Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 75-82 recibido 26-02-2015 revisado 26-04-2015 aceptado 08-05-2015

Explorando Twitter mediante la integración de información estructurada y no estructurada

Exploring Twitter by Combining Structured and Unstructured Information

Juan M. Cotelo Universidad de Sevilla jcotelo@us.es
Fermín Cruz Universidad de Sevilla fcruz@us.es
F. Javier Ortega Universidad de Sevilla javierortega@us.es
José A. Troyano Universidad de Sevilla troyano@us.es

Resumen: En este artículo mostramos cómo es posible sacar partido de la información estructurada que proporciona la red social Twitter. Los textos escritos en Twitter son cortos y de baja calidad, lo que dificulta la aplicación de técnicas y herramientas que tradicionalmente se han venido usando para procesar textos en lenguaje natural. Sin embargo, Twitter ofrece mucho más que los 140 caracteres de sus mensajes para trabajar. En el ecosistema Twitter hay muchos objetos (*tweets*, *hashtags*, usuarios, palabras, ...) y relaciones entre ellos (co-ocurrencia, menciones, re-tuiteos, ...) que ofrecen innumerables posiblidades de procesado alternativo a las técnicas clásicas de PLN. En este trabajo hemos puesto nuestra atención en la tarea de clasificación de *tweets*. Sólo usando la información de la relación *Follow* hemos conseguido un clasificador que iguala los resultados de un clasificador basado en bolsas de palabras. Cuando usamos las *features* de los dos modelos, el resultado de la clasificación mejora en más de 13 puntos porcentuales con respecto a los modelos originales lo que demuestra que ambos clasificadores aportan informaciones complementarias. También hemos aplicado la misma filosofía a la tarea de recopilación del corpus con el que hemos trabajado, usando una técnica de recuperación dinámica basada en relaciones entre entidades Twitter que nos ha permitido construir una colección de *tweets* más representativa.
Palabras clave: Recuperación de tweets, clasificación de tweets, información estructurada y no estructurada

Abstract: In this paper we show how it is possible to extract useful knowledge from Twitter structured information that can improve the results of a NLP task. Tweets are short and low quality and this makes it difficult to apply classical NLP techniques to this kind of texts. However, Twitter offers more than 140 characters in their messages to work with. In Twitter ecosystem there are many objects (tweets, hashtags, users, words, ...) and relationships between them (co-occurrence, mentions, re-tweets, ...) that allow us to experiment with alternative processing techniques. In this paper we have worked with a tweet classification task. If we only use knowledge extracted from the relationship *Follow* we achieve similar results to those of a classifier based on bags of words. When we combine the knowledge from both sources we improve the results in more than 13 percentual points with respect to the original models. This shows that structured information is not only a good source of knowledge but is also complementary to the content of the messages. We also have applied the same philosophy to the task of collecting the corpus for our classification task. In this case we have use a dynamic retrieval technique based on relationships between Twitter entities that allows us to build a collection of more representative tweets.
Keywords: Tweets retrieval, tweets categorization, structured and unstructured information

ISSN 1135-5948

1 Introducción

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, Ferragina, y Scaiella, 2012; Schulz et al., 2014), en especial para determinar la polaridad de las opiniones siendo ésta una de las tareas sobre Twitter más estudiadas por la comunidad científica (Agarwal et al., 2011; Montejo-Ráez et al., 2014; Fernández et al., 2014; Pla y Hurtado, 2014) la extracción de *topics* (Lau, Collier, y Baldwin, 2012; Chen et al., 2013), la identificación de perfiles (Abel et al., 2011), la geolocalización (Han, Cook, y Baldwin, 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* 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 trabajan son cortos, con una estructura gramatical en ocasiones poco definida, 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 y 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 handicap, pero afortunadamente hay maneras de resolver tareas sobre textos sin prestar mucha atención a los textos en sí. Por ejemplo, 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). La pregunta que ha motivado este trabajo es ¿se podría hacer algo parecido con ciertas tareas sobre Twitter? La respuesta es claramente sí. Twitter, además de sus defectos en cuanto a la calidad de los textos, tiene una gran virtud: los textos están acompañados de mucha información estructurada que relaciona a múltiples entidades de distinta naturaleza. Y es posible diseñar soluciones a muchas tareas que contengan un componente no estructural (mediante el análisis del contenido de los mensajes) y otro componente estructural (mediante el análisis de los datos y relaciones asociados a los mensajes).

