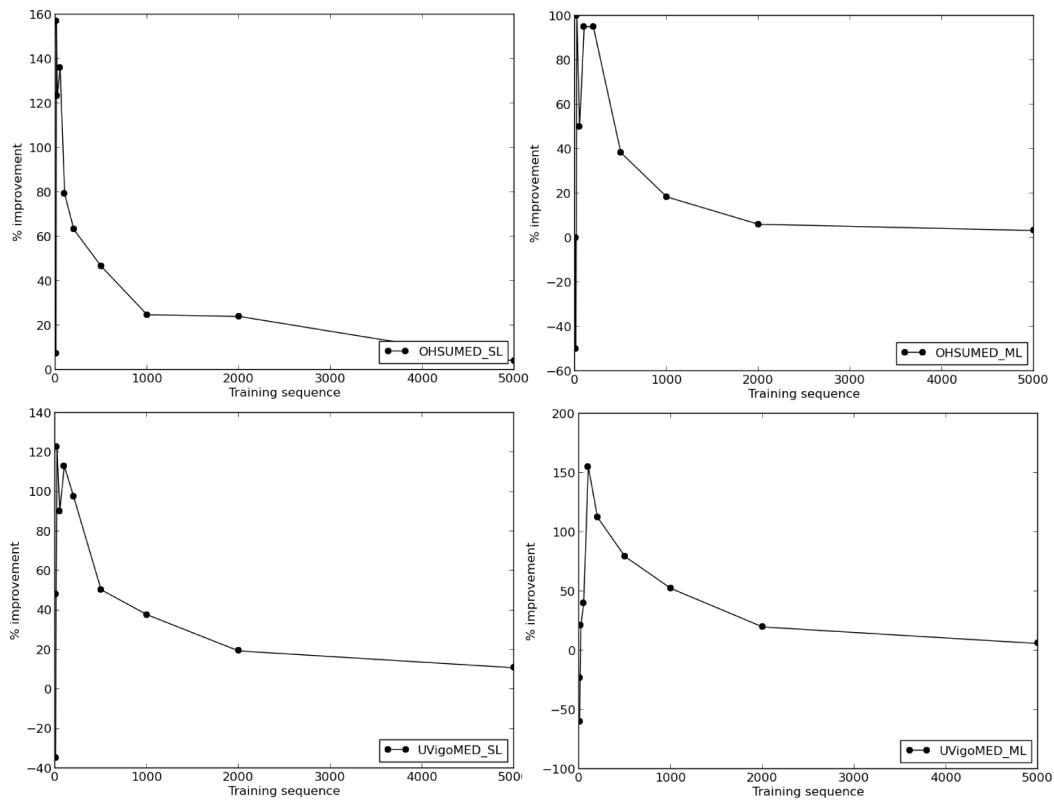


Figure 8 F1-score percentage improvement for single-labelled OHSUMED, multi-labelled OHSUMED, single-labelled UVigoMED and multi-labelled UVigoMED according to training sequence length variation.

Table 4 Hamming loss, precision, accuracy, recall and F1-score for BoW and BoC varying the length of the training sequence in multi-labelled UVigoMED corpus.

	5	10	20	50	100	200	500	1,000	2,000	5,000
BoW	HL	0.087	0.090	0.068	0.064	0.065	0.065	0.061	0.059	0.056
	P	0.069	0.033	0.114	0.107	0.160	0.328	0.489	0.544	0.573
	A	0.001	0.010	0.090	0.011	0.019	0.034	0.083	0.136	0.182
	R	0.024	0.031	0.015	0.016	0.029	0.062	0.152	0.229	0.312
	F1	0.010	0.013	0.014	0.020	0.029	0.064	0.150	0.223	0.311
BoC	HL	0.086	0.071	0.062	0.061	0.060	0.060	0.056	0.054	0.052
	P	0.001	0.140	0.225	0.186	0.411	0.536	0.589	0.601	0.606
	A	0.022	0.014	0.009	0.017	0.043	0.081	0.156	0.199	0.217
	R	0.021	0.021	0.014	0.023	0.069	0.138	0.282	0.364	0.414
	F1	0.004	0.010	0.017	0.028	0.074	0.136	0.269	0.340	0.373

categorisation facilitates access to those biomedical articles that are really of interest, thus reducing the time needed to find them.

