UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

Diseño e Implementación de un Sistema de Clasificación Afectiva de Opinión y Relevancia

PROYECTO FIN DE CARRERA
INGENIERÍA DE TELECOMUNICACIÓN

Autora: Blanca Galego Pascual
Tutor: Julio Villena Román
Junio 2010

Titulo: Diseño e Implementación de un Sistema de Clasificación Afectiva de Opinión y Relevancia

Autora: Blanca Galego Pascual

Tutor: Julio Villena Román

EL TRIBUNAL

Presidente:

Pedro Muñoz Merino

Secretario:

Isaac Seoane Pujol

Vocal:

Jessica Rivero Espinosa

Realizado el acto de defensa del Proyecto Fin de Carrera el día 10 de Junio de 2010 en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de:

Fdo: Presidente Fdo: Secretario Fdo: Vocal

Agradecimientos

A Julio, mi tutor, por todo el tiempo y esfuerzo dedicado, a mi familia, por el apoyo constante, a todos los compañeros de carrera, por todos los buenos momentos compartidos y por supuesto, a mis amigos, por las escapadas de fin de semana, los tan necesarios viajes, las llamadas y los correos y simplemente por hacer estos largos años mucho más llevaderos.

Resumen

Dentro de una sociedad en la que cada vez hay más información y el acceso a la misma está cada vez más extendido, la capacidad para trabajar con ella se ha vuelto una tarea imprescindible. Parte del hecho de poder trabajar de forma eficiente con esta información pasa por ser capaces de organizarla o clasificarla adecuadamente, en especial teniendo en cuenta el incremento en volumen que ha experimentado en formato electrónico.

Esta reciente disponibilidad de información digital ha permitido que la clasificación automática sea algo abordable y que por lo tanto se hayan podido desarrollar diversas técnicas para su implementación. La clasificación automática, sobre todo de cara a detectar la relevancia de la información, ha sido un área de investigación muy estudiada durante los últimos años aunque es la clasificación afectiva de textos la que ha cobrado más importancia recientemente.

El objetivo de este Proyecto de Fin de Carrera es investigar y desarrollar un prototipo de clasificación de relevancia y de clasificación de opinión, evaluado sobre el corpus de textos obtenido del foro NTCIR-6.

El análisis de opinión consiste en determinar si un documento dado expresa o no una opinión mientras que el análisis de relevancia se centra en determinar si un documento es relevante a un tema predeterminado, y por tanto dando una medida de fiabilidad de la opinión que expresa.

NTCIR son una serie de talleres de evaluación diseñados para mejorar la investigación sobre tecnologías de acceso a la información. En ellos se plantean distintas tareas sobre diversas áreas dentro del acceso a la información y proporcionan los corpus necesarios para llevarlas a cabo.

Este proyecto se basa en la tarea de análisis multilingüe de opinión (MOAT) de NTCIR-6. Esta tarea proporciona noticias en inglés, japonés, chino tradicional y chino simplificado, aunque el proyecto se centrará únicamente en el análisis del corpus en inglés.

A partir de estos datos se diseñará e implementará un sistema de clasificación afectiva de opinión y relevancia, pudiendo comprobar las dificultades asociadas a cada uno de los análisis según la naturaleza de los mismos.

Índice de Contenidos

Índice de Figuras

Índice de Tablas

1 Introducción

1.1 Motivación

Dentro de una sociedad en la que cada vez hay más información y el acceso a la misma está cada vez más extendido, la capacidad para trabajar con ella se ha vuelto una tarea imprescindible. Parte del hecho de poder trabajar de forma eficiente con esta información pasa por ser capaces de organizarla o clasificarla adecuadamente, en especial teniendo en cuenta el incremento en volumen que ha experimentado en formato electrónico.

Los artículos de prensa se han convertido en una parte muy significativa de la información disponible en Internet, siendo la fuente de información más frecuentemente actualizada y a la que acceden un rango más diverso de usuarios. El volumen de información que genera la prensa es enorme, lo que supone una dificultad añadida a la hora de explorarla y analizarla.

Todos estos factores han provocado el auge de dos áreas de trabajo como son la Recuperación de Información (IR, *Information Retrieval*) y la Clasificación de Texto (TC, *Text Categorization*). La IR se centra en la localización y acceso a los recursos relevantes para un tema en concreto mientras que la TC clasifica la información obtenida según su contenido, de forma que sea más sencillo trabajar con ella.

Es precisamente la reciente disponibilidad de información digital lo que ha permitido que la clasificación automática sea algo abordable y que por lo tanto se hayan podido desarrollar diversas técnicas para su implementación. La clasificación automática, sobre todo de cara a detectar la relevancia de la información, ha sido un área de investigación muy estudiada durante los últimos años. Sin embargo, no es el caso de la clasificación afectiva de textos, que ha cobrado más importancia recientemente.

La clasificación afectiva ha sido considerada un área dentro de la clasificación de documentos durante mucho tiempo. Mientras que el problema de clasificación clásica consiste en decidir a qué tema pertenece un documento dentro de un conjunto de temáticas posibles, en la clasificación afectiva lo que se quiere determinar es la subjetividad, si ese texto presenta opiniones, ya sean positivas o negativas, sobre el tema en cuestión. Esto esta muy unido al análisis de relevancia, un caso particular de clasificación clásica que será de vital importancia para escoger los eliminar ruido dentro de la clasificación afectiva.

El análisis de relevancia aplicado a documentos consiste en determinar el grado de relevancia que tiene un documento respecto a un tema dado, o en otras palabras, cómo de significativo ese para dicho tema. Este análisis de relevancia, al ser combinado con el análisis de opinión dará una idea de lo fiable que es la opinión expresada por un documento para cierto tema.

El auge del análisis de opinión viene en parte avivado por el interés tanto de empresas como de administraciones públicas de analizar, filtrar o detectar las opiniones expresadas por sus clientes o por los ciudadanos, algo que no se limita a los entornos creados para ello, sino que cada vez más aparecen en entornos independientes como pueden ser los blogs o los foros.

1.2 Objetivos

El objetivo de este Proyecto de Fin de Carrera es investigar y desarrollar un prototipo de clasificación de relevancia y de clasificación afectiva. En él, dada una consulta sobre un determinado tema, el sistema busca en el conjunto de documentos indexados y devuelve la lista de frases de aquellos documentos que, por un lado, proporcionan información relevante sobre dicha consulta, y, por otro, el grado de opinión (o subjetividad) que conllevan.

Por ejemplo, ante la consulta "*series cancellation*", el sistema debería devolver como relevantes por ejemplo las frases "*The series cancellation was announced yesterday*" y "*Cancelling the series is the network biggest mistake to date*", indicando que la segunda de ellas expresa una opinión del redactor de la noticia.

Este objetivo coincide con el de la tarea MOAT (Multilingual Opinion Analysis) del foro NTCIR-6, así que para entrenar y evaluar el sistema, se va a hacer uso del corpus de noticias y la evaluación allí proporcionado.

NTCIR son una serie de talleres de evaluación diseñados para mejorar la investigación sobre tecnologías de acceso a la información. Para ello, plantean distintas tareas sobre diversas áreas dentro del acceso a la información y proporcionan los corpus necesarios para llevarlas a cabo. Este proyecto se basa en la tarea de análisis multilingüe de opinión de NTCIR-6. Esta tarea proporciona noticias en inglés, japonés, chino tradicional y chino simplificado, pero el proyecto se centrará únicamente en el análisis del corpus en inglés. El corpus esta compuesto por artículos de periódicos agrupados por tema sobre el que tratan, de forma que además del análisis afectivo intrínseco asociado a cada texto, también se pueda analizar la relevancia de dichos artículos para cierta consulta dada.

Los objetivos fundamentales de este proyecto son los siguientes:

- Ser capaces de extraer la información relevante de los textos y obtener una representación estructurada de los mismos de forma que se facilite su procesado y análisis.
- Lograr que el conocimiento adquirido a partir de documentos ya categorizados permita desarrollar un sistema de clasificación válido y eficiente ante una consulta de un usuario.
- Entender y analizar los problemas derivados del análisis de la carga afectiva de un texto a la hora de implementar un clasificador automático de la misma.

1.3 Estructura de la Memoria

La memoria de este proyecto se divide en cinco capítulos principales en los que se tratarán distintos aspectos del mismo y cuyo contenido se detalla brevemente a continuación:

- Capítulo 1: Introducción

  En este capítulo se hablará de los fundamentos y motivaciones sobre las que se apoyan los clasificadores automáticos de textos, se enumerarán los objetivos perseguidos en este Proyecto Fin de Carrera y se describirá la estructura de la memoria realizada.

- Capítulo 2: Estado del Arte

  Se dará una visión general de los sistemas de clasificación automáticos de texto, sus características generales y algunas de las técnicas empleadas; se hace especial hincapié en la clasificación de opinión tanto por su relevancia para el proyecto como por su importancia dentro de la clasificación automática. Asimismo se hablará de NTCIR, su origen, algunas de sus características e información relevante de cara a la resolución del escenario planteado.

- Capítulo 3: Diseño del Sistema

  En el tercer capítulo se describirá con detalle todas las decisiones de diseño así como los pasos seguidos para la implementación del sistema desarrollado en este proyecto. Se diferenciarán dos líneas principales, una relativa al análisis de relevancia y otra relativa al análisis de opinión.

- Capítulo 4: Implementación del Sistema

  Una vez visto el diseño el sistema, se pasará a ver en detalle cómo se ha implementado, por lo general a través de diagramas de flujo de forma que se puedan ver fácilmente los distintos módulos usados para implementar el sistema, su funcionalidad básica y las dificultades y soluciones asociadas encontradas durante este proceso.

- Capítulo 5: Evaluación

  En este capitulo se detallarán los escenarios y experimentos llevados a cabo para evaluar el rendimiento del sistema, incluyendo resúmenes comprensibles de los resultados obtenidos durante dichas evaluaciones y análisis de los mismos.

- Capítulo 6: Conclusiones y Trabajos Futuros

  En el último capítulo se explicarán las conclusiones a las que se ha llegado durante la realización del proyecto. Por otro lado, se presentarán distintas líneas de investigación en las que seguir trabajando ya sea de cara a la mejora del sistema o simplemente para aportar nuevas ideas dentro del ámbito de la clasificación de textos.

- Anexo A: Temas de NTCIR-6

  Este anexo contiene el fichero de temas proporcionado junto al corpus de noticias de NTCIR-6. En el aparece toda la información de cada uno de los temas de la que dispone el sistema y que utilizará en la clasificación de relevancia.

- Anexo B: Lista de Palabras de Parada

  En este anexo se incluye una versión reducida de la lista de palabras de parada empleada durante el procesado de texto. La lista original utilizada contiene cada palabra 3 veces: en mayúsculas, en minúsculas, y con la primera letra en mayúsculas. En el anexo se ha dejado solamente la versión en minúsculas.

- Bibliografía: Referencias

  Se han incluido todas las referencias utilizadas en este proyecto por orden de aparición en esta memoria.

2 Estado del Arte

2.1 Introducción

Durante los últimos veinte años, la rápida expansión que ha experimentado globalmente Internet ha hecho posible la reducción de la complejidad del acceso a todo tipo de información. Cada vez hay más fuentes de contenidos y un volumen de datos disponibles muchísimo mayor, generando una gran cantidad de información a explorar y analizar.

Esta nueva disponibilidad y volumen de la información puede llegar a ser un inconveniente, de ahí la necesidad de nuevos métodos mediante los cuales se pueda filtrar y estructurar dicha información; como consecuencia, la organización de la información de forma automática ha pasado a ser una tarea de vital importancia y la gestión eficiente se ha convertido en algo imprescindible. Todo esto ha provocado que cada vez sea más necesario disponer de herramientas que permitan automatizar esta clasificación.

Las tareas que llevan a cabo la gestión automática de documentos basándose en su contenido se conocen comúnmente como Recuperación de la Información, en inglés *Information Retrieval* (IR). A través de IR se determina qué documentos de una colección dada son más relevantes para la consulta que realiza un usuario, lo que permite eliminar información inútil. Por otro lado, las tareas mediante las que se asignan categorías predeterminadas a documentos escritos se conoce como Clasificación de Texto o *Text Categorization* (TC). La TC permite mejorar las prestaciones de sistemas IR ya que es una forma de filtrar la información relevante a un tema [1].

La clasificación automática de texto se puede definir como un conjunto de algoritmos, técnicas y sistemas capaces de asignar un documento a una o varias categorías o grupos de documentos asociados según su afinidad temática. Para ello se emplean técnicas de Aprendizaje Automático (ML, *Machine Learning*) y de Procesamiento del Lenguaje Natural (NLP, *Natural Language Processing*).

El NLP estudia los problemas inherentes al procesamiento y manipulación de lenguajes naturales mediante ordenadores. Por lenguaje natural se entiende el lenguaje hablado o escrito por humanos para propósitos generales de comunicación. Mediante este estudio se pretende desarrollar herramientas y técnicas que permitan que los ordenadores puedan entender y manipular el lenguaje al igual que lo entienden y utilizan los humanos.

Se basa en numerosas disciplinas tales como las ciencias de la información, lingüística, matemáticas, inteligencia artificial, psicología, etc. De la misma forma, las aplicaciones para las que resultará de gran utilidad son muchas y variadas: traducción, procesamiento y resumen de textos escritos, interfaces de usuario, reconocimiento de voz, etc.

2.2 Clasificación Automática de Textos

Como ya se ha comentado, gracias al auge de la información en formato digital, el ámbito del Aprendizaje Automático aplicado al NPL se ha desarrollado mucho en los últimos años, dando lugar una gran cantidad de modelos y técnicas de aprendizaje.

Las distintas técnicas de aprendizaje se aplican en la fase de entrenamiento del clasificador, en la que se utilizarán un conjunto de documentos previamente clasificados dentro de una categoría. Cualquiera de los algoritmos que se van a ver más adelante necesitan una presentación ordenada de los documentos. Una de las más utilizadas es la representación basada en el Modelo de Espacio Vectorial, donde cada documento se transforma en un vector de palabras a las que se le asigna un peso en función de su importancia dentro del documento.

Una vez que el clasificador haya sido entrenado, se clasificarán un grupo de textos distintos a los usados para el entrenamiento de los cuales también se tiene la asignación de categorías; de esa forma se podrá medir la efectividad del mismo.

Algunas de las aplicaciones de la clasificación automática de textos son la indexación automática de textos, filtrado de textos, clasificación de páginas Web, filtrado de correos electrónicos no deseados (*spam*) o clasificación de noticias.

A continuación se verán algunas de las principales características del Modelo de Espacio Vectorial, seguido de las principales clasificaciones para los distintos tipos de clasificadores que puede haber y algunas de las técnicas más significativas empleadas para diseñarlos.

2.2.1 Modelo de Espacio Vectorial

Como ya se ha dicho, el Modelo de Espacio Vectorial, es una de las principales técnicas para representar documentos de forma estructurada.

Este modelo fue presentado por Salton en 1975 [2] y usado por primera vez en el sistema de recuperación de información SMART. Algunas de sus principales ventajas sobre modelos anteriores (principalmente el modelo booleano) son las siguientes:

1. Es un modelo simple, basado en álgebra lineal.
2. Los pesos de los términos no tienen que ser binarios.
3. Permite una medida continua de similitud entre consultas y documentos.
4. Permite ordenar los documentos según su posible relevancia.
5. Permite recuperar documentos que solo tienen correspondencia parcial a una consulta.

Por otro lado, y teniendo en cuenta el volumen de datos con los que se va a trabajar, desde el punto de vista computacional es un algoritmo bastante lento y tiene como principales desventajas tener que ser calculado por completo cuando se añade un nuevo término y necesitar acceso a todos los términos, no solo a los de la consulta.

Alguna de sus otras limitaciones son las siguientes:

1. Los documentos muy largos dificultan el cálculo de la similitud, ya que implican realizar productos escalares con vectores de alta dimensionalidad.
2. Falsos negativos: documentos con un contenido similar pero con distinto vocabulario dan un producto escalar muy bajo, aunque esto no es tanto una limitación del modelo como de los sistemas IR basados en palabras claves.
3. Falsos positivos, por lo general derivados de las fases de parsing (por lo que tampoco exactamente limitación del modelo). Ejemplos: *Marching: March + ing, Therapist: the + rapist o the + rap + ist*.

Algunas de las medidas que se pueden tomar para mitigar estos efectos son:

- identificar palabras claves representativas para cada documento.
- eliminar todas las palabras de parada y términos muy comunes.
- lematizar los términos.
- limitar el espacio vectorial a sustantivos y unos pocos adjetivos y verbos descriptivos.
- usar técnicas de mapeo por tema.
- calcular sub-vectores en los documentos largos.
- ignorar documentos por debajo de un umbral del valor del coseno.

Un problema importante de los modelos de vectores de términos es que asumen que los términos son independientes entre sí, cuando normalmente este no va a ser el caso. Los términos pueden estar relacionados por:

1. Polisemia: términos que se usan para expresar cosas distintas en diferentes contextos. "*Driving a car*" y "*driving results*", es un ejemplo. Provoca que documentos que no tienen nada que ver tengan un valor alto de similitud entre sí. Afecta a la precisión.
2. Sinonimia: términos diferentes con los que se expresa lo mismo. Hacen que dos frases muy parecidas semánticamente tengan una similitud baja. Como ejemplo se puede tomar "*aircraft*", "*airplane*" y "*plane*". Afecta a la cobertura.

Todos estos puntos tendrán que tenerse en cuenta a la hora de elegir el modelo con el representarán los documentos ya que según los datos con los que se trabaje, y el tipo de clasificación que se esté haciendo, afectarán más o menos al sistema. Otros detalles de diseño e implementación se verán con más profundidad en apartados posteriores (ver apartado 3.2.1).

2.2.2 Tipos de Clasificadores

Los clasificadores pueden ser dividirse en distintos grupos según una serie de criterios relativos a las características de los mismos. Las principales clasificaciones son las siguientes:

- Clasificación supervisada o no supervisada, según existan o no categorías definidas a priori.
- Clasificación paramétrica o no paramétrica, según se usen o no parámetros estadísticos en el cálculo del clasificador.
- Clasificación simple o múltiple, según el número de categorías en las que se pueda clasificar un documento.
- Clasificación centrada en la categoría o centrada en el documento, según el enfoque que se de a la clasificación a raíz de la información inicial disponible.

A continuación se verá en detalle en qué consiste cada una de ellas.

2.2.2.1 Clasificación Supervisada y no Supervisada

Esta clasificación se basa en si existen o no categorías conceptuales definidas a priori en el sistema.

Cuando estas categorías ya han sido creadas, se está ante una clasificación supervisada en la que se irá asignando cada uno de los documentos a las categorías. Como su propio nombre indica requiere elaboración manual o intelectual de las categorías y forzará la presencia de una etapa de entrenamiento en el clasificador.

Estos clasificadores buscan elaborar un patrón representativo de cada una de las categorías entrenadas, de forma que al aplicar una función determinada se pueda estimar la similitud entre el documento que se quiere clasificar y los distintos patrones definidos. Estos patrones se definen a partir de un conjunto de documentos que han sido previamente clasificados, el denominado conjunto de entrenamiento que se usará en la etapa homónima.

Por otro lado, cuando no existen categorías creadas a priori, se estará ante la clasificación no supervisada, en la que los documentos serán clasificados automáticamente teniendo en cuenta únicamente su contenido.

2.2.2.2 Clasificación Paramétrica y no Paramétrica

Este sistema de clasificación se basa en el uso o no de parámetros estadísticos a la hora de calcular el clasificador.

En la clasificación paramétrica, se supone para los datos un modelo estadístico conocido, y se usa la fase de entrenamiento para calcular los parámetros estadísticos del mismo. El conjunto de pruebas se utiliza para determinar la capacidad de generalización del clasificador [3].

La clasificación no paramétrica supondrá un modelo general que se pueda aproximar a cualquier distribución. Se subdivide en dos categorías, una primera categoría en la que lo que se busca es identificar patrones para cada una de las categorías, generalmente mediante vectores de términos con pesos (como por ejemplo el algoritmo de Rocchio [4]), y una segunda categoría en la que se clasifica a partir de la similitud de los documentos con distintos patrones [5] (como por ejemplo el algoritmo kNN).

2.2.2.3 Clasificación Simple y Múltiple

Esta clasificación utiliza el número de categorías bajo las cuales se puede clasificar un documento para realizar una clasificación. Habrá dos categorías: clasificación simple y clasificación múltiple.

La clasificación simple es aquella en la que cada documento del sistema puede pertenecer a una única categoría, o en otras palabras, las categorías son mutuamente exclusivas entre sí. El ejemplo más claro de este caso es la clasificación binaria, en la cual un documento si no pertenece a una categoría, pertenece a su complementaria. En estas clasificaciones, el centroide más similar indica a qué categoría hay que asignar el documento.

La clasificación múltiple es aquella en la cual un documento puede recibir un número variables de categorías comprendido entre cero y el número total de categorías en el sistema. En estas clasificaciones, es un umbral de similitud el que indica a qué categorías hay que asignar el documento.

Una clasificación múltiple en la que hay N categorías, se puede transformar con facilidad en N clasificaciones simples, siempre y cuando se cumpla que las categorías sean estocásticamente independientes entre sí. Esto se traduce en que un sistema diseñado para clasificación simple puede ser utilizado para realizar una clasificación múltiple mientras que el caso contrario no será posible [1].

2.2.2.4 Clasificación Centrada en la Categoría y Centrada en el Documento

Una vez haya sido construido el clasificador, su utilización puede recibir dos enfoques distintos, según los datos de los que se disponga inicialmente.

El primer enfoque es la clasificación centrada en la categoría (CPC, *Category-Pivoted Classification*) en la que se parte de un documento, y se intentará encontrar todas las categorías a las que éste pertenece.

El segundo enfoque es la clasificación centrada en el documento (DPC, *Document-Pivoted Classification*), en la que se parte de la categoría, y se intentará encontrar todos los documentos que se clasifican dentro de ella.

Como es de suponer, no en todos los escenarios se tendrá la misma información disponible, por lo que cada uno de los enfoques será más apropiado para distintas situaciones. Un ejemplo en el que CPC es más recomendable es cuando el número de categorías dentro de una clasificación ya realizada cambia; en ese escenario, lo que se querrá es averiguar qué documentos pertenecen a la nueva categoría. Por otro lado, DPC va a ser el enfoque a elegir en escenario en el que se dispone de los documentos uno a uno y durante un período considerable.

2.2.3 Técnicas de Clasificación

Como ya se ha mencionado, existen distintas técnicas que se pueden utilizar para diseñar un clasificador, cada una de ellas basadas en distintas principios y teorías. La elección de la técnica a utilizar vendrá condicionada por las características tanto del entorno como de los datos con los que se va a trabajar, los recursos disponibles y el enfoque que se le vaya a dar al sistema.

A continuación se habla en más detalle de los seis principales grupos de técnicas.

2.2.3.1 Algoritmos Probabilísticos

Los algoritmos probabilísticos son una técnica de clasificación supervisada que se basan en la teoría probabilística, concretamente en el teorema de Bayes, que permite estimar la distribución de probabilidad de un evento aleatorio A dado B a partir de la distribución de probabilidad condicional de B dado A y de la distribución de probabilidad marginal de A. Expresado matemáticamente:

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{P(B)}$$ **ECUACIÓN 1**

Al problema de la clasificación se aplica una versión simplificada de este teorema llamada *Naïve Bayes*, en la que se realizan suposiciones importantes de independencia probabilística. Dentro de este escenario, lo que se estima es la probabilidad de que un documento pertenezca a una categoría.

Que un documento pertenezca o no a una categoría dependerá de las características del mismo, que en el caso de que se esté aplicando el modelo vectorial como representación del documento, serán los términos que contengan. En la fase de entrenamiento se podrá obtener la probabilidad de que los distintos términos aparezcan en los documentos de cada categoría a partir de su frecuencia de aparición en los mismos.

Si se particulariza el teorema de Bayes a un escenario en el que C_i es la categoría a la que pertenecería el documento y $\overrightarrow{d_j} = \{w_{1j},...,w_{|T|j}\}$ es el documento representado a través de un vector con los pesos de los términos que incluye, entonces la probabilidad de que un documento pertenezca a la categoría C_i es la siguiente [1]:

$$P(C_i \mid \overrightarrow{d_j}) = \frac{P(C_i) \cdot P(\overrightarrow{d_j} \mid C_i)}{P(\overrightarrow{d_j})}$$ **ECUACIÓN 2**

En esta expresión las dos probabilidades marginales representan la probabilidad de que un documento $\overrightarrow{d_j}$ elegido al azar, pertenezca a la categoría C_i. La probabilidad condicional del segundo término es la que representan más problemas de cálculo y para la cual habrá que hacer las suposiciones de independencia ya mencionadas. Concretamente se supondrá que dos coordenadas cualesquiera del vector son variables aleatorias estadísticamente independientes, y por lo tanto se podrá expresar matemáticamente de la siguiente forma:

$$P(\overrightarrow{d_j} \mid C_i) = \prod_{k=1}^{|T|} P(w_{kj} \mid C_i)$$ **ECUACIÓN 3**

Esta aproximación se realiza por la tendencia del número de documentos a ser muy elevado. De esto se puede deducir que el cálculo de la prioridad va a ser un punto crítico dentro de este tipo de algoritmos. Uno de los problemas que presentan estos algoritmos son las estimaciones incorrectas de las probabilidades cuando el conjunto de entrenamiento no es lo suficientemente grande. Para intentar mitigar estos errores y evitar distorsiones se usan técnicas de suavizado [6].

A partir de las probabilidades calculadas y siendo la composición de términos o características de un documento nuevo conocida, se puede estimar la probabilidad que tiene de pertenecer a cada una de las categorías. Este tipo de algoritmos son fáciles de implementar, rápidos y eficaces, de ahí su popularidad [6].

2.2.3.2 Algoritmo de Rocchio

El algoritmo de Rocchio [4] se basa en las mismas ideas utilizadas en la realimentación por relevancia. Tras una iteración inicial, se decide qué documentos son o no relevantes y se construye una nueva consulta a partir de estos documentos señalados. El algoritmo proporciona un

sistema para construir el vector de la nueva consulta, recalculando los pesos de sus términos y aplicando coeficientes de pesos distintos a la consulta inicial a los documentos marcados como relevantes y a los documentos marcados como no relevantes.

En resumen, el algoritmo ayuda a construir patrones para cada una de las categorías de documentos. En la fase de entrenamiento, al tener ya documentos etiquetados (es por tanto clasificación supervisada), se puede aplicar el modelo vectorial para construir vectores asociados a cada una de las categorías usando como ejemplos positivos los documentos etiquetados en esa categoría en la fase de entrenamiento, y como ejemplos negativos los etiquetados en el resto de categorías.

La expresión utilizada para construir el vector de la categoría C_i a partir de un conjunto de documentos de entrenamiento T es la siguiente:

$$w_{ki} = \beta \cdot \sum_{d_j \in POS_i} \frac{w_{kj}}{|POS_i|} - \gamma \cdot \sum_{d_j \in NEG_i} \frac{w_{kj}}{|NEG_i|}$$

ECUACIÓN 4

En donde los significados de los distintos elementos de la misma son los siguientes:

- $C_i = \{w_{1i}, w_{2i}, ..., w_{ri}\}$ es la representación vectorial de la categoría.
- w_{ki} es el peso del término t_k en el documento d_j.
- $POS_i = \{d_j \in T_r \mid \phi(d_j, c_i) = True\}$
- $NEG_i = \{d_j \in T_r \mid \phi(d_j, c_i) = False\}$
- β y γ son los parámetros que se utilizan para ajustar la importancia de los ejemplos positivos y negativos.

La representación vectorial o patrón de cada una de las categorías se puede considerar el centroide representante de las mismas. El clasificador indicará las distancias entre el documento que se quiere clasificar y los centroides de las categorías, y será clasificado en la categoría a la que tenga mayor similitud (es decir, menor distancia).

2.2.3.3 Algoritmos del Vecino más Próximo y Variantes

El algoritmo del vecino más próximo (*Nearest Neighbor*, NN) es el principal representante de las técnicas basadas en ejemplares o instancias. Este tipo de técnicas utilizan ejemplos ya clasificados para asignar los nuevos elementos a las categorías correspondientes en función de la distancia a las instancias seleccionadas. Las instancias seleccionadas no tienen por qué ser necesariamente datos clasificados correctamente, ya que dependiendo del método utilizado se pueden tomar los más representativos, los clasificados incorrectamente en una clasificación inicial, etc.

El algoritmo de los k vecinos más cercanos (*k-Nearest Neighbors*, kNN) es la generalización de NN y como se ha dicho, el principal representante de este tipo de algoritmos gracias a su sencillez conceptual y a su eficacia.

Este algoritmo lo que hace es tomar los k documentos más parecidos al documento a clasificar, independientemente de su categoría, y la categoría con más presencia dentro de esos k elementos más próximos será a la que se asigne al documento que se está clasificando.

La forma más habitual de determinar la proximidad es a través de la distancia Euclídea:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{|C|} (x_{ik} - x_{jk})^2}$$

ECUACIÓN 5

donde $|C|$ son las categorías del sistema, y cada documento está escrito en función de $|T|$ términos de la siguiente forma: $x_j = \{w_{1j}, \ldots, w_{|T|j}\}$.

En la Figura 1 se puede ver un ejemplo simplificado: en la parte de la izquierda se ha representado el conjunto de datos que se usará para clasificar, representado con círculos si pertenece a la categoría 1, y con triángulos si pertenece a la categoría 2. En la derecha se representan los dos documentos que van a ser clasificados, A y B, y junto a ellos, se señalan las tres muestras más próximas.

En el caso del documento A, al tener dentro de sus tres muestras más cercanas a dos de la categoría 1, el documento será asignado a esa categoría. Siguiendo un razonamiento análogo B será asignado a la categoría 2.

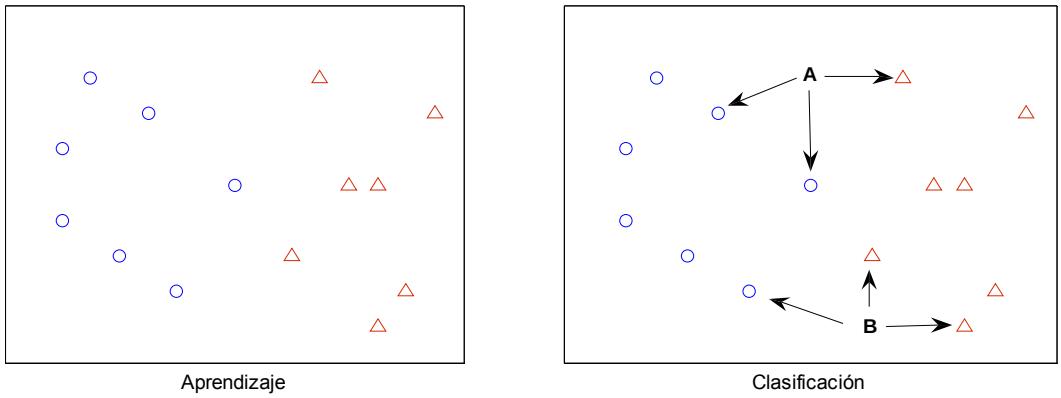

Figura 1: Algoritmo kNN.

Al no estar haciendo ninguna suposición sobre la distribución de las variables, se trata de un método no paramétrico. KNN es especialmente eficaz en escenarios en los que los documentos son heterogéneos y difusos y el número de categorías posibles es alto.

2.2.3.4 Árboles de Clasificación

Los árboles de clasificación son uno de los métodos de aprendizaje inductivo supervisado no paramétrico más utilizado [7]. Es una forma de representación del conocimiento muy sencilla, de ahí su popularidad en numerosos ámbitos a pesar de carecer de la expresividad de otros métodos tales como las redes semánticas o la lógica de primer orden.

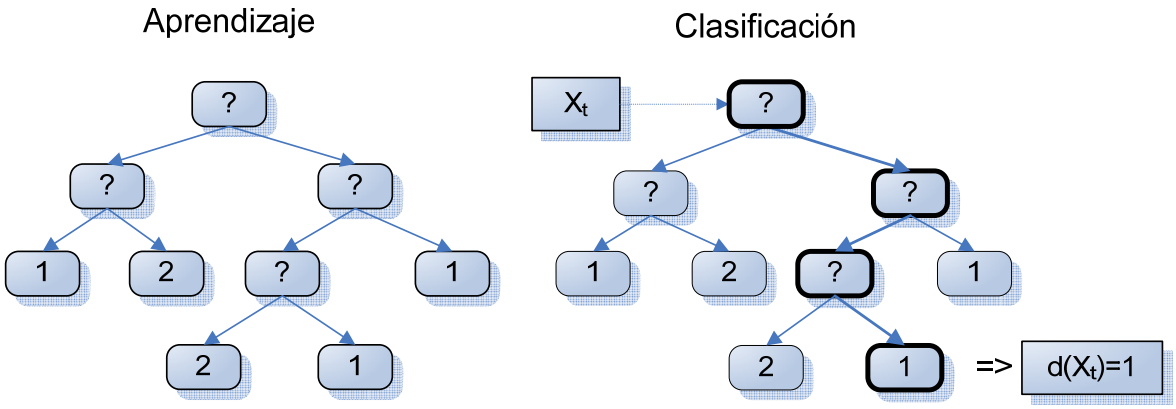

Figura 2: Árboles de clasificación.

Un árbol es la estructura resultante de la partición recursiva del espacio de representación a partir del conjunto de prototipos (los documentos). Esta partición recursiva se traduce en una organización jerárquica que se puede representar mediante una estructura de árbol. Dentro de estos árboles, cada nodo interior contiene una pregunta sobre un atributo concreto, que dará lugar a un nuevo nodo por cada posible respuesta. Las hojas del árbol serán las decisiones.

Para clasificar patrones, se realizan una serie de preguntas relativas a sus atributos, empezando por el nodo raíz y siguiendo el camino marcado por las respuestas a las preguntas de los nodos intermedios hasta llegar a la hoja. La etiqueta asociada a esa hoja será el valor que se le asigne al patrón.

La metodología a seguir se resume en los dos pasos que se pueden ver en la figura:

1. Aprendizaje, en el cual se construye el árbol a partir de un conjunto de prototipos (lo que sería el conjunto de documentos de entrenamiento). Es la fase más importante ya que de ella dependerá el resultado de la clasificación.
2. Clasificación, en la cual se etiqueta un patrón independiente del aprendizaje. Los pasos para clasificarlo son los ya descritos: se responden todas las preguntas desde el nodo raíz hasta la hoja con la etiqueta siguiendo el camino marcado por las respuestas a los nodos intermedios.

Como se ha mencionado, la construcción de los árboles es la fase más importante, por lo que se han desarrollado diversos algoritmos para mejorar la automatización de este proceso. Uno de los más famosos es ID3, desarrollado por Quinlan [8] y basado en el concepto de la ganancia de información. Este método presenta problemas en la clasificación de atributos continuos, por lo que el mismo Quinlan propuso más adelante como extensión el algoritmo C4.5 [9] que a día de hoy sigue siendo de los más utilizados.

2.2.3.5 Redes Neuronales

Las redes neuronales son un modelo que intenta simular la estructura y aspectos funcionales de las redes de neuronas biológicas. Para ello combina neuronas artificiales de forma que a través de su interconexión colaboren para obtener un resultado final. Una de sus aplicaciones es el reconocimiento de patrones, de ahí su utilidad en clasificación automática [6].

Como se puede ver en la imagen situada más a la izquierda de la Figura 3, una neurona es un dispositivo formado por una serie de entradas y una única salida [10]. Cada neurona acepta como entrada las salidas procedentes de otras neuronas, de forma que la entrada efectiva a la neurona sea la suma ponderada de las entradas. Cada una de las neuronas tiene un estado de activación, un valor comprendido entre 0 y 1 que indica si la neurona está o no activada. Cualquier valor distinto de cero implica que la neurona está activada y la salida de una neurona es el estado de activación, por lo tanto, siempre que una neurona esté activa, estará generando una salida.

La tarea que realiza cada neurona es sencilla: recibe información de entrada de otras neuronas o del exterior y la utiliza para generar una señal de salida que propaga a otras unidades.

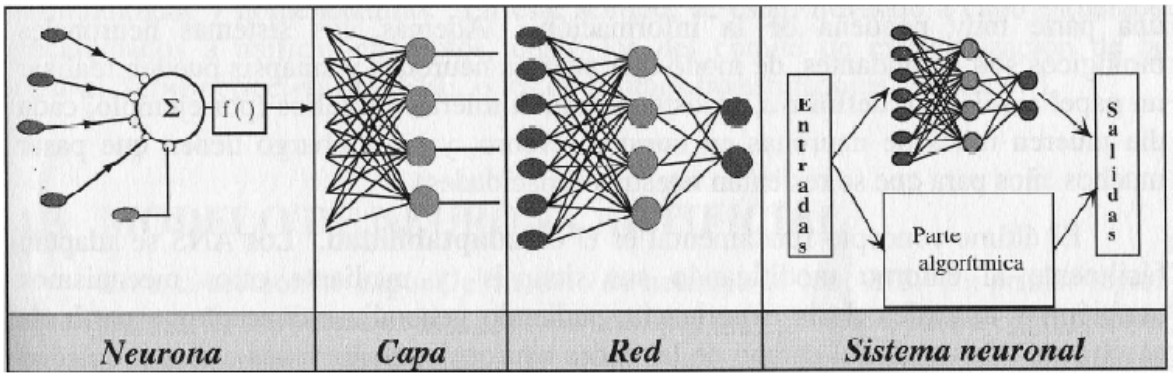

Figura 3: Descomposición de una red neuronal.

Las interconexiones entre neuronas pueden tener asignados pesos de forma que cada una de las entradas puede tener más o menos fuerza. Estos pesos son los que se podrán ajustar en fases de entrenamiento de cara a obtener la salida deseada.

Por lo general las neuronas se organizan en capas, que interconectadas entre sí formarán las redes neuronales. Hay tres tipos principales de capas: las compuestas por unidades de entrada, que reciben información del exterior, las compuestas por unidades de salida, que generan la salida del problema (en el caso de la clasificación, son las que mapean las categorías) y las unidades ocultas, que son todas las capas entre la entrada y la salida y que se llaman así por no ser visibles desde el exterior (ver tercera imagen de la Figura 3).

2.2.3.6 Máquinas de Vectores Soporte

Las máquinas de vectores soporte (*Support Vector Machines*, SVM) son un conjunto de algoritmos de clasificación y regresión basados en la minimización del riesgo estructural ("*Structural Risk Mnimization*", SRM), un principio originado en la teoría estadística del aprendizaje [11].

El objetivo en estos algoritmos es obtener el hiperplano óptimo que da lugar al mayor margen de separación entre las clases.

Como se puede ver en la Figura 4, puede haber varios clasificadores lineales, pero solo uno de ellos maximizará la distancia entre él y el punto más cercano de cada una de las clases [12].

En la figura los clasificadores son lineales, y sencillos de calcular, pero normalmente esto no será así.

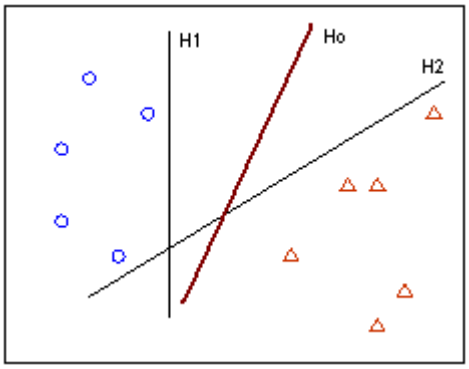

Figura 4: Hiperplano de separación óptimo.

La idea de este principio es encontrar una hipótesis h para la cual, a partir de una cota sobre el riesgo esperado, *R(h)* (tasa de error medio sobre el conjunto de test) se concluye que para un conjunto de entrenamiento dado, será necesario minimizar el riesgo empírico $R_{emp}(h)$ (tasa de error media sobre el conjunto de entrenamiento) y la dimensión VC (*Vapnik Chervonenkis*) del espacio de hipótesis [10].

El riesgo empírico se calcula mediante la siguiente expresión:

$$R_{emp}(h) = \frac{1}{2n}\sum_{i=1}^{n}\left|y_i - f(x_i, h)\right| \qquad \textbf{ECUACIÓN 6}$$

Donde los distintos parámetros representan lo siguiente:

- *n* es el tamaño del conjunto de entrenamiento
- $f(x_i, h)$ es la salida del clasificador para el vector de entrenamiento x_i.
- $y_i \in \{-1,1\}$ es la etiqueta correspondiente al vector x_i.

Por otro lado, el riesgo esperado se calcula como sigue:

$$R(h) = \int \frac{1}{2}\left|y - f(x,h)\right| \cdot dP(x,y) \qquad \textbf{ECUACIÓN 7}$$

En esta expresión *dP(x,y)* es desconocido. En 1995, Vapnik demostró que, con una probabilidad de 1-η con $0 \leq \eta \leq 1$ existe una expresión para la cual se puede conseguir una cota superior del riesgo esperado. La expresión es la siguiente:

$$R(h) \leq R_{emp}(h) + \sqrt{\frac{d\left(\ln\frac{2n}{d}+1\right) - \ln\frac{\eta}{4}}{n}}$$ **ECUACIÓN 8**

Donde *d* es un entero no negativo conocido como la dimensión *Vapnik Chervonenkis* (VC), y proporciona una medida de capacidad de los algoritmos de clasificación estadística. Se define como la cardinalidad de mayor conjunto de puntos que el algoritmo puede separar.

Las dos ideas fundamentales para construir un clasificador SVM son la transformación de la entrada en un espacio de alta dimensión y el cálculo del hiperplano separador óptimo. La transformación inicial se realiza a través de la elección de una función kernel adecuada y es una de las principales dificultades del método. Esta transformación se realiza porque al trabajar en un espacio de alta dimensión, las clases se podrán considerar con alta probabilidad linealmente separables, algo que hará el cálculo del hiperplano separador óptimo menos costoso computacionalmente hablando. El hiperplano viene determinado por unas pocas observaciones, los denominados vectores de soporte precisamente por ser las que determinan la forma del hiperplano.

La forma final de la regla de clasificación para el caso binario (clases +1 y -1) quedaría así:

$$f(x) = b + \sum_i \alpha_i \cdot K(x, x_i)$$ **ECUACIÓN 9**

Donde *b* y α_i son parámetros que el clasificador aprende en la fase de entrenamiento, y $K(x,x_i)$ es el valor de la función kernel para los puntos *x* y x_i. Si $f(x)$ es mayor que el umbral, entonces el valor estimado para el punto *x* será +1, y en el caso contrario, -1.

2.3 Clasificación Afectiva

Aunque inicialmente la clasificación afectiva de documentos (conocida como *opinion mining, sentiment classification* o *sentiment analysis*) se ha considerado una disciplina dentro de la clasificación de documentos, en los últimos años es un área que ha experimentado un creciente interés en sí misma.

La clasificación basada en opinión se centra en determinar si en los textos a analizar se expresan opiniones positivas o negativas. Si se identifican esos dos posibles resultados como dos categorías en las que clasificar un documento, se ve claramente por qué es considerada una tarea dentro de la clasificación de documentos. La dificultad añadida que tiene este tipo de clasificación radica en la subjetividad tanto de los documentos como del concepto a analizar, algo que obliga a plantear soluciones distintas a las que se utilizan en la clasificación de documentos clásica.

En la clasificación basada en opinión entran en juego no solamente los fenómenos léxicos, sintácticos y semánticos del lenguaje sino también los pragmáticos. Esto deriva en que las técnicas de procesamiento de lenguaje natural van a tener mucho más peso en este tipo de clasificación.

Este tipo de análisis se va a utilizar en multitud de contextos, desde análisis de blogs a técnicas de inteligencia de negocios. Estos son algunos ejemplos de aplicaciones concretas:

- Búsquedas en foros y blogs: como ámbitos cada vez más relevantes en los medios de comunicación, y más utilizados a la hora de expresar opiniones, resulta muy útil realizar búsquedas rápidas de las opiniones expresadas en estos medios. Un ejemplo muy claro de cómo un porcentaje muy importante de las opiniones se encuentran en este tipo de medios son los últimos lanzamientos tecnológicos, donde se tiene como ejemplos más recientes el anuncio de iPad de Apple y la activación en Gmail de Google Buzz.
- Política: en un área donde las opiniones tanto de los votantes como de los propios políticos son importantes; el estudio de la opinión se puede usar con facilidad o bien para dar a los votantes una visión más clara de qué opina cada político de los distintos temas o bien a los

distintos partidos para estudiar la opinión popular de cara a nuevas propuestas o temas de actualidad.

- Clasificación de correos electrónicos y distribución priorizada: cada vez un mayor número de los servicios de atención al cliente se gestionan a través de correo electrónico. La posibilidad de poder clasificar de forma automática estos correos, dando por ejemplo una mayor prioridad a los que expresen opiniones negativas de cara a darles prioridad para evitar un descontento mayor de los clientes.
- Inteligencia de negocio o inteligencia competitiva: como se ha visto en los anteriores puntos, es muy sencillo utilizar el análisis de opinión en aplicaciones de inteligencia, entre las cuales se incluye la inteligencia de negocio y mercado con el consiguiente interés empresarial que generan.

  Este tipo de análisis puede ser extremadamente útil de cara a evaluar la opinión de los usuarios tanto de servicios como de productos, aunque no está libre de dificultades derivadas del distinto formato en el que se pueden encontrar los documentos. Estas dificultades junto con la necesidad de filtrar la información para acceder a la que es realmente relevante son las mismas a las que se enfrentaría un clasificador manual, de ahí que encontrar soluciones robustas para este tipo de problemas pueda facilitar sustancialmente la labor en numerosos ámbitos.
- Aplicaciones como subcomponente tecnológico: existen numerosas aplicaciones donde se pueden obtener mejores resultados a partir de la utilización de sistemas de análisis afectivo y minería de opinión. Es en estos escenarios donde se entiende que este tipo de sistemas son subcomponentes tecnológicos de otras aplicaciones. Un ejemplo muy claro son las aplicaciones de recomendaciones: cuando un elemento reciba mucho feedback negativo dejará de ser recomendado.

Una de las aplicaciones más habituales y que se va a ver dentro del proyecto es la clasificación afectiva aplicada a la relevancia. Identificar qué documentos son relevantes para cierto tema es algo con una carga subjetiva importante y que juega un factor determinante en la calidad de los resultados que se obtendrán para otro tipo de clasificaciones.

En resumen, la utilización de aplicaciones de minería de opinión o clasificación afectiva se puede extender a todos los ámbitos en los que se desee mejorar la interacción entre los humanos y los ordenadores, algo que hoy en día con el auge de los servicios a través de Internet, está presente prácticamente en cualquier área.

2.3.1 Particularidades

A raíz de diversos análisis llevados a cabo en el campo de la clasificación automática de textos se han podido identificar características específicas del análisis afectivo. A continuación se verán algunas de estas características.

2.3.1.1 Presencia de Términos frente a Frecuencia

Como ya se ha visto en apartados anteriores, los documentos se describen a través de sus características más representativas, ya sea con un vector de las mismas o mediante cualquier otra técnica. Cuando se utiliza el vector, es habitual que cada una de las dimensiones se corresponda a un término individual. Tradicionalmente en IR se usa la frecuencia de aparición para representar en el vector cada uno de estos términos, aunque en clasificación afectiva se han obtenido mejores resultados cuando en lugar de medir la frecuencia de aparición del término, simplemente se tiene en cuenta su presencia [13]. Esto se traduce en que los vectores que representarían el documento tendrían elementos binarios, donde uno indica que el término aparece en el documento y cero que no aparece.

Este resultado es significativo ya que marca una diferencia entre la categorización estándar y la clasificación de la polaridad. Mientras que la relevancia de un tema sí que es más probable que

sea enfatizada por la frecuencia de aparición de palabras clave, el sentimiento no suele ser realzado mediante la repetición de los mismos términos.

2.3.1.2 Categorías Gramaticales y Sintaxis

Las categorías gramaticales han sido algo muy utilizado en la minería de opinión ya que hasta cierto punto se puede considerar como una forma de desambiguación del sentido de las palabras.

Seguramente la categoría más importante y sobre la que más se ha investigado sea la de los adjetivos. Se ha detectado en diversas investigaciones una correlación importante entre la presencia de adjetivos en una frase y la subjetividad de la misma [14]. Estas conclusiones han influido para que en posteriores líneas de investigación los adjetivos hayan sido usando como características relevantes para la clasificación de sentimientos.