La Figura 1 muestra algunos de los objetos y relaciones más importantes que podemos extraer de Twitter. No están todos los objetos, y ni siquiera están todas las posibles relaciones entre los cuatro objetos destacados, pero aún así da una idea del potencial de esta información si se utiliza convenientemente.

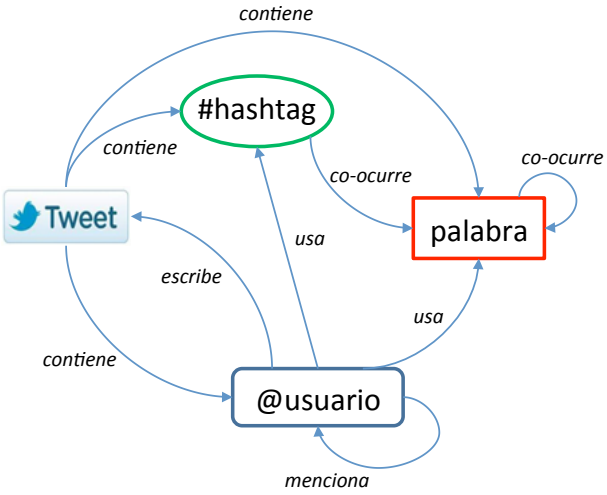

Figura 1: Algunas de las relaciones de Twitter.

En este trabajo pretendemos mostrar cómo el uso combinado de información estructurada y no estructurada beneficia la resolución de dos tareas relacionadas con Twitter: la recuperación de *tweets* y la clasificación de *tweets*. En ambos casos se usan técnicas que hacen uso tanto del contenido de los mensajes como de toda la estructura que los rodea.

El resto del trabajo se organiza de la siguiente manera. En la sección 2 se presenta el marco de trabajo en el que vamos a realizar las experimentaciones, se trata del análisis de afinidades políticas en Twitter. En la sección 3 se explica cómo se aprovecha un grafo de relaciones entre elementos de Twitter para recopilar el corpus con el que trabajaremos. En la sección 4 se muestra cómo se pueden mejorar los resultados de una clasificación de *tweets* haciendo uso de información sobre relaciones entre los autores. Por último, en la sección 5 se extraen las conclusiones.

2 El marco de trabajo: análisis de afinidades políticas en Twitter

El ámbito político es uno de los más utilizados por los estudios que usan Twitter como fuente de datos. La política siempre da que hablar y eso también se traslada a las redes sociales. Tanto los eventos importantes en el calendario político (p.e. las elecciones) como las noticias que ocurren día a día, provocan un aluvión de mensajes de personas, más o menos influyentes en la red, que se posicionan y opinan con respecto a la actualidad política. Hay trabajos que usan Twitter para resolver distintas tareas relacionadas con la política como la predicción de resultados (Tumasjan et al., 2010), la clasificación de usuarios (Pennacchiotti y Popescu, 2011) o la aplicación de técnicas de análisis de sentimientos con respecto a partidos políticos (Mejova, Srinivasan, y Boynton, 2013).

Para este trabajo hemos elegido, dentro del dominio político, una tarea de clasificación múltiple que permita determinar la postura de los mensajes con respecto a los dos grandes partidos del panorama político español: PP y PSOE. Para cada uno de los dos partidos hemos definido tres posibles categorías: a favor, en contra y neutral. Con esta configuración, cada *tweet* puede ser clasificado en una de las nueve categorías que se corresponden con el producto cartesiano de los ejes correspondientes a cada partido.