Comparing the proposed approach to other similar approaches in the literature is not an easy task, due to the lack of biomedical literature classification systems that

use a general-purpose semantic annotator, the variety of corpora—and subsets of them—employed, the variety of classification algorithms employed, and the different performance measures used. The only work that uses a general-purpose semantic annotator to classify biomedical literature is [Huang & Milne \(2012\)](#), who classify a subset of MEDLINE—Med100, without specifying whether it is single or multi-label—using a KNN algorithm, and with a proportion of training documents similar to our work (83%). The authors report a F1-score—they do not specify whether it is macro or micro—about 53%. Regarding the use of domain-specific semantic annotators to create representations of documents in biomedical literature classification tasks, we can cite the work of [Zhou, Zhang & Hu \(2008b\)](#), where the authors classify, using a Naïve Bayes algorithm, a subset of OHSUMED corpus comprising only 7,400 documents of the year 1991 that belong to just one category from a total of 14—instead of the 23 that comprises the original OHSUMED corpus—obtaining a macro F1-score of 64%, and using as training sequence 33% of documents of the corpus; and [Yetisgen-Yildiz & Pratt \(2005\)](#), where the authors use an SVM to classify a non-standard subset of OHSUMED corpus composed of 179,796 titles of biomedical articles belonging to 1,928 MeSH categories—without specifying if it is single or multi-label—providing micro F1-score values of 57%.

Finally, the study leaves open lines to future research. The work presented in this paper may be extended by applying the classifier and document representation proposed to the classification of medical histories and patient records, using the proposed document representation along with other classification strategies and algorithms, experimenting with other semantic annotators, and conducting more experiments with other corpora. Another possible future line is the design and development of a software application that allows the visualisation of the documents classified according to the proposal presented in this paper. Thus, users may interact with the application and perform exploratory searches through the categories in which documents are classified. In addition, this will allow us to receive input from users about the results of the classification.

CONCLUSIONS

This study presents the benefits of using a Wikipedia-based bag-of-concepts document representation and its application to the SVM classification algorithm to classify biomedical literature into a predefined set of categories. The experiments conducted showed that the BoC representation outperforms the classical BoW representation by up to 157% for the single-label problem and up to 100% for the multi-label problem for the OHSUMED corpus. In addition, we created a purpose-built corpus—UVigoMED—that comprises biomedical articles belonging to MEDLINE of year 2014, in which the performance of the classifier using the BoC representation outperforms BoW by up to 122% for the single-label problem and up to 155% in the multi-label problem.

In consequence, we conclude that a Wikipedia-based bag-of-concepts document representation is superior to a baseline BoW representation when it comes to classifying biomedical literature. This is especially true when training sequences are short.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Research partially supported by the Galician Regional Government under project GRC2013-006 (Consolidation of Research Units) and through REDPLIR (Red Gallega de Procesamiento del Lenguaje y Recuperacion de Informacion)—R2014/034. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

Galician Regional Government: GRC2013-006.

REDPLIR (Red Gallega de Procesamiento del Lenguaje y Recuperacion de Informacion): R2014/034.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Marcos Antonio Mouriño García and Roberto Pérez Rodríguez conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Luis E. Anido Rifón contributed reagents/materials/analysis tools, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

UVigoMED Corpus: <http://itec-sde.net/UVigoMED.zip>.

REFERENCES

- Aronson AR.** 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *AMIA Annual Symposium Proceedings* 17–21.
- Blei DM, Ng AY, Jordan MI.** 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.
- Blizard WD.** 1988. Multiset theory. *Notre Dame Journal of Formal Logic* 30(1):36–66 DOI [10.1305/ndjfl/1093634995](https://doi.org/10.1305/ndjfl/1093634995).
- Bloehdorn S, Hotho A.** 2004. Boosting for text classification with semantic features. In: *WebKDD*. Vol. 3932. Springer, 149–166 DOI [10.1007/11899402_10](https://doi.org/10.1007/11899402_10).
- Bodenreider O.** 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32:D267–D270 DOI [10.1093/nar/gkh061](https://doi.org/10.1093/nar/gkh061).
- Dai M, Shah NH, Xuan W, Musen MA, Watson SJ, Athey BD, Meng F.** 2008. An efficient solution for mapping free text to ontology terms. In: *AMIA summit on translational bioinformatics, San*

Francisco CA, vol. 21. Available at <http://knowledge.amia.org/amia-55142-tbi2008a-1.650887/t-002-1.985042/f-001-1.985043/a-041-1.985157/an-041-1.985158?qr=1>.

- Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R.** 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* **41**:391–407 DOI [10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASI1>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9).
- Egozi O, Markovitch S, Gabrilovich E.** 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems* **29**(2):1–38 DOI [10.1145/1961209.1961211](https://doi.org/10.1145/1961209.1961211).
- Elkin PL, Cimino JJ, Lowe HJ, Aronow DB, Payne TH, Pincetl PS, Barnett GO.** 1988. Mapping to MeSH: the art of trapping MeSH equivalence from within narrative text. In: *Proceedings of the annual symposium on computer application in medical care*. 185–190.
- Gabrilovich E, Markovitch S.** 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: *Proceedings of the 20th international joint conference on artificial intelligence*. 1606–1611.
- Gabrilovich E, Markovitch S.** 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research* **34**:443–498.
- Godbole S, Sarawagi S.** 2004. Discriminative methods for multi-labeled classification. In: *Advances in knowledge discovery and data, Lecture notes in computer science*, vol. 3056. 22–30 Available at <http://www.springerlink.com/index/maa4ag38jd3pwrc0.pdf>.
- Harris ZS.** 1968. *Mathematical structures of language*.
- Hearst M, Dumais S, Osman E, Platt J, Scholkopf B.** 1998. Support vector machines. *Intelligent Systems and their Applications, IEEE* **13**(4):18–28 DOI [10.1109/5254.708428](https://doi.org/10.1109/5254.708428).
- Huang L, Milne D.** 2012. Learning a concept-based document similarity measure. *Journal of the American Society for Information Science and Technology* **63**:1593–1608 DOI [10.1002/as.22689](https://doi.org/10.1002/as.22689).
- Joachims T.** 1998. Text categorization with support vector machines: learning with many relevant features. In: *Machine learning: ECML-98*. Vol. 1398. New York: Springer, 137–142. Available at <http://link.springer.com/chapter/10.1007%2FBFb0026683>.
- Jonquet C, Shah NH, Musen MA.** 2009. The open biomedical annotator. *Summit on Translational Bioinformatics 2009*:56–60.
- Kang N, Afzal Z, Singh B, Van Mulligen EM, Kors JA.** 2012. Using an ensemble system to improve concept extraction from clinical records. *Journal of Biomedical Informatics* **45**:423–428 DOI [10.1016/j.jbi.2011.12.009](https://doi.org/10.1016/j.jbi.2011.12.009).
- Kim H, Howland P, Park H.** 2005. Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research* **6**:37–53.
- Landauer TK, Dumais ST.** 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* **104**(2):211–240 DOI [10.1037/0033-295X.104.2.211](https://doi.org/10.1037/0033-295X.104.2.211).
- Levelt WJ.** 1993. *Speaking: from intention to articulation, ACL-MIT press series in natural-language processing*, vol. 1. Cambridge: MIT Press. Available at <https://mitpress.mit.edu/books/speaking>.
- Lipscomb CE.** 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* **88**(3):265–266.
- Lowe HJ, Barnett GO.** 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association* **271**:1103–1108 DOI [10.1001/jama.1994.03510380059038](https://doi.org/10.1001/jama.1994.03510380059038).