El hecho de que los adjetivos tengan esta relación directa con la subjetividad de la frase no implica que el resto de categorías no ayuden a identificar expresiones de opinión o sentimiento. Es muy sencillo encontrar muestras en otras categorías como pueden ser los sustantivos (por ejemplo: *furia*) o los verbos (por ejemplo: *querer*) en los que la palabra es un indicador importante de sentimiento.

Por otro lado, las relaciones sintácticas entre las distintas palabras de un texto son un elemento que se ha intentado incorporar al análisis afectivo de texto ya que su análisis, aunque supone profundizar más en el lenguaje con el que se está tratando, puede ayudar a generar patrones gramaticales a partir de los cuales identificar frases como indicadores de subjetividad [14].

2.3.1.3 Negación

La negación es un factor muy importante a la hora de determinar la opinión, especialmente porque mediante las técnicas habitualmente utilizadas en la clasificación automática de textos, una frase que difiere de otra únicamente en la negación y que por tanto tienen que pertenecer a clases opuestas en la clasificación afectiva, no son lo suficientemente distintas.

Un claro ejemplo de cómo las técnicas estándar usadas en la categorización clásica no son adecuadas es el uso de fases de eliminación de palabras de parada en la fase previa de procesado de texto. Es bastante común encontrar palabras con subjetividad muy fuerte tales como “no” o “nunca” dentro de la lista de palabras de parada eliminadas de la representación final del documento, evitando que se pueda tener en cuenta su peso semántico en la clasificación.

Una forma de tratar las negaciones es como características de segundo orden; en ese caso, a pesar de que en una representación inicial como puede ser el vector de características no se tendrían en cuenta, se generaría una segunda representación donde sí lo hiciesen. Otra alternativa es codificar la negación en las palabras cercanas a la misma de forma que la negación esté integrada en la representación inicial.

Otra de las dificultades que presenta la negación es la posibilidad de expresarla de formas sutiles tales como el sarcasmo y la ironía, ambas bastante difíciles de detectar.

2.3.2 Consideraciones

En la clasificación afectiva, de la misma forma que tiene ciertas características específicas a la misma que se salen de lo visto en la clasificación automática de textos, también existen ciertas consideraciones que hay que tener en cuenta a la hora de llevarla a cabo que son más importantes que para casos de categorización clásicos.

2.3.2.1 Consideraciones de Dominio

La precisión de la clasificación afectiva puede estar influencia de forma significativa por el dominio de los documentos sobre los que se está trabajando. El motivo es que una misma frase, dependiendo del contexto en el que se encuentre puede expresar sentimientos distintos. Por ejemplo, el adjetivo “previsible”, dentro de un contexto de crítica cinematográfica o literaria tiene

connotaciones negativas mientras que en un contexto alrededor de un experimento, tiene un sentimiento más positivo. Se puede ir más allá y considerar que no son solo palabras sino hasta expresiones las que cambian según el dominio en el que se empleen.

Este tipo de diferencias aumentan la dificultad cuando el clasificador es entrenado con un conjunto de datos pertenecientes a un dominio mientras que los documentos a clasificar pertenecen a otro distinto.

Si bien es cierto que la noción de opinión positiva y negativa es algo bastante constante en los distintos dominios, en general, tanto la subjetividad como las opiniones son muy sensibles al contexto en el que se encuentran y siempre van a depender del dominio que se esté tratando.

2.3.2.2 Consideraciones de Tema

Ya se ha hablado de las consideraciones a tener en cuenta cuando se trata con documentos pertenecientes a distintos dominios, pero hay que destacar que incluso dentro de un mismo documento puede haber variaciones considerables que pueden afectar el análisis afectivo y que por tanto deben tenerse en cuenta.

Una posible solución para integrar sentimiento y tema es realizar un análisis inicial sobre el tema, y una vez realizado, analizar los resultados obtenidos para el sentimiento [1]. De forma alternativa se puede modelar conjuntamente tema y sentimiento, ya sea de forma simultánea o uno antes que el otro.

Aun teniendo el tema en cuenta, cuando se está trabajando con un documento, no todas las frases que lo componen van a estar englobadas dentro del tema, por lo tanto es un aspecto más que hay que tener en cuenta. Algunos estudios lo que han hecho es dividir el proceso en dos pasos, un primer paso en el que se determina si cada frase pertenece o no al tema, y una segunda frase en la que el análisis de opinión se aplica únicamente sobre las frases que se han determinado pertenecientes al tema. Todo esto se basa en la suposición de que si una frase expresa opinión y se ha considerado que pertenece al tema, entonces esa opinión será relativa a él [15].

Siguiendo esta línea, también hay que contemplar la posibilidad de que haya más de un tema dentro de un mismo documento, como por ejemplo sería un documento que compare dos productos o servicios. Yendo a un caso un poco más extremo, dentro del tema tratado en el documento se puede estar hablando de distintos aspectos del mismo, dando lugar a sub-temas que también habría que tratar por separado. En ambos casos existe la posibilidad de ignorar los temas secundarios y evaluar únicamente la opinión sobre el tema principal (en algunos casos los temas principales están definidos desde el principio), aunque es más apropiado identificar todos los temas incluidos y determinar las opiniones para cada uno de ellos [15].

2.3.2.3 Consideraciones de Lenguaje y Estructura

Uno de los principales problemas a la hora de evaluar sentimientos y opiniones es que éstos pueden ser expresados de formas muy sutiles que resultan muy complicadas de analizar de forma automática. Recursos como el sarcasmo o la ironía ya han sido mencionados previamente como ejemplos de situaciones que no ayudan a extraer sentimientos, pero éstos son solo algunas de las formas del lenguaje que van a dificultar la extracción de opiniones. Sin necesidad de llegar a recursos como los mencionados, es perfectamente factible expresar opinión de forma que no se puedan identificar con claridad palabras clave.

Por lo general, la información suele seguir una estructura global predefinida, que aunque no tiene por qué darse siempre, a mayor o menor nivel es fácil de identificar. Un ejemplo muy abundante son aquellos fragmentos en los que, aunque en la mayor parte del párrafo hay una polaridad determinada dominante, con una última frase, se niega todo lo dicho de forma que la polaridad final del fragmento es la contraria a la que se puede asumir inicialmente.

Se puede ilustrar fácilmente con el siguiente texto:

> "*La premisa de la serie era original, tenía un reparto de calidad y gracias al renombre del creador, la cadena no reparó en gastos a la hora de promocionarla. Sin embargo, cuando finalmente se estrenó, fue un estrepitoso desastre.*"

Si se plantea esta situación en análisis de texto (aparecen muchas palabras de las consideradas claves) y en análisis de opinión (aparecen muchas palabras claves de una polaridad), se puede ver con claridad como usando la misma técnica para ambos casos, en uno se determinaría correctamente que el documento es relevante, mientras que en el análisis de opinión, se consideraría que el documento tienen la polaridad equivocada.

Teniendo en cuenta un análisis de más bajo nivel, se puede ver la dependencia clara que va a tener el análisis afectivo tanto del orden con estructuras tan sencillas como "A es mejor que B" y "B es mejor que A": su representación vectorial sería la misma, la única variación es el orden de los elementos, y sin embargo precisamente esto puede cambiar por completo la opinión de la misma.

Se ha podido observar la importancia de resumir el sentimiento de un documento. Concretamente se ha comprobado cómo evaluar las N últimas frases de un artículo resumen mejor el sentimiento del mismo que usar las N primeras frases o las N frases determinadas más subjetivas [13].

2.3.3 Técnicas Utilizadas

En este apartado se verá un breve resumen de algunas de las técnicas que se utilizan en el análisis de documentos de opinión, aunque antes de hablar de ellas hay que mencionar los modelos psicológicos emocionales más relevantes que se aplican al mundo computacional.

El primer modelo es el basado en categorías emocionales, y es seguramente el más intuitivo de los que se van a mencionar. La primera definición de estas categorías fue dada por Plutchik y Kellerman en 1980 y constaba de ocho emociones básicas. Esta definición fue mejorada más adelante dando lugar al modelo Circumplex [16], que usa una circunferencia en la que se representan las características emocionales a partir de dos ejes: valencia (positiva o negativa) y agitación (alta o baja).

Existe un modelo parecido al Circumplex llamado dimensiones emocionales, en el que se cuantifica las dimensiones de valencia, activación y control mediante un vector de tres elementos.

Por último existe un último modelo, el modelo OOC [17], que en contraposición a lo visto para los otros modelos, presenta una jerarquía cognitiva de emociones evitando el uso de categorías y dimensiones. Concretamente se basa en el uso de eventos, agentes y objetos y define veintidós categorías emocionales frente a los veintiocho estados afectivos del modelo Circumplex.

Las distintas técnicas que se utilizan en clasificación afectiva de textos se diferencian principalmente en el enfoque que cada una de ellas dan a la detección y clasificación de emociones en textos. A continuación se habla de los principales tipos de técnicas.

2.3.3.1 Técnicas Basadas en IR

Una de las técnicas empleadas para determinar la orientación semántica de un texto es realizar un análisis "*Pointwise Mutual Information and Information Retrieval*" o PMI-IR.

Turney describió un clasificador no supervisado basado en la opinión en el que se decide si el documento tiene carácter positivo o negativo en base a la orientación semántica de los términos que lo componen [14]. La orientación de cada uno de los términos se calcula mediante el algoritmo PMI-IR, en el que se estima la Información Mutua Puntual (PMI) entre los términos y unos representantes unívocos denominados semillas.

La idea del algoritmo es que expresiones que tienen asociada una opinión positiva van a aparecer mucho más frecuentemente cerca de términos con un clara orientación semántica positiva que cerca de términos con una clara orientación negativa. Dicho de otra forma, las palabras que expresan sentimientos parecidos, tienden a co-aparecer en un mismo texto mientras que si tienen sentimientos contrarios no suelen aparecer juntas.

Dos ejemplos de semillas de polaridades opuestas que se podrían usar como representantes serían "*excelente*" y "*horrible*". Ambas tienen connotaciones opuestas claramente definidas y atiende a razón que en cualquier contexto las palabras que salen con ellas tengan también una orientación semántica análoga.

2.3.3.2 Técnicas Basadas en Clasificación de Textos

Dentro de las técnicas basadas en clasificación de textos aplicadas a la clasificación afectiva hay dos tipos principales, las Máquinas de Vectores Soporte (SVM) y Análisis de Semántica Latente (*Latent Semantics Analysis,* LSA).

Las máquinas de vectores de soporte son de las técnicas más utilizadas dentro de la clasificación temática de grandes colecciones de texto y se encuentra en el desarrollo de muchos clasificadores emocionales. Un ejemplo de ello es el clasificador emocional de blogs desarrollado por Leshed y Kaye [18].

Por otro lado, como ejemplo de análisis de semántica latente, Turney y Littman presentan un sistema de identificación de la polaridad en el que la polaridad de cada palabra se calcula mediante la diferencia entre su similitud con un grupo de palabras negativas y otro de palabras positivas [19].

Ambas técnicas tienen como problema fundamental la necesidad de un volumen elevado de datos en las fases de pruebas y entrenamiento para asegurar un buen funcionamiento.

2.3.3.3 Técnicas Basadas en Diccionarios Afectivos

Estas técnicas se basan en buscar las palabras afectivas de un texto contenidas en un diccionario de vocablos afectivos construido previamente.

Uno de los algoritmos más destacados es *Emotional Keyword Spotting* (EKS) debido a su sencillez de implementación y bajo coste computacional. En él, la orientación de la opinión del texto se determina haciendo una media de los valores emocionales de cada una de las palabras clave que se hayan detectado dentro del texto [20]. Dentro de este tipo de algoritmos es común encontrarse la denominada afinidad léxica, en la cual se exporta la emoción de las palabras clave a sus palabras más cercanas [21].

Tanto si se utiliza afinidad léxica como si no, este tipo de métodos no detectan los cambios de polaridad debido a particularidades del texto tales como la negación [22], lo que supone una desventaja importante. A pesar de esto, gracias a las propiedades de sencillez y coste computacional mencionadas, este tipo de técnicas son bastante populares en aplicaciones de tiempo real.

2.3.3.4 Otras Técnicas

Como se ya se ha mencionado en numerosas ocasiones, la investigación sobre la clasificación de texto afectivo ha experimentado un crecimiento significativo en los últimos años, por lo tanto es lógico que se haya estudiado el resultado de diversas técnicas, no todas englobables en los tipos vistos.

Se han obtenido resultados favorables usando un sistema en el que primero se analiza la subjetividad de las frases de un texto, y una vez determinada, se analiza la polaridad de las opiniones. Esta técnica utiliza sistemas Bayesianos, Máquinas de Soporte Vectorial y modelos de Máxima Entropía.

Por otro lado, Liu, Lieberman y Selker [23] extrajeron conceptos de una voluminosa base de conocimiento del sentido común; de esta forma se pueden detectar emociones en frases donde a priori no existe una emoción definida de forma explícita. Aunque esto es muy útil, se trata de una técnica muy compleja debido a todo el tratamiento semántico que hay que hacer de la base de conocimiento. Otro sistema complejo, el desarrollado por Ovesdotter, Roth y Sproat [24], utiliza

palabras afectivas, parámetros del texto tales como la temática, longitud de frase, etc. para predecir la emoción del texto dentro del ámbito de la lectura de cuentos.

Cabe destacar que la mayoría de las técnicas empleadas actualmente se centran en analizar primero a nivel de frase, de forma que luego se puedan realizar estimaciones sobre grupos de frases y luego pasar a estimar los párrafos dentro del texto a partir de las polaridades buscadas [25]. Por lo general, el enfoque utilizado para determinar la polaridad de un documento a partir de sus componentes se basa en un tipo de suma semántica ponderada de los sentimientos individuales presentes; en caso de ambigüedad en las polaridades identificadas, se utilizan modelos de abstracción conceptuales.

Otro de los enfoques utilizado en la detección de sentimientos se basa en identificar frases comparativas en un texto. Ya se ha hablado de las dificultades derivadas de este tipo de estructuras y lo abundantes que pueden ser en ciertos escenarios (en críticas de servicios o productos, es muy común comparar con otros servicios/productos). La idea es encontrar para cada frase todas las reglas de polaridad que se satisfacen, y escoger aquella que tenga una confianza más alta para clasificarla.

2.4 NTCIR

Como ya se ha mencionado en la introducción del proyecto, el corpus a utilizar es el de uno de los talleres del proyecto NTCIR[1]. NTCIR viene de "*NII Test Collection for IR Systems*", un proyecto organizado por la institución de investigación académica japonesa NII, "*Nacional Institute of Informatics*" [26].

NTICR Workshop son una serie de talleres de evaluación diseñados para mejorar la investigación sobre tecnologías de acceso a la información (IA), incluyendo IR, respuesta de preguntas, producción de resúmenes, extracción, etc. Sus objetivos son:

1. Fomentar la investigación sobre tecnologías de Acceso a la Información proporcionando colecciones de gran tamaño para experimentos y una infraestructura común de evaluación para permitir las comparaciones entre los distintos sistemas.
2. Proporcionar un foro para grupos de investigación interesados en la comparación entre sistemas y el intercambio de ideas de investigación en un ambiente informal.
3. Investigar métodos de evaluación de técnicas de Acceso a la Información y métodos para construir conjuntos de datos a gran escala reutilizables para experimentos.

En cada taller de evaluación se suele proporcionar colecciones de prueba (conjuntos de datos que se pueden usar para experimentos) y procedimientos unificados de evaluación para los resultados de los experimentos. Cada grupo participante investiga y lleva a cabo experimentos usando unos datos comunes proporcionados por los organizadores de NTCIR.

La importancia de tener colecciones de prueba de gran escala reutilizables en IA ha sido ampliamente reconocida, hasta el punto que los talleres de evaluación son reconocidos como un nuevo estilo de proyectos de investigación que facilitan dicha investigación proporcionando datos y un foro de intercambio de ideas y tecnología.

El primer taller de NTCIR comenzó en noviembre de 1998, y desde entonces se han organizado ocho talleres, incluyendo NTICR8 que comenzó en mayo de 2009 y no ha finalizado todavía. Los talleres de NTCIR hacen hincapié en el trabajo con lenguajes asiáticos, estando disponible para la gran mayoría de tareas conjuntos de datos en inglés, japonés, chino (tradicional y simplificado), y en algunos casos, coreano.

[1] http://research.nii.ac.jp/ntcir/index-en.html

Para este proyecto se ha partido de los datos proporcionados para NTCIR-6 pero se han tenido en cuenta algunos de los resultados de NTCIR-7 por tratarse del taller más reciente en el momento del comienzo del proyecto y por tanto el que contenía técnicas más actuales.

Cada uno de los talleres de NTCIR se divide en distintos módulos de sub-tareas:

- NTCIR-6:
  - *Cross-Lingual Information Retrieval* (CLIR)
  - *Cross-Language Question Answering* (CLQA)
  - *Patent Retrieval* (PATENT)
  - *Question Answering* (QAC)
  - *Pilot Task (Opinion)*
  - *Multimodal Summarization for Trend Information* (MuST)
- NTCIR-7:
  - *Advanced Cross-Lingual Information Access*
  - *User Generated Contents*
    - *Multi-lingual Opinion Analysis Task* (MOAT)
    - *Cross-Lingual Information Retrieval over Blog* (CLIR-B)
  - *Focused Domains*
    - *Patent Mining Task (PAT MN)*
    - *Patent Translation Task (PAT MT)*
  - *Multimodal Summarization of Trends* (MuST)

Este proyecto se va a centrar en la tarea común a los dos talleres relativa al análisis de opinión. En ella hay dos análisis bien diferenciados, el relativo a relevancia, que incluye como única subtarea identificar si el documento es relevante a una consulta dado, y el análisis de opinión. Dentro de este último habrá tres subtareas distintas: identificar la subjetividad de un documento (única tarea obligatoria para la participación en el taller), identificar la polaridad de la opinión en el caso de que la tenga y el sujeto que da dicha opinión.

Como se puede intuir por el nombre de algunas de las tareas, se va a disponer de datos en distintos idiomas En el caso del módulo de análisis de opinión en NTCIR-6 se facilitaron datos en inglés, japonés y chino; en NTCIR-7 se disponía de corpus en inglés, en japonés, en chino tradicional y en chino simplificado. En este proyecto se realizarán todos los análisis sobre el corpus proporcionado en inglés.

Es importante destacar que el enfoque a dar al sistema implementado para llevar a cabo estos análisis no está definido en la tarea, por lo que dentro de los sistemas participantes hay una diversidad importante de soluciones en las que se priorizan diversos aspectos o simplemente se investigan los resultados alcanzados con técnicas concretas.

Por lo general, es muy habitual ver sistemas que utilizan variantes del sistema clásico de IR para implementar el análisis de la relevancia, aunque no son las únicas técnicas utilizadas. A continuación se verán a grandes rasgos algunas de las técnicas implementadas por los participantes de NTCIR:

- Expansión de la consulta: en la que mediante diversas técnicas se incrementa el número de términos incluidos en la consulta. Algunas de las formas de realizar este incremento es mediante realimentación [27] o usando motores de búsqueda web. Asimismo es muy

común utilizar probabilidades logarítmicas para controlar el peso de los nuevos términos con los que se expande la consulta [28].

- Uso de varianza para definir la consulta: una forma de determinar cuál va a ser la lista de términos realmente relevantes para representar una consulta será utilizando la varianza de la frecuencia relativa de los mismos. Un término muy frecuente en una consulta y muy poco frecuente en el resto, será mejor representante [29].
- Patrones: se obtienen a partir de palabras claves de la descripción del tema que define la relevancia y mediante una expansión a través de búsqueda web, patrones que se puedan considerar inherentemente relevantes al tema. Al comparar, si un documento se corresponde a cualquiera de los patrones definidos, se considerará relevante [30].
- Distancia y similitud: en lugar de usar únicamente la medida de relevancia, se genera una medida compuesta por la medida de la similitud del documento al tema y la distancia del documento al documento más relevante. Se aplicará un umbral a partir del cual los documentos se consideren relevantes [31].
- Uso de ventanas: las ventanas se usarán para determinar bloques de documentos, los cuales estarán representados mediante un vector, y por tanto se podrá calcular la distancia con el tema. Si todos los bloques que contienen cierto documento tienen un valor medio de distancias mayor que cierto umbral, entonces el documento será relevante [32].

Se puede apreciar cómo muchas de ellas son variantes o combinaciones de técnicas vistas en apartados anteriores.

Como ya se ha mencionado, la tarea propuesta por NTCIR se centra en el estudio de la opinión e incluye varias subtareas dedicadas a estudiar distintos aspectos de la misma. En este proyecto únicamente se va a tratar el análisis de la subjetividad por lo que solamente se hablará de las soluciones propuestas para esa subtarea.

Dentro de las soluciones se observan dos claras tendencias: el uso de diccionarios afectivos para la definición de características mediante las que representar los documentos y el uso de SVM para la clasificación.

De las SVM ya se ha hablado con anterioridad aunque cabe mencionar que al ser un método bastante complejo es muy habitual utilizar implementaciones ya hechas para el ámbito del reconocimiento de patrones tales como SVM *Light*[2] o SVM *Tiny*[3].

La gran diferencia entre las distintas soluciones que utilizan SVM está en cómo se escogen las características con las que se define el documento. Estas características no son siempre las mismas pero algunas de ellas son usadas a menudo en los distintos sistemas. Algunas de las características más comunes en las soluciones vistas son:

1. Partes del discurso o categoría gramatical (POS, *Part Of Speech*).
2. Verbos de comunicación.
3. Elementos que contienen polaridad, ya sean palabras o frases.
4. Adverbios que suelen aparecer en expresiones de opinión o adverbios que aparecen junto a un verbo.
5. Entidades de nombre.

[2] http://svmlight.joachims.org/

[3] http://chasen.org/~taku/software/TinySVM/

6. Dependencias en la frase.
7. Modismos.
8. Marcadores de construcción.

Muchas de estas características no son inmediatas, por lo que habrá que realizar análisis previos sobre los documentos para obtenerlas. Por ejemplo, para encontrar palabras con polaridad asociada es muy común usar diccionarios afectivos; para encontrar las entidades de nombre existen los denominados extractores de entidades de nombre (*Named Entity Extractors*, NER); las dependencias de frase se pueden obtener utilizando analizadores sintácticos, etc.

Es importante tener en cuenta que el uso de estas características no limita como se podrán clasificar los documentos, por lo que se usan tanto en soluciones con las ya mencionadas SVM como con otras formas de clasificar. En algunos casos se ha considerado que la combinación de algunas de las características es suficiente para etiquetar una frase como subjetiva. Un ejemplo sería considerar que una frase tiene opinión si incluye una expresión o palabra de opinión y una entidad que dé opinión [33][34].

El hecho de que estas características como representación de los documentos permitan usar distintas técnicas de clasificación, favorece la comparación de las mismas, encontrándose varias comparaciones entre usar o no técnicas de aprendizaje máquina, o incluso alguna comparación entre el rendimiento entre varias de ellas como puede ser la comparación entre Naïve Bayes (del que te también se ha hablado y segundo método máquina más popular entre las soluciones propuestas) y SVM [35].

Por supuesto no todas las soluciones propuestas se limitan a los escenarios descritos, por lo que a continuación se hablará brevemente de algunas de las que no lo hacen:

- Modelo de máxima entropía. El modelo de máxima entropía utiliza características típicas como n-gramas y categorías gramaticales para representar los documentos pero también asume independencia entre cada una de las frases, algo que no se cumple cuando se trata de realizar un análisis afectivo. Para evitar esta suposición, se realiza un análisis global de la subjetividad del documento mediante SVM; este resultado se añade como una característica adicional en el modelo de máxima entropía [30] de forma que se tenga en cuenta al clasificar.
- Información local y global: como se ha mencionado en el punto anterior, es una suposición bastante generalizada que si en un documento la mayoría de las frases expresan una opinión, otra frase dentro de ese documento tendrá una probabilidad mayor de tener opinión [35], lo que implica añadir información del documento a la clasificación de subjetividad de cada una de las frases.

  Una forma de hacerlo es calculando la medida de opinión total a partir de dos medidas distintas, una local, calculada por ejemplo con diccionarios afectivos y otra global, a partir de las estadísticas del documento. Otra opción puede ser, en vez de tener en cuenta información del documento completo, utilizar ventanas configurables para definir bloques sobre los cuales se hacen análisis subjetivos de cara a tener una visión un poco más amplia que una única frase [32].
- Palabras con opinión: Una alternativa a obtener las palabras que expresan opinión de los diccionarios afectivos es seleccionar esta lista de palabras del propio conjunto de entrenamiento, de forma que la lista de palabras que se va a utilizar esté limitada a palabras que aparezcan en el corpus. Dentro de las soluciones vistas, hay técnicas para completar estas listas, ya sea de forma manual [29] (seleccionado las palabras asociadas a las palabras más frecuentes de los temas), o a través de algoritmos como EM (*Expectation – Maximization*) [27].

Dos de los puntos que se destacan en muchas de las soluciones propuestas son los problemas asociados a los dominios y la falta de portabilidad de los sistemas debido a la particularización de la solución para ciertos idiomas. Ambos puntos se han visto previamente dentro de las consideraciones a tener en cuenta en la clasificación afectiva.

Muchos de estos sistemas obtienen palabras claves o características para la fase de clasificación a partir de los documentos de los datos entrenamiento. Sacar unas características fiables, en muchos casos necesita un volumen de datos considerable que no siempre va a estar disponible, mucho menos en el mismo dominio.

En NTCIR-7 se daba una colección de entrenamiento bastante reducida, por lo que muchos de los participantes recurrieron a los datos disponibles de NTCIR-6 para entrenar sus sistemas. Ambas colecciones pertenecen a distintos dominios, por lo que desde la fase de entrenamiento se está añadiendo un error al sistema que va a ser especialmente acusado en la clasificación afectiva. Para evitar esto existen algoritmos tales como SCL (*Structural Correspondence Learning*) que intentan minimizar las discrepancias entre dominios [33][34][36].

Por otro lado, hay que recordar que la idea inicial de la tarea propuesta por NTCIR es realizar un análisis de opinión multilingüe, por lo que cabe pensar que el idioma elegido será un factor relevante en el análisis, diseño y rendimiento del sistema implementado.

Se ha visto cómo varias de las soluciones implementadas utilizan características ya sea morfológicas o estructurales particulares del idioma, algo que limita la solución. Un ejemplo claro es el visto en [37], donde se usa que en japonés características tales como el modo, el tiempo verbal o el aspecto (bastante utilizadas como características para determinar subjetividad) se encuentran en expresiones al final de la frase. Parece bastante intuitivo que al tratarse de una característica particular del lenguaje, su aplicación a cualquier otro, inglés por ejemplo, no daría los mismos beneficios.

Siempre que se trabaje con distintos idiomas va a haber diferencias importantes que se van a hacer notar a la hora de clasificar. Partiendo de que no todas las características que se han visto van a ser igual de significativas para los distintos idiomas, es muy razonable asumir que para un mismo sistema se van a obtener resultados bastante diferentes.

De la misma forma, las herramientas que se usan para obtener las distintas características no siempre están disponibles en todos los idiomas, por lo que compatibilizar un mismo sistema para que trabaje con todas las herramientas necesarias supondrá una dificultad añadida.

3 Diseño del Sistema

En este apartado se describirá en detalle el diseño seguido para la implementación del sistema, incluyendo tanto diagramas generales que den una visión global del mismo como detalles sobre todas las decisiones de diseño tomadas a lo largo de su implementación.

Una de las primeras decisiones es determinar el tipo de aprendizaje que se utilizará para obtener los distintos clasificadores. En este caso, se dispone de una clasificación previa del conjunto de datos según las distintas clases o categorías, por lo que se utilizará aprendizaje supervisado. Una de las desventajas de este tipo de aprendizaje es la dependencia de los resultados tanto del clasificador elegido como del conjunto de datos que se usan para entrenar los distintos clasificadores.

Los datos de entrada de los que parte el sistema tienen dos secciones bien diferenciadas: los artículos que se quieren analizar (ubicados en la carpeta “NTCIR6-Opinion-EN”) y los temas con los que están relacionados (en la carpeta “NTCIR6-Topics”).

Los temas son dados en un listado, el archivo “NTCIR6OpinionTopics_EN.txt” (ver Anexo A – Temas de NTCIR-6), donde cada uno de ellos tendrá el formato del siguiente ejemplo:

```
<TOPIC>
<NUM>001</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>Time Warner, American Online (AOL), Merger, Impact</TITLE>
<DESC>Find reports about the impact of AOL/Time Warner merger.</DESC>
<NARR>
<BACK>Time Warner and American Online (AOL) announced a merger on January 10th, 2000. The market value was estimated at $US350 billion making it the biggest merger in the US.</BACK>
<REL>Comments on AOL/Time Warner merger's effects on Internet and entertainment media businesses are relevant. Descriptions of the development of the AOL/Time Warner merger are partially relevant. Information about the total amount and the transformation of ownership structure are irrelevant.</REL>
</NARR>
<CONC>Time Warner, American Online, AOL, Gerald Levin, merger, M&A, Merger and Acquisition, media, entertainment business</CONC>
</TOPIC>
```

Figura 5: Ejemplo de tema.

Y en donde el significado de cada una de las etiquetas es el siguiente:

<TOPIC>: define el comienzo y el final del tema.
<NUM>: identificador del tema.
<SLANG>: lenguaje fuente del tema (EN, CH, JP).
<TLANG>: lenguaje destino del tema.
<TITLE>: representación concisa de la petición de información, compuesta de un nombre o de un grupo nominal.
<DESC>: pequeña descripción del tema. Breve descripción de la información necesaria compuesta de una o dos frases.
<NARR>: Marca el comienzo de la información narrada relativa al tema.
<BACK>: antecedentes del tema descrito.
<REL>: interpretación más en profundidad sobre la petición de información y los nombres propios, lista de lo que es y no es relevante, requisitos especiales o limitaciones de los documentos relevantes.

El número total de temas proporcionado es 28. Debido al enfoque dado en este proyecto, la única información del tema que será utilizada será el identificador, el título y la descripción (<NUM>,

<DESC> y <TITLE>). De estos dos últimos es donde se sacará la información que va a constituir la consulta asociada a cada uno de los temas.

La segunda parte de los datos son los artículos. Cada uno de los 28 temas tendrá asociado un número de artículos, distinto para cada uno de los temas. Cada artículo es dado en un único fichero con el formato de la siguiente figura.

```
<DOC>
<DOCNO>EN-200001142A02OA305</DOCNO>
<LANG>EN</LANG>
<HEADLINE>[EDITORIAL] A major merger</HEADLINE>
<DATE>2000-01-14</DATE>
<TEXT>
<STNO>0001</STNO>
 The news of Time Warner and America Online's (AOL) plans to merge reminds us that major changes are sweeping the information industry.
<STNO>0002</STNO>
But this deal certainly does not come as a surprise to those who work in this field.
</TEXT>
</DOC>
```

Figura 6: Fragmento de artículo.

Donde cada una de las etiquetas representa lo siguiente:

<DOC>: determina el comienzo del documento.
<DOCNO>: es el identificador del documento, único dentro del conjunto de todos los datos.
<HEADLINE>: titular del artículo.
<DATE>: fecha de publicación del artículo.
<TEXT>: determina el comienzo del texto del artículo.
<STNO>: abreviatura para “Sentence Number”. Es el número identificador para cada una de las frases que componen el artículo. Estos identificadores son solo válidos dentro del artículo.

Como se puede ver en las etiquetas, para facilitar el procesado de los artículos, el contenido de cada uno de ellos viene separado por las frases que lo componen. Esto es de gran ayuda a la hora de determinar la unidad de análisis del sistema, o dicho de otra forma, con qué nivel de detalle se analizarán tanto la relevancia como la opinión.

Estas unidades de análisis se asignarán de forma aleatoria, y siempre asegurando que se mantengan las proporciones en los distintos temas, a los conjuntos en los que se va a dividir el corpus. Se trabaja con tres conjuntos, el de entrenamiento, que se utilizará para entrenar los clasificadores a partir de las etiquetas previamente asignadas, el conjunto de pruebas, con el que se determinarán los valores óptimos para los distintos parámetros a usar en el sistema, y el conjunto de validación que será utilizado para obtener los valores finales del sistema a partir de los cuales se evaluará el rendimiento del mismo.

La estructura que va a seguir el sistema no va a diferir mucho de la encontrada normalmente en los sistemas clasificadores de texto (ver Figura 7). Las tareas de encontrar documentos relevantes o documentos con opinión se pueden caracterizar como una clasificación binaria de documentos relevantes y no relevantes o en el caso del clasificador de opinión, documentos con opinión o sin ella.

El proceso básico que se va a seguir, y que determinará el procesado de los datos de entrada es el siguiente: cada una de las frases de los artículos se considerará un documento a clasificar y la consulta con respecto a la cual se evaluará el grado de opinión expresado será la obtenida de la información proporcionada para el tema al que esté asociado.

En la Figura 7 se ve una primera etapa denominada "*Procesado Inicial*". Es precisamente en esta etapa donde se extraerá la información necesaria de cada uno de los documentos y de cada tema y se les dará el formato necesario para poder ser usados por el clasificador.

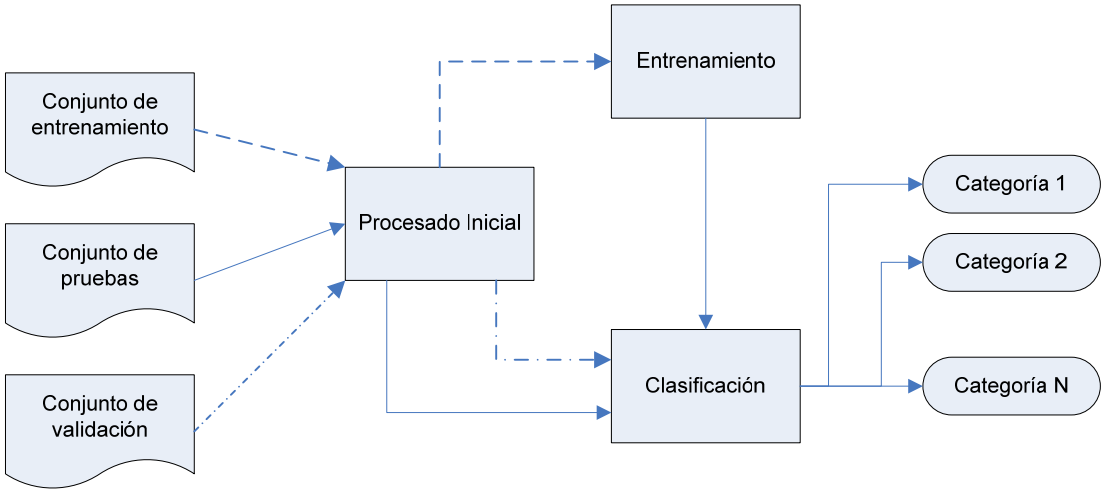

Figura 7: Diagrama de bloques básico de un clasificador de texto.

La segunda etapa es la denominada "*Entrenamiento*", y en ella, como indica su nombre, se entrenará el sistema con la parte de datos que se determinen como el conjunto de entrenamiento de forma que el sistema aprenda de ellos y pueda desarrollar el clasificador que se usará para el resto de los datos en la etapa de clasificación.

La última etapa, "*Clasificación*", toma el clasificador desarrollado en la etapa anterior y lo utiliza para asignar cada uno de los documentos pertenecientes a los conjuntos de validación y pruebas a las categorías correspondientes.

El sistema diseñado en este proyecto va a tener una estructura muy parecida a la genérica de un clasificador de texto, aunque con la particularidad de que este sistema se aplicará a dos clasificaciones distintas, una según la relevancia y otra según el grado de opinión expresado. De esto se puede intuir que el sistema tendrá dos fases de entrenamiento distintas, una para cada uno de los clasificadores.

La parte de clasificación se podrá agrupar en una misma etapa, ya que al ser clasificadores independientes y compartir el formato de trabajo, no existe ningún impedimento para que se pueda hacer de forma simultánea.

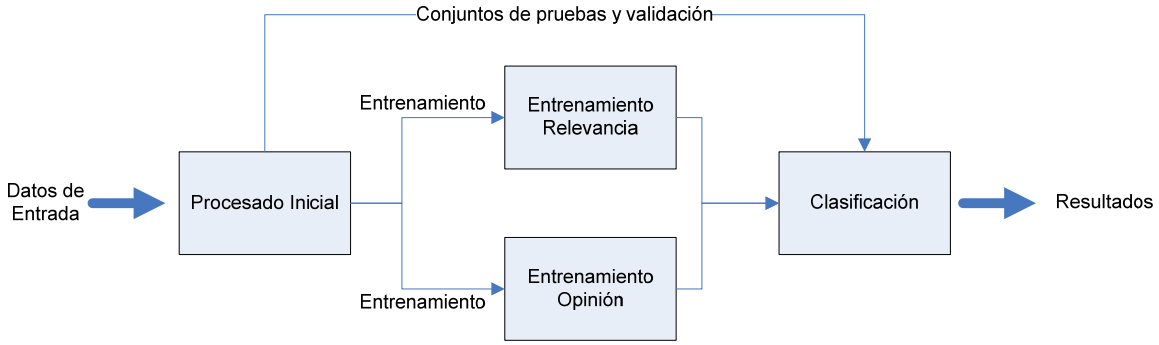

Figura 8: Diagrama de bloques simplificado del sistema.

A continuación se verá con más detalle cómo está implementada cada una de estas etapas del sistema.

3.1 Procesado Inicial

El procesado de los datos es el primer módulo del sistema y será en él donde se obtendrá de los datos proporcionados la información necesaria y en el formato adecuado para poder entrenar cada uno de los clasificadores.

Como se ha podido ver en la Figura 8, esta etapa es igual para las dos clasificaciones que se van a realizar en el sistema y común para todos los conjuntos de datos que se va a trabajar.

En esta fase de procesado inicial existen dos partes bien diferenciadas, una primera etapa de extracción de la información en la que se procesará directamente la información de partida de la que se dispone y una segunda etapa en la que se adaptará esa información extraída al formato necesario para que pueda ser utilizada en las fases posteriores del sistema.

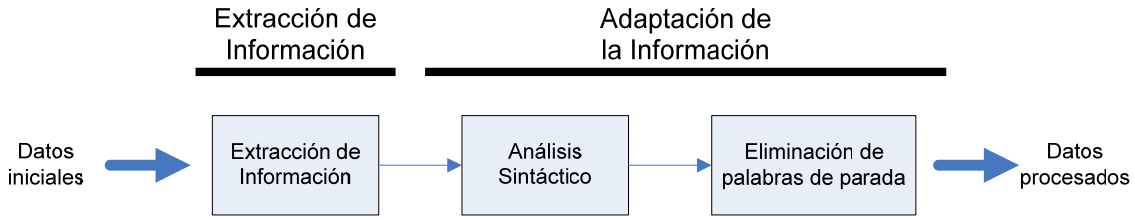

Figura 9: Diagrama de bloques del procesado inicial.

En la Figura 9 se puede ver que en el caso de la segunda etapa se realizarán dos tareas para alcanzar el resultado deseado. En los siguientes apartados se verá en detalle lo que implica la realización de cada una de las etapas.

3.1.1 Extracción de Información

Como se de esperar, no toda la información proporcionada inicialmente va a ser utilizada, por lo tanto el primer paso dentro del procesado de datos va a consistir en extraer la información relevante para el análisis que se va a llevar a cabo.

Ya se ha visto en el apartado anterior cómo los datos de entrada con los que se va a trabajar vienen en dos formatos distintos, por lo que el procesado no será exactamente igual para ambos a pesar de que la idea base sea la misma.

En el caso de los **artículos**, esta primera fase de la etapa de procesado consistirá en separar cada una de las frases de cada uno de los artículos en unidades independientes. Para cada artículo se proporciona un identificador del mismo, un titular, la fecha de publicación y el texto que lo compone separando las frases mediante etiquetas.

Si se parte del ejemplo de artículo visto en la Figura 6, este procesado consistirá en llegar a lo siguiente:

```
0001 - The news of Time Warner and America Online's (AOL) plans to merge
reminds us that major changes are sweeping the information industry.
0002 - But this deal certainly does not come as a surprise to those who
work in this field.
```

Esas frases van a ser la unidad de análisis que se va a tomar a lo largo de todo el sistema por lo que para facilitar el manejo de datos se tendrá un fichero independiente para cada una de ellas. El

hecho de usar cada frase como unidad de análisis viene facilitado tanto por la estructura de los artículos que claramente facilita su extracción como por la estructura de los resultados dados que habrá que usar en la fase de entrenamiento.

En el caso de los **temas** la decisión de qué parte de la información dada utilizar para definir la consulta será más complicada debido a la cantidad de información proporcionada. Como se ha visto en la Figura 5, para cada uno de los temas se proporciona al menos un identificador numérico, los lenguajes de origen y destino, el título del tema, una descripción y dos campos narrados: antecedentes e interpretación de qué se considera relevante para el tema (ver Anexo A – Temas de NTCIR-6).

Precisamente por esto, se ha decidido analizar los resultados que obtendrá el sistema para distintas definiciones de consulta, tomando los dos campos que se han considerado más similares a una consulta, la descripción del tema y su título.

Para hacerse una idea de qué información aporta cada uno de los datos mencionados y por qué se han seleccionado, a continuación se incluye un ejemplo de cada uno de ellos para uno de los temas con los que se trabajará:

> ID de Tema: 001
>
> Título: Time Warner, American Online (AOL), Merger, Impact
>
> Descripción: Find reports about the impact of AOL/Time Warner merger.

Evaluar el título es algo muy inmediato ya que se aproxima mucho a la idea de consulta que se tiene, pero la decisión de usar también la descripción no es tan inmediata. La idea de evaluar los resultados proporcionados por la descripción del tema viene de tener en cuenta que, al ser una frase completa construida en lugar de solamente palabras clave como es el título, tras procesarla, al estar incluyendo en dicho procesado la eliminación de las palabras de parada, la consulta resultante de la descripción tendrá una lista de palabras clave mayor que la resultante del título.

Tras haber sido procesados, para cada tema se tendrá una carpeta con los ficheros correspondientes a cada una de las frases de los artículos relacionados con el mismo, y en la carpeta de temas, un fichero con la consulta asociada a ese tema.

3.1.2 Adaptación de la Información

Una vez extraída la información que se va a utilizar, se pasa a la parte en la que se tiene que adaptar esta información al formato adecuado para que pueda ser utilizada por el resto de las etapas del sistema.

Como se ha visto en la Figura 9, esta etapa consta de dos sub-etapas diferenciadas: una primera etapa en la que se lleva a cabo un análisis sintáctico de cada unidad de análisis para dejarla en su estado más básico, los lemas, y una segunda etapa donde se eliminarán las palabras de parada presentes.

Ambas etapas son tareas básicas de procesamiento de texto. La fase de reducir al estado más básico o lematización ayudará a identificar como una misma palabra a todas las derivadas de una misma raíz, eliminando una cantidad importante de variantes. La fase de eliminación de palabras de parada eliminará de las unidades de análisis palabras que no aportan información significativa a la unidad.

A continuación se verán ambas fases con más detalle.

3.1.2.1 Análisis Sintáctico

Como ya se ha mencionado, la primera tarea de la fase de adaptación de la información es la lematización o *stemming* en la que se analizará sintácticamente cada una de las unidades de análisis para obtener una lista de todos los lemas que la componen.

Este proceso permite agrupar las palabras que tienen un mismo significado de forma que haya menos palabras distintas en el conjunto de datos. Dos ejemplos que lo ilustran claramente son las conjugaciones verbales y los plurales de los nombres. Para el análisis que se va a realizar, aporta más información saber que *cantar* aparece dos veces en un documento, que saber que aparecen *cantaré* y *cantaría*. De forma análoga, no será útil considerar dos palabras distintas *cantante* y *cantantes* ya que su significado semántico es el mismo.

Para poder realizar esta lematización se utilizará un analizador sintáctico o "parser", Freeling 2.1 [1][38]. Freeling es un paquete de librerías que proporciona una serie de servicios de análisis de idiomas, algo que facilita su utilización por aplicaciones externas. Incluye un simple programa que actúa como interfaz básico a la librería y que permite el análisis de textos a través de la línea de comandos. Es precisamente esta última opción la que será utilizada.

Aparte de estas opciones, Freeling dispone de una demo online[4] mediante la cual se podrá ilustrar los resultados que se obtendrán al utilizar el interfaz básico proporcionado con el paquete de librerías.

La frase a analizar será "*Otherwise the nation will lag behind in the trend of globalization.*", obtenida de la frase con identificador "0025" del artículo "KT2000_09741" perteneciente al tema "001" del corpus.

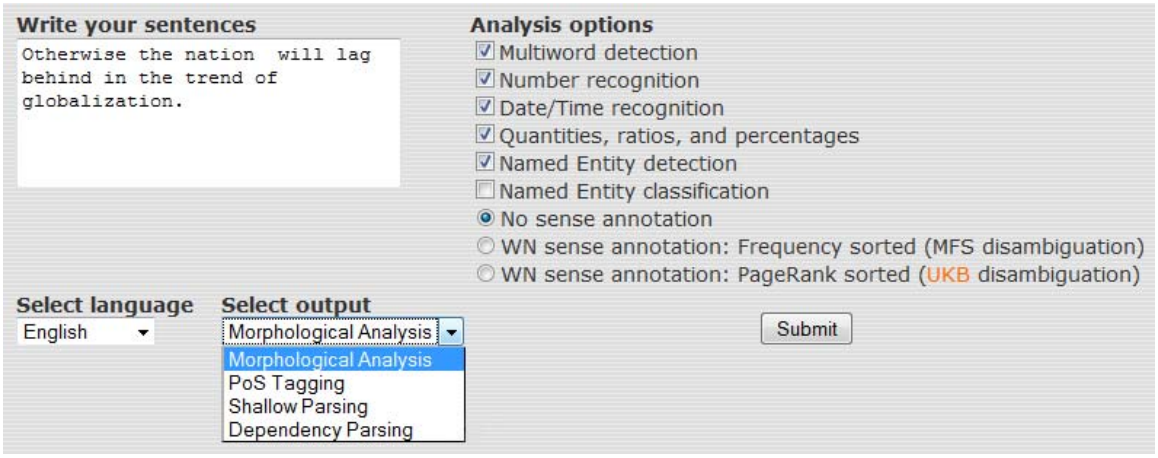

Figura 10: Opciones de análisis de Freeling Online.

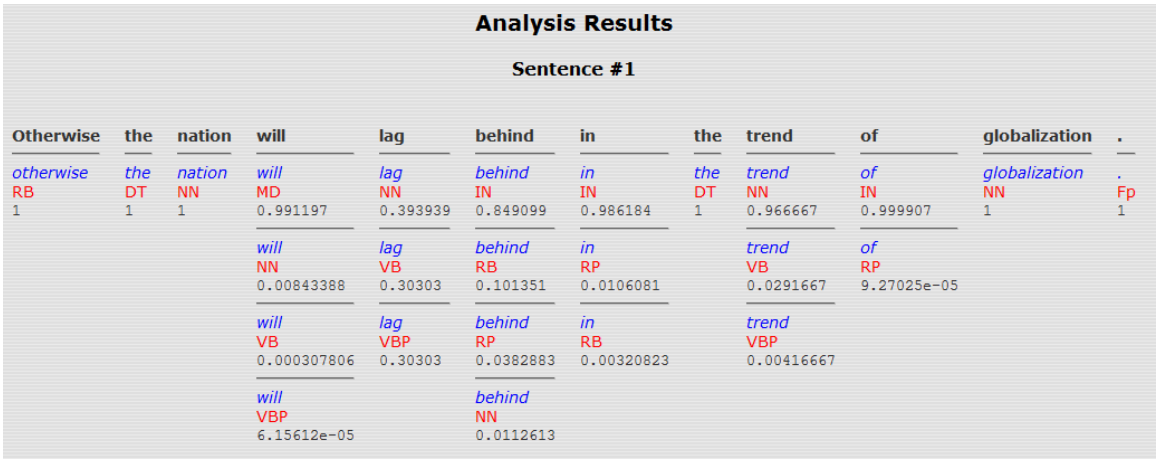

Figura 11: Resultado de análisis de Freeling Online.

[4] http://garraf.epsevg.upc.es/freeling/demo.php

Como se puede ver en la Figura 10, existen diversas opciones de configuración. Cuando se ejecute Freeling a través de la línea de comandos, será necesario especificar esta configuración mediante el fichero adecuado, aunque los resultados que aparecen de la configuración por defecto (ver Figura 11) proporcionan los datos que se van a necesitar, y por lo tanto no será necesario modificarla.