Para realizar nuestros experimentos necesitamos, en primer lugar, un corpus representativo de textos referentes a estos dos partidos políticos, que utilizaremos para evaluar distintas aproximaciones a la tarea de clasificación. Tanto el proceso de construcción del corpus, como la solución final planteada para la clasificación, nos brindan oportunidades de verificar la hipótesis principal del trabajo: el beneficio de integrar información estructurada y no estructurada a la hora de analizar los contenidos publicados en una plataforma como Twitter. En ambos casos (recuperación y clasificación de *tweets*) acabaremos apoyándonos en sendos grafos construidos a partir de ciertas relaciones de Twitter para mejorar el resultado de cada tarea. En el problema de la recuperación usaremos un grafo de usuarios y términos que nos ayudará a diseñar consultas flexibles con las que podremos adaptarnos a los nuevos temas y tendencias que continuamente aparecen en Twitter. En el caso de la clasificación demostraremos que con la ayuda de un grafo de usuarios se puede extraer información adicional que permite mejorar los resultados de la tarea.

3 Recuperación adaptativa de tweets

La manera más habitual de recopilar corpus de *tweets* es decidir un conjunto de términos (palabras, *hashtags* o nombres de usuarios) y usar la API de Twitter para lanzar consultas con esos términos. Este método es directo y se consigue recuperar todos los mensajes que se vayan escribiendo y que contengan esas palabras clave. Si esos términos son frecuentes, en poco tiempo se puede conseguir una buena colección de mensajes con la que trabajar. El problema de esta aproximación es que no tiene en cuenta una de las principales características de Twitter: la espontaneidad. Fijar de antemano el conjunto de palabras clave de las consultas nos limita sólo a los mensajes que contengan estos términos y nos podemos dejar fuera mensajes que tengan que ver con nuevos eventos que pueden aparecer en cualquier momento. En el caso de la política, si usamos un conjunto de términos predeterminados perdemos la posibilidad de capturar al vuelo términos que en un determinado momento captan la atención de la comunidad de usuarios interesada.

Nuestra aproximación a la recuperación pasa por tener consultas dinámicas, que evolucionan y que se adaptan a los temas que en cada momento toman más protagonismo en la conversación colectiva. Nuestras consultas estarán compuestas de dos tipos de términos, usuarios y etiquetas. Para no perder el foco se definen una serie de términos semilla que en nuestro caso serán las etiquetas #PP y #PSOE. A partir de ellos, y cada cierto tiempo, se construirán las consultas que incluirán

también aquellos términos que se consideren más relevantes en función del análisis de los mensajes recuperados en la consulta anterior. Este proceso contempla dos fases: construcción de un grafo de relaciones entre usuarios y etiquetas a partir de los textos recuperados en la consulta anterior, y aplicación de un algoritmo de *ranking* a dicho grafo para determinar los términos más relevantes en ese momento.

Las cuatro relaciones contempladas a la hora de construir el grafo de etiquetas y usuarios son:

- Usa: Relaciona un autor con una etiqueta. El autor del *tweet* usa la etiqueta en el texto del mensaje.
- Menciona: Relaciona un autor con otro autor. El primero de ellos usa el nombre del segundo en el mensaje.
- Re-tuitea: Relaciona un autor con otro autor. El primero de ellos reenvía un mensaje del segundo.
- Co-ocurre: Relaciona dos etiquetas entre sí. Ambas aparecen en un mismo mensaje.

Si se intenta añadir un arco ya existente al grafo, se incrementa en uno el peso de esa relación. Esta información será utilizada por el algoritmo de *ranking* para dar más importancia a las relaciones con frecuencias de aparición más altas.

La figura 2 muestra un sencillo ejemplo del tipo de grafo que se construye al procesar una secuencia de *tweets*. Todas las relaciones identificadas en cada uno de los mensajes son registradas mediante la creación de los correspondientes nodos o arcos del grafo.

