Departamento de Lenguajes y Sistemas Informáticos
Escuela Técnica Superior de Ingeniería Ingeniería
Universidad de Sevilla

Análisis de Contenidos Generados por Usuarios mediante la Integración de Información Estructurada y No Estructurada

Juan Manuel Cotelo Moya

Dirigida por Prof. Dr. José A. Troyano
y por Prof. Dr. Fermín L. Cruz
Universidad de Sevilla

Sevilla, Junio 2015

Índice general

Índice de figuras

Índice de cuadros

Agradecimientos

En una primera instancia, deseo expresar mi más profundo agradecimiento a José, no sólo desde la perspectiva del gran trabajo como director de esta tesis, sino también por el apoyo moral y comprensión que me ha proporcionado durante todo el trayecto, y en especial, en los momentos más duros cuando todo parecía negro.

A Fermín le agradezco todo el esfuerzo de codirección de esta tesis, siendo su paciencia y gran capacidad de análisis del detalle una habilidad inestimable respecto a la revisión tanto de los artículos como del borrador de esta memoria de tesis.

Quiero agradecer a los compañeros Javier, Fernando y Carlos por su participación en los trabajos realizados durante el transcurso de la investigación doctoral. También agradezco a Marc y a Tomás toda la compañía ofrecida, pues ellos han conseguido que trabajar en aquél despacho se hiciera mucho más llevadero.

Desde un punto de vista personal, agradezco enormemente la comprensión de mis amigos y compañeros, recalcando el apoyo y compañía ofrecido día a día durante todos estos años por parte de Mar, Sergio, Manu, Dani, Sonia, Ashur y Karim, los cuales me han proporcionado energía y motivación hasta en los peores momentos.

También doy las gracias a mis compañeros del ámbito musical, sobre todo a Enrique y Alejandro, pues las incursiones realizadas en este ámbito han sido un soplo de aire fresco imprescindible para seguir adelante.

Finalmente, en el plano familiar, agradezco todo el apoyo proporcionado por mis padres, Juan Manuel y Charo, y mi hermana, María, pues ellos me han instado a seguir adelante en todo momento.

Resumen

Los servicios de redes sociales han pasado a ser una parte fundamental del entramado social de los últimos años. Estas herramientas permiten a las personas crear, intercambiar o compartir información, ideas, imágenes y cualquier tipo de medio en comunidades virtuales y redes. El impacto de estos servicios de redes sociales sobre la sociedad ha sido tal, que han introducido cambios sustanciales sobre la comunicación a todos los niveles: individual, comunitaria, organizacional y empresarial. Todo este contenido generado por el usuario tiene un carácter viral del que carecen el resto de medios de comunicación, sirviendo como fuente de conocimiento para nuevas oportunidades de negocio. Es más, la simbiosis que existe entre los dispositivos móviles y estas redes sociales ha provocado que los contenidos generados por los usuarios incorporen nuevos factores como la localización del usuario y el momento exacto de creación y edición del mensaje o contenido en cuestión. Esto abre nuevos mercados potenciales que relacionan a las personas, las redes sociales, el mercado móvil y los eventos en tiempo real.

Al abordar este nuevo tipo de contenido, hay que comprender que las redes sociales nos otorgan la oportunidad de combinar dos aspectos fundamentales que los mensajes contienen: información estructurada con la no estructurada en forma, fundamentalmente, de textos cortos. La información estructurada nos proporciona conocimiento adicional que permite analizar el mensaje y al usuario dentro de un contexto específico de carácter social, temporal y/o espacial. Combinar significativamente ambos tipos de información puede resultar fundamental para un tratamiento efectivo de los mensajes.

En esta memoria de tesis, se explora la hipótesis consistente en que, al integrar el conocimiento proveniente de dos tipos de información de distinta naturaleza (estructurada y no estructurada) existentes en los mensajes de las redes sociales, se pueden resolver, de forma más efectiva y significativa, ciertas tareas relacionadas con el procesamiento de este tipo de contenidos. Para validar dicha hipótesis, se proponen una serie de tareas a resolver, siempre bajo el paradigma de la integración de ambos tipos de información: la recuperación temática de mensajes en redes de microblogging, la clasificación de opinión sobre los mensajes de estas redes y la caracterización de grupos de usuarios dentro de un contexto específico.

Cada tarea es tratada de forma individual, proporcionando una formalización para la misma, caracterizando los fenómenos más relevantes, proponiendo uno o varios métodos para abordarla, realizando una evaluación sobre ellos y explorando los resultados de forma consecuente. Las principales aportaciones se resumen en las siguientes propuestas: un método dinámico y adaptativo para generar consultas que son consumibles por un sistema de microblogging como Twitter, un sistema de normalización léxica altamente modular, un esquema de

integración para combinar modelos de características provenientes de información estructurada y no estructurada, y una aproximación para la caracterización de grupos de usuarios de las redes dentro de un contexto específico.

Capítulo 1

Introducción

Este capítulo de carácter introductorio tiene como objetivo presentar las motivaciones y el contexto del trabajo expuesto en esta memoria, plantear algunas hipótesis iniciales y esbozar las lineas generales sobres las aportaciones presentadas en esta tesis. La última sección de este capítulo consiste en la descripción de la estructura de esta memoria que además, sirve como guía de lectura para el documento; no sólo se enumeran y enuncian los capítulos de la memoria, sino que se incluye un breve resumen de cada uno de ellos.

1.1. Motivación y contexto

Los servicios de redes sociales han pasado a ser una parte fundamental del entramado social de los últimos años. Estas herramientas permiten a las personas crear, intercambiar o compartir información, ideas, imágenes y cualquier tipo de medio en comunidades virtuales y redes. El impacto de estos servicios de redes sociales sobre la sociedad ha sido tal, que han introducido cambios sustanciales sobre la comunicación a todos los niveles: individual, comunitaria, organizacional y empresarial.

Todo este contenido generado por el usuario tiene un carácter viral del que carecen el resto de medios de comunicación, sirviendo como fuente de conocimiento para nuevas oportunidades de negocio. Es más, la simbiosis que existe entre los dispositivos móviles y estas redes sociales ha provocado que los contenidos generados por los usuarios incorporen nuevos factores como la localización del usuario y el momento exacto de creación y edición del mensaje o contenido en cuestión. Esto abre nuevos mercados potenciales que relacionan a las personas, las redes sociales, el mercado móvil y los eventos en tiempo real.

Al abordar este nuevo tipo de contenido, hay que comprender que las redes sociales nos otorgan la oportunidad de combinar dos aspectos fundamentales que los mensajes contienen: información estructurada con la no estructurada en forma, fundamentalmente, de textos cortos. La información estructurada nos proporciona conocimiento adicional que permite analizar el mensaje y al usuario dentro de un contexto específico de carácter social, temporal y/o espacial.

La red social que ha sido foco de gran parte del trabajo plasmado en esta tesis ha sido Twitter, pues esta mezcla de red social y *microblogging* es un buen ejemplo de red de crecimiento exponencial con un grado de opinión alto, cuyos

mensajes son breves y además exhibe una topología de red interesante. Desde su aparición en 2006, Twitter se ha convertido, además de en un fenómeno social, en un proveedor de material de experimentación para la comunidad del Procesamiento del Lenguaje Natural.

Hay infinidad de trabajos que aprovechan los, escasos y de baja calidad, 140 caracteres para múltiples tareas de tratamiento de textos. Entre estas tareas se encuentran la clasificación de textos (Vitale et al., 2012; Schulz et al., 2014), en especial para determinar la polaridad de las opiniones siendo ésta una de las tareas sobre Twitter más estudiada por la comunidad científica (Agarwal et al., 2011; Montejo-Ráez et al., 2014; Fernández et al., 2014; Pla and Hurtado, 2014), la extracción de *topics* (Lau et al., 2012; Chen et al., 2013), la identificación de perfiles (Lau et al., 2012; Chen et al., 2013), la geolocalización (Han et al., 2014), y muchas otras tareas cuyo objetivo es sacar información en claro desde textos escritos en lenguaje natural.

Sin embargo, cuando uno se enfrenta al trabajo de leer y etiquetar *tweets* (mensajes escritos en Twitter) para conseguir un recurso de entrenamiento para una tarea PLN la pregunta que recurrentemente se viene a la cabeza es: ¿realmente se puede hacer PLN sobre Twitter con los problemas de cantidad y calidad que presentan sus textos? Lo cierto es que no se puede hacer un PLN de calidad si los textos con los que se trabaja son cortos, sin una estructura gramatical, llenos de errores ortográficos o de elementos extraños (como *emoticonos* o *ASCII-art*). Hay intentos de mejorar la calidad de los textos mediante técnicas de normalización (Han and Baldwin, 2011; Villena Román et al., 2013) que consiguen limpiar un poco los *tweets* de algunos fenómenos, pero estas técnicas tienen un límite y en muchos casos hay que tratar con textos que directamente "no tienen arreglo". Estos problemas de calidad son claramente un importante handicap, pero afortunadamente hay maneras de resolver tareas sobre textos sin prestar mucha atención a los textos en sí.

Cuando Google irrumpió en 1998 con su buscador ofreciendo una solución rápida y eficaz al problema de recuperación de documentos en Internet no lo hizo porque su sistema incluyese un tratamiento de textos especialmente bueno, sino porque aprovechó los hipervínculos para evaluar la calidad de las páginas independientemente de lo que contuviesen. Es decir, usó información estructurada (los hipervínculos) como clave para resolver un problema que tenía que ver con información no estructurada (recuperar textos relacionados con una consulta). Una de las ideas más importantes que es trasversal a todos los puntos del trabajo plasmado en esta memoria de tesis se puede relacionar a lo que Google hizo en su época: aprovechar que los textos de las redes sociales poseen información estructural además del contenido textual.

Twitter no es una excepción respecto a la información estructurada. A pesar de que la calidad de los mensajes en Twitter no suele ser alta, la información estructurada que existe en ellos, en forma de relaciones entre entidades como usuarios o mensajes, es muy alta. Por ello, es posible diseñar soluciones a muchas tareas, dentro de esta red social, que contengan un componente no estructural (analizando el contenido textual de las interacciones) y otro componente estructural (analizando los datos y las relaciones entre entidades a través de los mensajes).

1.2. Hipótesis

La hipótesis principal que es la fuerza motriz del trabajo representado en esta memoria de tesis es la siguiente:

Hipótesis 1 *Los contenidos en las redes sociales ofrecen la oportunidad de integrar conocimiento proveniente de dos tipos de información de distinta naturaleza (estructurada y no estructurada) para resolver, de forma más efectiva y significativa, ciertas tareas relacionadas con el procesamiento de este tipo de contenidos.*

Con el objetivo de validar esta hipótesis, se ha experimentado con dos tareas concretas para verificar de qué manera se pueden integrar ambos tipos de información, dando a lugar a las dos siguientes hipótesis:

Hipótesis 2 *Se pueden mejorar los resultados de una tarea de recuperación de información sobre una red social si se diseña un proceso o método que combine el análisis de información tanto estructurada como no estructurada.*

Usando el caso de la red social Twitter para explorar la validez de la hipótesis 2, se parte de las siguientes consideraciones respecto a la tarea de recuperación de información sobre esta red social:

- La interfaz de consultas provista por Twitter es muy básica; funciona a base de consultas formadas por un conjunto finito y estático de palabras clave que hay que definir a priori.
- Twitter es una red social que está en constante cambio, dotándole de un carácter temporal muy dinámico. Un método de consultas estático no es una opción viable para adaptarse a sucesos imprevistos durante el periodo de recuperación de datos.
- Encontrar un conjunto de palabras clave estático que defina la temática deseada no es sencillo, pues requiere de saber qué temas y cómo los usuarios van a referirse a ellos.

Teniendo en cuenta las consideraciones anteriores, se pueden solventar los problemas que acarrea abordar esta red social mediante el diseño de un método que incluya información estructural adicional para poder abarcar la temática de forma más eficiente, complementando a los posibles conceptos encontrados en los mensajes con la relevancia en la topología de la red. Además, este método deberá ser dinámico para abordar sucesos imprevistos, por lo que debe reajustarse automáticamente teniendo en cuenta los conceptos más relevantes en la red, volviendo a hacer uso de la información estructurada para ello.

Hipótesis 3 *Se pueden mejorar los resultados de una tarea de clasificación de documentos provenientes de una red social mediante el análisis y tratamiento conjunto de información tanto estructurada como no estructurada.*

La clasificación de documentos es una tarea que, aunque es bien conocida y en contextos tradicionales de PLN está abordada, falla bastante cuando el contexto a tratar son las redes sociales. Para explorar la validez de la hipótesis 3, hay que tener en cuenta las siguientes consideraciones respecto a los documentos encontrados en las redes sociales:

- Los documentos suelen contener un gran contenido de opinión personal respecto al tema tratado, siendo principalmente motivados por un hecho polémico de reciente actividad.

- Existe un gran componente estructural en los mensajes con opinión y, con frecuencia, son respuestas a otros mensajes provistos por usuarios de mayor relevancia dentro de esa red.

- Los usuarios tienden a generar comunidades implícitas respecto a cada temática estudiada, agrupándose de forma orgánica sin necesidad de contacto directo; se agrupan compartiendo ideas e intereses comunes de forma indirecta.

- Dado que los usuarios, mensajes y sus relaciones forman parte fundamental de una red social, el análisis de los mismos no puede realizarse por separado. Gran parte de la significancia de los mensajes se pierde si no se considera el contexto de la red en el que se realiza.

A la vista de las consideraciones expuestas, se comprende que no es una buena idea la de separar los mensajes del resto de entidades existentes en la red social. Cabe esperar que establecer algún tipo de mecanismo para detectar las comunidades sea un factor importante a la hora de analizar los mensajes; la respuesta de los usuarios respecto a un tema de opinión puede verse influida por las ideas de la comunidad a la que el usuario pertenece. Es necesario abordar dicho descubrimiento de comunidades mas allá del *clustering* tradicional, utilizando la información estructural de la red social.

Sustentando la hipótesis 3, la clasificación de los mensajes de una red social respecto a una temática en cuestión se ve enormemente influida por la naturaleza del usuario dentro de esta red social. No sólo es importante el contenido del mensaje, sino el contexto de la red en el que se hace y a que grupos se refiere. Por ello, la idea de incorporar dicha información estructural como factor relevante, es bastante intuitiva, siendo importante elaborar un esquema de combinación y representación para combinar ambos tipos de información a la hora de diseñar un sistema clasificador.

En otro orden e independientemente de las hipótesis presentadas, los mensajes provenientes de las redes sociales también tienen una interesante contrapartida de carácter transversal respecto a su información textual no estructurada. Estos mensajes suelen presentar, en muchas ocasiones, problemas de calidad respecto a la redacción del texto en cuestión. Tanto el auge de los dispositivos móviles como la relación que estos dispositivos tienen con las redes sociales, resultan en textos escritos a la carrera y en cualquier situación, primando la velocidad sobre la redacción del texto.

Si se quiere cierto tipo de garantía a la hora de aplicar técnicas de PLN sobre estos textos, es necesario realizar algún tipo de tratamiento previo más allá del preprocesado tradicional. En esta tesis, también se ha abordado una tarea de esta naturaleza consistente en la normalización de tweets, cuyo objetivo consiste en restaurar los textos a nivel léxico para conseguir una redacción de mayor calidad y mejorando potencialmente cualquier tarea posterior sobre los textos restaurados.

1.3. Resumen de la propuesta

En esta tesis se proponen una serie de técnicas y métodos para validar las diferentes hipótesis y situaciones planteadas, abordando la problemática de las tareas descritas dentro del contexto de las redes sociales. Cada problemática o tarea es tratada de forma individual, proporcionando una formalización para la misma, caracterizando los fenómenos más relevantes, proponiendo uno o varios métodos para abordarla, realizando una evaluación sobre ellos y explorando los resultados de forma consecuente.

En líneas generales, las aportaciones expuestas en esta memoria de tesis son las siguientes:

1. Se aborda el problema de la recuperación de información en Twitter, paso inicial para generar cualquier dataset para cualquier otra tarea. Se propone un método dinámico y adaptativo para generar consultas que son consumibles por su sistema, utilizando una representación de grafo para calcular la relevancia de los términos en función de los tweets recogidos. Esta aportación ha sido plasmada y publicada en los artículos Cotelo et al. (2012) y Cotelo et al. (2014).

2. Se analizan los principales fenómenos de error en la redacción de los mensajes recuperados y se propone un sistema de normalización léxica altamente modular, cuya filosofía es que cada módulo realice un análisis independiente y se elija la mejor corrección. Esta aportación ha sido plasmada y publicada en los artículos Cotelo et al. (2013) y Cotelo et al. (2015a).

3. Se propone un esquema de integración para combinar modelos de características provenientes de información estructurada y no estructurada, el cual mejora significativamente la tarea de clasificación de mensajes. Esta aportación ha sido plasmada en el artículo Cotelo et al. (2015b) que está bajo revisión en el momento de la redacción de esta memoria.

4. Se explora la caracterización de grupos de usuarios dentro de un contexto específico, proporcionando un novedoso método para el descubrimiento de usuarios basado en biclustering. Esta aportación ha sido plasmada en el artículo Cotelo et al. (2015d) que está bajo revisión en el momento de la redacción de esta memoria.

Además de lo expuesto, la obra Cotelo et al. (2015c) plasma el análisis conjunto de las tareas de recuperación de información y clasificación de opinión respecto al contexto político español de la red social Twitter durante un periodo determinado, detallando como se han relacionado ambos procesos y los resultados obtenidos.

1.4. Estructura de la tesis

La estructura de esta memoria de tesis es como sigue. En el capítulo *Recuperación temática de* tweets (2) se aborda la problemática de la extracción de datasets significativos de la red Twitter, observando que la API que Twitter ofrece es insuficiente para afrontar el carácter temporal y espontáneo de la red y proponiendo un método de generación de consultas dinámicas que evolucionan

y se adaptan a los temas que toman más protagonismo dentro de una temática especificada.

En el capítulo 3 (*Normalización de tweets*) se analizan y caracterizan los diferentes fenómenos de error asociados a la escritura textual de los tweets, siendo éstos frecuentemente escritos desde dispositivos móviles, sin redacción ni reflexión antes de la publicación del mensaje. A partir de la caracterización de los errores, se propone un sistema para normalizar léxicamente estos tweets, siendo éste de naturaleza extensible y modular, fácil de ampliar requiriendo poco esfuerzo manual tanto para configurarlo inicialmente como para adaptarlo a otros contextos.

En el capítulo 4 (*Combinación de información textual y estructural aplicada a la categorización automática de tweets*) se introduce al lector al análisis de afinidades políticas en Twitter, proponiendo una tarea de clasificación múltiple para determinar la postura de los mensajes respecto al panorama político español. Para determinar las posturas de los mensajes, se proponen dos grandes aproximaciones para la clasificación: basada en contenido y basada en estructura. Finalmente, se propone un esquema de integración de ambos tipos de modelos usando diferentes técnicas.

En el capítulo 5 (*Detección de comunidades de interés mediante Spectral Biclustering*) se explora otra faceta del análisis del contexto político en Twitter, solo que ésta vez con un enfoque menos individual, intentando caracterizar colectivos de usuarios en función de intereses comunes mediante la detección de comunidades implícitas sobre una temática en particular (siendo en este caso, el panorama político español). La principal contribución consiste en una novedosa aproximación basada en biclustering, motivada por el bajo rendimiento obtenido con las técnicas de clustering tradicional. Adicionalmente, se incluye un análisis empírico de las comunidades obtenidas sobre una selección de usuarios no famosos ni oficiales con mayor relevancia obtenidas mediante el método explicado.

Finalmente, en el capítulo 6 (*Conclusiones*) se realiza un resumen de las aportaciones y se describen las conclusiones más relevantes extraídas de la experimentación y los resultados de investigación.

En lugar de existir un capítulo que aborda el estado del arte de forma monolítica, se ha tratado el estado del arte relacionado a cada capítulo individualmente, incluyéndose éste en una sección dentro de cada capítulo llamada *Trabajos Relacionados*. Cada una de esta secciones analiza el estado del arte de la tarea abordada en el capítulo, dedicando apartados específicos a diversos enfoques y trabajos de especial interés.

Capítulo 2

Recuperación temática de *tweets*

2.1. Introducción

Recientemente, Twitter ha recibido mas atención por parte de la comunidad científica debido, en gran parte, a factores como su alto potencial como recurso para tareas tales como el análisis de opinión o el seguimiento de temas "candentes" o de moda en tiempo real y su relativo éxito relacionado al reciente uso generalizado de los *smartphones*.

Sin embargo, Twitter es una enorme red social en la cual buscar y recuperar mensajes, llamados *tweets*, que traten específicamente un tema o contexto no es una tarea nada fácil. Aunque Twitter proporciona una interfaz basada en palabras claves muy similar a la que proporcionan los modernos motores de búsqueda a los que estamos acostumbrados, esta interfaz es bastante limitada respecto a su utilidad y alcance, siendo más una simple herramienta de búsqueda para el usuario común que un punto de entrada para un proceso serio de extracción de datos. Mas aún, idear un conjunto de palabras clave que sea realmente efectivo es una tarea difícil por sí misma debido a dos factores fundamentales: la dificultad de establecer de antemano un conjunto de palabras clave adecuado para representar el tema en cuestión, y la muy cambiante naturaleza de la red, pues Twitter sufre de constantes interacciones y cambios en tiempo real.

Antes de intentar analizar cualquier tipo de información específica a un tema que haya sido obtenida de Twitter, el proceso de recuperación de datos debe ser abordado de forma muy sistemática, con vistas a asegurar un buen grado de calidad y significancia en los datos extraídos. Esta idea es la principal motivación para abordar una tarea que no había sido previamente definida: dado un tema, obtener todos los tweets relacionados con ese tema durante una ventana de tiempo bien definida. Una formalización de esta tarea es necesaria para poder identificar correctamente sus principales problemas, siendo así mucho más fácil idear un método sistemático para abordar esta nueva tarea. Por ello, en este capítulo se establece una formalización para esta tarea y se presenta un método general para abordarla, aprovechando la estructura de grafo subyacente de Twitter.

Como se observa en la sección 2.2, la ausencia de aproximaciones que aborden de una forma general y sistemática el proceso de la recuperación de datos en Twitter, asegurando un alto grado de calidad y significancia en los datos obtenidos, es una de las motivaciones más importantes del esfuerzo que hay detrás de este capítulo. Las contribuciones de este capítulo se resumen a continuación:

- Se introduce el problema de la recuperación de tweets altamente relacionados con un tema específico, haciendo hincapié en las peculiaridades que deben tenerse en cuenta al tratar con información proveniente de Twitter.
- Se propone un método general para abordar dicha tarea, el cual se aprovecha de la estructura de red subyacente de Twitter.
- Se realiza un proceso de experimentación y evaluación en el cual se compara el método propuesto con otras aproximaciones típicas.
- Se realiza un análisis sobre los datos obtenidos, discutiendo algunos factores interesantes respecto al propio método propuesto y sugiriendo ciertas directrices para optimizar más aún el método.

Este capítulo se organiza de la siguiente manera. En la sección 2.2 (*Trabajos relacionados*) se realiza una revisión del estado del arte actual relacionado con la temática del capítulo y se analizan algunos trabajos de mayor interés. En la sección 2.3 (*Definición de la tarea*) se introduce el problema y se propone una formalización de la tarea en sí. En la sección 2.4 (*Método propuesto*), se describe el método previamente mencionado para abordar la tarea. En la sección 2.5 (*Experimentación*) se describe el proceso de experimentación y se muestra una comparativa entre el método propuesto y las aproximaciones que típicamente se usan en las obras mencionadas. En la sección 2.6 (*Análisis de los resultados*), se realiza un análisis detallado sobre los datos recogidos por el método y se discuten algunas consideraciones importantes a tener en cuenta. Finalmente, en la sección 2.7 (*Conclusiones y trabajo futuro*) se exponen las conclusiones de este capítulo.

2.2. Trabajos relacionados

La mayoría de los trabajos de investigación que usan datos provenientes de Twitter simplemente confían en usar la herramienta básica de consulta y, dependiendo del problema abordado, esto puede resultar en una pérdida significativa de datos y/o incurrir en un inesperado aumento del ruido en los datos recuperados. Esta situación se debe principalmente a que la correcta recuperación y tratamiento de los datos no es su preocupación principal y la mayoría de los trabajos simplemente confeccionan listas de términos de forma artesanal que usan para construir las consultas que van a ser usadas en la herramienta de búsqueda básica.

Trabajos como Tumasjan et al. (2010); Gayo-Avello et al. (2011); Hong and Nadler (2011); Pennacchiotti and Popescu (2011); Congosto et al. (2011); Agarwal et al. (2011); Davidov et al. (2010a,b); Go et al. (2009); Jiang et al. (2011); Kim et al. (2009); Pak and Paroubek (2010); Silva et al. (2011); Tan et al. (2011) sólo hacen uso de este tipo de listas acorde a sus necesidades específicas, normalmente seleccionando usuarios y *hashtags* (etiquetas) que, mediante

algún criterio *ad-hoc*, guardan cierta relación con el tema abordado en cuestión. Este tipo de solución es demasiado específica y la cobertura de datos suele ser insuficiente para la tarea abordada.

Un esfuerzo que es digno de mención, que proviene de la comunidad científica dedicada al *Information Retrieval(IR)*, es la edición del 2012 del *TREC Microblog track* (Soboroff et al., 2012). La línea presentada comparte ciertas similaridades con el objetivo principal presentado en este capítulo pero su foco es claramente distinto; la principal tarea de búsqueda presentada es la *Real-time Adhoc Task*, consistente en obtener la más reciente aunque relevante información a partir de una consulta dada. De ahí que el objetivo es responder a una consulta específica teniendo en cuenta el tiempo de la consulta, devolviendo tweets relevantes ordenados de más a menos reciente. Por motivos técnicos, en lugar de usar la propia red Twitter, esas búsquedas se realizan sobre un corpus ya previamente generado y diseñado por el organismo americano *NIST (National Institute of Standards and Technology)*.

2.2.1. Recuperación de información basada en preferencias de usuarios

En la obra Golbeck and Hansen (2011) se presenta la idea de la recuperación de tweets personalizada mediante la creación de filtros en base a las preferencias de usuario, todo ello dentro del contexto político. Para ello, presenta una metodología completamente *ad-hoc* que estima el sesgo político de la audiencia de los medios de comunicación más relevantes, usando dichos valores estimados para la creación de filtros de tweets, mejorar la experiencia del usuario y, en resumen, abordar la recuperación de tweets desde un punto de vista distinto e interesante.

Los autores estiman la preferencia política de las audiencias de los diferentes medios de comunicación usando información disponible de los informes oficiales de valoración política de los congresistas y utilizando a los usuarios seguidores de esos congresistas y consumidores de los medios de comunicación como enlace.

El método que proponen para determinar la preferencia política de la audiencia de los medios se basa en tres pasos:

1. Se establecen las puntuaciones base a priori sobre el conjunto de usuarios del grupo semilla, siendo en el caso específico explorado por los autores, los congresistas de los estados unidos. Usando el informe oficial de la asociación *Americans for Democratic Action (ADA)* referente al año 2009 (for Democratic Action, 2009), se asigna a cada congresista c un valor $p_c \in [0, 1]$ indicando cómo de liberal es ese congresista, siendo 0 un congresista muy conservador y 1 un congresista muy liberal. El uso de esta medida específica se debe a que ésta es una medida muy aceptada para valorar la posición política.

2. Se calculan los valores de preferencia política a los usuarios seguidores de los congresistas, usando para ello los valores de los congresistas anteriormente calculados. Los autores parten de la suposición que, de media, la preferencia política de un seguidor cualquiera es reflejada por los valores ideológicos de aquellos congresistas a los que siguen. Por ello, establecen que para cada seguidor f, el valor de preferencia política inferido p_f consiste en la media aritmética de los valores de pertenencia asignados a los

congresistas que el usuario f sigue, quedando que si C_f es el conjunto de congresistas a los que f sigue, $p_f = \frac{\sum_{c \in C_f} p_c}{|C_f|}$. Los valores de p_f también se encuentran en el intervalo $[0, 1]$ y tienen un significado análogo a p_c.

3. Calcular los valores de preferencia política sobre el conjunto de cuentas de usuario objeto de la investigación, usando para ello los valores de los usuarios seguidores calculados anteriormente. De manera análoga al paso anterior, se calculan los valores de los medios de comunicación, siendo el valor p_m del medio de comunicación m como la media aritmética de los valores de preferencia de los usuarios seguidores de los congresistas que siguen al medio m específicamente.

   Sin embargo, la representación política está demasiado sesgada debido a que la distribución de seguidores está muy desequilibrada; aunque la representación política de la cámara era principalmente Demócrata, los congresistas republicanos tenían un número desproporcionado de seguidores, dejando a la otra clase política claramente poco representada. Cualquier cálculo directo induciría incorrectamente a que la mayoría de los medios serían mucho mas conservadores de lo que realmente son.

   Para solventar esta situación, los autores proponen un esquema de muestreo con validación cruzada, donde las clases están equilibradas respecto a los resultados de las elecciones del congreso[1], usando $k = 10$ *folds* y ponderando el resultado final sobre todos los *folds* muestreados.

La figura 2.1 consiste en un esquema del proceso descrito, mostrando el carácter secuencial del método y los elementos que intervienen.

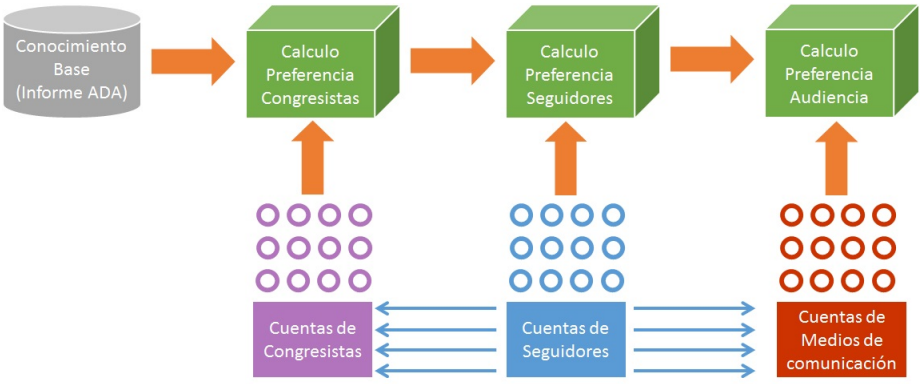


Figura 2.1: Esquema del método presentado en Golbeck and Hansen (2011)

Uno de los resultados interesantes que obtienen los autores mediante el método propuesto es que la distribución de valores de preferencia de los usuarios en el contexto político es claramente bimodal, argumentando que una de las razones principales radica en que los datos originales de los que parten también exigen bimodalidad.

Esto contrasta con otros estudios que aseguran que la distribución de la opinión política es exhibe cierta normalidad, siendo la mayoría de la población

[1] vigentes respecto al momento de la escritura del artículo

moderada. Otro factor importante a tener en cuenta se debe a que los usuarios que suelen seguir a los congresistas son usuarios políticamente conscientes y activos, características que conllevan a una serie de ideologías mas radicales, desmarcándose claramente de los sectores mas moderados.

Este método utiliza, aunque muy limitadamente, la topología de red del contexto político que exploran, consiguiendo un modelo básico para caracterizar al grupo objetivo mediante los propios usuarios, partiendo de un conocimiento base previo. Este modelo tiene utilidad dentro de la recuperación de información pues permite realizar un ajuste o filtrado sobre un conjunto obtenido previamente, mejorando la calidad y el enfoque del subconjunto resultante. Incluso se puede ajustar para personalización o análisis de sesgo.

Sin embargo, este método no es general ni dinámico, requiere de cierto tipo de información base inicial que no siempre se puede disponer y no tiene en cuenta los contenidos de los usuarios ni cómo éstos reaccionarían a eventos imprevistos, pues sólo tiene en cuenta la topología estática en un momento determinado del tiempo.

2.2.2. Métodos con *pseudo-relevance feedback*

Relevance feedback es una característica que exhiben algunos métodos de *IR*, basada en la idea de utilizar los resultados relevantes de una consulta inicialmente dada para realizar otra consulta (o refinar una existente), descartando los no relevantes en el proceso. Existen tres tipos principales de *relevance feedback*: explícito, implícito y "pseudo" o ciego. Tanto el explícito como el implícito requieren de intervención humana, mientras que el "pseudo" es completamente automático.

En nuestro caso, los únicos métodos que interesan son aquellos con *pseudo-relevance feedback* o "ciegos", dado que el carácter iterativo y de refinamiento se asemeja, en términos muy generales, al método dinámico adaptativo que se propone.

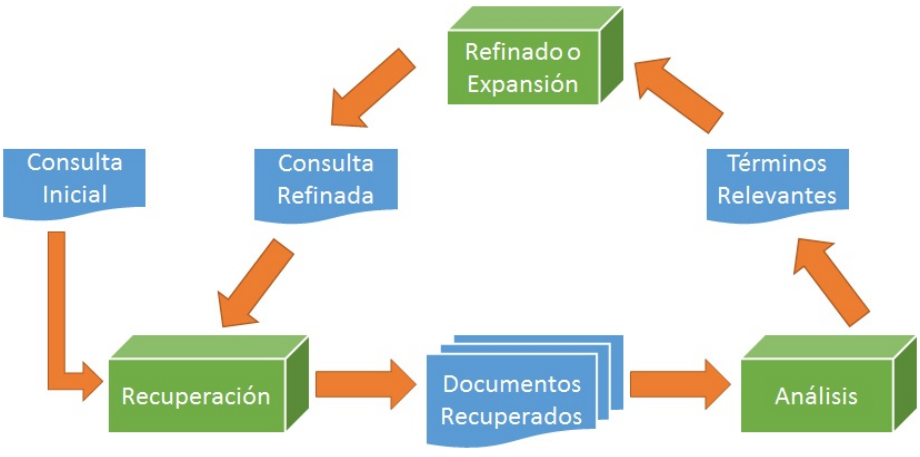


Figura 2.2: Esquema general de un método con *pseudo-relevance feedback*

Aunque los detalles explícitos dependen del método en cuestión, el proceso general consiste en un análisis automático local de los datos y las consultas.

En términos generales, el sistema parte de un conjunto definido de documentos, los analiza, elige los mejores y utiliza esa información para refinar la consulta. La figura 2.2 muestra un esquema general de un proceso iterativo con *pseudo-relevance feedback*.

Estos métodos suelen funcionar bastante bien y, mediante la expansión de consultas, pueden mejorar la cobertura de las consultas realizadas inicialmente y mejorar el rendimiento global. No obstante, este efecto se basa enormemente en la calidad del análisis y la selección, y cualquier procedimiento de este tipo puede derivar en resultados erróneos y/o no deseados si el proceso no es controlado correctamente.

Sin embargo, nuestro método difiere en puntos bastante esenciales; todo el proceso de análisis y refinamiento se basa en un análisis topológico de los elementos estructurales de los mensajes, como se explica en la sección 2.4. Puede encontrarse mas información sobre modelos con "pseudo-relevance feedback" o efectos similares en obras como Lee et al. (2008), Tao and Zhai (2006), Cao et al. (2008) y Lavrenko and Croft (2001).

2.3. Definición de la tarea

El objetivo final de la tarea tratada en este capítulo consiste en la recuperación de todos los mensajes de los usuarios en Twitter (llamados *tweets*) relacionados con un tema específico y que han sido enviados durante una ventana de tiempo en concreto. Para entender la complejidad esta tarea, hay que aclarar tres conceptos importantes que la definen: Twitter, tema y tiempo.

En Twitter, los tweets poseen ciertas características especiales a considerar. En primer lugar, existe una restricción estricta respecto a la longitud de los tweets, siendo 140 caracteres este límite superior, cosa que no es del todo inusual pues la brevedad de los mensajes es una parte fundamental en cualquier red de micro-blogging. Como contrapartida, restricciones de longitud muy frecuentemente conllevan fenómenos como palabras acortadas, abundancia de acrónimos y técnicas similares para poder lidiar con este límite impuesto, resultando en una pérdida de calidad general respecto a la redacción del texto. Incluso herramientas acortadoras de URL se han convertido en una necesidad cuando se desea incluir hipervínculos en el tweet.

En segundo lugar, los tweets sólo pueden contener lo que se denomina *texto plano*, lo que significa que el contenido no posee formato de ningún tipo. Twitter proporciona adicionalmente un par de constructos especiales (también representados en texto plano) para especificar interrelaciones entre usuarios. Uno de ellos, es la mención directa a otros usuarios en un tweet usando el carácter '@' como prefijo al nombre del usuario mencionado en cuestión.

El otro tipo de constructo son los denominados comúnmente *"hashtags"*, los cuales son palabras o inclusos expresiones (palabras concatenadas sin usar espacios) prefijadas por el carácter '#'. Se usan principalmente para marcar el mensaje y participar en una especie de discusión *ad-hoc* sin mecanismo de moderación sobre un tema bastante específico, funcionando como etiquetas de metadatos sin administración por parte de la plataforma. Más aún, estos hashtags son promocionados por usuarios individuales y algunas veces actúan como "señales" o indicadores, pudiéndose vagamente relacionar entre sí tweets que hayan usado hashtag, estableciéndose así un criterio de agrupación.

En tercer lugar, Twitter sólo ofrece el uso de consultas directas para poder obtener los mensajes de su plataforma. Las consultas deben estar formadas por palabras claves (ya sean palabras, hashtags o nombres de usuario) y su composición es bastante simple, asemejándose mucho a las interfaces ofrecidas por los típicos motores de búsqueda web. Aunque simple e intuitivo, este tipo de interfaz restringe drásticamente y dificulta el diseño de aproximaciones para resolver la tarea de la recuperación temática de tweets.

El concepto de *tema* es sencillo de entender de forma intuitiva pero es difícil establecer una definición exacta. En términos generales, podríamos definir un tema como el sujeto u objecto principal de una discusión entre usuarios. Asimismo, la realización del tema deseado podría ser difícil por sí misma, encontrando dificultades en la "transformación" del concepto abstracto a algo que puede ser consumido por sistemas de información.