De estos resultados, se puede intuir que no toda la información proporcionada por Freeling será de utilidad en este sistema, por lo que habrá que procesar éstos resultados de forma adecuada. Para el ejemplo de la Figura 11, el fichero resultante contendrá la siguiente lista de palabras: *otherwise*, *the*, *nation*, *will*, *lag*, *behind*, *in*, *the*, *trend*, *of*, *globalization*.

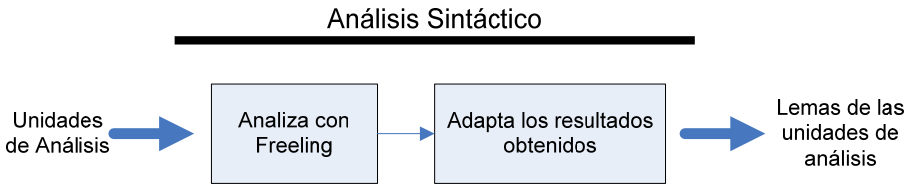

Figura 12: Diagrama de bloques de la fase de análisis sintáctico.

Esta fase de procesado adicional dará como resultado las unidades de análisis obtenidas tras la extracción de información pero en su forma más básica, es decir, los lemas. En la Figura 12 se puede ver el diagrama de bloques resultante al tener en cuenta las dos etapas descritas.

3.1.2.2 Eliminación de palabras de parada

Tras la fase de lematización, la etapa final del procesado inicial es la eliminación de las palabras de parada.

Las palabras de parada son una colección de palabras y caracteres que serán eliminados antes de la fase de análisis. Estas palabras son eliminadas porque no aportan información a los datos con los que se está trabajando, ya sea por la falta de significado semántico de las mismas o porque su presencia en el idioma que se está tratando es muy común y por tanto no aportan información.

No existe una lista de palabras de parada única para cada idioma, por lo que existe cierta libertad a la hora de seleccionar qué palabras se incluyen en esa lista. En este caso, a la lista de palabras estándar definida para el inglés se añadirá un versión de la palabra con la primera letra en mayúsculas, una tercera copia con todas las letras en mayúsculas, las cifras del 0 al 9, las letras del alfabeto y en el caso de que tengan, sus caracteres con distintos acentos (por ejemplo ä, è, etc.).

Una versión simplificada de la lista de palabras de parada utilizada se puede encontrar en el Anexo B – Lista de Palabras de Parada.

3.2 Matriz de Entrenamiento

Construir la matriz de entrenamiento es el primer paso y el más importante de la etapa de entrenamiento del clasificador de relevancia, ya que es la herramienta mediante la cual se van a representar todos los documentos. Precisamente por ser el método de representación de los documentos es un punto clave en el sistema y se ha decidido tratarlo como un apartado independiente al clasificador.

La matriz que se va a calcular representará vectorialmente el contenido de todos los documentos, dando una visión global de la aparición de los distintos términos en todo el conjunto de datos de cada uno de los temas.

El cálculo de la matriz supondrá la parte más importante de carga computacional del sistema debido al tamaño de los datos y por tanto al tamaño de las matrices generadas. Esta forma de representación de lo documentos según los términos que aparecen en los mismos se corresponde al Modelo de Espacio Vectorial [1]:

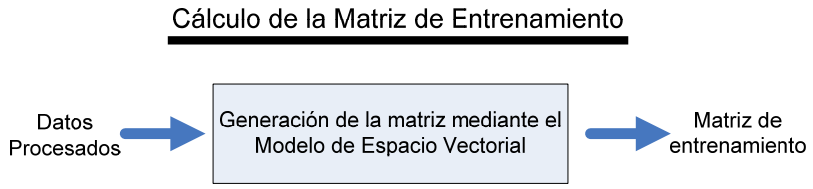

Figura 13: Diagrama de bloques del cálculo de la matriz de entrenamiento.

3.2.1 Modelo de Espacio Vectorial

Como ya se ha visto en apartados anteriores, el Modelo de Espacio Vectorial es uno de los modelos más extendidos y utilizados en recuperación de la información. Se basa en la representación de cada uno de los documentos en función de las palabras claves que contienen, de forma que se pueda saber con facilidad su presencia e importancia en los mismos para una consulta dada.

Estas características se obtendrán a partir del peso otorgado a cada término en cada documento siguiendo la siguiente expresión:

$$\omega_i = tf_i * \log\left(\frac{D}{df_i}\right) = tf_i * IDF_i$$ **ECUACIÓN 10**

En ella, tf_i representa la frecuencia del término en el documento y que por tanto representa la información local del mismo, D es el número de documentos con el que se está trabajando, y df_i es la frecuencia global del término, es decir, el número de documentos en los que aparece el término. Estos dos términos, cuando se dividen y se les aplica un logaritmo (como se puede ver en la segunda igualdad de la ecuación) forman la frecuencia inversa del documento, IDF_i (*Inverse Document Frequency*), la parte de la expresión que nos va a dar información global del término.

Para ver con más claridad tanto el cálculo de la matriz como los posteriores cálculos que serán realizados, se plantean los siguientes documentos y consulta como ejemplo [39]:

D_1 = "*Shipment of gold damaged in a fire*"

D_2 = "*Delivery of silver arrived in a silver truck*"

D_3 = "*Shipment of gold damaged in a truck*"

Q = "*Gold silver truck*"

De ellos, y suponiendo que son procesados como se ha descrito en el apartado de Análisis Sintáctico, se obtendría la siguiente tabla de términos, sobre los cuales se calculará los distintos factores de la ecuación de cálculo de pesos:

Tabla 1: Ejemplo de cálculo de la matriz de entrenamiento.

	tf_i			Cálculo IDF_i			Cálculo $w_i = tf_i * IDF_i$			
Términos	**Q**	**D_1**	**D_2**	**D_3**	**df_i**	**D/df_i**	**IDF_i**	**Q**	**D_1**	**D_2**	**D_3**
a	0	1	1	1	3	3/3 = 1	0	0	0	0	0
arrived	0	0	1	1	2	3/2 = 1,5	0,1761	0	0	0.1761	0,1761
damaged	0	1	0	0	1	3/1 = 3	0,4771	0	0,4771	0	0
delivery	0	0	1	0	1	3/1 = 3	0,4771	0	0	0,4771	0
fire	0	1	0	0	1	3/1 = 3	0,4771	0	0,4771	0	0
gold	1	1	0	1	2	3/2 = 1,5	0,1761	0,1761	0,1761	0	0,1761
in	0	1	1	1	3	3/3 = 1	0	0	0	0	0
of	0	1	1	1	3	3/3 = 1	0	0	0	0	0
silver	1	0	2	0	2	3/1 = 3	0,4771	0,4771	0	0.9542	0
shipment	0	1	0	1	2	3/2 = 1,5	0,1761	0	0,1761	0	0,1761
truck	1	0	1	1	2	3/2 = 1,5	0,1761	0,1761	0	0.1761	0,1761

De la primera columna se puede intuir que en algún punto del proceso será necesario crear un índice de términos del conjunto de documentos con los que se está trabajando. Tras tener ese índice, es fácil ver cómo en las siguientes cuatro columnas se representan con facilidad el contenido tanto de los documentos como de la consulta a modo de coordenadas en formato vector.

En las columnas correspondientes a "Cálculo IDF_i", se pueden ver los cálculos intermedios que se realizarán para poder calcular los pesos correspondientes tanto para los documentos como para la consulta. Estos pesos se pueden ver en las cuatro columnas de la derecha de la Tabla 1.

Un detalle importante a considerar es que en el ejemplo no se han eliminado las palabras de parada (*a*, *in*, *of*), aunque se puede ver cómo se les asigna pesos muy bajos. En este ejemplo las tres palabras de parada aparecen en todos los documentos por lo que siempre tienen peso cero, pero en un escenario con más documento, en el momento que apareciesen en todos menos en uno o dos documentos, el peso dejaría de ser nulo y por lo tanto añadiría ruido al sistema. Precisamente por esto, y como forma de mejorar los resultados del algoritmo se realiza como última fase del procesado inicial la eliminación de palabras de parada.

La matriz resultante con la que se trabajaría en este caso es la siguiente:

$$\begin{pmatrix} & a & arrived & damaged & delivery & fire & gold & in & of & silver & shipment & truck \\ Q & 0 & 0 & 0 & 0 & 0 & 0.1761 & 0 & 0 & 0.4771 & 0 & 0.1761 \\ D_1 & 0 & 0 & 0.4771 & 0 & 0.4771 & 0.1761 & 0 & 0 & 0 & 0.1761 & 0 \\ D_2 & 0 & 0.1761 & 0 & 0.4771 & 0 & 0 & 0 & 0 & 0.9542 & 0 & 0.1761 \\ D_3 & 0 & 0.1761 & 0 & 0 & 0 & 0.1761 & 0 & 0 & 0 & 0.1761 & 0.1761 \end{pmatrix}$$

Figura 14: Ejemplo de matriz de entrenamiento.

En lo siguientes apartados se verá cómo se va a utilizar esta matriz para el cálculo de la relevancia y cómo afectará al de opinión.

3.3 Clasificador de Relevancia

Seguramente una de las decisiones más importantes a la hora de estimar la relevancia de los distintos artículos es determinar a través de qué medida se va a decidir si un artículo es o no relevante al tema al que pertenece.

Como se ha visto en el apartado de la Matriz de Entrenamiento, representar estos textos matemáticamente es una práctica muy extendida en el NPL en general y en IR en particular. El uso del Modelo de Espacio Vectorial proporciona la gran ventaja de poder aplicar operaciones matemáticas a dichas representaciones.

La operación más habitual, sobre todo de cara al cálculo de la relevancia de documentos, es el cálculo de la similitud entre una consulta dada y una serie de documentos. Una vez calculado este valor, tendremos una medida para cada una de las unidades de análisis que permitirá entrenar un clasificador en función de ese valor. Estas distintas etapas de entrenamiento y clasificación se corresponden a las vistas para el modelo básico de clasificación de texto y son las que constituirán el clasificador de relevancia.

3.3.1 Similitud

Una vez elegida la medida de similitud como el valor mediante el cual se estimará si son o no relevantes cada uno de las unidades de análisis a una consulta dada, habrá que plantearse cómo calcular esta similitud.

Como se ha visto en otros apartados, no hay una única forma de calcular esta medida. Las dos más habituales son la distancia euclídea y la distancia del coseno aunque será esta última la que se utilizará en este sistema ya que proporciona resultados ligeramente mejores según investigación en el campo.

La expresión para el cálculo de esta similitud será pues la siguiente:

$$Sim(Q, D_i) = \cos(\theta_{D_i}) = \frac{\sum_i w_{Q,j} \cdot w_{i,j}}{\sqrt{\sum_j w_{Q,j}^2} \cdot \sqrt{\sum_i w_{i,j}^2}} = \frac{Q \bullet D_i}{|Q| * |D_i|}$$

ECUACIÓN 11

En dicha expresión, *Q* representa la consulta para cada uno de los temas y D_i cada uno de los documentos con los que se va a comparar para ver su relevancia (las unidades de análisis).

Siguiendo con el ejemplo que se ha visto para el cálculo de la matriz de entrenamiento (ver Tabla 1), se aplicará la ecuación vista de la similitud para comprobar a pequeña escala el tipo de resultados que se obtendrá aplicando esta medida.

El primer paso será el cálculo de los módulos que aparecen en el denominador. Se puede ver en la ecuación como de los dos términos que hay en el dominador, uno de ellos es el módulo de la consulta que se está haciendo (y por lo tanto será constante para las distintas unidades de análisis) mientras que el segundo es el módulo del documento con el que se está comparando esta última. En otras palabras, será necesario calcular tanto el módulo de la consulta como el de cada uno de los documentos.

Tomando como referencia la matriz de la Figura 14, cada uno de los módulos será la raíz de la suma de los cuadrados de los elementos de cada una de las filas:

$$|Q| = \sqrt{\left(0.1761^2 + 0.1761^2 + 0.4771^2\right)} = \sqrt{0.28964683} = 0.5382$$
$$|D_1| = \sqrt{\left(0.4771^2 + 0.4771^2 + 0.1761^2 + 0.1761^2\right)} = \sqrt{0.5173} = 0.7192$$
$$|D_2| = \sqrt{\left(0.1761^2 + 0.4771^2 + 0.4771^2 + 0.9542^2 + 0.1761^2\right)} = \sqrt{1.2001} = 1.0955$$
$$|D_3| = \sqrt{\left(0.1761^2 + 0.1761^2 + 0.1761^2 + 0.1761^2\right)} = \sqrt{0.1240} = 0.3522$$

Tras los módulos se calcula el producto escalar de la consulta con cada uno de los documentos:

$$Q \bullet D_1 = 0.1761^2 = 0.0310$$
$$Q \bullet D_2 = 0.4771 * 0.9542 + 0.1761^2 = 0.4862$$
$$Q \bullet D_3 = 0.1761^2 + 0.1761^2 = 0.0620$$

Una vez calculado tanto los productos escalares como los módulos, ya se puede calcular la similitud.

$$Sim(Q, D_1) = \cos(\theta_{D_1}) = \frac{0.0310}{0.5382 * 0.7192} = 0.0801$$
$$Sim(Q, D_2) = \cos(\theta_{D_2}) = \frac{0.4862}{0.5382 * 1.0955} = 0.8246$$
$$Sim(Q, D_3) = \cos(\theta_{D_3}) = \frac{0.0620}{0.5382 * 0.3522} = 0.3271$$

Que esté normalizado implica que todos los valores estarán entre 0 y 1, por lo tanto se considerará que los documentos con valores de similitud más próximos a 1 serán los más parecidos a la consulta.

Una vez calculada la similitud, se pueden ordenar los documentos de más a menos relevante, siendo los más relevantes los que mayor valor de similitud tienen. En este caso el documento más relevante es D_2, mientras que el menos relevante es D_1 por lo que el documento más relevante a la consulta "*Gold silver truck*" es "*Delivery of silver arrived in a silver truck.*"

Se puede ver claramente cómo a pesar de que en el documento D_3, al igual que en D_2, aparecen dos de las tres palabras buscadas, que una de las palabras, "*silver*", aparezca dos veces en D_2, afecta en gran medida al peso de ese término en el documento y por tanto al resultado de similitud. Es un indicador claro de como en relevancia, la frecuencia de aparición de los términos afecta al resultado.

Una vez se haya calculado el valor de similitud para todos los documentos, existirá una fase adicional de adaptación de formato. Es importante tener en cuenta que tras el cálculo de similitud es cuando se comienza a utilizar la separación lógica de conjuntos realizada, por lo facilitará la labor que los resultados obtenidos tengan un formato que tenga esto en cuenta.

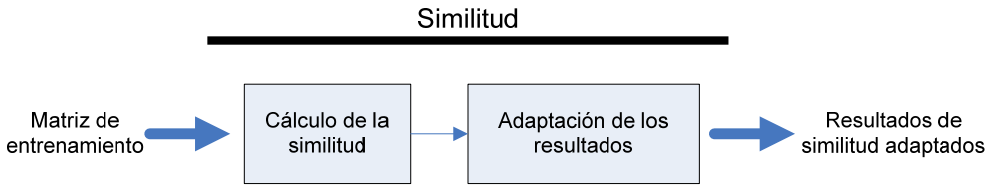

Figura 15: Diagrama de bloques del cálculo de la similitud.

En la Figura 15 queda reflejada esta necesidad de una segunda etapa de adaptación de los resultados.

3.3.2 Cálculo del Umbral

Una vez se haya calculado la similitud entre al consulta dada y cada una de las frases de cada uno de los documentos y esos resultados hayan sido adaptados, se pasará a calcular el umbral lineal únicamente a partir de los resultados obtenidos del conjunto de entrenamiento. Mediante este umbral se clasificarán los documentos pertenecientes a los conjuntos de pruebas y validación de forma que se pueda evaluar el rendimiento del mismo. Como ya se ha explicado anteriormente, se puede calcular el clasificador de esta forma gracias a tener las etiquetas para toda la colección de documentos.

También cabe recordar que tanto para el clasificador de relevancia como para el de opinión, las dos categorías a las que puede pertenecer una unidad de análisis van a ser exclusivas, de forma que un documento únicamente puede pertenecer a una de ellas (clasificación simple).

El umbral se va a calcular usando una media ponderada de las distintas categorías y sus centroides. Estos centroides son los representantes de las categorías y por lo tanto habrá uno por categoría (Relevantes o YES y No Relevantes o NO). Se calcularán mediante la media del valor de la medida que se está usando, es decir la similitud.

$$\eta_{relevancia} = \frac{\left(C_{relevantes} \cdot N_{relevantes} + C_{no\,Relevantes} \cdot N_{no\,Relevantes}\right)}{N_{total}}$$

ECUACIÓN 12

Donde C_i es el centroide de la categoría *i* y N_i el número de muestras total, de elementos relevantes o de elementos no relevantes. El haber introducido el número de muestras de las distintas clases hace que se tenga en cuenta la posibilidad de que las dos categorías no tengan un número parecido de elementos, y por lo tanto en ese caso, que la proximidad del umbral a un centroide u otro sea proporcional a ese número.

El asignar peso al centroide según el número de muestras asociadas a él va a tener un impacto significativo en el resultado. Las muestras de similitud nula no modifican el centroide de la categoría a la que pertenecen por lo que en los escenarios en los que el número de estas muestras no sea despreciable, calcular el umbral de una forma u otra tendrá un impacto importante.

Esto se ilustra claramente en el ejemplo representado en la Figura 16. En él se ha representado quince muestras, seis de ellas pertenecientes a la categoría YES y las otras nueve de la categoría NO.

Tabla 2: Valores Similitud.

Relevantes	No Relevantes
0,4	0
0,6	0,1
0,5	0,3
0,2	0
0	0,1
0,2	0,2
	0
	0
	0

Tabla 3: Formas de calcular el umbral.

	Relevantes	No Relevantes
Centroides	0,3167	0,1734
Umbral Media	0,19723	
Umbral Ponderado	0,1734	

Si representamos estos valores de similitud en un diagrama de dispersión y añadimos tanto los centroides como los dos umbrales calculados, se ve claramente cómo afecta a los resultados. Esto es lo que está representado en la Figura 16.

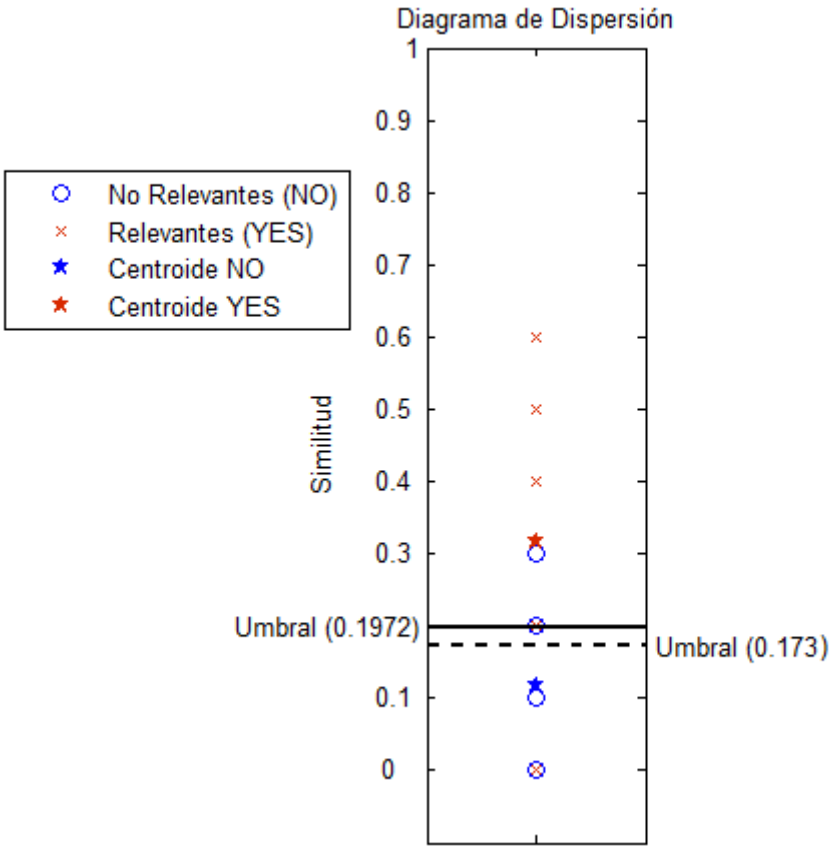

Figura 16: Ejemplo cálculo umbral.

En los datos con los que se va a trabajar habrá una cantidad bastante importante de unidades de análisis con valor de similitud cero, por lo que la diferencia entre usar un umbral u otro será muy significativa. Se usará siempre el umbral calculado a través de la media ponderada y además, por el significado asociado a la medida de la similitud, la interpretación del mismo será siempre la siguiente:

$$\begin{cases} S_i \geq \eta_{relevancia} \Rightarrow i \in YES \\ S_i < \eta_{relevancia} \Rightarrow i \in NO \end{cases}$$

Una vez calculado el umbral de decisión, ya se pueden evaluar los resultados, por lo que con esta última etapa se da por finalizado la fase de entrenamiento del clasificador de relevancia, con el siguiente diagrama de bloques resultante.

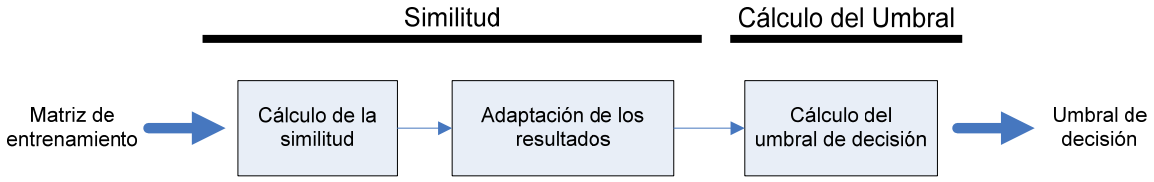

Figura 17: Diagrama de bloques del clasificador de relevancia.

3.3.3 Expansión de Términos

Durante las primeras evaluaciones del sistema realizadas, una de las fuentes de errores más significativas que se detectaron fue la cantidad de unidades de análisis con valor de similitud nulo. Como ya se ha visto, el valor de la similitud tiene como valor mínimo cero, por lo que independientemente del valor del umbral calculado, todas las unidades con similitud nula son categorizadas como no relevantes. Para intentar mitigar esto, se decide evaluar los efectos de realizar una expansión de términos en las consultas añadiendo al sistema una fase de realimentación.

Las expansiones de términos parten del concepto de ontología, es decir, la idea de crear módulos conceptuales donde palabras que pueden estar relacionadas con un mismo tema o dominio se agrupen, haciendo más sencillo obtener palabras relevantes a la hora de realizar una búsqueda [40]. Una de las formas de ontología más simples es la sinonimia, que da una relación semántica directa entre palabras.

Existen herramientas con este tipo de clasificaciones ya hechas, como es por ejemplo "*Wordnet*"[5]. La información que proporcionan estas herramientas es muy amplia y estudiar en condiciones su efecto sobre el cálculo de la relevancia y la opinión queda fuera del alcance de este proyecto. En su lugar, se utilizarán los datos de los que ya se disponen en el sistema para limitar la expansión a una relación directa contenida en el propio corpus y evitar tener que hacer consideraciones de adaptación de dominio.

Concretamente se hará uso de la similitud calculada para cada una de las unidades de análisis y de los pesos asociados a los distintos términos dentro de las matrices de entrenamiento asociadas a cada consulta.

Se considerará que los documentos determinados más relevantes a la consulta, tendrán palabras relevantes a la misma que no se han tenido en cuenta a la hora de calcular la similitud por no estar presentes en la consulta original. Por ello, y tomando como referencia el peso asociado a los términos en la matriz de entrenamiento, se generará una nueva consulta a la que se añadirán los N términos con mayor peso dentro de cada tema para los M documentos más relevantes del mismo. Tanto M como N serán parámetros configurables, mediante los que se evaluarán distintas configuraciones y así poder obtener la que da mejores resultados para el sistema.

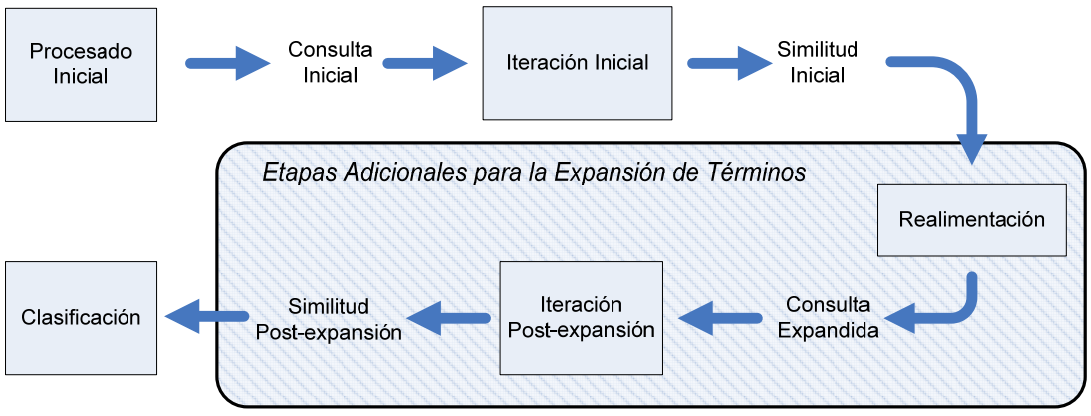

Figura 18: Esquema funcional del sistema con expansión de términos.

[5] http://wordnet.princeton.edu/

Se puede intuir que el número de veces que se puede extender la consulta es tan alto como se quiera, pero como se ha mencionado, el objetivo no es hacer un estudio detallado de los efectos de la expansión de términos, por lo que se hará una única expansión.

En la figura se puede ver cómo se integran las etapas necesarias para hacer esta expansión de términos dentro del esquema global del sistema (ver Figura 8). Se puede ver claramente cómo afecta solamente a la fase de entrenamiento.

En la figura se distinguen dos iteraciones distintas: la "Iteración Inicial" y la "Iteración Post-expansión". Estas dos etapas engloban las tareas vistas en el apartado de Matriz de Entrenamiento y el subapartado sobre Similitud dentro del Clasificador de Relevancia y su principal diferencia radica en la consulta que se recibe como entrada en cada una de ellas y por tanto, en el valor de similitud que generan. La forma de obtener la similitud será exactamente igual en ambas.

Dentro de la Figura 18, aparecen delimitadas las etapas que constituyen el diagrama de bloques de la fase de expansión de términos. En él se puede ver la etapa de "Iteración Post-expansión" de la que ya se ha hablado y una etapa previa de realimentación.

En esta etapa es donde se creará la nueva consulta ("Consulta Expandida" en el esquema de la Figura 18) con los términos elegidos. Como ya se ha dicho, el número de términos añadidos vendrá determinado por los dos parámetros configurables mencionados: M, que determinará los documentos más relevantes a la consulta inicial de los que se obtendrán términos y N, que será el número de términos de cada uno de esos documentos que se añadirán a la consulta original.

De la misma forma en la que se usa la similitud para seleccionar los documentos a usar, para decidir qué términos se van a añadir se utilizan los pesos calculados en la matriz de entrenamiento de la iteración inicial, seleccionando los términos según su peso.

A continuación se incluye un ejemplo de cómo afecta este proceso a una de las consultas.

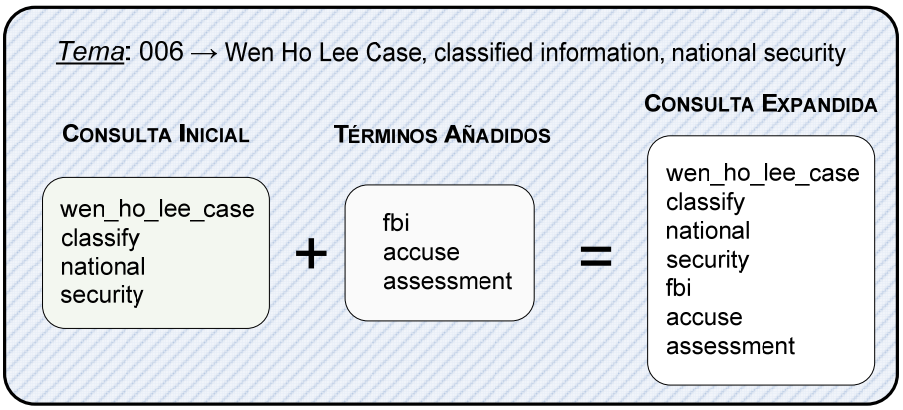

Figura 19: Ejemplo expansión de términos.

En la Figura 19 se ve la evolución que sigue la consulta del tema "006" durante el proceso de expansión. Junto al identificador del tema del que se ha obtenido el ejemplo, se ve el título asociado al mismo, uno de los campos a partir del cual se ha determinado que se obtendrá la consulta.

La consulta resultante tras realizar el procesado descrito en Procesado Inicial aparece como "Consulta Inicial", y junto a ella figuran los términos que se añadirán a la misma tras realimentar el sistema tomando del documento más relevante (M = 1) y los tres términos con mayor peso (N = 3).

Al haber añadido términos a la consulta, aumentan las probabilidades de que alguno de los documentos que inicialmente tenían similitud nula deje de tenerla por la presencia de nuevos términos.

3.4 Clasificador de Opinión

Al igual que en el clasificador de relevancia, el primer paso en el clasificador de opinión será determinar qué medida se utilizará para clasificar las unidades de análisis como unidades con o sin opinión.

La opinión es un concepto asociado al valor afectivo de las palabras, por lo que a diferencia de lo que sucedía con la relevancia, la mera presencia de palabras claves dentro de las unidades de análisis no bastará para determinar si tienen o no opinión. En su lugar se utilizará un diccionario afectivos, concretamente el "*General Inquirer*"[6] para identificar dichos valores afectivos.

La aplicación de este diccionario al escenario con el que se está trabajando supondrá añadir una etapa de procesado de texto adicional a la realizada en la fase de Procesado Inicial para preparar los datos para el cálculo de la opinión. Una vez se haya calculado la opinión, será necesario entrenar el clasificador. En este caso, en vez de usar únicamente un clasificador lineal como se hizo para la relevancia, el umbral de decisión se calculará también usando una etapa previa de agrupamiento mediante el algoritmo k-medias.

Todas estas fases están identificadas en el siguiente diagrama de bloques:

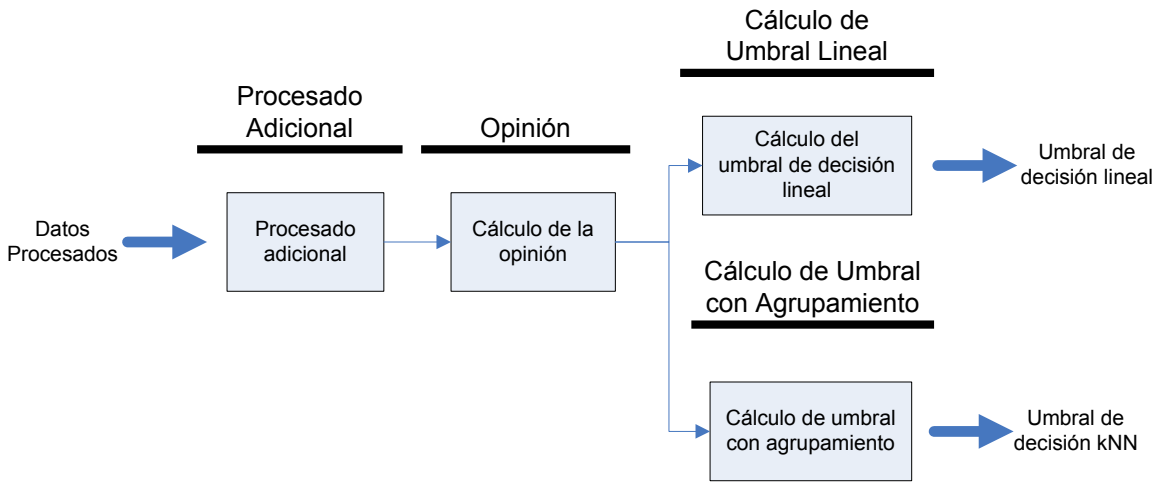

Figura 20: Diagrama de bloques del clasificador de opinión.

En los siguientes apartados se verá con más detalle cada uno de estos puntos y las implicaciones de cada uno de ellos.

3.4.1 Procesado Adicional

Como ya se ha mencionado, para poder dotar a las unidades de análisis del significado afectivo necesario para poder estimar si tienen o no opinión se va a realizar mediante el uso del diccionario semántico "*General Inquirer*" [41][42].

Este diccionario consta de 11895 entradas, cada una de ellas con las etiquetas que definen sus posibles valores semánticos según sus significados. Dentro de estas etiquetas, para el análisis

[6] http://www.wjh.harvard.edu/~inquirer/homecat.htm

que se va a realizar en este proyecto solo se tendrán en cuenta seis de ellas. La siguiente tabla es un resumen de las etiquetas que se van a considerar, su significado y el número de entradas etiquetadas en el diccionario con cada una de ellas. En la Tabla 5 se verán algunos ejemplos de palabras que tienen asignadas las etiquetas escogidas.

Tabla 4: Etiquetas de "*General Inquirer*" usadas.

Etiqueta	Significado	Número de entradas	Símbolo
Positiv	Palabras positivas	1915	P
Negativ	Palabras negativas	2291	N
Strong	Palabras que implican fuerza	1902	+
Weak	Palabras que implican debilidad	755	-
Yes	Palabras que indican directamente acuerdo, incluyendo expresiones[7]	20	+
No	Palabras que indican directamente desacuerdo	7	-

Tabla 5: Ejemplos de palabras con las etiquetas objetivo.

Etiqueta	Ejemplos
Positiv	*Able, ecstatic, prolific…*
Negativ	*Difficult, confess, foolish...*
Strong	*Powerful, bound, force, …*
Weak	*Light, meek, obsolete…*
Yes	*Right, mean (by all means), okay…*
No	*Disagree, wrong, long (no longer)…*

Se ve que el número de entradas utilizado será bastante menor que el total de entradas del diccionario original (4545 de 11895). Por este motivo y de cara a facilitar el proceso de asignación de las etiquetas a los datos procesados se añadirá una etapa que procese el diccionario.

El diccionario procesado se utilizará en la fase de asignación de etiquetas para filtrar los lemas que hay en cada una de las unidades de análisis de forma que queden únicamente las palabras que tengan alguno de los valores afectivos asociados a las etiquetas especificadas en la Tabla 4.

En esta tabla también se pueden ver los símbolos que van a representar los valores afectivos asociados a cada etiqueta. Cabe mencionar que de cara a la interpretación de estos valores

[7] Como por ejemplo la palabra "*course*", también incluida con la acepción que toma en la expresión "*of course*".

afectivos, se considerará que los valores asociados a las etiquetas "*Positiv*" y "*Negativ*", por ser independientes de su contexto, serán más fuertes que los del resto de etiquetas.

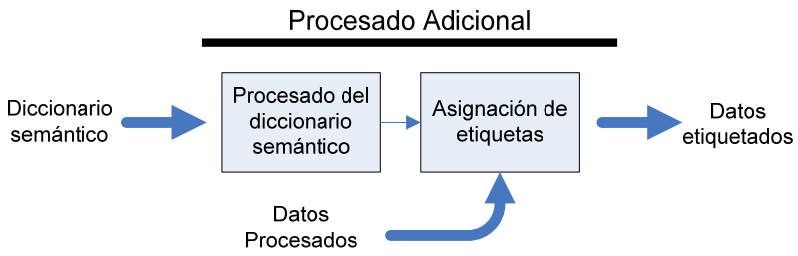

Figura 21: Diagrama de bloques del procesado adicional.

3.4.2 Opinión

Una vez se tenga un valor semántico asociado a los lemas resultantes para cada una de las unidades de análisis, hay que elegir cómo obtener un valor numérico a partir del cual se pueda calcular un umbral de decisión mediante el que clasificar las unidades de análisis.

Por lo tanto, la primera decisión que habrá que tomar será qué valores numéricos se asignarán a cada uno de los símbolos con los que se representa el valor semántico de cada lema. Tras ver los resultados obtenidos en otros escenarios se decide probar distintas configuraciones.

Las configuraciones elegidas se podrán dividir en dos tipos, dependiendo del número de valores que se puedan asignar. El primero de ellos habrá un valor por símbolo (o etiqueta de palabra) por lo que se trabajará con cuatro valores; el segundo tipo, en vez de asignar un valor por símbolo, asigna valores a las combinaciones más significativas, por lo que se trabajará con seis valores.

Dentro del primer tipo se evaluarán dos configuraciones que se diferenciarán en el valor que toman las etiquetas de palabra con valor semántico fuerte, es decir "P" y "N".

En las dos configuraciones se dará siempre un valor mayor a "P" y "N" respecto a las etiquetas "+" y "-" ya que como se ha mencionado previamente, su valor semántico es más independientemente del contexto en el que se encuentren.

En la Tabla 6 se pueden ver los cuatro valores que se asignan en las dos configuraciones del primer tipo.

Tabla 6: Configuraciones para cuatro valores (4tags).

Etiqueta de palabra	4tags (2)	4tags (3)
P	+2	+3
N	-2	-3
+	+1	+1
-	-1	-1

En esta misma tabla se pueden ver los distintos valores numéricos o pesos asignados en las dos configuraciones. El hecho de incrementar el valor máximo asignado en "*4tags (3)*" respecto a la otra configuración hará que el rango de posibles valores de opinión resultantes aumente, separando las muestras entre sí y por tanto facilitando la decisión alrededor de la frontera.

Asimismo, hay que tener en cuenta que la decisión de asignar un valor numérico positivo a la etiqueta asociada a la fuerza de un lema irá en detrimento de las palabras con valor semántico negativo.

En el segundo tipo de configuración, como se ha mencionado, se tendrán en cuenta seis combinaciones de etiquetas a la hora de asignar valores numéricos. Esta configuración no asigna un valor numérico a las etiquetas correspondientes a los valores semánticos débiles, sino que les da un valor según la etiqueta de valor semántico fuerte a la que acompañen, evitando así favorecer a uno de los valores fuertes como ocurre con las configuraciones del primer tipo.

Se puede ver con más claridad en la Tabla 8, concretamente en las columnas correspondientes a las dos configuraciones del primer tipo. En ellas, para las palabras que tienen valor semántico negativo, la asignación de valor numérico dependiendo de si tienen connotaciones de debilidad o fuerza no se hace correctamente, resultando en que una palabra fuerte negativa tenga un valor numérico mayor que una palabra débil negativa.

El asignar valores según combinaciones de etiquetas se puede plantear como asignar un valor fijo a las etiquetas “P” y “N” mientras que “+” y “-” tomarán un valor según el valor semántico fuerte al que acompañen. En la Tabla 7 se puede ver cómo se aplicaría esto, y las etiquetas resultantes para las seis combinaciones que se van a contemplar.

En la columna correspondiente a esta configuración de seis valores de la Tabla 8 se aprecia claramente cómo se obtienen resultados más coherentes para las palabras con valor semántico negativo.

Tabla 7: Configuraciones para seis etiquetas (6tags).

Etiqueta de palabra	Valor Fijo	Valor Variable	Valor Total
P+	+2	+1	+3
P	+2	0	+2
P-	+2	-1	+1
N-	-2	+1	-1
N	-2	0	-2
N+	-2	-1	-3

Tabla 8: Ejemplo configuraciones del cálculo de opinión.

Palabra Ejemplo	Etiquetas	4tags (2)	4tags (3)	6tags
Rival	N+	**-1**	**-2**	**-3**
Difficult	N	**-2**	**-3**	**-2**
Delay	N-	**-3**	**-4**	**-1**
Modest	P-	1	2	1
Offer	P	2	3	2
Major	P+	3	4	3

Hay que tener en cuenta que los valores negativos van a ser sólo una parte de la muestra de unidades de análisis, por lo que será interesante ver los resultados obtenidos con los distintos métodos.

Una vez se haya asignado un valor numérico, w_{word}, acorde al valor semántico asignado en la etapa de Procesado Adicional a cada una de las palabras que componen las unidades de análisis, se calculará la opinión de la unidad, w_{ud}, realizando para cada una de ellas la media de dichos valores numéricos respecto al número de palabras con valor semántico en la unidad, *N*.

$$w_{ud} = \frac{1}{N} \cdot \sum_{N} w_{word}$$ **ECUACIÓN 13**

Al igual que ocurre cuando se calcula la similitud, será necesario añadir una etapa posterior a este cálculo para adaptar los resultados obtenidos a un formato que pueda ser utilizado fácilmente por las siguientes etapas.

Esta nueva etapa está reflejada en el diagrama de bloques de la Figura 22. Cabe resaltar que estos resultados estarán ya clasificados según los distintos subconjuntos con los que se va a trabajar.

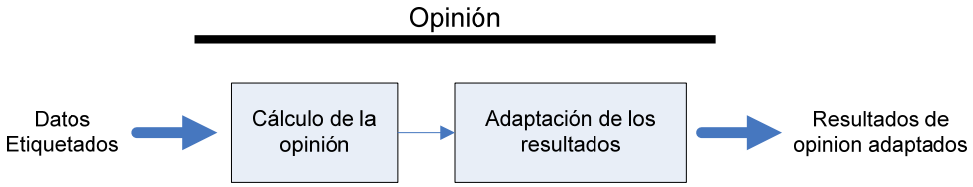

Figura 22: Diagrama de bloques del cálculo de la opinión.

3.4.3 Cálculo del Umbral

Para estudiar la clasificación según la opinión de las unidades de análisis se usarán dos métodos distintos. El primero de ellos y el más sencillo, será el un clasificador lineal análogo al utilizado para la clasificación de la relevancia; en el segundo, se añadirá una etapa de agrupamiento previa al cálculo del umbral de clasificación.

3.4.3.1 Umbral Lineal

Como se acaba de mencionar, el cálculo de este umbral lineal va a ser análogo al Cálculo del Umbral que se ha visto para el clasificador de relevancia.

Se partirá únicamente de los resultados obtenidos para el conjunto de entrenamiento y, a través de una media ponderada de las distintas categorías (en este caso documentos con opinión o YES y documentos sin opinión o NO), se calculará el umbral de decisión. En este caso, la medida que se utilizará es la opinión que se ha calculado para cada una de las unidades de análisis.

La expresión para calcular el umbral es la siguiente:

$$\eta_{opinión} = \frac{\left(C_{conOpinión} \cdot N_{conOpinión} + C_{\text{sin } Opinión} \cdot N_{\text{sin } Opinión}\right)}{N_{total}}$$ **ECUACIÓN 14**

Donde C_i es el centroide la categoría *i* y N_i el número de muestras total, de elementos con opinión o de elementos sin ella. El centroide de cada categoría se calculará haciendo la media de la suma de los valores de opinión de todas las muestras pertenecientes a dicha categoría.

Las consideraciones que se hicieron en el cálculo del umbral de relevancia sobre la asignación de pesos igualitaria para evitar que una categoría tenga más peso que otra en el cálculo del umbral se aplican exactamente igual para el umbral de opinión.

Asimismo, por la naturaleza de los valores numéricos que se ha asignado a cada una de las etiquetas según tuviesen valor semántico positivo o negativo, la interpretación del umbral será siempre:

$$\begin{cases} |O|_i \geq \eta_{opninión} \Rightarrow i \in YES \\ resto \Rightarrow i \in NO \end{cases}$$

3.4.3.2 Umbral Lineal con Etapa de Agrupamiento

El segundo método de clasificación de la opinión va a ser añadiendo una etapa de agrupamiento antes de realizar la clasificación.

El añadir esta etapa al cálculo del umbral de decisión, reducirá el coste computacional. Al disponer de una medida de proximidad para todos los documentos del conjunto de entrenamiento, la implementación de esta etapa adicional de agrupamiento es muy sencilla y poco costosa.

El algoritmo a utilizar será una versión modificada de k-medias, ya que no se usará la distancia euclídea sino la medida de similitud calculada para la clasificación de la relevancia. Hay que recordar que esta similitud se calculaba a partir de la distancia coseno.

Se considerará que las k unidades de análisis más relevantes para cada uno de los temas del corpus de entrenamiento serán representantes adecuados de cada uno de ellos, y por tanto calcular el umbral de decisión se limitará a procesar k muestras por cada tema en lugar de todas las muestras del conjunto de entrenamiento.

De esas k muestras para cada tema, se calculará un valor de opinión representante de las mismas (un centroide) que será promediado con el resto de representantes del resto de los temas para calcular un valor global que constituirá el umbral de decisión de opinión.

$$\eta_{opinion} = \frac{1}{N_T} \cdot \left(N_{conOp} \cdot \frac{1}{k} \sum_k w_{word} \Big|_{conOp} + N_{\sin Op} \cdot \frac{1}{k} \sum_k w_{word} \Big|_{\sin Op} \right)$$

ECUACIÓN 15

Donde N_T es el número total de muestras, N_{conOp} es el número de muestras con opinión y N_{sinOp} es el número de muestras sin opinión.

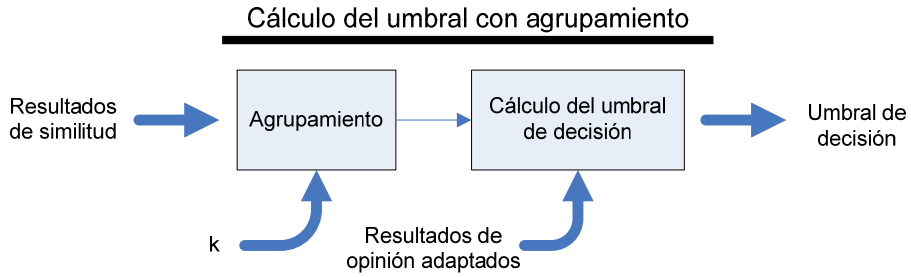

Figura 23: Diagrama de bloques del cálculo del umbral con agrupamiento.

El hecho de utilizar un algoritmo que requiere un parámetro de configuración como es k, hará que haya que analizar los distintos valores posibles de cara a escoger el que dé los mejores resultados.

Al igual que en el cálculo del umbral de decisión sin etapa de agrupamiento, la medida que se está utilizando para calcular la opinión tiene un significado intrínseco que hace que el sentido de la decisión sea siempre el siguiente:

$$\begin{cases} |O|_i \geq \eta_{opninión_K} \Rightarrow i \in YES \\ resto \Rightarrow i \in NO \end{cases}$$

Cabe destacar que al utilizar la medida de similitud para la selección de las k unidades de análisis a utilizar para cada tema, influirá en los resultados obtenidos si se utiliza o no expansión de términos para el cálculo de la similitud.

3.5 Resumen

Al comienzo de este capítulo se mostró un diagrama de bloques genérico de las principales etapas que se iban a encontrar en este sistema. A lo largo de los distintos apartados, se han ido analizando cada una de las etapas identificadas, explicando cómo se realizará cada una de ellas e identificando las distintas fases que implicarán.

