Evaluación de aproximaciones lingüísticas para la asignación de temas a tuits

David Vilares, Miguel A. Alonso and Carlos Gómez-Rodríguez

Grupo LyS, Departamento de Computación, Universidade da Coruña
Campus de A Coruña s/n, 15071, A Coruña, España
{david.vilares, miguel.alonso, carlos.gomez}@udc.es

Resumen Empresas y organizaciones están empezado a interesarse en monitorizar lo que los usuarios opinan sobre ellas en Twitter ya que los tuits constituyen una buena fuente de información para conocer la percepción que la sociedad tiene sobre su área de negocio. Para ello, primero es necesario discriminar las opiniones no relacionadas, dada la gran cantidad de mensajes que se publican a diario en esta red social. En este trabajo presentamos un enfoque basado en procesamiento de lenguaje natural para la clasificación de tuits en función de su temática y evaluamos cómo la información morfológica, sintáctica y semántica puede ayudar a identificar dichos temas. Los resultados experimentales confirman que nuestro enfoque mejora los resultados obtenidos por otros sistemas bajo las mismas condiciones.

Palabras clave: Clasificación de tópicos, Twitter, Procesamiento del lenguaje natural.

1. Introducción

Twitter es una red de *micro blogging* que ha experimentado un gran éxito en los últimos años. En ella, los usuarios publican mensajes de hasta 140 caracteres denominados *tuits*. En ocasiones estos mensajes simplemente reflejan conversaciones y trivialidades, pero cada vez es más habitual encontrar opiniones acerca de productos, servicios, eventos y otros muchos temas. Ello ha despertado el interés de muchas empresas y organizaciones, que ven en este sitio una fuente de información para monitorizar lo que se comenta sobre ellas. Conocer cuál es la percepción de la sociedad acerca de su negocio, descubrir cuáles son los puntos fuertes y débiles de sus productos, o identificar comparaciones con la competencia, son algunos de sus objetivos. Uno de los principales problemas reside en discriminar los mensajes útiles entre la gran cantidad de información que se puede encontrar en este medio, ya que en Twitter se publican actualmente cerca de 500 millones de mensajes diarios, escritos por más de 100 millones de usuarios. Todo ello, sumado a la variedad de temas tratados, convierten a esta red social en un medio ruidoso. Aunque el propio sitio web de Twitter dispone de herramientas de filtrado para seleccionar aquellos mensajes que estén escritos en un

idioma específico o contengan determinadas palabras clave, esto no es suficiente cuando se desea monitorizar una serie de tuits que traten sobre una temática más general.

En este artículo estudiamos y evaluamos cómo distintos enfoques basados en conocimiento lingüístico pueden ayudar a identificar las temáticas de las que trata un tuit escrito en castellano. El problema se ha abordado desde un enfoque de clasificación multi-etiqueta, dado que en un mismo mensaje puede hacerse referencia a varios temas. En concreto, el artículo estudia cómo el conocimiento morfológico, psicométrico y semántico influye en la creación de clasificadores temáticos supervisados. Además, se evalúa cómo la extracción de relaciones sintácticas entre términos puede mejorar el rendimiento de las aproximaciones puramente léxicas mediante la utilización de tripletas sintácticas generalizadas. La utilidad práctica del trabajo viene avalada por los buenos resultados obtenidos en el Taller de Análisis del Sentimiento en la SEPLN [1], donde un sistema inicial que implementaba este enfoque obtuvo el mejor resultado.

El resto del artículo se organiza como sigue. La sección 2 introduce el trabajo relacionado con la clasificación temática en Twitter. En la sección 3 se describe la base teórica de nuestro enfoque. Los experimentos se presentan en la sección 4. Por último, la sección 5 resume las conclusiones y líneas de trabajo futuras.