Con respecto al concepto de la temporalidad, otra cosa a tener muy en cuenta es el dinamismo temporal de la propia red. Twitter es una red social que exhibe constantes cambios en tiempo real y su comunidad reacciona muy rápidamente a cualquier evento, creando nuevas tendencias, temas y hashtags constantemente. Esto dificulta considerablemente el seguimiento de temas a lo largo de un periodo de tiempo.

Todos estos factores convierten la recuperación temática de tweets en una tarea nada trivial pero muy interesante: generar consultas que aseguren la recuperación de tweets de alta significancia a lo largo del tiempo. Una vez expuesta esta explicación intuitiva de la tarea, se propone una formalización de la misma (Definición 2.1) que servirá como apoyo para el resto de este capítulo.

Definición 2.1 *Dadas las siguientes disposiciones:*

- *Sea T el conjunto de todos los tweets existentes en Twitter. Dado que T crece continuamente en el tiempo, se denota como T^t al conjunto de todos los tweets existentes en el instante t.*
- *Sea Q el conjunto de todas las consultas que se pueden hacer sobre T. Cada consulta $q \subseteq Q$ es un conjunto específico de palabras clave cuya ejecución en el instante t devuelve $T_q^t \subseteq T^t$ que es el conjunto de tweets en T^t que contienen al menos una de las palabras claves de q.*
- *Sea $T_{tema}^t \subseteq T^t$ el conjunto de todos los tweets existentes en Twitter en el instante t que guardan relación sobre un tema específico.*
- *Sea $Q_{tema}^t \subseteq Q$ tal que $Q_{tema}^t = \left\{q_i : T_{q_i}^t \supseteq T_{tema}^t\right\}$.*

Dado un instante t, la tarea de recuperación temática de tweets consiste en:

1. *Caracterizar un $q_i \in Q_{topic}^t$, preferiblemente con un $|T_{q_i}^t|$ pequeño, que limite el número de tweets recuperados que no guarden relación con el tema en cuestión.*
2. *Seleccionar los tweets del resultante $T_{q_i}^t$ que pertenezcan a T_{topic}^t, eliminando los tweets que no están relacionados con el tema en cuestión.*

Esta formalización realiza una división natural de la tarea en dos subtareas, cada una relacionada con los puntos (1) y (2) de la definición 2.1 respectivamente: *generación de consultas* y *filtrado*. Aunque ambas tareas están claramente

relacionadas, esta división es conveniente pues permite enfocarse en cada subtarea independiente, siendo mas sencillo y generando soluciones más eficientes para cada uno de los desafíos por separado.

Además, cada tarea cumple diferentes roles dentro del sistema, afectando al rendimiento del mismo de muy diferente manera. La subtarea *generación de consultas* fundamentalmente afecta a la capacidad global de recolección del sistema, debido a que el conjunto de tweets obtenidos en primera instancia depende exclusivamente de la consulta generada. Generar consultas muy generalistas puede incrementar esa capacidad de recolección pero introducirá una gran cantidad de ruido innecesariamente.

Por otra parte, la subtarea de *fitrado* es la que realiza un proceso de selección *a posterior* sobre los tweets obtenidos en primera instancia, intentando minimizar el ruido y afectando en gran medida a la precisión final del todo el sistema. Sin embargo, un filtrado muy exigente podría descartar tweets que guardan relación con la temática. La subtarea de filtrado se analiza y aborda en la sección 2.6.2, separadamente de la subtarea de generación de consultas.

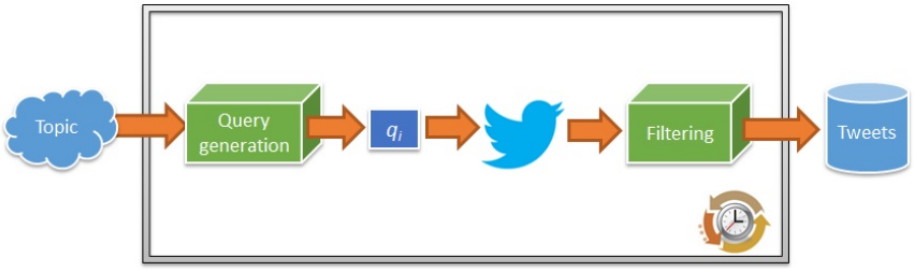


Figura 2.3: Recuperación temática de tweets

La figura 2.3 muestra los elementos principales mencionados en esta sección. Como nota aclaratoria, el concepto del "tiempo" está representado en el diagrama mediante la metáfora del reloj.

La tarea en su conjunto no sólo es difícil de caracterizar, sino también de resolver y evaluar. Las peculiaridades técnicas de Twitter, la dificultad del proceso de realización del tema deseado en cuestión a algo más concreto (como un conjunto de palabras clave), la propia naturaleza de los usuarios y de sus mensajes generados junto con el comportamiento temporal de la tarea son las dificultades prominentes a tener en cuenta y resolver.

2.4. Método propuesto

En esta sección se propone un método para abordar la tarea de la recuperación temática de tweets, siendo este un método para resolver específicamente la subtarea de la *generación de consultas*. Tal y como hemos mencionando en la sección anterior, la generación de consultas es principalmente responsable de la cobertura global de la tarea, así que el foco principal de este método es el de aumentar la cobertura lo mejor posible sin incurrir en una inaceptable pérdida de precisión.

Dado que Twitter es realmente un grafo gigantesco con varios tipos de relaciones entre sus elementos constituyentes, no es nada descabellado tomar una

aproximación basada en grafos para representar y analizar sus contenidos. En esencia, el método consiste en construir una representación de grafo a partir de los tweets tomando los hashtags y usuarios como nodos, realizar un ranking de relevancia sobre dichos nodos, seleccionar los mejores y generar una consulta con ellos, realizándose todo este proceso de forma iterativa.

2.4.1. Construcción del grafo

Para construir la representación de grafo a partir una colección de tweets recopilados, se parte de la idea de generar nodos y aristas usando los elementos que componen un tweet: palabras, usuarios y hashtags. En términos generales, un usuario puede simplemente escribir texto, mencionar a un usuario, retuitear otro tweet y hacer uso de los hashtags que quiera. La figura 2.4 muestra los elementos posibles en un tweet y las algunas de las relaciones que pueden existir entre ellos.

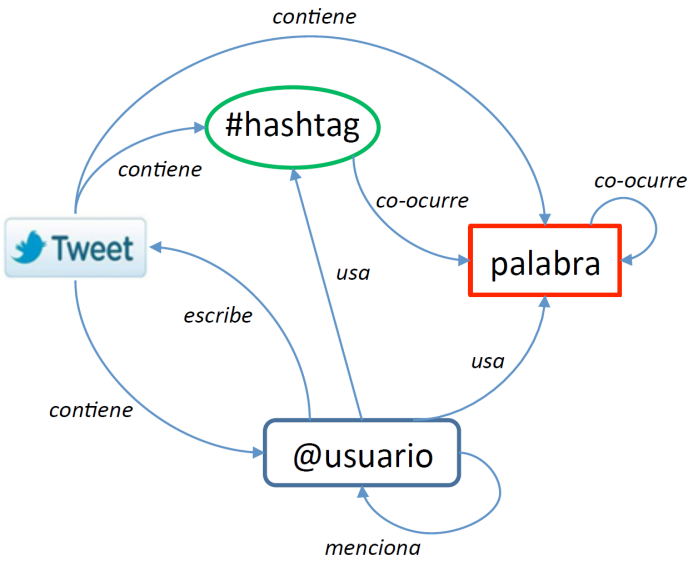


Figura 2.4: Grafo de elementos y relaciones posibles en un tweet

El método propuesto no se basa en la información texual general de un tweet, sino que hace uso de los elementos estructurales posibles: usuarios y hashtags. Teniendo en cuenta estos elementos estructurales, en los tweets podemos encontrar y definir las siguientes relaciones:

- *Menciones*: el autor de un tweet incluye el nombre de cualquier otro usuario en los contenidos de ese tweet prefijándolo con el carácter '@'. Este tipo relación entre usuarios se utiliza cuando alguien quiere referirse a un usuario en especial o responder a algunos de sus tweets.

- *Retweets*: el contenido de este tweet es esencialmente el mismo que el de otro tweet al que hace referencia, prefijado dicho contenido con algun tipo de fraseo especial como "`RT @usuario_original:`", haciendo mención al usuario original o utilizando el botón `retweet` que ofrece la interfaz de

usuario de Twitter. Se puede ver como subtipo específico de la relación de *mención* en la que también se replican las relaciones encontradas en el tweet original, debido a que los contenidos son prácticamente iguales.

- *Uso sencillo de hashtag*: el autor del tweet incluye un hashtag de su elección en los contenidos de ese tweet, generando así una relación entre el usuario y el hashtag.
- *Coocurrencia de hashtags*: El autor incluye más de un hashtag en los contenidos del tweet. Este caso es claramente una extensión del caso anterior, añadiendo una nueva relación entre los dos o más hashtags que aparecen.

Partiendo de estas relaciones observadas, se construye un grafo estructural que representa la actual topología de red entre usuarios y hashtags, donde los nodos son los hashtags y los usuarios, y las aristas entre ellos representan las relaciones encontradas en los contenidos de los tweets recuperados.

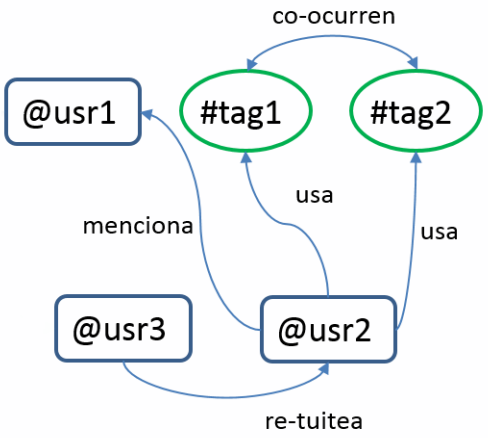


Figura 2.5: Ejemplo de multiples relaciones en un solo tweet

Las *menciones* y los *retweets* son representados mediante arcos dirigidos entre usuarios mientras que los *usos sencillos de hashtags* son representados mediante arcos dirigidos de nodos usuarios a nodos hastags. La *coocurrencia de hashtags* se repsesenta estableciendo un arco no dirigido[2] entre ambos *hashtags*. Si algún arco ya existe, su peso se ve incrementado para reforzar ese tipo de relación. Es posible que un tweet exhiba mas de una instancia específica de relación, por ejemplo, varias menciones y uso de varios hashtags. La figura 2.5 muestra este fenómeno de múltiple instanciación de relaciones, haciendo referencia al siguiente tweet de ejemplo: `@usr3:‘‘RT @usr2: ejemplo con @usr1 #tag1 #tag2’’`.

Para aclarar el proceso de construcción del grafo, a continuación se muestra un ejemplo minimalista del proceso paso a paso, usando la secuencia de tweets mostrada en la figura 2.6. La tabla 2.1 muestra el desglose individual de las relaciones que cada tweet alberga en este ejemplo.

[2]Un arco no dirigido se puede interpretar como si existieran dos arcos dirigidos entre los dos nodos en cuestión, cada uno en una dirección diferente.

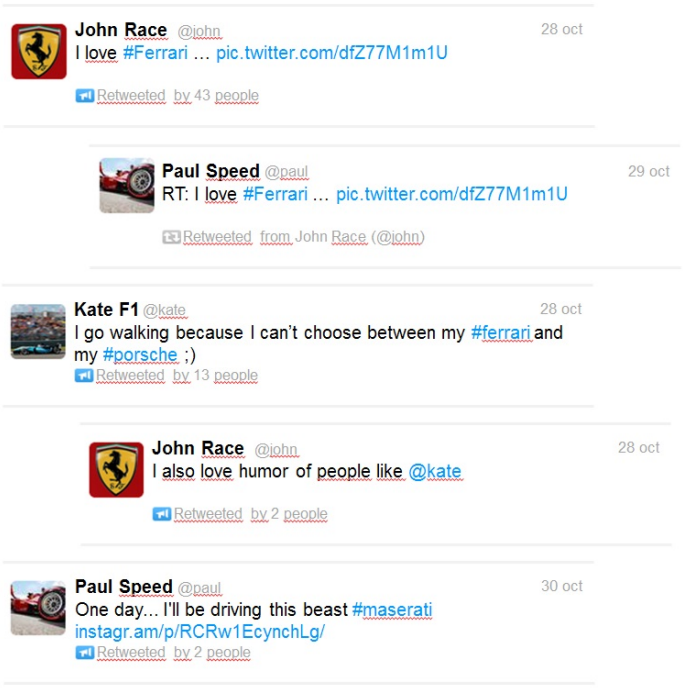


Figura 2.6: Muestra de interacción de usuarios en Twitter relacionada con Ferrari

El proceso completo de construcción del grafo también es mostrado en la figura 2.7, siendo cada subfigura parte de la secuencia del proceso. El proceso se realiza como sigue a continuación:

Tweet 1: Crear los nodos `@john` and `#ferrari` y establecer un arco dirigido de `@john` a `#ferrari`.

Tweet 2: Crear el nodo `@paul`, establecer un arco dirigido de `@paul` a `@john` e incrementar el peso de arco existente de `@john` a `#ferrari`.

Tweet 3: Crear el nodo `@kate`, establecer un arco dirigido de `@kate` a `#ferrari` y establacer otro de `@kate` a `#porsche`.

Tweet 4: Establecer un arco dirigido de `@john` a `@kate`.

Tweet 5: Crear el nodo `#maserati` y establecer un arco dirigido de `@paul` a `#maserati`.

En este ejemplo, el grafo resultante de representar a estos 5 tweets posee 6 nodos y 7 aristas con una baja densidad de $D = 0{,}233$. Un ejemplo del mundo real poseería millones de nodos, decenas de millones de aristas y sería mas denso.

2.4.2. Análisis del grafo

Después de que el grafo haya sido construido, se le aplica un algoritmo de ranking de relevancia, obteniendo los nodos mas relevantes del grafo teniendo

Tweet	Autor	Relaciones	Contenidos
1	@john	uso de hashtag	`I love #Ferrari ... pic.twitter.com/dfZ77M1m1U`
2	@paul	retweet	`RT: I love #Ferrari ... pic.twitter.com/dfZ77M1m1U`
3	@kate	uso de hashtag	`I go walking because I can't choose between my #ferrari and my #porsche ;)`
4	@john	mención	`I also love humour of people like @kate`
5	@paul	uso de hashtag	`One day... I'll be driving this beast #maserati instagr.am/p/RCRw1EcynchLg/`

Cuadro 2.1: Desglose de la interacción de usuarios perteneciente a la figura 2.6

en cuenta la topología del mismo. El algoritmo de ranking elegido en cuestión es el ampliamente conocido *PageRank*.

PageRank (Page et al., 1999) es un algoritmo para generar rankings de relevancia originalmente pensado para medir la importancia en Internet de una página web de acuerdo a los enlaces que apuntan a ella, pero el algoritmo es lo suficientemente general para su aplicación en otros contextos diferentes. Dado un grafo $G = (V, E)$ donde V es un conjunto de vértices y E un conjunto de aristas dirigidas entre dos vértices, se definen de antemano dos operaciones $In(V_i)$ y $Out(V_i)$ para calcular, respectivamente, el conjunto de aristas entrantes y salientes al vértice V_i. A partir de estas dos operaciones básicas, podemos definir la puntuación PageRank de un vértice dado de acuerdo a la siguiente fórmula:

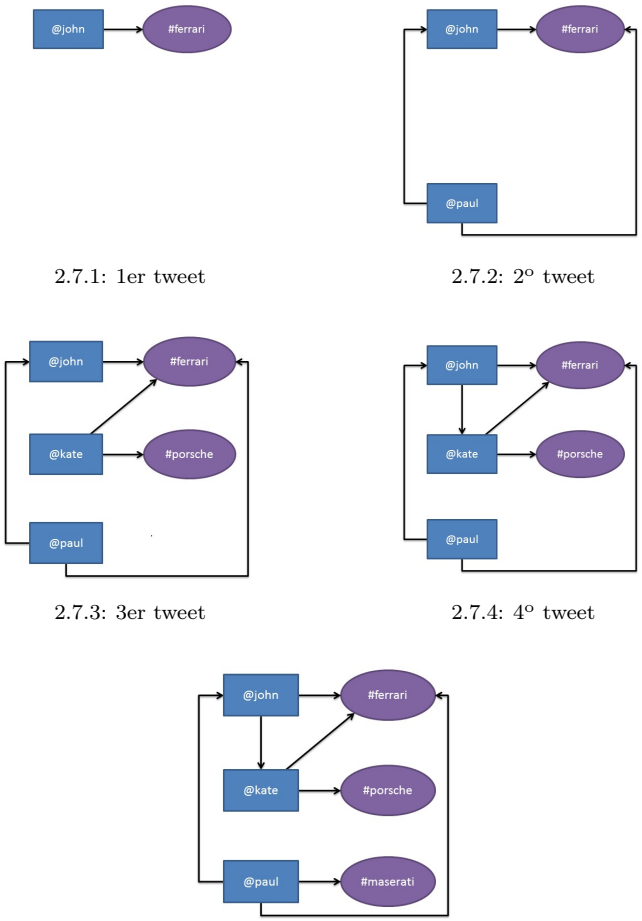
$$PR(V_i) = (1 - d) + d \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} PR(V_j), \tag{2.1}$$

donde d es un factor de atenuación que permite la correcta convergencia del algoritmo. En el contexto original de la navegación en Internet, este factor representa la probabilidad de que un usuario acceda a una página desde un enlace de la página actual, siendo $(1 - d)$ la probabilidad de que un usuario salte a una página aleatoria no enlazada a la página actual. En la obra original, teniendo en cuenta el contexto para el cual se definió el algoritmo, se recomienda un valor de 0,85 para el factor d.

Estableciendo inicialmente valores arbitrarios para las puntuaciones de los nodos, se alcanza un punto de convergencia mediante la aplicación iterativa de la fórmula, siendo el criterio de parada que la mayor diferencia de puntuaciones para cada nodo entre dos iteraciones consecutivas sea inferior a un valor umbral. Una vez que el algoritmo ha terminado, la puntuación de cada nodo representa la importancia de dicho nodo dentro de la red, pudiéndose establecer un ranking y utilizarse como criterio para la toma de decisiones.

2.4.3. Descripción del método

Los procesos de construcción y análisis del grafo son elementos fundamentales de nuestro método, de ahí que hayan sido definidos previamente antes de



2.7.1: 1er tweet

2.7.2: 2º tweet

2.7.3: 3er tweet

2.7.4: 4º tweet

2.7.5: 5º tweet

Figura 2.7: Proceso de construcción del grafo paso a paso del ejemplo provisto

entrar la propia definición del método. En esta sección se propone un método iterativo el cual, a cada paso, analiza los datos recolectados en la iteración anterior para generar una consulta apropiada para el siguiente paso. En esta sección también se especifica cómo construir la consulta exactamente y cuáles de los datos recolectados se usan para la construcción del grafo a cada paso.

La figura 2.8 muestra el proceso completo como una especialización de la definición de la tarea previamente presentada en la sección 2.3 y mediante la figura 2.3 de dicha sección. Nótese que este método está puramente enfocado a resolver la subtarea de *generación de consultas* y no se realiza etapa de filtrado alguna. Adicionalmente, el método se describe completamente mediante el algoritmo 1.

En cada una de las iteraciones, se realiza el proceso de generación y análisis de grafo partiendo de los tweets previamente obtenidos durante una ventana de tiempo determinada. Se puede elegir cualquier ventana de tiempo, variando desde varios minutos hasta días, pero se observó experimentalmente que una

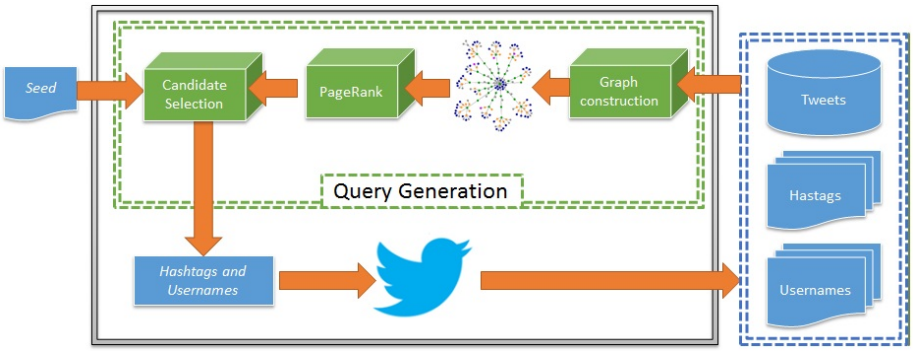


Figura 2.8: Método propuesto para resolver la tarea de recuperación temática de tweets

ventana de 60 minutos es una elección muy adecuada.

Al no tener iteración anterior, la primera iteración del método carece de tweets previamente capturados, necesitando algún tipo parámetro inicial. En este caso, se usa un conjunto *semilla* de palabras clave, siendo lo más apropiado que este conjunto semilla represente el conjunto central del tema seleccionado de forma adecuada.

Después de realizar el paso de construcción del grafo con los datos de la iteración anterior, se usa PageRank para confeccionar un ranking de relevancia sobre los nodos del grafo, el cual se va a utilizar como lista tentativa de candidatos. El proceso de selección de candidatos se realiza simplemente eligiendo los mejores nodos y descartando el resto, estableciendo algún punto como umbral en la lista, cuyo valor se determina de antemano. Para este caso, se determina que el punto de corte es un parámetro entero k, representando el tamaño máximo del conjunto de elementos seleccionados, fijo de antemano al inicio del método.

Finalmente, generar la consulta consumible por Twitter es un paso bastante directo. Los elementos seleccionados por la etapa anterior son usados como palabras clave y todos son concatenados usando el típico operador *or*. En esencia, usamos los mismos elementos estructurales propios de Twitter (usuarios y hashtags) para la construcción de las consultas pero teniendo en cuenta la topología de red subyacente y la naturaleza cambiante de Twitter.

Como ya se ha comentado previamente en el apartado 2.2.2, el método propuesto guarda cierta semejanza con los métodos que poseen o se basan en el efecto *pseudo-relevance feedback* usado en *IR* y a nivel muy esencial tienen elementos en común. Sin embargo, nuestro método difiere muy significativamente del modelo *pseudo-relevance feedback* en varios puntos clave: construcción de grafo y análisis topológico en lugar de usar un modelo *vector space*, tener en cuenta los elementos estructurales en lugar de sólo el contenido y la realización de las consultas usando esos elementos estructurales.

2.5. Experimentación

Para poder probar el rendimiento del método propuesto en este capítulo, se ha aplicado sobre un tema en concreto, realizado un análisis sobre el conjunto

ALGORITMO 1: Método propuesto

Data: Términos semillla s, ventana de análisis w, tiempo de iteración i

begin
  $Terminos \leftarrow \{s\}$
  $T \leftarrow \emptyset$
  repeat
    $T \leftarrow T \cup$ `AplicarConsultaTwitter(`$Terminos$, i`)`
    $G \leftarrow$ `CrearGrafoVacío()`
    $T_w \leftarrow$ tweets de T obtenidos en las últimas w iteraciones
    foreach *tweet* $t \in T_w$ *tales que el término* s *aparece en* t **do**
      $user \leftarrow$ usuario que escribió t
      $tags \leftarrow$ hashtags que aparecen en t
      $mentions \leftarrow$ menciones de usuario que aparecen en t
      if $user \notin G$ **then**
        actualiza G con un nodo que represente a $user$
      foreach $tag \in tags$ **do**
        if $tag \notin G$ **then**
          actualiza G con un nodo que represente tag
        if *arco dirigido* $a = (user, tag) \notin G$ **then**
          añadir a G el arco dirigido a
        incrementa el peso del arco existente $a = (user, tag)$
      foreach $mentioned\_user \in mentions$ **do**
        if $mentioned\_user \notin G$ **then**
          actualiza G con un nodo que represente a $mentioned\_user$
        if *arco dirigido* $a = (user, mentioned\_user) \notin G$ **then**
          añadir a G el arco dirigido a
        incrementa el peso del arco existente $a = (user, mentioned\_user)$
      foreach $tag_1, tag_2 \in tags$ *tales que* $tag_1 \neq tag_2$ **do**
        if *arco no dirigido* $a = (tag_1, tag_2) \notin G$ **then**
          añadir a G el arco no dirigido a
        incrementa el peso del arco existente $a = (tag_1, tag_2)$
    $R \leftarrow$ `PageRank(`G`)`
    $Terminos \leftarrow$ primeros n términos de R
  until *fin de la sesión de recuperación de tweets*

de datos obtenido y generado una comparativa con otras aproximaciones más simples y con diferentes parámetros de configuración.

2.5.1. Entorno experimental

El tema seleccionado para la recolección de datos fue el evento *Campeonato Europeo de Fútbol de la UEFA 2012*, coloquialmente llamado *Eurocopa 2012* o *Euro2012*. Este evento es una competición deportiva muy importante en el continente europeo y muy seguida por todos los medios de comunicación, especialmente por los medios de deporte.

El evento fue abordado desde la perspectiva española, siendo solo de interés los tweets en Español. El término central usado como semilla para el método fue el famoso hashtag `#vamosespaña`. Este hashtag era bien recibido en la comunidad española, siendo un término muy común para animar a la selección española de fútbol que participaba en el evento.

La recolección de tweets se hizo durante la duración de todo el evento, que duró desde el 6 de junio del 2012 al 3 de julio del 2012. Para ello se usaron tres métodos diferentes en paralelo, que se describen a continuación:

- **Palabra clave central**: Usar la palabra clave central elegida, siendo en este caso el término `#vamosespaña`, como único término de consulta durante todo el evento. Esta aproximación simplista es usada como *baseline* comparativo respecto al resto de métodos durante el análisis experimental.
- **Lista estática ad-hoc de palabras claves**: Está aproximación consiste en hacer uso de una colección de palabras clave elegidas por un experto (siendo estas usualmente hashtags) como consulta estática a lo largo de todo el evento. Para probar este método, se ha elegido el siguiente conjunto de palabras clave: `#vamosespaña`, `#nohaydossintres`, `#eurocopa`, `#eurocopa2012`, `#laroja`. Estos términos fueron elegidos por su alta relevancia y muy baja ambigüedad respecto al evento.
- **Método dinámico propuesto**: Se usa como semilla inicial para el método la palabra clave central `#vamosespaña`, variando el tamaño máximo de palabras claves inferidas hasta un valor de $k = 20$.

2.5.2. Muestreo y etiquetado

Evaluar completamente un dataset de estas características sería una tarea titánica, pues requeriría de revisar manualmente del orden de cientos de miles o incluso millones de tweets; el dataset del caso partícular tratado en este capítulo está compuesto por unos 3 millones de tweets.

En su lugar se escoge una muestra aleatoria significativa que, acorde al teorema central del límite, estableciendo un nivel de significancia estadística $\alpha = 0{,}05$ y una cota superior de error del 1 %, consta de 10000 mensajes aunque solo se requieren estrictamente 9604 mensajes. Esta muestra permite inferir con precisión las propiedades del dataset original y es usada durante todo el proceso de evaluación en lugar del dataset original.

Para poder estimar la precisión, tres revisores etiquetaron manualmente cada mensaje si era relevante o no, muy similar a un proceso de clasificación binaria. La tabla 2.2 muestra los datos de precisión acorde a cada revisor y el índice de

consenso entre revisores. Los valores de consenso y el estimador κ (Fleiss, 1971) indican que la tarea de etiquetado está bien definida y que las anotaciones hechas por los revisores son lo suficientemente fiables para usar la media armónica de los valores de precisión obtenidos individualmente como medida de precisión para posteriores usos en este proceso de evaluación.

Medida	Revisor 1	Revisor 2	Revisor 3
Precision	0.8882	0.8292	0.8715
Consenso con Revisor 1	1.0	0.9290	0.9651
Consenso con Revisor 2	0.9290	1.0	0.9375
Consenso con Revisor 3	0.9651	0.9375	1.0
Consenso completo	0.9158		
Estimador κ	0.7627		

Cuadro 2.2: Precisión calculada y valores de consenso

2.5.3. Evaluación

Mediante la muestra anotada, se puede estimar la precisión (proporción de los tweets recuperados que son relevantes) con un error inferior al 1 %, siendo esta debidamente calculada e incluida en este estudio. Sin embargo, el cómputo de la cobertura, llamada *recall* en la literatura, presenta grandes dificultades en este estudio.

Para explicar por qué el cálculo exacto de la cobertura no es factible, es necesario entender en que consiste exactamente: mide cuantos documentos relevantes han sido en comparación al total de documentos relevantes que existen. En este caso, significaría que se debería saber el número exacto de tweets existentes en todo Twitter que están relacionados con el tema seleccionado, cosa que no es para nada factible.

Mas aún, el muestreo no es una técnica viable para obtener una estimación correcta sobre el número de tweets existentes relacionados con la temática. Aunque Twitter proporciona un flujo muestral con una distribución pseudoaleatoria de mensajes, es extremadamente raro que algún mensaje del tema en cuestión aparezca en ese flujo, dado que cualquier temática a elegir es insignificante en comparación con el total de temáticas que pueden darse en toda la red Twitter en un momento determinado.

Si consideramos el tamaño máximo evaluado para el conjunto de palabras clave en este entorno experimental ($k = 20$), el dataset resultante posee alrededor de 3 millones de mensajes y en Twitter se generan mil millones de mensajes semanales de media, dejando muy poco margen para obtener una muestra significativa de mensajes relevantes totales. Por ello, se consideró dejar de lado el uso la métrica *recall* exacta para el proceso de evaluación y su uso para cualquier otra medida.

En su lugar, se ha definido una medida alternativa llamada *cobertura de dataset* o *dataset recall*, la cual es similar a la medida de *recall* pero a nivel de dataset, solo teniendo en cuenta todo los tweets relevante obtenidos como estimación de los todos los tweets relevantes existentes. Esta alternativa permite medir la capacidad de cobertura del método respecto a diferentes parámetros.

Al igual que con la medida de precisión, esta medida produce cualquier valor $x \in \mathbb{R} : x \in [0, 1]$.

La tabla 2.3 muestra una comparativa de los resultados de rendimiento del método propuesto respecto al baseline y el método de lista estática ad-hoc usado en la mayoría de las aplicaciones. Como se ha mencionado antes, los términos propuestos para el método de lista estática fueron aquellos que tuvieron alta relevancia dentro de la comunidad española respecto al evento (`#vamosespaña,#eurocopa,#eurocopa2012,#laroja`), mientras que para el método dinámico propuesto, se muestran valores de rendimiento para los parámetros $k = 10$ y $k = 20$.

Método	Precisión	Dataset Recall
Baseline (sólo el término central)	0,9726	0,1504
Lista estática seleccionada	0,9721	0,2698
Método dinámico con $k = 10$	0,9274	0,8034
Método dinámico con $k = 20$	0,8713	1,0000

Cuadro 2.3: Comparativa de rendimiento entre métodos

En líneas generales, se observa que nuestro método obtiene un mayor volumen de datos a expensas de un ligero descenso en la precisión aunque elegir correctamente el parámetro k no es trivial y depende mucho del problema en cuestión. Esta pérdida de precisión se debe en parte a la inclusión no deseada de usuarios con un alto número de seguidores, cuyos mensajes tienen un gran impacto en la red social pero son altamente "ruidosos", debido a que suelen escribir sobre una gran variedad mensajes de temática muy variada y son usuarios muy mencionados [3].

2.6. Análisis de los resultados

Una manera de analizar efectivamente el rendimiento del método consiste en la observación de su comportamiento respecto al tamaño del conjunto de palabras claves dinámicamente generado, reflejando el grado de bondad de las consultas generadas y que efectos tienen sobre el dataset.

2.6.1. Tamaño del conjunto de palabras clave

La figura 2.9 muestra el comportamiento de la precisión del dataset respecto a diferentes valores de tamaño máximo del conjunto de términos (k) y es fácil de observar que la pérdida de precisión está claramente asociada al incremento del tamaño máximo del conjunto de términos. Este comportamiento no es inesperado, pues el método intenta aumentar el volumen del conjunto de datos mediante la introducción de nuevos términos, siendo esta una forma usual de introducir ruido en el dataset. Sin embargo, esta pérdida de precisión es relativamente baja, siendo muy estable para determinados intervalos del parámetro k.

[3] Cuando se hace una consulta a Twitter que contiene un nombre de usuario, se obtienen tanto los tweets escritos por ese usuario como los que los tweets que hacen mención a este.

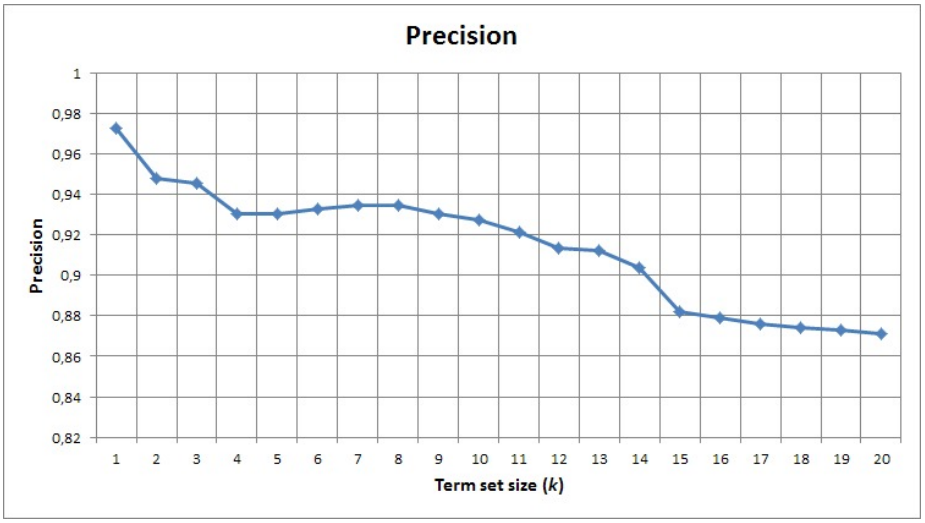


Figura 2.9: Precisión del dataset respecto al tamaño máximo del conjunto de términos (k)

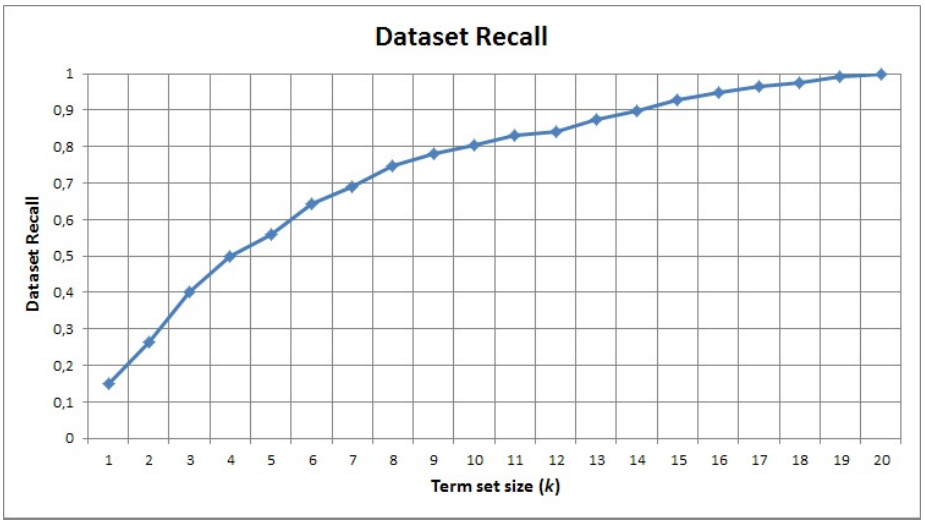


Figura 2.10: Cobertura de dataset o *Dataset recall* respecto al tamaño máximo del conjunto de términos (k)

Como se ha mencionado anteriormente, el cómputo de la medida *recall* no es factible, habiendo utilizado la cobertura de dataset o *dataset recall* en su lugar. Aunque no es un substituto real de la medida original, el uso de esta medida alternativa se considera como muy apropiado, debido a que el foco principal del método es el de incrementar el volumen de datos obtenidos sin incurrir una alta pérdida en la precisión. La figura 2.10 muestra el comportamiento de la medida alternativa *dataset recall* respecto a diferentes valores del parámetro k.

Un método mejor para identificar el "mejor" valor de k consiste en ordenar ambas métricas en conjunto, similar a una figura de *frontera de Pareto*, siendo

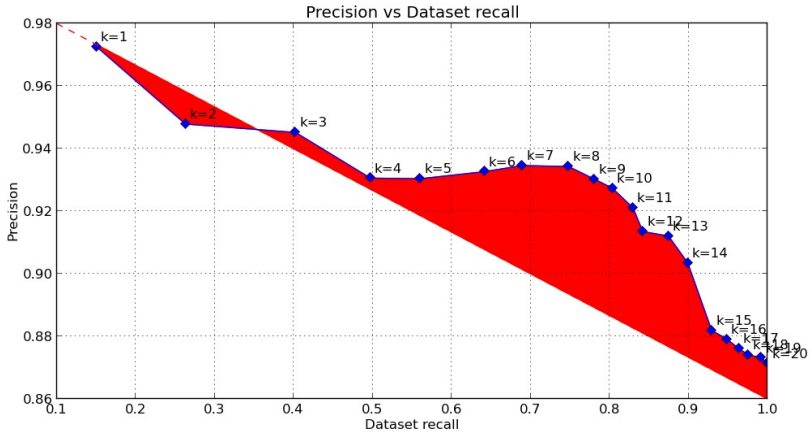


Figura 2.11: Precisión y *dataset recall* respecto a diferentes valores k, mostrando una curva de *frontera de Pareto*.

la figura 2.11 un claro ejemplo de este tipo de combinación. En ella se muestra una curva de *eficiencia paramétrica*, siendo cada punto de la curva la precisión y el *dataset recall* para cada valor de k, observándose que la curva exhibe un buen compromiso entre precisión y volumen de datos para un valor de $k = 8$.

Ese punto en concreto ($k = 8$) consigue una precisión del 93,43 % y una cobertura de dataset del 74,78 %, resultando en un incremento de volumen de 5,32 veces respecto al baseline con sólo un descenso en la precisión del 3,5 %. A partir de este punto, la curva empieza a caer considerablemente, por lo que puede no ser conveniente incrementar el valor de k. Por supuesto, esta conclusión no es aplicable a todos los casos: el valor de k óptimo depende de la naturaleza de la temática elegida y de las necesidades de la aplicación en concreto.