- Medelyan O, Witten IH, Milne D.** 2008. Topic indexing with Wikipedia. In: *Proceedings of the AAAI WikiAI workshop*. 19–24.
- Milne D, Witten IH.** 2013. An open-source toolkit for mining Wikipedia. *Artificial Intelligence* 194:222–239 DOI 10.1016/j.artint.2012.06.007.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E.** 2012. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* 12:2825–2830.
- Phan X-H, Nguyen L-M, Horiguchi S.** 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: *Proceedings of the 17th international conference on World Wide Web - WWW '08*. 91–100. Available at <http://portal.acm.org/citation.cfm?doid=1367497.1367510>.
- Porter MF.** 1980. An algorithm for suffix stripping. *Program* 14(3):130–137 DOI 10.1108/eb046814.
- Rigutini L, Maggini M, Liu B.** 2005. An EM based training algorithm for cross-language text categorization. In: *Proceedings—2005 IEEE/WIC/ACM international conference on web intelligence, WI 2005*, vol. 2005. 529–535.
- Sahlgren M.** 2008. The distributional hypothesis. *Italian Journal of Linguistics* 20(1):33–54.
- Sahlgren M, Cöster R.** 2004. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In: *Proceedings of the 20th international conference on computational linguistics*. Available at <http://dl.acm.org/citation.cfm?id=1220425>.
- Salton G, Wong A, Yang CS.** 1975. A vector space model for automatic indexing. *Communications of the ACM* 18(11):613–620 DOI 10.1145/361219.361220.
- Schapire RE, Singer Y.** 2000. BoosTexter: a boosting-based system for text categorization. *Machine Learning* 39:135–168 DOI 10.1023/A:1007649029923.
- Sebastiani F.** 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34(1):1–47 DOI 10.1145/505282.505283.
- Settles B.** 2010. Active learning literature survey. *Machine Learning* 15(2):201–221.
- Stock WG.** 2010. Concepts and semantic relations in information science. *Journal of the American Society for Information Science and Technology* 61(10):1951–1969 DOI 10.1002/asi.21382.
- Täckström O.** 2005. An evaluation of bag-of-concepts representations in automatic text classification. Doctoral dissertation, KTH, 1–72. Available at http://www.nada.kth.se/utbildning/grukth/exjobb/rapportlistor/2005/rapporter05/tackstrom_oscar_05150.pdf.
- Tsao Y, Chen KY, Wang HM.** 2013. Semantic naïve Bayes classifier for document classification. In: *International joint conference on natural language processing*. 1117–1123. Available at <http://www.aclweb.org/anthology/I/I13/I13-1158.pdf>.
- Tsoumakas G, Katakis I.** 2007. Multi-label classification: an overview. *International Journal of Data Warehousing and Mining* 3:1–13 DOI 10.4018/jdwm.2007070101.
- Vivaldi J, Rodríguez H.** 2010. Using Wikipedia for term extraction in the biomedical domain: first experiences. *Procesamiento del Lenguaje Natural* 45:251–254.
- Wang P, Hu J, Zeng H-J, Chen Z.** 2008. Using Wikipedia knowledge to improve text classification. *Knowledge and Information Systems* 19(3):265–281 DOI 10.1007/s10115-008-0152-4.
- Wang P, Hu J, Zeng H-J, Chen L, Chen Z.** 2007. Improving text classification by using encyclopedia knowledge. In: *Seventh IEEE international conference on data mining (ICDM 2007)*. 332–341.

- Yang Y.** 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval* 1(1):69–90 DOI 10.1023/A:1009982220290.
- Yetisgen-Yildiz M, Pratt W.** 2005. The effect of feature representation on MEDLINE document classification. *AMIA Annual Symposium Proceedings* 849–853.
- Zhang Z, Phan X-H, Horiguchi S.** 2008. An efficient feature selection using hidden topic in text categorization. In: *Advanced information networking and applications-workshops*.
- Zheng B, McLean DC, Lu X.** 2006. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC Bioinformatics* 7:58 DOI 10.1186/1471-2105-7-58.
- Zhou X, Zhang X, Hu X.** 2006. MaxMatcher: biological concept extraction using approximate dictionary lookup. In: *PRICAI 2006: trends in artificial intelligence, Lecture notes in computer science*, vol. 4099. Springer, 1145–1149. Available at http://link.springer.com/chapter/10.1007%2F978-3-540-36668-3_150#page-1.
- Zhou X, Zhang X, Hu X.** 2008a. Semantic smoothing for Bayesian text classification with small training data. In: *Proceedings of the international conference on data mining*. 289–300. Available at <http://pubs.siam.org/doi/abs/10.1137/1.9781611972788.2>.
- Zhou X, Zhang X, Hu X.** 2008b. Semantic smoothing for Bayesian text classification with small training data. In: *Proceedings of the SIAM international conference on data mining, SDM 2008*. 289–300. Available at <http://pubs.siam.org/doi/abs/10.1137/1.9781611972788.26>.

Wikipedia-based cross-language text classification

A Bag of Concepts Approach
for Biomedical Document
Classification Using Wikipedia
Knowledge: Spanish-English
Cross-language Case Study