2. Trabajo relacionado

La categorización temática de textos ha sido tradicionalmente enmarcada como una aplicación de la clasificación mediante técnicas de aprendizaje automático. Partiendo de una colección de documentos preanotados para cada categoría, se construye un modelo que aprende a diferenciar entre cada una de ellas. Hasta hace unos años, esta tarea se centraba principalmente en analizar textos largos. No obstante, con el éxito de Twitter, ha crecido el interés en realizar tareas de categorización sobre micro-textos. Sin embargo, en este ámbito la literatura relacionada es escasa y la mayor parte de los estudios emplean colecciones de documentos en inglés. Para el castellano, se dispone del corpus TASS 2013, presentado en el Taller de Análisis del Sentimiento en la SEPLN [1]. Se trata una colección de tuits escritos en castellano, anotados con las temáticas que se tratan en cada uno de ellos: *cine, fútbol, economía, entretenimiento, literatura, música, política, deportes* (salvo fútbol), *tecnología* y *otros*. Varios autores han realizado experimentos sobre este corpus, lo que facilita la comparación de diferentes enfoques. Por cuestiones de espacio, nos limitaremos a dar una breve descripción de aquellos que participaron en la competición TASS 2013.

Pla y Hurtado [2] proponen una cascada de clasificadores binarios SMO [3] para cada temática con el fin de determinar si un tuit pertenece a dicho tema o no. En caso de que esta cascada no asigne ningún tema a un tuit, se utilizará una segunda cascada de clasificadores libSVM [4]. Dado que cada clasificador libSVM proporciona un valor de confianza de que un tuit pertenezca a una temática dada, se asignará finalmente a cada tuit sometido a esta segunda fase una única temática: aquella que presenta mayor confianza.

Cordobés et al. [5] proponen una técnica basada en similitud de grafos para identificar las temáticas de las que trata un tuit. En este enfoque, cada palabra constituye un vértice del grafo. Una conexión entre dos vértices (arco) representa que esos dos elementos aparecen conjuntamente en algún tuit. A cada arco se le asigna un peso que representa la frecuencia de aparición conjunta de ambos términos. Para reducir la dispersión, las palabras son normalizadas a su raíz gramatical. El conjunto de entrenamiento se emplea para construir un grafo para cada tema, uniendo los grafos obtenidos para los tuits de esa categoría. Después, se construye un grafo para cada tuit del conjunto de test y se busca cuál de los grafos de referencia es más similar, mediante técnicas basadas en [6]. Su propuesta contempla que un tuit solo puede ser asignado a una categoría, aunque en él se traten varios temas.

Castellano González et al. [7] aplican un enfoque basado en técnicas de Recuperación de Información que construye modelos del lenguaje en base a la Divergencia Kullback-Leibler, de tal modo que el contenido de cada tuit es utilizado como si fuese una consulta al índice así construido. Sus resultados sugieren que la indexación de todas las palabras es fundamental para obtener un buen rendimiento, ya que las entidades nombradas sólo ayudan en la clasificación de un pequeño número de tuits.

Montejo-Ráez et al. [8] convierten los términos de un tuit en una representación vectorial siguiendo un esquema de pesos td-idf, tras aplicar un proceso de normalización, que sirve de punto de partida para construir una matriz términos-temáticas que ayude a clasificar las temáticas. La aproximación no obtuvo un buen rendimiento, atribuido al reducido tamaño del conjunto de entrenamiento.

Rufo Mendo [9] propone aplicar un modelo bayesiano, Naive Bayes Complement (NBC) con co-entrenamiento. Sin embargo, lejos de mejorar los resultados, el modelo del NBC con co-entrenamiento empeoró el rendimiento del modelo NBC original.

3. Clasificación de temáticas múltiples en Twitter

La identificación de temáticas en Twitter debe ser abordada como una tarea multietiqueta, dado que un mismo tuit puede referirse a varios temas. Por ejemplo, un tuit donde se critica la política económica del gobierno debería ser etiquetado tanto en *política* como en *economía*. Para ello, proponemos una estrategia *uno contra todos*: dados n temas, construimos n clasificadores binarios, donde cada uno de ellos distingue una temática i, con $i \in \{1 \dots n\}$, del resto del conjunto de categorías j, donde $j \in \{1 \dots n\}$ y $j \neq i$. Para crear cada uno de los clasificadores binarios, utilizamos la implementación de SMO [3] incluida en WEKA [10]. Nuestro enfoque se basa en alimentar esos clasificadores con conocimiento lingüístico que se obtiene de los tuits mediante la utilización de diversas técnicas de Procesamiento del Lenguaje Natural (PLN). En primer lugar se lleva a cabo un proceso de normalización *ad-hoc*. Después se procede a realizar un análisis morfológico y sintáctico, que sirven de punto de partida para la extracción de características lingüísticas. Es importante señalar que nuestro

enfoque se caracteriza por no hacer uso de ningún tipo de meta-información. No se consideran los datos proporcionados por el usuario en su perfil y tampoco se ha realizado un análisis de los enlaces externos que puedan aparecer. Ello permite que nuestro enfoque pueda ser fácilmente adaptable a otros medios sociales. A continuación describimos cada una de las fases más detalladamente.