2.6.2. Filtrando usuarios inherentemente ruidosos

Como se ha mencionado con anterioridad, algunos de los usuarios incluidos contribuyen enormemente a la pérdida de precisión observada, induciendo una considerable cantidad de ruido. Estos usuarios son, con frecuencia, personas muy famosas y existen dos buenos ejemplos en nuestro dataset: El piloto español de Fórmula Uno Fernando Alonso (`@alo_oficial`) y el cantante y compositor español Alejandro Sanz (`@alejandrosanz`).

Estos usuarios fueron inicialmente seleccionados por el método porque explícitamente animaron a la selección española de fútbol, pero han permanecido en la topología debido a su alta relevancía dentro de Twitter. Alrededor del 44,62 % de los mensajes obtenidos relacionados con `@alejandrosanz` guardan relación con el tema explorado, siendo el resto una considerable cantidad de ruido. El caso de los mensajes relacionados con el usuario `@alo_oficial` es significativamente peor, pues solo el 15,80 % de los mensajes son relevantes, siendo la mayoría de sus mensajes una fuente de ruido.

Si filtramos sólo estos dos usuarios ruidosos del dataset, podríamos obtener un aumento en la precisión muy significativo a expensas de una ligera pérdida

de volumen. En efecto, filtrando estos usuarios ruidosos, el método propuesto es capaz de incrementar la precisión por encima del 91 % para $k = 20$ y por encima del 96 % usando el propuesto $k = 8$.

La tabla 2.6.2 contiene una comparativa mostrando las diferencias de rendimiento entre la versión del método con y sin filtrado adicional.

Método	Precisión	D. Recall	Prec. diff	D. Recall diff
$k = 8$	0,9343	0,7478	-	-
$k = 10$	0,9274	0,8034	-	-
$k = 20$	0,8714	1,0000	-	-
$k = 8$ c/filtrado	0,9604	0,8298	+2,61 %	−1,74 %
$k = 10$ c/filtrado	0,9521	0,8609	+2,48 %	−1,77 %
$k = 20$ c/filtrado	0,9185	0,9488	+4,72 %	−1,88 %

Cuadro 2.4: Comparativa de resultados de rendimiento del método incluyendo el filtrado de usuarios ruidosos

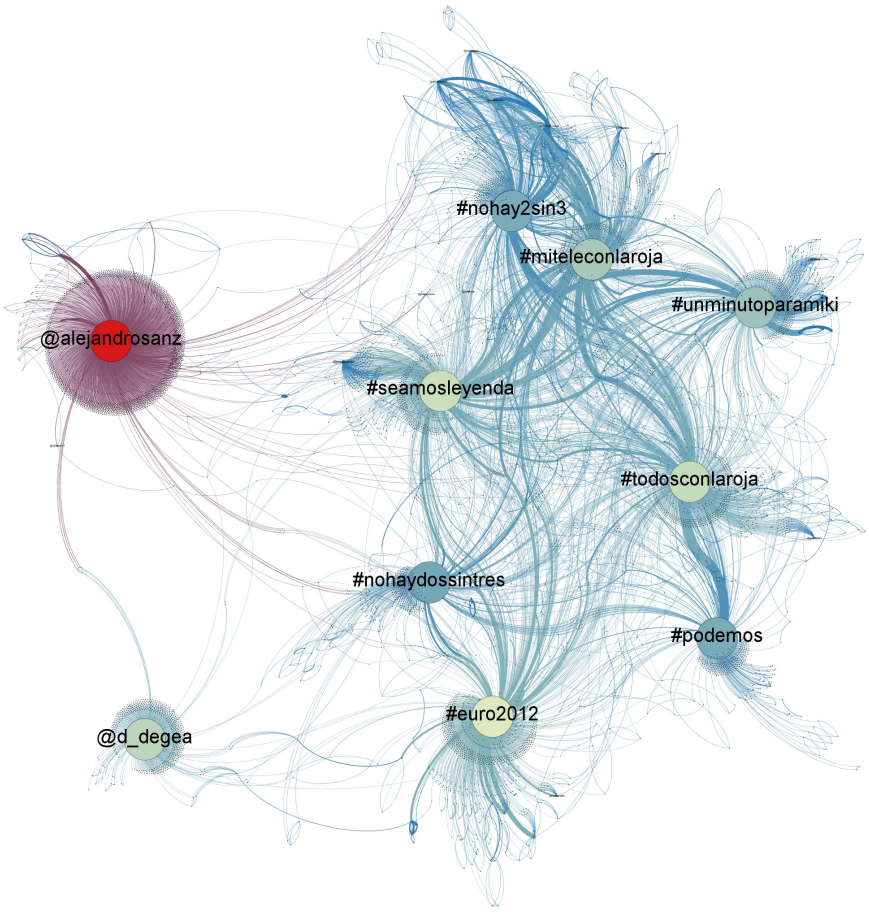


Figura 2.12: Grafo de relevancia correspondiente al partido entre España y Portugal durante el evento Euro2012.

La figura 2.12 muestra el grafo de relevancia correspondiente al partido entre España y Portugal durante el evento Euro2012, siendo visualmente sencillo detectar las entidades mas relevantes en el grafo y el grado de inteconexión existente. Debido a que el término central usado (`#vamosespaña`) es demasiado relevante y está extremadamente conectado con el resto de elementos, por motivos de claridad, se ha omitido del grafo.

Se observa que el usuario `@alejandrosanz` está presente en el grafo como elemento relevante aunque exhibe el comportamiento de un *"exterior hub"* o *"núcleo periférico"*; posee una gran comunidad pero está sufre de una vaga conexión con el resto de elementos del grafo, indicando una baja relación con el resto de términos y de la temática en general. Esto se correlaciona con el ruido previamente asociado a este usuario y es bastante intuitivo que eliminar este tipo de "núcleos periféricos" del grafo es una buena aproximación para una etapa de filtrado.

2.7. Conclusiones y trabajo futuro

En este capítulo, se ha provisto de una formalización para una tarea, hasta el momento no muy bien definida, que es de importancia para hoy en día: la tarea de la recuperación temática de tweets. A partir de esta definición, se idea y propone un método general que aborda la tarea, haciendo hincapié en el análisis estructural del grafo subyacente, apostando por una aproximación topológica en lugar de una puramente textual.

La evaluación muestra la efectividad del método propuesto, observando que al elegir un buen punto de eficiencia paramétrica ($k = 8$ en este caso) se obtiene un aumento considerable en el volumen del dataset recuperado sin incurrir en mucho ruido; se obtiene un dataset 5,32 veces mas grande con una pérdida de precisión del $3,5\,\%$. Además, se explora la idea de filtrar usuarios ruidosos a posteriori y se observa que, simplemente al eliminar dos usuarios particularmente famosos y ruidosos, se obtiene un incremento considerable en la precisión.

Como trabajo futuro, una línea de mejora interesante sería la consideración del uso de otros tipos de relaciones adicionales para la fase de construcción del grafo. Otra línea de mejora sería la de idear ciertas modificaciones sobre el algoritmo de relevancia *PageRank* que pueden ser útiles para el cálculo de relevancia dentro de este contexto. Otra vía de trabajo sería la de abordar la subtarea de filtrado de forma explícita, probablemente realizando análisis sobre el contenido como complemento al análisis estructural actual, aunque el proceso de analizar el texto de un tweet es una tarea por sí misma.

Capítulo 3

Normalización de tweets

Como ya sabemos, Twitter es una red social de éxito generalizado, donde millones de personas expresan ideas continuamente sobre cualquier tema, siendo una gran fuente de información. Sin embargo, la mayoría de los tweets suelen estar redactados con prisa, sin revisión y con un alto grado de abreviación, convirtiéndolos en textos nada apropiados para el procesamiento de lenguaje natural tradicional.

En este capítulo se aborda este fenómeno de forma directa, generando una caracterización de los problemas textuales encontrados y proponiendo un sistema para mejorar la calidad textual de estos tweets. Este sistema propuesto es de naturaleza extensible y modular, siendo fácil de ampliar y requiriendo poco esfuerzo manual tanto para configurarlo inicialmente como para adaptarlo a otros contextos.

3.1. Introducción

Uno de los desafíos mas importantes que afrontamos hoy en día consiste en determinar cómo procesar y analizar la enorme cantidad de información que se puede extraer de Internet, especialmente de sitios que son redes sociales como Twitter, donde millones de personas expresan ideas y opiniones diariamente sobre una enorme variedad de temas distintos.

Los textos escritos en Twitter, se caracterizan por tener una corta longitud (140 caracteres como máximo), siendo textos muy pequeños en comparación con el tamaño de los textos en géneros más tradicionales. Además, la mayor parte de estos tweets se escriben desde dispositivos móviles tales como *smartphones* o *tablets*, siendo escritos con prisa (incluso ciertos usuarios escriben casi en tiempo real, como si se tratara de un sistema de mensajería instanánea) y sin ningún tipo de revisión antes del envío, claramente favoreciendo la velocidad a expensas de la calidad y exactitud de la redacción.

Consecuentemente, los usuarios de este tipo de redes han desarrollado una nueva forma de expresión que incluye el uso generalizado de abreviaturas similares a las usadas en los SMS, variantes léxicas, repetición o ausencia intencional de caracteres y uso de emoticonos, amén de otras técnicas.

Pero los fenómenos mencionados anteriormente provocan una serie de dificultades inherentes al procesado de estos textos usando análisis NLP tradicional.

Las herramientas NLP actuales suelen tener problemas a la hora de procesar y entender estos textos caracterizados por ser cortos y con mucho ruido, siendo estos textos originalmente poco apropiados para tareas como *Opinion Mining (Minería de opiniones)*, *Topic Modelling (Modelado de temática)* o cualquier otra tarea de caracterización.

En términos generales, cualquier tarea NLP que utilice tweets como datos de entrada se benificiaría enormemente de un proceso de normalización, pues las técnicas usadas en NLP suelen ser bastante sensibles a la calidad y la longitud de los textos de entrada. *Text Summarization* (Resumen automático de textos;Jabeen et al. (2013)) y *Ontology-based Sentiment Analysis* (Análisis de subjetividad basado en ontologías;(Kontopoulos et al., 2013)) son ejemplos de tareas que confían en la normalización de tweets para su correcto funcionamiento.

En este capítulo, en lugar de enfocarnos en una técnica específica para solventar la tarea de normalización, se propone un sistema totalmente modular que se fundamenta en la combinación de varios módulos expertos independientes. En lugar de tener módulos para abordar categorías de error amplias, cada uno de los módulos está diseñado para abordar un fenómeno de error frecuente pero muy específico, incrementando inherentemente la precisión del módulo y los costes de diseño e implementación el mismo. En esencia, el sistema se comporta como un "grupo experto": cuando se encuentra un término *OOV* (*Out of Vocabulary* o fuera de vocabulario), uno o más módulos pueden proponer una corrección para ese término, realizándose un ranking sobre estas proposiciones y eligiendo la mejor. De esta manera, en lugar de confiar en una sola técnica multipropósito, el sistema permite que cada módulo se implemente usando las técnicas o aproximaciones que se desee.

Este tipo de aproximación tiene varias ventajas. La ventaja más clara es su fácil expansión ante un nuevo tipo de fenómeno de error no visto anteriormente mediante la adición de nuevos módulos específicos que aborden este fenómeno. Otra ventaja es la robustez del sistema ante fenómenos de error difíciles y/o ambiguos gracias al sistema de ranking. Además de lo dicho, los costes de diseño y construcción son reducidos en comparación con otras aproximaciones mientras que los resultados experimentales muestran muy buen rendimiento.

En la sección 3.2 (*Trabajos relacionados*) se realiza una revisión del estado del arte actual relacionado con la temática del capítulo y se analizan algunos trabajos de mayor interés, describiendo la arquitectura y el funcionamiento de cada sistema de normalización propuesto. En la sección 3.3 (*Caracterización del problema*), se realiza una caracterización del problema a partir de la anotación de una muestra estadísticamente significativa de un corpus compuesto por 3.1 millones de tweets. El resultado de esta caracterización es una distribución estadística de los diferentes fenómenos de error que pueden encontrarse en Twitter. En la sección 3.4 (*Arquitectura del sistema propuesto*), se describe la arquitectura altamente modular del sistema propuesto, la cual se divide en diferentes etapas: *preprocesado*, *detección*, *generación de candidatos* y *selección de candidatos*. En la sección 3.5 (*Recursos utilizados*), se identifican varias grandes categorías conceptuales para los términos y se generan los recursos léxicos modularmente teniendo en cuenta estas categorías. Además, se describe individualmente el proceso específico y el coste asociado a la generación de cada léxico. En la sección 3.6 (*Evaluación del sistema*), se realiza una evaluación del sistema desde diferentes perspectivas y se proponen varias métricas para realizar dicho

proceso de evaluación. No sólo se mide el rendimiento del sistema completo, sino que también se mide el rendimiento con diferentes módulos activados (cada uno abordando diferentes fenómenos de error). Además, se proporciona un análisis de rendimiento respecto a los diferentes fenómenos de error y un análisis sobre la etapa de selección de candidatos, sirviendo este análisis como base para una extensión cuyo objetivo es el de mejorar y ajustar dicho proceso de selección de candidatos. Finalmente, en la sección 3.7 (*Conclusiones y trabajo futuro*), se hace un resumen de los esfuerzos mostrados en este capítulo, se revisan las aportaciones principales del capítulo y se proponen diversas opciones para mejorar el sistema y varias direcciones para ampliar la investigación.

3.2. Trabajos relacionados

Aunque en este capítulo se abordan textos escritos en español, la mayoría de los trabajos existentes que abordan el problema de la normalización léxica tratan con SMS o textos provenientes de redes sociales escritos principalmente en inglés. A pesar de que el español y el inglés son lenguajes diferentes, muchas de las ideas y procesos utilizados en las aproximaciones de normalización léxica dirigidas al inglés pueden ser adaptados para textos en español.

En la obra Eisenstein (2013) se presenta un estudio sobre el mal uso del lenguaje en Internet. Este estudio revisa y critica diferentes tipos de aproximaciones NLP que tratan con este problema, dividiéndolas en dos categorías inicialmente disjuntas: *normalización* y *adaptación al dominio*. El objetivo de las obras de *normalización* radica en mejorar la calidad de los textos, convirtiendo los términos OOV en palabras, expresiones o locuciones válidas. Las obras enfocadas a la adaptación del dominio se centran en adaptar herramientas NLP existentes para que puedan procesar este tipo de lenguaje *"mal formado"* o *"ruidoso"*. Las aproximaciones para normalizar texto "ruidoso" suelen utilizar técnicas basadas en reglas (Sidarenka et al., 2013), modelos estadísticos de lenguaje (Yang and Eisenstein, 2013) o una mezcla de ambos (Costa-Jussa and Banchs, 2013).

Para resolver el problema de las abreviaturas SMS, la obra Pennell and Liu (2011) propone un método bifase. La primera fase consiste en un modelo de traducción automática que, en lugar de ser un sistema a nivel de palabra, es un sistema nivel de caracteres que aprende asociaciones entre dichos caracteres (letras y símbolos, tanto elementos sueltos como en grupo). La segunda fase está diseñada para añadir información del contexto con el fin de refinar la normalización de las abreviaturas, usando un modelo de lenguaje a nivel de palabra (similar a la idea de modelo sensible al ruido presentada en Shannon (1948)).

En Han and Baldwin (2011) se lleva a cabo un estudio sobre los términos OOV que se dan en la red Twitter, analizando los usos poco ortodoxos del lenguaje que se dan dentro de la red social. Teniendo en cuenta las observaciones de este estudio, se propone una técnica no supervisada enfocada en la normalización de términos OOV compuestos por una sola palabra, teniendo en cuenta variaciones morfofonémicas de palabras y el contexto en el que estas variaciones ocurren. En Han et al. (2012), los mismos autores abordan el mismo problema de normalización mediante el uso de un léxico generado a partir de pares de términos OOV y *IV* (*In vocabulary* o dentro de vocabulario), utilizando un algoritmo de similaridad morfofonémica entre pares para generar un ranking. En la obra Han et al. (2013) se realiza una extensión a al trabajo presentado en Han

and Baldwin (2011) con una explicación más detallada y una experimentación más completa.

Aunque la mayoría de los trabajos en el ámbito de la normalización están desarrollados para el inglés, existen varios estudios interesantes para el español. En los siguientes apartados analizaremos algunos de estos trabajos.

3.2.1. Transductores de estados finitos

Los *Transductores de Estados Finitos* o *Finite State Transducers (FST)* son unos formalismos análogos a los *Autómatas Finitos* que en lugar de trabajar sobre una cinta de entrada y devolver un estado de aceptación o rechazo, producen una salida en una cinta de salida adicional. La figura 3.1 muestra un ejemplo simple de un FST generado para transformar las palabras *data* y *dew* en sus representación de pronunciación, mostrando que *data* tiene distintas variantes de pronunciación.

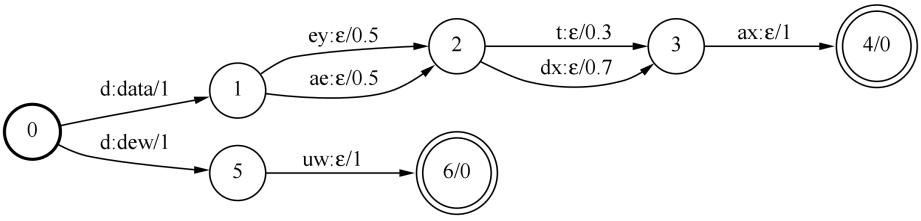


Figura 3.1: Ejemplo de un FST de pronunciación sobre las palabras inglesas *data* y *dew*

Aunque los fundamentos teóricos de estos autómatas son bien conocidos (Berstel and Boasson, 1979; Berstel and Reutenauer, 1988), hasta que no se realizaron avances significativos sobre los FST con transciones ponderadas (ver Mohri (2009)) no han surgido herramientas para la creación, optimización y minimización de dichos FSTs.

La aparición de toolkits que permiten trabajar con FSTs de forma eficiente ha logrado un incremento considerable en la popularidad de estos formalismos, utilizándose con éxito en tareas de NLP como el reconocimiento y síntesis del habla, reconocimiento óptico de carácteres y traducción automática.

En la obra Porta and Sancho (2013), se propone una interesante aproximación para la normalizacición de textos muy fundamentada en la utilización de FSTs. Incialmente, los autores realizan un breve análisis sobre los diferentes fenómenos de error comunes que son responsables de la generación de OOVs.

Una vez realizado este análisis, generan una serie de reglas de transformación para cada uno de los siguientes fenómenos: *logogramas y pictogramas*, *acrónimos y omisión de letras*, *variantes reconocidas*, *variantes fonéticas*, *yuxtaposiciones*, *ausencia de acentuación* y *eliminación de duplicados*. Además, se genera un FST general de edición basado en la distancia Levenshtein, haciendo uso de diversas fuentes léxicas como el Diccionario de la Real Academia Española (DRAE) y un léxico secundario compuesto por las 100k palabras mas frecuentes del inglés según el British National Corpus, pues es frecuente que los usuarios españoles hagan uso de palabras comunes procedentes del inglés dentro de los textos escritos.

Dado que los transductores construidos a partir de las reglas generan candidatos de corrección a nivel léxico, no son capaces de tener en cuenta el contexto dentro de la oración, siendo, a priori, cualquier candidato generado válido desde este punto de vista. Por ello, los autores proponen el uso de un modelo de lenguaje para determinar el grafo de palabras correcto respecto a las múltiples variantes posibles propuestas. Haciendo uso de un corpus de páginas web *Wacky*, se genera un modelo de lenguaje de trigramas con *back-off* para dotar de cierto conocimiento estádistico a la hora de seleccionar un candidato de corrección dentro de su contexto.

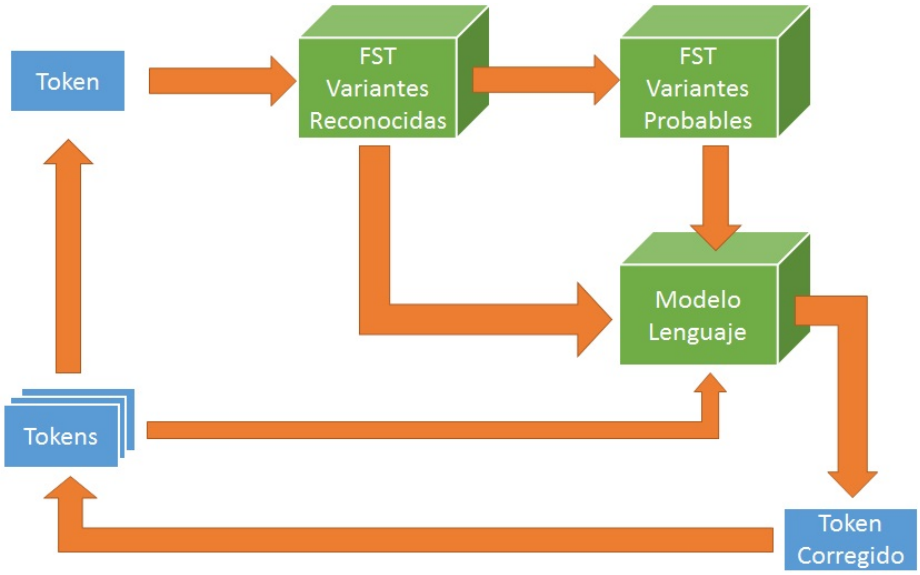


Figura 3.2: Funcionamiento del sistema de normalización propuesto en Porta and Sancho (2013)

La arquitectura del sistema propuesto se basa en combinar los FSTs en dos grandes FST (uno de variantes reconocidas y otro de variantes posibles), ponerlos en dos etapas secuenciales y utilizar el modelo de lenguaje para validar los candidatos dentro del contexto. Los tokens que no son analizados por el primer FST son revisados por el segundo más general. Todo el tokenizado se ha hecho utilizando Freeling. La figura 3.2 muestra a la arquitectura general del sistema de normalización propuesto y su funcionamiento.

3.2.2. Enfoques de carácter estrictamente léxico

En la obra Gamallo et al. (2013b) se propone una aproximación cuyo enfoque es puramente léxico, compuesto principalmente por varios diccionarios de diferente naturaleza y un conjunto de reglas de transformación simples.

Los autores identifican tres tipos de errores que son muy frequentes en el lenguaje español: *capitalización incorrecta*, *repetición intencionada de caracteres* y *errores de confusión y acentuación*. Para estos tres tipos de errores, han diseñado manualmente un conjunto específico de reglas de transformación.

Para el resto de los fenómenos de error posibles, los cuales corresponden a un conjunto hetereogéneo de problemas, usan una estrategia genérica: búsqueda en recursos léxicos y selección del mejor candidato de corrección. En este caso, los recursos léxicos consisten en tres diccionarios construidos de distinta forma y con un objetivo concreto:

- **Diccionario de normalización (ND)**: Este diccionario está compuesto por 824 variantes léxicas erróneas con sus respectivas formas aceptadas. Cada entrada de este diccionario consiste en un par "variante-corrección" donde se indica explícitamente la corrección para cada variante. Este diccionario incluye tanto palabras mal escritas, como emoticonos extraídos de la Wikipedia y abreviaturas aceptadas en la RAE.

- **Diccionario estándar (SD)**: Este diccionario está constituido por todas las formas automáticamente generadas de los lemas encontrados en el DRAE. Para las formas verbales, los autores han utilizado el conjugador Cilenis (Gamallo et al., 2013a), mientras que para los sustantivos y adjetivos han usado un conjunto de reglas morfológicas específicas. El conjunto final de formas generadas es cuenta con un total 778.149 formas.

- **Diccionario de nombres propios (PND)**: Este diccionario de nombres propios y términos de dominio específico ha sido automáticamente construido mediante a través de un recurso enciclopédico, en este caso, la Wikipedia. Usando un volcado de la Wikipedia, identifican los nombres de los artículos que corresponden a personas, lugares y organizaciones, aplicando un proceso de tokenización y descartando las variantes que se encuentran en el diccionario estándar (SD): El total de entradas de este diccionario es de 107.980 unigramas.

Además de los diccionarios, se genera un modelo de lenguaje basado en *Part-of-speech* (clases de palabras a nivel morfosintáctico; abreviado como *POS*) para la etapa de selección de candidatos. Tanto el proceso de tokenización como el sistema de etiquetado POS han sido realizados mediante FreeLing.

Usando los recursos y las reglas descritas anteriormente, el método de normalización que proponen funciona de la siguiente manera para cada token:

1. Se determina si el token es una OOV mediante una búsqueda en los diccionarios. Si el token no se encuentra en ninguno de estos recursos, el token es considerado OOV.

2. Se generan las variantes de corrección primarias usando las reglas de transformación. Puede generarse más de una variante primaria, aunque se pondera positivamente a aquella que se encuentre en alguno de los recursos léxicos existentes.

3. Si no se genera ninguna variante de corrección primaria para el token OOV, se genera una larga lista de variantes secundarias generadas que se encuentren a distancia edicion 1 respecto al OOV original, usando para ello, los recursos léxicos.

4. Se selecciona el mejor candidato usando un modelo del lenguaje respecto a una ventana de tamaño 4 (2 tokens a la izquierda y a la derecha del

token OOV). En cualquier caso, si no se ha generado ninguna variante de corrección válida (ya sea primaria o secundaria), el token OOV se considera correcto, dejándose sin modificar.

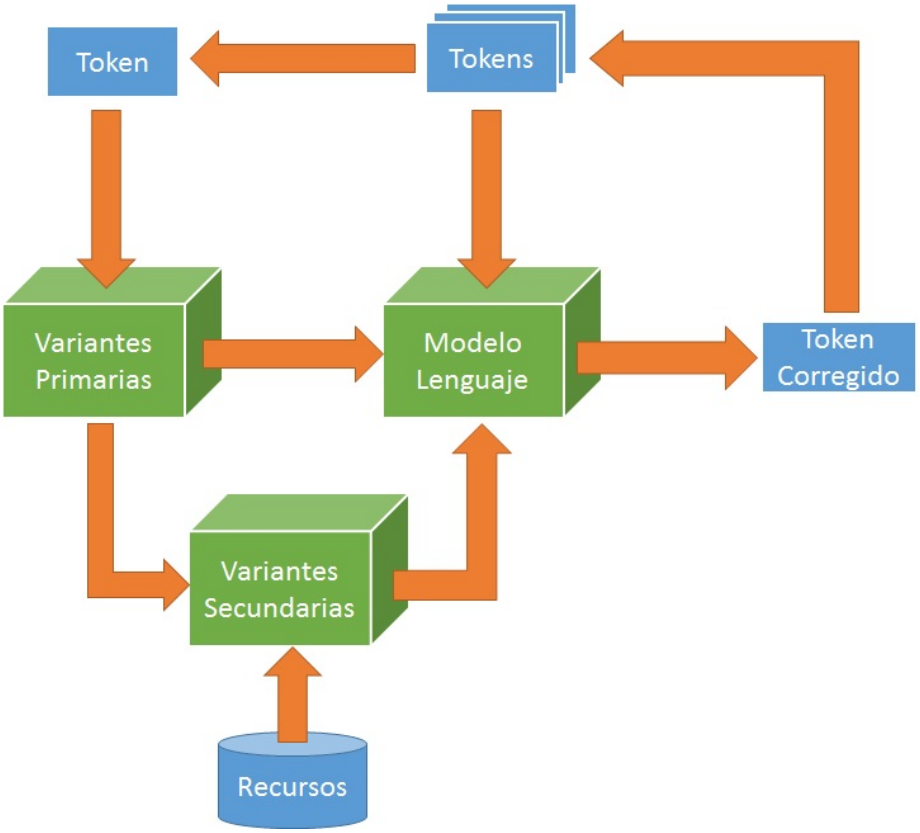


Figura 3.3: Funcionamiento del sistema de normalización propuesto en Gamallo et al. (2013b)

La figura 3.2 muestra a la arquitectura general del sistema de normalización propuesto, los componentes que interactúan y su funcionamiento general. En el proceso de experimentación incluido en su trabajo, los autores llegan a la conclusión de que uno de los puntos claves de su sistema consiste en la separación de la generación de variantes en las dos etapas en cascada, pues si se combinan ambas etapas en una sola etapa no se prima a los candidatos generados para solventar los errores comunes y el rendimiento baja considerablemente.

3.2.3. Sistemas modulares y multicomponente

En Ageno et al. (2013) se propone una aproximación basada en diferentes módulos de procesamiento que actúan independientemente para generar candidatos para cada palabra desconocida. Al igual que en otras aproximaciones, los autores realizan un análisis superficial sobre los fenómenos de error encontrados y diseñan un sistema modular en el cual cada módulo está enfocado a resolver una problemática en particular. Existen tres grandes grupos de módulos:

- **Módulos de expresiones regulares**: Estos módulos son colecciones independientes de expresiones regulares que se encargan de proponer candidatos de corrección a patrones muy recurrentes como emoticonos, algunas onomatopeyas, ciertas abreviaturas y/o errores muy frecuentes. Cada uno de estos módulos propone una única solución para cada caso.

- **Módulos unipalabra**: Cada módulo perteneciente a esta clase usa un recurso léxico propio de palabras simples y un conjunto de medidas de distancias de edicion con el fin de encontrar candidatos similares al token OOV. Las distancias usadas son la distancia de edición tradicional, de similitud fonológica y de distancia física en un teclado.

- **Módulos multipalabra**: Estos módulos parten de la información proporcionada por el resto de módulos y toman una decisión basada en el contexto de la OOV. En concreto, los autores utilizan tres módulos: un módulo con un diccionario multipalabra, un módulo con un etiquetador POS y un módulo que analiza las palabras concatenadas. Estos módulos actúan secuencialmente, cada uno utilizando información adicional para complementar, filtrar o descartar candidatos propuestos por el resto de módulos.

Los autores han utilizado FreeLing como framework para implementar los módulos y como herramienta de preprocesado y tokenizado. Además, utilizan el toolkit FOMA (Hulden, 2009) para realizar búsquedas aproximadas eficientes sobre los diferentes recursos utilizados por cada módulo.

Cada módulo posee un conjunto de recursos léxicos específico. Los módulos de expresiones regulares utilizan un lista de acrónimos conocidos y abreviaturas comunmente usadas en tweets, una lista de emoticonos y una lista de onomatopeyas extraidas del DRAE.

Los módulos unipalabra utilizan un diccionario propio del lenguaje español, un diccionario propio del lenguaje inglés, un diccionario de variantes morfológicas generadas a partir del diccionario español, una lista de nombres propios que incluye algunos diminutivos y un diccionario de entidades unipalabra tales como localizaciones, organizaciones, personas, canales de televisión, programas, productos y otros medios de comunicación.

Para los módulos multipalabra, sólo se ha generado un diccionario de entidades multipalabra, el cual es de naturaleza similar al diccionario de entidades unipalabra pero con la diferencia que los términos de las entidades son exclusivamente multipalabra.

El método de normalización propuesto consiste en obtener una primera tanda de candidatos mediante los módulos de expresiones regulares y unipalabra, refinar dicha tanda con los módulos multipalabra y aplicar un esquema de votación simple para seleccionar el mejor candidato. La figura 3.4 muestra a la arquitectura general del sistema de normalización propuesto y su funcionamiento.

3.3. Caracterización del problema

Para caracterizar el problema y evaluar el sistema de normalización propuesto en este capítulo se ha utilizado un dataset compuesto por 3.1 millones de

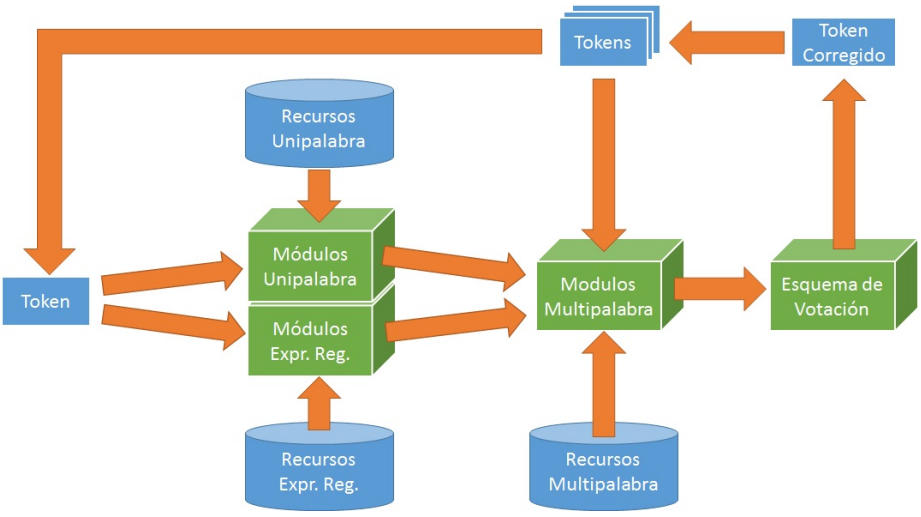


Figura 3.4: Funcionamiento del sistema de normalización propuesto en Ageno et al. (2013)

tweets escritos en español que están altamente relacionados con el *Campeonato Europeo de Fútbol de la UEFA 2012*, coloquialmente llamado *Eurocopa 2012* o *Euro2012*. El dataset ha sido generado usando el proceso dinámico de recuperación de tweets descrito inicialmente en la obra Cotelo et al. (2014) y revisado en el capítulo 2. Este dataset es muy útil debido a que contiene una gran cantidad de tweets en español que exhiben los típicos problemas de calidad y redacción que hemos mencionado anteriormente. Como dato adicional, este dataset ya había sido recolectado y analizado para la evaluación del método de recuperación descrito en el capítulo 2, por lo que no había necesidad de realizar otro proceso de recuperación si los datos pueden ser reutilizados con otro fin.

Como paso inicial al esfuerzo realizado en este capítulo, se ha llevado a cabo un análisis sobre la distribución de términos de nuestro dataset, utilizando varios vocabularios como fuente de conocimiento existente para la determinación de términos válidos. La tabla 3.1 y la figura 3.5 representan los resultados de dicho análisis, mostrando que sólo 68,76 % de los términos encontrados fueron reconocidos dentro del vocabulario común del lenguaje español, denominando esta proporción como *Language IV* (*In-Vocabulary* o dentro del vocabulario del Lenguaje).

Una parte significativa de los términos no pertenecientes al vocabulario del lenguaje español fueron OOV mientras que el resto de términos detectados pertenecían a un vocabulario de propósito especifico (*Specific IV*) o algo distinto a lo que comúnmente consideramos una palabra (fechas, hashtags, menciones, numerales, etc).

Antes de realizar ningún proceso de normalización, se procedió a realizar una caracterización de los fenómenos de errores existentes causantes de los términos OOV. Como no es viable realizar este tipo de caracterización sobre los 3.1 millones del dataset original, se obtuvo una muestra estadísticamente significativa del dataset ($\alpha = 0{,}05$, error $= 1\,\%$) compuesta por tweets que tuvieran al menos un término OOV. Cada uno de estos tweets fue analizado manualmente, siendo

Tipo		Detección	Descripción
Language IV		68,76 %	Términos validos encontrados en el vocabulario común del lenguaje español
Rest		31,24 %	Términos no encontrados en el vocabulario común del lenguaje español
	OOV	28,82 %	Términos no encontrados en ningún otro recurso
	Specific IV	15,82 %	Términos validos encontrados en algún recurso de propósito específico (género, dominio, etc. . .)
	REGEX	53,33 %	Términos válidos reconocidos mediante expresiones regulares (menciones, hashtags, fechas, etc. . .)

Cuadro 3.1: Distribución de términos detectados en el dataset

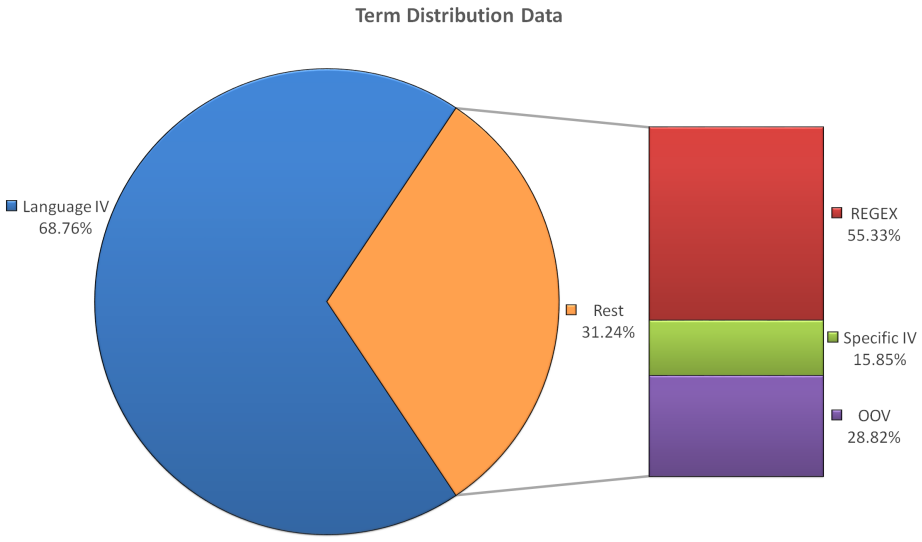


Figura 3.5: Distribución de términos detectados en el dataset

cada término OOV etiquetado indicando el fenómeno de error asociado y su forma correcta. Se dieron algunos casos de fenómenos múltiples en una misma OOV. Esta muestra manualmente anotada se usa como dataset de evaluación durante todo el proceso de experimentación.

La caracterización de los diferentes fenómenos de error da lugar a las siguientes categorías:

- *Errores comunes de ortografía* (`ORT`): segmentación de palabras incorrecta, ausencia o mal uso de acentuación, uso incorrecto de las mayúsculas y faltas de ortografía más comunes.