Una vez que se ha construido el grafo, se le aplica una adaptación del algoritmo PageRank (Page et al., 1999) que tiene en cuenta los pesos de los arcos a la hora de calcular la relevancia de los nodos. Cada cierto período de tiempo se calcula un nuevo grafo a partir de los mensajes de la consulta anterior y se lanza una nueva consulta con los términos más relevantes según el *ranking*. El período de tiempo dependerá de la actividad de la comunidad relacionada con los términos semilla, en el caso de PP y PSOE observamos que 60 minutos era un buen período. En cuanto al número de términos del *ranking* con los que quedarnos, tras hacer varias pruebas, vimos que entre 5 y 15 se conseguía una captura de

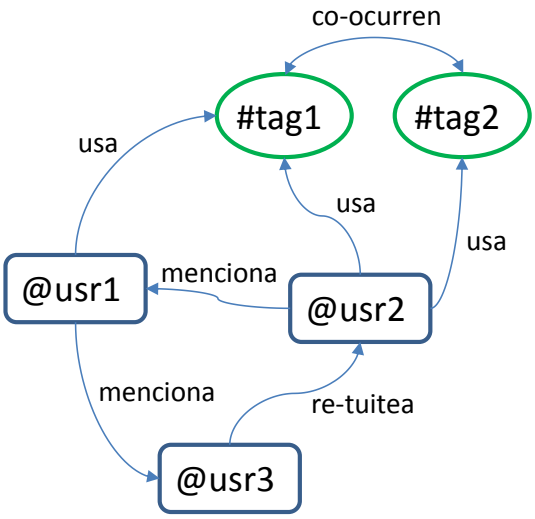

Figura 2: Ejemplo sencillo del tipo de grafo de etiquetas y usuarios usado en el proceso de recuperación.

buena calidad sin la introducción de mensajes no relacionados con las semillas y elegimos como umbral de corte el 10.

El objetivo final de esta tarea es disponer de una colección de *tweets* anotados que nos permita evaluar la siguiente tarea de clasificación. Para ello se lanzó el proceso de recuperación durante una semana y del conjunto total de mensajes se extrajo una muestra aleatoria de 3000 *tweets*. Dichos mensajes fueron anotados manualmente, registrando en cada caso la categoría a la que pertenecía de las nueve definidas en nuestro esquema.

Con idea de evaluar también el proceso de recuperación, se anotó si los *tweets* de la muestra eran relevantes, o no, para la temática PP/PSOE. El resultado de esta anotación fue que un 92,25 % de los mensajes estaban relacionados con la temática. Hay que destacar que muchos de los *tweets* no relevantes se deben a la utilización de términos de búsqueda, a priori fiables, que resultan ser ambiguos (por ejemplo la etiqueta #PP es usada por el programa de televisión chileno *Primer Plano*). Por tanto, la pérdida de precisión no es achacable en su totalidad al posible ruido introducido por el método de recuperación dinámico. Por ejemplo, en el caso de la etiqueta #PP sólo el 96,60 % de los *tweets* recuperados de forma directa con la consulta estática #PP son relevantes.

Esta pérdida de precisión del método dinámico de recuperación está acompañada con una mayor capacidad de recuperación de *tweets* interesantes, cifrada en un incremen-

to en volumen del 125,93 % con respecto a los mensajes que se hubiesen recuperado sólo con los términos semilla #PP y #PSOE mediante consultas estáticas.

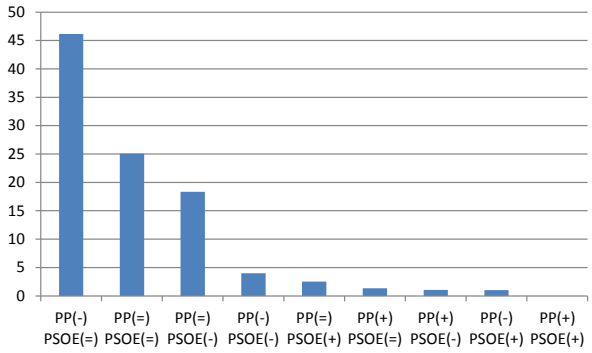

Figura 3: Distribución de porcentajes en el corpus anotado para cada una de las nueve categorías de la tarea de clasificación.

4 Clasificación de la afinidad política

En esta sección explicaremos las diferentes aproximaciones que hemos seguido a la hora de abordar la tarea de clasificación. En total hemos probado tres modelos distintos: uno basado en el contenido de los mensajes, otro basado en la estructura y otro que integra ambas informaciones.