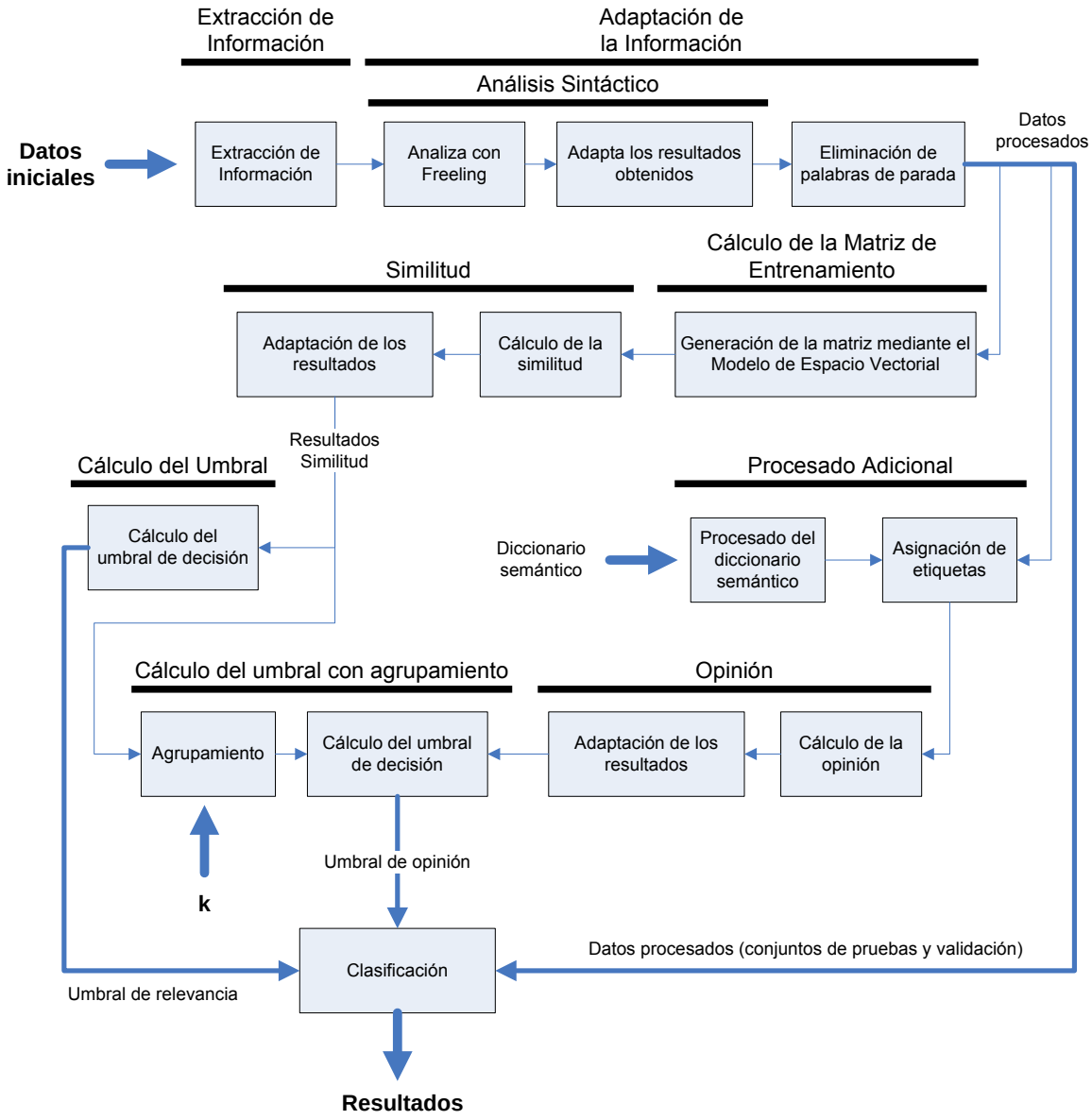

Figura 24: Diagrama de bloques detallado del sistema.

Para cada una de estas etapas principales se han incluido diagramas de bloques en los que se puede ver de forma general las principales tareas a llevar a cabo, las entradas que tendrá el sistema en cada etapa y las salidas que producirán.

En la Figura 24 se ha incluido el resultado de enlazar todos estos diagramas de bloques a bajo nivel que se han ido viendo, dando lugar a un esquema bastante fiel a lo que luego será la implementación del sistema.

En la figura, se han identificado tanto las distintas tareas, como las etapas a las que pertenecen. Asimismo, las entradas más relevantes del sistema aparecen claramente marcadas, así como los datos necesarios para realizar la clasificación final sobre los conjuntos de pruebas y validación que son los que finalmente darán el rendimiento del sistema.

Cabe destacar que para el cálculo de opinión se ha incluido la opción en la que se usa una etapa de agrupamiento, ya que se puede considerar que no usarla equivale a usar una etapa de agrupamiento con cero como número de elementos a agrupar, y por tanto el esquema es fácilmente adaptable a este escenario.

Asimismo, tampoco se ha tenido en cuenta la variante del sistema en la que se realiza una expansión de términos, ya que un esquema detallado no aporta más información que la ya facilitada en el esquema funcional con expansión de términos incluido en la Figura 18.

4 Implementación del Sistema

A continuación se explicará en detalle cómo se ha llevado a cabo la implementación de cada una de las etapas de las que se ha hablado en el apartado anterior. La estructura a seguir será prácticamente la misma en cada uno de los apartados: se verá con detalle los distintos ficheros de código escritos para llevar a cabo las distintas fases del sistema y las particularidades asociadas a cada una de las tareas realizadas.

Todo el código escrito para este proyecto se ha realizado en PHP, y se ha seleccionado, aparte de por su sencillez como lenguaje de programación, por las posibilidades que da a la hora de trabajar con cadenas de texto; PHP es un lenguaje orientado al manejo de páginas web dinámicas, por lo que la manipulación y presentación de texto es un aspecto básico del mismo.

Para cada uno de los programas, siempre que proceda se incluirá un diagrama de flujo que describa en líneas generales los pasos dados para implementar su funcionalidad y un esquema de las entradas y salidas del mismo.

Debido al volumen de datos que se maneja y las numerosas subcarpetas que se generarán durante el procesado, el esquema de las entradas y salidas ayudará a ver con más claridad qué datos se están usando en cada fase del proceso y a identificar las dependencias entre las distintas tareas.

4.1 Procesado de los Datos

Antes de proceder a explicar en detalle cómo se ha implementado el módulo del procesado de datos, es importante saber cuáles son los datos de entrada del sistema por lo que lo primero que se verá es una descripción de los mismos. Una vez se haya visto la entrada de la que se parte, se pasará a explicar en detalle cómo se han implementado las distintas fases que se identificaron para el procesado de datos en el correspondiente apartado de diseño.

A la hora de llevar a cabo las distintas etapas del proceso, el resultado de cada una de ellas se irá guardando en distintas carpetas de forma que en cualquier punto del análisis se disponga de los pasos intermedios del procesado.

Como se ha mencionado, esto junto a la cantidad de datos con la que se trabaja, supondrá una estructura de carpetas compleja, por lo que en cada una de las etapas de la implementación se especificará los cambios que se van viendo en la estructura base de carpetas.

Además, la inclusión de estas estructuras facilitará que en todo momento sea fácil localizar las entradas y salidas a las que se hace referencia.

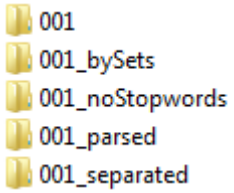

Figura 25: Estructura de carpetas tras procesado de datos.

Tras la fase de procesado de datos se pasará de tener una única carpeta por tema con los artículos relativos al mismo a tener la estructura de carpetas mostrada en la Figura 25 y particularizada para la consulta obtenida del tema con identificador “001”.

En los datos de entrada del sistema se distinguen dos secciones: los artículos que se quieren analizar (ubicados en la carpeta “NTCIR6-Opinion-EN”) y los temas con los que están relacionados (en la carpeta “NTCIR6-Topics”). Ya se ha visto anteriormente la composición de cada uno de ellos, por lo que a continuación se pasará a ver en detalle cómo se ha implementado el procesado a realizar sobre ellos.

4.1.1 Extracción de Información

Ya se ha mencionado que no va a ser utilizada toda la información proporcionada tanto para los temas como para los artículos, por lo tanto el primer paso dentro del procesado de datos es extraer la información relevante para el análisis que se va a llevar a cabo.

En el caso de los temas, a partir de cada uno de ellos se generará un archivo que tendrá por nombre el número identificador del tema y que contendrá la consulta asociada. Como ya se ha explicado previamente, se van a evaluar distintos tipos de consultas, por lo que ésta podrá ser la descripción, el título o una combinación de ambos.

Para los artículos, cada frase se convertirá en un fichero cuyo nombre estará compuesto del nombre del artículo seguido del identificador de la frase. De esta forma se conseguirá que cada uno de los ficheros resultantes tenga un nombre único en todo el conjunto de los datos. Asimismo, la triada formada por el identificador de tema, el nombre del artículo y el identificador de frase permitirá identificar unívocamente cualquier unidad de análisis en todo el sistema.

Como el procesado de los dos tipos de datos es distinto, se harán de forma independiente, cada uno de ellos a través de un programa distinto. Los dos programas usados son:

- *process_docs.php*
- *process_topics.php*

Ambos ficheros tiene estructuras muy similares (ver Figura 26 y Figura 27), ya que la funcionalidad básica de ambos es buscar en ficheros de texto las cadenas que delimitan cada uno de los campos y obtener así los datos de los que obtener los campos que se han determinado necesarios para el análisis que se va a realizar.

Como ya se ha mencionado, es importante tener en cuenta que se va a manejar un volumen de información muy alto, por lo que es vital tener una estructura de datos que lo facilite. Con este propósito, cada paso de este procesado inicial, tanto de los datos como de los temas será almacenado en carpetas diferentes de forma que una vez finalizado todo el proceso se puedan ver los resultados intermedios de cada uno de los pasos que se han dado.

Inicialmente, los artículos pertenecientes a cada uno de los temas estarán en una carpeta nombrada con el identificador numérico del tema, cada uno de ellos en archivos independientes. Esta colección de carpetas, va a ser la entrada de *process_docs.php*. La salida será volcada en una carpeta cuyo nombre será el identificador numérico del tema al que pertenecen los artículos, seguido de la cadena "_separated".

En esta carpeta se encontrará un archivo por cada una de las frases de todos los artículos pertenecientes al tema.

Ya se ha explicado en la descripción de los datos de entrada del sistema como cada una de las frases viene precedida de su identificador local (único solo en el artículo), que determina el identificador de la frase.

El programa *process_docs.php* se basa en la búsqueda de esa cadena para generar cada uno de los archivos y asignarles el correspondiente identificador. De la posición de la etiqueta (<STNO> de apertura y </STNO> de cierre) y del hecho que es un dato conocido que los identificadores de frase van a ser siempre de longitud fija (cuatro cifras), se obtendrá, tanto la frase como el identificador de la misma.

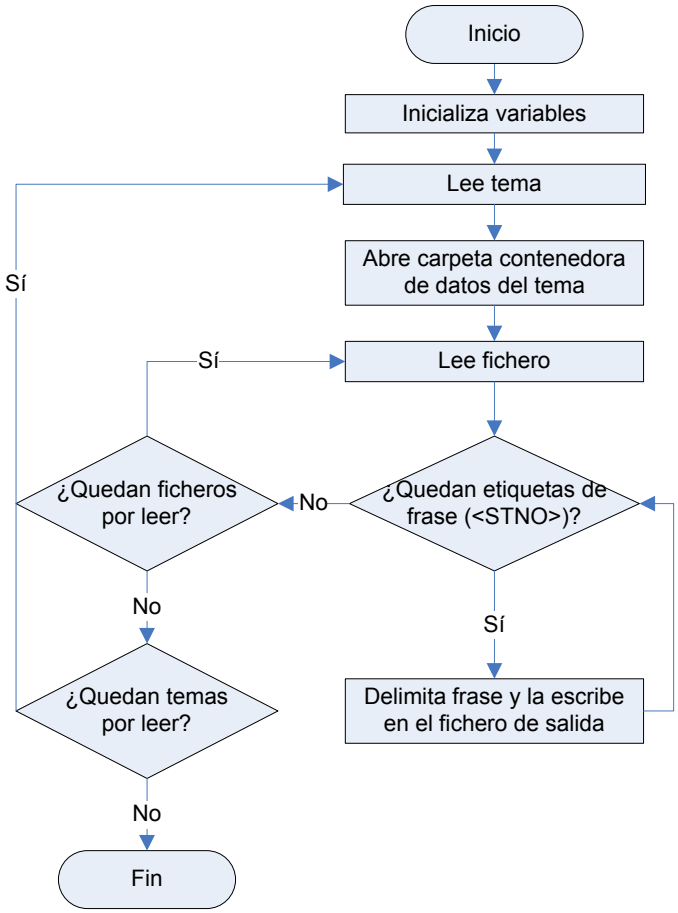

Figura 26: Diagrama de flujo de *process_docs.php*

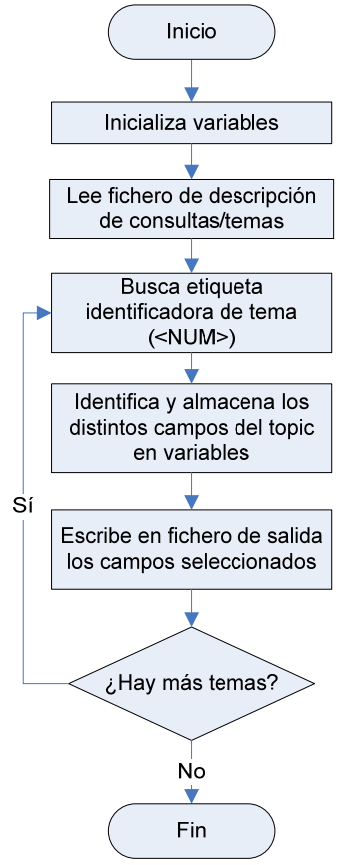

Figura 27: Diagrama de flujo de *process_topics.php*

Cada una de las frases se volcará en un fichero en cuyo nombre figurará el nombre del artículo al que pertenece y su identificador dentro del mismo, haciendo la combinación unívoca en todo el corpus.

La Figura 28 muestra un ejemplo de la nomenclatura descrita tomado directamente de los datos, seleccionando como ejemplo el tema “001” y uno de los artículos del mismo.

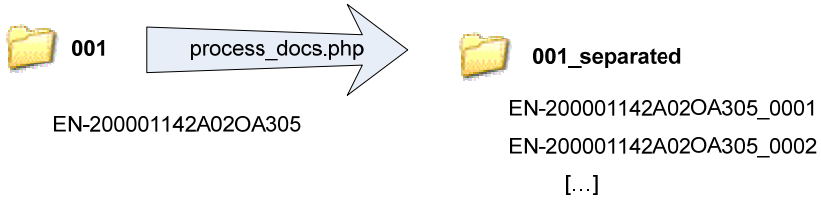

Figura 28: E/S *process_docs.php*.

A diferencia de los artículos, toda la información perteneciente a los temas se encuentra en un único fichero (el ya mencionado “NTCIR6OpinionTopics_EN.txt” incluido en el Anexo A – Temas de NTCIR-6), por lo que *procesa_topics.php* tiene una parte mucho menor dedicada a la lectura de ficheros y directorios. Además, proporcionalmente, se va a seleccionar bastante menos información sobre cada uno de los temas que en los artículos, donde se extrae prácticamente toda la información de cada fichero.

Como se ve en la Figura 27, el primer paso es identificar cada uno de los temas para poder extraer la información necesaria de cada uno de ellos. En el formato de entrada, existe la etiqueta <TOPIC> que delimita la información de cada tema, pero como los distintos datos que van a ser extraídos tienen etiquetas individuales, serán estas etiquetas las que serán buscadas.

Es importante comprobar que la etiqueta que se va a usar para determinar el número de temas del sistema esté presente en todos ellos, o en otras palabras, que el número de etiquetas <TOPIC> sea igual al de la etiqueta escogida puesto que si no lo fuese se perdería información.

El primer campo que va a ser utilizado, será el que determina el identificador de los temas, la etiqueta <NUM>, y mediante la que se podrá obtener el número de tres cifras que identificará unívocamente a cada tema.

Una vez ha sido identificado el tema, se buscarán los campos del mismo que se van a usar para definir la consulta: el título y la descripción. El título viene determinado por la etiqueta <TITLE> y la descripción por la etiqueta <DESC>. Ambas se guardarán en variables independientes para poder disponer de las dos a la hora de elegir los datos de salida con los que se va a trabajar.

Como ya se ha mencionado, la entrada de este programa consiste en un único archivo de texto con la información de todos los temas. La salida va a ser una única carpeta llamada “000”, donde se guardará un archivo por tema encontrado en el fichero de entrada. Cada uno de esos ficheros se nombrará con el identificador del tema correspondiente y contendrá la información que se elija como consulta asociada al tema.

En la fase de evaluación se estudiarán tres posibilidades: usar solamente el título, usar solamente la descripción, o ambos.

En la Figura 29 se puede ver, siguiendo el mismo formato que en figuras anteriores, la entrada del sistema y la salida, tomando valores reales como ejemplo. En la salida se puede ver la carpeta donde se guardan cada uno de los ficheros, los ficheros y el contenido de dos de ellos.

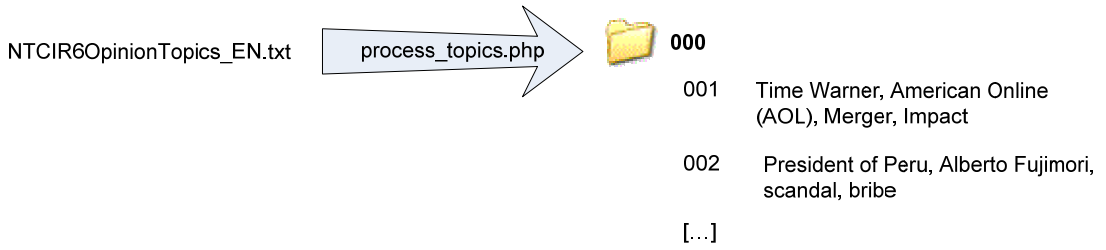

Figura 29: E/S *process_topics.php*.

4.1.2 Adaptación de la Información

Una vez extraída la información que se va a utilizar, se pasa a tratarla sintácticamente. Este análisis se dividirá en dos etapas: el análisis sintáctico y la eliminación de palabras de parada. Ambas etapas serán comunes tanto para el tratamiento de los artículos como para las consultas definidas para cada tema.

Como ya se ha explicado anteriormente, para la primera etapa, el análisis sintáctico, se usará el analizador Freeling [38] tanto para el análisis de los artículos como para los temas.

Los ficheros usados para esta parte son los siguientes:

- *parse_sentences.php*
- *parse_topics.php*
- *getLemma.php*

Al igual que ocurría en el apartado anterior, los ficheros *parse_sentences.php* y *parse_topics.php* van a ser muy parecidos en estructura ya que ambos realizarán la misma operación, uno para las frases de cada uno de los artículos y el otro para las que definen cada una de las consultas asociada a cada temas (ver Figura 30 y Figura 31).

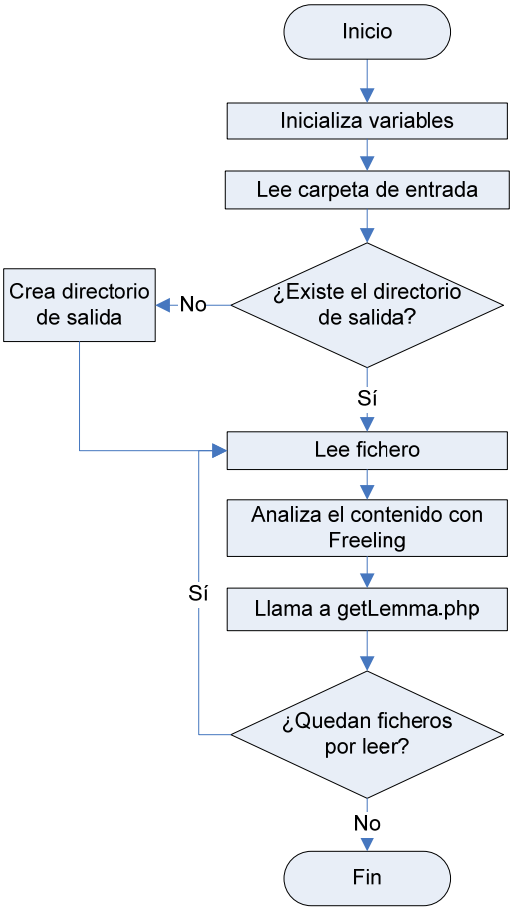

Figura 30: Diagrama de flujo de *parse_sentences.php*.

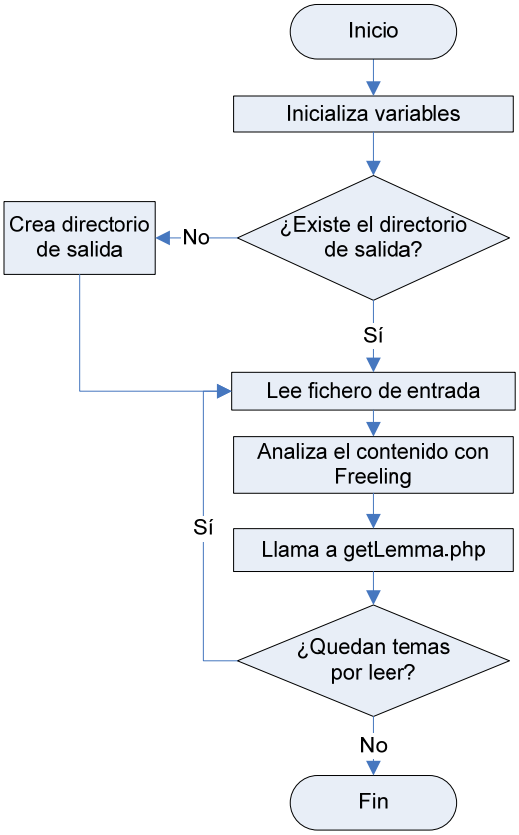

Figura 31: Diagrama de flujo de *parse_topics.php*.

El programa *getLemma.php* es mediante el cual se procesará la información obtenida de Freeling, por lo que se llamará tanto desde *parse_sentences.php* como desde *parse_topics.php* inmediatamente después de haber usado Freeling.

Freeling usa la entrada estándar y será de esa forma como se le pasará la frase a analizar. De forma análoga, la salida usada para mostrar los resultados del análisis es también la estándar por lo que se usará este detalle en la implementación de *getLemma.php* para evitar tener que guardar la información en un fichero antes de procesarla.

El programa *parse_sentences.php,* tratará el análisis de las frases en las que se ha dividido cada uno de los artículos y por lo tanto va a recorrer cada una de las carpetas creadas en la fase de extracción de información. La salida del programa se guardará en carpetas llamadas con el identificador numérico del tema al que pertenece la frase seguido de la cadena "_parsed".

Dentro de cada carpeta, se encontrará un fichero por cada una de las frases del artículo, llamado exactamente igual que el fichero correspondiente de entrada. En cuanto a contenido se refiere, en el fichero de salida se encontrarán, los lemas o parte básica de cada una de las palabras de la frase de entrada.

En la Figura 32 se puede ver el resultado del análisis sintáctico de la primera frase del artículo "EN-200001142A02OA305" del tema "001".

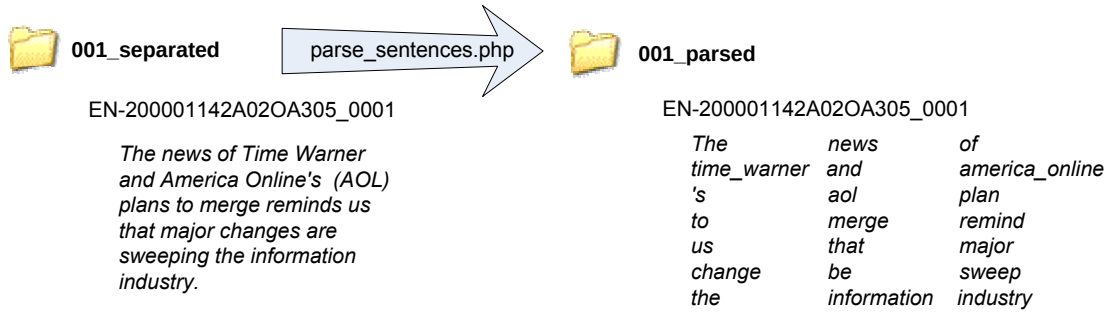

Figura 32: E/S *parse_sentences.php*.

Por otra parte, *parse_topics.php* va a ser muy parecido a *parse_sentences.php*. Leerá cada uno de los ficheros que se encuentran en la carpeta de consultas "000", en donde se guardó la información de cada uno de los temas. El contenido de cada uno de esos ficheros, la consulta, es lo que se le pasará de entrada a Freeling y al igual que en *parse_sentences.php*, la salida se guardará en un fichero llamado igual que el de entrada pero en un carpeta denominada "000_parsed".

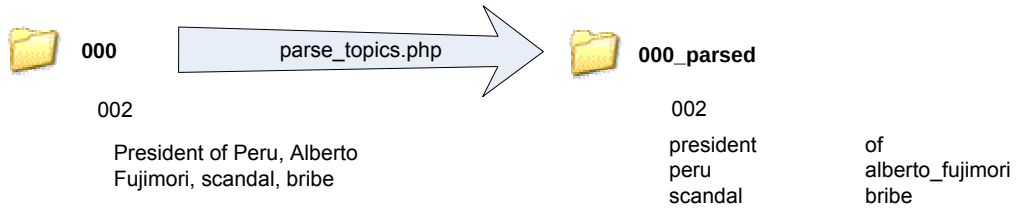

Figura 33: E/S *parse_topics.php*.

Tanto *parse_sentences.php* como *parse_topics.php* realizaran la llamada a Freeling que devolverá el análisis sintáctico de la frase seleccionada como entrada. Ya se ha mencionado que la forma de pasar los datos a Freeling será a través de la entrada estándar por lo que justo antes de invocar Freeling, se volcará en la salida estándar el contenido del fichero donde se encuentra la frase.

La llamada que se encuentra en el código (definida como una cadena) es la siguiente:

```
"cat ".$inFile."|analyze -f /home/bgalego/tools/en_ntcir.cfg | php
".$TOOLS."getLema.php>".$outFile
```

En ella se pueden observar tres partes distintas:

- *"cat ".$inFile*: que vuelca el contenido del archivo definido por *$inFile* en la salida estándar del sistema.
- *analyze –f /home/bgalego/tools/en_ntcir.cfg*: llamada a Freeling usando como entrada la estándar del sistema y como parámetro el fichero de configuración adecuado. El fichero de configuración, "en_config.cfg" es el fichero de configuración por defecto para inglés y en él se pueden especificar las distintas opciones de análisis de Freeling.
- *php getLemma.php > $outFile*: donde se procesan los resultados del análisis hecho por Freeling.

Para poder ver con más detalle exactamente qué procesado se realiza en *getLemma.php*, primero se verá un ejemplo del análisis sintáctico que proporciona Freeling. En concreto se usará la frase "0002" del artículo "EN-20000114202OA305" perteneciente al tema "001":

```
But this deal certainly does not come as a surprise to those who work in
this field.
```

Tras pasar esta frase a Freeling como entrada, el resultado obtenido es el siguiente:

```
But but CC 0.996024
this this DT 0.999415
deal deal NN 0.861219
certainly certainly RB 1
does do VBZ 1
not not RB 1
come come VB 0.610627
as as IN 0.835009
a 1 Z 0.99998
surprise surprise NN 0.868056
to to TO 1
those those DT 0.999297
who who WP 1
work work NN 0.647878
in in IN 0.986184
this this DT 0.999415
field field NN 0.959707
. . Fp 1
```

Como ya se ha explicado, la única parte que va a ser utilizada va a ser el lema, incluido en los resultados tras la palabra original. El programa *getLemma.php* lo que hará es extraer el lema, realizar un filtrado inicial sobre el mismo, y finalmente guardarlo en el fichero especificado (ver Figura 34).

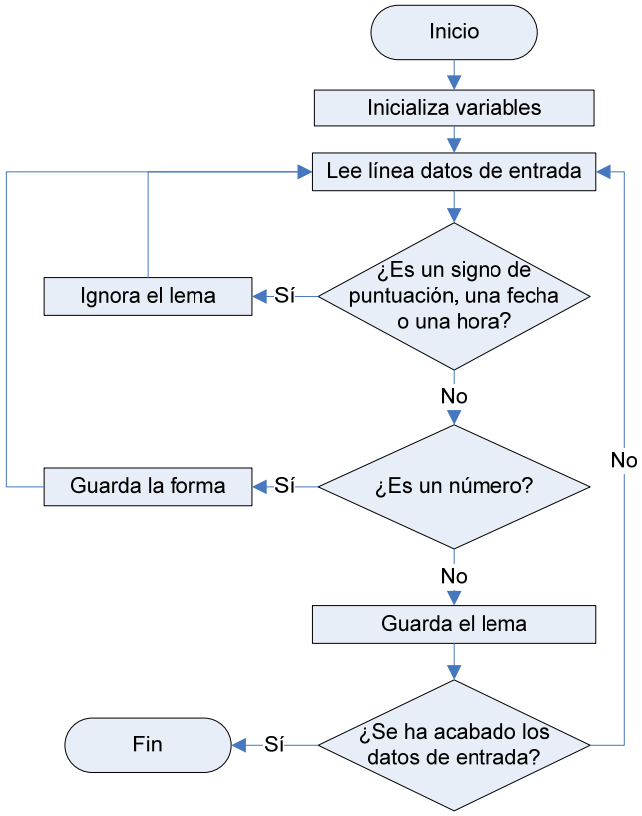

Figura 34: Diagrama de flujo de *getLemma.php*.

El filtrado que se hace sobre los lemas hace que no se guarden ni los signos de puntuación (categoría F) ni las fechas y horas (ambas categoría W). Asimismo, se asegurará que en el caso de los números (categoría Z) se guarde la palabra original en vez del lema ya que ésta se corresponde a la forma escrita del número y no a la cifra.

Una vez finalizada la fase de extracción de los lemas, se hará un último filtrado en el que se eliminarán las palabras de parada (ver Eliminación de palabras de parada), la segunda etapa descrita dentro de este apartado.

Como se ha venido haciendo hasta ahora, se usarán dos programas distintos, uno para las consultas y otro para la información de los artículos:

- *removeStopwords.php*
- *removeStopwordsTopics.php*

De nuevo parte de la diferencia entre ambos programas va a estar en la distinta estructura de los datos de entrada que se le pasa a cada uno de ellos.

En los dos programas la entrada va a ser el resultado del análisis sintáctico, en el caso de los temas, guardado en la carpeta “000_parsed” y para los artículos en las carpetas llamadas “tID_parsed” donde “tID” es el identificador del tema al que pertenece el artículo. La notación de la carpeta donde se van a guardar los datos de salida en los dos casos va a ser análoga a la que se viene utilizando: para los temas la salida se guardará en “000_noStopwords” y para los artículos, para cada tema, en carpetas llamadas “tID_noStopwords”.

En la Figura 35 y en la Figura 36 se puede ver todo esto con más claridad. Además en ambas figuras se incluye un ejemplo concreto del procesado que se realiza. En la Figura 35 el ejemplo es el procesado de la primera frase del artículo “EN_200001142A02OA305” del tema “001”.

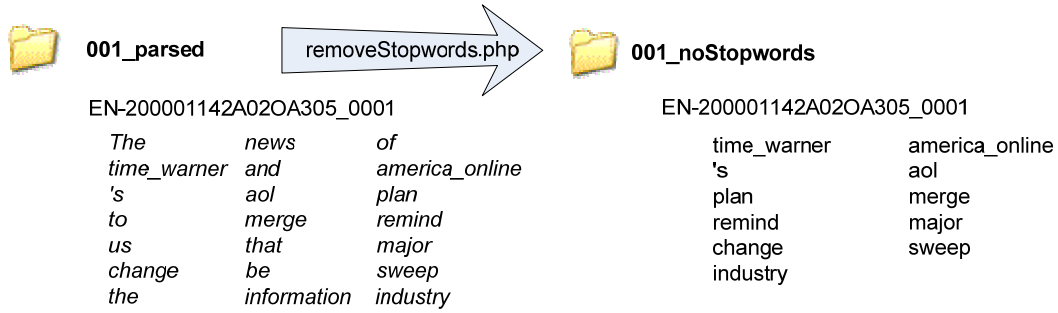

Figura 35: E/S *removeStopwords.php*.

Se puede observar que tras eliminar las palabras de parada, han desaparecido de la lista 11 de las 21 que había. Es importante recordar que las palabras de parada no son solamente las conjunciones, artículos, etc. sino todas aquellas palabras que son lo suficientemente comunes como para considerar que no aportan información adicional.

En la Figura 36 el ejemplo usado es la descripción del tema 002. Tras el procesado solamente ha sido eliminada una de las 6 palabras, la preposición “*of*”.

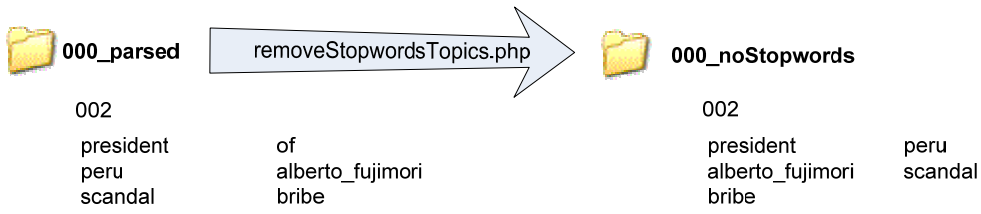

Figura 36: E/S *removeStopwordsTopics.php*.

Las siguientes figuras muestran los diagramas de *removeStopwordsTopics.php* y de *removeStopwords.php*. Aparte de la implementación del acceso a las carpetas tanto de entrada y de salida descritas, se puede ver con un poco más detalle cómo se determina si una palabra es o no de parada.

En los dos diagramas de flujo se referencia una lista de palabras de parada. Esta lista es el fichero “*stopwords_en.txt*”, ubicado en la carpeta de herramientas, y la cual se usará para determinar qué palabras no son relevantes para los análisis que se van a realizar. Se puede ver una versión simplificada de la misma en Anexo B – Lista de Palabras de Parada.

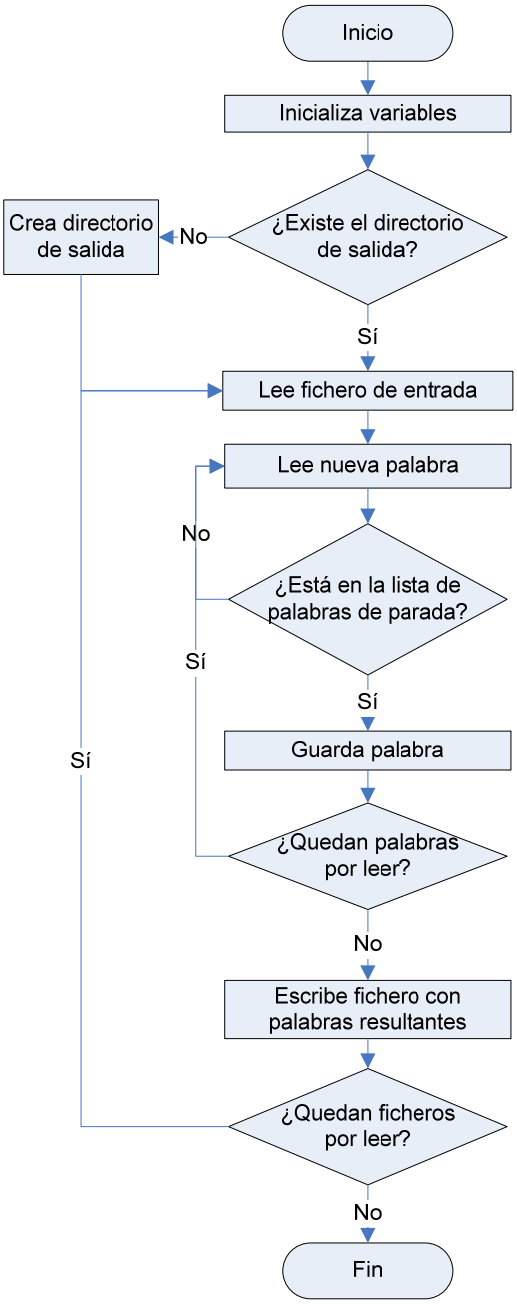

Figura 37: Diagrama de flujo de *removeStopwordsTopics.php*.

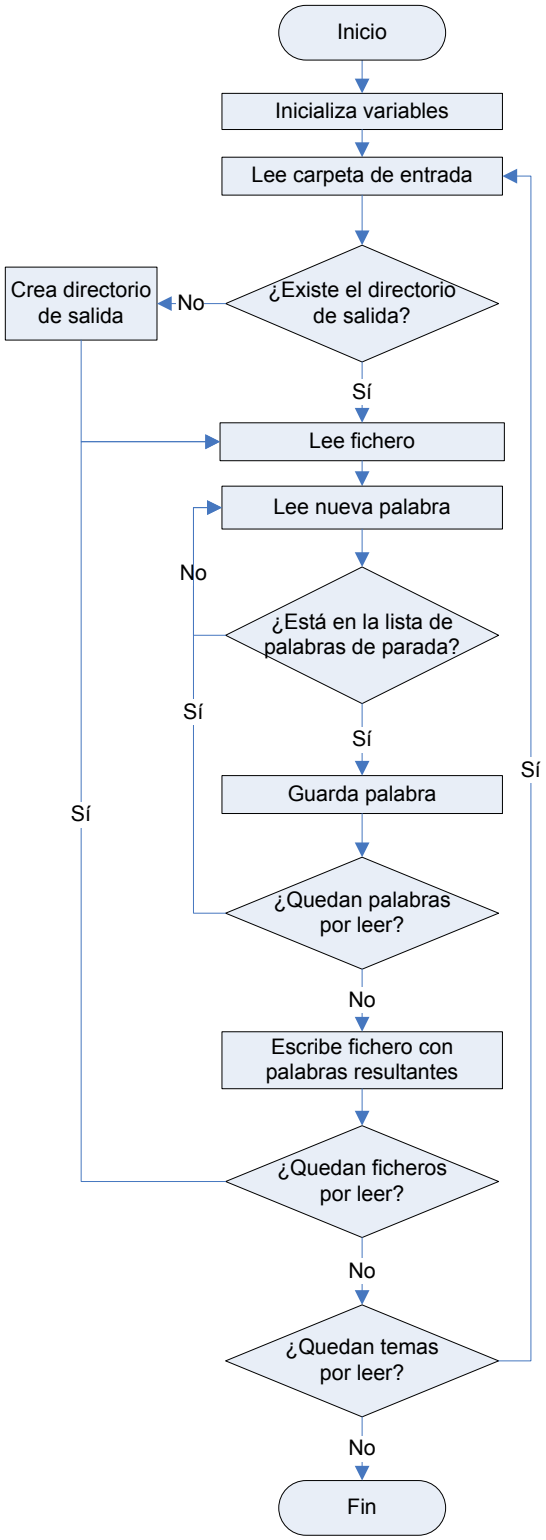

Figura 38: Diagrama de flujo de *removeStopwords.php*.

4.1.3 Separación de Datos en Conjuntos

Se ha descrito al principio de este proyecto que para intentar mitigar los efectos derivados de las dependencias con los datos se va a usar validación cruzada. Esto implica dividir los datos en tres conjuntos distintos, el conjunto de entrenamiento, el conjunto de pruebas y el conjunto de validación. Esta separación no es necesaria hacerla hasta que se llega al cálculo de las medidas a partir de las cuales se va a clasificar, ya que todo el procesado previo es común a todos los conjuntos.

En este caso y debido al volumen de los datos, la división que se realizará será únicamente lógica, creando tres listados distintos, uno por cada uno de los conjuntos en los que se quiere dividir el corpus. En ese listado aparecerán los nombres de los archivos procesados, es decir el resultado obtenido tras la etapa de extracción de información. La elección de usar el nombre de los ficheros en ese punto del procesado se debe a que se trata de un identificador unívoco en todo el sistema. Es precisamente por este motivo por lo que se mantendrá durante el resto del análisis.

La separación por conjuntos se realizará por temas, creando una carpeta llamada con el identificador del tema seguido de la cadena "_bySets" y donde se encontrarán tres ficheros: "*trainingIndex.txt*", "*testIndex.txt*" y "*validationIndex.txt*". Cada uno de los ficheros contendrá la lista de frases de ese tema que pertenecen a cada uno de los conjuntos.

En la Figura 39 se puede ver con más detalle el esquema de las entradas y salidas del programa que va a realizar esta tarea, *logicRandDivision.php*, con los datos del primer tema como ejemplo.

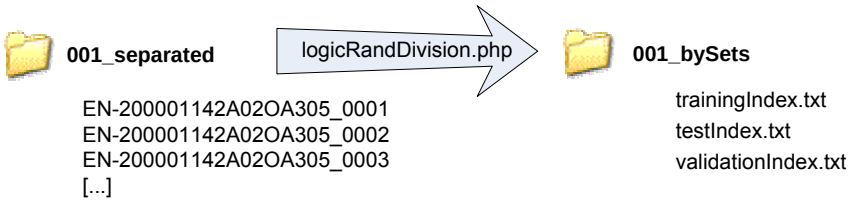

Figura 39: E/S *logicRandDivision.php*.

La proporción de muestras que tendrá cada uno de los conjuntos será la siguiente: 50% para el conjunto de entrenamiento, 25% para el conjunto de pruebas y 25% para el de validación.

Es importante remarcar que al estar mirando las proporciones para cada uno de los temas, cuando el número de ficheros en el mismo no es exactamente proporcional a la división que estamos haciendo, por redondeo se da prioridad al conjunto de entrenamiento ante los otros dos y al conjunto de pruebas ante el de validación. Esto hace que la proporción final obtenida, aunque es lo suficientemente buena, no sea exactamente la teórica.

Tabla 9: Proporciones reales de la división por conjuntos.

	Conjunto de entrenamiento	Conjunto de Pruebas	Conjunto de Validación
Muestras Teóricas	3712	1856	1856
Muestras Reales	3718	1859	1847
Porcentaje Real (%)	50.08%	25.04%	24.87%

El programa *logicRandDivision.php* leerá la información obtenida tras la fase de extracción de información de cada uno de los temas y repartirá esa información en los tres conjuntos de los que se ha hablado. La Figura 40 detalla este proceso.

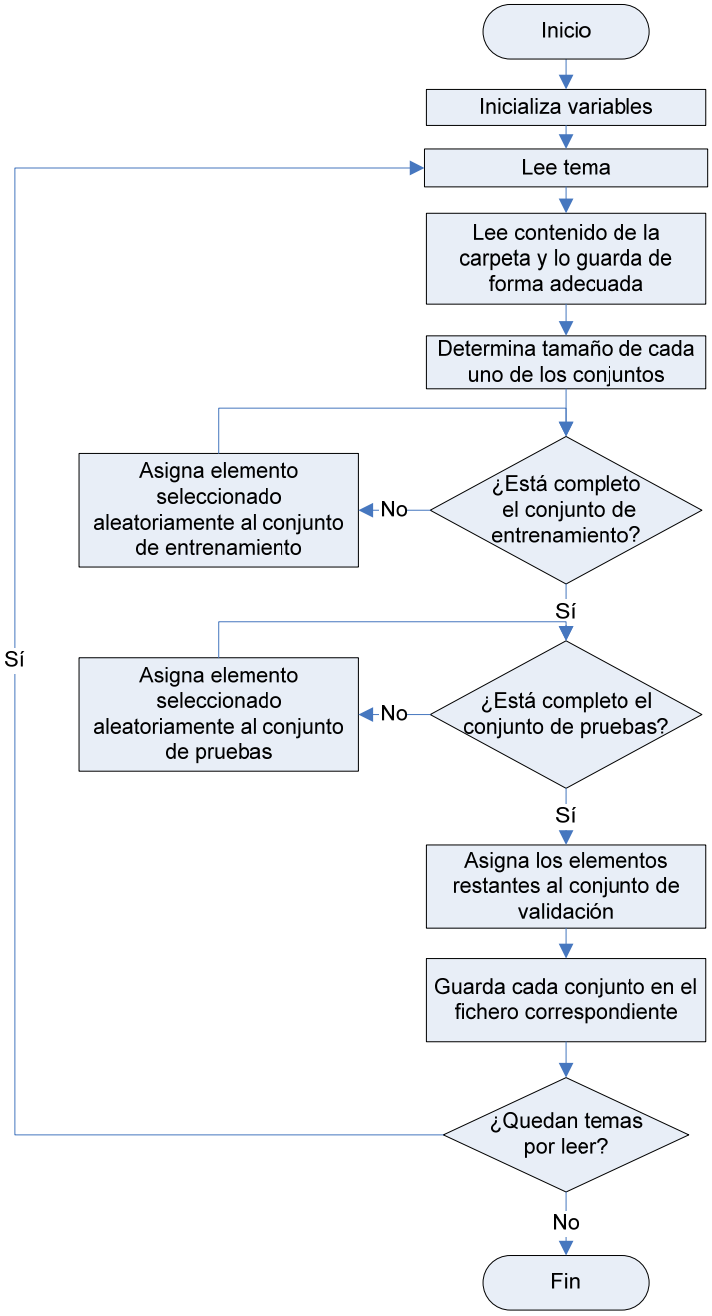

Figura 40: Diagrama de flujo de *logicRandDivision.php*.

La asignación de una prioridad mayor de un conjunto sobre otro tiene lugar cuando se determina el tamaño del conjunto.

Es importante remarcar en este punto que se van a realizar innumerables lecturas de ficheros y búsquedas sobre los resultados de esas lecturas a lo largo de todo el análisis. Por este motivo, se ha intentado durante todo el proceso, dejar un espacio en blanco al principio de cada fichero, de forma que se pueda evitar que el resultado de las búsquedas sea cero, y dé lugar a confusiones con NULL.

4.2 Clasificador de Relevancia

Una vez finalizada la fase de procesado, se pasa a la etapa de entrenamiento de la cual se obtendrá el umbral de relevancia. Dentro de esta etapa hay dos fases bien definidas: el cálculo de la similitud mediante la matriz de entrenamiento y el cálculo del umbral de relevancia del clasificador. Habrá una tercera fase mediante la que se implementará una expansión de términos de la consulta, aunque no es necesaria para obtener un valor del clasificador.

Todo este proceso va a añadir más carpetas a la estructura que ya se describió al comienzo de la etapa de procesado de datos.

En la Figura 41 se puede ver cómo se van a añadir dos carpetas adicionales a las ya creadas en la etapa anterior.

Cabe mencionar que estas dos carpetas surgen de la decisión de almacenar por separado los datos generados en las distintas iteraciones de la expansión de términos.

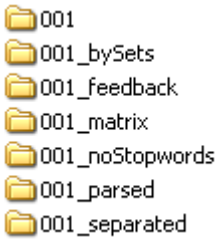

Figura 41: Estructura de carpetas tras cálculo del clasificador de relevancia.

4.2.1 Modelo de Espacio Vectorial

En apartados anteriores se ha visto a través de ejemplos cómo se realiza la representación matemática de los documentos a partir de los términos que contienen. Gracias a esta representación se podrá calcular la similitud, medida escogida para clasificar los documentos según su relevancia. Cada unidad de análisis tendrá un valor de similitud que será usado en fases posteriores, tanto para calcular el umbral de decisión como para clasificar las unidades de los conjuntos de pruebas y de validación.

Dentro de la implementación se diferenciarán dos fases, la primera en la que se calcula la matriz de pesos de cada uno de los temas del corpus y una segunda fase en la que se calculan las medidas de similitud necesarias.

4.2.1.1 Matriz de pesos

El cálculo de la matriz de pesos será independiente para cada uno de los temas, por lo que cada uno de ellos tendrá una carpeta en la que se guardarán todos los ficheros relevantes al cálculo de la misma. El nombre de esta carpeta se obtendrá a partir del identificador del tema al que hace referencia seguido de la cadena "_matrix".

Los programas utilizados en esta fase son los siguientes y serán fácilmente trazables a los distintos pasos seguidos en el ejemplo visto en Modelo de Espacio Vectorial:

- *generateWordsIndex.php*
- *generateDocsIndex.php*
- *mapWords.php*
- *obtainWeights.php*
- *redoMatrix.php*

Los dos primeros ficheros, *generateWordsIndex.php* y *generateDocsIndex.php* simplemente van a generar dos índices para cada tema, uno con los documentos que pertenecen al mismo y otro con las palabras que aparecen en el mismo. Esto se lleva a cabo porque la matriz que va a ser generada tendrá tantas columnas como palabras y tantas filas como documentos.

Como se puede ver en Figura 42, *generateDocsIndex.php* listará los ficheros que se encuentran en la carpeta resultante de realizar el procesado inicial (llamada con el identificador de tema seguido de la cadena "_noStopwords"), es decir las unidades de análisis asociadas al tema.

Cada unidad tendrá asociada en el índice generado un identificador numérico de documento.

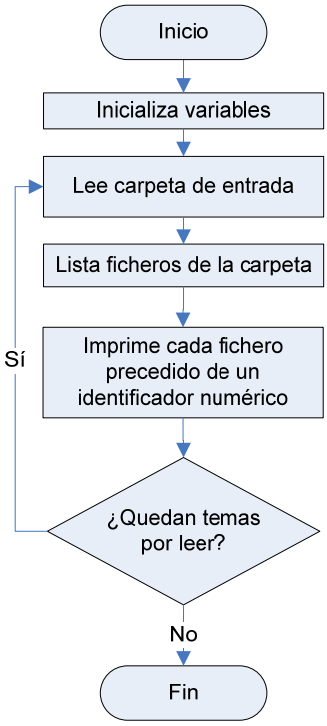

Figura 42: Diagrama de flujo de *generateDocsIndex.php*.