- *Lenguaje móvil* (`TXT`): acrónimos *adhoc*, abreviaturas, omisión de letras, omisión de elementos morfosintácticos, uso de logogramas y pictogramas.
- *Confusión homofónica* (`HOMO`): cambios fonológicos comunes y otras variantes de escritura de origen fonológico.
- *Onomatopeyas no identificadas* (`ONO`): onomatopeyas muy poco frecuentes o variantes de onomatopeyas que no son fácilmente reconocibles.
- *Repetición de caracteres* (`REP`): repetición innecesaria de caracteres, utilizada normalmente para enfatizar.
- *Arte ASCII* (`ASC`): Uso especial y creativo de los caracteres disponibles para reflejar situaciones, emociones, estados de ánimo o para expresar cualquier otro tipo de comunicación no verbal e incluso con fines puramente estéticos.
- *Flexiones y variantes libres*: Variantes de escritura aceptadas normalmente vinculadas a una región o colectivo específico. No es un fenómeno de error en el sentido estricto de la palabra, pero a efectos del proceso de detección, son considerados *a priori* como términos OOV.
- *Términos foráneos* (`FT`): Términos de origen extranjero, provenientes de otros idiomas, que no tienen por qué ser aceptados en el contexto del lenguaje español.

La tabla 3.2 ilustra esta caracterización y proporciona varios ejemplos para cada fenómeno. Se observa que la mayoría de los errores que se pueden encontrar encajan en alguna de las grandes categorías de error descritas anteriormente, estando claramente relacionados con el estilo de escritura rápido e informal asociado a Twitter, comúnmente propiciado por la escritura usando un dispositivo móvil.

3.4. Arquitectura del sistema propuesto

Las obras de normalización existentes están mayormente diseñadas para abordar un caso en concreto, siendo muy costoso adaptarlas para otro dominio o lenguaje. La aproximación que se propone y evalúa en este capítulo toma un camino diferente: un sistema que está específicamente enfocado en ser flexible, modular, tolerante ante problemas difíciles y con poca carga de trabajo manual, comparado con aproximaciones más tradicionales. En términos generales, todos los módulos del sistema hacen uso de uno o más elementos generales que componen el sistema:

- **Recursos y fuentes de conocimiento**: léxicos, corpus y cualquier otro tipo de recurso lingüístico, incluyendo recursos que contienen conocimiento específico sobre el medio usado o el dominio abordado. Los recursos utilizados por el sistema se describen en la sección 3.5.
- **Reglas**: reglas para el manejo automático de fenómenos comúnmente encontrados en Twitter, como repetición excesiva de caracteres, acrónimos o errores de carácter homofónico.

Fenómeno	Prop.	Ejemplos
Errores comunes de ortografía	27,51 %	sacalo → sácalo, trapirar → transpirar, ...
Lenguaje móvil	07,92 %	x2 → por dos, q → que, aro → claro, ...
Confusión homofónica	08,52 %	kasa → casa, caxo → cacho, ...
Onomatopeyas no identificadas	05,96 %	jajajajjajja → ja, jum → um, ...
Repetición de caracteres	15,25 %	siiiiiiiii → si, quiiiiieeeeroooo → quiero, ...
Arte ASCII	13,80 %	♠ ❤ oO._.Oo ...
Variantes libres	06,90 %	gatino, besote, bonico, ...
Otros errores	06,73 %	htt, asdfasdfasdf, ...
Términos foráneos	04,00 %	flow, ftw, great, lol, ...
Fenómeno múltiple	03,41 %	diass → días, artooo → rato, ...

Cuadro 3.2: Caracterización de los fenómenos de error encontrados en el dataset

- **Análisis léxico**: la distancia léxica tradicional permite a los analizadores abordar errores ortográficos comunes.

El sistema analiza cada término a nivel léxico, determinando si son términos OOV o no, utilizando los recursos y técnicas disponibles. Si el sistema determina que el término en cuestión es OOV, éste genera un conjunto de candidatos de corrección y finalmente selecciona el mejor candidato de corrección para cada caso. La figura 3.6 muestra los componentes, cómo se interconectan y el flujo de procesado del sistema.

Conceptualmente hablando, el sistema se divide en diferentes etapas: *preprocesado*, *detección*, *generación de candidatos* y *selección de candidatos*. Las etapas son abordadas en detalle a lo largo los apartados que vienen a continuación.

3.4.1. Preprocessado y detección

Como ocurre en la mayoría de los sistemas de procesado existentes en el campo del NLP, el primer paso del sistema propuesto consiste en realizar un preprocesado sobre los textos en bruto. El módulo de preprocesado se encarga de realizar el tratamiento inicial típico en todo análisis léxico, generando un flujo de *tokens* para cada tweet, teniendo en cuenta términos especiales como constructos de Twitter (hashtags y nombres de usuario), numerales, fechas, elementos URL, emoticonos y algunos tipos de ordinales. El correcto tratamiento de estos elementos especiales es fundamental para que cualquier sistema de este tipo funcione correctamente.

Además de lo descrito anteriormente, el módulo de preprocesado también se encarga de las tareas de tratamiento y limpieza relacionadas con la codificación de caracteres, corrigiendo caracteres mal codificados, descartando elementos

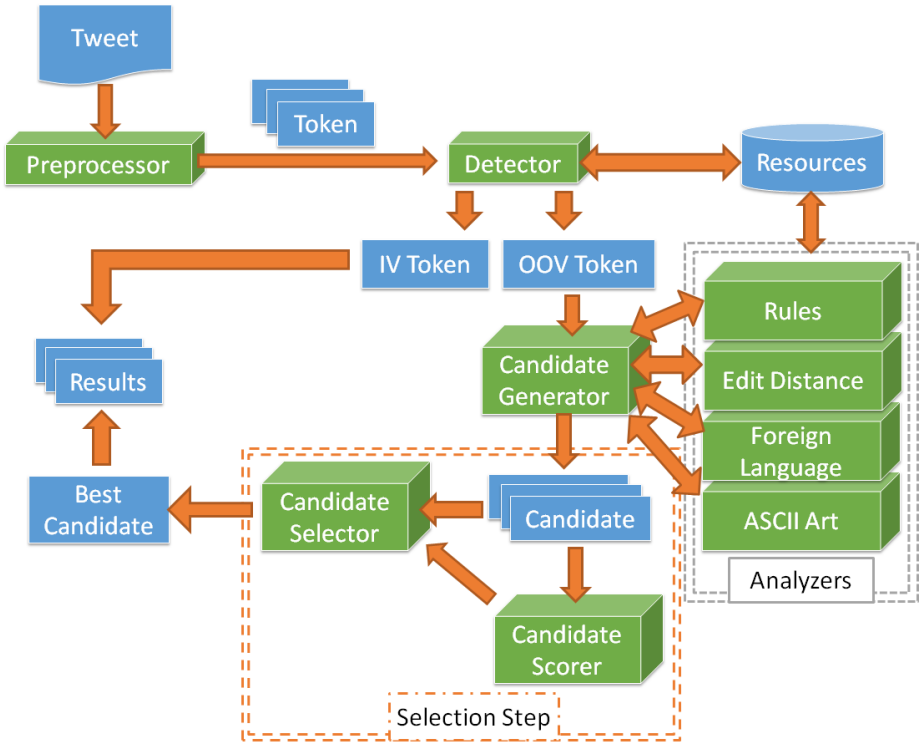


Figura 3.6: Arquitectura del sistema con sus diferentes etapas de procesado

Unicode no estándar y normalizando todos los caracteres a una misma forma canónica. De lo contrario, el resto de las etapas tendrían problemas al realizar los correspondientes análisis a nivel de carácter.

La siguiente etapa consiste en detectar, a grosso modo, la naturaleza de cada token resultante. El módulo de detección intenta determinar si un token es un término OOV o no, comprobando si pertenece a algún recurso externo (o fuente de conocimiento equivalente) o si se asemeja a construcciones conocidas como nombres de usuario, hashtags, fechas, numerales, direcciones URL, etc. Como recursos externos para la detección se usan un conjunto de léxicos, vocabularios y otro tipo de recursos textuales (ver sección 3.5, cada uno proporcionando formas conocidas y aceptadas en el lenguaje español y Twitter, incluyendo emoticonos, locuciones, variantes y coloqualismos. Para la detección de construcciones especiales, se usa una aproximación basada en reglas y expresiones regulares.

Después de esta etapa de detección, cada token detectado como término OOV es llevado a la siguiente etapa de la cadena de procesado mientras que el resto simplemente se marcan con la naturaleza detectada en cuestión y son enviados al final de la cadena.

3.4.2. Generación de candidatos

Después de que un token sea marcado como OOV, se envía al siguiente módulo de la cadena de procesado: el módulo generador de candidatos. Este módulo controla la etapa de generación de candidatos y está enlazado a varios

módulos *analizadores* que son los que realmente realizan el grueso del trabajo de la generación de candidatos.

Dado un token detectado como OOV, el módulo de generación de candidatos ordena a cada analizador a realizar proceso de suposición del error, dando cada módulo su respuesta en forma de candidatos propuestos para corregir dicho token. El proceso subyacente específico varía dependiendo del analizador, cada uno especializándose en algún fenómeno de error en particular y haciendo uso de cualquier tipo de recurso externo para realizar esta labor. Cada módulo analizador provee algún tipo de mecanismo para generar puntuaciones básicas sobre los candidatos, indicando el grado de confianza otorgado por el analizador para cada corrección propuesta.

La arquitectura modular permite tener un número arbitrario de analizadores distintos, otorgando total flexibilidad a la hora de diseñar el sistema en cuestión para cualquier contexto. Para esta obra en cuestión, se han implementado los siguientes módulos analizadores:

El módulo *Edit distance* (distancia de edición) tiene un funcionamiento muy similar al esquema de sugerencias basado en distancias de edición, una técnica muy utilizada en la mayoría de los correctores ortográficos, aunque una de las mayores diferencias es que está diseñado para tener en cuenta múltiples léxicos y vocabularios en lugar de uno solo. Este módulo es el más general del sistema, abordando los fenómenos de error más comunes, siendo el analizador que trata los *errores ortográficos comunes (ORT)* mucho mejor que ningún otro módulo implementado. La distancia entre cadenas de caracteres utilizada es la distancia *Damerau-Levenshtein* (Damerau, 1964; Levenshtein, 1966) la cual, dadas dos cadenas cualesquiera, expresa el número de operaciones léxicas de caracteres (inserción, eliminación, substitución o transposición adyacente) que transformaría una cadena en la otra. Por ejemplo, la distancia Damerau-Levenshtein entre las palabras *puerta* y *perro* es 3: una operación de eliminación y dos operaciones de substitución. Para la implementación del módulo se basa en el uso varios léxicos y *autómatas de Levenshtein* autómata (Schulz and Mihov, 2002). Los valores de confianza asignados a los candidatos generados se encuentran en el intervalo $[1, 3]$, siendo el valor de confianza inversamente proporcional a la distancia entre el candidato propuesto y el OOV; la idea es que un candidato es menos probable cuanto más mayor es su distancia respecto al OOV.

El módulo analizador *Tranformation rules* (reglas de transformación) contiene una colección de reglas creadas manualmente por un experto. Estas reglas tienen como objetivo inyectar directamente conocimiento "humano experto" al sistema y cada una representa un tipo de error *"bien definido"*. Usando estas reglas, el módulo genera un conjunto de candidatos mediante la aplicación de reglas cuyo patrón coincide con el token OOV, realizando una transformación léxica sobre el token original de acuerdo a cada regla coincidente y generando de esta forma un candidato de corrección. Es técnicamente posible generar más de un candidato debido a múltiples coincidencias de patrón, pero la cantidad de casos normalmente se limita a unos pocos. Estas reglas están diseñadas para abordar fenómenos que el módulo de distancia de edición no es capaz de tratar correctamente tales como la *Repetición de caracteres* o el *Lenguaje móvil*. El proceso de creación de reglas se describe en profundidad en la sección 3.5 y la tabla 3.3 muestra algunos ejemplos de regla, usando el lenguaje de expresiones

regulares de `Python`[1]. Este módulo siempre asigna un valor máximo de confianza 3 para los candidatos generados.

Patrón	Tr. Léxica	Ejemplos	Fenómeno
^[ck]n$	con	kn → con	Lenguaje Móvil
		cn → con	
x([aeiouáéíóú])	ch\1	xaval → chaval	Lenguaje Móvil
		coxe → coche	
((\w)(\w))\1+(\2|\3)?	\g<1>	sisisisi → si	Repetición caracteres
		nonono → no	
^t[qk]m+$	te quiero mucho	tkm → te quiero mucho	Lenguaje Móvil
		tqm → te quiero mucho	

Cuadro 3.3: Extracto de las reglas de transformación usadas en el sistema

El módulo de detección *Foreign language* (lenguaje foráneo) es el encargado de identificar si el token OOV en cuestión pertenece a un lenguaje distinto del español y si el módulo detecta que ese token pertenece a otro lenguaje, se marca dicho token como *termino foráneo*. Merece la pena recalcar que el módulo no propone ningún tipo de candidato de corrección; términos que pertenecen a otros lenguajes deben ser marcados como tales pero no corregidos. El analizador utiliza un módulo de detección de lenguaje basado en trigramas implementado en `Python 3` (Phi-Long, 2012) como motor de detección subyacente. Los valores de confianza asignados a los candidatos generados se hayan comprendidos en el intervalo $[1, 3]$ y el valor es directamente proporcional al valor de confianza proporcionado por el estimador del motor de detección subyacente.

El módulo de detección de *ASCII Art* (arte ASCII) intenta identificar si un token OOV es algun tipo de arte ASCII, un emoticono no registrado o una variante de uno ya conocido. Funciona mediante la mezcla de varias expresiones regulares cuidadosamente generadas y una lista de emoticonos comunes. Este módulo se comporta de forma similar al módulo de detección de lenguaje foráneo en el sentido que sólo marca los tokens. Este módulo siempre asigna el valor máximo de confianza 3 para los candidatos generados.

Después de que los candidatos sean propuestos por los analizadores, el módulo de generación de candidatos realiza un paso de filtrado y validación, con la intención de eliminar candidatos incorrectos y/o duplicados, todo ello acorde a un conjunto de reglas de validación y recursos lingüísticos.

Recapitulando, la etapa de generación de candidatos funciona de la siguiente manera:

1. Dado un token OOV, el módulo encargado del proceso general de generación de candidatos envía dicho token a los módulos analizadores.

2. Cada módulo analizador puede o no proponer uno o más candidatos como posibles correcciones para dicho token OOV.

3. El módulo del proceso general de generación de candidatos filtra y valida los candidatos propuestos por los módulos analizadores.

[1] Ver la documentación oficial *The Python 3 Standard Library*, módulo *re*, accesible desde `https://docs.python.org/3/library/re.html`

4. El conjunto de candidatos resultante es enviado a la siguiente etapa de la cadena de procesado.

3.4.3. Selección del candidato y resumen del proceso completo

Esta es la etapa final de la cadena de procesado donde, para cada token OOV, se elige la mejor corrección posible entre los candidatos propuestos.

Para poder elegir el mejor candidato, se asigna un valor numérico a cada usando el módulo *Candidate Scorer* (puntuador de candidatos), generando puntuaciones mediante la normalización de los valores de confianza asignados a los candidatos. Después de esto, el módulo *Candidate Selector* (selector de candidatos) ordena la lista de candidatos por puntuación y elige el mejor, utilizando varios criterios en cascada para resolver empates:

1. *Puntuación*: Se elige el candidato con la mejor puntuación y en caso de empate se recurre al siguiente criterio.

2. *Frecuencia dentro del corpus*: Se elige el candidato cuya frecuencia de aparición dentro del corpus generado a partir del dataset. La mayoría de casos se resuelven en este criterio, pero si existen uno o mas candidatos con la misma frecuencia de aparición, se recurre al último criterio.

3. *Frecuencia del fenómeno de errro asociado*: Se elige el candidato cuyo fenómeno de error asociado sea el más frecuente. En el muy raro caso de que dos términos distintos de igual puntuación, tengan misma frecuencia dentro del corpus y resuelvan el mismo fenómeno, se escoge uno cualquiera.

El flujo de trabajo del sistema completo puede ser resumido tal y como sigue: el sistema genera un flujo de tokens a partir de un tweet usando el módulo de preprocesado. Para cada token, se determina si es un token OOV o no usando el módulo de detección. Si el token es un token IV, no se realiza procesamiento posterior alguno debido a que ya se considera una *forma válida*. De lo contrario, si el token es un token OOV, el modulo de generación de candidatos crea una lista tentativa de candidatos mediante el uso de los analizadores previamente descritos. Como paso final, el módulo selector de candidatos elige el mejor candidato para la corrección del token usando las puntuaciones que le proporciona el módulo puntuador de candidatos.

La tabla 3.4 muestra algunos ejemplos de salida del sistema, mostrando la corrección final propuesta por el sistema para cada uno de los tweets de entrada de ejemplo. Por motivos de claridad, sólo se muestra el texto final resultante, omitiendo toda información respecto a los tokens y a los candidatos.

3.5. Recursos utilizados

Tal y como se ha mencionado anteriormente, para realizar el proceso de evaluación del sistema propuesto se ha considerado el evento *Euro2012* como escenario de normalización. A lo largo de esta sección se describen el proceso y el coste de generación para cada uno de los recursos utilizados por el sistema para abordar el evento en cuestión.

Tweet original	Corrección propuesta
RT @axestrella7: fds estupendo: Partidzo d España con mi amr. tqmmmmmm :)	RT @axestrella7: Fin de semana estupendo: partidazo de España con mi amor. Te quiero mucho :)
iiiiiiiker iiiiiker iiiiiiiker!!!! q crack *.* #VamosEspaña	Íker Íker Íker! Qué crack *.* #VamosEspaña
@SergioRamos ers un crack menudopenalty mas bien tirado demostrando q lo de la otra vez solo fue mala suerte	@SergioRamos eres un crack menudo penalty mas bien tirado demostrando que lo de la otra vez sólo fue mala suerte

Cuadro 3.4: Ejemplos de salida del sistema propuesto

Previo al proceso de generación de recursos, se realizó un análisis sobre los textos para determinar el vocabulario utilizado en los mismos. Los resultados obtenidos inducen a la idea de que el vocabulario total puede dividirse en diferentes categorías conceptuales independientes entre sí y cada una de ellas dedicada a abordar un aspecto diferente del léxico usado en los textos.

Los resultados llevan a identificar tres grandes categorías conceptuales: *Language* (lenguaje), *Genre* (género) and *Domain* (dominio). Los términos que pertenecen a la categoría *Language* (*Lang* de forma abreviada) son aquellos que pertenecen, en un sentido general y amplio, al lenguaje español, no siendo asociados a ningún contexto específico, medio de comunicaciones o dominio en particular. Las formas aceptadas que son usadas con mas frecuencia caen dentro de esta categoría. La categoría *Genre* está compuesta por términos cuyo uso está típicamente confinado al contexto de Twitter e Internet, siendo normalmente terminología específica o expresiones muy utilizadas por los usuarios de estos medios. La categoría *Domain* está compuesta por elementos comunes del domino en cuestión: entidades, términos futbolísticos y otros términos referentes al evento *Euro2012* tales como nombres de jugadores, equipos o estadios donde se celebran los partidos.

Como resultado directo de esta categorización conceptual, se han generado y utilizado varios léxicos especializados para las diferentes etapas de detección y análisis en el sistema propuesto. Teniendo en cuenta las categorías observadas, cada léxico se genera independientemente del resto y enfocándose en cubrir un aspecto específico, mejorando la flexibilidad y la facilidad de adaptación a otros contextos distintos que difieran en género y/o dominio. Todos estos recursos léxicos se componen de texto sin formato y cada línea es una entrada distinta.

Acorde con los experimentos realizados, el sistema puede ser configurado desde cero para un nuevo escenario (combinación de lenguaje, género y dominio) mediante la inversión de aproximadamente 20 horas de esfuerzo manual. Una vez que este trabajo inicial se ha realizado, adaptar el sistema para otro domino distinto (cualquier dominio no relacionado con fútbol y el evento *Euro2012*) sólo requiere de generar el nuevo recurso del dominio en cuestión.

La cantidad específica de tiempo requerida para generar este nuevo domino es muy dependiente del domino en cuestión. En algunos casos, como cabe esperar, este esfuerzo puede reducirse si se adaptan recursos ya existentes o si se hace uso de procedimientos automáticos para acelerar la tarea de construc-

ción del recurso. En términos generales, en el campo del NLP, es normal el uso de estos métodos automáticos para generar recursos, reduciendo enormemente los tiempos requeridos. Como contrapartida, es común el realizar un proceso de validación y refinamiento manual sobre los léxicos generados con técnicas automáticas, pues estas técnicas suelen introducir algo de ruido en el recurso generado.

Léxico	Entradas	Descripción
Lang	1250796	Formas comunes del lenguaje Español. Basado en los diccionarios del `LibreOffice`.
Genre	60	Formas comunes relacionadas con Twitter e Internet. Generado manualmente.
Domain	2710	Términos relacionados con fútbol y el evento Euro2012. Generado manualmente.
Emoticons	320	Emoticonos de uso común. Generado manualmente.

Cuadro 3.5: Léxicos usados por el sistema propuesto

La tabla 3.5 muestra las estadísticas generales de todos los léxicos generados y utilizados. La generación de los léxicos fue realizada con, relativamente, poco esfuerzo y costes asociados. A continuación se detalla los procesos y costes de la generación de léxicos:

- **Lang**: Los diccionarios de español de LibreOffice (y OpenOffice) están en el formato de la herramienta `Hunspell`. Por ello, fue necesario utilizar las herramientas `munch/unmunch` (provistas por el propio paquete Hunspell) para extraer todas las formas, aplicando transformaciones morfológicas durante el todo el proceso. Este proceso automático tardó unos 90 minutos.

- **Genre**: Este léxico de tamaño reducido fue generado a mano y contiene las formas más usadas que están relacionadas con Twitter (incluyendo algunos préstamos lingüísticos y palabras inglesas). Se tardó unos 75 minutos en hacer el léxico.

- **Domain**: Este léxico de tamaño moderado fue generado mediante la agregación de varias listas recuperadas automáticamente de fuentes deportivas especializadas (jugadores, equipos, países y localizaciones) a una lista de entidades y términos futbolísticos comunes. Llevó unas 4 horas realizar los procesos necesarios para generar el léxico: recuperación de las listas, preprocesado necesario de las listas y generación de los términos futbolísticos comunes.

- **Emoticons**: Este léxico de reducido tamaño fue generado a través de varias listas de emoticonos bien conocidos, siendo la más extensa de estas listas la extraida de la Wikipedia[2]. Este proceso requirió de unos 60 minutos, debido a que fue necesario un proceso de revisión manual sobre todos los emoticonos incluidos. Aunque conceptualmente hablando, los emoticonos pertenecen a la categoría de *Genre* (género), se generaron y mantuvieron independientemente por motivos de simplicidad.

[2] `http://en.wikipedia.org/wiki/List_of_emoticons`

Para el módulo de reglas de transformación léxicas, se ha generado manualmente un conjunto de 71 reglas. A pesar de que el proceso fue significativamente más lento que el de la generación de léxicos, sólo se necesitó de 15 horas para componer y refinar el conjunto final de reglas, siendo un coste relativamente bajo para reglas hechas a medida que abordan diversos fenómenos de error variados entre sí. Como dato adicional, el conjunto de reglas está únicamente vinculado al lenguaje en cuestión, por lo que puede ser usado para otros dominios y/o géneros.

El módulo detector de lenguaje foráneo utiliza un modelo de lenguaje de trigramas de caracteres como estrategia principal y utiliza información de diccionarios adicionales como estrategia de respaldo para el caso de que no haya datos suficientes en el tweet para estimar su lenguaje eficazmente.

3.6. Evaluación del sistema

En esta sección se realiza la evaluación del sistema propuesto, midiendo el rendimiento del mismo bajo diferentes perspectivas:*modulo*, *fenómeno* y *etapa selección de candidato*. Se definen varias medidas de rendimiento, cada una diseñada para evaluar un aspecto relevante en concreto de la cadena de procesamiento. Estas medidas se describen y formalizan en el siguiente apartado (ver 3.6.1). La muestra anotada del dataset usada durante este proceso de evaluación es la misma que la descrita previamente en el apartado 3.3.

3.6.1. Medidas de rendimiento

Como se ha mencionado previamente, para poder medir correctamente el rendimiento del sistema se proponen diversas medidas específicamente diseñadas para evaluar diferentes aspectos del sistema. Se describe cada métrica a continuación:

- *Candidate Coverage* o cobertura de candidatos: mide cuántas veces un término OOV ha sido cubierto por el sistema, existiendo la corrección apropiada dentro del conjunto de candidatos propuesto. En esencia, mide la capacidad del sistema para generar soluciones correctas, aunque no sean necesariamente elegidas como corrección definitiva.
- *Selection Precision* o precisión de selección: mide cuántas veces el candidato correcto ha sido correctamente elegido del conjunto de candidatos propuesto, siempre y cuando el candidato correcto se halle dentro del conjunto de candidatos propuesto. Mide cómo de exacta y precisa es la etapa de selección de candidatos.
- *Accuracy* o exactitud: mide cuántas veces un token OOV ha sido correctamente abordado, significando que el sistema genera y elige la correcta corrección para ese término OOV. Esta medida está directamente relacionada con la métrica *accuracy* encontrada en el campo de *Information Retrieval* y mide cómo de bien el sistema realmente corrige satisfactoriamente los términos OOV encontrados.

Además de la explicación de las métricas propuestas, se provee una formalización para ellas tal y como sigue:

Definición 3.1 *Dada las siguientes provisiones:*

- *Sea* $T_{dataset}$ *el conjunto de todos los tweets pertenecientes al dataset de evaluación.*
- *sea* OOV_t *el conjunto de los términos OOV detectados en el tweet* $t \in T_{dataset}$.
- *Sea* C^t_{oov} *el conjunto de candidatos generados por el sistema para un token OOV* $oov \in OOV_t$ *detectado en el tweet* $t \in T_{dataset}$.
- *Sea* $corr^t_{oov}$ *la corrección etiquetada para el token* $oov \in OOV_t$ *detectado en* $t \in T_{dataset}$.
- *Sea* $csel^t_{oov}$ *el candidato seleccionado por el sistema del conjunto* C^t_{oov}.

Las medidas para analizar el rendimiento del sistema son las que siguen:

$$Candidate\ Coverage = \frac{\sum_{t \in T_{dataset}} |\{corr^t_{oov} : corr^t_{oov} \in C^t_{oov},\ oov \in OOV_t\}|}{\sum_{t \in T_{dataset}} |\{oov : oov \in OOV_t\}|}$$

$$Selection\ Precision = \frac{\sum_{t \in T_{dataset}} |\{csel^t_{oov} : csel^t_{oov} = corr^t_{oov},\ csel^t_{oov} \in C^t_{oov},\ oov \in OOV_t\}|}{\sum_{t \in T_{dataset}} |\{corr^t_{oov} : corr^t_{oov} \in C^t_{oov},\ oov \in OOV_t\}|}$$

$$Accuracy = \frac{\sum_{t \in T_{dataset}} |\{csel^t_{oov} : csel^t_{oov} = corr^t_{oov},\ csel^t_{oov} \in C^t_{oov},\ oov \in OOV_t\}|}{\sum_{t \in T_{dataset}} |\{oov : oov \in OOV_t\}|}$$

3.6.2. Evaluación del sistema respecto a los módulos

En este apartado se muestra el rendimiento del sistema con diferentes módulos activados incrementalmente, observando cómo se comporta el mismo a medida que se van añadiendo módulos uno a uno. Además, se analiza la *contribución individual* de cada módulo al rendimiento general del sistema.

La figura 3.7 y la tabla 3.6 muestran los valores de rendimiento del sistema de normalización respecto al dataset de evaluación. Se observa que la medida *accuracy* mejora significativamente a medida que se van activando más módulos, apreciándose un incrementando del número total de candidatos generados por término OOV.

Este aumento de rendimiento se debe principalmente al rol que tienen los módulos sobre generación de candidatos. Cada módulo contribuye independientemente con sus proposiciones de corrección, incrementando el número de candidatos diferentes a ser considerados en cada conjunto de candidatos y por ello, teniendo un gran impacto sobre la cobertura del sistema (*candidate coverage*).

Conjuntos de candidatos mayores, sobre todo si los candidatos provienen de diferentes módulos, tienden a contener la corrección correcta debido a que cada candidato ha sido generado para resolver un fenómeno de error en particular, abordando una mayor casuística. Este aumento de la cobertura al activar más módulos, tiene como contrapartida la introducción de ruido adicional y mayores requisitos computacionales.

Modulos Activos	Candidatos / OOV	Candidate Coverage	Selection Precision	Accuracy
Lang (referencia)	43,01	48,07 %	65,92 %	31,68 %
Lang, Genre + Domain (GD)	94,00	51,16 %	64,05 %	32,77 %
Lang, GD, Transformation Rules (TR)	94,21	70,48 %	90,13 %	63,52 %
Lang, GD, TR, ASCII Art & Emoticons (ASCII)	94,91	83,62 %	**91,68** %	76,66 %
Full System (Lang, GD, TR, ASCII, Foreign language)	95,22	**91,65** %	88,20 %	**80,83** %

Cuadro 3.6: Rendimiento del sistema con diferentes módulos activados incrementalmente

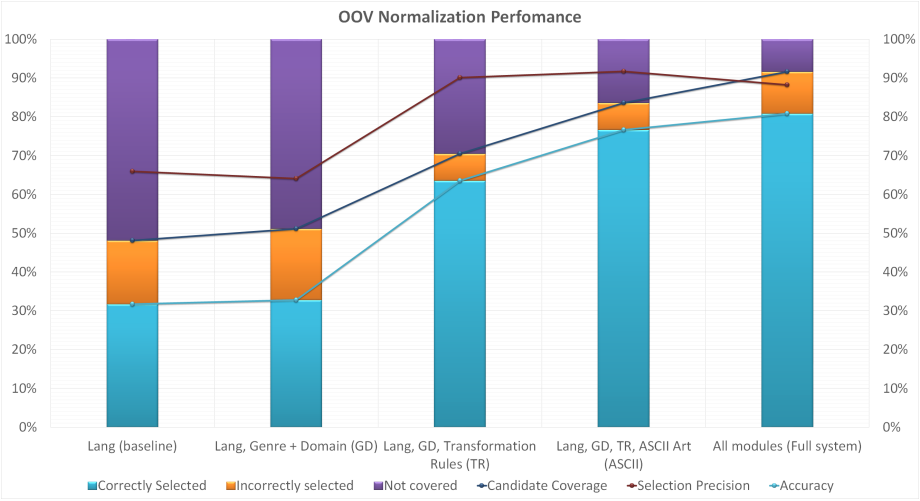


Figura 3.7: Rendimiento del sistema con diferentes módulos activados

La tabla 3.7 muestra cómo los distintos módulos contribuyen a generar el candidato correcto respecto a cada tipo de fenómeno de error. Hay que tener en cuenta que varios módulos pueden proponer al mismo candidato de forma independiente (causando solapamiento) y algunos fenómenos se logran cubrir completamente, dando a lugar que la suma de las columnas no tiene por que ser 100 %.

Merece la pena destacar que se ha usado un umbral de distancia máxima $k \leq 2$ para el módulo de distancia de edición, generando así candidatos que sólo estén a distancia 2 como máximo, debido a que la mayoría de los candidatos correctos se encuentran a esa distancia. Aunque es cierto que seleccionando un umbral k mayor se incluyen más candidatos, la mayoría de estos candidatos

Módulos	ASC	HOMO	ONO	ORT	FT	REP	TXT
Edit distance	00,00 %	**92,11** %	06,38 %	**86,31** %	00,00 %	57,98 %	15,79 %
Transform rules	00,00 %	**92,11** %	**63,38** %	56,65 %	00,00 %	**99,16** %	**53,95** %
ASCII Art	**94,44** %	00,00 %	00,00 %	00,00 %	00,00 %	00,00 %	00,00 %
Foreign language	00,00 %	00,00 %	00,00 %	00,00 %	**100,00** %	00,00 %	00,00 %
Estrategia defecto	05,56 %	00,00 %	34,04 %	00,00 %	00,00 %	00,00 %	00,00 %

Cuadro 3.7: Contribución de los módulos a la generación del candidato correcto respecto a cada fenómeno de error

adicionales no suelen ser soluciones válidas, teniendo valores de confianza bajos y no siendo seleccionados.

3.6.3. Rendimiento del sistema respecto a los fenómenos de error

En este apartado se analizan los diferentes valores de rendimiento del sistema con todos los módulos activados (*full system*) respecto a cada uno de los tipos de fenómenos de error subyacentes. No todos los tipos de fenómenos son igualmente difíciles de resolver y en esta sección se analiza esta situación, proporcionando una explicación sobre el comportamiento de cada fenómeno y las peculiaridades que cada uno presenta.

Fenómeno	Selection Precision	Candidate Coverage	Accuracy
ASC	92,22 %	100,00 %	92,22 %
HOMO	100,00 %	92,11 %	92,11 %
ONO	78,72 %	100,00 %	78,72 %
ORT	85,90 %	86,31 %	74,14 %
FT	62,96 %	100,00 %	62,96 %
REP	99,15 %	99,16 %	98,32 %
TXT	79,59 %	64,47 %	51,32 %

Cuadro 3.8: Rendimiento del sistema completo respecto a diferentes fenómenos de error

El sistema completo rinde de forma distinta dependiendo del fenómeno de error subyacente al que se enfrenta en cada término OOV, siendo algunos fenómenos más fáciles de normalizar que otros. La tabla 3.8 y la figura 3.8 detallan el rendimiento del sistema al completo (todos los módulos activados) respecto a cada fenómeno de error.

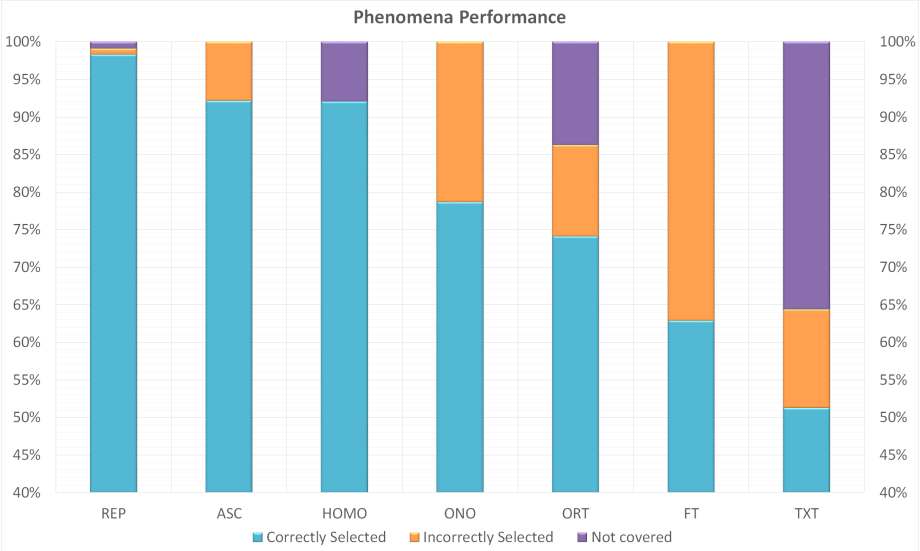


Figura 3.8: Rendimiento del sistema completo respecto a diferentes fenómenos de error

Algunos fenómenos de error como `ASC`, `HOMO` y `REP` son mas sencillos de tratar debido a que son fenómenos mejor comprendidos y definibles y, por ello, abordados con mucha efectividad. Otros fenómenos como `ONO` y `ORT`, también son fenómenos bien comprendidos pero son más difíciles de normalizar debido a factores como una menor cobertura por parte de los módulos y que son fenómenos fácilmente confundibles con otros.

El fenómeno `FT` obtiene peores resultados de rendimiento debido a que es fácilmente confundible con el fenómeno `ORT`. Es común encontrar que los usuarios substituyen palabras españolas por términos "equivalentes" de otros lenguajes y algunas veces estos términos son muy similares a términos válidos en español. Esto conlleva a un alto grado de confusión entre los fenómenos `FT` y `ORT` que resulta en una baja precisión en la etapa de selección de candidatos.

El fenómeno `TXT` es el que obtiene el rendimiento mas bajo. Es difícil de abordar correctamente, pues nuevos acrónimos de carácter específico son generados constantemente y para cada término válido del lenguaje existen varias variantes abreviadas tipo SMS.

3.6.4. Capacidad de adaptación del sistema

En este apartado se evalúa la flexibilidad del sistema al adaptarse éste a un nuevo dominio radicalmente distinto al usado en la configuración inicial, realizando los experimentos sobre un corpus perteneciente a un dominio diferente.

Por consiguiente, se ha obtenido otro dataset sobre un tema de diferente naturaleza: tweets escritos en español durante la presentación del borrador final con las modificaciones a la ley que regula el aborto en España, conocida coloquialmente como *ley del aborto*. El periodo en cuestión abarca desde el 20 de Diciembre del 2013 hasta el 23 de Diciembre del 2013. La reforma propuesta causó un gran impacto sobre la población española y los partidos políticos ma-

yoritarios se posicionaron activamente respecto a la propuesta. El proceso de generación del dataset del *aborto* fue el mismo que el usado en para el dataset *Euro2012*: el método de recuperación dinámica descrito en Cotelo et al. (2014).

Dado que el lenguaje y la plataforma permanecen siendo iguales, normalizar los tweets de este dataset sólo requirió de la generación de un nuevo recurso de dominio (*Domain*), reutilizando directamente los recursos y parámetros del experimento sobre el dataset *Euro2012*. Este nuevo recurso de dominio contiene los nombre de los políticos y ministros relevantes, términos específicos sobre el tema del aborto y otros términos directamente relacionados con el sistema legislativo español. Componer este recurso sólo costó una hora de esfuerzo manual, siendo un coste de adaptación muy bajo.

Se procedió de manera muy similar a la hora de generar el dataset de evaluación: se obtuvo una muestra estadísticamente significativa del dataset *aborto* ($\alpha = 0{,}05$, error = 1 %), cuyos tweets pertenecientes a la muestra debían tener al menos un término identificado como OOV. Este dataset de evaluación fue anotado manualmente de la misma forma que para el experimento *Euro2012*.