4.1 Clasificación basada en el contenido

En la figura 3 se observa cómo se distribuye la muestra de *tweets* anotados en las nueve categorías de la tarea de clasificación. Los resultados muestran en general que la percepción de la política por parte de los usuarios de Twitter es bastante negativa. Por destacar sólo algunos datos, el 69,49 % de los mensajes son negativos con respecto a alguno de los dos partidos (el 4 % son negativos para los dos), mientras que sólo el 5,96 % son positivos con respecto a uno de los dos partidos (y no hay ningún mensaje que sea positivo para los dos).

Dado que la colección está bastante sesgada hacia las tres categorías mayoritarias (que suman el 89,54 % de los mensajes, hemos realizado dos tipos de experimentos: con todas las categorías (lo hemos denominado *Tarea-9*) y sólo con los mensajes de las tres categorías mayoritarias (*Tarea-3*).

La primera aproximación probada es la de la clasificación basada en el contenido. Con un modelo clásico de bolsa de palabras obtuvimos un resultado del 61,97 % de *accuracy* para la *Tarea-9* y de un 68,36 % para la *Tarea-3*. Los resultados se obtuvieron con un clasificador SVM y aplicando validación cruzada sobre el corpus de entrenamiento. Son resultados bastante bajos incluso para la tarea reducida en la que sólo hay que decidir entre tres categorías, lo que da idea de la dificultad del corpus. Esta dificultad de decidir en base al contenido se debe, entre otros factores, a los fenómenos usados para expresar la afinidad o rechazo a un partido, la falta de calidad de los textos y el hecho de que en muchos mensajes no haya menciones explícitas a los partidos PP y PSOE a pesar de que sí se refieren a ellos.

4.2 Clasificación basada en la estructura

Una vez que tenemos los resultados de la clasificación basada en contenidos el siguiente paso es intentar verificar la hipótesis de que la información estructurada que proporciona Twitter es de utilidad para una tarea como ésta. Analizando las distintas informaciones y relaciones disponibles, nos decidimos por aquellas relacionadas con los autores. La idea es intentar obtener un perfil de los autores de los mensajes que refleje la tendencia política del mismo. Si se asume que esta tendencia política va a ser relativamente constante, esta información puede ser de gran ayuda para complementar la que aporta el propio mensaje. Para determinar esa tendencia nos apoyamos en la relación *Follow* de Twitter, que nos da una información muy valiosa sobre los intereses de los usuarios.

A partir de los autores de los mensajes de nuestro corpus, recuperamos a todos los usuarios seguidos por éstos, lo que nos da como resultado un grafo de seguidores. Nuestra intuición era que en ese grafo hay suficiente información como para agrupar a los usuarios de nuestro corpus en grupos con la misma tendencia política.

El problema recuerda al de detección de comunidades en una red (Girvan y Newman, 2002; Shen et al., 2009) aunque las técnicas clásicas para resolver esta tarea no son directamente aplicables por la particular estructura de nuestro grafo. Por lo general estas técnicas identifican grupos de nodos densamente conectados (Ball, Karrer, y Newman, 2011) y en nuestro caso lo que necesitamos es

agrupar los nodos de los autores que siguen a un grupo similar de usuarios (llamaremos a estos usuarios *referentes*). Esta diferenciación entre los usuarios autores de mensajes y usuarios referentes nos da una estructura de grafo bipartito como la que se ilustra en el esquema de la figura 4.

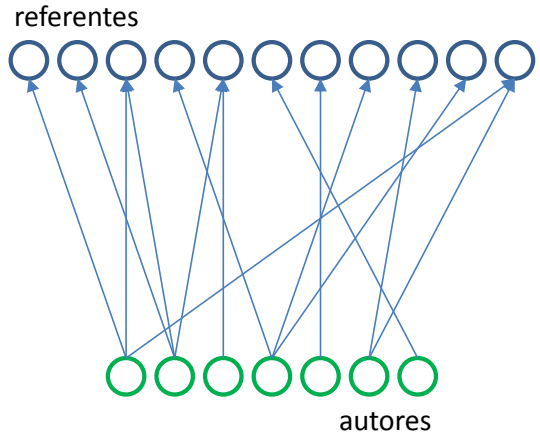

Figura 4: Ejemplo sencillo del grafo bipartito de usuarios usado en el proceso de clasificación.