En la Figura 43 se puede apreciar con claridad cómo la estructura del programa que genera el índice de palabras es bastante más compleja que la que lista los documentos. Parte de esta complejidad viene dada por el hecho de que, donde antes los nombres de los ficheros eran únicos, una misma palabra puede aparecer y aparecerá varias veces en las distintas unidades de análisis de un tema y por lo tanto tendrá que ser controlado.

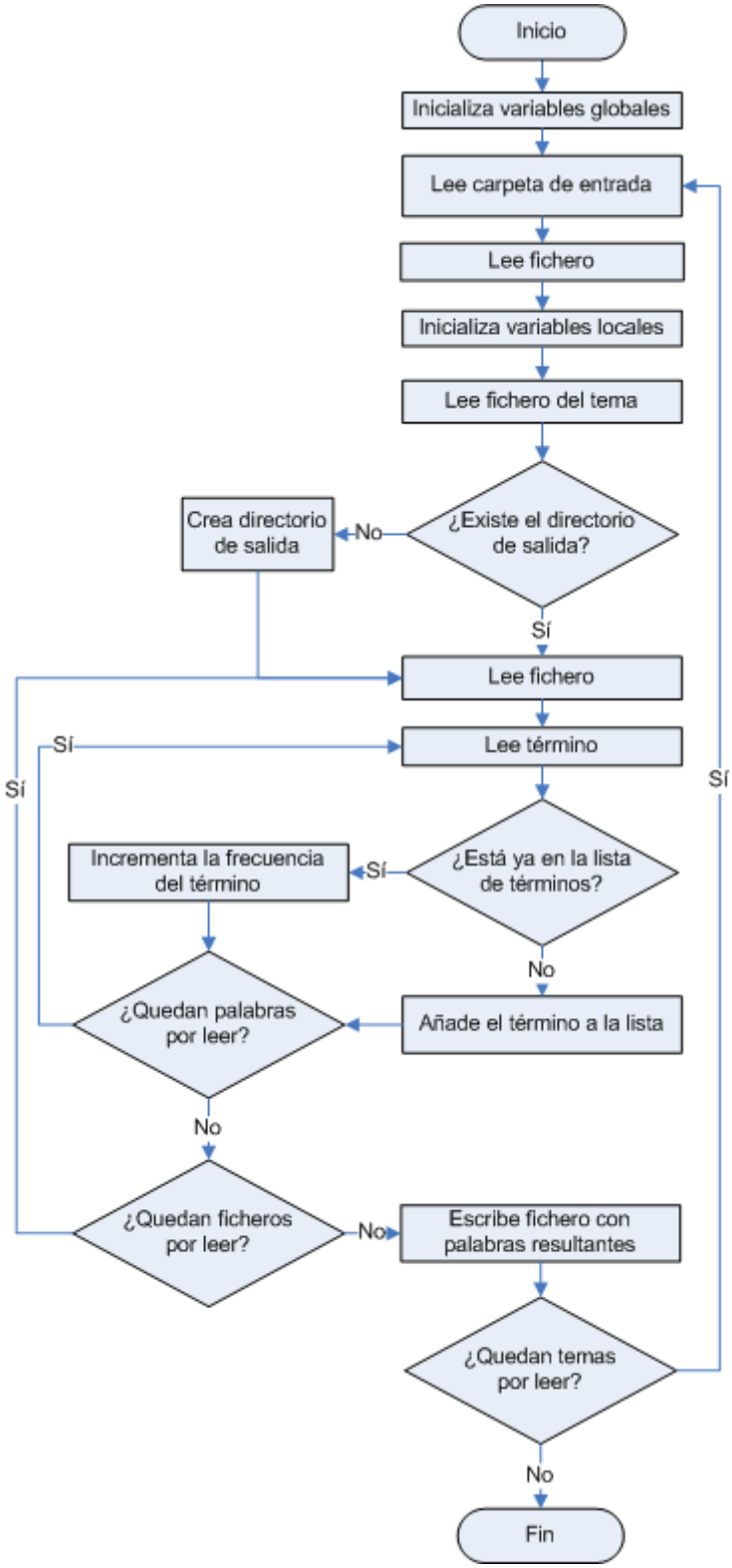

Figura 43: Diagrama de flujo de *generateWordsIndex.php*.

Teniendo esto en cuenta, se aprovechará no solo para listar las palabras que aparecen en un tema sino también para anotar la frecuencia de aparición de cada término en el tema, dato que como se ha visto, se utilizará para el cálculo de los pesos.

Por lo tanto la salida resultante de *generateWordsIndex.php* será un fichero en el que cada línea tendrá un identificador de palabra seguido de la propia palabra, y la frecuencia de aparición de esa palabra en el tema, todos ellos separados entre sí por un tabulador. Su nombre será "*indexWords.txt*".

Es importante remarcar que en este listado de palabras se están incluyendo las que aparecen en la consulta elegida para cada tema ya que a la hora de generar la matriz tiene que ser representado con un vector como si de un documento más se tratase, por lo que *generateWordsIndex.php* tendrá como entrada tanto la información relativa a cada uno de los temas como la del tema al que pertenecen. Todo esto se puede ver en la siguiente figura.

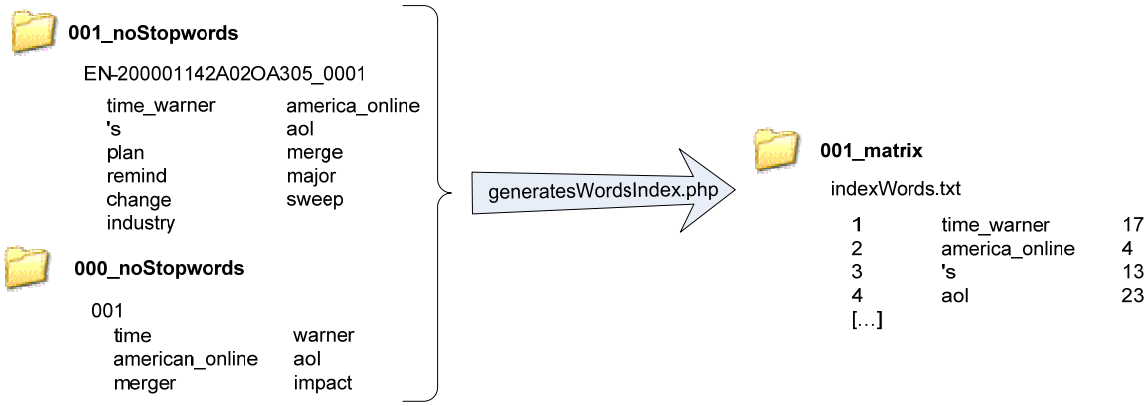

Figura 44: E/S *generateWordsIndex.php*.

También es importante notar cómo los dos programas, a pesar de ser completamente independientes el uno del otro, al haberse establecido que *generateWordsIndex.php* es el único de los dos que va a comprobar si la carpeta en la que se deben crear los documentos de salida existe, será el primero en ejecutarse. Al hacerlo, comprueba si existe ya la carpeta de salida, y si lo hace, la borra y la crea de nuevo. Es una forma de asegurarse que en el momento que se empieza a trabajar con una matriz, todos los datos van a ser los actuales y no los que quedan de ejecuciones anteriores.

Por lo tanto, cuando se ejecute *generateDocsIndex.php* ya existirá la carpeta identificada por el identificador de tema al que hace referencia y la cadena "_matrix" y dentro de ella el índice de las palabras. El fichero de salida se llamará "*indexDocs.txt*" y contendrá un identificador del documento y el nombre del mismo, separados entre sí por un tabulador.

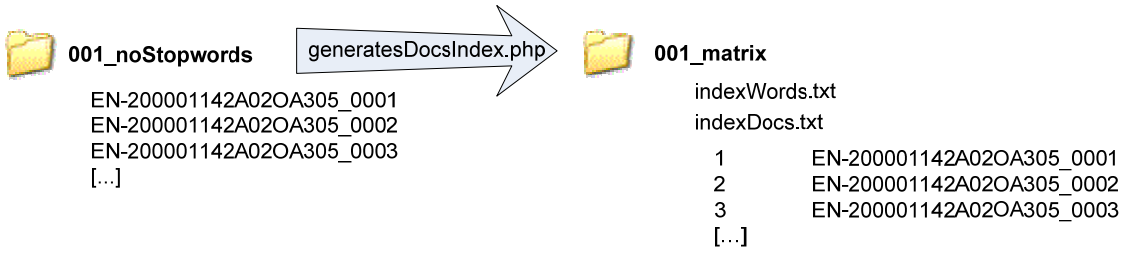

Figura 45: E/S *generateDocsIndex.php*.

Una vez se tengan los dos listados, el siguiente paso es hacer un mapeo de los términos en los documentos, de forma que accediendo directamente a la matriz se pueda saber automáticamente en qué documentos aparece cierto término.

Al igual que ha ocurrido en *generateWordsIndex.php*, en este programa hay que tener en cuenta la consulta. Una vez haya sido incluida en el mapeado, se tratará como un documento más y no será necesario volver a acceder a ella más que a través de lo disponible en el propio mapeado.

Inicialmente y debido a cómo se realiza la lectura, las filas se corresponderán a cada uno de los términos del mapa y las columnas a los documentos en los que éstas se pueden encontrar. Es importante destacar que esta notación es temporal ya que de cara a los análisis, será más práctico tener la configuración transpuesta a la descrita.

Así mismo para evitar tener que cargar más delante de nuevo la información obtenida al generar el listado de términos, se incluirá la información ahí descrita en el mapa.

La matriz resultante tendrá el siguiente orden de columnas: *Identificador de término, término, frecuencia global, tema, primer documento,[...], último documento*.

En otras palabras, si el tema contiene N términos, y M documentos, las dimensiones de la matriz de resultante serán (N)x(M+4).

En la Figura 46 se puede ver el proceso seguido para obtener el resultado que se acaba de describir.

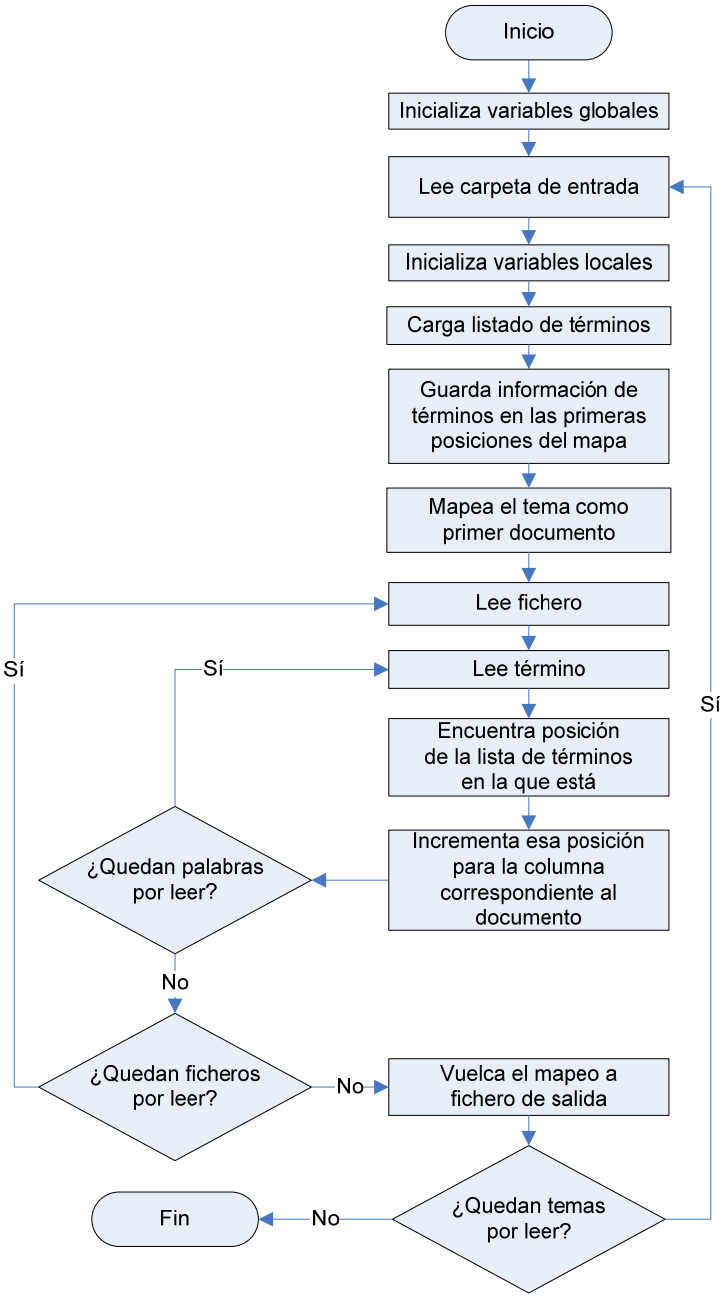

Figura 46: Diagrama de flujo de *mapWords.php*.

Respecto a las entradas que va a tener este programa, ya se ha mencionado que se va a tener en cuenta la consulta como un documento más, por lo que será necesario acceder a ella. Asimismo, se ha mencionado que la información incluida en "*indexWords.txt*" va a ser incluida en la matriz, por lo que ese fichero constituirá la tercera y última entrada de *mapWords.php* (ver Figura 47).

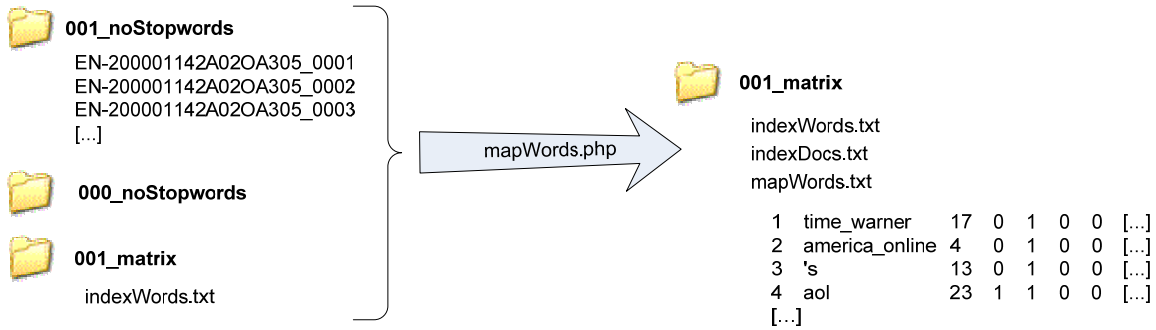

Figura 47: E/S *mapWords.php*.

Respecto a la salida, la matriz resultante será volcada en un fichero llamado "*mapWords.txt*" dentro de la carpeta donde se están guardando todos los pasos relativos al cálculo de la matriz. Dentro de la matriz, los elementos estarán separados entre sí por un tabulador.

Una vez se ha realizado el mapeado, lo único que queda para tener la matriz de entrenamiento es calcular los pesos según el modelo de espacio vectorial ya explicado. Se realizará mediante el programa *obtainWeights.php*.

En este programa se aplicará la expresión vista en la ecuación 10 del apartado 3.2.1:

$$\omega_i = tf_i * \log\left(\frac{D}{df_i}\right) = tf_i * IDF_i$$ donde IDF_i es la Frecuencia Inversa del Documento.

Todos estos datos serán obtenidos como sigue:

tf_i = es el número de veces que aparece una palabra en un documento en concreto, es decir el dato incluido en la matriz de mapeado.

D = es el número de documentos que se tiene para cada matriz. Como se genera una matriz por tema, será el número de unidades de análisis que tiene cada tema. Se obtendrá del número de columnas del mapeo realizado, ya que se sabe que el número de columnas es el número de documentos más el tema más las tres columnas de datos obtenidas de *indexWords.txt*. En otras palabras D = número de columnas – 4.

df_i = es el número de documentos en los que aparece la palabra, es decir la denominada frecuencia global y la tercera columna de la matriz volcada en *mapWords.txt*

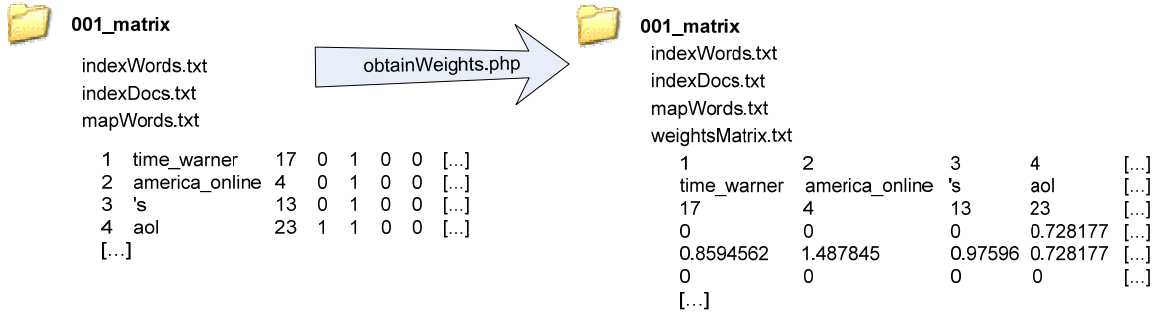

Figura 48: E/S *obtainWeights.php*.

Se observa que todos los datos, como se ha mencionado anteriormente, se pueden obtener de la matriz de mapeado generada, por lo que será la única entrada de *obtainWeights.php*. La salida será un fichero llamado "*weightsMatrix.txt*" guardado en la misma carpeta donde se han guardado los cálculos anteriores de este apartado (ver Figura 48).

Ya se ha mencionado con anterioridad pero es importante hacer hincapié en que la matriz de entrenamiento resultante va a ser transpuesta a la descrita en la fase de mapeado. Esto se debe a que resultará más cómodo en futuros análisis tener los datos de esa manera.

Esta transposición se traduce en que la matriz de pesos tendrá tantas columnas como términos distintos haya en el tema, y tantas filas como la suma del número de unidades de análisis, más una fila adicional para el mapeado de la información de la consulta, más otras tres filas iniciales con la información obtenida en "*indexWords.txt*".

Los pasos seguidos para obtener la matriz de entrenamiento se pueden ver en el diagrama de flujo de la Figura 49.

El último programa a mencionar relativo al cálculo de la matriz de pesos, es una función auxiliar llamada *redoMatrix.php*.

El cálculo de la matriz va a ser una tarea repetitiva a lo largo de todas las simulaciones que se van a realizar para evaluar el sistema. Además, las cuatro funciones que se han descrito se van a llamar siempre en bloque, en el mismo orden, y no van a dar ningún resultado intermedio que resulte de interés para la evaluación de resultados.

Por estos motivos resulta extremadamente práctico tener una sola función que llame a las cuatro mencionadas y que por tanto genere directamente la matriz de entrenamiento. Esta es precisamente la función de *redoMatrix.php*.

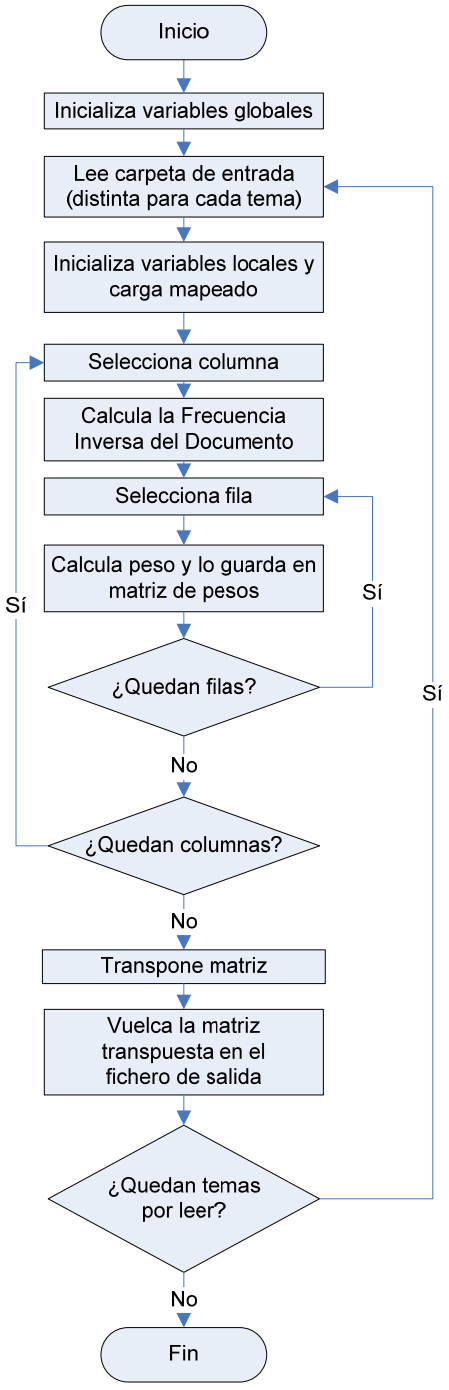

Figura 49: Diagrama de flujo de *obtainWeights.php*.

4.2.1.2 Cálculo de la Similitud

Como se ha visto en apartados anteriores, el cálculo de la similitud va a determinar los resultados que se van a obtener en el clasificador de relevancia, y en menor medida va a afectar los resultados de opinión cuando se emplee una etapa de agrupamiento.

Como ya se indicó en el diagrama de bloques de la Figura 15, el cálculo de similitud tiene dos partes, la dedicada propiamente al cálculo de la similitud, y una segunda fase de preparación de los resultados de acuerdo a los conjuntos en los que se ha dividido el corpus. Los programas usados son los siguientes:

- *getSimilarity.php*
- *prepareResultsSim.php*

El programa *getSimilarity.php* va a realizar el cálculo de la similitud, o lo que es lo mismo, va a calcular la distancia entre cada uno de los documentos referentes a cada tema, y la consulta asociada. Es precisamente por esta necesidad de comparar la consulta con los documentos por lo que en el proceso de generar la matriz de pesos se incluye la consulta como primer documento.

Este módulo calculará la distancia mencionada implementando la siguiente expresión:

$$Sim(T, D_i) = \cos(\theta_{D_i}) = \frac{\sum_i w_{T,j} \cdot w_{i,j}}{\sqrt{\sum_j w_{T,j}^2} \cdot \sqrt{\sum_i w_{i,j}^2}}$$

En donde *T* hace referencia al tema y D_i a cada uno de los documentos asociados al mismo.

El resultado de estos cálculos será un valor de similitud para cada uno de los documentos, es decir, siguiendo la notación de la expresión que se acaba de ver, *i* valores. Estos valores se imprimirán en un fichero de salida llamado "*similitud.txt*", precedidos del identificador del documento al que hacen referencia.

Este documento se guardará junto a toda la información de la matriz y será la única salida de *getSimilarity.php*. La entrada será únicamente la matriz de pesos almacenada en "*weightsMatrix.txt*".

Esta relación de entradas y salidas se puede ver gráficamente en la Figura 51.

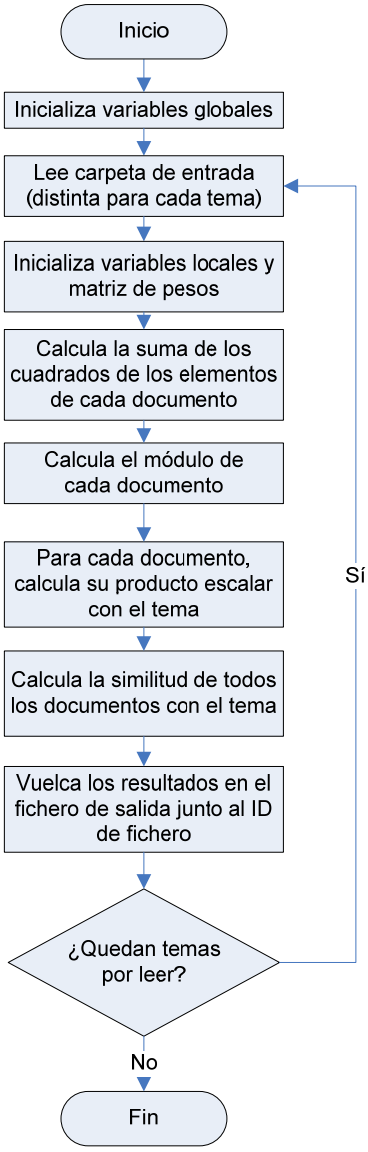

Figura 50: Diagrama de flujo de *getSimilarity.php*.

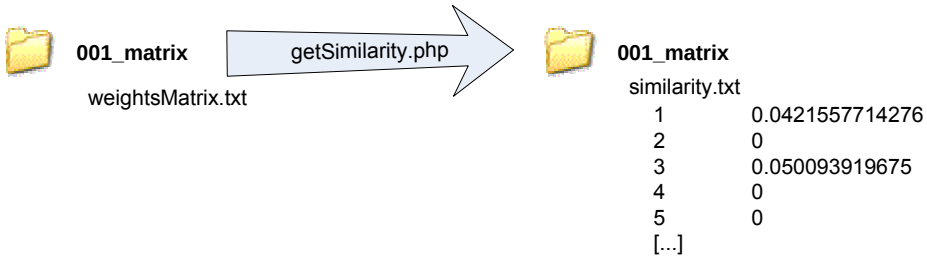

Figura 51: E/S *getSimilarity.php*.

Una vez se haya finalizado esta fase del proceso, en el resto de las etapas los datos ya van a estar en los distintos conjuntos en los que se separa el corpus, por lo que en la etapa de adaptación de formato de la similitud, se dividirán los resultados obtenidos. Esta división evitará lecturas adicionales en etapas posteriores. Esto es lo que se va a llevar a cabo con el programa *prepareResultsSim.php*.

La información incluida en los tres ficheros que se van a generar para cada uno de los conjuntos es la siguiente:

1. identificador del tema al que hace referencia.
2. nombre del artículo al que pertenece el fichero.
3. identificador de la frase en el fichero.
4. valor de similitud calculado.

Tanto la información incluida como el formato escogido para estos ficheros vienen determinados por lo que se necesitará a la hora de buscar los resultados en la anotación dada del corpus, ya sea para el cálculo del umbral o para la fase de evaluación.

Los campos son separados por comas y los tres ficheros se guardarán en una carpeta llamada con la cadena "_bySets" tras el identificador del tema al que hace referencia.

En la Figura 52 se muestra el diagrama de flujo de este programa, que al igual que la mayoría de código que se ha visto hasta ahora, procesa de forma independiente cada tema.

Se puede ver un ejemplo, tanto de la notación del documento como de las entradas y salidas del programa en la Figura 53.

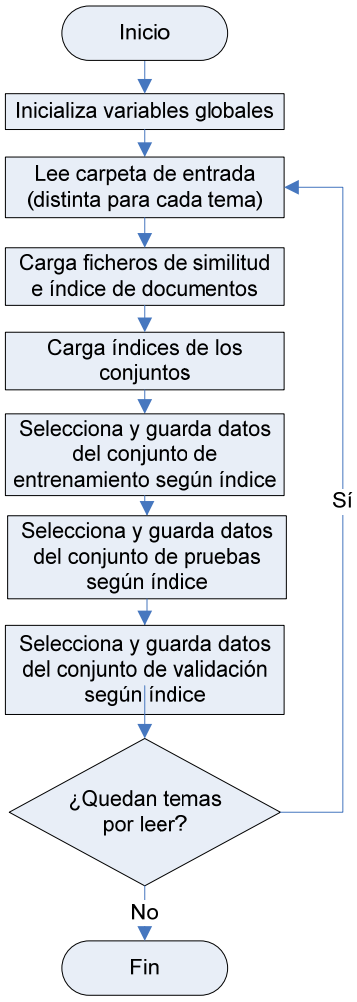

Figura 52: Diagrama de flujo de *prepareResultsSim.php*.

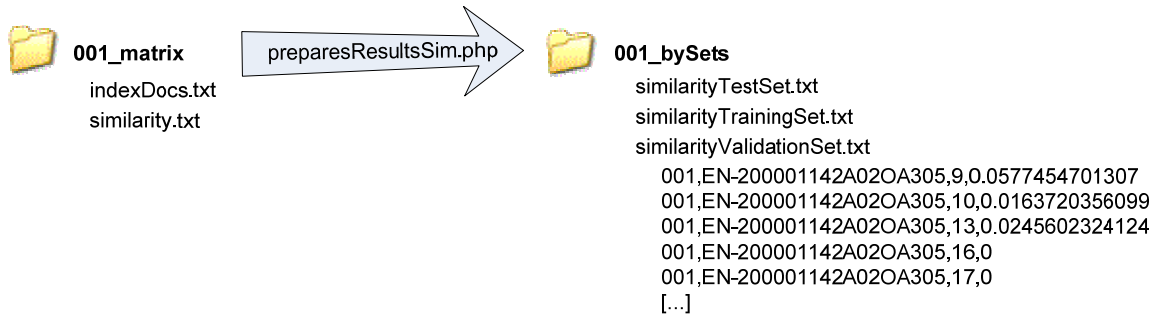

Figura 53: E/S *prepareResultsSim.php*.

4.2.2 Cálculo del Umbral

Tras tener la similitud para todos los datos y dividida en los distintos conjuntos, se pasa a calcular el umbral de decisión que se va a tomar para estimar si una unidad semántica, en este caso, cada una de las frases, es o no relevante.

Como ya se ha explicado, para calcular este umbral se usarán únicamente los resultados calculados para el conjunto de entrenamiento y la anotación de resultados de la que se dispone. Todo esto se realizará en el programa *getRelevanceThreshold.php*.

Figura 54: E/S *getRelevanceThreshold.php*.

Como se puede ver en la Figura 54, las dos únicas entradas del programa van a ser los ya mencionados resultados de similitud del conjunto de entrenamiento y la anotación con los resultados de todos los conjuntos. La salida será un único archivo llamado "*relevanceThreshold.txt*" donde se guardará el valor del umbral calculado y el valor de la decisión cuando la similitud es mayor que ese valor. Recoger este último valor no es realmente necesario ya que como se ha explicado previamente, por las características de la medida, el umbral tendrá siempre un sentido definido.

En el ejemplo usado en la Figura 54, el umbral resultante indica que se considerará que un documento es relevante si su valor de similitud es mayor que *0,0193249311619*.

Para calcular este valor lo que se hace es encontrar en la anotación dada el valor de relevancia para cada uno de las unidades de análisis que figuran como documentos pertenecientes al conjunto de entrenamiento. Dependiendo de su valor de relevancia, su valor de similitud calculado se sumará a una de las dos variables en las que se almacenarán los valores totales de similitud tanto para el grupo de documentos relevantes como para los no relevantes.

De esta forma, una vez se haya comprobado el valor de similitud de todos los documentos del conjunto de entrenamiento, se sabrá cuántos de ellos son relevantes, y que valor suma la similitud que se ha calculado a través de la matriz de entrenamiento para ellos. Esto será análogo para los no relevantes.

A diferencia de lo que se ha visto hasta ahora, la suma del valor de simulación de los documentos relevantes incluirá los elementos relevantes para todos los temas, en lugar de ser particular para cada uno de ellos, y por lo tanto los dos sumatorios que se obtengan (uno por categoría) serán un valor global relativo al total del conjunto de entrenamiento.

Estos dos valores se dividirán entre el número de documentos de la categoría encontrados, convirtiéndose en dos medias globales del valor de similitud para cada una de las categorías del sistema.

A partir de estas dos medias globales y usando una media ponderada, se calculará el umbral de decisión del sistema. En apartados anteriores ya se ha explicado el por qué de una media ponderada en lugar de una media simple (ver apartado 3.3.2).

En el cálculo genérico de un umbral, habría que comparar el número de muestras de cada una de las categorías entre las que se va a decidir para determinar qué valor se toma cuando una muestra es mayor que el umbral, pero en este caso, como ya se ha mencionado, el sentido va implícito.

En la Figura 55 se muestra el diagrama de flujo del programa en el que se lleva a cabo el proceso que se acaba de describir.

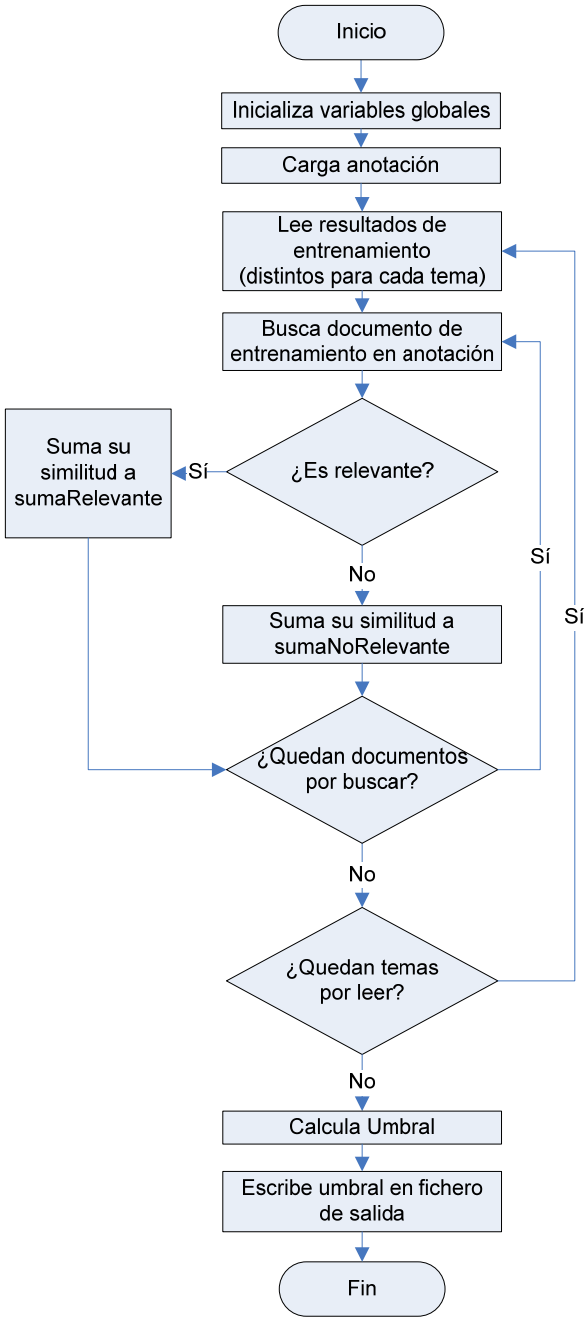

Figura 55: Diagrama de flujo de *getRelevanceThreshold.php*.

4.3 Expansión de Términos

Como ya se ha explicado en el correspondiente apartado de diseño (Expansión de Términos), el hecho de utilizar realimentación en el sistema lo que hace es expandir el número de términos de la consulta definida para cada tema, dando lugar a una nueva consulta más extensa que la anterior y que se usará en lugar de la obtenida inicialmente.

El criterio que se va a seguir para determinar que términos se añaden al nuevo tema será la similitud y los pesos calculados en una primera iteración, lo que implica que este punto es directamente dependiente de la iteración inicial sin realimentación.

La redefinición de la consulta (la consulta expandida) se hará mediante *feedback.php* que recibirá dos parámetros de entrada, el primero determinando el número de documentos (M) de los que se cogerán términos y un segundo parámetro que determinará el número de términos (N) por cada uno de esos documentos.

En el ejemplo incluido en la Figura 57 (correspondiente al ya visto en la Figura 19) se han utilizado los valores M = 1 y N = 3, por lo que solamente se añadirá a cada consulta los tres términos con mayor peso del documento con mayor similitud a la consulta inicial. Es importante recalcar que no se tendrán en cuenta los documentos con similitud nula o los términos con pesos nulos y por lo tanto, no será posible afirmar que se vayan a añadir siempre M*N términos.

En la Figura 57 se pueden ver las distintas entradas de *feedback.php*: la consulta inicial de cada uno de los temas, los resultados del conjunto de entrenamiento para poder ver los valores de similitud asignados a cada uno de ellos y la matriz de pesos y el índice de documentos a partir de los cuales se obtendrán los pesos que se usarán para escoger los términos a añadir. Precisamente en estas entradas se puede ver la dependencia directa con una iteración previa que se ha mencionado antes.

En la Figura 56 está el diagrama del flujo del programa. Cada uno de los temas se tratará independientemente ya que todos los datos de entrada son independientes para cada tema.

Es importante destacar que el resultado de *feedback.php* va a dejar los resultados en una estructura igual a la que se tiene en la carpeta “000_noStopwords”, de forma que las modificaciones que haya que hacer para poder realizar la iteración con la expansión de términos sean mínimas.

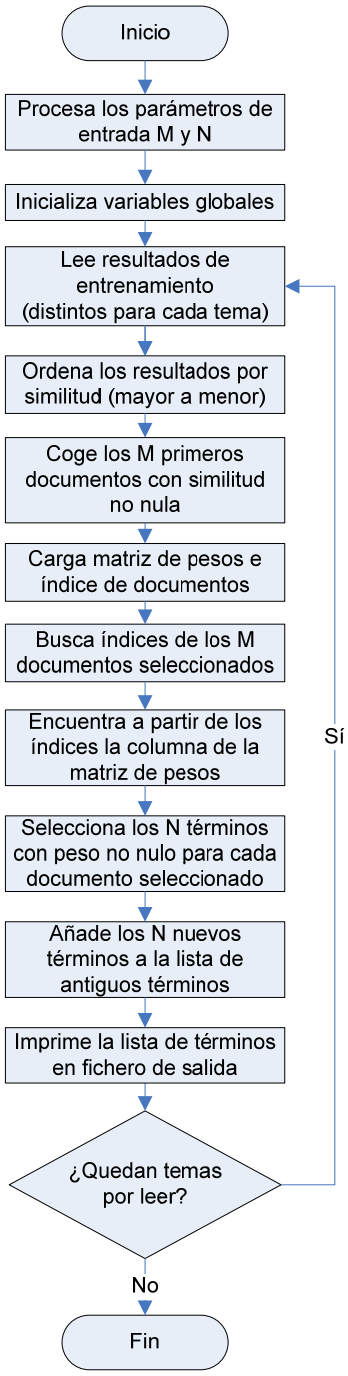

Figura 56: Diagrama de flujo de *feedback.php*.

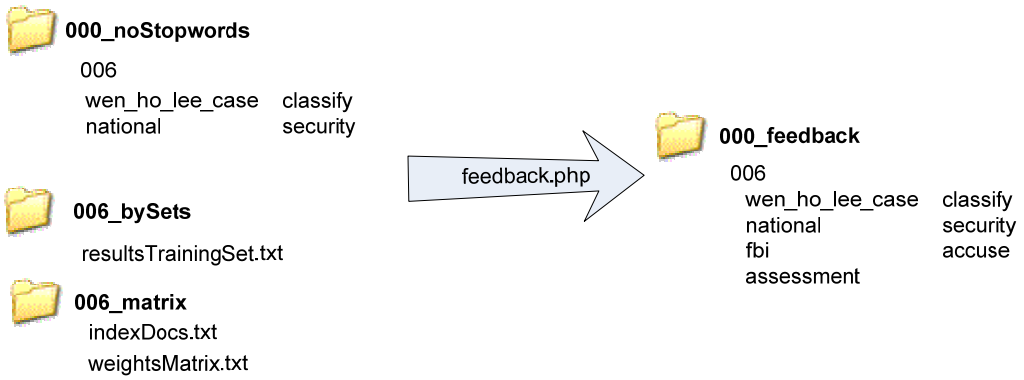

Figura 57: E/S *feedback.php*.

Cuando se haya ejecutado una primera iteración sin realimentación, la iteración inicial, será cuando se pueda ejecutar *feedback.php* y generar así las consultas expandidas. Con estas nuevas consultas el proceso a seguir para la iteración con realimentación va a ser prácticamente igual que el de la iteración inicial. Los dos únicos detalles que hay que controlar es que todo los programas que tomen como entrada la definición de la consulta, tomen la carpeta donde se encuentra la consulta expandida y que los nuevos resultados que se van a calcular no sobrescriban los de la iteración inicial.

Es importante mencionar que este último detalle no es tanto una necesidad del sistema para funcionar correctamente como una forma de facilitar las distintas evaluaciones que se van a realizar con posibles combinaciones de los parámetros M y N. Al no sobrescribir los resultados de la iteración inicial, para evaluar estas distintas combinaciones, solo será necesario realizar la iteración post-expansión ya que los resultados previos serán validos para todas ellas.

Con esto en mente, durante la ejecución se crearán dos carpetas alternativas a las de la iteración inicial en las que se guardarán los datos de la iteración con realimentación. Como ya se vio en la Figura 41, habrá una carpeta adicional de realimentación por cada tema: “tID_feedback”, para guardar tanto los datos relativos a la creación de la matriz de pesos y el cálculo de la similitud como para los resultados divididos por conjunto al que pertenecen.

En esta carpeta se guardarán los resultados en la iteración con realimentación de todo lo descrito en los apartados de “Modelo de Espacio Vectorial” y “Cálculo del Umbral”.

Por lo tanto, tan solo adaptando estos dos puntos (cómo hacerlo está indicado en el código pertinente), ejecutar la primera iteración con realimentación se hace siguiendo la misma secuencia de ejecución que para la iteración normal.

4.4 Clasificador de Opinión

El cálculo del clasificador de opinión va a ser muy parecido al ya utilizado para la relevancia, algo que facilitara su implementación.

Precisamente por esto, la estructura de carpetas resultantes (ver Figura 58) va a ser prácticamente igual que la vista en el apartado anterior.

Se puede observar cómo la única diferencia es la aparición de la subcarpeta llamada “001_tagged”. Esta carpeta va a formar parte del procesado adicional de la información que se realiza para poder obtener una medida numérica asociada a cada una de las unidades de análisis.

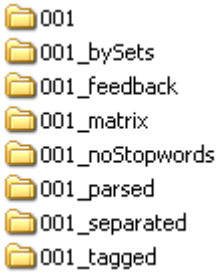

Figura 58: Estructura de carpetas tras el cálculo del umbral de opinión.

La implementación del clasificador de opinión estará dividida en tres partes: una primera parte dedicada al ya mencionado procesado adicional necesario, una segunda parte para el cálculo de la medida de opinión, y una última parte relativa al cálculo del umbral.

4.4.1 Asignación de Etiquetas

Como ya se ha explicado en el apartado de diseño Procesado Adicional, para el cálculo de opinión se añade una nueva etapa al final del procesado de datos mediante la cual se asignarán a los lemas resultantes del procesado inicial etiquetas relacionadas con su valor semántico.

Las etiquetas en cuestión dan información sobre el valor afectivo de las palabras que han quedado, de forma que luego se pueda obtener una medida de opinión de cada una de las unidades de análisis. La información semántica se obtiene mediante el diccionario semántico *General Inquirer*" que ayudará a identificar las palabras con connotaciones positivas o negativas y la fuerza asociada (ver Tabla 4). *General Inquirer* ofrece muchas más posibilidades de las que van a ser utilizadas, por se procesará el diccionario antes de utilizarlo. Se hará mediante *processInquirerDic.php* (ver Figura 59).

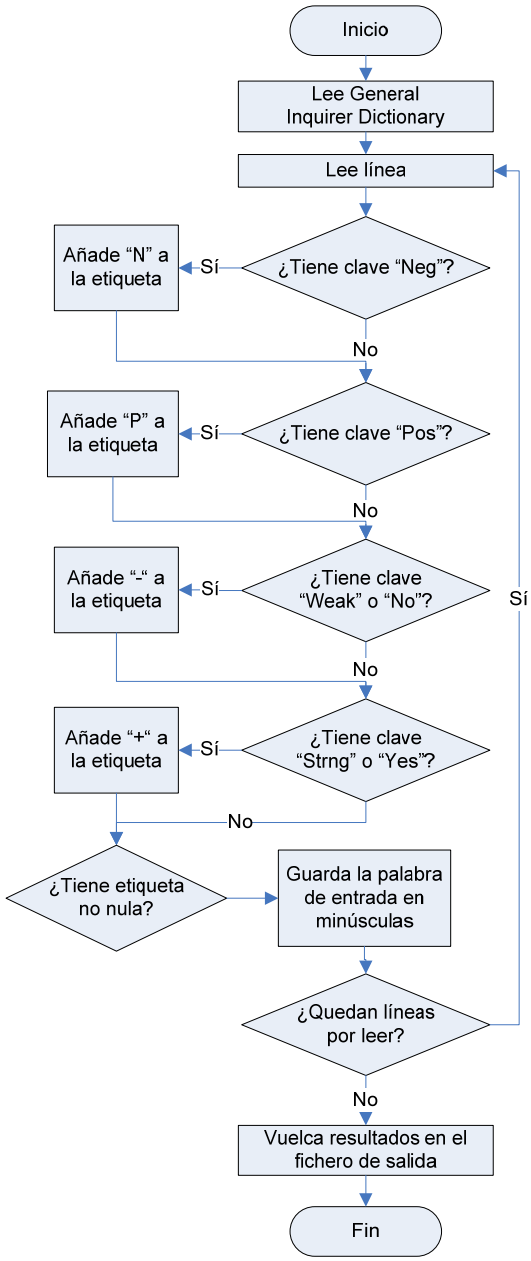

Figura 59: Diagrama de flujo de *processInquirerDic.php*.

Este código se ejecutará una única vez para obtener una versión simplificada del diccionario semántico. La entrada será "*inqdic.txt*" que tendrá el siguiente formato:

```
ABSORB#1 H4Lvd Means Modif  |
ABSORB#2 Lvd SUPV
ABSORBENT H4 Pos Incr IndAdj Modif |
ABSORPTION H4 Pos Affil Incr Noun  |
ABSTAIN Lvd TRANS SUPV
```

El primer valor es la palabra entrada, que en el caso de que haya varias acepciones de la misma vendrá repetida identificando cada una mediante "#" y un número al final de la misma. Tras la palabra se encuentra el origen de la misma, y a partir de ahí aparecerán las distintas etiquetas que puede tener relativas a los posibles valores semánticos de la palabra.

Una vez se haya procesado, los mismos valores que se han visto más arriba, quedarán como:

absorbent	*P*
absorption	*P*

Se puede observar que las únicas palabras que se guardarán son las que tienen una de las seis etiquetas elegidas. La relación entre etiqueta y símbolo usado es la especificada en la Tabla 4.

El fichero de salida resultante será "*inquirer_en.txt*" y en el habrá 4545 palabras de las 11895 del diccionario original, lo que supone aproximadamente un 38% del total. Este será el documento que se utilizará para asignar etiquetas a las palabras resultantes tras el procesado inicial hecho sobre el corpus.

Cabe resaltar que las cuatro etiquetas seleccionadas, Positivo (P), Negativo (N), Fuerte / Sí (+) y Débil / No (-), no son exclusivas entre ellas por lo que se podrán encontrar varias combinaciones.

Para asignar las distintas etiquetas a los datos relativos a cada tema se usará *assignTags.php*, que tomará como entradas el diccionario semántico procesado ("*inquirer_en.txt*") y el resultado del procesado inicial, es decir las carpetas con nomenclatura "tID_noStopwords".

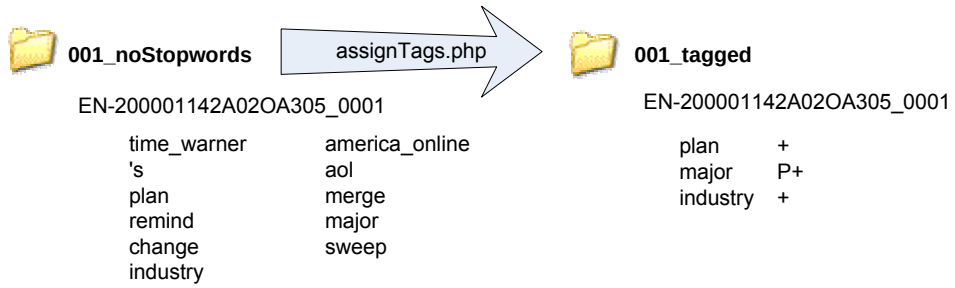

Figura 60: E/S *assignTags.php*.

En la Figura 60 se pueden ver las entradas mencionadas y también que el resultado de este procesado se va a guardar en carpetas llamadas con el identificador del tema al que hacen referencia seguido de la cadena "_tagged". También es importante remarcar algo que se ve con claridad en el ejemplo: en la carpeta resultante, únicamente se van a guardar aquellas palabras que tengan valor semántico. Se guardarán en el fichero de salida seguidas de un tabulador más la etiqueta semántica asociada.