Fenómeno	Selection Precision	Candidate Coverage	Total Accuracy
`ASC`	76,67 %	100,00 %	76,67 %
`HOMO`	95,45 %	95,65 %	91,30 %
`ORT`	87,50 %	87,13 %	76,24 %
`REP`	100,00 %	100,00 %	100,00 %
`TXT`	96,63 %	94,98 %	86,96 %
Total	94,40 %	91,81 %	82,71 %

Cuadro 3.9: Rendimiento del sistema sobre el dataset de evaluación del *aborto*.

La tabla 3.9 muestra el rendimiento obtenido por el sistema (incluyendo un desglose por fenómeno) al configurarse para abordar un nuevo dominio. El rendimiento global obtenido por el sistema se encuentra dentro del mismo rango de valores obtenido por el experimento anterior, sólo existiendo una ligera mejora de rendimiento (+1,88 %) respecto al otro experimento. Esta diferencia puede ser atribuida a que la calidad de los tweets en el dataset *aborto* es mayor y que los fenómenos de error encontrados son ligeramente más fáciles de resolver.

Se concluye que con un mínimo esfuerzo manual adicional (alrededor de 1 hora) para generar el recurso adicional, el sistema propuesto se adapta muy satisfactoriamente al nuevo dominio.

3.6.5. Ajustando la etapa de selección de candidatos

En este apartado se realiza un análisis sobre la etapa de selección de candidatos de la cadena de procesado, el cual sienta la bases para una modificación de la actual etapa de selección. Esta modificación consiste en la inclusión de una sub-etapa de clasificación *a posteriori* que permite un refinado aún mayor de la etapa de selección de candidatos. Este refinamiento conlleva a un significante número de mejoras de rendimiento.

La etapa de selección de candidatos, como se ha visto previamente en la

sección 3.4, es directa y sencilla: consiste en generar un ranking sobre los candidatos usando los valores de confianza provistos por los módulos analizadores y seleccionar el candidato que mejor puntuación obtenga. El objetivo final de este paso es la correcta elección del candidato, siendo este la corrección correcta para el término OOV dado.

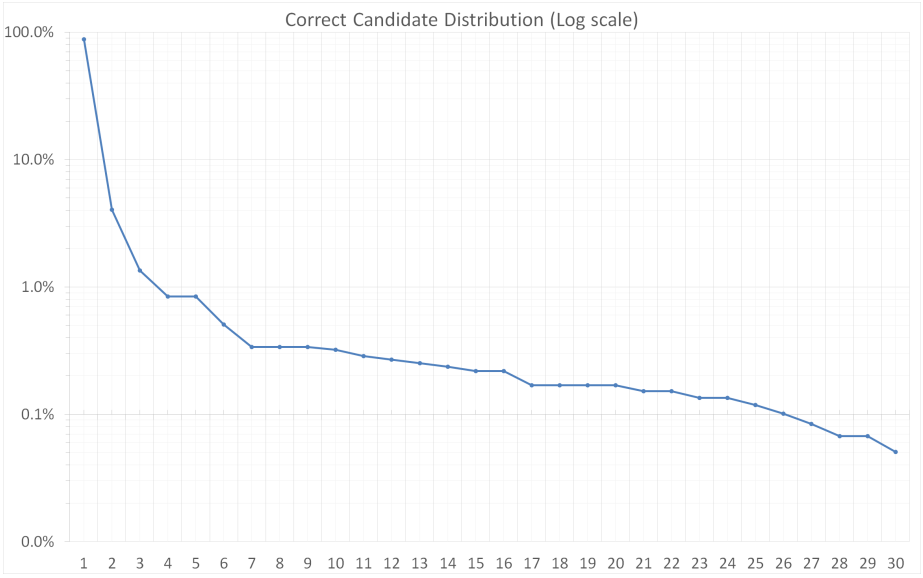


Figura 3.9: Distribución de la posición de ranking del candidato correcto

El rendimiento general de la etapa de selección de candidatos se estima mediante la medida *Selection Precision* definida con anterioridad en la sección de evaluación. Teniendo en cuenta la actual etapa de selección de candidatos, la medida simplemente cuenta el número de candidatos correctos que han sido colocados como primeros dentro del ranking. La figura 3.9 muestra la distribución de las posiciones donde el candidato correcto ha sido colocado dentro del ranking por el módulo de selección de candidatos.

Es fácil observar que ésta distribución se asemeja a una distribución *ley de potencias* o *power-law*: la mayoría de los candidatos correctos se encuentran entre las primeras posiciones del ranking seguidos de una cola larga. Como consecuencia de este tipo de distribuciones, teniendo en cuenta sólo los primeros $n = 5$ posiciones del ranking, más del 95 % de los candidatos correctos son cubiertos.

Candidatos considerados	Cota superior Selection Precision	Cota superior Accuracy
Posición $p \in [1, 3]$	93,60 %	85,78 %
Posición $p \in [1, 5]$	95,28 %	87,33 %
Todos	100,00 %	91,65 %

Cuadro 3.10: Rendimiento maximal teórico para un proceso de candidato de selección perfecto

La tabla 3.10 muestra los valores maximales de rendimiento teóricamente alcanzables si el proceso de selección de candidatos fuera totalmente perfecto. Observando los valores, se infiere que al incluir un hipotético proceso que genere un reranking preciso sobre las primeras posiciones del ranking original, se aumentaría el rendimiento total del sistema.

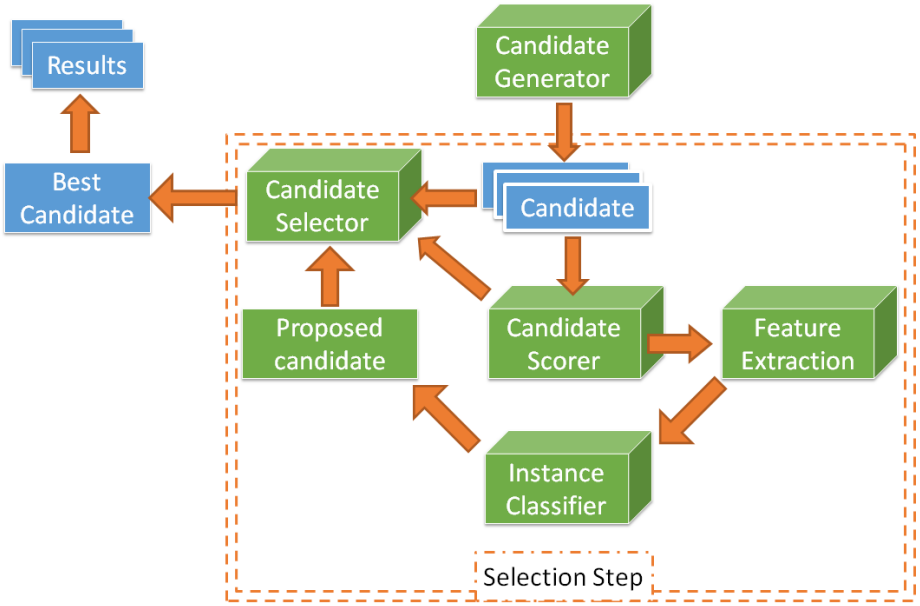


Figura 3.10: Inclusión de la etapa de clasificación *a posteriori* a la etapa de selección del sistema propuesto

Teniendo en cuenta estos resultados, se propone una variante del proceso de selección de candidatos original que consiste en incluir un proceso auxiliar basado en clasificadores. Esta extensión ajusta y afina el proceso de selección, sugiriendo rerankings de los primeros n elementos del ranking basado en puntuaciones original. La figura 3.10 muestra cómo la extensión se incluiría dentro del proceso de selección original, modificando la arquitectura del sistema.

Para el proceso clasificador de la extension, se ha elegido un clasificador *Random Forest*. Este clasificador consiste en un meta-estimador que entrena independientemente un número finito de árboles de decisión sobre subconjuntos de los datos de entrenamiento, suavizando los resultados, mejorando la capacidad predictiva y reduciendo el sobreajuste.

Esta etapa de clasificación *a posteriori* intenta maximizar el rendimiento del proceso de selección de candidatos del sistema, aumentando los resultados de la medida *Selection Precision*. Si nos basamos en los datos experimentales discutidos anteriormente, se observa que es más que adecuado el considerar sólo los $n = 5$ primeros candidatos para este proceso.

La tabla 3.11 y la figura 3.11 muestran el rendimiento del sistema con la extensión propuesta. Se ha utilizado un esquema de validación cruzada con $k = 10$ durante el proceso de evaluación del clasificador para evitar el sobreajuste. Al introducir esta extensión en la etapa de selección, se observa un incremento significativo en el rendimiento total del sistema, debido a que esta extensión

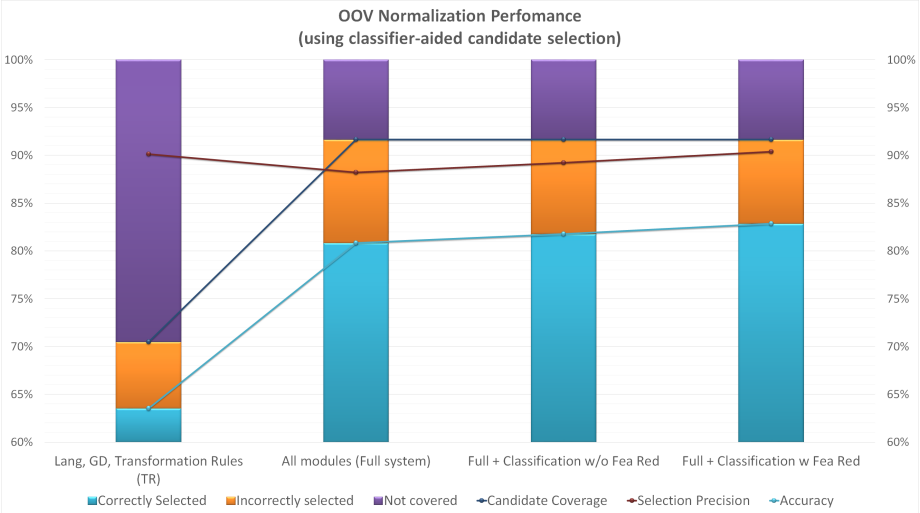


Figura 3.11: Rendimiento del sistema con una etapa de clasificación *a posteriori*

Módulos activos	Candidate Coverage	Selection Precision	Accuracy
Lang, GD, Transformation Rules (TR)	70,48 %	90,13 %	63,52 %
Todos (Full system)	91,65 %	88,20 %	80,83 %
Todos + Clasificador	91,65 %	89,21 %	81,76 %
Todos + Clasificador + Reducción Características	**91,65** %	**90,39** %	**82,84** %

Cuadro 3.11: Comparativa de rendimiento incluyendo la extensión propuesta

realmente refina el ranking propuesto inicialmente por el módulo.

Si se realiza un proceso automático de reducción de características, el sistema consigue un incremento de rendimiento en la selección de candidato del 88,20 % al 90,39 %, que resulta en una mejora total del sistema pasando del 80,83 % al 82,84 %. Esta mejora es bastante significativa; es muy difícil superar la barrera del 90 % en la medida de *Selection Precision* debido a la dificultad de la tarea en sí a estos niveles de rendimiento.

3.7. Conclusiones y trabajo futuro

A lo largo de este capítulo se ha presentado una novedosa aproximación modular basada en recursos para abordar la tarea de la normalización léxica de tweets. La principal idea detrás de esta aproximación consiste en que el sistema diseñado en cuestión se comporte como un "grupo de expertos", logrando un sistema más tolerante con los errores difíciles y más sencillo de ampliar que las aproximaciones encontradas comúnmente en la literatura, las cuales recurren a sólo una técnica específica para resolver el problema.

El sistema propuesto en cuestión posee una arquitectura extensible compuesta por módulos independientes, cada uno de ellos enfocado a una tarea específica o un fenómeno de error en concreto. Esta capacidad de especialización permite rebajar lo costes para hacer cualquier módulo a la vez que se incrementa la eficacia del sistema. Además, expandir el sistema para abordar otros fenómenos de error distintos sólo requiere la adición de módulos específicamente construidos para resolver dicho fenómeno.

Esta combinación de arquitectura modular y de recursos de carácter ligero hace que esta aproximación sea muy fácil de adaptar a otros dominios, géneros u otros escenarios a muy bajo coste de implementación adicional. Una vez que el sistema está implementado para algún lenguaje en concreto, los costes de adaptación son muy reducidos; en nuestro caso, una vez creado los recursos de lenguaje, sólo costó 4 horas de trabajo manual generar los recursos para que abordara el género y el dominio en cuestión.

Se ha mostrado que al aumentar el número de módulos de análisis especializados en el sistema, este experimenta un incremento global sobre el rendimiento a expensas de introducir algo de ruido, causando que la etapa de selección de candidatos sea ligeramente más difícil. Para paliar este efecto, se propone una mejora sobre la etapa de selección de candidatos original consistente en la introducción de un clasificador automático para realizar un ajuste sobre los rankings obtenidos. Los resultados muestran que el uso de este proceso de clasificación *a posteriori* incrementa significativamente el rendimiento del sistema debido a un aumento de la precisión en la etapa de selección.

Considerando la mejora mencionada, el rendimiento total del sistema es bastante significativo: obtiene más de del 82 % de *accuracy* comparado con el 31 % obtenido por el baseline.

El sistema propuesto también posee ciertas desventajas. Tal y como se ha mencionado previamente, aumentar el número de módulos que participan en el sistema "pseudo-democratico" de generación de candidatos conlleva un aumento de ruido y dificulta el proceso de decisión. Aunque este ruido inducido es inherente a este tipo de esquema de funcionamiento e impone una vaga cota superior sobre el rendimiento del sistema, los módulos pueden ser diseñados para reducir el solapamiento en la medida de lo posible, generando menos ruido y causando una menor confusión.

Como en toda aproximación léxica, un factor importante a tener en cuenta es el correcto diseño e implementación de la etapa de preprocesado. El procesado, tokenización y detección de OOV son actividades vitales que deben hacerse lo mejor posible para el correcto funcionamiento del sistema; todo el resto del sistema depende de la fase de preprocesado y detección.

El sistema puede ser mejorado en varias direcciones. Actualmente se utilizan expresiones regulares para tokenizar y tratar los textos de los tweets. Incluir un módulo segmentador en la etapa de preprocesado resultaría en una mejor tokenización y, por ende, en una posible mejora del sistema.

Cada token OOV se aborda y analiza a nivel léxico de forma independiente, no teniendo en cuenta el resto del tweet. Tanto la generación y la selección de candidatos podría beneficiarse mucho si dispusieran de cierta información contextual, permitiendo el descarte automático de candidatos que no serían factibles debido a restricciones de carácter morfosintáctico.

Otra dirección de investigación interesante consiste en mejorar el sistema de puntuación de candidatos y utilizar mejores métodos de selección. En lugar de

usar heurísticas *ad-hoc* para generar valores de confianza, los módulos podrían hacer uso de métodos de aprendizaje automático para establecer estos valores de confianza, haciendo que la etapa de selección sea más precisa y sencilla.

Capítulo 4

Combinación de información textual y estructural aplicada a la categorización automática de tweets

En este capítulo se explora la tarea de categorización de tweets, abordando específicamente la determinación de la opinión política de los usuarios dentro de su contexto político. Para ello, en lugar de confiar únicamente en el análisis del contenido textual de los tweets, tal y como lo hacen la mayoría de trabajos existentes, se extrae conocimiento complementario de la topología de la red social y se realiza un análisis de la información estructural provista por las relaciones entre tweets y usuarios.

A lo largo del capítulo se realiza una discusión sobre el análisis de ambos tipos de contenido (textual y estructural), aplicando diferentes aproximaciones para obtener diferentes modelos y realizando una evaluación sobre ellos. También se discute la idea de integrar información de ambos aspectos de los tweets, combinando modelos provenientes de diferentes tipos de conocimiento mediante varios métodos. Después de la evaluación de estos métodos de combinación, en la cual se obtienen buenos resultados, se realiza una discusión sobre las principales ventajas e inconvenientes a tener en cuenta.

4.1. Introducción

Categorizar mensajes de Twitter es una tarea interesante y valiosa. Como ya hemos comentado, en este capítulo se aborda la categorización de tweets, enfocada específicamente a la determinación de la opinión política de los tweets escritos dentro de un contexto político en particular. Una colección de tweets no solo contiene información textual, sino que también ofrece cierta información estructural debido a la relaciones existentes entre mensajes y usuarios, formando una red implícita.

A lo largo de este capítulo se discute el análisis de ambos tipos de contenido, se aplican diferentes propuestas a cada tipo de contenido para generar modelos

de características y se proponen varios métodos para combinar con éxito modelos de características provenientes de diferentes tipos de información en el proceso de clasificación.

En esta memoria de tesis se explora la tarea de categorización de tweets, la abordando determinación de la opinión política, específicamente dentro del contexto político español.

Extraer conocimiento del contenido estructural de los tweet y generar un modelo de características válido no ha resultado trivial. En este caso, requirió de la generación de una representación basada en grafos e inferir las comunidades emergentes dentro de él para generar un modelo de características los suficientemente expresivo.

La combinación de modelos extraídos provenientes de diferentes tipos de conocimiento (en este caso, modelos textuales con modelos estructurales) resultó ser bastante satisfactorio, obteniendo buenos resultados pero entrañando una serie de dificultades debido a que dichos modelos eran de naturaleza diferente. Para ello, se desarrolló una estrategia de combinación de *Ensemble Learning* que se llama *Multiple Pipeline Stacked Generalization*, diseñada específicamente para aprovecharse de la mezcla de modelos diferentes.

Este capítulo está estructurado de la siguiente manera: En la sección 4.2 (*Trabajos relacionados*) se realiza una revisión del estado del arte actual relacionado con la temática del capítulo y se analizan algunos trabajos de mayor interés.

En la sección 4.3 (*Definición de la tarea*), se define específicamente la tarea abordada en este trabajo y se realiza una caracterización del dataset en cuestión, describiéndose como se ha generado y las peculiaridades que posee.

En la sección 4.4 (*Extracción de conocimiento a partir del contenido textual*), se aborda la extracción de conocimiento del contenido textual explícito de los tweets, estudiando la aplicación de modelos tales como el *Bag-of-Words* (BoW) y su problemática, proponiendo un proceso de selección automática de características para una mejora en el rendimiento ofrecido por este modelo.

En la sección 4.5 (*Extracción de conocimiento a partir del contenido estructural*) se diserta sobre cómo extraer conocimiento útil sobre el información estructural de los tweets, explotando características topológicas de la subyacente red formada por los usuarios y los mensajes. Para ello, se realiza una aproximación de carácter topológico consistente en generar un grafo de amistad bipartito basado en la identificación de dos grandes tipos de usuarios (*creadores de contenidos* and *consumidores de contenido*). Tomando este grafo bipartito como base, se proponen dos modelos de características de comunidad diferentes, basados en el *Louvain Method* y la técnica *Spectral Biclustering* respectivamente.

En la sección 4.6 (*Estrategias de combinación de características*), se combinan modelos tanto estructurales como textuales de tres formas distintas: mediante la combinación directa, usando la técnica *Stacking Generalization* y mediante una variación propia de esta técnica que se ha denominado *Multiple Pipeline Stacked Generalization*.

Finalmente, en la sección 4.7 (*Conclusiones y trabajo futuro*), se resumen los esfuerzos realizados en este capítulo, se revisan los puntos principales del trabajo y se diserta sobre la importancia de la combinación de contenidos tanto textuales como estructurales para la categorización automática de tweets.

4.2. Trabajos relacionados

Existen varios trabajos que tratan con la clasificación de tweets, especialmente dentro del área de *Sentiment Analisys*, tratando de determinar opiniones en el contenido subjetivo de muchos de los tweets. La mayoría de estos trabajos encaran la clasificación de *polaridad*, decidiendo si un tweet expresa una actitud u opinión positiva, negativa o neutral respecto a un tema en cuestión. Por ejemplo, en Babour and Khan (2014), la clasificación de polaridad se realiza mediante el uso de léxicones de polaridad (recursos que consisten en listas de palabras positivas y negativas) y posteriormente ajustando la polaridades obtenidas teniendo el cuenta el contexto semántico en que dichas palabras aparecen. En Al-Osaimi and Badruddin (2014), la falta de léxicones de polaridad del lenguaje árabe es paliada mediante el uso de emoticonos que aparecen en los tweets, construyendo un clasificador de polaridad usando estos emoticonos. Dicha idea se extrae directamente de la obra Pak and Paroubek (2010), donde la misma técnica es aplicada a un corpus de tweets pero escritos en inglés. En otros trabajos como Xie et al. (2014), los autores diseñan un sistema para generar resúmenes automáticos de las opiniones de un usuario de Twitter mediante la integración de varias opiniones, tanto negativas como positivas, expresadas por los usuarios respecto a diferentes temas.

En todos estos trabajos, la clasificación del tweet es realizada únicamente en su contenido textual: la información estructural inherente a Twitter no es usada en ningún caso, como pudiera ser la relación o similitud entre tweets de diferentes usuarios o el uso común de etiquetas.

4.2.1. Clasificación de polaridad en política

Varios autores han estado interesados en el estudio del comportamiento de los usuarios de Twitter en relación con el dominio de la política. En la obra Small (2011), se realiza un análisis sobre los hashtags usados dentro del contexto de la política canadiense, intentando distinguir los diferentes objetivos para los cuales dichos hastags son usados en este contexto.

En Park (2013), se realizó un estudio sobre los diferentes tipos de usuarios existentes en Twitter, con el fin de caracterizar los, así llamados, *líderes de opinión*; estos líderes son bastante activos y tienden a buscar información, movilizar colectivos y expresar opiniones públicamente, ejerciendo una gran influencia dentro de la tendencia política de sus seguidores (idea que aplicaremos en nuestro trabajo, ver sección 4.5).

En el trabajo Tumasjan et al. (2010), se usó el recurso *LIWC2007* (Linguistic Inquiry and Word Count; Pennebaker Kahn et al. (2007)) para determinar características emocionales y cognitivas de los tweets en relación a las elecciones federales al parlamento nacional alemán en 2009. Los autores concluyen que existe una alta correlación entre los resultados obtenidos y métricas estadísticas poblacionales tales como concentración y participación. Incluso el mero número de tweets que mencionan a un candidato es un buen estimador de los resultados electorales.

Los autores de la obra Barclay (2014) también intentan medir si existe alguna correlación entre la actividad en redes sociales (Twitter y Facebook) y los resultados de las elecciones presidenciales de los EEUU del 2012. Los autores realizaron manualmente una clasificación de polaridad de los ciudadanos sobre

los tweets que mencionaban a ambos candidatos (Obama y Romney), páginas oficiales de dichos candidatos y los comentarios en estas páginas.

La conclusión final es que existe una correlación fuertemente positiva entre los datos obtenidos y los resultados electorales. Estas observaciones sobre la utilidad de la información en los tweets a la hora de realizar predicciones electorales o estimar otras variables políticas que tradicionalmente han sido estimadas mediante encuestas, hacen de las clasificación automática de tweets una tarea muy interesante dentro de los contextos políticos.

4.2.2. Clasificación de polaridad mediante propagación de etiquetas

Previamente se ha mencionado que la mayoría de trabajos de clasificación se basan en procesar el contenido textual de los tweets sin tener en cuenta la información estructural que sus mensajes poseen. Sin embargo, existen algunas pocas obras que hacen uso, ya sea total o parcial, de las relaciones entre usuarios, tweets y etiquetas existentes en los mensajes de la red.

Un buen ejemplo se da en la obra Speriosu et al. (2011), en la cual se realiza una transformación del contenido textual de los tweets a una representación basada en grafo de los tweets, la cual es combinada con el grafo de seguidores de los usuarios involucrados y se aplica un algoritmo de propagación de etiquetas para completar el grafo.

El punto de partida consiste en generar un grafo cuyos nodos representan a los tweets, palabras y usuarios, donde parte de ellos (conjunto semilla) deben estar inicialmente etiquetados. Se presentan varios módulos para obtener este etiquetado inicial y ruidoso:

- **EmoMaxEnt**: Haciendo uso de un clasificador de *Máxima Entropía (MaxEnt)* entrenado con un léxico de emoticonos, se generan prediciones de polaridad para cada tweet y se anota dicha polaridad en los nodos que representan los tweets.
- **Léxico de opinión**: Se crea un nodo por cada palabra de opinión encontrada dentro del léxico *OpinionFinder* (Wilson et al., 2005). Dado un tweet colocado en el grafo, por cada palabra de opinión existente en el contenido de dicho tweet, se añade una arista entre el tweet y el nodo que representa la palabra de opinión.
- **Dataset Anotado**: Se generan un conjunto de nodos tweet semilla ya anotados extraídos de un dataset de polaridad de tweets. Estos nodos son enlazados con el resto de elementos del sistema (usuarios, hasthags y palabras).

Una vez generado este grafo inicial, se realiza una serie de esquemas de ponderación (ver Speriosu et al. (2011) para una explicación en mayor profunidad) sobre las diferentes aristas entre los nodos de usuarios, hashtags, emoticonos y resto de tweets. El grafo resultante se usará como grafo base a la hora de aplicar el algoritmo de propagación de etiquetas.

La predicción se hace de una forma muy directa: se completa el grafo base con el mismo tipo de información de los tweets a clasificar (sólo que éstos carecen de polaridad base) y se aplica el algoritmo de propagación de etiquetas descrito en

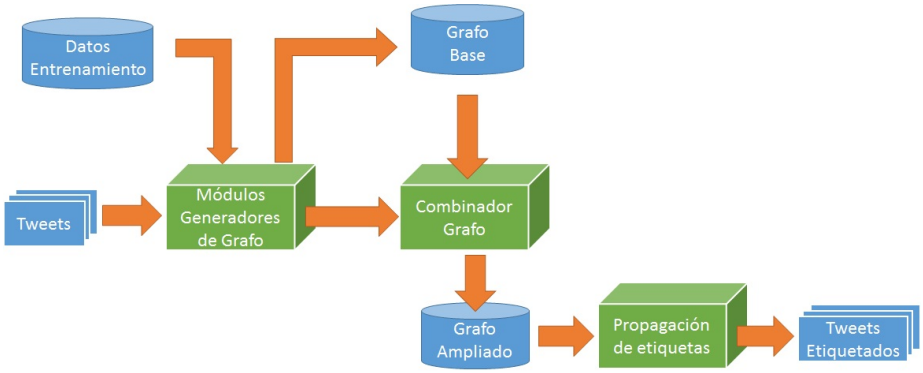


Figura 4.1: Arquitectura y funcionamiento del sistema de propagación de etiquetas descrito en Speriosu et al. (2011)

la obra Talukdar and Crammer (2009), obteniendo una puntuación de polaridad para cada nuevo nodo tweet sin etiquetar. La figura 4.1 muestra la arquitectura y el funcionamiento general del sistema descrito.

Los autores realizan una evaluación con tres corpus de diferente naturaleza para probar su propuesta, siendo uno de ellos relacionado con el dominio de la política. Sin embargo, los datasets no se comportan de forma equilibrada debido a que poseen grandes diferencias en el vocabulario respecto al conjunto de entrenamiento, obteniendo peores resultados en dos de ellos. Es más, los autores concluyen que su tratamiento del grafo de seguidores es claramente insuficiente, pues éste no logra conseguir mejoras significativas en el proceso de clasificación.

Aunque esta representación de grafo aplicada al contenido textual es efectiva y obtiene mejores resultados que otras aproximaciones que tratan el contenido textual en su forma más o menos original, esta propuesta sigue sin tener en cuenta gran parte del conocimiento de la topología de red existente y no explota significativamente el carácter estructural que la red subyacente ofrece.

4.2.3. Categorización mas allá de la polaridad

Aunque la determinación de polaridad básica (postivo, negativo, neutro) en la categorización de tweets suele ser el enfoque predominante, existen algunos trabajos que exploran otro tipo de características, tales como las emociones asociadas al tweet o el propósito detrás del mismo.

Un interesante trabajo que va mas allá de clasificar la polaridad de los tweets es el descrito en la obra Mohammad et al. (2014). En ella, los autores categorizan los tweets referentes a las elecciones presidenciales del 2012 de los EEUU respecto a diferentes aspectos: *polaridad* (positivo, negativo, neutro), *emoción* (felicidad, enfado, tristeza, ...), *propósito* (ridiculizar, proporcionar información factual, desahogarse, admirar, destacar errores o meteduras de pata, ...) y *estilo* (sarcasmo, hipérbole, aserción simple, ...).

Antes de diseñar ningún proceso de clasificación, los autores realizan un estudio preliminar sobre un conjunto de tweets que, por primera vez respecto a la bibliografía actual, fueron anotados manualmente para incluir este tipo de

información de forma conjunta.

Como primer paso de generación de dicho dataset, recuperaron el conjunto de tweets mediante una lista estática de 23 palabras clave usando la API básica de twitter, siendo 21 de éstos *hashtags* relacionados con las respectivas campañas. Esta aproximación es muy limitada, como se ha expuesto en el capítulo 2, y ha requerido de un filtrado a posteriori bastante exhaustivo para controlar tweets que no tienen nada que ver, tweets provenientes de otro idioma y retweets no deseados. El conjunto de tweets recuperados fue cercano a los $170k$, que es un número bastante bajo teniendo en cuenta que el periodo de recuperación fue de dos meses y el tamaño de la población de un país como EEUU.

Anotar manualmente tal cantidad de tweets no es una tarea factible. Por ello, el conjunto total de tweets anotado fue de 2000, exigiendo que cada uno de los tweets fuera de un usuario distinto. Lo interesante es que el proceso de anotación no fue realizado por el equipo investigador, sino que utilizaron un sistema de crowdsourcing combinando *CrowdFlower* y *Amazon Mechanical Turk*, donde usuarios anónimos pueden responder a una serie de preguntas en forma de cuestionarios u otro tipo de tareas. Para cada tarea propuesta a un usuario, se exponen dos cuestionarios: uno encargado de las emociones contenidas en los tweets y otro que relaciona las emociones, con un propósito o una temática.

Una vez anotado el componente emocional de los tweets, se procede a utilizar un sistema de clasificación automática para detectar las emociones y el propósito de los tweets usando el dataset anteriormente anotado. La figura 4.2 muestra el funcionamiento general del sistema de clasificación automática utilizado para cualquiera de las tareas, el cual consiste en generar un modelo de características con el cual representar los tweets, entrenar el clasificador SVM con el dataset anotado y generar predicciones ante tweets no vistos.

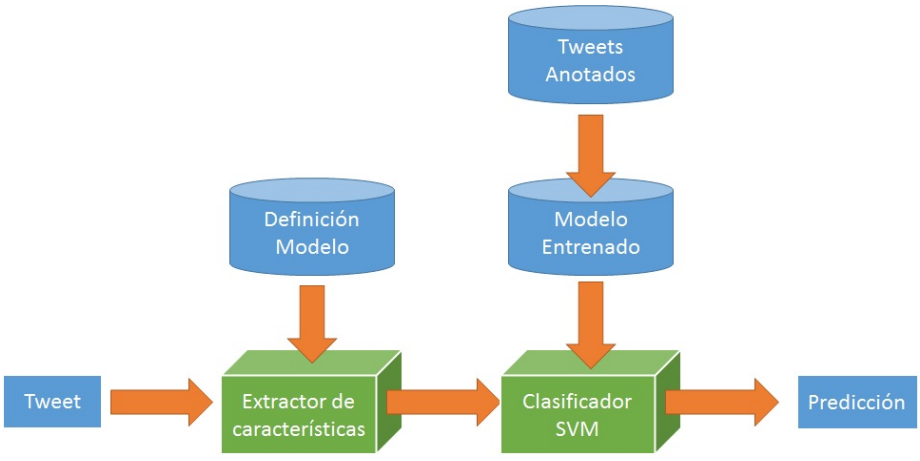


Figura 4.2: Esquema de funcionamiento general del sistema de clasificación automática descrito en Mohammad et al. (2014)

La detección de las emociones dentro de un tweets es dividida en dos subtareas: detectar el estado emocional y detectar el estímulo emocional. Para detectar el estado emocional, el cual es dividido en 8 categorías básicas (felicidad, tristeza, enfado, miedo, sorpresa, anticipación, confianza y disgusto), utilizan diferentes características en su modelo:

- **Presencia de N-gramas de palabras**: Se indican los unigramas y bigramas usados en el tweets, lematizados usando el *Porter's stemmer* (Porter, 1980).
- **Símbolos de puntuación**: Número de secuencias contiguas de signos de puntuación, tales como puntos, comas, signos de exclamación, signos de interrogación y combinaciones de ellos.
- **Palabras alargadas**: Número de palabras alargadas cuyo carácter final ha sido repetido más de tres veces, e.g., "sooo" y "mannnnnnnn".
- **Emoticonos**: Presencia o ausencia de emoticionos positivos o negativos. El emoticono y su polaridad se determinan mediante una expresión regular simple.
- **Léxicones de emoción**: Usan un léxico de asociación palabra-emoción (Mohammad and Turney, 2010) para comprobar si el tweet analizado tiene palabras de carácter emocional. Este léxico contiene anotaciones de emoción, hechas manualmente, para un total de 14k tipos de palabras. Cada palabra tiene asociada una emoción de las 8 emociones básicas expuestas anteriormente, por lo que un tweet puede llegar a tener varias palabras de diferentes emociones. Las características extraídas son el número de palabras que están asociadas a cada emoción básica dentro del tweet.
- **Características de negación**: Se examinan los tweets para comprobar si contienen negaciones. Si al realizar un análisis de dependencia sobre la estructura del tweet, un negador se encuentra cerca de una palabra de emoción (determinada por el léxico de emoción), éste modifica a la palabra de emoción.
- **Características de posición**: Se incluye un conjunto de características para capturar si los términos descritos anteriormente aparecen al principio o al final de tweet.
- **Características combinadas**: Aunque los modelos no lineales, tales como el SVM con un kernel no lineal, son capaces de capturar interacciones entre diferentes características de un modelo, los autores generan un conjunto de características combinadas, marcando explícitamente cuando dos sub-características relevantes aparecen a la vez.

El modelo de características descrito es generado a partir del dataset mencionado anteriormente y es aplicado para entrenar un clasificador SVM con el fin de abordar la tarea de detección de emociones.

De forma análoga a la detección de emoción, la detección de estímulo se realiza mediante un clasificador SVM sobre 8 categorías de estímulo (los 2 candidatos a gobierno, los 2 partidos políticos, el proceso electoral, alguna otra institución, algún otro individuo o simplemente sin especificar) partiendo el siguiente modelo de características: presencia de N-gramas de palabras, presencia de hashtags, características léxicas (referencia a cualquiera de las categorías de estímulo), características de posición y características combinadas.

Para identificar automáticamente el propósito de un tweet, se entrena otro clasificador SVM sobre 11 categorías de intención, representando cada tweet

como un vector del siguiente modelo de características: presencia de N-gramas de palabras, número de ocurrencias de elementos *Part of Speech (POS)*, presencia de clusters de palabras (palabras pertencientes a cada uno de los 1000 clusters de palabras provistos por la herramienta Twitter NLP tool; ver Gimpel et al. (2011)), cantidad de palabras totalmente en mayúsculas, cantidad de palabras de emoción, cantidad de secuencias consecutivas de símbolos de puntuación, cantidad de palabras alargadas y hashtags, presencia de emoticonos y características de negación.

Aunque los modelos de características son muy detallados y complejos, los resultados obtenidos tanto en las tareas de categorización de emoción son mejores respecto a la tarea de caracterización de propósito, pues los resultados no llegan a superar el 50 % de acierto. El proceso de detección automática de emoción es capaz de llegar al 56,84 % mientras que el proceso de detección automática de estímulo obtiene resultados similares.

La principal razón del bajo rendimiento de estas tareas se debe a que poseen un nivel de detalle demasiado alto y no existe un gran consenso entre los anotadores, provocando este desequilibrio entre tareas y el bajo rendimiento. Es más, los revisores humanos superan al método automático en más de un 30 % del rendimiento, llegando casi al 90 % de acierto en algunos casos y por lo que es cuestionable que esta aproximación sea del todo apropiada para la tarea en cuestión.

4.3. Definición de la tarea

Como ya hemos mencionado previamente, en esta capítulo se aborda la tarea de determinar la opinión política de los tweets escritos en español cuyos contenidos estén altamente relacionados con algún aspecto del contexto político español. Como en muchos otros países, la situación política española está claramente dominada por un puñado de fuerzas políticas, en particular por dos partidos políticos mayoritarios: el conservador, liberal y cristiano-demócrata *Partido Popular (PP)* y el social-demócrata *Partido Socialista Obrero Español (PSOE)*. Desde la transición española a la democracia, estos partidos son los únicos que han tomado gobierno en el país y son los partidos que van a ser el objeto de estudio para la tarea.

Se ha generado una colección de tweets que hacen referencia a alguno de los partidos en cuestión (PP o PSOE), utilizando el método de recuperación dinámica de tweets explicado en Cotelo et al. (2014) y expandido en el capítulo 2.

A partir de la colección de tweets completa, compuesta por más de 100k tweets, se ha generado el dataset final compuesto por una muestra de 3000 tweets manualmente anotados que hacen referencia al gobierno actual (el Partido Popular en el momento de la recolección de datos) o al partido de la oposición (PSOE).

Cualquier tweet de este dataset puede expresar una postura positiva, negativa o neutral respecto a los partidos PP y PSOE, así que se define la tarea de categorización como clasificar los tweets en alguna de las 9 categorías combinatorias (el producto cartesiano de las posibles posturas respecto al PP y al PSOE independientemente). Durante el proceso de anotación manual, cada tweet fue marcado indicando la postura política de dicho tweet respecto a las categorías

mencionadas.

	PSOE positivo	PSOE negativo	PSOE neutral
PP positivo	00,00 %	01,07 %	01,36 %
PP negativo	01,02 %	04,00 %	46,14 %
PP neutral	02,51 %	18,83 %	25,07 %

Cuadro 4.1: Distribución de la opinión política dentro del dataset.

La tabla 4.1 muestra la distribución de la opinión política de los tweets del dataset y se observa un fenómeno interesante. La mayoría de los usuarios españoles son bastante sectarios respecto a cualquier tema relacionado con la política, dejando a la mayoría de categorías con una muy baja representación; más del 90 % del dataset pertenece a 3 de las 9 clases de opinión política posible. Más aún, los usuarios rara vez apoyan positivamente los esfuerzos provenientes de un partido mayoritario, siendo la mayoría de estos tweets comentarios con poca opinión o críticas muy negativas contra alguno de los partidos, aunque rara vez contra ambos.