Con esta topología, más que identificar si los usuarios están bien conectados, lo que nos interesa es identificar si existen patrones en la forma de seguir a los usuarios referentes. En este sentido hemos usado la matriz de adyacencia de autores y consumidores para agrupar a los autores similares en los mismos grupos. Usando técnicas de clustering hemos obtenido un total de 5 grupos de usuarios, de modo que para cada autor podemos calcular su grado de pertenencia a cada uno de estos grupos. Usando los grados de pertenencia del autor de un *tweet* como *features* para nuestra tarea de clasificación obtenemos un 59,97 % de precisión para la tarea completa y un 68,75 % para la tarea reducida. Es decir, sólo usando información sobre las personas a las que sigue el autor de un texto somos capaces de obtener resultados similares a los que obtiene el clasificador basado en bolsa de palabras.

4.3 Integración de ambos modelos

Una vez que hemos observado que la capacidad de clasificación del contenido y de las relaciones *Follow* es similar, la pregunta natural es, ¿serán complementarias? Para responderla hemos realizado dos experimentos de combinación. En el primero de ellos hemos construido un *dataset* en el que se incluye, para cada mensaje, las *features* provenientes del modelo de bolsa de palabras (experimento *BOW*) y las de la afinidad a cada una de las comunidades detectadas (experimento *Grafo*). En el segundo esquema de combinación hemos aplicado la técnica de *Stacking* (Wolpert, 1992) que permite aplicar un clasificador de segundo nivel sobre las salidas de una serie de clasificadores base. Para ello hemos usado las *features* de los modelos *BOW* y *Grafo* para entrenar a su vez a cuatro clasificadores base (SVM, *Naive Bayes*, Máxima Entropía y *Random Forests*). En la tabla 1 se muestran los resultados de los dos esquemas de combinación, junto con los de los modelos originales y los de un *baseline* que consiste en elegir la categoría más frecuente en ambas tareas de clasificación. Los resultados de todos los experimentos se han obtenido mediante validación cruzada de 10 iteraciones.

Modelo	**Tarea-9**	**Tarea-3**
Baseline	46,37 %	51,53 %
BOW	61,97 %	68,36 %
Grafo	59,97 %	68,75 %
BOW+Grafo	64,79 %	71,98 %
Stacking	75,10 %	81,14 %

Tabla 1: Resultados de *accuracy* de los diferentes modelos entrenados para la clasificación de *tweets*.

Tal y como se observa en la tabla, la información estructurada no sólo es de la "misma calidad" que el contenido de los mensajes, al obtenerse con ella resultados similares, sino que además es complementaria. Cuando se integran en un clasificador *features* de ambos modelos el resultado de la clasificación mejora a los dos clasificadores base. Para la *Tarea-9* esta mejora es de casi 3 puntos porcentuales si se integran directamente las *features* y en más de 13 puntos si se usa una técnica más sofisticada de combinación que permite sacar mayor ventaja de las visiones complementarias que aporta cada tipo de información.

5 Conclusiones y trabajo futuro

En este trabajo hemos mostrado la utilidad de la información estructurada proporcionada por un medio social como Twitter a la hora de procesar los mensajes de los usuarios. Este tipo de aproximaciones permiten que las técnicas PLN usadas a la hora de resolver

ciertas tareas sobre textos puedan ser complementadas con otras informaciones e indicios. Este apoyo es especialmente interesante a la hora de procesar contenidos escritos por usuarios que por lo general suelen ser de una pobre calidad lingüística.