3.1. Procesamiento de lenguaje natural para el tratamiento de tuits

Preprocesado. Twitter, como otras redes sociales, dispone de diversas expresiones y símbolos propios. Nuestro preprocesador se centra en normalizar algunos de los elementos no gramaticales más habituales que pueden afectar negativamente al rendimiento de una aproximación basada en PLN:

- *Nombres de usuario*: Se detectan los nombres de los usuarios de Twitter, eliminando el símbolo '@' y convirtiendo la primera letra en su correspondiente mayúscula. De esta manera podrán ser tratados adecuadamente durante el resto del proceso, al ser identificados como nombres propios.
- *Hashtags*: Si el hashtag aparece al final o al principio del tuit, se elimina completamente el hashtag. En esos casos se asume que el usuario simplemente desea etiquetar su tuit con un evento específico. Aunque en principio podrían parecer útiles para identificar la temática, este tipo de hashtags suelen referirse a sucesos muy concretos, y por lo tanto su utilidad no persiste a lo largo del tiempo. En otro caso, solo el *'#'* es eliminado. Cuando un hashtag aparece entre palabras, se asume que el término aporta información morfológica y sintáctica (por ejemplo, *'El #iphone es caro.'*).
- *Signos de puntuación*: En entornos web, y en particular en Twitter; donde los usuarios tienen muy limitado el espacio para expresar sus argumentos, es común que los usuarios no respeten las normas ortográficas sobre signos de puntuación. Frases como *'[...] me gusta,pero [...]'* o *'Ayer llegó las 10:00.Y hoy apareció tarde otra vez'* son dos posibles ejemplos de fallos ortográficos habituales. El algoritmo detecta mediante expresiones regulares este tipo de situaciones, incorporando espacios en blanco cuando competa.

Análisis morfológico y lematización. Dada una oración $O = w_1 w_2 ... w_n$, la etapa de análisis morfológico consiste en asignar a cada palabra w_i una etiqueta morfológica e_i que indica su categoría gramatical, lo que habitualmente se denomina *etiqueta de grano grueso*. Opcionalmente también puede incluir información sobre género, número o forma verbal, lo que conforma una *etiqueta de grano fino*. También permite obtener el *lema* o forma canónica de una palabra, esto es, la forma que aparecería como entrada en un diccionario convencional. Tradicionalmente, los etiquetadores morfológicos han trabajado sobre textos gramaticalmente correctos. Con el éxito de la redes sociales, han empezado a surgir herramientas para realizar un análisis morfológico adaptado a este tipo de medios [11]. Sin embargo, la mayoría de estos recursos han sido desarrollados para el inglés, no estando disponibles para el castellano. Para entrenar el etiquetador se ha utilizado el corpus Ancora [12], una colección de textos periodísticos

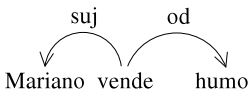

Figura 1. Árbol de dependencias para la frase *'Mariano vende humo'*. *Suj* se refiere a la función sintáctica *sujeto* y *od* al *objeto directo*

anotada con información morfológica y sintáctica, que ha sido usada para construir un etiquetador siguiendo la propuesta de Brill [13]. El 90 % de la colección se empleó para construir el modelo y el de 10 % restante se utilizó a modo de test. Con el fin de adaptar este corpus de textos periodísticos a un entorno web, la sección de entrenamiento fue expandida: cada oración fue duplicada sin que contuviese ningún acento con el fin de que el etiquetador pudiese manejar correctamente frases que contuviesen palabras sin sus tildes. Es por ello por lo que se optó por no usar un modelo pre-entrenado sobre el corpus estándar. Se obtuvo un rendimiento del 95.71 %, coherente con el estado del arte para el castellano.