Esta baja representación de las otras seis clases de opinión política fue la razón principal para evaluar una versión simplificada o "reducida" del problema en conjunción con la versión "completa". En lugar de tener en cuenta todas las clases, el problema reducido solo considera tweets cuya opinión política es una de las tres clases con mayor representación: totalmente neutral, PP negativo/PSOE neutral y PP neutral/PSOE negativo.

En resumen, la tarea de determinar la opinión política de los tweets es realizada sobre dos versiones del mismo dataset, la completa y la reducida. Después del proceso de determinación de opinión política, toda evaluación se realiza mediante la comprobación directa de las opiniones políticas anotadas manualmente en el dataset y las inferidas por los procesos de clasificación.

4.4. Extracción de conocimiento a partir del contenido textual

Tratar con el contenido textual de los tweets difiere, en varios aspectos, del procesamiento de textos típico. Por un lado, hay que tener mucho cuidado durante el proceso de *tokenizado* pues los tweets suelen contener elementos especiales de gran relevancia y carga semántica como los *hashtags* y las menciones de usuario. Por otro lado, los tweets frecuentemente están "contaminados" con otros elementos que podrían calificarse como no relevantes y con muy poca carga semántica, tales como arte ASCII, numerales y ordinales, compuestos de fecha/hora y URLs. Estos elementos deben ser cuidadosamente detectados y eliminados sin alterar el resto de elementos relevantes del texto.

Por ello, el correcto procesamiento del contenido textual de los tweets es crucial para cualquier análisis posterior y durante el análisis del contenido textual. Los tweets han sido cuidadosamente procesados teniendo en cuenta los puntos anteriores, además de realizar un procesado mas tradicional como eliminación de palabras huecas y elementos de puntuación, usando el mismo proceso presentado en la sección 3.4.1.

4.4.1. Modelo Bag-of-Words estándar

El primer modelo de características propuesto en este paper es el conocido modelo *Bag-of-Words (BoW)*. Este modelo simplifica cada documento representándolo como una *bolsa* o multiconjunto de las palabras que lo componen, ignorando cualquier regla gramatical u ordenación de las palabras pero teniendo en cuenta la multiplicidad de las palabras dentro del documento. Cuando los documentos de este modelo se representa vectorialmente, se asemeja a una representación de histograma.

A pesar de su simplicidad, este modelo de características es ampliamente usado en diferentes aplicaciones las cuales requieren de un vector de características para entrenar clasificadores y los resultados obtenidos suelen ser adecuados. En el caso particular que se aborda, se considera que este modelo es aceptable debido a que la naturaleza del contenido textual de los tweets es breve y de muy baja complejidad gramatical.

Se ha evaluado el rendimiento del modelo sin variante alguna como una primera aproximación a la tarea, incluyendo varios *dummy baselines* por motivos puramente comparativos. Estos *dummy baselines* no son modelos propiamente dichos, sino estrategias extremadamente simplistas de clasificación que intentan comprobar la sensatez de los resultados obtenidos. En este caso, se han implementado el estimadores aleatorios (tanto uniforme como estratificado) y el estimador que clasifica siempre como la clase mas frecuente. La tabla 4.2 muestra los valores de rendimiento de la métrica *accuracy* (exactitud completa) con validación cruzada respecto a las dos versiones de la tarea propuesta, usando para ello un clasificador *Support Vector Machine (SVM)* con búsqueda de hiperparámetros. Todo el proceso de validación cruzada se ha realizado *estratificadamente* con $k = 10$ *folds* o pliegues, preservando la proporción de clases en todos los pliegues y minimizando la posible asimetría de clases.

Modelo de Características	Problema Completo	Problema Reducido
Dummy Aleatorio uniforme	12,73 %	34,86 %
Dummy Aleatorio estratificado	31,32 %	36,37 %
Dummy Clase mas frecuente	46,13 %	51,24 %
Bag-of-Words	61,97 %	68,36 %

Cuadro 4.2: Exactitud con validación cruzada del modelo Bag-of-Words tradicional.

A pesar de que el rendimiento del modelo BoW es bastante superior a los baselines presentados, los resultados no son demasiado buenos. El modelo BoW supera al estimador *Dummy Clase mas frecuente* en $\approx +15\,\%$ para el problema completo y en $\approx +17\,\%$ para el problema reducido, consiguiendo un éxito moderado (entre el 61 % y el 69 % de exactitud).

4.4.2. Modelo BoW con selección automática de características

Una manera muy popular para mejorar el modelo BoW consiste en aplicar un esquema de ponderación *TF-IDF* a los documentos en formato vectorial, otorgando mayor relevancia a ciertas palabras y mejorando el rendimiento general del modelo. Debido a la naturaleza del contenido textual de los tweets, al aplicar el esquema de ponderación *TF-IDF* sólo se obtiene una moderada mejora de los resultados (obteniendo un 70,42 % en la tarea completa y un 77,45 % en la tarea reducida).

Sin embargo, al probar otras técnicas para mejorar el modelo BoW se ha observado que el margen de mejora es mucho mayor que el ofrecido por el esquema *TF-IDF*. Esto se debe a que la coocurrencia a nivel de palabra dentro de los tweets es muy baja y la longitud de los tweets es demasiado pequeña en comparación al número de documentos existentes en el dataset. Cualquier sistema de ponderación basado en la distribución de palabras de los documentos esta sometida a esta limitación (siendo *TF-IDF* un ejemplo de este tipo de esquemas).

Un análisis superficial sobre los datos vectoriales del modelo BoW revelan un problema importante que está asociado al contenido textual. La dimensionalidad proporcionada por el modelo BoW es demasiado elevada ($\approx 3k$ características) mientras que el número de palabras/tweet es del orden de 10,41. Esto hace que la matriz de datos obtenida sea extremadamente dispersa, obteniendo una densidad de valores no nulos del $\approx 3.34 \times 10^{-3}$ y los clasificadores tengan muchas dificultades a la hora de entrenarse y, por consiguiente, obtienen un rendimiento pobre.

Para paliar esta situación, se aplica un proceso de reducción de dimensionalidad sobre el modelo BoW. En términos generales, existen dos grandes familias de técnicas para abordar dicha reducción de dimensionalidad: basadas en *dataset transformation* y basadas en *feature selection*.

Las basadas en *Dataset transformation (Transformación del dataset)* intentan reducir la dimensionalidad del modelo mediante transformaciones sobre el conjunto de datos, combinando dimensiones para lograr un menor número de ellas. Estas técnicas reducen el problema a uno de factorización matricial (*PCA*, *Kernel PCA*, *SVD* o *NNMF*) o realizan un proceso de *Manifold Learning* (*LLE*, *LTSA*, *IsoMap*, *Spectral Embedding* or *MDS*).

Aproximaciones basadas en problemas de factorización matricial intentan explicar la máxima varianza posible utilizando el menor número de componentes independientes aunque tienden a ignorar cualquier otra información que no se puede observar mediante la covarianza. Este tipo de aproximaciones son muy útiles si el dataset puede interpretarse de forma efectiva como la mezcla de señales interdependientes o si existen conjuntos de características con una alta correlación y se pueden transformar en dimensiones independientes.

Estas aproximaciones consiguieron resultados moderadamente satisfactorios, superando a la mejora obtenida mediante *TF-IDF* en el caso de la factorización mediante *NNMF* (obteniendo un 72,42 % en la tarea completa y un 78,91 % en la tarea reducida). Sin embargo, como veremos más adelante, las técnicas basadas en *feature selection* proporcionaron un rendimiento mucho mayor.

Por otro lado, las aproximaciones basadas en *Manifold Learning* (o *inferencia de variedades*) realizan una reducción no lineal sobre el conjunto de datos,

basándose en que la dimensionalidad existente en el dataset es artificialmente alta. Cada técnica en particular funciona de forma diferente, pero la idea principal radica en construir una representación de baja dimensionalidad usando una función de coste encargada de que los datos transformados retengan parte de las propiedades locales observada de los datos originales de mayor dimensionalidad. Este tipo de técnicas son útiles cuando se observa que los datos poseen una estructura claramente no lineal y que existe cierto grado de relación (no necesariamente lineal) entre las diferentes dimensiones.

Ahora bien, las aproximaciones basadas en inferencia de variedades no funcionaron correctamente sobre el modelo BoW; la mayoría de las aproximaciones obtuvieron resultados significativamente peores. Los mejores resultados se obtuvieron al aplicar *IsoMap*, consiguiendo un 57,45 % en la tarea completa y un 64,20 % en la tarea reducida. Estos resultados son claramente inferiores a los obtenidos mediante el modelo BoW sin transformar o modificar.

Aunque las técnicas de transformación del dataset suelen funcionar bastante bien y mejorar el rendimiento general, no todas obtuvieron resultados satisfactorios. Aunque en el caso de las técnicas de factorización matricial, se obtuvieron rendimientos superiores al modelo BoW, los resultados obtenidos fueron significativamente inferiores en comparación a los obtenidos mediante las técnicas basdas en *Feature Selection*.

Las técnicas basadas en *Feature Selection* (*Selección de características*) tienen una filosofía distinta, pues se basan en la suposición general de que los datos contienen características (dimensiones) que son redundantes o irrelevantes, provocando un claro perjuicio al proceso de clasificación. Estas técnicas encajan mejor con el modelo BoW pues, como hemos mencionado anteriormente, el alto grado de dispersión de los datos y el hecho de que muchas palabras no tienen por qué ser relevantes para la tarea hacen que muchas características no sean informativas.

Existen una gran cantidad de técnicas de selección de características, aunque la mayoría recurren a la *búsqueda en los espacios de características*, *métricas de filtrado* o *métodos embebidos*. Para el caso explorado en este capítulo, se comprobó que los métodos embebidos funcionaron mucho mejor que el resto, siendo dignos de mención los métodos basados *Norma L1* y las técnicas basadas en *Árboles de Decisión*.

La selección de características basada en *Norma L1* utilizan modelos lineales normalizados al espacio $L1$, dando lugar a soluciones dispersas con muchos coeficientes nulos. El proceso de selección consiste en elegir características con el mayor número de coeficientes no nulos. Sin embargo, los modelos lineales dependen mucho de los parámetros de regularización (parámetros que regulan la dispersión de los resultados) y no existe una regla o criterio general para elegir los mejores parámetros de regularización, haciendo que el proceso de selección de características sea muy sensible a la parametrización.

Los *árboles de decisión* pueden calcular valores de importancia sobre las características (mediante el cómputo de la entropía o la impureza Gini), siendo estos valores de importancia muy útiles a la hora de descartar características irrelevantes. Bosques de árboles de decisión con un alto número de estimadores tienden a recuperar con mucho éxito el conjunto de características significativas y no exhiben los mismos problemas de parametrización que los métodos basados en la *Norma L1*, siendo muy estables a la hora de su uso.

Finalmente, se definió un proceso de selección automática de característi-

cas para transformar el modelo BoW, haciendo uso de un *Bosque de árboles extremadamente aleatorizados* con un gran número de estimadores. Este tipo de bosque es similar al *Bosque de árboles aleatorizados* tradicional pero difiere en la manera de elegir los umbrales para cada subconjunto de datos aleatorios: los umbrales se eligen al azar dentro del conjunto de características candidatas y se eligen los mejores en lugar de simplemente buscar las características mas discriminativas. Esta variación permite reducir aún mas la varianza del modelo interno.

Modelo de Características	Problema Completo	Problema Reducido
Dummy Clase mas frecuente	46,13 %	51,24 %
Bag-of-Words	61,97 %	68,36 %
BoW-AFS	77,37 %	88,38 %

Cuadro 4.3: Exactitud con validación cruzada del modelo BoW con selección automática de características (AFS).

La tabla 4.3 muestra los valores de la métrica *accuracy* cuando se aplica la selección automática de características (AFS) al modelo BoW original justo antes de ser usado para entrenar el clasificador SVM. Con este paso de selección automática de características, el modelo BoW experimenta una enorme mejora de rendimiento ($\approx$ desde un +15 % hasta un +20 % respecto al modelo BoW simple) en ambas versiones del problema, llegando hasta un 88 % de exactitud en la versión reducida del problema. Se observa que con este paso de selección de características, el modelo BoW consigue un buen rendimiento y es apropiado como modelo de características para la clasificación de tweets a partir de contenido textual.

4.5. Extracción de conocimiento a partir del contenido estructural

La naturaleza estructural de los tweets es una fuente de conocimiento muy interesante y relevante, aunque este aspecto de los tweets es, con frecuencia, pasado por alto. Además, extraer conocimiento a partir de la información estructural encontrada en una colección de tweets es un proceso para nada trivial. A pesar del hecho de que existen constructos especiales dentro de los tweets que establecen algunas relaciones entre elementos, tales como las menciones de usuario o el uso explícito de *hashtags*, la red subyacente es muy rica y compleja, requiriendo de esfuerzo adicional y métodos específicos para abordar dicha complejidad.

Para realizar la extracción de dicho conocimiento, se opta por extraer la información topológica de la red subyacente, usando para ello todo un proceso de transformación que parte de la colección de tweets original para generar una representación de grafo bipartito, en la cual los nodos son los usuarios de la red y las aristas sus relaciones de amistad directas.

La figura 4.3 muestra, de forma esquemática, las diferentes etapas del proceso. Se parte de la colección de tweets, se genera un grafo de amistad, de éste

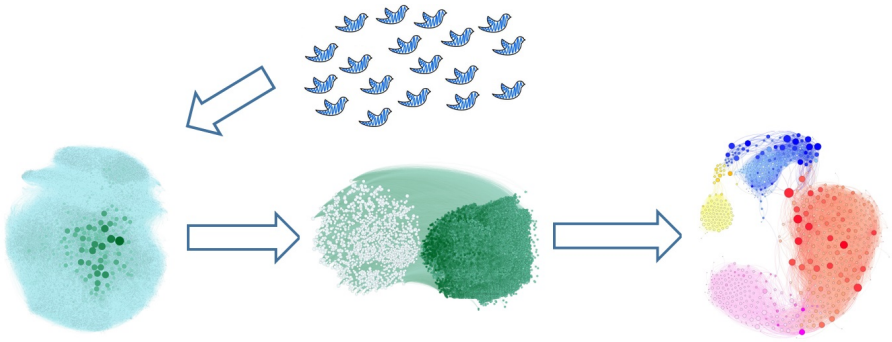


Figura 4.3: Esquema del proceso de extracción de conocimiento a partir del contenido estructural

un grafo bipartito y, finalmente, una representación espectral de mayor utilidad que depende de la técnica utilizada en cuestión. Todo el proceso de generación hasta el punto del grafo bipartito se encuentra detallado en la sección 5.4.1.

Dicho esto, se abordará el análisis de la información estructural extraída mediante dos aproximaciones diferentes pero claramente orientadas a interpretar la información topológica extraíble de una colección de tweets: una aproximación de *granularidad gruesa* basada en la aplicación de la bien conocida técnica *Louvain method* (Blondel et al., 2008) y otra aproximación de mayor poder expresivo basada en la aplicación de la técnica *Spectral Biclustering* (Kluger et al., 2003).

La primera aproximación de carácter topológico consiste en generar un modelo de características mediante el modelo de comunidades extraído por el *Método Louvain*. En líneas generales, se genera un grafo de similaridad y de él se obtiene un modelo de comunidades mediante la aplicación de este método. El proceso de generación del modelo de comunidades mediante el método Louvain se encuentra detallado en la sección 5.4.2, explicando cómo se genera el grafo de similaridad, qué métrica se ha usado y qué criterios se han utilizado para el proceso del grafo bipartito.

Usando este modelo de comunidad, se ha generado un modelo de características en el cual cada usuario se representa mediante un vector de afinidad sobre las comunidades resultantes. Para cada usuario que aparece en el dataset, se calcula la proporción de relaciones de amistad respecto a cada comunidad, obteniendo un vector cuya suma de valores es 1.

En la tabla 4.4 se muestra la exactitud de esta aproximación topológica, además de los valores de exactitud para las demás aproximaciones. Se observa un ligero incremento en el rendimiento al aplicar una etapa de selección automática de características pero es muy difícil mejorar más este modelo con este tipo de técnicas debido al bajo número de características que son usadas. Es muy interesante que una aproximación estructural de este tipo adquiera este poder predictivo, considerando que este modelo ignora por completo lo que los usuarios expresan en sus tweets, realizando predicciones simplemente mediante la estructura de red.

Análogamente al caso anterior, la segunda aproximación de carácter topológico consiste en generar un modelo de características mediante el modelo de

comunidades extraído por el *Spectral Biclustering*. Esta aproximación intenta paliar las deficiencias existentes en el modelo anterior, clasificando todos los nodos del grafo bipartito en conjunto, en lugar de hacerlo en diferido mediante un grafo de similaridad.

A grandes rasgos, la técnica consiste en generar una representación matricial del grafo bipartito, calcular los biclusters (comunidades de nodos creadores y consumidores de contenido; ver apartado 5.4.3) y generar un modelo de características a partir de estos biclusters usando una métrica de relevancia intracluster.

Para cada creador de contenido i y su bicluster más representativo b_i, se calcula su peso *intra-bicluster* $w_i = \frac{\sum_{j \in b_i} M_{i,j}}{|j \in b_i|}$, siendo éste la proporción de seguidores directos dentro del mismo bicluster asignado al creador de contenido i. Estos pesos representan la relevancia de los distintos creadores de contenido dentro de la comunidad representada por su bicluster.

Una vez generado el modelo con los pesos, para cada usuario que aparece en el dataset, se genera un modelo de características similar al usado en la aproximación anterior pero con una excepción: la medida de pertenencia a cada comunidad (bicluster) se hace mediante la suma de los pesos intra-cluster de los generadores de contenido que pertenecen a ese bicluster.

Modelo de Características	Problema Completo	Problema Reducido
Bag-of-Words	61,97 %	68,36 %
Afinidad Comunidades	50,07 %	56,04 %
Afinidad Comunidades-AFS	50,28 %	56,51 %
Relevancia Biclustering	59,97 %	68,75 %
Relevancia Biclustering-AFS	61,11 %	69,60 %

Cuadro 4.4: Exactitud con validación cruzada de las aproximaciones basadas en topología de red

En la tabla 4.4 se muestran los resultados obtenidos por este modelo de características usando el algoritmo de *Spectral Biclustering* para obtener las comunidades y la relevancia de los miembros dentro de esas comunidades. Esta aproximación funciona sustancialmente mejor que la aproximación estructural previa y obtiene resultados similar al modelo BoW, mostrando que la estructura de red subyacente es, por sí misma, muy útil. Merece la pena destacar que la etapa de selección automática de características también mejora ligeramente los resultados aunque al poseer un número reducido de características, el rendimiento no puede mejorar mucho más, situación que comparte con la aproximación topológica anterior.

4.6. Estrategias de combinación de características

En las secciones anteriores se han mostrado cómo se ha extraído conocimiento tanto del contenido estructural como del contenido textual de un tweet,

abordándose cada tipo de contenido de forma independiente y consiguiendo diferentes resultados. En esta sección se evalúa la idea de mezclar los mejores modelos de características provenientes de diferentes tipos de conocimiento, explorando varias formas de combinar dichos modelos para mejorar aún mas los resultados.

4.6.1. Combinación directa

La primera estrategia probada consistió en la combinación directa de ambos modelos de características en un modelo de mayor tamaño. La operación de combinación utilizada fue simplemente la concatenación directa de ambos modelos resultando en otro modelo vectorial de características. Además, se ha aplicado una etapa de selección de características de dos formas: *a priori*, combinando modelos ya previamente reducidos, y *a posteriori*, reduciendo el conjunto de características final después de combinar ambos modelos originales.

Modelo de Características	Problema Completo	Problema Reducido
BoW-AFS	77,37 %	88,38 %
Biclustering-AFS	61,11 %	69,60 %
Combinación	64,79 %	71,98 %
Combinación-AFS a posteriori	75,24 %	86,68 %
Combinación-AFS a priori	77,40 %	89,07 %

Cuadro 4.5: Exactitud con validación cruzada de la combinación directa de características

La tabla 4.5 muestra los valores de rendimiento obtenidos por las diferentes versiones de combinación directa. Esta estrategia de combinación tiene dificultades para obtener mejores resultados en la mayoría de las variantes, siendo la variante con modelos pre-reducidos la única que obtiene resultados realmente significativos. En el problema reducido, el resultado obtenido por esta variante es superior y en la versión completa, el resultado es similar al mejor modelo individual.

Se observa que los clasificadores suelen tener dificultades a la hora de abordar el modelo generado mediante concatenación directa. Los modelos de características incluidos representan diferentes tipos de conocimiento que confunden a los clasificadores pues estos modelos pueden presentar discrepancias al mismo nivel y estar en desacuerdo respecto a las clases inferidas.

4.6.2. Stacked Generalization

El método de combinación directa usado en el apartado anterior no obtuvo muy buenos resultados y está claro que se necesitó de otros esquemas de combinación que fueran capaces de manejar datos de diferente naturaleza.

Los métodos de *Ensemble learning* usan múltiples algoritmos de aprendizaje para obtener mejor rendimiento predictivo. Aunque existen varios de estos métodos que usan el mismo modelo de características y/o algoritmos, los métodos de *Ensemble learning* tienden a obtener muchos mejores resultados si existe

una diversidad significativa entre los modelos. Muchos métodos de *Ensemble learning* tienden a promocionar tal diversidad entre los modelos que combinan, pero los métodos que son interesante para este caso en cuestión son aquellos que permiten la combinación de diferentes modelos y algoritmos.

El método *Stacked generalization* o *Stacking* (Wolpert, 1992; Breiman, 1996) implica entrenar un meta-clasificador sobre las salidas o predicciones de varios otros clasificadores. La idea en cuestión radica en que el meta-clasificador aprende cómo de bien cada uno de los clasificadores son capaces de aprender los datos de entrenamiento.

Modelo de Características	Problema Completo	Problema Reducido
Combinación	64,79 %	71,98 %
Combinación-AFS a priori	77,40 %	89,07 %
Stacked Generalization	79,99 %	89,22 %

Cuadro 4.6: Exactitud con validación cruzada del método *Stacked Generalization*

La tabla 4.6 muestra la exactitud obtenida por el método de *Stacked Generalization*, mostrando que este método funciona significativamente mejor que estrategia de combinación directa utilizada en el apartado anterior. Experimentalmente, se determinó que el método funcionaba mejor cuando se combinaban el modelo BoW con AFS y el modelo de *Spectral Biclustering sin AFS*.

4.6.3. Multiple Pipeline Stacked Generalization

En la sección anterior se observa que la técnica de *Stacking* es capaz de combinar con éxito modelos de características generados de fuentes muy diferentes. Sin embargo, los clasificadores individuales usados en la técnica exhiben los mismos problemas de aprendizaje que en la combinación directa, pues dichos clasificadores usan todo el modelo de características completo de forma muy similar a la combinación directa. Aunque el meta-clasificador hace lo posible para mitigar esta situación y logra cierto éxito en ello, se ha ideado una variación sobre la técnica de *Stacking* que intenta abordar específicamente esta problemática.

El método en cuestión es llamado *Multiple Pipeline Stacked Generalzation* y es similar al *Stacking* tradicional pero es capaz de procesar cada modelo de características original en *pipelines* (flujos de procesado) independiente, en lugar de usar el conjunto de características completo en cada clasificador. Cada modelo de características es procesado de forma separada en cada *pipeline*, teniendo cada uno de ellos su propio conjunto de clasificadores individuales con sus parámetros, permitiendo potencialmente un mejor ajuste a cada modelo.

Un resultado directo de este procesamiento independiente es que el modelo de nivel 1 es significativamente mayor. Siendo N clases, M modelos y K_m clasificadores por modelo de características m, el modelo de nivel 1 tendrá $|N| \times \sum_{i \in M} |K_i|$ características, a diferencia del modelo de nivel 1 de la técnica *Stacking* tradicional que tiene $|N| \times |K|$ características ($|K|$ clasificadores independientes). El resto del proceso (etapa de aprendizaje en el meta-clasificador

y evaluación) es idéntico al modelo tradicional.

Cada *pipeline* de trabajo fue configurado usando los mejores clasificadores previamente probados en cada modelo de características por separado, teniendo en cuenta también las diferentes versiones del problema. Para la versión completa del problema, se observó experimentalmente que la combinación de SVM-C, Random Forests, Logistic Regression y Multinomial Naive Bayes funcionó bien para ambos modelos de características originales. Sin embargo, en la versión reducida del problema, la mejor combinación encontrada fue SVM-C, Random Forest y Logistic Regression para el modelo BoW, mientras que para el modelo de *Spectral Biclsutering* fue la combinación de SVM-C y Random Forest.

Modelo de Características	Problema Completo	Problema Reducido
Combinación-AFS a priori	77,35 %	89,07 %
Stacked Generalization	79,99 %	89,22 %
Multiple Pipeline Stacked Generalization	82,36 %	91,22 %

Cuadro 4.7: Exactitud con validación cruzada de la variante propuesta Multiple Pipeline Stacked Generalization

La tabla 4.7 muestra la exactitud de la variante propuesta *Multiple Pipeline Stacked Generalization* con los mismos modelos usados durante la evaluación de *Stacked Generalization*. Se observa que la variación propuesta obtiene un rendimiento significativamente mejor en ambas versiones del problema, siendo una manera mucho mas efectiva de combinar ambos modelos de características.

4.7. Conclusiones y trabajo futuro

A lo largo de este capítulo, se ha propuesto una aproximación a la tarea de categorización de tweets dentro del contexto político, basada en la idea de combinar nos fuentes de conocimiento diferentes: el contenido textual de los tweets y la información estructural que la red social subyacente ofrece. Este enfoque difiere del tradicional enfoque sobre el contenido textual que se puede encontrar con frecuencia en la literatura actual.

Partiendo de una colección de tweets generados usando el método de recuperación dinámica explicado en Cotelo et al. (2014) y expandido en el capítulo 2, se ha generado un dataset compuesto por 3000 tweets escritos en español que hacen referencia directa al gobierno actual (referente al periodo de la recolección del dataset) o su oposición, clasificando dichos tweets en opiniones Positiva, Negativa o Neutra respecto a cada partido, generando un total de nueve clases. La distribución de estas nueve clases de clasificación es bastante desequilibrada, por lo que se evalúa en paralelo una versión "reducida" del problema, en la cual simplemente se consideran las tres categorías mas frecuentes.

Después de preprocesar y tokenizar los tweets, se ha generado un modelo de características basado en el bastante establecido modelo *Bag-of-Words*. A pesar de su simplicidad, obtiene un éxito moderado y supera a todos los baselines propuestos. Se observó que la aplicación de un esquema de ponderado *TF-IDF* no mejora los resultados. Después de un análisis superficial, se observó

que este modelo sufría problemas de dimensionalidad, por lo que se le aplicó una etapa de selección de características basada en el uso de *Bosque de árboles extremadamente aleatorizados*. Con esta reducción de características, el modelo BoW experimenta un considerable aumento en el rendimiento, convirtiendo este modelo inicial en uno ligeramente apropiado para la clasificación de tweets a través de información textual.

El análisis de la información estructural posee un enfoque distinto al ser dicha información de muy distinta naturaleza. A partir de la idea de que los usuarios forman comunidades implícitas cuyos miembros tienden a compartir intereses comunes sin la necesidad de contacto directo, se ha ideado una estrategia basada en grafos para caracterizar dichos usuarios mediante el descubrimiento de estas comunidades implícitas.

Se construye un grafo de amistad bipartito que es usado como base para la generación de dos modelos de características estructurales diferentes. Uno de los modelos estructurales se genera a partir de las comunidades detectadas usando el *Louvain method* sobre un grafo de similaridad construido a partir del grafo de amistad y usando la medida *Dice* como medida de similaridad. EL otro modelo se construye mediante la aplicación de la técnica *Spectral Biclustering* sobre la matriz de amistad transformada y calculando un modelo de comunidades difuso y mas complejo. Aunque ambas aproximaciones estructurales son interesantes, el modelo de características basado en la técnica de *Spectral Biclustering* es claramente superior; obtiene resultados similares al modelo BoW mientras que los valores de rendimiento obtenido a través de la estrategia basada en el *Louvain method* son claramente inferiores.

Después de extraer conocimiento, se discute la idea de mezclar los mejores modelos de características extraídos de ambos tipos de contenido. Para mezclar estos modelos de características, se ha recurrido a métodos de combinación que, de alguna forma, son capaces de tratar con modelos de características provenientes de diferentes tipos de conocimiento. Se han probado tres esquemas de combinación diferentes: combinación directa, *Stacked Generalization* y una variante propia de *Stacked Generalization* llamada *Multiple Pipeline Stacked Generalzation*. Los resultados muestran que la variación propuesta funciona significativamente mejor que cualquier otro modelo de características independiente y esquema de combinación, siendo esta variación bastante efectiva a la hora de combinar modelos de características de diferentes tipos de contenido.

Se puede concluir que mezclar tanto conocimiento textual como estructural es una buena aproximación para determinar la orientación política de los tweets. Extraer conocimiento y generar buenos modelos de características es mucho más difícil para el contenido estructural que para el contenido textual, y hacer uso de ambos modelos de características resultó ser nada trivial, pues la combinación directa de los modelos no se comportaba de manera correcta debido a que los modelos eran de diferente naturaleza. Mas aún, se debe tener un cuidado especial a la hora de combinar modelos de ambos tipos de contenido. La aproximación mixta propuesta trata estos problemas cuidadosamente y ha demostrado ser efectiva y significativa.

Capítulo 5

Detección de comunidades de interés a través de las relaciones de amistad mediante Spectral Biclustering

Como ya se ha mencionado anteriormente, el análisis de los tweets es una tarea interesante y en el capítulo 4 se explora y aborda la tarea de caracterizar y categorizar los tweets. En este capítulo, en lugar de abordar la tarea de análisis con un enfoque individual o pormenorizado (tweet a tweet), se explora la idea de caracterizar colectivos de usuarios en función de intereses comunes, abordándose la tarea de la detección de comunidades implícitas y *ad-hoc* respecto a un tema en cuestión.

Como contribución de mayor peso, se propone una interesante y novedosa aproximación para descubrir estas comunidades implícitas que consiste en modelar la tarea de detección de comunidades como un problema de clasificación en bloque o *biclustering*. Además de la aproximación propuesta, se incluye como *baseline* exigente una aproximación adicional de carácter más convencional que se basa en modelar la tarea como un problema de maximización de la *modularidad*.

Después de presentar éstas aproximaciones, se establecen varios métodos de evaluación tanto extrínsecos como intrínsecos. Una vez analizada la viabilidad y el rendimiento de la contribución mediante los métodos de evaluación, se realiza un análisis de carácter empírico sobre una selección de usuarios extraídos de las comunidades inferidas mediante la técnica de *Spectral Biclustering*. Estos usuarios analizados tiene en común que son los usuarios más relevantes dentro de su comunidad pero no son organismos oficiales ni celebridades de nivel nacional.

Una vez llegado al final del capítulo, se realiza una discusión sobre los resultados mostrados y se concluye que la aproximación basada en la idea del *biclustering* es superior a la aproximación convencional basada en *modularidad*, ofreciendo un modelo más rico y consiguiendo mejores resultados.

5.1. Introducción

A lo largo del capítulo 4 se explora la tarea del análisis de los tweets desde un enfoque individual o pormenorizado. Sin embargo, en este capítulo se explora el caracterizar a los usuarios desde un enfoque a mayor escala, abordándose la tarea de detección de comunidades de usuarios en Twitter relacionados con el tema del contexto político español. La detección de tales comunidades orientadas políticamente da lugar a una serie de recursos que son un buen complemento a las fuentes de información más tradicionales usadas en el análisis político. El tamaño potencial del conjunto de datos inferido por este proceso es varios órdenes de magnitud mayor y suele cubrir a otros sectores demográficos que las encuestas y otros métodos más tradicionales no son capaces de cubrir.

En este capítulo, se propone una aproximación no supervisada para revelar estas comunidades *ad-hoc* e implícitas de usuarios que comparten posturas políticas similares, a pesar de que, con frecuencia, tales usuarios no interactúan explícitamente con otros miembros de esta comunidad ni declaran directamente su ideología política en sus perfiles o en sus tweets. Mediante el uso de un grafo de amistad directo, se propone una aproximación basada en modelar la tarea como un problema de biclustering para este proceso de descubrimiento no supervisado de comunidades.

Por otra parte, no sólo se describe la aproximación propuesta, sino que también se contrasta el rendimiento de la aproximación descrita respecto a una aproximación de carácter más convencional. Se ha seleccionado un conocido y eficiente método de *maximización de la modularidad*(Newman and Girvan (2004), Newman (2006); ver 5.4.2) como punto de comparación exigente pues este método es capaz de procesar redes muy grandes y es bastante bueno a la hora de detectar comunidades de gran escala dentro de las redes. Los resultados de evaluación muestran que la aproximación propuesta es superior al *baseline* contrastado.

Este capítulo se organiza tal y como sigue a continuación: En la sección 5.2 (*Trabajos relacionados*) se realiza una revisión del estado del arte actual relacionado con la temática del capítulo y se analizan algunos trabajos de mayor interés.

En la sección 5.3 (*Definición de la tarea*) se define la tarea abordada en este capítulo, especificando tanto el objetivo principal como la colección de datos (dataset) analizada.

En la sección 5.4 (*Detección de comunidades usando Spectral Biclustering*) se describen las dos aproximaciones estudiadas en este capítulo. La primera consiste en un método basado en la maximización de la modularidad, el cual es propuesto como *baseline* exigente, mientras que el segundo método es la novedosa contribución basada en *biclustering*.

En la sección 5.5 (*Evaluación de las aproximaciones*), se realiza una evaluación de ambas propuestas mediante métodos de evaluación intrínsecos y extrínsecos. Además, se incluye una análisis de carácter empírico sobre una selección de los usuarios no famosos ni oficiales que son más relevantes a partir de una lista de dichos usuarios extraída de las comunidades inferidas por la aproximación basada en biclustering.

Finalmente, en la sección 5.6 (*Conclusiones y trabajo futuro*) se hace un resumen de los esfuerzos y se revisan tanto los resultados como las principales contribuciones expuestas.

5.2. Trabajos relacionados

La detección de comunidades on-line ha sido investigada en un buen número de obras desde diferentes puntos de vista, cada uno de ellos intentando determinar o inferir la estructura de una red dada con el fin de agrupar nodos de la red en comunidades. Los métodos con mayor representación dentro de la bibliografía son aquellos basados en el análisis enlaces, mientras que, a medida que se cambia el paradigma y se incluyen elementos estocásticos, vamos encontrando métodos que hibridan la información estructural con información estadística, hasta el punto de encontrar métodos totalmente estocásticos.

Por ello, de los diferentes tipos de aproximaciones disponibles en la bibliografía, se analizan propuestas interesantes para la detección de comunidades respecto a tres grandes grupos: métodos basados en análisis de enlaces, métodos que incluyen cierta información estadística a un modelo estructural y métodos bayesianos generativos totalmente estócásticos.

5.2.1. Métodos basados en análisis de enlaces

Los métodos de análisis basados en enlaces son los más comunes, pues son fácilmente aplicables sobre una representación de grafo cualquiera y su funcionamiento es bastante intuitivo. Dado que, con frecuencia, las redes sociales proporcionan información estructural de forma relativamente explícita, es sencillo transportar el contexto analizado a una representación de grafo cuyos nodos son usuarios.

Por ejemplo, en el trabajo Tseng (2005) se propone una aproximación que usa el algoritmo *HITS* (Kleinberg, 1999) para realizar un análisis basado en enlaces. Esta obra describe un sistema de detección de comunidades sobre la *blogosfera* que realiza un recorrido *random-walk* sobre el grafo de la red para detectar las comunidades formadas por los principales blogs del ranking.

Dicho grafo se construye usando blogs como nodos y conectándolos entre sí en función de sus respectivas citas y aplicando varias iteraciones del algoritmo HITS para calcular las puntuaciones de los nodos. Merece la pena destacar que la propuesta encontrada en Gibson et al. (1998) comparte la misma idea de usar el algoritmo HITS para inferir las comunidades de usuarios.

En otros trabajos como Lim and Datta (2012b,a), los autores abordan explícitamente la detección de comunidades de usuarios en una red social, siendo en este caso la red Twitter. Partiendo de la suposición de que los usuarios con intereses similares deberían seguir a los mismos usuarios famosos y celebridades, definen un método aglomerativo usando solo enlaces de carácter topológico para detectar las comunidades mediante el uso del método *Clique Percolation* (Derényi et al., 2005) y el algoritmo *Infomap* (Rosvall and Bergstrom, 2008). Este método, al basarse en el análisis de elementos altamente seguidos como las celebridades y usuarios famosos, está enfocado al análisis de fenómenos como el marketing viral, no siendo esta aproximación relevante para detectar comunidades que aborden otra temática.

Posteriormente, los autores proponen una extensión algo más general que intenta detectar comunidades de usuarios con un alto grado de interacción entre ellos. Esta extensión, llamada *Highly Interactive Community Detection (HICD)*, incluye el uso de enlaces implícitos como menciones entre usuarios y retweets, relaciones más sutiles que la de seguir a otro usuario.

Sin embargo, este método es incapaz de agrupar usuarios cuyos intereses latentes sean iguales pero no tengan alta interacción, algo bastante común en redes como Twitter: la mayoría de los usuarios siguen a otros relevantes que comparten ideas, pero no suelen establecerse relaciones entre usuarios no conocidos simplemente por el hecho de seguir a gente similar.

En el trabajo Yang and Leskovec (2015), los autores proponen un método de detección de comunidades basándose en lo que ellos definen como *Ground-truth communities*. Estas comunidades son comunidades funcionales explícitamente etiquetadas cuyos miembros comparten el mismo rol, afiliación o cualquier otro atributo. Bajo esta noción, los autores analizan varias definiciones estructurales de lo que es una comunidad dentro de una red, dando pié a diversas aproximaciones basadas en grafos y enlaces bajo cada una de estas definiciones.