Hemos elegido la tarea de clasificación de *tweets* en el ámbito político para verificar esta hipótesis. Aplicando técnicas clásicas de modelo vectorial obtuvimos una precisión del 61,97 % en la tarea de clasificar los mensajes según su afinidad o rechazo con respecto a los partidos PP y PSOE. Este resultado es prácticamente igualado (con un 59,97 %) con un clasificador que usa sólo como *features* información extraída de las relaciones entre usuarios de Twitter. La primera conclusión, por tanto, es que para una tarea de clasificación de contenidos la información estructurada de la red es casi tan útil como los mensajes en sí.

Tras experimentar con técnicas de combinación, conseguimos alcanzar el 75,10 %, superando ampliamente los resultados de los clasificadores iniciales. La segunda conclusión de nuestro trabajo es, por tanto, que el conocimiento aportado por la vía estructural resulta ser muy complementario con respecto al que se puede extraer mediante el análisis de los contenidos.

Esta misma filosofía de integración de información estructurada y no estructurada ha sido también puesta en práctica en el proceso de recuperación de *tweets* que hemos desarrollado para construir el corpus para la tarea de clasificación. En este caso hemos utilizado un método dinámico que hace uso de relaciones entre distintas entidades de Twitter para crear consultas que en cada momento se adaptan a los términos de interés con respecto a una determinada temática.

Estamos convencidos de que este tipo de aproximaciones será de utilidad en muchas tareas, no sólo sobre Twitter, sino en cualquier medio en el que se disponga informaciones de distinta naturaleza. Como línea de trabajo futuro tenemos pensado explorar este espacio de nuevas tareas, medios sociales y tipos de informaciones estructuradas disponibles en estos medios.

Agradecimientos

Este trabajo ha sido financiado a través de los proyectos ATTOS-ACOGEUS (TIN2012-38536-C03- 02) y AORESCU (P11-TIC-7684 MO).

Bibliografía

Abel, F., Q. Gao, G.J. Houben, y K. Tao. 2011. Semantic enrichment of twitter posts for user profile construction on the social web. En *The Semanic Web: Research and Applications*. Springer, pági-nas 375–389.

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

Ball, B., B. Karrer, y M. Newman. 2011. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):36–103.

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

Fernández, J., Y. Gutiérrez, J.M. Gómez, y P. Martınez-Barco. 2014. Gplsi: Supervised sentiment analysis in twitter using skipgrams. *SemEval 2014*, páginas 294–298.

Girvan, M. y M. EJ Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

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

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

Lau, J.H., N. Collier, y T. Bald-win. 2012. On-line trend analysis with topic models: \# twitter trends detection

topic model online. En *COLING*, páginas 1519–1534. Citeseer.

Mejova, Y., P. Srinivasan, y B. Boynton. 2013. Gop primary season on twitter: popular political sentiment in social media. En *Proceedings of the sixth ACM international conference on Web search and data mining*, páginas 517–526. ACM.

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

Page, L., S. Brin, R. Motwani, y T. Winograd. 1999. The page-rank citation ranking: Bringing order to the web.

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

Pla, F. y L.F. Hurtado. 2014. Sentiment analysis in twitter for spanish. En *Natural Language Processing and Information Systems*. Springer, páginas 208–213.

Schulz, A., E. Loza Mencía, T. T. Dang, y B. Schmidt. 2014. Evaluating multi-label classifica-tion of incident-related tweets. *Making Sense of Microposts (# Microposts2014)*, páginas 26–33.

Shen, H., X. Cheng, K. Cai, y M. Hu. 2009. Detect overlapping and hierarchical community structure in net-works. *Physica A: Statistical Mechanics and its Applications*, 388(8):1706–1712.

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

Villena R., J., S. L. Serrano, E. Martínez Cámara, y J. C. González Cristóbal. 2013. Tass-workshop on sentiment analysis at sepln. *Procesamiento del Lenguaje Natural,* 50:37–44.

Vitale, D., P. Ferragina, y U. Scaiella. 2012. Classification of short texts by deploying topical annotations. En *Advances in Information Retrieval*. Springer, páginas 376–387.

Wolpert, D. H. 1992. Stacked generalization *Neural networks*, 5(2):241–259.