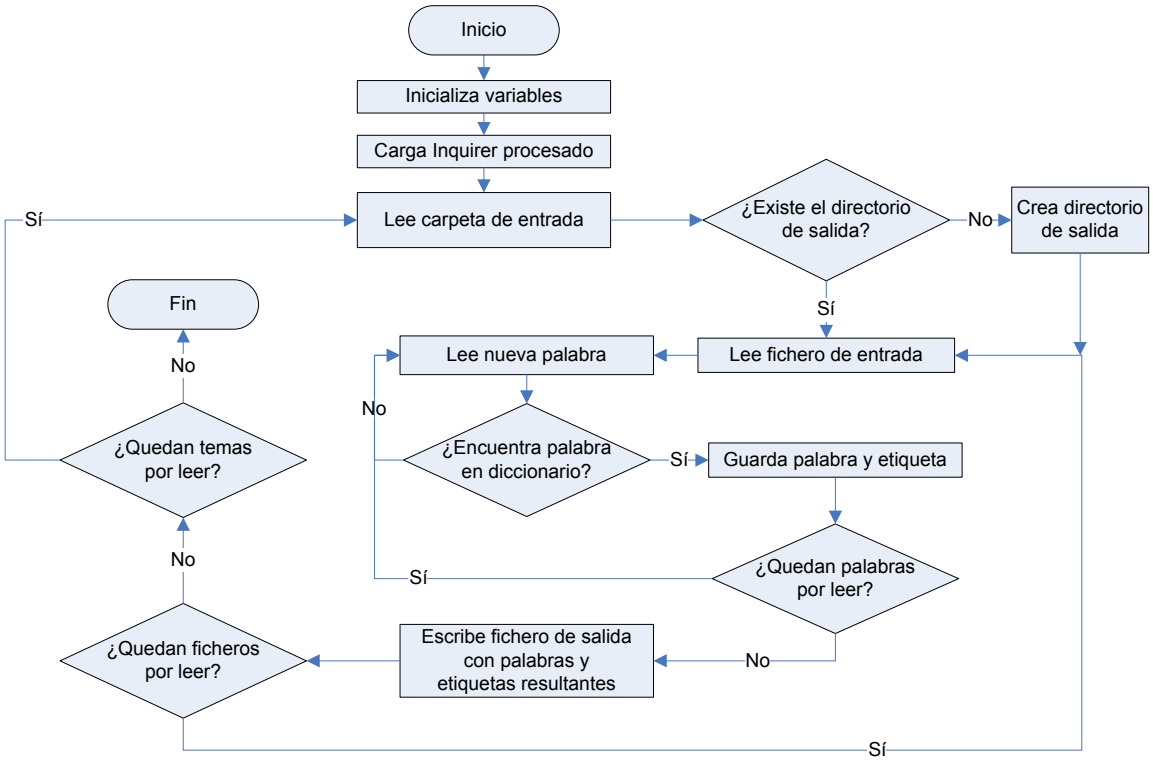

Figura 61: Diagrama de flujo de *assignTags.php*.

También se puede ver cómo la estructura va a ser bastante parecida a otros de los programas vistos en la parte de procesado inicial, como por ejemplo *removeStopwords.php*. En ambos programas se leerán los datos de entrada y dependiendo en si aparecen en una lista determinada, serán o no volcados en los ficheros de salida. En la Figura 61 se puede ver el diagrama de flujo, que es muy parecido al ya visto en la Figura 35.

4.4.2 Cálculo de Opinión

Una vez etiquetados los datos, el siguiente paso es calcular el valor numérico que se usará para determinar la opinión y por tanto para clasificar. Se hará asignando un valor numérico o bien a las distintas etiquetas que se han asignado o a una combinación de las mismas, según la configuración de cálculo de opinión elegida (ver apartado de diseño sobre la Opinión).

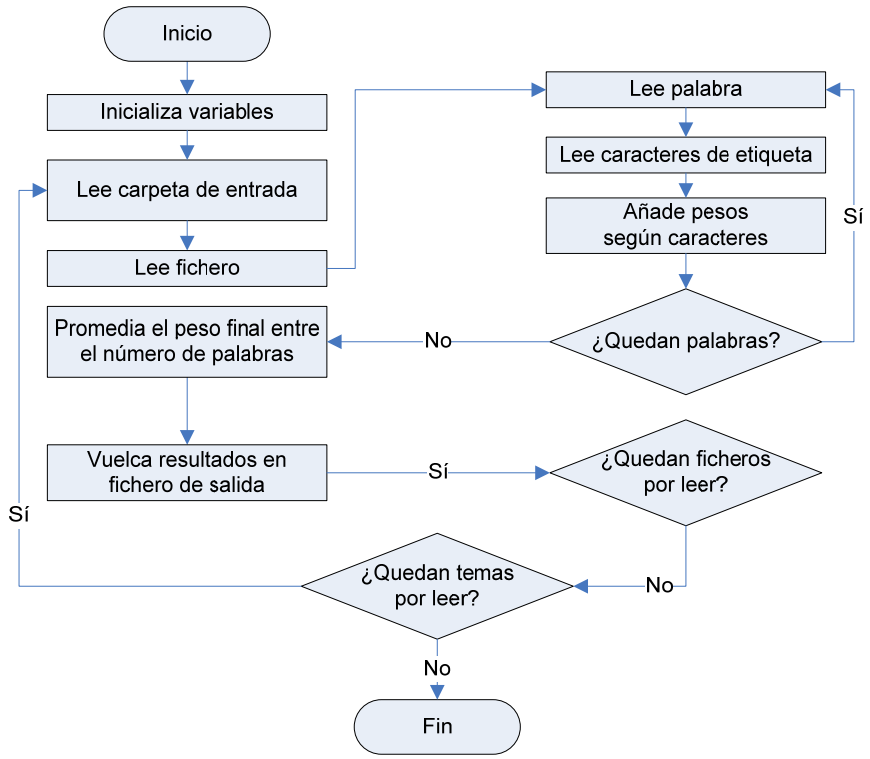

Figura 62: Diagrama de flujo de *getOpinion.php*.

Como se puede ver en el diagrama flujo de *getOpinion.php* en la Figura 62, cada uno de los caracteres presentes en la etiqueta de una palabra va a tener asignado un peso determinado. Para cada unidad de análisis se obtendrá un único valor, que será el resultado de promediar los valores numéricos asignados a cada uno de los lemas según las etiquetas asignadas.

Este cálculo es independiente de cómo se asigna el valor numérico a cada etiqueta, lo que permitirá probar distintas configuraciones del cálculo de opinión con facilidad, cambiando los valores asignados o bien a cada una de las cuatro etiquetas o bien a las combinaciones más significativas seleccionadas.

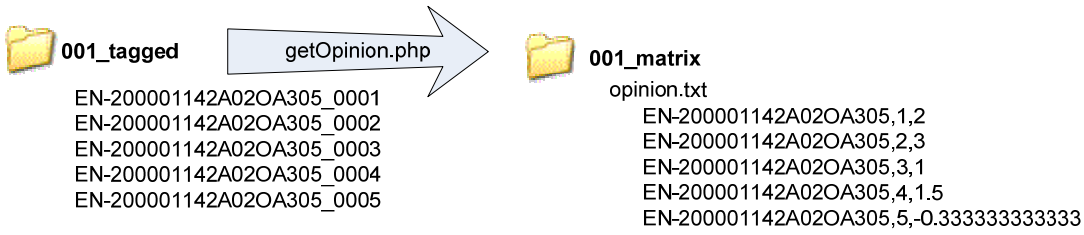

Figura 63: E/S *getOpinion.php*.

El valor calculado se volcará en un fichero de salida llamado "*opinion.txt*" que se guardará en la misma carpeta en la que está la información relativa al cálculo de la similitud ("*tID_matrix*"). El fichero contendrá el nombre del artículo, el identificador de la frase y el valor calculado de la opinión.

Una vez calculada la opinión, y de forma análoga a lo realizado para la similitud, se dividirán los resultados por conjuntos mediante *prepareResultsOp.php*. Como se puede ver en la Figura 65 a pesar de que la salida va a tener la misma estructura, la entrada de la que se parte para reorganizar los datos es diferente, de ahí que no se utilice un único programa para esta adaptación de los resultados.

Si se compara la Figura 64 con la Figura 52, se puede observar cómo la diferencia radica en la necesidad de cargar el índice de documentos generado para la matriz para identificar cada valor de similitud con el fichero al que se corresponde.

Cabe mencionar que ya que el cálculo de la opinión no va a depender en ningún caso de la similitud, en ningún momento se verá afectado por el uso o no de una expansión de términos para el cálculo de la relevancia, y por lo tanto sus resultados siempre se guardarán en la carpeta principal de resultados divididos por conjunto ("tID_bySets").

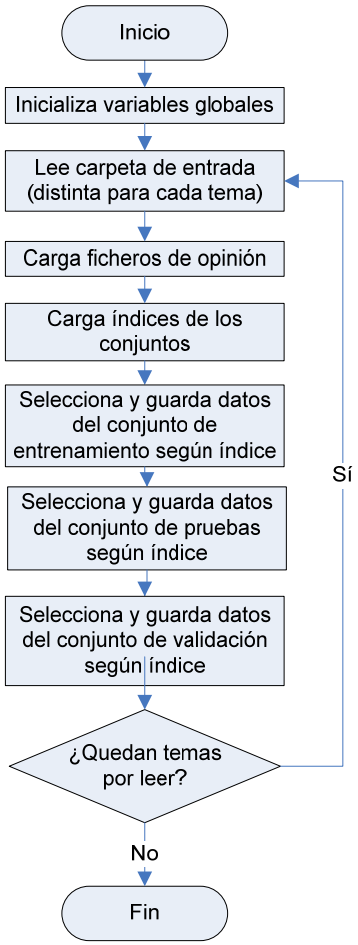

Figura 64: Diagrama de flujo de *prepareResultsOp.php*.

Los resultados de salida, como se puede ver en la Figura 65 serán tres documentos: "*opinionTestSet.txt*", "*opinionTrainingSet.txt*" y "*opinionValidationSet.txt*". En cada uno de ellos se volcará el identificador de tema, el nombre del artículo, el identificador de la frase de la que se trata y el valor de opinión calculado para el documento descrito por esos tres datos.

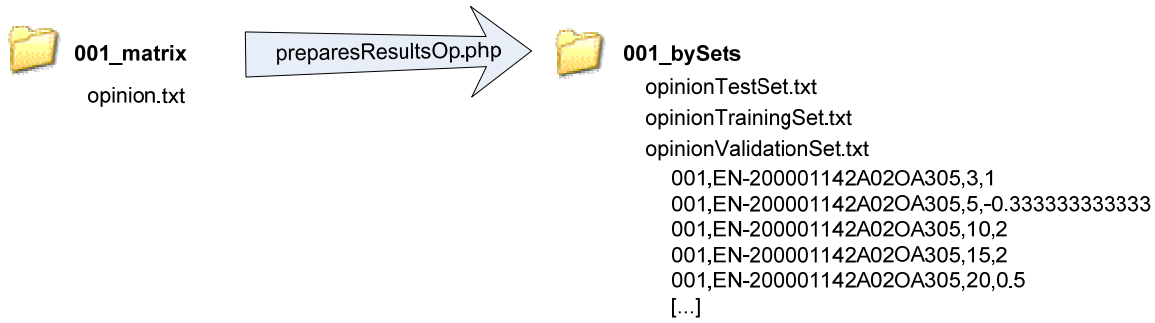

Figura 65: E/S *prepareResultsOp.php*.

4.4.3 Cálculo del Umbral

Una vez separados los resultados de la medida de opinión por conjuntos, se pasa a calcular el umbral que se va a usar para tomar la decisión sobre la subjetividad de una frase.

Como se ha visto en previos apartados, se usará una fase de agrupamiento para determinar qué documentos van a tenerse en cuenta a la hora de calcular el umbral de decisión. Por lo tanto el cálculo del umbral se dividirá en dos partes, una primera etapa, la de agrupamiento, en la que se escogerán los k documentos a utilizar de cada consulta así como la información necesaria de cada uno de ellos, y una segunda parte en la que se realizará el cálculo del umbral. Se implementarán con *kNN.php* y *getOpinionThreshold.php* respectivamente.

Como se puede ver en la Figura 66, las únicas entradas que va a tener *kNN.php* serán los resultados tanto de opinión como de similitud del conjunto de entrenamiento. De estos dos documentos se obtendrá la triada que define unívocamente un fichero (identificador de tema y nombre del artículo a los que pertenece más el identificador de frase) y su valor asociado de similitud y opinión.

Son precisamente esos cinco datos los que se escribirán en el fichero de salida que se generará. Se llamará "*kNN.txt"* y se guardará en la carpeta en la que contiene la información relativa a la matriz generada (es decir en en "tID_matrix" si la iteración es sin realimentación y en "tID_feedback" cuando sea con realimentación).

El parámetro k se pasará como parámetro de entrada a *kNN.php*, siendo absolutamente necesario que sea un valor válido. En la Figura 66 se puede ver un ejemplo de salida de *kNN.php* cuando se determina mediante el parámetro k que se utilice los tres documentos más similares a la consulta.

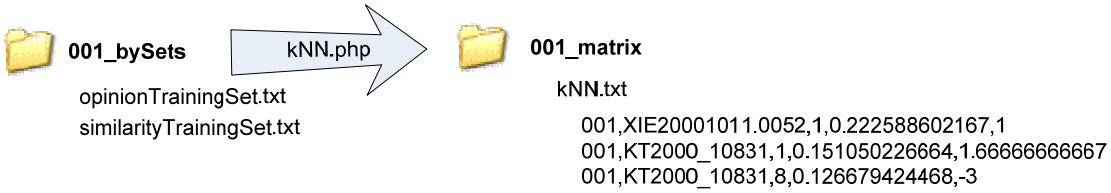

Figura 66: E/S *kNN.php*.

Para el cálculo del umbral, siempre va a ser necesario que se ejecute previamente *kNN.php*, por lo tanto, para automatizar un poco más la ejecución y evitar errores, *kNN.php* se llamará al principio de *getOpinionThreshold.php* en vez de hacerlo de forma independiente.

Este programa va a ser bastante parecido al utilizado para calcular el umbral de relevancia, por lo que tanto su diagrama de flujo (Figura 69) como sus entradas y salidas van a ser muy similares.

En la Figura 67 se puede ver cómo tanto la entrada relativa a la anotación como la salida son análogas en formato y ubicación a las vistas para *getRelevanceThreshold.php*. Como cabe esperar, la salida se llamará de forma diferente, "*opinionThreshold.txt*" concretamente, pero el formato de su contenido será exactamente igual: dos líneas, la primera con el valor del umbral calculado y la segunda con el sentido del umbral.

El umbral para la opinión, de forma análoga al de relevancia, se calculará mediante la media ponderada de la suma del valor de opinión de los documentos anotados como documentos con opinión y su homóloga para los documentos anotados sin opinión.

Figura 67: E/S *getOpinionThreshold.php*.

Como se puede ver en el diagrama de flujo de la Figura 69, se ha tenido en cuenta la posibilidad de que el parámetro k sea cero. Esto se ha hecho para poder habilitar la opción de calcular el umbral de decisión sin que haya una etapa de agrupamiento previa, de forma análoga al cálculo del umbral de relevancia. Así pues, k=0 implicará que no se va a usar agrupamiento, y que por tanto se tomarán todos los datos del conjunto entrenamiento para el cálculo del mismo.

Cuando se calcule el umbral con k=0, la entrada de los datos cambiará. En vez de usar "*kNN.txt*", se cargarán los resultados de opinión del conjunto de entrenamiento, por lo que siguiendo el ejemplo de la Figura 67, en lugar de acceder a "*/001_matrix/kNN.txt*", se accederá al archivo *"opinionTrainingSet.txt"* dentro de la carpeta "*/001_bySets/*".

Así mismo, hay que tener en cuenta que el formato de "*kNN.txt*" es distinto al de los resultados guardados en el fichero "*opinionTrainingSet.txt*", aunque el efecto será mínimo ya que la información no incluida en los resultados del conjunto de entrenamiento respecto a lo incluido en kNN es el valor de similitud, que no se usa para calcular el umbral.

Cabe resaltar que aunque el cálculo del valor de opinión es independiente de la expansión de términos, el de la similitud no lo es y por lo tanto afectará a la etapa de agrupamiento. Esto implica que habrá que adecuar el código para que admita las dos opciones (con y sin realimentación).

Respecto a las salidas de los dos diagramas de flujo (Figura 68 y Figura 69), *kNN.php* generará una salida por consulta, mientras que *getOpinionThreshold.php* creará un único fichero común para todo el corpus.

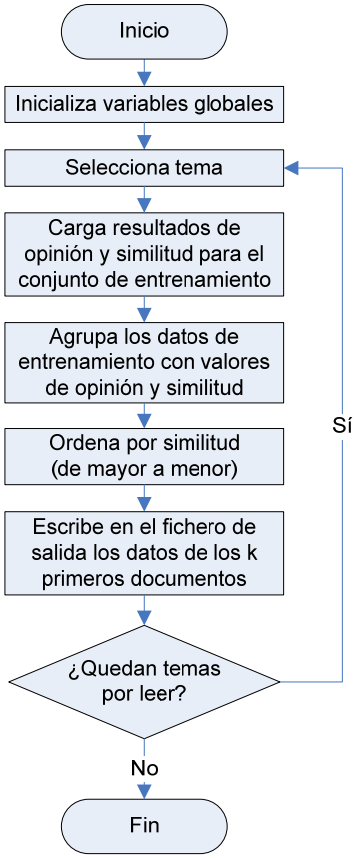

Figura 68: Diagrama de flujo de *kNN.php*.

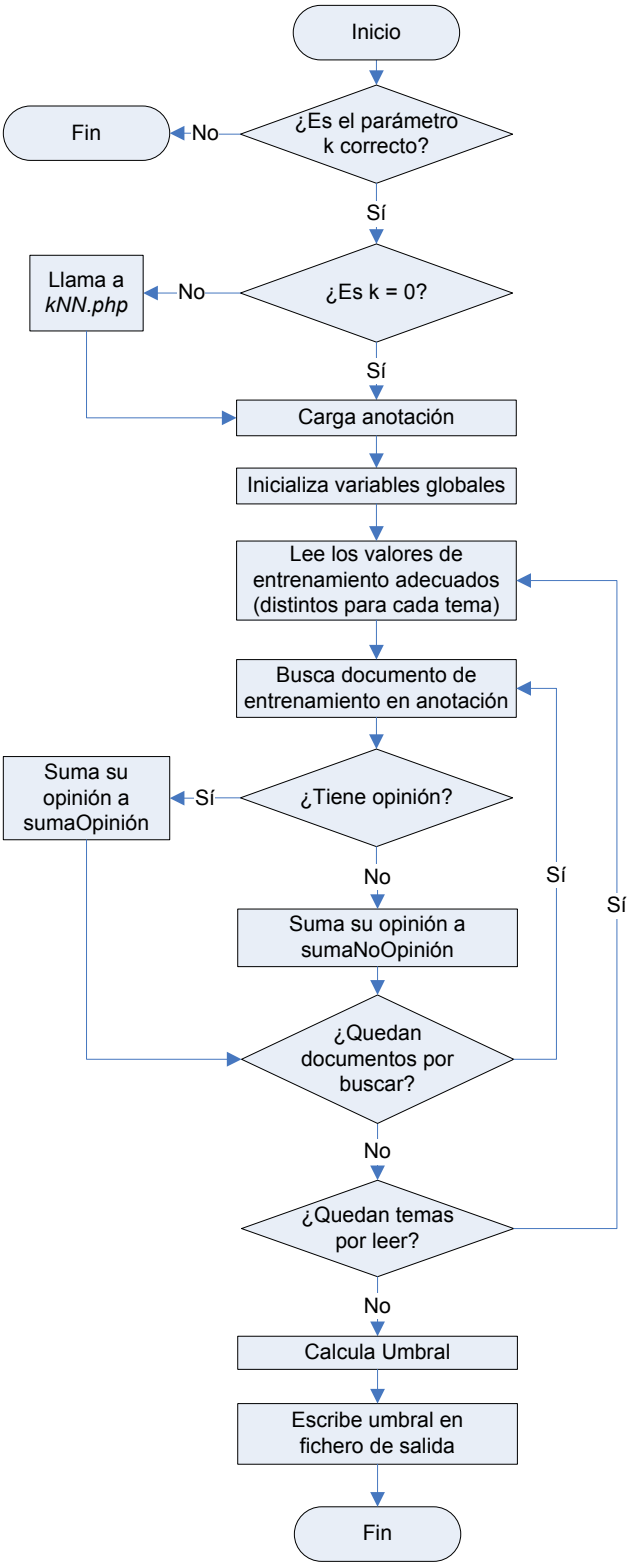

Figura 69: Diagrama de flujo de *getOpinionThreshold*.

4.5 Evaluación de los Resultados

Una vez se hayan calculado tanto el umbral de relevancia como el de opinión, se evaluarán los resultados que generan usando los dos conjuntos de datos que no han sido utilizados todavía, el conjunto de pruebas y el conjunto de validación.

La evaluación de ambos conjuntos va a ser igual, por lo tanto se utilizará un único programa, *runEvaluation.php*, que podrá ser particularizado para cada uno de los conjuntos mediante parámetros de entrada.

El primero de ellos será obligatorio y determinará qué conjunto se va a evaluar. Tendrá dos únicos valores posibles: "-t" para evaluar el conjunto de pruebas y "-v" para evaluar el conjunto de validación.

Por defecto siempre se evaluará la relevancia pero habrá un segundo parámetro opcional que permitirá elegir si se quiere evaluar también la opinión. Este parámetro se corresponderá al parámetro k del algoritmo kNN mediante el que se quiera calcular el umbral de opinión y tendrá que ser por tanto un valor numérico.

Las entradas de *runEvaluation.php*, serán la anotación de resultados del conjunto que se esté evaluando, los resultados tanto de similitud y de opinión del conjunto a evaluar, y los umbrales calculados para las dos medidas.

La salida del programa será volcada a la salida estándar por lo que los resultados se mostrarán únicamente por pantalla.

Los datos de salida que se volcarán son los necesarios para obtener la matriz de confusión y el número de errores totales cometidos en la estimación.

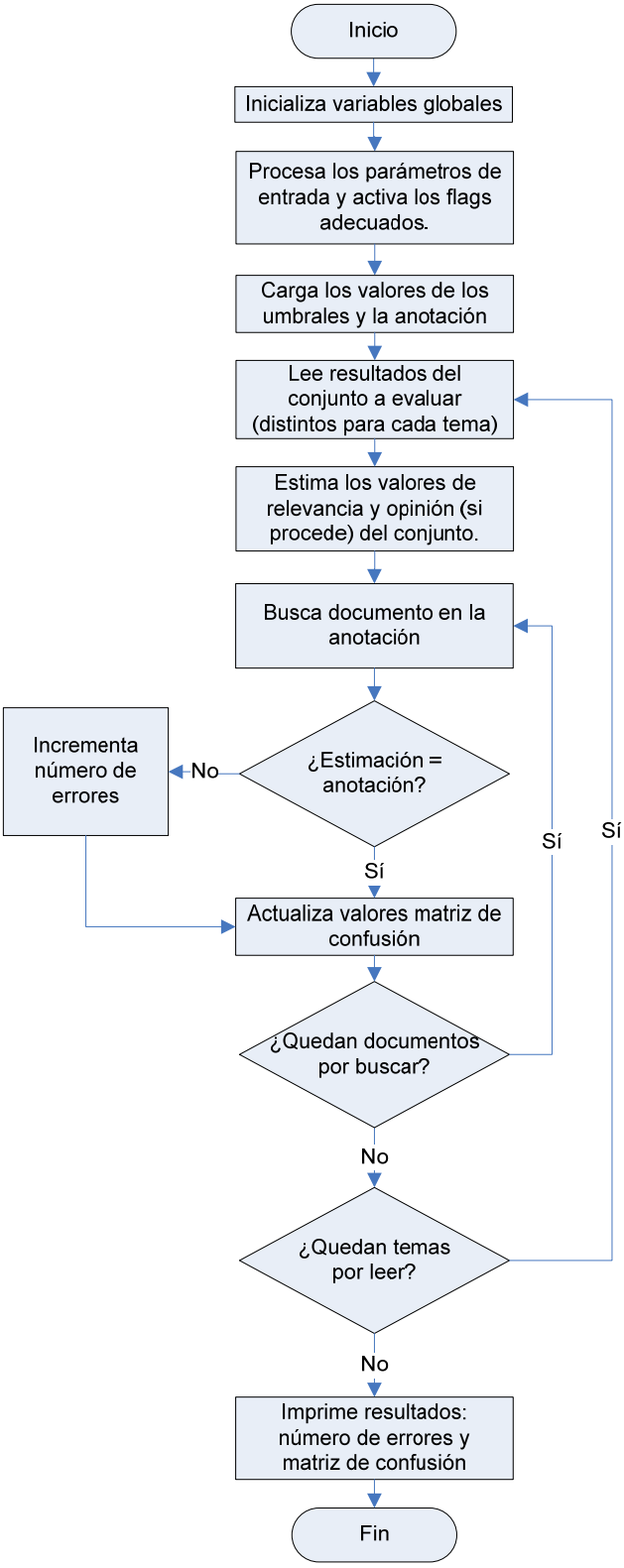

Figura 70: Diagrama de flujo de *runEvaluation.php*.

5 Evaluación

En este apartado se pasa a exponer los resultados obtenidos en las distintas evaluaciones realizadas dando detalles sobre las medidas utilizadas y los escenarios bajo los cuales se han tomado. Asimismo, se comentarán a medida que se expongan las distintas conclusiones que se obtienen de ellos.

5.1 Características del Corpus

Antes de ver los resultados es importante realizar una descripción detallada del corpus con el que se está trabajando y de cómo sus características van a afectar a los resultados obtenidos.

Se parte de la gran ventaja de tener un corpus proporcionado por NTCIR-6, algo que facilita enormemente la labor pues evita el tener que crear una colección de datos a medida para realizar los distintos análisis. Al mismo tiempo, el hecho de ser datos predefinidos hace que no se pueda tomar ninguna decisión respecto a sus características y por tanto que no se pueda asegurar ciertas condiciones de igualdad respecto a la distribución de los resultados.

Como ya se ha mencionado anteriormente el corpus está formado por 28 temas distintos (ver Anexo A – Temas de NTCIR-6), cada uno de ellos identificados mediante un número de tres cifras y a los que hay asociados un número de artículos.

El número de artículos no es uniforme sino que depende de cada tema como se puede ver en la siguiente figura.

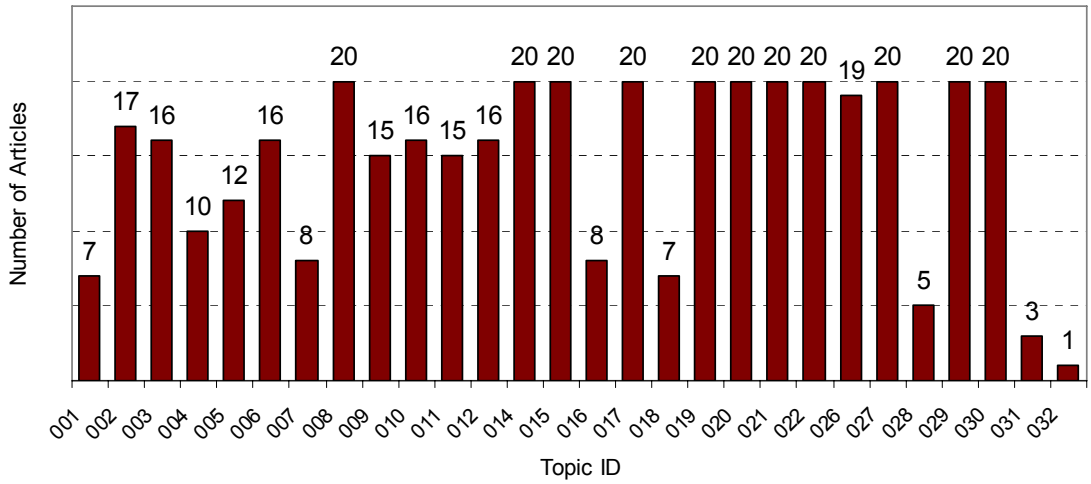

Figura 71: Número de artículos por tema.

Es importante tener en cuenta que el hecho de que varios temas tengan el mismo número de artículos asociados no quiere decir que la cantidad de información sea parecida ya que, como se puede observar en la Figura 72, el número de frases presentes en cada uno de los artículos puede ser muy distinta y es precisamente la frase lo que se ha elegido como unidad de análisis.

La primera conclusión que se puede sacar de esta información es que al dividir los subconjuntos con los que se va a trabajar de forma proporcional para cada uno de los temas, los temas con mayor número de frases tendrán más unidades de análisis en el conjunto de entrenamiento, y por tanto más peso en el umbral calculado.

El número total de frases del corpus es de 7424. Para cada una de ellas, se dispone además de información sobre si es o no relevante, si tiene o no opinión, y en caso de que la tenga, la polaridad de la misma y quién da esa opinión.

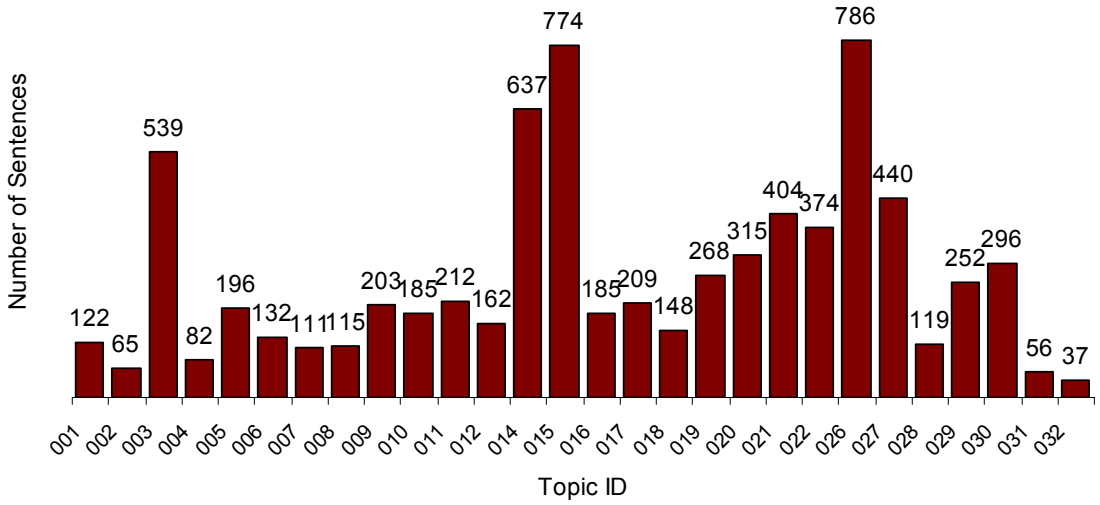

Figura 72: Número de frases por tema.

Como se ha dicho, aparte de relevancia y opinión los resultados dan detalles sobre la polaridad y el sujeto que da la opinión, pero estos dos últimos datos, al no entrar dentro del alcance de este proyecto no serán utilizados.

A continuación se pasa a analizar la proporción en los resultados de las dos categorías tanto en la clasificación de relevancia como en la de opinión.

Tabla 10: Proporción de categorías en los resultados.

	Número de Frases	Porcentaje (%)
Relevantes	3704	49,89%
No Relevantes	3720	50,11%
Con Opinión	1972	26,56%
Sin Opinión	5452	73,44%

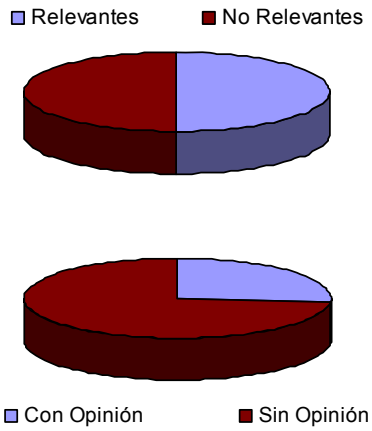

Figura 73: Distribución de clases en el corpus.

Se ve con claridad cómo para la relevancia, la cantidad de frases que son relevantes respecto a las que no es bastante parecido, una característica buena de cara al entrenamiento del umbral. La proporción en la opinión es bastante más dispar, donde solamente una de cada cuatro frases tendrá opinión.

Como ya se ha mencionado en apartados anteriores, el corpus se dividirá en los siguientes conjuntos:

- Conjunto de entrenamiento: supondrá el 50% del total del corpus, y se utilizará en la fase de entrenamiento de los dos clasificadores del sistema para obtener los correspondientes umbrales de decisión.
- Conjunto de pruebas: supondrá el 25% del total de unidades de análisis del corpus y se utilizará para hacer una primera evaluación de los resultados y para ajustar los distintos parámetros configurables del sistema según la característica del mismo que se quiera optimizar.
- Conjunto de validación: estará formado por el 25% del corpus restante y se utilizará para obtener los valores finales de las evaluaciones para los valores de los parámetros fijados con el conjunto de pruebas.

Cada uno de estos conjuntos tiene un tamaño lo suficientemente grande como para que se mantengan las proporciones vistas en la Figura 73.

La asignación de las distintas unidades de análisis a un subconjunto u otro se hará de forma aleatoria para mitigar la dependencia de los datos elegidos. Además se realizará mediante un muestreo estratificado, por lo que se mantendrá en cada uno de los conjuntos las proporciones de documentos pertenecientes a cada tema.

5.2 Medidas de Evaluación

Las medidas de evaluación que se utilizarán son las tradicionales en los sistemas de Recuperación de Información: la precisión (P) y la cobertura o *recall* (R). Ambas medidas se basan en un sistema compuesto de una colección de documentos y una consulta sobre los mismos.

La precisión proporcionará el número de resultados relevantes a la consulta planteada (no hay que confundir esta relevancia con el valor de relevancia que se ha tratado a lo largo del proyecto) mientras que la cobertura dará el número de resultados relevantes a la consulta encontrados respecto a todos los resultados relevantes del corpus.

Ambas medidas se pueden describir matemáticamente de la siguiente forma:

$$precisión = \frac{\left|\{documentos\_relevantes\} \cap \{documentos\_encontrados\}\right|}{\left|\{documentos\_encontrados\}\right|}$$ **ECUACIÓN 16**

$$cobertura = \frac{\left|\{documentos\_relevantes\} \cap \{documentos\_encontrados\}\right|}{\left|\{documentos\_relevantes\}\right|}$$ **ECUACIÓN 17**

Estas medidas se pueden particularizar para cada una de las clases del sistema, por lo que se puede obtener un par de medidas por categoría más una medida global obtenida de promediar las de todas las categorías del sistema.

Para poder calcular con facilidad estas medidas, se usará la matriz de confusión, en la que se representan tanto en las filas como en las columnas las clases en las que se va a dividir el sistema, considerando que las filas son los resultados obtenidos y las columnas los correctos.

Tabla 11: Matriz de confusión.

		Resultado Correcto	
		YES	NO
Resultado Obtenido	YES	tp (*true positive*)	fp (*false positive*)
	NO	fn (*false negative*)	tn (*true negative*)

En términos de esta matriz se podrá calcular todas las medidas descritas. Si consideramos que vamos a medir la precisión para las dos clases del sistema, tendremos en función de la matriz lo siguiente:

$$precisión_{YES} = P_{YES} = \frac{tp}{tp + fp} \qquad \text{ECUACIÓN 18}$$

$$cobertura_{YES} = R_{YES} = \frac{tp}{tp + fn}$$

$$precisión_{NO} = P_N = \frac{tn}{tn + fn} \qquad \text{ECUACIÓN 19}$$

$$cobertura_{NO} = R_N = \frac{tn}{tn + fp}$$

Existe una medida que combina el par formado por la precisión y cobertura, de forma que mediante un único valor se pueda saber la calidad de unos resultados. Se trata de la medida-F. Su forma más general es la siguiente:

$$F_\beta = \frac{(1+\beta^2)}{\beta^2} \cdot \frac{precisión \cdot cobertura}{precisión + cobertura} \xrightarrow{\beta=1} F_1 = 2 \cdot \frac{precisión \cdot cobertura}{precisión + cobertura} \qquad \text{ECUACIÓN 20}$$

La dependencia el parámetro β permite dar distinto peso a la precisión respecto a la cobertura. Uno de los valores más comunes es β = 1, que otorga el mismo peso a las dos medidas.

Hay dos métodos para realizar el cálculo de la medida-F, *micro-averaging* y *macro-averaging* [38].

- *Micro-averaging* calculará la precisión y la cobertura para todas las categorías del sistema juntas, es decir a partir de unos valores de verdaderos positivos, falsos positivos y falsos negativos globales. Una vez calculada esta precisión y esta cobertura, se calculará una medida-F con ellas siguiendo la expresión vista en la última ecuación vista.

$$\left.\begin{aligned} P^\mu &= \frac{tp}{tp + fp} = \frac{\sum_{i=1}^{M} tp_i}{\sum_{i=1}^{M} (tp_i + fp_i)} \\ R^\mu &= \frac{tp}{tp + fn} = \frac{\sum_{i=1}^{M} tp_i}{\sum_{i=1}^{M} (tp_i + fn_i)} \end{aligned}\right\} \Rightarrow F^\mu = 2 \cdot \frac{P^\mu \cdot R^\mu}{P^\mu + R^\mu} \qquad \text{ECUACIÓN 21}$$

- *Macro-averaging* calculará una medida-F para cada una de las M categorías, es decir una precisión y una cobertura para cada categoría. Una vez se haya calculado una medida-F por categoría, se hará la media entre las medidas-F de todas ellas. Este método es el que se va a usar en la fase de evaluación.

$$\left.\begin{aligned} P_i^M &= \frac{tp_i}{tp_i + fp_i} \\ R_i^M &= \frac{tp_i}{tp_i + fn_i} \end{aligned}\right\} \Rightarrow F_i^M = 2 \cdot \frac{P_i^M \cdot R_i^M}{P_i^M + R_i^M} \Rightarrow F^M = \frac{\sum_{i=1}^{M} F_i^M}{M}$$ **ECUACIÓN 22**

El método de *micro-averaging* da el mismo peso a todas los documentos mientras que *macro-averaging* da el mismo peso a todas las categorías. En este caso, al estar siempre en una decisión binaria va a influenciar menos, pero en el caso en el que hubiese más categorías *micro-averaging* favorecería a las categorías más comunes mientras que *macro-averaging* favorecería a las menos comunes.

Aparte de la media de la medida-F como resultado global de una simulación también se usará en determinados casos la medida de exactitud, que proporcionará una idea de cuántas predicciones se hacen correctamente, independientemente de la categoría a la que pertenecen. Si se escribe en función de la matriz de confusión vista en la Tabla 11 queda la siguiente expresión:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$ **ECUACIÓN 23**

Al estar usando una división lógica de los conjuntos con los que se trabaja, para dar fiabilidad a los resultados y en la medida en la que se pueda eliminar el componente de dependencia de las muestras tomadas, se realizarán varias repeticiones del mismo escenario. El resultado final será la resultante de calcular la media truncada de los resultados obtenidos, es decir la media de las muestras resultantes tras eliminar la que proporciona el mejor resultado y la que proporciona el peor.

Al hablar de mejor y peor resultado es importante tener en cuenta que el objetivo en la toma de estas decisiones va a ser siempre optimizar la detección de la categoría YES por lo que se optimizará el valor de la precisión calculada para dicha categoría. Hay que tener en cuenta que tanto esta decisión como el truncamiento de los datos de cara al cálculo de la media se hará sobre los datos del conjunto pruebas mientras que los mostrados en los resultados serán los obtenidos para el conjunto de validación.

5.3 Resultados de la Evaluación

En este apartado se incluirán los resultados más importantes y significativos de la fase de evaluación del sistema diseñado. Se describirán los escenarios empleados, las decisiones tomadas en cuanto a opciones de simulación y se incluirán tablas y gráficos comprensibles de los resultados obtenidos.

5.3.1 Relevancia

La primera fase es el análisis de la relevancia. Este análisis se va a dividir principalmente en dos fases: una fase inicial en la que se obtienen los primeros resultados del sistema, y una segunda fase en la que se intenta mejorar estos resultados a través de una expansión de términos implementada mediante realimentación.

5.3.1.1 Primera Evaluación de la Relevancia

La primera decisión que se tomará a lo largo del proceso de evaluación es la relacionada con qué definición de la consulta se va a tomar para realizar todos los posteriores análisis. Esta decisión se realizará sobre los resultados obtenidos para las primeras simulaciones de estimación de la relevancia.

Tras ver las definiciones proporcionadas para cada uno de los temas, se determinó evaluar los resultados que se obtenían con tres opciones distintas: usando solamente el campo descripción (<DESC>), usando solamente el campo título (<TITLE>), y usando una combinación de ambos.

Para cada una de esas opciones se realizan varias simulaciones a partir de las cuales se obtienen resultados tanto para el conjunto de pruebas como para el de validación, aunque en este caso como ya se ha mencionado, el de pruebas solo se utilizará para decidir qué valores se truncan de la media y obtener el valor correspondiente del conjunto de validación.

A continuación se verán dos grupos de medidas para las distintas consultas definidas. El primer conjunto estará compuesto de las medidas particulares para cada una de las categorías que se obtienen directamente a partir de la matriz de confusión: precisión y cobertura para las dos clases del sistema (ver Tabla 11).

El otro conjunto englobará las medidas de carácter más global e incluirá la medida-F para cada una de las categorías del sistema, la medida-F global calculada como ya se ha mencionado mediante *macro-averaging* y la exactitud del sistema.

Tabla 12: Medidas según definición de consulta para el conjunto de pruebas.

	Precisión YES	Cobertura YES	Precision NO	Cobertura NO
Desc	0,6781	0,1620	0,5205	**0,9223**
Desc&Title	0,6076	0,1394	0,5126	0,9101
Title	**0,6815**	**0,1856**	**0,5316**	0,9142
	Medida-F YES	**Medida-F NO**	**Medida-F Global**	**Exactitud**
Desc	0,2615	0,6654	0,4635	0,5395
Desc&Title	0,2268	0,6558	0,4413	0,5237
Title	**0,2917**	**0,6722**	**0,4820**	**0,5519**

En la tabla que se acaba de mostrar de los resultados para el conjunto de pruebas, aparece resaltado el resultado más alto para cada una de las medidas tomadas. Se puede ver claramente cómo salvo para la cobertura para la categoría NO, el resto de medidas son siempre mejores cuando la definición de la consulta está formada únicamente por el título del tema.

A continuación se ven gráficamente la medida-F para cada una de las categorías (primera fila), la medida-F global (segunda fila, izquierda) y la exactitud (segunda fila, derecha), esta vez para el conjunto de validación.

Tabla 13: Medidas según definición de consulta para el conjunto de validación.

	Precisión YES	Cobertura YES	Precision NO	Cobertura NO
Desc	0,6725	0,1574	0,5282	**0,9249**
Desc&Title	0,6236	0,1405	0,5131	0,9151
Title	**0,7089**	**0,1921**	**0,5337**	0,9115

	Medida-F YES	Medida-F NO	Medida-F Global	Exactitud
Desc	0,2551	0,6724	0,4638	0,5450
Desc&Title	0,2293	0,6575	0,4434	0,5259
Title	**0,3023**	**0,6760**	**0,4892**	**0,5575**

Se puede ver cómo se mantiene la proporción de los resultados al cambiar de conjunto. Como se ha determinado que se optimizará para la precisión de la categoría YES, se toma para el resto de la evaluación el título como campo que determinará la consulta.

5.3.1.2 Expansión de Términos

Una vez que ya se ha determinado la definición de la consulta que va a ser utilizada, se pasará a intentar mejorar los datos obtenidos mediante una expansión de términos realizada mediante realimentación.

Tabla 14: Medidas finales para definición de consulta resultante.

Medida-F YES	Medida-F NO	Medida-F Global	Exactitud
0,3024	0,6760	0,4892	0,5576
Precisión YES	**Cobertura YES**	**Precisión NO**	**Cobertura NO**
0,7089	0,1921	0,5338	0,9215

El proceso de expansión de términos, como se ha explicado anteriormente tiene dos parámetros configurables, el número de documentos de los que se van a tomar términos (M), y el número de términos a utilizar de cada uno de ellos (N). Esta combinación dará lugar a un número de términos que se añadirán a las definiciones de las consulta.

Es importante resaltar que el número inicial de términos presente en las consultas es bastante pequeño (99 términos en total para los 28 temas) por lo que a la mínima expansión que se realice se obtendrá un crecimiento considerable. Por ejemplo, la expansión mínima, un término del documento más similar a la consulta, teniendo en cuenta que la media de términos por consulta está entre 3 y 4, supondría en el mejor de los casos un crecimiento del 25%.

La expansión de términos realizada se reflejará en las nuevas similitudes calculadas. Al ampliar significativamente el número de términos presentes en la consulta, más documentos tendrán una similitud distinta de cero. Este cambio en las similitudes se puede ver en los diagramas de dispersión incluidos en la siguiente figura.

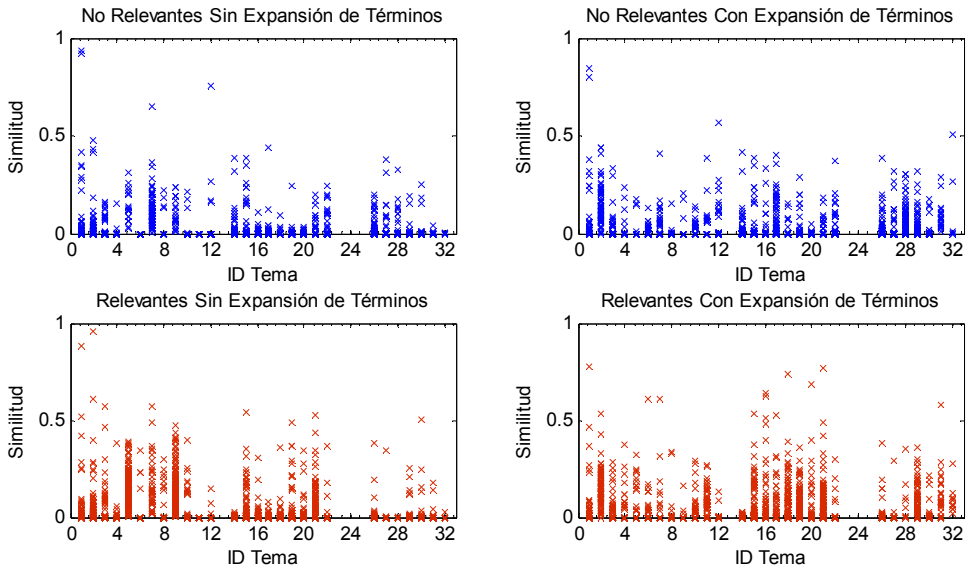

Figura 74: Diagrama de dispersión tras la expansión de términos (M=15, N=7).

Se puede apreciar cómo en los diagramas de la columna derecha los datos están más expandidos, habiendo una concentración ligeramente mayor para valores de similitud distintos de cero respecto a los representados en la columna de la izquierda (es decir los datos sin expansión de términos).

En la evaluación se realizarán simulaciones para distintas combinaciones de estos dos parámetros y se anotará cuántos términos se ganan para cada una de ellas, de forma que se pueda comparar el valor obtenido para las Medidas-F de las dos categorías con las que se está trabajando. Es importante recordar que el crecimiento del número de términos añadidos no va a ser proporcional a M*N ya que únicamente se tendrán en cuenta los documentos con similitud no nula y los términos con peso no nulo, de ahí la necesidad de saber exactamente cuántos documentos se añaden para cada combinación.

Esta relación entre términos añadidos y resultados es precisamente esto lo que se puede observar en la Figura 75. Todas estas medidas se realizan con el conjunto de pruebas, que como ya se ha mencionado, es el que se utiliza para hallar los valores óptimos de los distintos parámetros del sistema.

Cada una de las combinaciones simuladas (34 en total) fueron repetidas varias veces, cada una de ellas con una separación lógica de conjuntos distinta para minimizar la dependencia de los datos escogidos. El valor resultante de estas simulaciones está calculado como ya se ha mencionado: tomando la media de las muestras resultantes tras eliminar los considerados mejor y peor resultado tomando, como se hará a lo largo del proyecto la medida-F de la categoría YES como factor determinante.