En consecuencia, los autores analizan diferentes las diferentes definiciones, categorizando éstas en cuatro grandes grupos y encontrando que dos de estos grupos son los que mejor rendimiento obtienen en cuanto a la identificación de comunidades *ground-truth*. Usando estos resultados, los autores extienden un algoritmo de clustering local a un método heurístico de detección de comunidades libre de parámetros que es capaz de escalar fácilmente a redes de mayor tamaño.

5.2.2. Clustering de comunidades basado en significancia estadística

La mayoría de las técnicas de detección de comunidades realizan una asignación unitaria de usuarios a comunidades sobre grafos no dirigidos sin importar la direccionalidad de las relaciones o la calidad de los clusters obtenidos.

Aunque varios algoritmos, como los basados en modularidad, buscan optimizar una medida de bondad para buscar los clusters, es raro que éstas realicen algún tipo de refinado a posteriori u optimización local para mejorar la calidad o la significancia de los clusters.

En este apartado se presenta una aproximación interesante, llamada *OSLOM (Order Statistics Local Optimization Method)* y descrita en la obra Lancichinetti et al. (2010), cuyo objetivo es abordar grafos dirigidos, ponderados y que permite cierto grado de solapamiento y jerarquía en las comunidades detectadas, mediante el análisis de la significancia estadística de los clusters.

Primero, los autores definen una forma para medir la significancia estadística de un cluster dado por su algoritmo, la cual se define como la probabilidad de encontrar dicho cluster en un modelo aleatorio nulo, una clase de grafos que no presentan ninguna estructura de comunidad. Para ello, utilizan el modelo de configuración descrito en Molloy and Reed (1995) como modelo nulo, el cual está pensado para generar redes aleatorias dada una distribución de grado (número de vecinos dado un vértice). Este modelo es básicamente el mismo que el modelo nulo utilizado para definir la modularidad (Newman and Girvan, 2004; Newman, 2006).

La figura 5.1 muestra la situación de análisis de un subgrafo de ejemplo C dentro del modelo nulo al incluir un vértice i. En este caso, se analizaría la probabilidad de que i tenga un número deseado de vecinos en el subgrafo C al incluirse, sabiendo que el modelo nulo posee un grado interno prefijado. Con este tipo de medidas, se puede estimar la significancia de añadir ciertos nodos

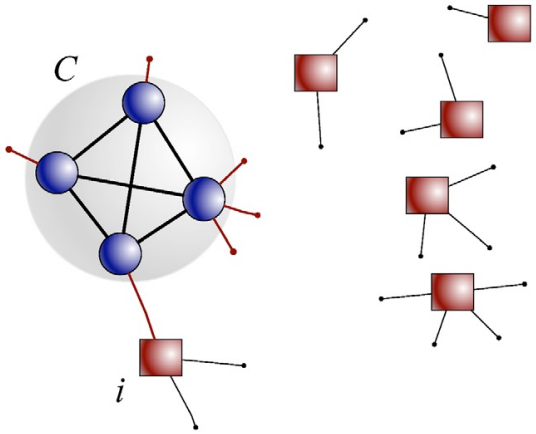


Figura 5.1: Esquema de evaluación al incluir el vértice i al subgrafo C en el modelo nulo

del grafo original al grafo nulo y cómo contribuyen al modelo nulo de forma no aleatoria, indicando que los nodos añadidos incluyen información relevante.

Una vez establecido el concepto de significancia estadística, el siguiente paso que establecen los autores es el de optimizar la puntuación de significancia estadística a través de la red al dividirla en clusters. Para ello, primero determinan como se optimiza la puntuación de un cluster dado: primero se explora la posibilidad de añadir vértices extras al subgrafo C (que representa el cluster) y después se refina mediante la eliminación de vertices no significativos.

El método OSLOM final sigue el siguiente esquema de tres fases:

- Se realiza una búsqueda de clusters significativos hasta que alcanza convergencia; para ello se utiliza un parámetro de puntuación mínima.
- Se analiza el conjunto resultante de clusters, intentando detectar su estructura interna para posibles uniones de clusters.
- Detecta una posible estructura jerárquica entre los clusters, proporcionando información adicional.

Dada que la naturaleza del sistema es de inclusión y refinamiento, para acelerar el proceso no se suele partir desde cero, sino que se puede iniciar el proceso con información adicional o con una partición generada por otro algoritmo de clustering.

La figura 5.2 muestra el funcionamiento general del método propuesto, donde se observa la naturaleza iterativa del mismo, con el fín de contrarrestar el carácter estocástico del método de análisis y convergencia.

El método es bastante versátil, siendo utilizando en la obra Greene et al. (2012) para detectar comunidades temáticas en Twitter a partir de listas de usuarios. En dicha obra, se construye un grafo de similaridad cuyos nodos representan las listas de usuarios y se establecen aristas entre nodos si dichas listas tienen usuarios en común.

Dadas los listas L_1, L_2 y el conjunto de elementos comunes $C = |L_1 \cap L_2|$, se utiliza como medida de similaridad la significancia estadística de la proba-

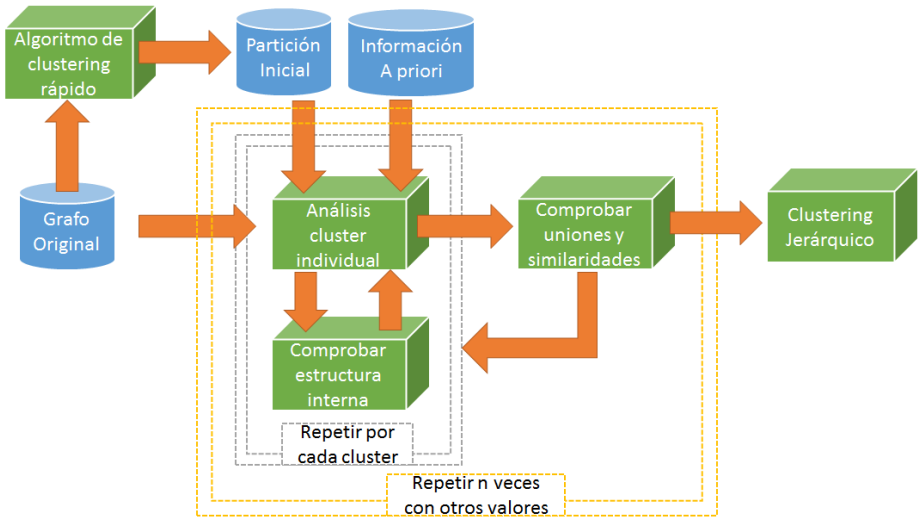


Figura 5.2: Funcionamiento del método iterativo de detección de comunidades OSLOM

bilidad de observar, al menos, $|C|$ elementos de L_1 en una lista cualquiera de tamaño $|L_2|$. Las aristas cuya medida de similaridad se inferior a un valor umbral especirficado son descartadas.

Después, se aplica el método OSLOM sobre el grafo de similaridad, obteniendo un clustering sobre las listas de usuarios, cada conjunto de listas de usuario siendo un tema de interés en potencia.

Los autores realizan una batería de pruebas con redes tanto artificiales como reales, y aunque los resultados son bastante buenos en términos generales, el tiempo de computación del método es muy alto, creciendo prohibitivamente respecto al número de vértices a considerar para incluir o eliminar de un cluster. Este método no es, por sí solo, factible para redes grandes.

Los autores exploran el usar el método OSLOM como refinamiento de una partición anterior, logrando así solventar parcialmente este problema para usar el método en redes grandes. Es obvio que el resultado final está bastante condicionado por la técnica inicial usada y en la experimentación mostrada, usan como punto de partida las particiones generadas por el *método Louvain* (Blondel et al., 2008). Como nota final, los propios autores mencionan que necesitan otro método para seleccionar los vertices en el proceso de análisis individual de clusters.

5.2.3. Modelos generativos

La mayoría de las aproximaciones para la detección de comunidades se basan en aplicar alguna técnica de análisis de redes para generar un clustering, ya sea mediante clustering aglomerativo, particionado de corte mínimo, medidas de centralidad y métodos similares.

Estos métodos, aunque consiguen buenos resultados en muchos casos, suelen realizar un particionado directo de las comunidades de usuarios y no tratan todos los enlaces de la red por igual, no haciendo distinción entre los tipos de

relaciones o diferentes interacciones en una red social.

Esta situación es una de las principales motivaciones que tienen los autores de la obra Sachan et al. (2012), en la cual implementan un modelo bayesiano generativo llamado *TUCM (Topic User Community Model)*, cuyo objeto es el de descubrir comunidades de usuarios latentes.

Este modelo, que consiste en una variación del bien conocido modelo generativo usado para descubrir diferentes temáticas en un conjunto de textos escritos *Latent Dirichlet Allocation (LDA)* (Blei et al., 2002), parte de la suposición de que los usuarios que interactúan más frecuentemente entre sí tienden a pertenecer algún tipo de comunidad latente común aunque no sea explícita.

Es más, el modelo intenta describir de forma conjunta a los usuarios y los temas explorados mediante los mensajes y sus diferentes tipos, permitiendo que un usuario pueda pertenecer a varias comunidades y estar interesado en diferentes temas, aunque no se determinan grados de pertenencia o interés respecto a las comunidades y los temas, respectivamente.

La idea base radica en representar a los usuarios mediante las interacciones de dichos usuarios dentro de la red social, determinando que la distribución de interacciones de un usuario se representa como una mezcla aleatoria de las variables latentes de las comunidades.

De forma general, el mensaje de un usuario viene determinado por dos elementos: la comunidad y la temática. Dado un usuario, la temática de la que quiere hablar y la comunidad asociada a la temática, la comunidad determina el tipo de interacción (tweet simple, retweet o respuesta) y la temática determina el conjunto de palabras usadas en ese mensaje.

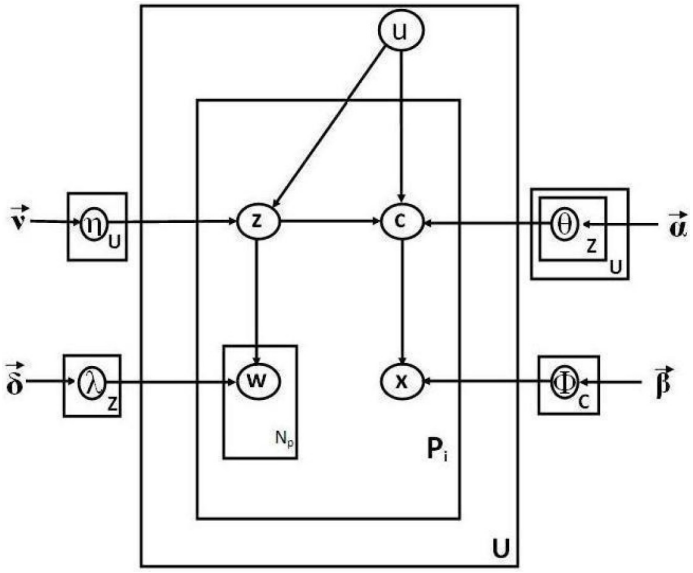


Figura 5.3: Representación gráfica del modelo TUCM

La figura 5.3 muestra el proceso generativo completo incluyendo las variables latentes y parámetros. Es interesante destacar que, dado que el cómputo de las probabilidades exactas de las distribuciones posteriores sobre el conjunto completo de hiperparámetros es un problema intratable, los autores utilizando

inferencia aproximada usando un método basado en *Gibbs sampling*. La descripción completa del proceso generativo específico se encuentra en la obra original Sachan et al. (2012), pues describir cómo funciona un proceso bayesiano completo queda fuera del alcance de esta memoria de tesis.

En la obra se determina que los resultados obtenidos por el modelo TUCM eran apropiados para abordar una red con capacidades de *broadcast* como Twitter y el rendimiento computacional de TUCM respecto a otros modelos similares es superior debido a que TUCM escala mucho mejor.

Sin embargo, el modelo TUCM requiere de cierta parametrización muy dependiente del contexto, pues como en LDA, hay que establecer el número de comunidades y temas a priori. Esto implica que cada modelo generado no sólo es extremadamente dependiente de los datos sino de la interpretación del responsable del experimento, requiriendo de un buen juicio a la hora de determinar estos parámetros para poder obtener un rendimiento apropiado.

5.3. Definición de la tarea

Una fuente de conocimiento interesante aparece una vez que se consideran los usuarios como elementos estructurales de una red en lugar de individuales de una red social. Estos usuarios establecen relaciones entre ellos, tanto explícitas como implícitas, dando lugar a la creación orgánica e implícita de grupos usuarios en las cuales los miembros del grupo comparten intereses comunes a pesar de que la mayoría de los usuarios del grupo no se conocen directamente o tienen algún tipo de contacto directo.

A raíz de esta observación, podemos definir como *comunidad* a un grupo de usuarios cuyos miembros comparten intereses comúnes respecto a una temática, con independencia a la interacción directa entre sus miembros o el reconocimiento expreso de pertenencia a la comunidad por parte de los mismos.

Detectar comunidades de usuarios dentro de una red dada es una tarea muy amplia. En este capítulo se aborda la tarea de detectar tales comunidades de usuarios dentro de la red Twitter, pero en lugar de extraer comunidades genéricas a gran escala, la detección se centra en detectar comunidades subyacentes respecto a un tema específico. Estas comunidades de granularidad fina son muy interesantes porque son capaces de revelar posturas respecto al tema abordado en cuestión y ayudan a identificar qué usuarios apoyan tales posturas.

Para esta obra en cuestión, se analiza la situación política española mediante la detección de comunidades de usuarios españoles cuyos mensajes guarden relación con el contexto político español. Sin embargo, no nos es factible recuperar y analizar la red compuesta por el conjunto completo de usuarios españoles debido a restricciones de carácter técnico; el tamaño de la red sería enorme y Twitter limita la cantidad total de información recuperada. Por lo tanto, se opta por analizar una muestra de los tweets pertinentes usando un método de generación de consultas que proporciona una cobertura mayor en comparación a hacer consultas simples sobre el sistema que proporciona Twitter.

Se construye una colección de tweets escritos en español durante la presentación del borrador final de las enmiendas a realizar sobre la ley que regula el aborto en España, conocida coloquialmente como *Ley del Aborto*. El periodo de la presentación en cuestión está comprendido entre el 20 de diciembre del 2013 y el 23 de diciembre del 2013. La reforma propuesta causó un gran impacto sobre

la población española y todos los partidos políticos mayoritarios se posicionaron de forma activa respecto a esta cuestión. El método en cuestión usado para recuperar estos tweets es método de recuperación dinámica explicado en Cotelo et al. (2014) y en el capítulo 2. Como ya se ha mencionado anteriormente, este método garantiza un alto volumen de tweets, introduce poco ruido y es capaz de reaccionar ante eventos imprevistos relacionados con la temática durante el periodo de recuperación de datos.

Al igual que en otros países, y como se menciona en el capítulo 4, la situación política española se encontraba claramente dominada por un puñado de fuerzas políticas, en particular por dos partidos políticos mayoritarios: el conservador, liberal y cristiano demócrata *Partido Popular (PP)* y el social demócrata *Partido Socialista Obrero Español (PSOE)*. Desde la transición española a la democracia, estos partidos son los únicos que han tomado gobierno en el país y se han seleccionado términos relacionados a esos partidos como conjunto semilla para el método de recuperación. Se han recuperado un total de 20251 tweets que involucran de forma directa a más de $180k$ usuarios.

A partir de esta colección de tweets, se extraen los usuarios que aparecen directamente en el contenido de cualquier tweet, ya sea el autor o cualquier otro usuario mencionado en el tweet. En complemento a los usuarios que aparecen, se recopilan las listas de amigos directos (aquellos a quien siguen), obteniendo de forma efectiva sus vecinos más inmediatos. Partiendo de la lista conjunta de usuarios (autores y mencionados) y sus respectivos amigos, se compone un grafo de amistad directa con miles de nodos (usuarios) y millones de aristas (relaciones de amistad directa).

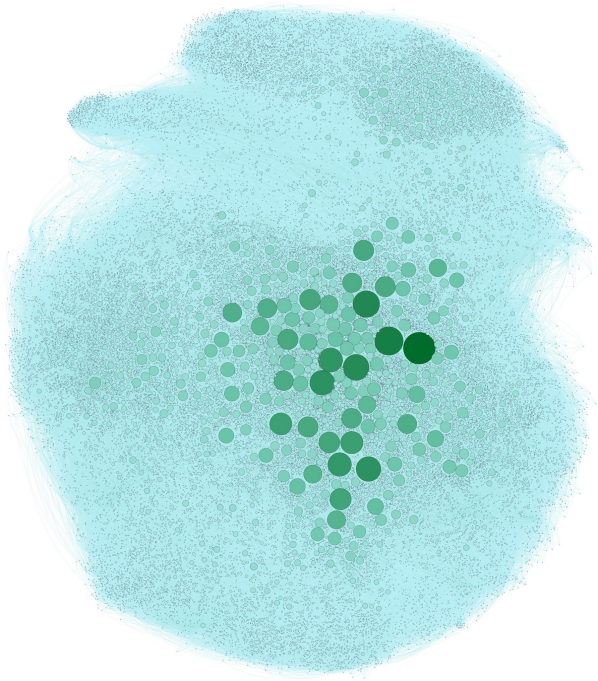


Figura 5.4: Grafo de amistad directa generado a partir de la colección de tweets obtenida

En la figura 5.4 se muestra el grafo de amistad resultante, el cual contiene más de $500k$ usuarios y más de $1,1M$ relaciones de amistad. Por motivos de claridad, los nodos son visualmente escalados y coloreados por su *grado* de entrada (número de relaciones de amistad entrantes).

Este grafo de amistad alberga usuarios que son de naturaleza muy variada; la mayoría de los usuarios son cuentas oficiales de medios de comunicación, gente políticamente activa, miembros oficiales de partidos políticos y gente corriente sin relación aparente con las organizaciones políticas, cubriendo la mayoría de los actores de contexto sociopolítico.

Este grafo de amistad de carácter político es una buena fuente de partida para la detección de comunidades de interés en Twitter respecto a la situación política española y por ello, es usado como base para el resto del trabajo presentado en este capítulo.

5.4. Detección de comunidades usando Spectral Biclustering

Identificar la estructura de comunidad subyacente que existe en una red de usuarios puede ser una tarea difícil y computacionalmente costosa. En esta sección se propone una aproximación basada en *biclustering* para detectar las comunidades de interés dentro de Twitter respecto a una temática específica.

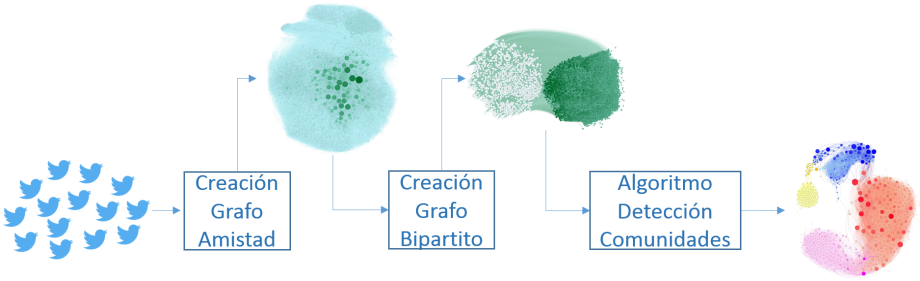


Figura 5.5: Esquema general del proceso de detección de comunidades.

La figura 5.5 consiste en un esquema general sobre el proceso utilizado para la detección de comunidades, consistente en varios pasos: creación del un grafo de amistad directa a partir de los tweets, transformación del grafo de amistad a un grafo bipartito y aplicación de una técnica de detección de comunidades usando como recurso base dicho grafo bipartito.

En primer lugar, se detalla el proceso de construcción del grafo bipartito que va a ser utilizado como fuente de conocimiento para las aproximaciones presentadas en esta sección. Este grafo bipartito expresa una ordenación topológica sobre los usuarios de gran interés y utilidad, y se construye a partir del grafo de amistad mencionado en la sección 5.3.

En segundo lugar, antes de entrar en detalle sobre la aproximación propuesta basada en biclustering, se describe una aproximación bastante más convencional que es incluida y usada como punto de comparación exigente y de relevancia. Esta aproximación consiste en modelar la tarea en cuestión como un problema de *Maximización de la Modularidad* y aplicar el conocido *Método Louvain*

(Blondel et al., 2008). La inclusión de esta aproximación específica radica en que las técnicas de maximización de la modularidad son apropiadas para la detección y descubrimiento de comunidades a gran escala, siendo el Método Louvain indicado por su eficiencia para abordar redes de gran tamaño.

En tercer lugar, se detalla la novedosa aproximación basada en biclustering para la extracción de la estructura de comunidad a partir del grafo de amistad mencionado anteriormente. La idea radica en modelar el proceso de detección de comunidades como una tarea de *biclustering* y hacer uso del algoritmo *Spectral Biclustering* (Kluger et al., 2003) para calcular los correspondientes biclusters, los cuales serán usados para componer las comunidades.

5.4.1. Creación del grafo bipartito

Para lograr detectar comunidades implícitas y poder caracterizar a los usuarios a través de dichas comunidades, se ha desarrollado una aproximación basada en grafos. Esta aproximación se basa en la idea de que se puede establecer y calcular algún tipo de medida de similaridad entre los usuarios si se analizan sus relaciones de amistad, siendo este análisis la base para identificar las comunidades de usuarios que comparten intereses comunes.

El primer paso consiste en la construcción de un grafo dirigido de amistad directa utilizando todos los usuarios que aparecen en la colección de tweets y sus respectivos amigos (usuarios a los cuales ellos siguen de forma directa). Este grafo resultante puede llegar a ser enorme y muy difícil de tratar computacionalmente hablando. El grafo de amistad utilizado a lo largo de este capítulo es el que se menciona al final de la sección 5.3.

Sin embargo, muchos de los nodos (usuarios) sólo tienen una arista entrante y ninguna saliente, por lo que no aportan mucho al grafo en sí y se consideran nodos irrelevantes. Por lo tanto, este grafo de amistad en bruto es podado, eliminando nodos irrelevantes que posean un bajo número de aristas entrantes y ninguno saliente. En lugar de establecer el umbral manualmente, se ha elegido el valor correspondiente al percentil $q = 95$ sobre la distibución del número de aristas entrantes, determinando el grado de aristas entrantes para nodos sin aristas salientes debe encontrarse dentro del 5 % superior de la distribución.

Este grafo de amistad podado posee dos tipos de usuarios: usuarios originales que son autores de algún tweet en el dataset original y los nuevos usuarios detectados a través de las relaciones de amistad existentes en la red. Después de un análisis superficial, se detecta que la mayoría de los usuarios originales son principalmente *consumidores de contenido* mientras que los nuevos usuarios detectados son, principalmente, *generadores de contenido*. Los nodos denominados generadores de contenido son elementos cuya función principal en la red es la de generar contenido potencialmente relevante y siendo, con frecuencia, fuentes de opinión politica . Estos usuarios corresponden normalmente a cuentas de medios de comunicación relevantes, usuarios políticamente activos y cuentas oficiales de los partidos políticos; estos usuarios suelen tener un gran número de seguidores, aunque no es una condición necesaria. Los consumidores de contenido forman el resto de los nodos de la red y estos usuarios "consumen" los contenidos ofrecidos por los generadores de contenido, recibiendo dicho contenido porque los consumidores siguen a dichos generadores. Es interesante ver que estos roles no son mutuamente excluyentes.

A partir del grafo de amistad, se construye un grafo bipartito que exprese este

comportamiento de creadores y consumidores, teniendo en cuenta que algunos nodos pueden, potencialmente, tener ambos roles y poseer tanto aristas entrantes como salientes. Estos nodos que exhiben los dos roles son transformados a dos nodos distintos, cada uno de ellos teniendo solo aristas entrantes (creador de contenido) o aristas salientes (consumidor de contenido). Cualquier nodo que se encuentre aislado después del proceso de podado o del proceso de transformación a grafo bipartito, es descartado y eliminado del grafo.

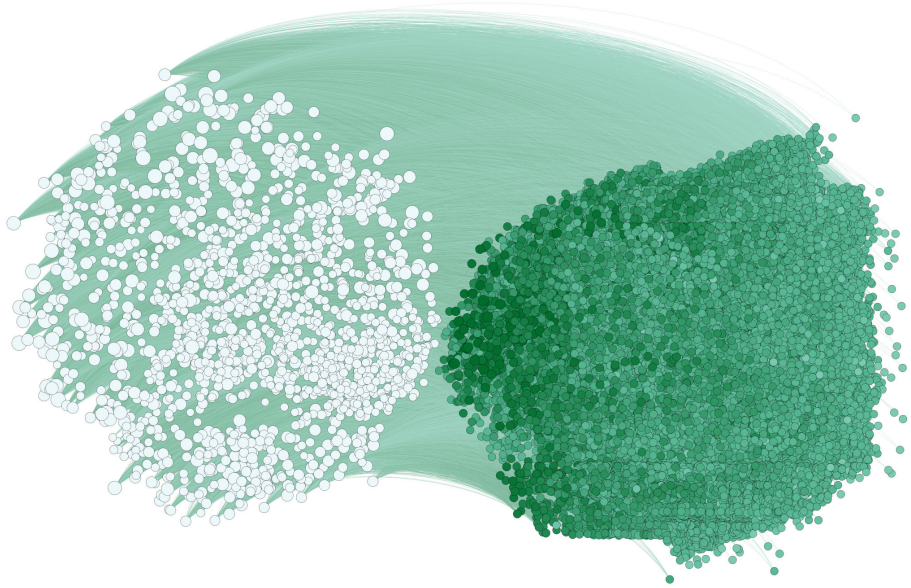


Figura 5.6: Grafo bipartito construido a partir del grafo de amistad directa.

La figura 5.6 muestra el grafo bipartito resultante donde los nodos son coloreados por grado de entrada y distribuidos por su clase bipartita (los consumidores de contenido se distribuyen hacia a la derecha mientras que los creadores de contenido son distribuidos a la izquierda).

5.4.2. El Método Louvain: un baseline exigente

La *Modularidad* (Newman and Girvan, 2004; Newman, 2006) es una función de recompensa o beneficio[1] diseñada para medir la calidad de una división de la red en comunidades en concreto. Una partición de la red con un alto grado de modularidad exhibe una red de interconexiones intra-comunidad muy densa y poca densidad de conexiones extra-comunitarias. En esencia, cualquier método de partición basado en modularidad intenta encontrar una partición de la red en comunidades que maximicen la función de modularidad a nivel global.

No obstante, realizar una búsqueda de fuerza bruta sobre el espacio de todas las particiones posibles de una red es intratable para la mayoría de los casos excepto los más triviales, debido a que el espacio de búsqueda crece exponencialmente con el tamaño de la red; se trata de un problema *NP-duro*. Por lo tanto,

[1] Una función de recompensa puede verse como una función objetivo a maximizar en lugar de minimizar, como en una función de coste o pérdida

la mayoría de los algoritmos recurren a métodos de optimización aproximada para buscar una partición, haciendo uso de algoritmos como los *algoritmos voraces*, métodos *quasi-Newton* o métodos de *optimización espectral*, consiguiendo soluciones aproximadas al problema de maximización de la modularidad con diferentes grados de velocidad y precisión.

Para este trabajo en concreto, se usa el popular *Método Louvain* (Blondel et al., 2008) para aproximar la función de modularidad. Este método iterativo optimiza la modularidad de las comunidades localmente hasta que no se consigue mejorar la modularidad global mediante perturbaciones del estado de la partición actual. Este método suele ser mejor que otros métodos en términos de tiempo de computación, lo cual permite el análisis de redes de gran tamaño en un tiempo razonable y consigue modelos muy precisos a la hora de abordar comunidades *ad-hoc* con una estructura de comunidades conocida.

El siguiente paso consiste en generar una matriz de similaridad sobre los nodos creadores de contenidos a partir de la información del grafo bipartito. La idea radica en que dos nodos creadores de contenido son similares si tienen consumidores de contenido en común. Se ha elegido la medida *Dice* como una medida de similaridad apropiada para comparar nodos creadores mediante sus conjuntos de consumidores. Usando esta medida, se construye un grafo ponderado de similaridad entre los nodos creadores de contenido cuyo peso de las aristas es igual a la coeficiente Dice entre los conjuntos de nodos consumidores de contenido que siguen a dichos nodos creadores.

Como detalle técnico, cualquier grafo de similaridad construido de esta forma es un *grafo completo*, significando que el grafo tiene $\frac{n(n-1)}{2}$ aristas, no mostrando de forma evidente alguna característica de carácter topológico y la mayoría de los algoritmos de detección de comunidades pueden tener dificultades cuando son aplicados a los grafos completos. Mas aún, la mayoría de las aristas tendrán pesos nulos o valores muy cercanos a cero, indicando que los esos nodos poseen casi ninguna similaridad.

Por ello, este grafo de similaridad ha sido procesado para revelar algún tipo de estructura oculta. Cualquier arista que represente un valor de similaridad bajo se considera como no informativa, así que se eliminan todas las aristas con un peso menor a cierto valor umbral especificado. Para este caso en concreto, se determinó que pesos inferiores a 0,35 indicaban una similaridad muy baja entre nodos creadores de contenido. Este umbral es considerado como apropiado pero no demasiado estricto, pues los valores de similaridad, al haber sido compuatados mediante la medida Dice, se encuentran en el rango de valores comprendido en el intervalo $[0, 1]$.

Al procesar este grafo de similaridad de la forma anteriormente descrita, el grafo resultante puede no seguir siendo un grafo conexo y las diferentes componentes conexas deben ser inspeccionadas individualmente aunque, en la mayoría de los casos, la componente conexa con el mayor número de nodos es la única relevante. Después de seleccionar las componentes conexas y elegir las relevantes, se ha aplicado el conocido método de detección de comunidades *Louvain method* (Blondel et al., 2008) al grafo de similaridad, asignando a cada nodo creador de contenido una comunidad en concreto. A los nodos consumidores de contenido se les asigna la comunidad en función de la proporción de relaciones de amistad respecto a cada comunidad de nodos creadores de contenidos.

El método Louvain obtiene mejores resultados (tanto en tiempo como en

valor de categorías) respecto a métodos similares como los que se describen en las obras Clauset et al. (2004) y Wakita and Tsurumi (2007). En la obra Rotta and Noack (2011) se mejora el método Louvain con un proceso de refinamiento multinivel pero tiene como contrapartida que es significativamente más lento, restringiendo el tamaño de las redes a analizar.

A pesar de todo lo expuesto, cualquier aproximación basada en la modularidad, sufre lo que se denomina *límite de resolución* y suelen dar resultados insatisfactorios cuando se quiere encontrar comunidades por debajo de cierta escala, dependiendo del tamaño total de la red. Por ello, estos métodos son precisos a la hora de detectar comunidades a gran escala (respecto a la red en concreto) pero no son apropiados para recuperar comunidades a menor escala.

Dado que el enfoque de este trabajo consiste en la obtención de comunidades a gran escala de usuarios políticamente sesgados extraídos de una red de gran tamaño, se considera que el método Louvain es bastante apropiado para la obtención de este tipo de comunidades dentro de esta tesis.

Merece la pena mencionar que esta aproximación (como muchas otras aproximaciones existentes) no considera que exista cierto solapamiento en las comunidades, mientras que realmente, muchos usuarios son capaces de exhibir rasgos políticos de diversas comunidades.

5.4.3. La aproximación propuesta basada en biclustering

Aunque el Método Louvain, como se ha mencionado anteriormente, es bastante apropiado para revelar comunidades a gran escala dentro de redes de usuarios, se percibe que este método está más o menos limitado; es insuficiente para descubrir el conocimiento estructural subyacente. La mayor desventaja presentada por la aproximación anterior es que todo el análisis de comunidades se realiza en "diferido", desacoplando de forma moderada los consumidores de contenido de los creadores de contenido en lugar de analizarlos de forma conjunta.

Por ello, en este apartado, se propone una aproximación diferente para revelar la estructura de comunidades subyacente, abordando el análisis conjunto de ambos tipos de nodo mediante el modelado de la tarea como un problema de *biclustering*, una manera novedosa de abordar esta tarea si recurrimos a la literatura actual.

El proceso de *Biclustering*, también llamado *Co-clustering* o *clustering bimodal*, es un proceso de minería de datos diseñado para realizar un clustering (o agrupamiento) simultaneo de filas y columnas de una matriz dada (Hartigan, 1972; Mirkin, 1998). Un *bicluster* es un subconjunto de la matriz original cuyas filas presentan un comportamiento similar respecto a sus columnas y viceversa. Sin embargo, la definición exacta de "comportamiento similar" depende del algoritmo de biclustering utilizado. Como requisito para aplicar cualquier técnica de biclustering sobre la tarea abordada en este capítulo, es necesario transformar la representación basada en grafo de los usuarios de la red a una representación matricial tradicional.

A partir del grafo bipartito reducido mencionado en el apartado 5.4.1, se genera una matriz de amistad M donde $M_{i,j} = 1$ si y solo sí, el creador i es seguido por el consumidor j (o lo que es lo mismo, j tiene una relación de amistad directa con i). Esta matriz está lista para ser usada como entrada para cualquier técnica de biclustering y tiene el beneficio de que cualquier bicluster extraído

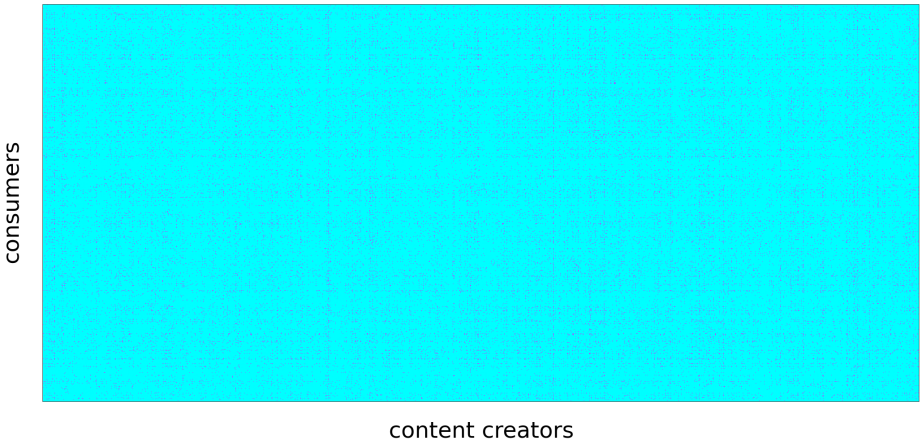


Figura 5.7: Matriz de amistad sin estructura obtenida del grafo bipartito.

del proceso representa una comunidad entre usuarios (tanto creadores como consumidores) que exhiben comportamiento similar. La figura 5.7 muestra la matriz de amistad resultante lista para ser utilizada como entrada para cualquier técnica de biclustering. A priori, no se observa ningún patron o estructura en la matriz, pero tras aplicar la técnica de biclustering aparecerá dicha estructura subyacente, tal y como se observa en la figura 5.8.

Existe una variedad de técnicas de biclustering, cada una de ellas con sus peculiaridades y su fundamento subyacente, pero describirlas a todas ellas queda fuera del alcance de esta tesis. El estudio Madeira and Oliveira (2004) es una buena fuente para el que esté interesado en saber más sobre técnicas de biclustering. Para el problema concreto abordado en esta tesis, se ha elegido el algoritmo *Spectral Biclustering* (Kluger et al., 2003) porque es capaz de generar un modelo de comunidades difuso que permite diferentes grados de pertenencia a diferentes biclusters en lugar de un típico modelo inyectivo.

La técnica *Spectral Biclustering* se basa en la idea de que la matriz de datos posee una estructura tipo tablero de ajedrez oculta y las n filas y m columnas pueden ser particionadas en $n \times m$ biclusters. Cada fila pertenecerá a m biclsuters y cada columna a n biclusters con diferentes grados de pertenencia. El algoritmo termina usando estos $m \times n$ biclusters para calcular el bicluster más representativo para cada elemento fila y cada elemento columna.

La figura 5.8 muestra una matriz de amistad cuyas filas y columnas han sido reordenadas después de aplicar el algoritmo de Spectral Biclustering. Tanto filas como columnas se han agrupado acorde a su bicluster más representativo y se han dibujado las líneas divisorias entre biclusters por motivos de claridad. En la figura se observa que la varianza en los datos dentro de cada bicluster es baja, indicando una alta correlación entre los creadores de contenido y los consumidores dentro de cada bicluster.

La figura 5.9 muestra un grafo de similaridad de los nodos creadores de contenido extraidos del modelo generado a partir del grafo biparito y la aproximación de biclustering explicada en esta sección. Los nodos han sido escalados respecto al número de seguidores directos en el grafo bipartito y coloreados en función del bicluster más representativo al que pertenecen. La relevancia intra-cluster se

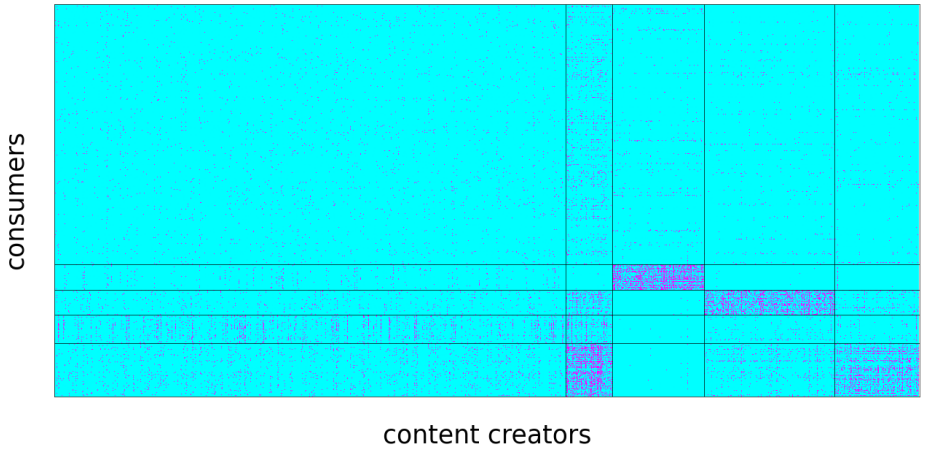


Figura 5.8: Matriz de amistad reordenada después de aplicar Spectral Biclustering.

ha representado mediante la alteración de la saturación del color de cada nodo, siendo los nodos más relevantes dentro del cluster aquellos con mayor valor de saturación. La métrica de similaridad usada es la distancia del coseno calculada sobre el modelo de pertenencia de los nodos creadores de contenido generado por el algoritmo de biclustering. Dos nodos están conectados si su similaridad es muy alta (más de 0,90).

Los nodos rojos son creadores de contenido afines al partido *PSOE* o a su ideología socialdemócrata mientras que los nodos de color morado son creadores de contenido afines a la ideología de izquierdas pero no están tan de acuerdo con la ideología de la democracia social, siendo afines a otros movimientos como el comunismo, el anarquismo social o el socialismo verde.

Los nodos azules son creadores de contenido afines al partido *PP*, su ideología cristiano-democrática y centro-derecha o cualquier otra ideología de derechas tales como son el conservadurismo, el liberalismo económico o el monarquismo. Los nodos amarillos son famosos, personas de la sociedad y prensa amarilla con baja relevancia política pero con un gran número de seguidores.

Es interesante que los nodos de gran tamaño que se encuentran entre las comunidades roja y azul son los medios de comunicación de tirada nacional *El País*, *El Mundo*, *20m* y el *ABC*, que hacen de puente entre las ideologías socialistas y centro-derecha, mientras que los medios totalmente afines a la izquierda o a la derecha se encuentran muy dentro de sus respectivas comunidades.

Esta aproximación de biclustering permite la generación de un modelo de mayor calidad que es inherentemente difuso, puede manejar comunidades de menor escala mejor que el Método Louvain y su coste computacional es sólo ligeramente superior.

5.5. Evaluación de las aproximaciones

A lo largo de esta sección se evalúan las dos aproximaciones propuestas en la sección anterior para la detección de comunidades. Medir el grado de éxito o el

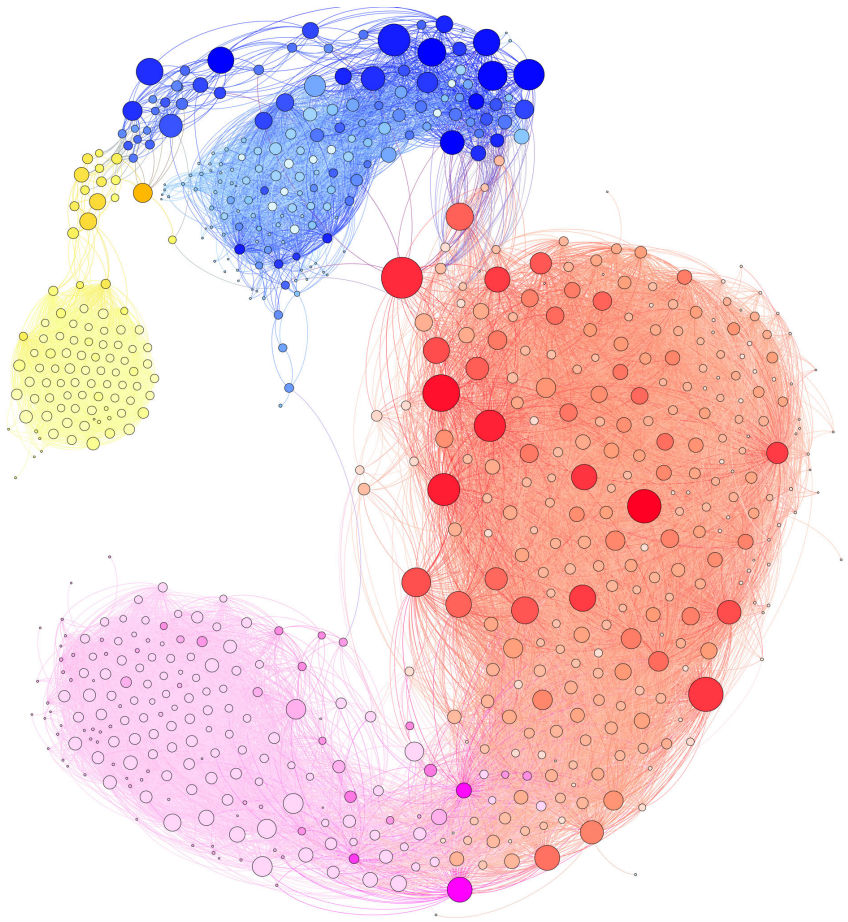


Figura 5.9: Grafo de similaridad de los creadores de contenidos extraído de la aproximación de Spectral Biclustering

rendimiento de alguna de estas aproximaciones no es una tarea fácil, puesto que no se posee un *gold standard* de referencia para evaluar la tarea definida. En su lugar, se proponen otras alternativas para medir la bondad de las propuestas.

Los apartados 5.5.1 y 5.5.2 proporcionan una comparación entre el método propuesto basado en Spectral Biclustering y la aproximación más tradicional basada en el Método Louvain. El proceso de evaluación descrito en el apartado 5.5.1 está basada en una métrica intrínseca mientras que una tarea extrínseca es usada como medida indirecta en el proceso de evaluación descrito en el apartado 5.5.2. Finalmente, en el apartado 5.5.3 se describe la realización de un análisis de carácter cualitativo sobre las comunidades extraídas por la aproximación de Spectral Biclustering.

5.5.1. Coeficiente Silhouette

Dado que no existe información real y fidedigna sobre las comunidades existentes en el dataset, no se dispone de los valores de referencia para comparar

y el proceso de evaluación debe ser realizado sobre el propio modelo. El *coeficiente Silhouette* (Rousseeuw, 1987) es un conocido método para la evaluación y validación de clusters de datos, estimando cómo de bien situado se encuentra cada objeto dentro de su cluster asignado.

Dado una única muestra, su distancia media intra-cluster a y su distancia media al cluster más cercano b son calculadas usando una distancia métrica específica, como la distancia euclídea, la distancia manhattan o la distancia del coseno. El coeficiente s para una única muestra se define como $s = \frac{b-a}{max(a,b)}$ donde $-1 \leq s \leq +1$. Para evaluar el modelo, se calcula el coeficiente para cada muestra existente en el conjunto de datos y se aplica una medida de tendencia central (como la media o la mediana) sobre los coeficientes para puntuar el modelo. Los valores de los coeficientes Silhouette se encuentran -1 y $+1$, donde -1 indica un clustering totalmente incorrecto y $+1$ un clustering muy bien definido (denso y muy bien delimitado). Valores alrededor de 0 indican que los clusters sufren de demasiado solapamiento.

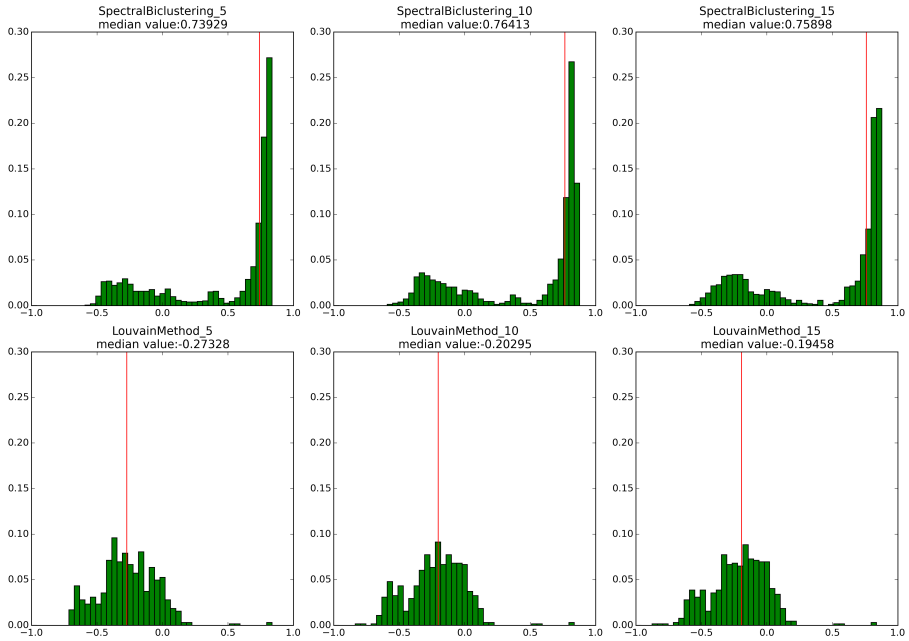


Figura 5.10: Distribución de los coeficientes Silhouette

Usando la distancia euclídea cuadrada como métrica y la mediana como medida de centralidad, se han calculado los coeficientes silhouette para ambas aproximaciones respecto a 5, 10 y 15 clusters. La figura 5.10 muestra las distribuciones de puntuaciones (histogramas) y cada mediana se muestra como una linea vertical roja.

Se observa que los coeficientes para el Método Louvain están bien distribuidos alrededor del valor medio y son mayoritariamente negativos, mostrando un clustering de mala calidad con cierto solapamiento. Aumentar el número de clusters mejora ligeramente el valor central, pero incrementa significativamente la desviación típica y los resultados generales se mantienen estables.

Las medianas de los coeficientes para la aproximación de Spectral Biclustering son mucho mayores y la mayoría de los coeficientes son mayores que 0,5. Es interesante que la distribución de valores exhibe bimodalidad, siendo los coeficientes negativos menos del 30 % del total y agrupados alrededor de −0,3. Esto indica un clustering mayoritariamente denso con algo de solapamiento; variar el número de clusters no cambia los valores significativamente.

A la vista de los resultados, se observa que la aproximación basada en Spectral Biclustering ofrece clusters con un mayor grado de densidad y delimitación que que la aproximación basada en el Método Louvain, dando a lugar a un clustering de, potencialmente, mejor calidad.

5.5.2. Evaluación extrínseca

Otra forma de analizar el rendimiento de cualquiera de las aproximaciones consiste en medir cómo de bien sus resultados contribuyen a otra tarea de distinta naturaleza. Para el caso abordado en esta tesis, las salidas de las aproximaciones son utilizadas como modelos de características en una tarea de clasificación de opinión sobre los tweets escritos en español, en concreto la tarea y el dataset descritos en el capítulo 4.

Cualquier tweet de este dataset puede expresar cualquier postura positiva, negativa o neutral respecto a los partidos PP y PSOE, definiendo las clases de opinión posibles como el producto cartesiano de cualquiera de las tres posturas respecto a cada partido, haciendo un total de 9 categorías. El dataset fue manualmente anotado, indicando la postura política de cada tweet de acuerdo a las categorías mencionadas.

Ahora bien, como ya se describe en las sección 4.3, la mayoría de los españoles son respecto a cualquier tema político, dejando la mayoría de las categorías de clasificación con muy baja representación; más del 90 % del dataset corresponde a tres de las nueve categorías de opinión política. Esta situación lleva a considerar dos versiones del dataset: una "completa" con todas las clases y una "reducida" con solo las tres categorías de mayor representación política.

Para cada aproximación de detección de comunidades previamente explorada y usando sus respectivos modelos de comunidades, a cada usuario en el dataset se le calcula la proporción de amigos que pertenecen a cada comunidad, obteniendo un vector de valores cuya suma es 1. El conjunto de todos estos vectores es el modelo de características correspondiente a esa aproximación. Por motivos de comparación, también se incluyen en esta comparativa tanto el modelo estándar *Bag-of-Words* como el clasificador *Dummy aleatorio estratificado*.

Modelo de Características	Problema Completo	Problema Reducido
Dummy Aleatorio estratificado	31,32 %	36,37 %
Bag-of-Words	61,97 %	68,36 %
Método Louvain	50,07 %	56,04 %
Spectral Biclustering	60,18 %	68,75 %

Cuadro 5.1: Exactitud con validación cruzada para diferentes modelos de características

De acuerdo a este método de evaluación extrínseca, la tabla 5.1 muestra que la aproximación de Spectral Biclustering consigue mejores resultados que el Método Louvain para esta tarea de clasificación. Cabe destacar que el modelo de características basado en Spectral Biclustering, a pesar de sólo basarse en la estructura topológica de la red, obtiene un rendimiento comparable al modelo de características Bag-of-Words basado en contenido textual.

5.5.3. Análisis cualitativo de usuarios políticamente relevantes

En los apartados anteriores, se ha analizado el rendimiento de la aproximación basada en Spectral Biclustering y se ha comparado respecto a la aproximación del Método Louvain, mostrando que la aproximación basada en biclustering es significativamente mejor. En este apartado se realiza otro tipo de evaluación, consistente en un análisis cualitativo cuyo objetivo es el de medir la capacidad de identificar usuarios relevantes dentro de las comunidades detectadas obtenidas por parte de la aproximación basada en biclustering.

Partiendo del mismo modelo de comunidades generado mediante Spectral Biclustering en el apartado 5.5.1, se han elegido aquellas comunidades que contienen las cuentas oficiales de los partidos PP y el PSOE. De ambas comunidades, se han elegido los primeros 10 usuarios relevantes que no eran cuentas políticas oficiales, medios de comunicación oficiales ni usuarios directamente afiliados con el correspondiente partido político de su cluster.

El objetivo de este estudio en concreto fue el de encontrar individuos que fueran afines a alguna ideología correspondiente a uno de los partidos políticos mayoritarios pero que no fueran miembros ampliamente conocidos del partido o tuvieran una estrecha relación con algún medio de comunicación mayoritario. De esta forma, se puede medir si el método es capaz de identificar y agrupar correctamente gente corriente que comparta la misma ideología política, a pesar de que pueden que no hayan interactuado entre sí y no profesen públicamente su afiliación política.

Para cada uno de estos usuarios, se han obtenido los 100 tweets más recientes de su *timeline* cuyo contenido tuviera relación directa con la política; estos tweets fueron manualmente anotados de la misma forma que aquellos usados en la tarea descrita en el apartado 5.5.2. Después de este proceso de anotación, se calcula la afinidad de cada usuario respecto al partido mayoritario de su respectiva comunidad (o *cluster group*), siendo la medida de afinidad definida tal como se explica a continuación. La relevancia intra-cluster de los usuarios es directamente proporcionada por el algoritmo de biclustering.

Definición 5.1 *Dado el usuario u, se definen los siguientes puntos:*

- *Sea x uno de los partidos políticos mayoritarios. Se define* $\text{pos}_u(x)$ *y* $\text{neg}_u(x)$ *como el número de veces que el usuario u expresa una opinión respecto al partido político x en cualquiera de los tweets, siendo ésta positiva o negativa respectivamente.*

- *Sea PP y PSOE los partidos políticos mayoritarios de estudio, se define el apoyo o* support *a dichos partidos por parte del usuario u como* $\text{sc}_u(PP) = \text{pos}_u(PP) + \text{neg}_u(PSOE)$ *y* $\text{sc}_u(PSOE) = \text{pos}_u(PSOE) + \text{neg}_u(PP)$.

A partir de lo expuesto, se definen las puntuaciones de afinidad para el usuario u respecto a ambos partidos políticos como:

$$\underset{u}{\text{aff}}(PP) = \frac{\text{sc}_u(PP)}{\text{sc}_u(PP) + \text{sc}_u(PSOE)}$$
$$\underset{u}{\text{aff}}(PSOE) = \frac{\text{sc}_u(PSOE)}{\text{sc}_u(PP) + \text{sc}_u(PSOE)}$$

Usuario	Afinidad	Relevancia
PSOE_USER#1	100,00 %	68,27 %
PSOE_USER#2	100,00 %	65,56 %
PP_USER#1	97,30 %	71,61 %
PSOE_USER#3	95,45 %	61,07 %
PP_USER#2	90,91 %	74,31 %
PSOE_USER#4	87,50 %	54,61 %
PSOE_USER#5	85,00 %	66,51 %
PP_USER#3	83,33 %	65,69 %
PSOE_USER#6	69,87 %	70,65 %
PP_USER#4	66,67 %	64,85 %

Cuadro 5.2: Valores de afinidad de los 10 usuarios con mayor relevancia política

La tabla 5.2 muestra los valores de relevancia intra-cluster y afinidad para los usuarios elegidos. Los usuarios han sido anonimizados antes del proceso de anotación manual por motivos de privacidad. Se puede observar que los usuarios obtienen altos valores de afinidad (respecto al partido mayoritario de su comunidad) a la vez que exhiben una buena correlación entre los valores de relevancia y afinidad. Esto indica que la aproximación principal propuesta en este capítulo es capaz de recuperar miembros de la comunidad altamente afines a la ideología de su comunidad a pesar de que la gran mayoría de estos usuarios no están afiliados a ningún partido político ni relacionados con ningún medio de comunicación políticamente sesgado.

5.6. Conclusiones y trabajo futuro

En este capítulo, se ha propuesto una idea original como aproximación para detectar comunidades de usuarios de interés dentro de Twitter. Modelando la tarea como un problema de biclustering y aplicando la técnica de *Spectral Biclustering*, se consigue una manera efectiva de abordar la tarea de detección de comunidades sobre redes de gran tamaño. Más concretamente, se han centrado los esfuerzos en extraer comunidades subyacentes en la red de usuarios españoles dentro del contexto político español.

Dado que analizar la red completa de usuarios españoles en Twitter es técnicamente intratable, se propone como alternativa el uso de una colección de tweets obtenidos usando el método de recuperación dinámica de tweets explicado en Cotelo et al. (2014) y expandido en el capítulo 2. A partir de esta

colección de tweets, se ha generado una muy útil representación de la red de usuarios consistente en un grafo de amistad directo, cuyos usuarios han sido extraídos directa e indirectamente de la colección de datos recuperado. Este grafo de amistad directa sirve como base para la mayoría del análisis y proceso de evaluación.

Para poder probar la propuesta de forma efectiva, se ha implementado otro método basado en la *maximización de la modularidad* como baseline exigente. Este tipo de aproximaciones obtienen buenos resultados a la hora de detectar comunidades a gran escala y sus requisitos computacionales son relativamente bajos, haciéndolos apropiados para el análisis de grafos de gran tamaño. De las técnicas de maximización de la modularidad disponibles, se ha optado por el conocido *Método Louvain* pues obtiene buenos resultados mientras que sus requisitos computacionales son mucho menores en comparación a otras técnicas.

A través de métodos de evaluación tanto intrínsecos como extrínsecos, se observa que el modelo de comunidades generado por la aproximación basada en *Spectral Biclustering* es muy superior al generado por el baseline propuesto. No sólo funciona mejor, sino que proporciona mejores modelos que permiten comunidades solapadas, pertenencia difusa a comunidades, comunidades más pequeñas y proporciona información adicional sobre la relevancia intra-cluster de los miembros de una comunidad. A pesar de que la técnica de biclustering elegida es ligeramente más lenta que el Método Louvain, sus requisitos de memoria son similares y los resultados obtenidos son mucho mejores.

Además de los métodos de evaluación propuestos, se realiza un análisis cualitativo de los usuarios políticamente relevantes mediante el uso del modelo de comunidades provisto por la aproximación de biclustering. Usando la información de relevancia intra-cluster del modelo, se pudo observar que el método era capaz de identificar y correctamente agrupar usuarios ordinarios que comparten la misma ideología política a pesar de que esos usuarios no han interactuado entre ellos y no profesan públicamente ningún tipo de afiliación política.

Finalmente, se concluye que la propuesta se comporta con éxito, siendo capaz de abordar grandes redes que suelen ser abordadas por aproximaciones tradicionales de maximización de la modularidad mientras que obtiene mejores resultados y proporciona un modelo mucho más rico.

Capítulo 6

Conclusiones

Como se ha mencionado a lo largo de esta memoria de tesis, las redes sociales son un elemento que ha pasado de ser un simple servicio a formar parte fundamental de la vida de las personas, experimentando un enorme auge en los últimos años. La facilidad de creación de contenido por parte de los usuarios, la sencillez con la que éste es compartido de forma efectiva, la inmediatez de las plataformas y la simbiosis con los dispositivos móviles de nueva generación (*smartphones*), han introducido cambios muy importantes sobre cómo las personas interactúan entre sí.

Este éxito generalizado por parte de las redes sociales mayoritarias, donde millones de personas opinan, comentan y comparten a diario ideas sobre cualquier tema, hace que las redes sociales sea un objetivo como fuente potencial de información a analizar. Dado que los mensajes generados por los usuarios, y en general contenidos de cualquier tipo, rara vez son objetivos, hace que la información potencial a extraer haya atraído la atención de un diversos sectores como un elemento lucrativo. Además de enfoques más obvios como campañas de marketing, análisis de la opinión pública, análisis de marcas o análisis de perfiles de usuarios, las empresas participan activamente dentro de estas redes, no sólo como elemento dinamizador, sino además como soporte técnico y atención al cliente.

Sin embargo, la información que se puede encontrar en este tipo de redes difiere bastante en lo que respecta a los tipos de información que se encuentran tradicionalmente en la mayoría de contextos, requiriendo de un enfoque holístico a la hora de analizar los contenidos de las redes sociales; es necesario explorar tanto el análisis del contenido en sí como el papel que éste o su autor desempeña dentro de la red.

En esta memoria de tesis, se recoge el proceso investigador realizado para enfrentarse a las diferentes facetas del análisis de mensajes en redes sociales, todo ello bajo el enfoque holísitco de integrar tanto el aspecto estructural de los contenidos y los usuarios como el contenido no estructurado de los mensajes. En primer lugar, se aborda la tarea de recuperación de los datos respecto a una temática, pues es imperativo generar datasets que estén los suficientemente enfocados a una temática dado que la variedad existente en las redes es prácticamente infinita.

Un aspecto a tener en cuenta, que es independiente a la temática elegida, es que los mensajes suelen ser mal redactados y de baja calidad, por lo que una

de las tareas abordadas es la de normalizar, a nivel léxico, los mensajes. Esta normalización busca paliar la baja calidad que pueda existir en los mensajes, intentando restaurar sus contenidos para que sea más efectivo analizarlos.

La clasificación de mensajes respecto a la opinión que contienen es otra de las tareas abordada que, además de ser interesante, tiene implicaciones adicionales a nivel socio-político muy relevantes. Siguiendo el tema central de la tesis, se aborda la tarea extrayendo modelos de ambos aspectos (estructurado y no estructurado) de los mensajes e integrándolos de forma efectiva usando un esquema de combinación diseñado para ello.

Muy interrelacionado con la tarea anterior, se ha tratado la detección de comunidades implícitas de usuarios respecto a una temática en concreto. Revelar este tipo de comunidades de interés contrasta respecto al enfoque pormenorizado de la tarea de clasificación, pues la naturaleza de ésta es a mucha mayor escala. Ahora bien, a la hora de caracterizar los colectivos de usuarios respecto a intereses comunes, hay que tener en cuenta que muchos de esos usuarios no se conocen ni interactúan de ninguna forma entre sí, resultando en un enfoque muy interesante respecto a muchas tareas de análisis.

Teniendo en cuenta lo expuesto anteriormente, las principales aportaciones contenidas en esta memoria de tesis son las siguientes:

1. El diseño de un método dinámico de recuperación de información respecto a una temática concreta sobre redes sociales, apropiada sobre una red social (Twitter) que presenta un comportamiento muy dinámico, pero que no ofrece herramientas lo suficientemente apropiadas para esta tarea.

2. Una caracterización de los fenómenos de error y peculiaridades encontrados en los mensajes recuperados de las redes sociales, principalmente de redacción pobre y escritos desde un dispositivo móvil.

3. Un sistema de normalización léxica altamente modular para restaurar la calidad de los mensajes recuperados. Este sistema combina diferentes componentes de normalización, es fácilmente extensible, su coste de implantación es bajo y tanto los recursos como los componentes son reutilizables.

4. Una metodología para categorizar mensajes de opinión de las redes sociales dentro del contexto específico, explorando aproximaciones basadas tanto en información estructurada como no estructurada y proponiendo un esquema de integración basado en *Stacking* para combinar modelos de características de diferente naturaleza.

5. Un método para resolver la tarea de descubrimiento y caracterización de las comunidades implícitas y *ad-hoc* de usuarios dentro de un contexto específico, usando un enfoque combinado de representación de grafo y *Spectral Biclustering*.

Las diferentes aportaciones han sido un resultado directo del diseño e implementación de los experimentos detallados en esta memoria de tesis, los cuales han logrado validar las hipótesis iniciales presentadas en la sección 1.2. Aunque conclusiones de forma detallada e individual respecto a cada experimentos, han sido presentadas en sus respectivos capítulos, podemos destacar algunas de las conclusiones principales extraídas de los resultados obtenidos:

1. En la tarea de recuperación de mensajes, el análisis del grafo de términos subyacente construido a partir de la información estructurada existente en los mensajes es una técnica mucho mas versátil y proporciona mucha mas cobertura que usar simplemente una consulta estática de términos. Además, usando sólo un conjunto de términos semilla significativo y acotado, se consigue una recuperación de mensajes de gran volumen centrada en un tema representado por ese conjunto semilla, con poco ruido introducido y que responde a eventos inesperados en el tiempo.

2. Respecto a la normalización de mensajes ruidosos, se comprueba que, tendiendo en cuenta los fenómenos de error observados, la idea de utilizar componentes independientes que funcionan como "grupo de expertos" es una forma muy efectiva para abordar dichos mensajes ruidosos. Además, como resultado adicional de esta idea, la arquitectura modular resultante exhibe un alto grado de especialización en sus componentes, siendo esta especialización beneficiosa en términos de coste, extensibilidad y adaptación a otros dominios

3. La elaboración de un esquema de integración de modelos de características extraídos de información tanto estructurada como no estructurada es fundamental para la tarea de clasificación de opinión de mensajes generados por los usuarios. Usar sólo información no estructurada es ineficaz, ya que los textos de los mensajes carecen de la cantidad de contenido necesaria para las técnicas de PLN comunes.

4. La detección de comunidades de usuarios de interés en las redes sociales requiere de técnicas diferentes a las de clustering tradicional, pero con una representación de grafo usando la información estructurada de los mensajes y la técnica de Spectral Biclustering, se consiguen unos resultados muy satisfactorios; las comunidades comunidades de usuarios obtenidas a través de este procedimiento son muy acertadas, probando que la aproximación topológica mostrada es viable en grandes grafos y obtiene resultados muy fidedignos.

A partir del desarrollo de la investigación expuesta en esta memoria de tesis, se han generado varios trabajos en forma de artículos de investigación. Los resultados publicados relacionados con la tarea de recuperación de información se encuentran en los trabajos Cotelo et al. (2012) y Cotelo et al. (2014), mientras que los resultados relacionados con la tarea de normalización se encuentran en los trabajos Cotelo et al. (2013) y Cotelo et al. (2015a). También se ha publicado el trabajo Cotelo et al. (2015c), el cual combina resultados provenientes de la recuperación de información y la clasificación de opinión en el contexto político español. Los resultados referentes a las tarea de clasificación de opinión y detección de comunidades se encuentran en las obras en proceso de revisión Cotelo et al. (2015b) y Cotelo et al. (2015d) respectivamente.

Bibliografía

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics.

Ageno, A., Comas, P. R., Padró, L., and Turmo, J. (2013). The talp-upc approach to tweet-norm 2013. In *Proceedings of the tweet normalization workshop at SEPLN 2013*. Sociedad Española para el Procesamiento del Lenguaje Natural.

Al-Osaimi, S. and Badruddin, K. M. (2014). Role of emotion icons in sentiment classification of arabic tweets. In *Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems*, pages 167–171. ACM.

Babour, A. and Khan, J. I. (2014). Tweet sentiment analytics with context sensitive tone-word lexicon. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 392–399. IEEE.

Barclay, F. P. (2014). Political opinion expressed in social media and election outcomes-us presidential elections2012. *Journal on Media & Communications (JMC)*, 1(2).

Berstel, J. and Boasson, L. (1979). Transductions and context-free languages. *Ed. Teubner*, pages 1–278.

Berstel, J. and Reutenauer, C. (1988). *Rational series and their languages*, volume 12. Springer-Verlag Berlin.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2002). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24(1):49–64.

Cao, G., Nie, J.-Y., Gao, J., and Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 243–250, New York, NY, USA. ACM.

Chen, Y., Amiri, H., Li, Z., and Chua, T.-S. (2013). Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 43–52. ACM.

Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

Congosto, M. L., Fernández, M., and Egido, E. M. (2011). Twitter y política: Información, opinión y ¿predicción? *Cuadernos de Comunicación Evoca*, 4.

Costa-Jussa, M. R. and Banchs, R. E. (2013). Automatic normalization of short texts by combining statistical and rule-based techniques. *Language resources and evaluation*, 47(1):179–193.

Cotelo, J., Cruz, F., Troyano, J., and Ortega, F. (2015a). A modular approach for lexical normalization applied to spanish tweets. *Expert Systems with Applications*, 42(10):4743 – 4754.

Cotelo, J. M., Cruz, F. L., Enríquez, F., and Troyano, J. A. (2015b). Tweet categorization by combining content and structural knowledge. Submitted to *Information Fusion* (INFFUS); Under Review.

Cotelo, J. M., Cruz, F. L., and Troyano, J. A. (2013). Resource-based lexical approach to tweet-norm task. In *Tweet-Norm@ SEPLN*, pages 20–24.

Cotelo, J. M., Cruz, F. L., and Troyano, J. A. (2014). Dynamic topic-related tweet retrieval. *Journal of the Association for Information Science and Technology*, 65(3):513–523.

Cotelo, J. M., Cruz Mata, F., and Troyano Jiménez, J. A. (2012). Generación adaptativa de consultas para la recuperación temática de tweets.

Cotelo, J. M., Cruz Mata, F., and Troyano Jiménez, J. A. (2015c). Explorando twitter mediante la integración de información estructurada y no estructurada. *Procesamiento del Lenguaje Natural*, (55).

Cotelo, J. M., Ortega, F. J., Troyano, J. A., and Vallejo, C. G. (2015d). Detecting communities of interest in twitter using spectral biclustering over friend relationships. Submitted to *Knowledge and Information Systems* (KAIS); Under Review.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Davidov, D., Tsur, O., and Rappoport, A. (2010a). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Davidov, D., Tsur, O., and Rappoport, A. (2010b). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Derényi, I., Palla, G., and Vicsek, T. (2005). Clique percolation in random networks. *Physical Review Letters*, 94(16):160202.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369.

Fernández, J., Gutiérrez, Y., Gómez, J. M., and Martínez-Barco, P. (2014). Gplsi: Supervised sentiment analysis in twitter using skipgrams. *SemEval 2014*, pages 294–298.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

for Democratic Action, A. (2009). Annual voting records.

Gamallo, P., García, M., Muñoz, M., and del Río, I. (2013a). Learning verb inflection using cilenis conjugators. *THE EUROCALL REVIEW*, 21(1):12–19.

Gamallo, P., García, M., and Pichel, J. R. (2013b). A method to lexical normalisation of tweets. In *Proceedings of the tweet normalization workshop at SEPLN 2013*. Sociedad Española para el Procesamiento del Lenguaje Natural.

Gayo-Avello, D., Metaxas, P. T., and Mustafaraj, E. (2011). Limits of electoral predictions using twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Gibson, D., Kleinberg, J., and Raghavan, P. (1998). Inferring Web communities from link topology. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems links, objects, time and space—structure in hypermedia systems - HYPERTEXT '98*, pages 225–234, New York, New York, USA. ACM Press.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. In *Processing*.

Golbeck, J. and Hansen, D. (2011). Computing political preference among twitter followers. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 1105–1108, New York, NY, USA. ACM.

Greene, D., O'Callaghan, D., and Cunningham, P. (2012). Identifying Topical Twitter Communities via User List Aggregation.

Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 368–378. Association for Computational Linguistics.

Han, B., Cook, P., and Baldwin, T. (2012). Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 421–432. Association for Computational Linguistics.

Han, B., Cook, P., and Baldwin, T. (2013). Lexical normalization for social media text. *ACM Trans. Intell. Syst. Technol.*, 4(1):5:1–5:27.

Han, B., Cook, P., and Baldwin, T. (2014). Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, pages 451–500.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337):123–129.

Hong, S. and Nadler, D. (2011). Does the early bird move the polls?: the use of the social media tool 'twitter' by u.s. politicians and its impact on public opinion. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, dg.o '11, pages 182–186, New York, NY, USA. ACM.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 29–32. Association for Computational Linguistics.

Jabeen, S., Shah, S., and Latif, A. (2013). Named entity recognition and normalization in tweets towards text summarization. In *Digital Information Management (ICDIM), 2013 Eighth International Conference on*, pages 223–227. IEEE.

Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. (2011). Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 151–160, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kahn, J. H., Tobin, R. M., Massey, A. E., and Anderson, J. A. (2007). Measuring emotional expression with the linguistic inquiry and word count. *The American journal of psychology*, pages 263–286.

Kim, E., Gilbert, S., Edwards, M. J., and Graeff, E. (2009). Detecting sadness in 140 characters: Sentiment analysis of mourning michael jackson on twitter. *Web Ecology*, 3.

Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46:668–677.

Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. (2003). Spectral biclustering of microarray cancer data: Co-clustering genes and conditions. *Genome Research*, 13:703–716.

Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert systems with applications*, 40(10):4065–4074.

Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2010). Finding statistically significant communities in networks. *PLoS ONE*, page 24.

Lau, J. H., Collier, N., and Baldwin, T. (2012). On-line trend analysis with topic models:\# twitter trends detection topic model online. In *COLING*, pages 1519–1534. Citeseer.

Lavrenko, V. and Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127, New York, NY, USA. ACM.

Lee, K. S., Croft, W. B., and Allan, J. (2008). A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 235–242, New York, NY, USA. ACM.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Lim, K. H. and Datta, A. (2012a). Finding twitter communities with common interests using following links of celebrities. In *Proceedings of the 3rd international workshop on Modeling social media - MSM '12*, page 25, New York, New York, USA. ACM Press.

Lim, K. H. and Datta, A. (2012b). Tweets Beget Propinquity: Detecting Highly Interactive Communities on Twitter Using Tweeting Links. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 214–221. Ieee.

Madeira, S. and Oliveira, A. (2004). Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 1(1):24–45.

Mirkin, B. (1998). *Mathematical classification and clustering: From how to what and why*. Springer.

Mohammad, S. M. and Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34. Association for Computational Linguistics.

Mohammad, S. M., Zhu, X., Kiritchenko, S., and Martin, J. (2014). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*.

Mohri, M. (2009). Weighted automata algorithms. In *Handbook of weighted automata*, pages 213–254. Springer.

Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., and Ureña-López, L. A. (2014). Ranked wordnet graph for sentiment polarity classification in twitter. *Computer Speech & Language*, 28(1):93–107.

Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Park, C. S. (2013). Does twitter motivate involvement in politics? tweeting, opinion leadership, and political engagement. *Computers in Human Behavior*, 29(4):1641–1648.

Pennacchiotti, M. and Popescu, A.-M. (2011). Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–438. ACM.

Pennell, D. and Liu, Y. (2011). A character-level machine translation approach for normalization of sms abbreviations. In *IJCNLP*, pages 974–982.

Phi-Long (2012). Python 3.3+ implementation of the language guessing module made by Jacob R. Rideout for KDE.

Pla, F. and Hurtado, L.-F. (2014). Sentiment analysis in twitter for spanish. In *Natural Language Processing and Information Systems*, pages 208–213. Springer.

Porta, J. and Sancho, J. L. (2013). Word normalization in twitter using finite-state transducers. In *Proceedings of the tweet normalization workshop at SEPLN 2013*. Sociedad Española para el Procesamiento del Lenguaje Natural.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–23.

Rotta, R. and Noack, A. (2011). Multilevel local search algorithms for modularity clustering. *J. Exp. Algorithmics*, 16:2.3:2.1–2.3:2.27.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0):53 – 65.

Sachan, M., Contractor, D., Faruquie, T. A., and Subramaniam, L. V. (2012). Using content and interactions for discovering communities in social networks. In *Proceedings of the 21st international conference on World Wide Web - WWW '12*, page 331, New York, New York, USA. ACM Press.

Schulz, A., Mencía, E. L., Dang, T. T., and Schmidt, B. (2014). Evaluating multi-label classification of incident-related tweets. *Making Sense of Microposts (# Microposts2014)*, pages 26–33.

Schulz, K. and Mihov, S. (2002). Fast string correction with levenshtein-automata. *INTERNATIONAL JOURNAL OF DOCUMENT ANALYSIS AND RECOGNITION*, 5:67–85.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 4:379–423.

Sidarenka, U., Scheffler, T., and Stede, M. (2013). Rule-based normalization of german twitter messages. In *Proc. of the GSCL Workshop Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation*.

Silva, I. S., Gomide, J., Veloso, A., Meira, Jr., W., and Ferreira, R. (2011). Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *Proceedings of the 34th international ACM SIsGIR conference on Research and development in Information*, SIGIR '11, pages 475–484, New York, NY, USA. ACM.

Small, T. A. (2011). What the hashtag? a content analysis of canadian politics on twitter. *Information, Communication & Society*, 14(6):872–895.

Soboroff, I., Ounis, I., and Lin, J. (2012). Overview of the trec-2012 microblog track. In *The Twenty-First Text REtrieval Conference Proceedings*.

Speriosu, M., Sudan, N., Upadhyay, S., and Baldridge, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63. Association for Computational Linguistics.

Talukdar, P. P. and Crammer, K. (2009). New regularized algorithms for transductive learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 442–457. Springer.

Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., and Li, P. (2011). User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1397–1405, New York, NY, USA. ACM.

Tao, T. and Zhai, C. (2006). Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 162–169, New York, NY, USA. ACM.

Tseng, B. L. (2005). Tomographic clustering to visualize blog communities as mountain views. In *In WWW 2005 Workshop on the Weblogging Ecosystem*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, pages 402–418.

Villena Román, J., Lana Serrano, S., Martínez Cámara, E., and González Cristóbal, J. C. (2013). Tass-workshop on sentiment analysis at sepln. *Procesamiento del Lenguaje Natural*, 50:37–44.

Vitale, D., Ferragina, P., and Scaiella, U. (2012). Classification of short texts by deploying topical annotations. In *Advances in Information Retrieval*, pages 376–387. Springer.

Wakita, K. and Tsurumi, T. (2007). Finding community structure in mega-scale social networks. *CoRR*, abs/cs/0702048.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5:241–259.

Xie, S., Tang, J., and Wang, T. (2014). Topic related opinion integration for users of social media. In *Social Media Processing*, pages 164–174. Springer.

Yang, J. and Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213.

Yang, Y. and Eisenstein, J. (2013). A log-linear model for unsupervised text normalization. In *EMNLP*, pages 61–72.