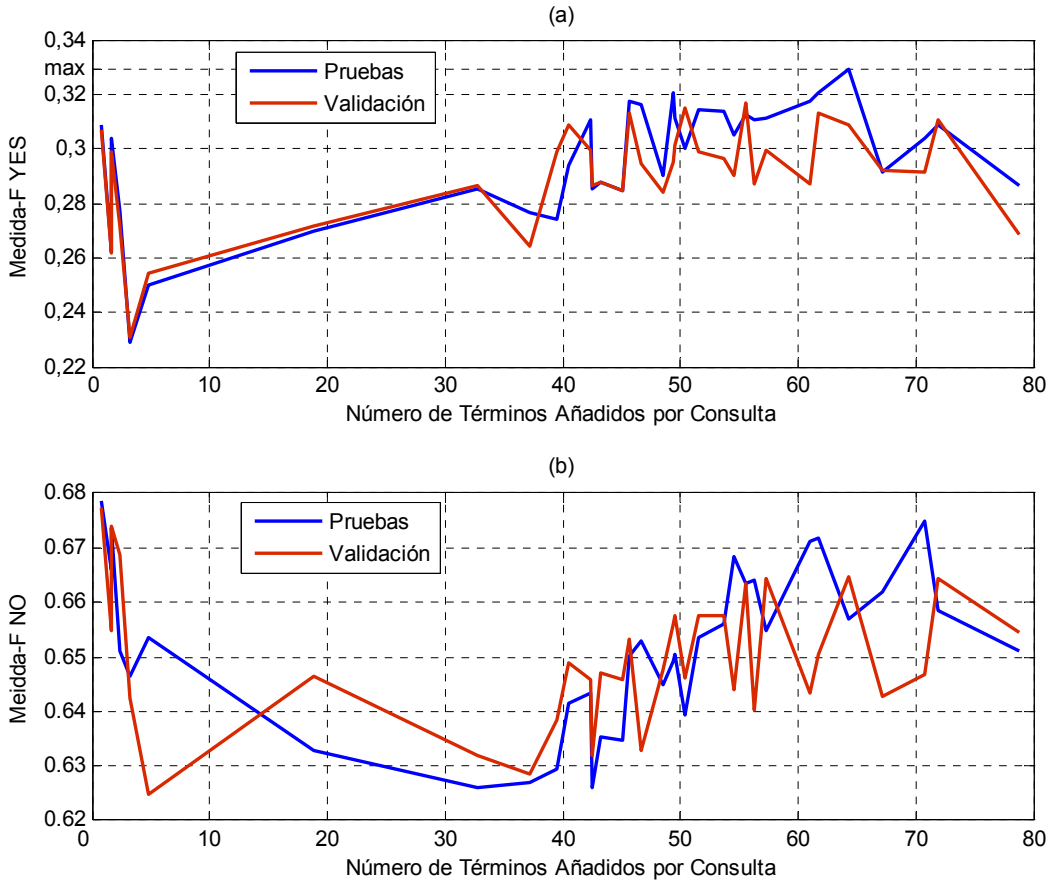

Figura 75: Medida-F por Categorías vs. Número de Términos Añadidos por Consulta.

Lo primero que se puede apreciar en esta gráfica es que a pesar de no ser exactamente iguales, el comportamiento para los dos conjuntos es bastante similar. Si se seleccionan los tres mejores valores según los resultados del conjunto de pruebas de medida-F YES (64, 62 y 49 términos añadidos por consulta), se obtiene la siguiente tabla.

Tabla 15: Mejores resultados de la expansión de términos.

Parámetros		Por Categoría		Globales	
Número de documentos	**Número de Términos**	**Medida-F YES**	**Medida-F NO**	**Medida-F Global**	**Exactitud**
13	6	0,3207	0,6493	0,4850	0,5374
14	7	0,3204	**0,6717**	**0,4961**	**0,5573**
15	7	**0,3291**	0,6569	0,4930	0,5460

Para cada una de las medidas se ha marcado qué combinación de parámetros da el mejor resultado. Como se ha hecho hasta ahora, se toma la medida-F YES como valor a optimizar en los datos relativos al conjunto de pruebas por lo que el mejor resultado es el obtenido para 15

documentos y 7 términos para cada definición de consulta (aproximadamente una media de 64 términos añadidos a cada consulta).

A continuación se muestran los valores del conjunto de validación correspondientes a los parámetros seleccionados como óptimos.

Tabla 16: Comparación de resultados tras expansión de términos.

	Medida-F YES	Medida-F NO	Medida-F Global	Exactitud
Sin Expansión	0,3024	0,6760	0,4892	0,5576
Con Expansión (15,7)	0,3085 (+2,02%)	0,6648 (-1,66%)	0,4866 (-0,53%)	0,5485 (-1,63%)
	Precisión YES	**Cobertura YES**	**Precisión NO**	**Cobertura NO**
Sin Expansión	0,7089	0,1921	0,5338	0,9215
Con Expansión (15,7)	0,6078 (-14,26%)	0,2066 (+7,55%)	0,5366 (+0,52%)	0,8732 (-5,24%)

Se puede ver que apenas hay diferencia entre los valores obtenidos para la expansión de términos y los que se habían obtenido sin ella, por lo tanto cabe plantearse si este procesado adicional tiene alguna influencia en los resultados.

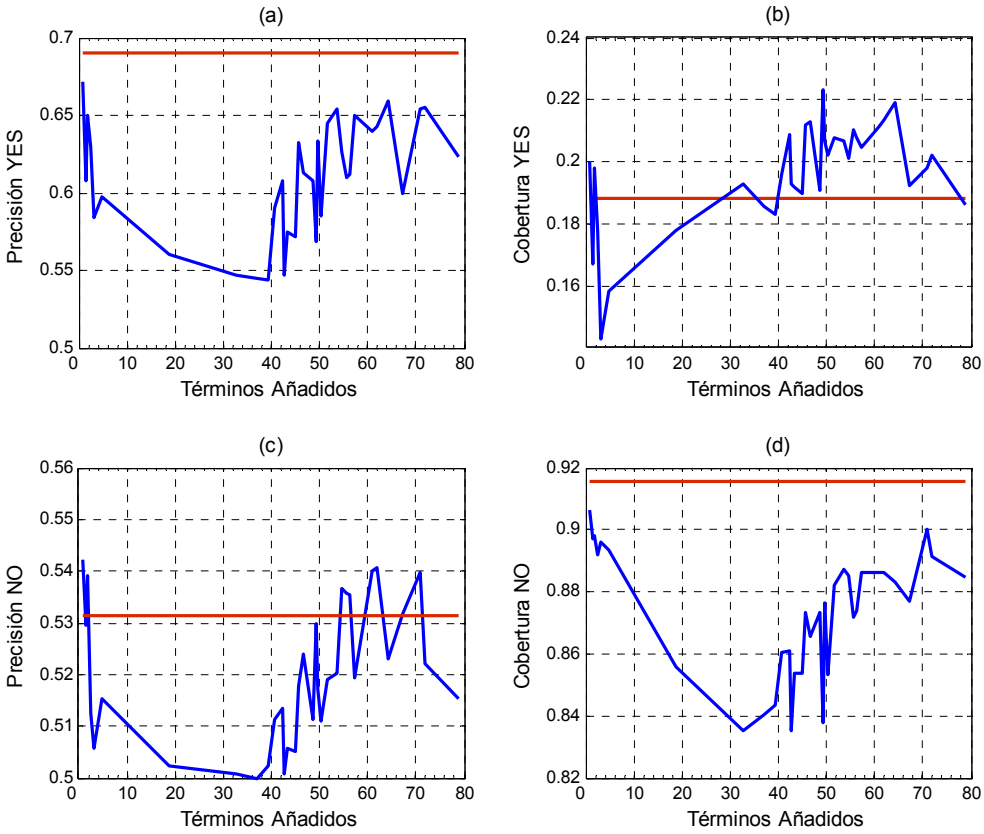

Figura 76: Influencia de la expansión de términos en la precisión y cobertura.

Para comprobar precisamente esto, en la Figura 76 se pueden ver la precisión y cobertura de las dos categorías con las que se está trabajando y su evolución según el número de términos añadidos. Se ha incluido además en cada una de las gráficas una línea continua que representa el valor de ese parámetro para el caso en el que no se usa la expansión de términos.

Se ve claramente cómo la medida más beneficiada por la expansión es la cobertura de la categoría YES, para la que se obtienen valores mejores en un rango considerable de valores de términos añadidos. La precisión de la clase NO también mejora, aunque de forma mucho más irregular. Por otro lado, tanto la precisión de YES como la cobertura de NO, reducen sus valores para cualquier número de términos añadido.

Estos comportamientos llevan a dos conclusiones: la primera es el claro compromiso que existe en el sistema a la hora de mejorar una de las medidas de una categoría. La mejora de una de ellas, conlleva obtener peores resultados para la otra, de ahí que elegir qué medida del sistema se va a priorizar a la hora de realizar las optimizaciones sea importante de cara a los resultados que se van a obtener.

La segunda conclusión es que, los resultados vistos en la Tabla 16 son tan parecidos entre sí debido a este equilibrio entre las medidas, que provoca que las medidas globales en los dos casos (con expansión y sin expansión) sean prácticamente iguales.

5.3.2 Opinión

La segunda fase de la evaluación es la relativa al análisis de opinión. Al igual que para el análisis de relevancia habrá dos fases bien definidas, una fase inicial enfocada en las distintas configuraciones utilizadas para el cálculo de la opinión, y una segunda fase en la que se verán la influencia en los resultados obtenidos al añadir una etapa adicional de agrupamiento antes de utilizar el umbral lineal.

Asimismo hay que tener en cuenta que mientras el cálculo básico de la opinión no depende de la medida de similitud usada para determinar la relevancia, el algoritmo usado en la etapa de agrupamiento sí que la utilizará, y por lo tanto el uso o no de expansión de términos influirá en los resultados.

5.3.2.1 Configuración del Cálculo de la Opinión

Como ya se ha mencionado, la primera fase de evaluación de resultados de la etapa del clasificador de opinión se va a centrar en evaluar las distintas configuraciones de cálculo de opinión.

Hay que recordar que se escogieron tres configuraciones distintas a comparar, "*4tags (2)*" y "*4tags (3)*" del primer tipo, en las que se asigna un valor a cada uno de los símbolos asociados a las etiquetas que definen el valor semántico, y "*6tags*", del segundo tipo, que en vez de asignar valores a los símbolos, se lo asigna a combinaciones de ellos.

Como se viene haciendo, para comparar los resultados de estas tres configuraciones se realizarán varias simulaciones con distintas divisiones por conjuntos y se aplicará la media truncada (ver apartado 5.2) para obtener así los valores de la medida-F de las dos categorías con las que se está trabajando para el conjunto de pruebas.

Estos resultados se obtienen usando un umbral lineal análogo al utilizado para obtener los valores de relevancia en apartados anteriores.

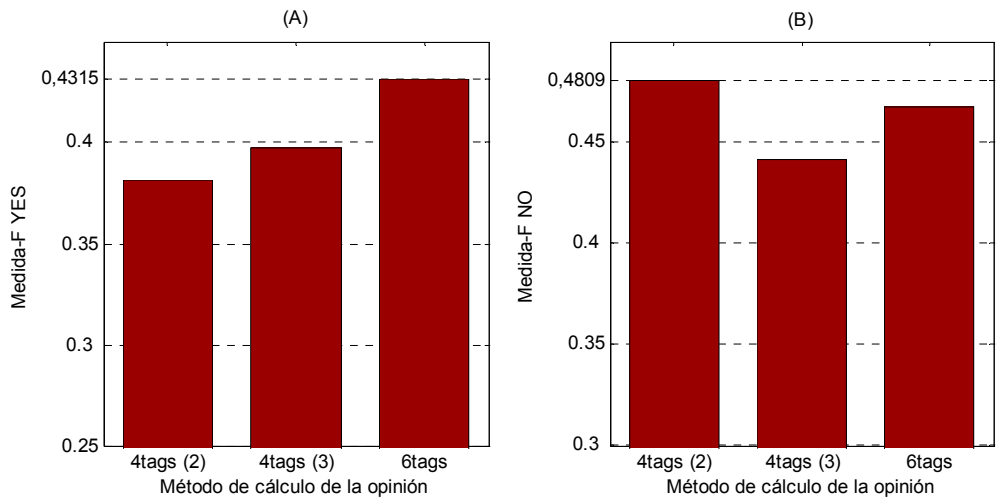

Figura 77: Medida-F por categoría según cálculo de opinión para el conjunto de pruebas.

Al igual que en casos anteriores se puede ver cómo dependiendo de la categoría a la que se vaya a dar prioridad, se optará por una de las configuraciones u otra. En este caso, se está priorizando la clase YES, por lo que se escogerá "*6tags*" como configuración para el cálculo de la opinión.

En la Tabla 17 se incluyen los resultados para el conjunto de validación de la configuración que ha dado un mejor resultado para el conjunto de pruebas.

Tabla 17: Medidas resultantes según definición de consulta para el conjunto de validación.

6tags			
Medida-F YES	**Medida-F NO**	**Medida-F Global**	**Exactitud**
0,42823	0,44571	0,437	0,4371
Precisión YES	**Cobertura YES**	**Precisión NO**	**Cobertura NO**
0,2927	0,7349	0,7484	0,2900

Para el resto de los análisis de opinión se usarán esta configuración y al haber utilizado para evaluar los resultados un clasificador lineal análogo al usado para la relevancia, estos mismos resultados valdrán para evaluar la influencia de añadir al clasificador una etapa de agrupamiento.

5.3.2.2 Usando Agrupamiento

Tras evaluar los resultados para un clasificador lineal, se pasará a ver los resultados obtenidos al añadir la fase de agrupamiento. Como ya se explicó, habrá un grupo por tema, y el número de documentos que se van a incluir en cada uno de ellos será configurable mediante el parámetro k, introducido por el usuario durante la ejecución. Debido a la forma de calcular el umbral esto es equivalente a reducir las muestras de entrenamiento a k grupos, siendo cada uno de los documentos seleccionados el representante de su grupo.

La primera simulación que se va a realizar es ver la variación de la medida-F para la categoría YES según el número de documentos que se incluyan en cada grupo. Dentro de la gráfica se incluye una línea continua que representa el resultado que se obtiene sin agrupamiento.

Se comprueba que la variación de resultados para valores de k bajos es bastante más acusada que para valores altos, por lo que el número de muestras tomadas en los distintos rangos de k se selecciona de acuerdo a esta circunstancia, realizando un mayor muestreo en las zonas con más variación en los resultados.

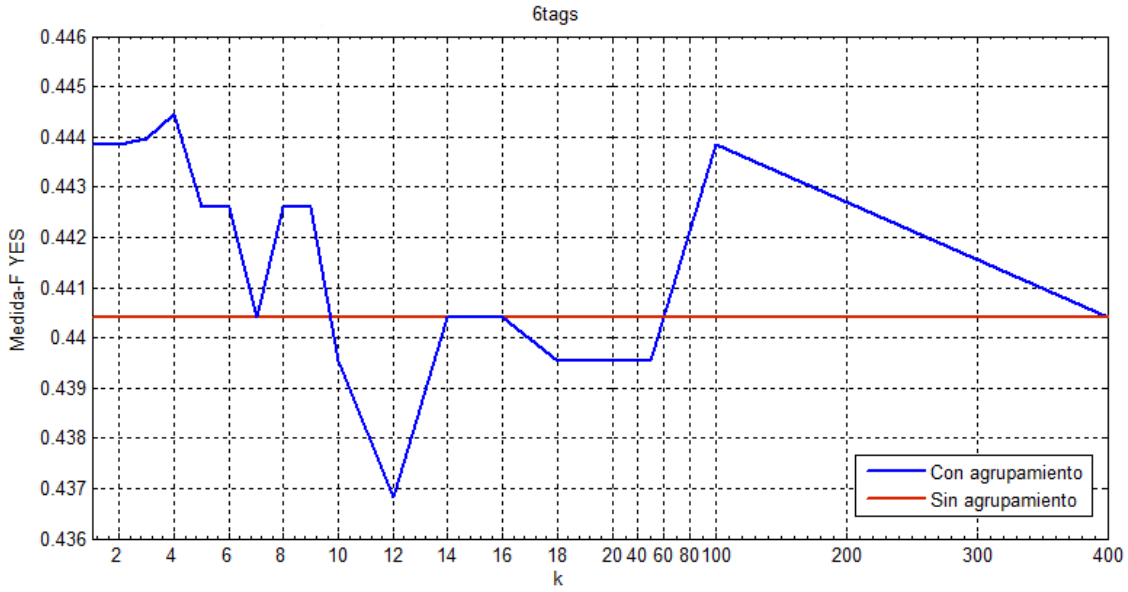

Figura 78: Medida-F YES según el número de grupos.

Es importante resaltar que en todas las simulaciones se comprobó que para un número lo suficientemente alto de k (concretamente k=400), el número de documentos tomados para el cálculo del umbral era el total de los documentos del conjunto de entrenamiento, y que por tanto el umbral resultante coincidía con el calculado sin la etapa de agrupamiento.

Tabla 18: Documentos de entrenamiento usados según número de grupos.

k	1	2	3	4	5	6	7	8	9
Nº Documentos	28	56	84	112	140	168	196	224	252
k	10	12	14	16	18	20	25	50	100
Nº Documentos	280	336	392	448	504	559	694	1338	2311

En la Tabla 18 se puede ver una referencia de la relación entre el número de grupos seleccionados y el número de documentos que se usan para el entrenamiento del umbral de decisión. Se puede ver que es a partir de k=20 grupos cuando el número de documentos deja de ser 28*k por no haber suficientes documentos para alguno de los temas.

Si se seleccionan los tres valores de k para los que se obtienen mejores resultados de la medida-F YES en el conjunto de pruebas, al comprobar los resultados para el conjunto de validación, se obtienen los siguientes resultados.

Tabla 19: Medidas según número de grupos.

k	% Documentos	Precisión YES	Cobertura YES	Precision NO	Cobertura NO
4	3,01 %	0,2908	0,7920	0,7936	0,2929
3	2,26 %	0,29037	0,7920	0,79276	0,29142
2	1,50 %	0,29151	0,7840	0,79264	0,30251
k	**% Documentos**	**Medida-F YES**	**Medida-F NO**	**Medida-F Global**	**Exactitud**
4	3,01 %	0,4254	0,4279	0,426635	0,4266
3	2,26 %	0,4249	0,4262	0,425554	0,4255
2	1,50 %	0,4250	0,4380	0,431487	0,4315

El porcentaje de documentos utilizado para obtener la medida, el 100%, se obtiene del número total de documentos que se usaría en el caso de no utilizar agrupamiento, es decir, el total de documentos de entrenamiento, 3718 documentos.

Usando la relación vista en la Tabla 18, los valores obtenidos para el valor k definido como óptimo (k_{opt} = 4) y los resultados de clasificar sin usar una etapa de agrupamiento, se puede generar la siguiente tabla resumen.

Tabla 20: Comparación resultados con y sin etapa de agrupamiento.

	Documentos Entrenamiento (%)	Medida-F YES	Medida-F NO	Medida-F Global	Exactitud
k = 0	100%	0,4282	0,4457	0,437	0,4371
k = 4	3,01 %	0,4254 (-0,66%)	0,4279 (-4%)	0,4266 (-2,37%)	0,4266 (-2,40%)

Se puede ver claramente que se obtienen valores prácticamente iguales a los obtenidos sin fase de agrupamiento, pero usando únicamente un 3% de los datos del conjunto de entrenamiento, y por tanto reduciendo enormemente la carga computacional del proceso.

5.3.2.2.1 Usando Expansión de Términos

La dependencia de la etapa de agrupamiento de la medida de similitud hace que el uso o no de expansión de términos en su cálculo afecte a los resultados obtenidos para el umbral de opinión.

Para ver este efecto, se realizan medidas análogas a las que se han visto en el apartado anterior, pero considerando la expansión de términos vista en el apartado 5.3.1.2. Como se determinó en dicho apartado se tomarán los parámetros para los que se obtuvo mejor valor de la

medida-F para la categoría YES. Así mismo, como se ha determinado en el apartado 5.3.2.1, la opinión se calculará usando la configuración "*6tags*".

En la gráfica de la Figura 79 se representan los resultados de medida-F para la categoría YES para distintos valores de k; además se representa mediante una línea continua el valor obtenido para el mismo escenario sin etapa de agrupamiento y poder así comparar ambos más fácilmente.

Se puede observar cómo para k=400 el valor obtenido converge con el valor obtenido para el escenario sin agrupamiento y cómo para los valores bajos de k existe una variación mayor en los resultados.

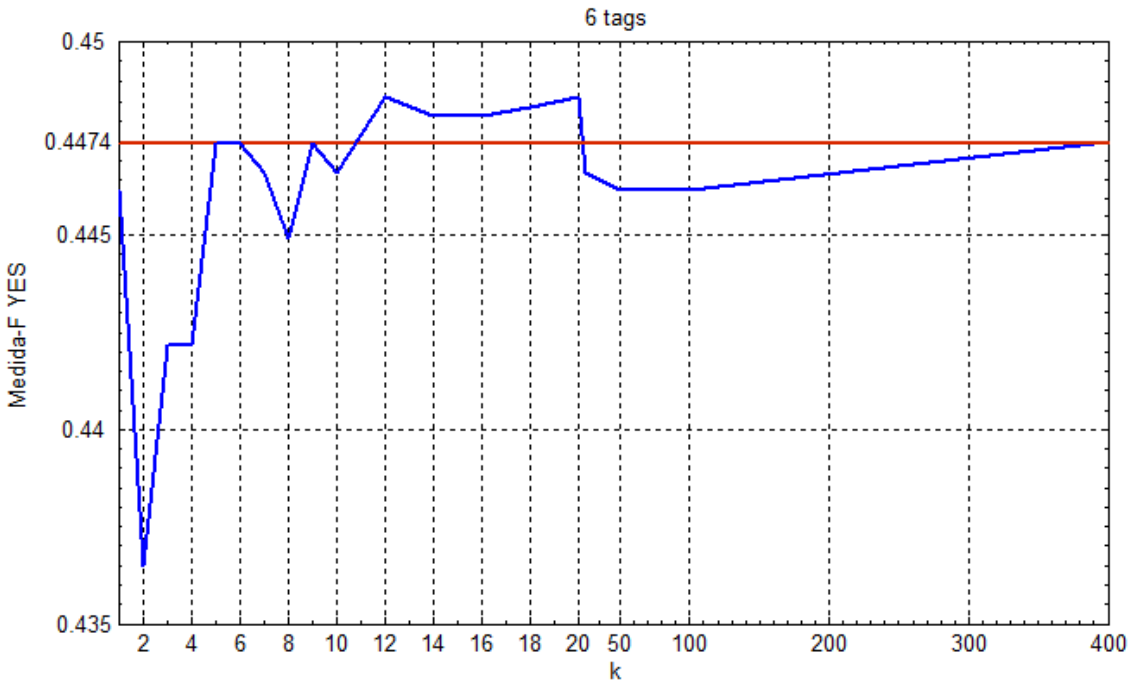

Figura 79: Medida-F YES según el número de grupos y con expansión de términos.

En la Figura 79 se puede ver cómo el rango de valores de k para los que se obtienen resultados por encima del obtenido para el escenario sin agrupamiento (13 valores de k en el intervalo [11, 23]), aunque es parecido al obtenido cuando no se usa expansión de términos (9 valores de k en el intervalo [1, 9]), el número de grupos necesarios para mejorar es bastante más alto. Es importante recordar que un número de grupos más alto supondrá un mayor número de documentos necesarios para entrenar el umbral.

Los siguientes resultados se corresponden a las tres mejores medidas identificadas en la Figura 79 evaluadas para el conjunto de validación.

Tabla 21: Medidas según número de grupos con expansión para el conjunto de validación.

k	Precisión YES	Cobertura YES	Precisión NO	Cobertura NO
20	0,2972	0,8287	0,8466	0,32533
12	0,2972	0,8287	0,8466	0,32533
18	0,2970	0,8287	0,8463	0,3246

k	Medida-F YES	Medida-F NO	Medida-F Global	Exactitud
20	0,4375	0,4700	0,453766	0,45425
12	0,4375	0,4700	0,453766	0,45425
18	0,437256	0,46923	0,453242	0,45371

Para ver cómo afecta la introducción de la expansión de términos a la etapa de agrupamiento, se comparan los resultados obtenidos para las cuatro medidas principales definidas en la siguiente tabla:

Tabla 22: Comparación resultados con y sin agrupamiento y expansión de términos.

	Documentos Entrenamiento (%)	Medida-F YES	Medida-F NO	Medida-F Global	Exactitud
k = 0	100%	0,4331	0,4818	0,45745	0,4585
k = 20	15,03 %	0,4375 (+1,02%)	0,470 (-2,45%)	0,4538 (-0,80%)	0,4542 (-0,94%)

De nuevo se puede ver cómo se obtienen resultados muy parecidos para la medida-F de la categoría YES sin necesidad de usar todos los datos de entrenamiento, y por lo tanto reduciendo la carga computacional. También se puede ver cómo esta reducción no es tan grande como en el caso en el que no se usaba expansión de términos, pero sigue siendo lo suficientemente significativo.

5.4 Comparación con los Participantes en NTCIR-6

Una vez vistos todos los escenarios y configuraciones evaluados, se pueden comparar a los resultados obtenidos por el resto de los participantes en NTCIR-6 [43].

Siguiendo el formato en el que se presentan los resultados en el resumen de la tarea dado por los organizadores de NTCIR-6 y teniendo en cuenta las distintas posibilidades que se han evaluado, se considerará que hay cuatro ejecuciones distintas en este proyecto:

- Relevancia sin realimentación y opinión sin agrupamiento (NFB – NGR)
- Relevancia con realimentación y opinión sin agrupamiento (FB – NGR)
- Relevancia sin realimentación y opinión con agrupamiento (NFB – GR)
- Relevancia con realimentación y opinión con agrupamiento (FB – GR)

Cabe mencionar que estas cuatro ejecuciones no implican cuatro resultados distintos para cada una de las clasificaciones, sino cuatro posibles configuraciones para el sistema. El ejemplo más claro, y ya explicado en apartados anteriores, es cómo al utilizar expansión de términos en el clasificador de relevancia no afectará a la clasificación de opinión siempre y cuando no haya etapa de agrupamiento.

En la siguiente tabla se incluyen los resultados de los participantes en NTCIR-6 así como el valor medio de los mismos de cara a facilitar la comparación. Hay que recordar que la única tarea de obligatoria para participar en NTCIR-6 MOAT era la clasificación de opinión, de ahí que algunos de los participantes no tengan resultados de relevancia.

Tabla 23: Resultados Participantes NTICR-6.

Grupo	RELEVANCIA			OPINIÓN		
	Precisión	Cobertura	Medida-F	Precisión	Cobertura	Medida-F
IIT-1	-	-	-	0,325	0,588	0,419
IIT-2	-	-	-	0,259	0,854	0,397
TUT	0,392	0,597	0,473	0,310	0,575	0,403
Cornell	-	-	-	0,317	0,651	0,427
NII	0,510	0,322	0,395	0,325	0,624	0,427
GATE	0,286	0,632	0,393	0,324	0,905	0,477
ICU-IR	0,409	0,263	0,320	0,396	0,524	0,451
Valor Medio	0,3993	0,4535	0,3953	0,3223	0,6744	0,4287

A continuación, y siguiendo el mismo formato que se ha visto para los participantes de NTCIR-6, se incluyen dos tablas, Tabla 24 y Tabla 25, en las que se resumen los resultados obtenidos por el sistema implementado en este proyecto para las configuraciones descritas junto al valor medio obtenido para cada medida para las cuatro configuraciones.

En la Tabla 24 se han particularizado los resultados para la categoría YES tanto en la clasificación de opinión como en la de relevancia. En la Tabla 25, los resultados incluidos son los globales del sistema, es decir, teniendo en cuenta todas las categorías de cada una de las clasificaciones.

Tabla 24: Resultados particularizados para las categorías YES.

Grupo	RELEVANCIA			OPINIÓN		
	Precisión YES	Cobertura YES	Medida-F YES	Precisión YES	Cobertura YES	Medida-F YES
NFB-NGR	0,7089	0,1921	0,3024	0,2927	0,7349	0,42823
FB-NGR	0,6078	0,2066	0,3985	0,2927	0,7349	0,42823
NFB-GR	0,7089	0,1921	0,3024	0,2908	0,7920	0,4254
FB-GR	0,6078	0,2066	0,3985	0,2972	0,8287	0,4375
Valor Medio	0,6584	0,1994	0,3054	0,2934	0,7726	0,4298

Si se comparan los valores de precisión y cobertura obtenidos en este proyecto en la clasificación de opinión, se puede ver cómo se ha obtenido una precisión ligeramente inferior a la media (en torno a un 10% más pequeña) mientras que la cobertura está por encima, siendo el

tercer mejor valor. Este compromiso entre ambos valores resulta en un valor de medida-F muy parecido al obtenido por los participantes.

En el clasificador de relevancia las diferencias entre los resultados de los participantes y los de este sistema son bastante más acusadas. En el sistema implementado el valor de la precisión está bastante por encima del obtenido para la cobertura. Esta tendencia no es igual para todos los participantes, resultando en un valor medio de precisión y cobertura parecido. Al comparar estos valores con los de este sistema, se observa que el valor de precisión que se obtiene es bastante mejor que el de cualquiera de los participantes mientras que el obtenido para la cobertura es bastante menor que cualquiera de ellos. En la medida-F, el buen valor obtenido para la precisión no es suficiente para contrarrestar el valor obtenido de cobertura, por lo que también se resulta estar por debajo de la media del de los participantes en NTCIR-6.

Tabla 25: Resultados globales.

Grupo	RELEVANCIA			OPINIÓN		
	Precisión Global	Cobertura Global	Medida-F Global	Precisión Global	Cobertura Global	Medida-F Global
NFB-NGR	0,62135	0,5568	0,4892	0,52055	0,51245	0,437
FB-NGR	0,5722	0,5399	0,4866	0,52055	0,51245	0,437
NFB-GR	0,62135	0,5568	0,4892	0,5422	0,54245	0,4266
FB-GR	0,5722	0,5399	0,4866	0,5719	0,577015	0,45376
Valor Medio	0,5968	0,5483	0,4879	0,5388	0,5361	0,4386

En la Tabla 25 se muestran los valores globales del sistema. Al compararlos con los valores cuando se particularizaba para la categoría YES, se nota una clara mejoría en las medidas que en ese caso eran muy bajas. Como se ha comentado en apartados anteriores, debido a la distribución del corpus con el que se ha trabajado, los resultados para la categoría NO son en general mejores, de ahí que la diferencia entre las medidas de cobertura y precisión sea menos acusada al evaluar el resultado global.

Para el clasificador de opinión, la diferencia entre una medida y otra es muy ligera, quedándose la medida-F ligeramente por encima de la obtenida por los participantes. En el caso del clasificador de relevancia se aprecia mucho más la influencia de la distribución del corpus, puesto que se pasa de tener una medida-F peor que cualquiera de las de los participantes a tener un valor mejor. El valor medio de medida-F Global es un 19% mejor que el valor medio de los participantes.

A continuación se incluyen varias gráficas mostrando histogramas con los mejores resultados de cada uno de los grupos participantes y los resultados obtenidos en este proyecto, tanto para el clasificador de relevancia como para el de opinión. Al igual que se ha hecho en el resto del apartado, se han tenido en cuenta por un lado los resultados particularizados para la categoría YES (Figura 80) y por otro lado, los resultados globales teniendo en cuenta todas las categorías de cada uno de los clasificadores (Figura 81).

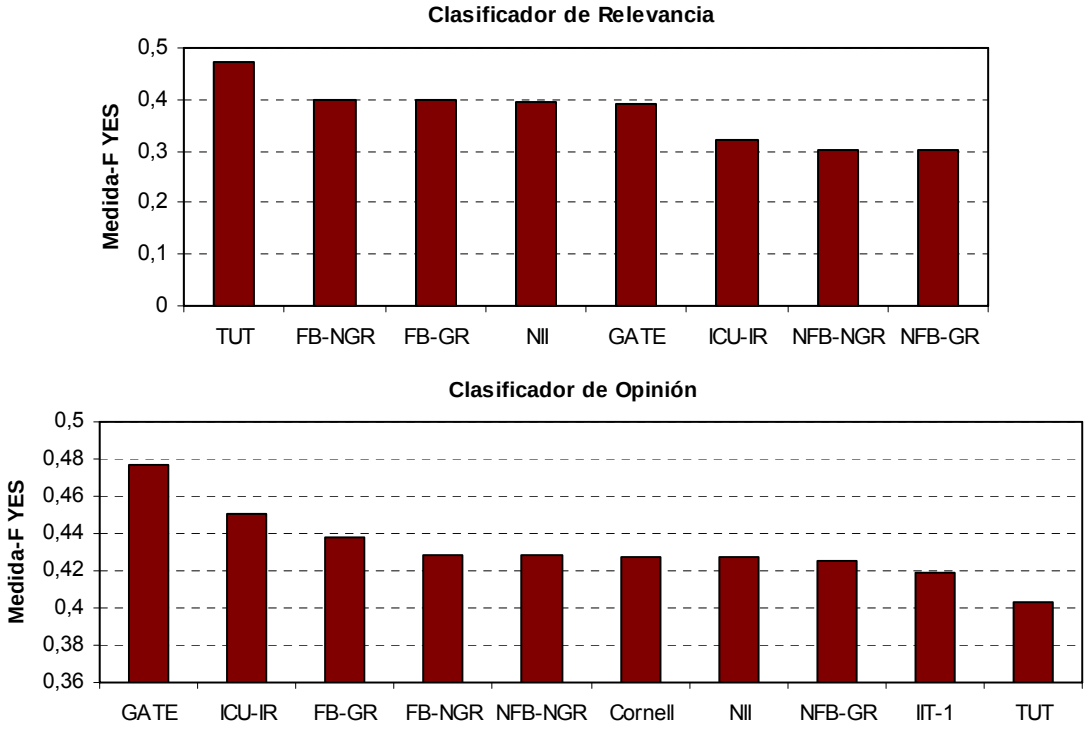

Figura 80: Comparación con grupos participantes en NTCIR-6 (Medida-F YES).

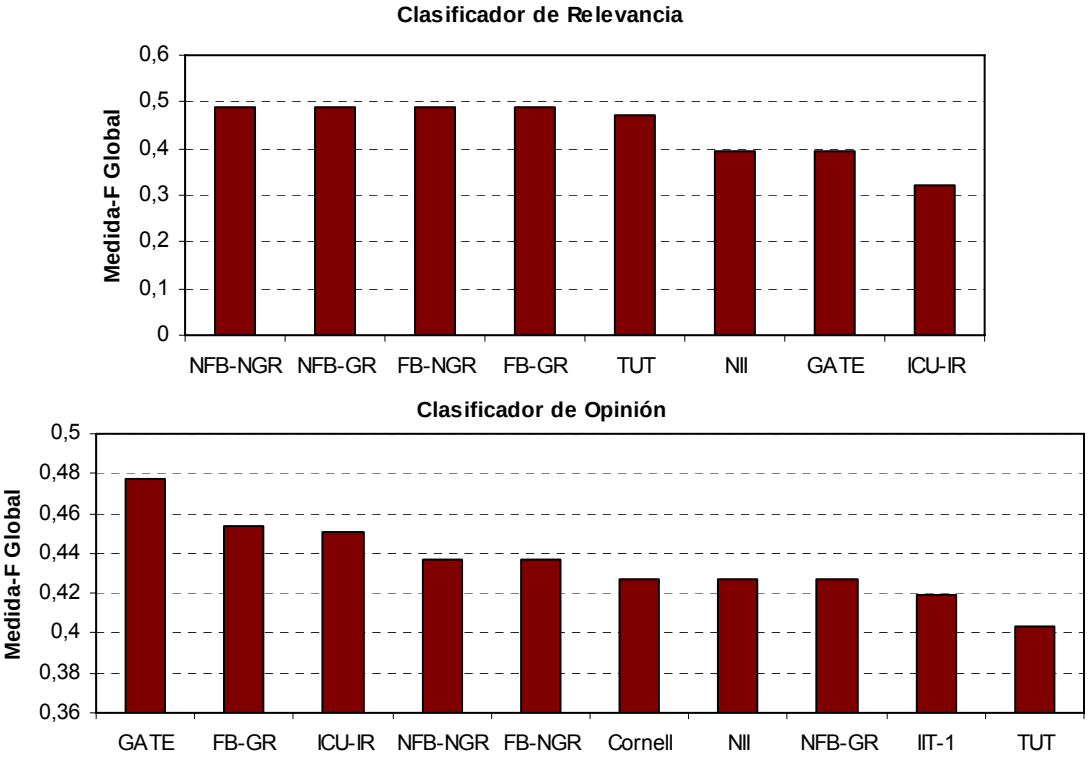

Figura 81: Comparación con grupos participantes en NTCIR-6 (Medida-F Global).

6 Conclusiones y Trabajos Futuros

6.1 Conclusiones

En este proyecto se ha realizado una investigación y se ha presentado una implementación de un sistema de clasificación afectiva doble: por un lado se evalúa la relevancia de un documento para un tema dado, y por otro lado se analiza la subjetividad del documento de forma que se pueda determinar si expresa o no una opinión. Teniendo en cuenta que este proyecto tiene un alcance limitado, en vez de plantear los dos clasificadores como sistemas independientes, se intenta que su implementación sea lo más parecida posible.

Al tratarse de una implementación inicial se han evaluado las distintas opciones de diseño básicas para los dos clasificadores.

El primer punto evaluado en el clasificador de relevancia es qué información es la más adecuada para generar la consulta del sistema de entre toda la proporcionada para cada uno de los temas. En dicho proceso de evaluación se ha llegado a la conclusión de que los mejores resultados se obtienen al usar el título del tema para definir la consulta.

Tabla 26: Resumen de los resultados obtenidos para el clasificador de relevancia.

	Medida-F YES	Medida-F NO	Medida-F Global	Exactitud
Sin expansión de términos	0,3024	0,6760	0,4892	0,5576
Con expansión de términos	0,3085 (+2,02%)	0,6648 (-1,66%)	0,4866 (-0,53%)	0,5485 (-1,63%)
	Precisión YES	**Cobertura YES**	**Precisión NO**	**Cobertura NO**
Sin expansión de términos	0,7089	0,1921	0,5338	0,9215
Con expansión de términos	0,6078 (-14,26%)	0,2066 (+7,55%)	0,5366 (+0,52%)	0,8732 (-5,24%)

Tras elegir la consulta del sistema y procesar todos los datos para dejarlos en un formato apropiado para trabajar, se ha utilizado el Modelo de Espacio Vectorial para representar cada uno de los documentos mediante un vector de los términos que contiene. Gracias a esta representación, calcular una medida de similitud entre la consulta definida y cada uno de los documentos es muy sencillo, por lo que se decide usar esa medida de similitud para clasificar los documentos como relevantes (categoría YES) o no relevantes (categoría NO).

Se utiliza un clasificador lineal generado a partir de los datos de entrenamiento y se evalúa mediante la clasificación de los conjuntos de pruebas y de validación. El conjunto de datos de pruebas será el que dé una primera estimación de los resultados, y en el caso de que haya que asignar valores a parámetros, determinará cuáles; el conjunto de validación, evaluado con los valores de parámetros definidos, es el que proporciona los resultados definitivos.

Tras la primera evaluación del clasificador, se observa el problema más significativo a la hora de determinar la relevancia: la medida escogida para ello, la similitud, en un porcentaje bastante importante de los documentos es nula, haciendo que siempre sean clasificados como no

relevantes. Esto se aprecia en los resultados de precisión y cobertura obtenidos para la categoría NO, de un 53% y un 92% respectivamente.

Este problema es la fuente principal de errores del clasificador de relevancia y da lugar a unos resultados pobres para la categoría YES (una medida-F de un 30,24% frente al 67,6% que se obtiene para la categoría NO). Estos resultados afectan a la medida global del sistema, quedándose la medida-F global en un 48,92% y la exactitud en un 55,76%.

Para intentar mitigar este problema, se decide realizar una expansión de términos de la consulta, razonando que al abarcar más términos, se obtendrán más valores de similitud no nulos. Como se vio inicialmente, existen bastantes formas de realizar esta expansión, aunque en este caso se ha decidido usar los resultados de similitud iniciales para añadir términos del propio conjunto de datos, evitando así problemas de dominio.

A nivel de implementación la expansión de términos supone añadir una etapa de realimentación tras la iteración inicial. Esta etapa es configurable, lo que permite evaluar la dependencia del sistema según el número de términos añadidos a las consultas. Los parámetros que se pueden configurar son el número de documentos de los que se extraerán términos, siempre siguiendo el orden de relevancia resultante de la evaluación inicial, y el número de términos que se añadirán a la nueva consulta de cada documento.

Se realizan evaluaciones para distintas combinaciones de estos dos parámetros, obteniendo los mejores resultados cuando se añaden a las consultas 7 términos de los 15 documentos más relevantes, lo que supone unos 64 términos por consulta.

Si se analizan los valores de precisión y cobertura de las categorías YES y NO, se aprecia una ligera mejora en la cobertura de la categoría YES y de la precisión de NO, pero en ningún caso son lo suficientemente significativas como para justificar el procesado adicional que exige esta etapa de realimentación. Los valores globales que se obtienen, en ningún caso son superiores a los obtenidos sin expansión de términos por lo que habría que, o bien desechar la opción de esta expansión, o plantearla de una forma distinta.

En el clasificador de opinión para medir la opinión se ha decidido utilizar una medida obtenida a partir de la información proporcionada por el diccionario afectivo "*General Inquirer*". El diccionario proporciona valores afectivos mediante etiquetas que se transformarán en un valor numérico de forma que se puedan operar y posteriormente clasificar. La primera decisión por tanto será determinar qué valores se asignará a dichas etiquetas.

Se definen dos métodos distintos, uno en el que se asignan valores por etiqueta (y del cual se evalúan variantes con distintos valores, *4tags(2)* y *4tags(3)*) y otro en el que los valores sean asignados a combinaciones de etiquetas (*6tags*). Tras una evaluación inicial utilizando un clasificador lineal análogo al utilizado en el clasificador de relevancia, se observan resultados claramente mejores para el método que considera combinaciones de etiquetas. En estas evaluaciones se obtienen resultados bastante parecidos para las dos categorías, obteniendo una precisión del 74,84% y una cobertura de un 29% para la categoría NO mientras que para la categoría YES se alcanza un 73,49% de cobertura y un 29,27% de precisión.

La discrepancia entre las dos medidas de cobertura y precisión hace que la medida-F global y la exactitud del sistema se queden ambas en torno al 43,7%.

Inicialmente, se planteó implementar el clasificador de opinión usando el algoritmo kNN, tanto por su sencillez conceptual como por los buenos resultados que proporciona. La implementación de este algoritmo hubiese supuesto un cambio significativo respecto al modelo del clasificador de relevancia, hasta el punto de convertirse en un sistema completamente diferente.

A raíz de esto, y tras los resultados observados en las primeras evaluaciones del clasificador de opinión con umbral lineal, se decidió integrar una etapa de agrupamiento previa a la clasificación mediante una versión modificada de k-medias. Con esta etapa lo que se hace es limitar el número de documentos usados para entrenar el clasificador lineal, usando la medida de similitud

calculada en el clasificador de relevancia. Dicho de otra forma, en lugar de utilizar todos los documentos de entrenamiento disponibles, y teniendo en cuenta que no siempre se dispondrá de volúmenes elevados de datos, se entrenará el clasificador de opinión usando los documentos determinados más relevantes.

Tabla 27: Resumen de los resultados obtenidos para el clasificador de opinión.

		Medida-F YES	Medida-F NO	Medida-F Global	Exactitud
Sin agrupamiento		0,42823	0,44571	0,437	0,4371
Con agrupamiento	**Sin expansión (k=4)**	0,4254 (-0,66%)	0,4279 (-4%)	0,4266 (-2,37%)	0,4266 (-2,40%)
	Con expansión (k=20)	0,4375 (+2,16%)	0,4700 (+5,45%)	0,4537 (+3,83%)	0,4542 (+3,92%)
		Precisión YES	**Cobertura YES**	**Precisión NO**	**Cobertura NO**
Sin agrupamiento		0,2927	0,7349	0,7484	0,2900
Con agrupamiento	**Sin expansión (k=4)**	0,2908 (-0,65%)	0,7920 (+7,77%)	0,7936 (+6,04%)	0,2929 (+1%)
	Con expansión (k=20)	0,2972 (+1,54%)	0,8287 (+12,76%)	0,8466 (+13,12%)	0,3253 (+12,18%)

Esta etapa de agrupamiento es configurable, pudiendo elegir cuántos documentos de cada tema se escogerán para calcular el umbral mediante el parámetro k. En la fase de evaluación se obtendrán resultados para distintos valores de dicho parámetro de cara a optimizar el sistema. Los mejores resultados se obtienen usando 4 documentos relevantes de cada uno de los temas para calcular el umbral de opinión de entrenamiento, haciendo un total de 112 documentos de entrenamiento.

Los resultados que se obtienen para estos documentos no suponen una mejora significativa en cuanto a las medidas globales del sistema, pero sí que muestran mejoras locales tanto en la cobertura de la categoría YES como en la precisión de la categoría NO. El detalle que resulta relevante es que añadiendo esta etapa y con ayuda de los datos obtenidos para el clasificador de relevancia, se ha reducido el número de documentos de entrenamiento necesarios.

Los 112 documentos escogidos suponen un 3% del total de documentos de entrenamiento que se usan cuando no existe la etapa de agrupamiento, por lo que el total de datos a procesar, y por tanto la carga computacional de la fase de entrenamiento del clasificador de opinión se ve reducida enormemente sin que empeoren los resultados.

Con esta etapa se consigue tanto reducir de forma muy significativa la complejidad de la fase de entrenamiento del clasificador de opinión como comprobar que la medida que se ha usado para determinar la relevancia es significativa.

No hay que olvidar que al estar usando la misma medida de similitud que la calculada para el clasificador de relevancia, la fase de agrupamiento se verá afectada por el uso de expansión de términos en el clasificador, por lo tanto se realiza un último experimento para comprobar cómo afecta a los resultados del clasificador de opinión cuando la similitud que se usa para agrupar es la generada con una consulta extendida.

En este caso, los mejores resultados se obtienen para 20 documentos relevantes, lo que supone un total de 559 documentos de entrenamiento. La tendencia es la misma que en el caso en el que no se usa expansión: las medidas globales no sufren cambios significativos, aunque sí que son ligeramente mejores que los obtenidos para la evaluación sin agrupamiento. Tanto la medida-F global como la exactitud se quedan en torno al 45,4% casi un 2% por encima de lo visto para el caso sin agrupamiento. Para las medidas de cobertura y precisión por categoría, de nuevo vuelve a mejorar la cobertura para la categoría YES y la precisión para la categoría NO.

En este caso, la reducción de la complejidad no es tan grande como en el caso anterior, pero sigue siendo muy significativa, ya que se usa en torno a un 15% de los datos de entrenamiento disponibles para conseguir resultados ligeramente mejores.

Tabla 28: Resumen comparativo con los participantes de NTCIR-6 de valores medios.

	OPINIÓN			RELEVANCIA		
	Precisión	**Cobertura**	**Medida-F**	**Precisión**	**Cobertura**	**Medida-F**
Participantes NTCIR-6	0,3223	0,6744	0,4287	0,3993	0,4535	0,3953
PFC (YES)	0,2934 (-8,96%)	0,7726 (+14,56%)	0,4298 (0,25%)	0,6584 (+64,88%)	0,1994 (-56,03%)	0,3054 (-22,74%)
PFC (Global)	0,5388 (+67,17%)	0,5361 (-20,51%)	0,4386 (+2,31%)	0,5968 (+49,46%)	0,5483 (+10,90%)	0,4879 (+23,42%)

Al comparar el sistema con los resultados obtenidos por los participantes en NTCIR-6 MOAT, se ha observado cómo para la clasificación de opinión los resultados de este sistema son ligeramente mejores. En el clasificador de relevancia existen diferencias más acusadas sobre todo entre en cuanto a las diferencias en los resultados de precisión y cobertura. En este caso se ve claramente cómo influyen los buenos resultados obtenidos para la categoría NO puesto que las medidas globales son bastante mejores que las obtenidas por los participantes.

Como resumen se puede decir que, a pesar de que los dos sistemas de clasificación no han dado resultados todo lo buenos que se hubiese querido, se consideran satisfactorios puesto que se han conseguido valores equiparables a los grupos de investigación punteros en el mundo del análisis afectivo dentro del esfuerzo y ámbito de un proyecto fin de carrera.

Además, teniendo en cuenta los datos de los que se partía y la limitación de intentar tener un sistema común para ambos clasificadores para reducir tanto la complejidad de implementación como para facilitar la identificación de errores, los resultados son muy positivos de cara a continuar la investigación sobre ambos temas en el futuro.

6.2 Trabajos Futuros

A lo largo del proyecto se han identificado numerosos puntos que no se han podido estudiar con la profundidad deseada por estar fuera del alcance del mismo y que son interesantes tanto como líneas de investigación independientes como de cara a mejorar las prestaciones del sistema.

A continuación se detallan algunos de estos puntos:

- Durante las numerosas evaluaciones del sistema se ha comprobado cómo la parte del sistema con mayor carga computacional es el cálculo de la matriz de entrenamiento utilizada para el cálculo de la similitud a través de cual se analiza la relevancia. Precisamente por esto, y de cara a optimizar el sistema, una mejora interesante sería una reducción de las dimensiones de la matriz. En el método implementado, cuando un término no aparece en un documento se guarda como un cero, algo que al estar trabajando con matrices dispersas supone un uso poco eficiente de memoria. Una forma de reducir las dimensiones de la matriz es omitiendo todas estas direcciones de memoria en las que no se guarda nada, de forma que no guardar un valor indique directamente que no aparece el término en el documento.
- Tanto para la obtención de los lemas en la fase de procesado inicial como en el cálculo de opinión se han usado dos herramientas muy concretas (el analizador semántico "*Freeling*" y el diccionario afectivo "*General Inquirer*" respectivamente), por lo que los resultados del sistema puedan estar definidos por ambas. Una posible línea de investigación sería la utilización de otras herramientas análogas que permitan ver el grado de influencia de las mismas en el sistema.
- Una de las desventajas de haber utilizado el mismo procesado inicial tanto para la clasificación de relevancia como para la de opinión es el hecho de que se incluya en la segunda la eliminación de palabras de parada. Dentro de las listas habituales de palabras de parada se incluyen palabras con un fuerte valor semántico por lo que sería una buena opción evaluar la calidad del clasificador eliminando esta fase del procesado inicial.
- En el análisis de opinión, a pesar de haber tenido en cuenta tanto valores semánticos positivos como negativos, en ningún momento se han diferenciado los resultados obtenidos según el signo de los mismos. Una posible línea de investigación sería utilizar estos valores para estimar la polaridad de la opinión detectada.
- De forma análoga, en las tareas planteadas por el taller de NTCIR, se contempla la posibilidad de analizar, en el caso de que haya una opinión, quién la expresa (*opinion holder*), por lo que es otra línea de investigación abordable.
- En las distintas técnicas utilizadas, se ha procurado no tener en cuenta particularidades del idioma, de forma que adaptar el sistema de un idioma a otro fuese sencillo. Los sistemas particularizados para un idioma suelen dar mejores resultados, especialmente en clasificación afectiva, por lo que podría investigarse características de idiomas.
- Se ha visto que existen numerosas técnicas aplicables en las distintas áreas que se han estudiado como por ejemplo el uso de características del documento como representación de los mismos en lugar del Modelo de Espacio Vectorial, por lo que un estudio comparativo de cara a ver qué da mejor resultado para un corpus dado es una opción muy interesante con vistas a optimizar el sistema.
- En esta misma línea, se han mencionado diferentes algoritmos de aprendizaje tales como las redes neuronales, SVM, etc., por lo que otra línea de estudio interesante sería utilizar cualquiera de estas otras técnicas.
- Desde el punto de vista de los idiomas, como se ha visto la tarea MOAT está orientada al análisis multilingüe, por lo que una posible línea de investigación es la aplicación del sistema a distintos idiomas, ya sean japonés y chino, de los que se dispone de corpus en MOAT o español, tras encontrar un corpus adecuado.

Anexo A – Temas de NTCIR-6

```
<TOPIC>
    <NUM>001</NUM>
    <SLANG>CH</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>Time Warner, American Online (AOL), Merger, Impact</TITLE>
    <DESC>Find reports about the impact of AO /Time Warner merger.</DESC>
    <NARR>
        <BACK>Time Warner and American Online (AOL) announced a merger on January 10th, 2000. The market value was estimated at $US350 billion making it the biggest merger in the US.</BACK>
        <REL>Comments on AOL/Time Warner merger's effects on Internet and entertainment media businesses are relevant. Descriptions of the development of the AOL/Time Warner merger are partially relevant. Information about the total amount and the transformation of ownership structure are irrelevant.</REL>
    </NARR>
    <CONC>Time Warner, American Online, AOL, Gerald Levin, merger, M&A, Merger and Acquisition, media, entertainment business</CONC>
</TOPIC>

<TOPIC>
    <NUM>002</NUM>
    <SLANG>CH</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>President of Peru, Alberto Fujimori, scandal, bribe</TITLE>
    <DESC>Find reports about Peru President Fujimori's bribery scandal in the 2000 election and his exile abroad after he was impeached by the Congress of Peru.</DESC>
    <NARR>
        <BACK>After President Fujimori won the 2000 election, riots began everywhere in Peru. The US government declared the election result to be invalid. Peru's media aired a tape showing Fujimori's staff trying to bribe the opposition party. Fujimori offered a written resignation to the Congress but they impeached him for moral decadence.</BACK>
        <REL>Reports on how Fujimori tried to manipulate the election or bribe the opposition party are relevant. Reports on responses and opinions of foreign governments such as US and Japan are partially relevant. Reports on the situation in Peru without mention of the presidential election in 2000 are irrelevant.</REL>
    </NARR>
    <CONC>President of Peru, Alberto Fujimori, bribe, freedom of the press, Japan, United States, US, extradite, impeach</CONC>
</TOPIC>

<TOPIC>
    <NUM>003</NUM>
    <SLANG>CH</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>Kim Dae Jun, Kim Jong Il, Inter-Korea Summit</TITLE>
    <DESC>Find reports on the Inter-Korea Summit between South Korean President Kim Dae Jun and North Korean leader Kim Jong Il in Pyongyang.</DESC>
    <NARR>
        <BACK>South Korean President Kim Dae Jun and North Korean leader Kim Jong Il held a summit in Pyongyang on June 13th, 2000. On the 14th, they signed a joint declaration on unification and reached an agreement on solving humanitarian issues such as repatriation of political refugees, reunifying separated North and South Korean families and providing funds from South Korea to North Korea to help the economy.</BACK>
        <REL>Reports on the concrete agreements made at the Inter-Korea summit about solving humanitarian issues, helping the economy of North Korea or speeding up the unification of North and South Korea are relevant. Reports about the opinions of other countries and their politicians are partially relevant. Reports mentioning only the problems between North Korea and South Korea but  not the implications and effects of this summit are irrelevant.</REL>
    </NARR>
```

<CONC>Kim Dae Jun, Kim Jong Il, Sunshine policy, Summit between North and South Korea, Inter-Korea Summit, Pyongyang</CONC>
</TOPIC>

<TOPIC>
<NUM>004</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>the US Secretary of Defense, William Sebastian Cohen, Beijing </TITLE>
<DESC>Find reports about the US Secretary of Defense, William S. Cohen, visit to Beijing in June, 2000.</DESC>
<NARR>
<BACK>The US Secretary of Defense, William S. Cohen, visited Beijing in June, 2000. Cohen's mission was to carry out a strategic dialogue with Beijing about his country's development of National Missile Defense (NMD) and Theater Missile Defense (TMD) systems and express concerns about the cross-strait relationship. This visit symbolized the normalization of US-China relations after the accidental May 1999 bombing of the Chinese embassy in Belgrade, Yugoslavia. It also reflected the gradual normalization and increased transparency in the US and Chinese military.</BACK>
<REL>Reports about the visit of US Secretary of Defense, William S. Cohen, and details on the meetings with Chinese leaders are relevant. Analysis or discussions about the political, military and diplomatic implications of Cohen's visit are partially relevant. Reports only on Cohen's itinerary are irrelevant. Reports on Cohen's opinions about the cross-strait relationship outside of this visit are also irrelevant.</REL>
</NARR>
<CONC>the US Secretary of Defense, William S. Cohen, visiting Beijing, Theater Missile Defense (TMD), National Missile Defense (NMD), cross-strait relationship, Taiwan problem, US-China relationship, military transparency, the 1999 bombing of Chinese embassy in Belgrade, Yugoslavia</CONC>
</TOPIC>

<TOPIC>
<NUM>005</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE> G8 Okinawa Summit </TITLE>
<DESC>Find reports on the G8 Okinawa Summit 2000.</DESC>
<NARR>
<BACK>The G8 Okinawa Summit was held in Nago City, Okinawa, and China was not invited. The participating countries planned to announce a statement about regional peace issues, and to publish a cooperative declaration that would focus on what the WTO should pay attention to like balance and tolerance toward developing countries. The G8 would also discuss the application and supervision of official development assistance (ODA) to developing countries.</BACK>
<REL>Discussions or analyses about the impact of the 2000 G8 Summit on global political and economic situations are relevant. Descriptions of the declarations or discussions about regional peace, WTO, ODA in the 2000 G8 Summit period are also relevant. Reports on other issues related to the 2000 G8 summit are also partially relevant. Reports only on the process of the 2000 G8 Summit or the arrangement of participants' journeys without discussing any content issue are irrelevant.</REL>
</NARR>
<CONC>The Group of Eight, G8 Summit, Official development assistance, ODA, Okinawa Summit, Regional peace issues, World Trade Organization, WTO </CONC>
</TOPIC>

<TOPIC>
<NUM>006</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>Wen Ho Lee Case, classified information, national security</TITLE>
<DESC>Find reports on the Chinese-American scientist, Wen Ho Lee who was suspected of stealing classified information about nuclear weapons from the US's Los Alamos National Laboratory. </DESC>
<NARR>

<BACK>The Chinese-American scientist, Wen Ho Lee, was accused of stealing classified information about US nuclear weapons. On September 13th, 2000 he plea bargained with federal prosecutors by pleading guilty to one felony count of downloading classified files about nuclear weapons. In exchange, federal prosecutors agreed to drop the remaining 58 counts. The judge then sentenced Lee to 278 days. On the 24th, the Albuquerque federal court, New Mexico agreed to release him on $1 million bail.</BACK>
<REL>Reports on the story and investigation of the Wen Ho Lee case are relevant. Discussions about the issues following the Wen Ho Lee Case such as racial discrimination, national security and so on are partially relevant. Descriptions of the support of Wen Ho Lee are irrelevant.</REL>
</NARR>
<CONC>Wen Ho Lee, Wen Ho Lee Case, Cox Report, Los Alamos National Laboratory, Nuclear warhead, Mishandling nuclear secrets, National security, Racial discrimination, Amnesty International</CONC>
</TOPIC>

<TOPIC>
<NUM>007</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>Ichiro, Rookie of the Year, Major League </TITLE>
<DESC>Find reports on Ichiro's first year in the Major League after his move from a Japanese league.</DESC>
<NARR>
<BACK>The Japanese professional baseball star, Ichiro had a distinguished record after he transferred to Seattle Mariners MLB. He won many titles and awards such as Ranks 1st in the Batting average Title, and the Rookie of the Year award..</BACK>
<REL>Report on Ichiro's records and awards in MLB are relevant. Descriptions of others' reflections and reactions to the above are partially relevant. Reports only on the Mariners' routine and playoff games are irrelevant.</REL>
</NARR>
<CONC>Ichiro Suzuki, Ranks 1st in Batting average, Rookie of the Year, Most Valuable Player, MVP, Gold Glove, the People's Honor Award, Fantasy Comparison, Ranks 1st in Stolen bases, Seattle Mariners, Orix Blue Wave</CONC>
</TOPIC>

<TOPIC>
<NUM>008</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>Jennifer Capriati, tennis</TITLE>
<DESC>Find reports on American female tennis player, Jennifer Capriati, who won some major games after a comeback and was once ranked the world's number one by the WTA.</DESC>
<NARR>
<BACK>The female tennis player, Jennifer Capriati, was the youngest star but left the tour due to personal problems. She made a comeback in 2001 and won the Australian Open and the French Open. As a result, she took the WTA's top spot in the ranking from Martina Hingis.</BACK>
<REL>Reports on Capriati's wins and record in any of four major tournaments; the Wimbledon, French Open, Australian Open and US Open, are relevant. Others' opinions or reactions to the above are partially relevant. Reports on Capriati's tennis career including personal reviews or special reports about the comeback are also partially relevant. Those only mentioning the competition details in which Capriati participated are irrelevant.</REL>
</NARR>
<CONC>Jennifer Capriati, Australian Open, French Open, Women's Tennis Association, world number one</CONC>
</TOPIC>

<TOPIC>
<NUM>009</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>EP-3 surveillance aircraft, F-8 fighter, aircraft collision</TITLE>

<DESC>Find reports on the midair collision of a US EP-3 surveillance aircraft and a Chinese F-8 fighter near Hainan Island.</DESC>
<NARR>
<BACK>A U.S. Navy EP-3 reconnaissance aircraft on a routine reconnaissance and surveillance mission near Hainan Island was intercepted and collided with a F-8 fighter jet aircraft of the People's Republic of China. The collision caused damage that resulted in an emergency landing of the EP-3 at the nearest airfield on Hainan Island and a drop into the sea of the F-8 fighter. After the accident, China asked the US to apologize and the US asked China to return the reconnaissance aircraft and the crew. This caused a stalemate in the US-China relationship for a while.</BACK>
<REL>Mention of the explanations for the collision of two military aircrafts such as the statements from the crew are relevant. Reports on the reflections or reactions around the world about the accident and the reactions from both sides are partially relevant. Mentions of only the function or the crew of EP-3 reconnaissance aircraft or F-8 fighter are irrelevant.</REL>
</NARR>
<CONC>EP-3 surveillance aircraft, reconnaissance aircraft, F-8 fighter, collision, military aircraft, Colin Powell, Jiang Zemin, President Bush</CONC>
</TOPIC>

<TOPIC>
<NUM>010</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>History Textbook Controversies, World War II</TITLE>
<DESC>Find reports on the controversial history textbook about the Second World War approved by the Japanese Ministry of Education.</DESC>
<NARR>
<BACK>The Japanese Ministry of Education approved a controversial high school history textbook that allegedly glosses over Japan's atrocities during World War Two such as the Nanjing Massacre, the use of millions of Asia women as "comfort women" and the history of the annexations and colonization before the war. It was condemned by other Asian nations and Japan was asked to revise this textbook.</BACK>
<REL>Reports on the fact that the Japanese Ministry of Education approved the history textbook or its content are relevant. Reports on reflections or reactions to this issue around the world are partially relevant. Content on victims, "comfort women", or Nanjing Massacre or other wars and colonization are irrelevant. Reports on the reflections and reactions of the Japanese government and people are also irrelevant.</REL>
</NARR>
<CONC>Ministry of Education, Japan, Junichiro Koizumi, textbook, comfort women, sexual slavery, Nanjing Massacre, annexation, colonization, protest, right-wing group, Lee Den Hui</CONC>
</TOPIC>

<TOPIC>
<NUM>011</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>Tobacco business, accusation, compensation</TITLE>
<DESC>Find reports related to accusations against Tobacco business and compensation awarded by the courts.</DESC>
<NARR>
<BACK>The U.S. tobacco giant, Philip Morris, has been ordered to pay compensation of $US 3 billion to a 56-year-old cancer patient. This was the largest compensation awarded to a single person in a tobacco related case. Other countries have had similar cases. -</BACK>
<REL>Reports on the content and accusations against the tobacco business and the amount of the judgments against the tobacco business are relevant. Reports on the ill effects of tobacco are partially relevant. Reports on the tobacco business related to stock price performance and the like are irrelevant.</REL>
</NARR>
<CONC>Tobacco, Tobacco business, Compensation, Tobacco ill effects</CONC>
</TOPIC>

<TOPIC>
<NUM>012</NUM>

```
    <SLANG>KR</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>Tiger Woods, sports star</TITLE>
    <DESC>Find documents about sports media or related enterprises recognizing Tiger Woods as a
    sports star.</DESC>
    <NARR>
      <BACK>During his four full years on the PGA Tour, Tiger Woods (25) was voted athlete of the year
      for the 3rd time. Sportsmen's changing views of golf was the reason he won their votes.</BACK>
      <REL>Documents about sports magazines or enterprises recognizing Tiger Woods as a sports star
      based on his record, skills or contribution to marketing are relevant. Documents about Tiger Woods'
      daily life or celebrity news outside of golf are irrelevant.</REL>
    </NARR>
    <CONC>Tiger Woods, golf, golf genius, PGA</CONC>
</TOPIC>

<TOPIC>
    <NUM>013</NUM>
    <SLANG>CH</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>"Chiutou" (Autumn Struggle), Appeal, Laborer, Protest, Taiwan</TITLE>
    <DESC>Find articles containing Taiwan laborers' appeal in the "Chiutou" (Autumn Struggle) protest and
    the laborer policies proposed by Government in 1998</DESC>
    <NARR>
        <BACK>The "Chiutou" (Autumn Struggle) protest of Taiwan laborers is held every November 12th.
        I would like to know the appeals the laborers proposed to the Council of Labor Affairs in Executive
        Yuan in 1998 and what laborer policy points the Council of Labor Affairs promised at that
        time.</BACK>
        <REL>The appeals of laborers are relevant. Feedback of the Council of Labor Affairs for appeals of
        the laborer policy points is relevant as well. The process of the protest is not relevant.</REL>
    </NARR>
    <CONC>Laborer, protest, Council of Labor Affairs, Appeal, Laborer Policy</CONC>
</TOPIC>

<TOPIC>
    <NUM>014</NUM>
    <SLANG>KR</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>Expert, Opinion, International Monetary Fund (IMF), Asian countries</TITLE>
    <DESC>Find expert critical opinion on the International Monetary Fund's (IMF) policy on Asian
    countries</DESC>
    <NARR>
        <BACK>The International Monetary Fund (IMF) presented many measures to deal with the
        economic crisis related to foreign exchange in Asian countries and Russia, and there are various
        opinions on these measures.</BACK>
        <REL>Expert critical opinion and criticism on the International Monetary Fund's (IMF)
        countermeasures on Asian countries is relevant.  Self-criticism by the IMF itself is also relevant.
        Critical opinions of government officials of the countries directly involved are partially relevant.
        Articles that simply describe the IMF's policy or negotiation between the IMF and the involved
        country are irrelevant.</REL>
    </NARR>
    <CONC>International Monetary Fund, IMF, foreign exchange crisis, economic crisis, Asia,
    influence</CONC>
</TOPIC>

<TOPIC>
    <NUM>015</NUM>
    <SLANG>KR</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>Teenager, Social Problem</TITLE>
    <DESC>Find articles dealing with a teenage social problem</DESC>
    <NARR>
```

<BACK>As materialism appears in many aspects of society, many incidents related to young teenagers are becoming a major social problem.</BACK>
<REL>Articles dealing with specific incidents or social problems related to teenagers (age 11 to 19) that show a summary or background story of an incident (problem) and information on the teenagers are relevant. Articles only addressing general criticisms on youth problems are irrelevant. Incidents or social problems where teenagers are mentioned but are not the main issue are partially relevant.</REL>
</NARR>
<CONC>teenager social problem, youth problem, youth, teenager, human traffic, runaway, robbery, suicide, sexual abuse</CONC>
</TOPIC>

<TOPIC>
<NUM>016</NUM>
<SLANG>KR</SLANG>
<TLANG>EN</TLANG>
<TITLE>Divorce, Family Discord, Criticisms</TITLE>
<DESC>Find articles describing criticisms on family discord such as divorce, separation, etc. </DESC>
<NARR>
<BACK>As a consequence of the society's tendency to form nuclear families and maximize the individual's freedom, divorces and separations are increasing heavily.</BACK>
<REL>Articles related to divorces and separation describing divorce (separation) statistics, reasons of divorce (separation), social problems such as child problems of divorced (separated) families, etc are relevant. Articles only recording someone's divorce, separation or remarriage are not relevant.</REL>
</NARR>
<CONC>divorce, separation, divorce statistics, divorce reason, child problem, family discord, beating, adultery</CONC>
</TOPIC>

<TOPIC>
<NUM>017</NUM>
<SLANG>KR</SLANG>
<TLANG>EN</TLANG>
<TITLE>China, Reaction, Taiwan, Diplomatic Relations</TITLE>
<DESC>Find articles that show China's reaction to Taiwan's establishment of diplomatic relations with foreign countries</DESC>
<NARR>
<BACK>Taiwan is making efforts to establish relations with foreign countries such as North Korea, Macedonia, etc. as well as expanding existing relations. China is showing negative reactions.</BACK>
<REL>Articles recording China's reaction to Taiwan's attempt to establish and strengthen diplomatic relations with foreign countries are relevant. Articles that interpret China's actions in international conferences or international organizations that are related to Taiwan's attempt to strengthen diplomatic relations are also relevant.</REL>
</NARR>
<CONC>Taiwan, Macedonia, amity, China, reaction, retaliation</CONC>
</TOPIC>

<TOPIC>
<NUM>018</NUM>
<SLANG>KR</SLANG>
<TLANG>EN</TLANG>
<TITLE>China, Stationing, Weapons, Taiwan</TITLE>
<DESC>Find articles showing China's military actions against Taiwan such as deployment of missiles, etc.</DESC>
<NARR>
<BACK>In order to keep alert against Taiwan, China is taking military actions against Taiwan such as deployment of missiles, etc.</BACK>
<REL>Articles showing China's military actions against Taiwan such as deployment of missiles near Taiwan and developing new arms are relevant. Articles containing Taiwanese government speeches, reaction, etc. that suggest the fact that China is taking military actions, even if the

articles do not show it directly, are relevant. China's usual military drills and relocations that are not military actions against Taiwan are irrelevant.</REL>
</NARR>
<CONC>China, Taiwan, arms disposition, military disposition, missile disposition</CONC>
</TOPIC>

<TOPIC>
<NUM>019</NUM>
<SLANG>KR</SLANG>
<TLANG>EN</TLANG>
<TITLE>Animal Cloning Technique</TITLE>
<DESC>Find articles introducing different countries' animal cloning techniques. </DESC>
<NARR>
<BACK>The birth of Dolly, the cloned sheep, at the Roslin research institute of England on July 5, 1996 is evaluated as the opening of a new world in genetics. Dolly was created through a method totally different from the existing one, which created identical twins by splitting the fertilized egg. The new method involved the removing of a somatic cell of a six year old ewe and transplanting a nucleus. Consequently, animal cloning using the same method is happening in many countries</BACK>
<REL>Articles about the present state of animal cloning techniques of different countries that include cloning techniques, cloned animals and research institutes or researchers are relevant. Articles containing contents of embryo experiments related to human cloning or criticism of human cloning are partially relevant. Articles introducing books related to animal cloning are irrelevant. Contents on illegally copying software are irrelevant.</REL>
</NARR>
<CONC>animal cloning, human cloning, cloning technique, DNA, gene, cloned sheep, Dolly, genetics, bioengineering</CONC>
</TOPIC>

<TOPIC>
<NUM>020</NUM>
<SLANG>JA</SLANG>
<TLANG>EN</TLANG>
<TITLE>Sexual Harassment, Lawsuits</TITLE>
<DESC>Find articles referring to incidents involving sexual harassment lawsuits.</DESC>
<NARR>
<BACK>Sexual harassment lawsuits occur frequently ever since suspicions were raised about President Clinton and a former trainee. In Japan as well, an incident in which Knock Yokoyama, former governor of Osaka sexually harassed a campaign worker became an issue and resulted in Mr. Yokoyama paying a fine of 10 million yen. </BACK>
<REL>Articles pertaining to incidents resulting in trials for sexual harassment are relevant. Articles commenting on sexual harassment trials are partially relevant. Articles solely concerned with sexual harassment are irrelevant.</REL>
</NARR>
<CONC>Sexual Harassment, Academic Harassment, Lawsuit, Trial, President Clinton, Monica Lewinski, Knock Yokoyama </CONC>
</TOPIC>

<TOPIC>
<NUM>021</NUM>
<SLANG>JA</SLANG>
<TLANG>EN</TLANG>
<TITLE>Olympic, Bribe, Suspicion</TITLE>
<DESC>Find articles pertaining to suspicions about bribes by IOC members involved in the selection of Olympic venues</DESC>
<NARR>
<BACK>Suspicions surfaced about members of the International Olympic Committee (IOC) accepting bribes in relation to the selection of venues for the Nagano and Salt Lake City Winter Olympics, and President Samaranch established a committee to investigate the situation. </BACK>
<REL>Articles pertaining to suspicions about monetary payments made in relation to the selection of Olympic venues are relevant. Articles that don't contain detailed information on the content of

```
        specific suspicions are partially relevant. Articles on suspicions of corruption involving non-IOC
        members involved in selecting Olympic delegates are irrelevant. </REL>
    </NARR>
    <CONC>Olympic, Five-Ring, International Olympic Committee (IOC), IOC Member, President
    Samaranch, Summon, Bribe, Purchase, Suspicion</CONC>
</TOPIC>

<TOPIC>
    <NUM>022</NUM>
    <SLANG>JA</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>North Korea, Daepodong, Asia, Response</TITLE>
    <DESC>Find articles on Asian nations' responses to North Korea's launching of a Daepodong
    missile.</DESC>
    <NARR>
        <BACK>In August 1998, North Korea launched a Daepodong 1 missile over Japan. The warhead
        landed off Sanriku, giving Japan a big shock. Although North Korea had given indications of the
        launch beforehand, the Japanese Government hadn't taken measures to deter it, and the
        announcement came after the event.</BACK>
        <REL>Articles on Asian nations' responses and reactions the North Korea's launch of the
        Daepodong missile are relevant. Articles on North Korean announcements, or comments made by
        unrelated countries on the impact of this incident are partially relevant. Articles that simply mention
        the relationship between North Korea and the United States or the Daepodong incident are
        irrelevant. </REL>
        <TERM>The Daepodong is an intermediate-range ballistic missile under development in the
        Democratic People's Republic of Korea (North Korea). It is said to have a range of 1700-2200
        kilometers. Its name derives from name of the place (Daepodong) from which the American
        reconnaissance satellite first photographed it.</TERM>
    </NARR>
    <CONC>The Democratic People's Republic of Korea (North Korea), Taepodong, Ballistic Missile, Off
    Sanriku, Punitive Measures, Prime Minister Keizo Obuchi, Kim Jong-il, Kim, Jong-il,
    Daepodong</CONC>
</TOPIC>

<TOPIC>
    <NUM>023</NUM>
    <SLANG>CH</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>Joining WTO </TITLE>
    <DESC>Find possible problems that industries will meet after Taiwan's joining WTO.</DESC>
    <NARR>It has taken Taiwan 10 years to get in to WTO. The Council For Economic Planning and
    Development, Chung-Hua Institution for Economic Research and Taiwan Institution for Economic
    Research evaluated the beneficial result of joining WTO. Related contents are supposed to include the
    evaluation contents, the advantages and disadvantages and the effects on agriculture, industry and
    business. If the documents only describe the opinions, comments, and attitudes of the America and
    other countries, or the political and diplomatic issues, they will be regarded irrelevant. </NARR>
    <CONC>Taiwan, WTO, agriculture, industry, benefits, economy, World Trade Organization</CONC>
</TOPIC>

<TOPIC>
    <NUM>024</NUM>
    <SLANG>CH</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>China Airlines Crash</TITLE>
    <DESC>Retrieve reports about China Airlines' crash while trying to land at Taoyan international
    airport.</DESC>
    <NARR>CI676 China Airlines crash on February 16th in 1998 is the most serious plane crash in Taiwan
    history. It even draws international attention. Related contents are supposed to include the death toll,
    accidents causes, who should take responsibility, the compensation from China Airlines. Take the
    accident in principal and ignore the introduction of victims.</NARR>
    <CONC>China Airlines, plane crash, black box, compensate, body count, death toll, jet crash</CONC>
</TOPIC>
```

<TOPIC>
<NUM>025</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>Province-refining</TITLE>
<DESC>Find the content of Province-refining enactment and Mr. James Soong's attitudes after the Province-refining</DESC>
<NARR>Taiwan Province and Taiwan province assembly have become history since December 20th, 1998. The temporary province function and organization regulation start applying on December 21st. Related contents include province-refining regulation and on what it is based, which regulations stop applying, which ones start applying, the purpose of province-refining, James Soong's reflections, attitudes and other related comments. Effects on individual because of province-refining will be regarded as irrelevant.</NARR>
<CONC>Province-refining, James Soong, Taiwan Province, chairman of Taiwan Province, Province assembly, budget</CONC>
</TOPIC>

<TOPIC>
<NUM>026</NUM>
<SLANG>JA</SLANG>
<TLANG>EN</TLANG>
<TITLE>Economic influence of the European monetary union</TITLE>
<DESC>Articles relating to economic influence of European monetary union.</DESC>
<NARR> The currency of eleven nations in Europe was unified by the resolution of the EU special summit meeting, and new currency, the "Euro", appeared conceptually in January, 1999. Although it will be 2002 when the bill and coin called the Euro actually appears on the market, a Euro credit card or a Euro cheque may be used for shopping, and the price in Euros is also displayed in retail stores. If an article is about the economic influence of this monetary union, it is relevant. If an article is about the realization of the monetary union in other areas, such as Asian countries, it is partially relevant.
</NARR>
<CONC>European monetary union, Euro, Euro currency, European Union, EU, European Economic Community, the European Central Bank, monetary and financial policy, economic gap</CONC>
</TOPIC>

<TOPIC>
<NUM>027</NUM>
<SLANG>JA</SLANG>
<TLANG>EN</TLANG>
<TITLE>President Kim Dae-Jung's policy toward Asia</TITLE>
<DESC>Articles relating to President Kim Dae-Jung's policy toward Asia</DESC>
<NARR>On February 25, 1998, in South Korea, Mr. Kim Dae-Jung, elected by presidential election at the end of 1997, was inaugurated as President, and the Kim Dae-Jung Administration was formally inaugurated. While President Kim regarded the new government as "the national government", he appealed against people concentrating on overcoming a serious economic crisis. He started structural adjustments, such as plutocracy reform. If an article describes President Kim Dae-Jung's policy toward Asia, it is relevant. If an article does not describe his policy towards Asia, but his posture towards a foreign country, or a plan, it is partially relevant.</NARR>
<CONC>Kim Dae-Jung, the President, the national government, policy toward Asia, economic crisis, economic reform, China, Taiwan, Japan</CONC>
</TOPIC>

<TOPIC>
<NUM>028</NUM>
<SLANG>EN</SLANG>
<TLANG>EN</TLANG>
<TITLE>Clinton scandals</TITLE>
<DESC>What was the reaction in Asia to the Clinton scandals?</DESC>
<NARR>Documents should present either specific opinions expressed in Asia about the Clinton scandal or discuss possible effects of the scandal and subsequent Congressional hearing on Asia. Not relevant are documents that just report information about these events.</NARR>
<CONC>the President of the USA, Bill Clinton, sex scandal, Asia, opinion</CONC>

```
</TOPIC>

<TOPIC>
    <NUM>029</NUM>
    <SLANG>EN</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>War crimes lawsuits</TITLE>
    <DESC>Provide information on lawsuits in Japan arising from war crimes committed by Japan during World War II.</DESC>
    <NARR>Documents should describe the progress or outcome of a specific civil lawsuit arising from war crimes committed by Japan during World War II.  This can include new lawsuits, verdicts, or public comments on a trial. Information on war crimes tribunals is not relevant. </NARR>
    <CONC>World War II, civil lawsuit, war crime by Japan, verdict</CONC>
</TOPIC>

<TOPIC>
    <NUM>030</NUM>
    <SLANG>EN</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>Nuclear power protests</TITLE>
    <DESC>Give information regarding protests against nuclear power.</DESC>
    <NARR>Documents should describe specific negative public reaction to nuclear power issues, such as a petition, protest or demonstration.  Nuclear power issues include generating power, building nuclear plants, and import/export of waste products. Articles regarding nuclear weapons development or testing are not relevant.</NARR>
    <CONC>nuclear power, protest against nuclear power, nuclear plant, waste product</CONC>
</TOPIC>

<TOPIC>
    <NUM>031</NUM>
    <SLANG>KR</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>College Admission Policy</TITLE>
    <DESC> Relevant documents should describe college admission policies (systems) and opinions of parents, students and teachers.</DESC>
    <NARR> The college admission policy is of great interest to the students who want to enter a college and to their parents. Relevant documents include an overall description of the policy or system that the government or a university has set up for college entrance, or discussions about specific issues on the policy or system, which include opinions of the stakeholders such as parents, students, or advisors from high schools or private institutions. Documents with opinions from the society as a whole alone, not from a specific stakeholder, are not relevant. </NARR>
    <CONC>college admission policies, college admissions system, parents, high school students, teachers</CONC>
</TOPIC>

<TOPIC>
    <NUM>032</NUM>
    <SLANG>KR</SLANG>
    <TLANG>EN</TLANG>
    <TITLE>Counseling for Youths</TITLE>
    <DESC>What are documents including the names of institutions for counseling youths' anguish and the guides for getting the counseling services?</DESC>
    <NARR>Relevant documents describe the names of counseling institutions for youths, details of youths' anguish, and methods by which a youth can receive counseling services, including telephone numbers, maps, and office hours. Expressions about youths' anguish can be general in relevant documents. If a document provides a detailed guide about how to access the institution without its name or has the name for the institution without a specific access point, it is partially relevant. A document without any method for accessing the institution is irrelevant.</NARR>
    <CONC>youth, new generation, counseling organization, counselor, life line, school violence</CONC>
</TOPIC>
```

Anexo B – Lista de Palabras de Parada

Esta lista de palabras de parada es una versión simplificada de la utilizada en el proyecto. En la lista original cada palabra aparece tres veces: en minúsculas, con la primera letra en mayúsculas y toda en mayúsculas. En esta lista solo se ha incluido la palabra en minúsculas.

-	although	backwards	correspond	enough
0	altogether	bareback	corresponding	entirely
1	always	barring	corresponds	especially
2	am	be	could	et
3	amain	became	couldn	etc
4	amid	because	course	even
5	amidst	become	current	ever
6	amiss	becomes	currently	every
7	amok	becoming	d	everybody
8	among	been	definitely	everyone
9	amongst	before	depending	everything
a	ampleforth	beforehand	describe	everywhere
á	an	behind	described	evidence
à	and	being	describes	exactly
â	anew	believe	describing	except
ä	another	believes	despite	excepting
able	any	belong	detail	exist
about	anybody	belonging	details	existed
above	anyhow	belongs	did	existing
according	anyone	below	didn	exists
accordingly	anything	beneath	didnt	expect
across	anytime	beside	different	expected
actually	anyway	besides	discuss	expecting
after	anyways	best	discussed	expects
afterwards	anywhere	better	discussing	experimentally
again	apart	between	discussion	explain
against	apiece	beyond	discussions	explained
ago	appear	both	ditto	explaining
ah	appeared	but	do	explains
aha	appearing	by	document	explicitly
ahead	appears	c	documents	express
ahem	appreciate	ç	does	expressed
ahoy	appropriate	called	doesn	expresses
aimed	apropos	came	doing	expressing
ain	are	can	doings	f
aint	aren	cannot	done	farther
alack	around	cant	dont	few
alas	arst	certain	doubtless	find
albeit	as	certainly	down	fine
alight	aside	clearly	downwards	follow
alike	ask	come	due	followed
all	asked	comes	during	following
allow	asking	coming	e	follows
allowed	asks	concerning	é	for
allowing	associated	consequently	è	forasmuch
allows	at	consider	ë	forever
almost	available	considerably	each	form
alone	away	considered	early	formed
along	awfully	considering	earnest	formerly
alongside	awhile	considers	eg	forming
aloud	awry	contain	eh	forms
already	b	contained	either	forth
alright	back	containing	else	forthwith
also	backward	contains	elsewhere	forward

forwards
found
from
fundamental
further
furthermore
g
general
get
gets
getting
give
gived
given
gives
giving
go
goes
going
gone
got
gotten
h
had
hadn
half
happen
happened
happens
hardly
has
hasn
have
haven
having
he
heavily
held
hello
hence
henceforth
her
here
hereabouts
hereafter
hereby
herein
heretofore
hereupon
herewith
hers
herself
hi
highly
him
himself
his
hither
hitherto
hold
holding
holds
home
hopefully
how
howbeit
however
i
ï
ie
if
ignore
ignored
ignores
ignoring
illicitly
immediate
importance
important
in
inasmuch
inc
include
included
includes
including
indeed
indicate
indicated
indicates
indicating
information
informations
inner
inside
insofar
instead
interest
into
involve
involved
involves
involving
inward
irrespective
is
isn
issue
it
its
itself
j
just
k
keep
keeping
keeps
kept
kind
know
knowing
known
knows
l
largely
last
late
lately
later
latter
latterly
lead
leaded
leading
leads
least
leave
leaving
left
legally
less
lest
let
like
liked
likely
likewise
listing
little
longer
look
looked
looking
looks
ltd
m
made
main
mainly
make
makes
making
many
may
maybe
mayed
maying
mays
me
meant
meantime
meanwhile
mention
mentioned
mentioning
mentions
merely
meself
might
mine
minus
more
moreover
most
mostly
much
must
my
myself
n
name
named
namely
names
naming
nd
near
nearby
nearer
nearly
necessary
need
needed
needing
needs
neither
never
nevertheless
new
newly
news
next
no
nobody
non
none
nonetheless
noone
nor
normally
not
nothing
notwithstanding
now
nowadays
nowhere
ñ
o
ó
ô
ö
õ
obviously
occurrence
occurrences
of
off
often
oh
ok
okay
on
once
one
ones
oneself
only
onto
onward
onwards
or
other
others
otherwise
ought
oughter
our
ours
ourselves
out
outright
outside
outward
outwards
outwith
over
overall
own
p
part
partially
particular
particularly
parts
past
pending
people
per
perhaps
permits
permitted
permitting
pertain
pertaining
pertains
pertinent
physically
piece
pieces
placed
plain
please
plus
posed
possible
presumably
pretty
previously
probably
prompt
prompted
provide
provided
provides
providing
q
quite
r
rather
rd

re
ready
really
reasonably
recently
recognise
recognised
recognises
recognising
recognize
recognized
recognizes
recognizing
refer
referred
referring
refers
regard
regarding
regardless
regards
related
relation
relations
relative
relatively
relevant
relevants
remain
remained
remaining
remains
report
reported
reporting
reports
respectively
right
s
said
same
saw
say
saying
says
secondly
see
seeing
seem
seemed
seeming
seemingly
seems
seen
sees
seldom
self
selves
sent
serious
seriously
several
shall
sharp
she
sheer
should
showed
simply
since
so
some
somebody
someday
somehow
someone
someplace
something
sometime
sometimes
somewhat
somewhere
sooner
speak
speaking
speaks
specific
specifically
specified
specifies
specify
specifying
spoke
stating
still
stories
story
straight
strongly
such
suggest
suggested
suggesting
suggests
suppose
supposed
supposes
supposing
sure
t
take
taken
takes
taking
tell
telling
tells
tend
tends
th
than
that
thats
the
thee
their
theirs
them
themselves
then
thence
there
thereabouts
thereafter
thereby
therefore
therefrom
therein
thereof
theres
thereto
thereunder
thereupon
these
they
think
this
tho
thorough
thoroughly
those
thou
though
through
throughout
thru
thus
till
tis
to
today
together
told
tomorrow
tonight
too
took
topic
topics
totally
toward
towards
tried
tries
truly
try
trying
twice
twofold
u
ü
un
under
unfortunately
unless
unlike
unlikely
until
unto
up
upon
upper
uppermost
upright
upside
upward
upwards
us
use
used
useful
uses
using
usually
v
various
ve
versus
very
via
vice
vs
w
want
wants
was
wasn
wass
we
well
went
were
weren
west
what
whatever
whatsoever
when
whence
whencesoever
whenever
where
whereabouts
whereafter
whereas
whereby
wherefore
wherein
whereupon
wherever
whether
which
whichever
while
whilst
whither
who
whoever
whole
wholesale
whom
whose
why
whyever
wide
will
willed
willing
wishing
with
within
without
won
worse
worst
would
wrong
x
y
yes
yesterday
yet
yonder
you
your
yours
yourself
yourselves

Referencias

Las referencias aquí incluidas están ordenadas por orden de citación en el texto de la memoria. Asimismo, las URLs citadas en el texto, principalmente como notas de pie de página, fueron comprobadas a día 10/05/2010.

[1] Fabrizio Sebastiani: "*Machine Learning in Automated Text Categorization*". Consiglio Nazionale delle Ricerche, Italy, 2002.

[2] G. Salton, A.Wong & C.S. Yang: "*A Vector Space Model for Automatic Indexing*", Cornell University, 1975.

[3] Kagan Tumer & Joydeep Ghosh: "*Order Statistics Combiners For Neural Classifiers*". Proceedings og the World Congreso of Neural Networks, 1995.

[4] J. J. Rocchio: "*Relevance feedback in information retrieval*". The SMART Retrieval System. Experiments in Automatic Document Processing, páginas 313–323. Prentice Hall, Englewoods Cliffs, N. J., 1971.

[5] Hrvoje Bacan, Igor Pandzic & Darko Gulika: "*Automated News Item Categorization*". Proceedings of the 19th Annual Conference of The Japanese Society for Artifical Intelligence, 2005.

[6] C.G. Figuerola, J.L. Alonso Berrocal, A. F. Zazo Rodriguez & E. Rodríguez: "*Algunas Técnicas de Clasificación Automática de Documentos*". Grupo REINA, Universidad de Salamanca, 2004.

[7] F.J. Cortijo Bon: "*Técnicas supervisadas II: Aproximación no paramétrica*". Universidad de la República, Uruguay, 2000.

[8] J.R. Quinlan: "*Induction of Decision Trees (ID3 algorithm)*". *Machine Learning J*., vol. 1, núm. 1, pp. 81-106 (Mar. 1986).

[9] J.R. Quinlan: "*Programs for Machine Learning"*, Morgan Kauffman, California, 1993.

[10] E. Cabello Pardos: "*Técnicas de Reconocimiento Facial mediante Redes Neuronales*". Tesis Doctoral, Universidad Rey Juan Carlos, 2004.

[11] Vladimir N. Vapnik: "*The Nature of Statistical Learning Theory*," 1995.

[12] Steve R. Gunn: "*Support Vector Machines for Classification and Regression*". University of Southhampton, UK, 1998.

[13] Bo Pang & Lillian Lee: "*Thumbs up? Sentiment Classification using Machine Learning Techniques*". Conference on Empirical Methods in Natural Language Processing, Junio de 2002.

[14] Peter D. Turney: "*Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*". 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Julio de 2002.

[15] Tetsuya Nasukawa & Jeonghee Yi: "*Sentiment analysis:capturing favourability using language processing*". International Conference on Knowledge Capture, 2003.

[16] J. Russell: "*A Circumplex model of affect*". Journal of personality and social psychology, 1980.

[17] A. Ortony, G. Clore, & A. Collins: "*The cognitive structure of emotions*". Cambridge University press, 1988.

[18] G. Leshed & J. Kaye: "*Understanding how bloggers feel: recognizing affect in blog posts.*" En Gary Olson y Robin Jeffries, editors, CHI Extended Abstracts, páginas 1019-1024, 2006.

[19] P. Turney & M. Littman: "*Measuring praise and criticism: Inference of semantic orientation from association.*". ACM Trans. Inf. Syst., 2003.

[20] V. Francisco, R. Hervás, & P. Gervás: "*Expresión de emociones en la síntesis de voz en contextos narrativos*". Simposio de Computación Ubicua e Inteligencia Ambiental, Septiembre, 2005.

[21] H. Liu, H. Lieberman, & T. Selker: "*A model of textual affect sensing using real-world knowledge*". En IUI '03: Proceedings of the 8th international conference on intelligent user interfaces, New York, NY, USA. 2003.

[22] V. Francisco & P. Gervás: "*Análisis de dependencias para la marcación de cuentos con emociones*". Procesamiento de Lenguaje Natural, Septiembre, 2006.

[23] Hugo Liu, Herny Lieberman & Ted Selker: "*A model of textual affect sensing using real-world knowledge*". 8th international conference on intelligent user interfaces, Miami, Florida (USA), 2003.

[24] C. Ovesdotter, D. Roth, & R. Sproat: "*Emotions from text: machine learning for text based emotion prediction*". Proceedings of HLT/EMNLP, Vancouver, Canadá, 2005.

[25] Soo-Min Kim & Eduad Hovy: "*Identifying and Analyzing Judgement Opinions*". Human Language Technology Conference, New York, 2006.

[26] NTCIR Project, Research Center for Information Resources, National Institute of Informatics, Japan (http://research.nii.ac.jp/ntcir/index-en.html, accedido el 10/05/2010).

[27] Yunping Huang, Yulin Wang & Le Sun:"*ISCAS at MOAT*". Institute of Software, Chinese Academy of Sciences, Beijing (China). NTCIR-7 Proceedings, 2008.

[28] Youngho Kim, Seongchan Kim & Sung-Hyon Myaeng: "*Extracting topic-related opinions and their targets in NTCIR-7*". Information and Communications University, Daejeon (Korea). NTCIR-7 Proceedings, 2008.

[29] Taras Zagiballov & John Carroll: "*Almost-Unsupervised Cross-Language Opinión Análisis at NTCIR-7*". Department of Informatics, University of Sussex, UK. NTCIR-7 Proceedings, 2008.

[30] Meng Xinfan & Wang Houfeng: "*Detecting Opinionated Sentences by Extracting Context Information*". Institute of Computational Linguistic, School of Information Science and Technology, Pekin University, China. NTCIR-7 Proceedings, 2008.

[31] Lun-Wei Ku, I-Chien Liu, Chia-Ying Lee, Kuan-hua Chen & Hsin-Hsi Chen: "*Sentence-Level Opinión Análisis by CopeOpi in NTCIR-7*". Dept. of Computer Science and Information Engineering, National Taiwan University, Taipei (Taiwan). NTCIR-7 Proceedings, 2008.

[32] Meng Xinfan & Wang Houfeng: "*Detecting Opinionated Sentences by Extracting Context Information*". Institute of Computational Linguistic, School of Information Science and Technology, Pekin University, China. NTCIR-7 Proceedings, 2008.

[33] Lizhen Qu, Cigdem Toprak, Nicklas Jakob & Iryna Gurevych: "*Sentence Level Subjectivity and Sentiment Analysis Experiments in NTCIR-7 MOAT Challenge*" Ubiquitous Knowledge processing Lba, Computer Science Dept., Technische Universität Darmstadt, Germany. NTCIR-7 Proceedings, 2008.

[34] Kang Liu & Jun Zhao: "*NLPR at MOAT in NTCIR-7*". National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. NTCIR-7 Proceedings, 2008.

[35] Bin Lu, Benjamin K. Tsou & Oi Yee Kwong: "*Supervised Approaches and Ensemble Techniques for Chinese Opinión Análisis at NTCIR-7*". Language Information Sciences Research Centre, City University of Hong Kong. NTCIR-7 Proceedings, 2008.

[36] Jungi Kim, Hun-Young Jung, Sang-Hyeob Na, Yeha Lee & Jong-Hyeok Lee: "*English Opinion Analysis for NTCIR at POSTECH*". Knowledge and Language Engineering Laboratory, Pham University of Science and Technology, Pohang (Republic of Korea). NTCIR-7 Proceedings, 2008.

[37] Daisuke Kobayashi, Hidetsugu nanba & Toshiyuki Takezawa: "*Extraction of opinion sentences using machine learning: Hiroshima city university at NTCIR-7 MOAT*". Hiroshima City University, Japan. NTCIR-7 Proceedings, 2008.

[38] Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró & Muntsa Padró: "*FreeLing 1.3: Syntactic and semantic services in an open-source NLP library"*. Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA. Genova, Italy. Mayo, 2006.

[39] D. A. Grossman & O. Frieder, "*Information Retrieval: Algorithms and Heuristics*", Springer, 2004 (second edition).

[40] Olga Vectomova & Ying Wang: "*A study of the effect of term proximity on query expansion*". *Journal of Information Science* **32** (4): 324–333, 2006.

[41] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie & asociados: "*The General Inquirer: A Computer Approach to Content Analysis*". The MIT Press, 1966.

[42] Edward Kelly & Philip Stone: "*Computer Recognition of English Word Senses*". North-Holland Linguistic Series, 1975.

[43] Yohei Seki, David Kirk Evans, Lun-Wei Ku & asociados: "*Overview of Opinion Analysis Pilot Task at NTCIR-6*". Toyohashi University of Technology, Japan & National Taiwan University, Taiwan, NTCIR-6 Proceedings, May, 2007.

[44] Julio Villena Román, Sara Lana Serrano and José C. González Cristóbal: "*MIRACLE at NTCIR-7 MOAT: First Experiments on Multilingual Opinion Analysis*". NTCIR-7 Proceedings, 2008.