Análisis sintáctico de dependencias. Dada una oración $O = w_1 w_2 ... w_n$, donde w_i representa la palabra en la posición i en dicha oración, el resultado de aplicar un análisis de dependencias a la oración resulta en un conjunto $G = \{(w_i, arco_{ij}, w_j)\}$, denominado *árbol de dependencias*. Cada elemento constituye una *tripleta de dependencias* que establece relaciones binarias entre pares de palabras: w_i es el término *padre*, w_j es el *dependiente* y $arco_{ij}$ representa la función sintáctica que relaciona ambos términos, conocida como *tipo de dependencia*. El resultado de aplicar un análisis de dependencias a la oración *'Mariano vende humo'* sería $G = \{(vende, sujeto, Mariano), (vende, objeto\ directo, humo)\}$. La figura 1 ilustra la representación gráfica de este sencillo árbol de dependencias. Se ha utilizado MaltParser [14] para construir un analizador dirigido por los datos, tomando Ancora como colección de referencia.

3.2. Extracción de características

De la información lingüística extraída tras aplicar las etapas de PLN se obtienen las características sobre las que trabajan los clasificadores supervisados. Proponemos varios conjuntos de características iniciales, que se detallan a continuación. Dados los problemas de dispersión que pueden presentar algunos de estos modelos, se aplica un filtro de ganancia de información con el fin de seleccionar las características relevantes. Como se ilustra en la sección experimental, algunos de estos modelos iniciales son combinados para mejorar el rendimiento.

N-gramas de términos. La utilización de n-gramas de términos para la creación de clasificadores supervisados constituye una buena línea de base en tareas de análisis de textos. En concreto, segmentar el texto por palabras para después utilizar cada una de ellas (unigramas) como un atributo de entrada al clasificador

constituye el modelo más simple que es posible construir siguiendo este enfoque. Uno de los principales problemas de la utilización de unigramas reside en su incapacidad de capturar correctamente el contexto. Una posible solución consiste en utilizar n-gramas de mayor longitud, como bigramas, donde los términos son agrupados consecutivamente de dos en dos. Ello aporta cierta información estructural acerca de las palabras, agrupándolas según el nivel de proximidad entre ellas. Además, en una lengua como el castellano, donde la variación morfológica de género y número es muy habitual, también puede resultar aplicar técnicas de lematización, para reducir la dispersión de las características. Por ello, nuestra sección experimental explora el rendimiento obtenido tanto como n-gramas de palabras como de lemas.

Etiquetas morfológicas. La utilización de información morfológica es útil en otras tareas de clasificación de textos, como la clasificación de subjetividad [15]. Nuestra hipótesis, para el caso concreto de la clasificación temática, es que este tipo de conocimiento no sería, por si mismo, un buen discriminante a la hora de diferenciar las temáticas de un tuit. Sin embargo, creemos que este tipo de información podría servir de ayuda como complemento a otros conjuntos de características. Por ejemplo, la utilización de nombres propios es más frecuente en dominios como el cine, donde existen muchas referencias a actores, directores o productores, que en otras áreas como la de los automóviles o los electrodomésticos; donde la referencia a características técnicas es más habitual.

Propiedades psicométricas. Las propiedades psicométricas hacen referencia a aspectos psicológicos y semánticos de las palabras. Para tenerlos en cuenta usamos Linguistic Inquiry and Word Count (LIWC) [16], un software que incluye una serie de lexicones para distintos idiomas, entre ellos el castellano. En él, se hace referencia a palabras que denotan aspectos psicológicos como la *ira* o el *enojo*, pero también asocia términos con temáticas muy concretas como *familia*, *televisión*, *trabajo* o *deportes*. Este estudio evalúa la efectividad de este tipo de características en el marco que nos ocupa. El mayor inconveniente de este tipo de recursos lingüísticos, desarrollados manualmente, es su limitada cobertura, problema ya comentado por otros autores [17]. Este problema se ve acentuado en entornos web, donde la calidad de los textos escritos es baja, afectando a la detección de términos que realmente reflejan propiedades psicométricas.

Tripletas sintácticas generalizadas. Las tripletas sintácticas generalizadas fueron presentadas en [18]. Su propuesta se basaba en generalizar bien el término *padre* o *dependiente* a su correspondiente categoría gramatical (nombre, adjetivo, verbo, etc). El objetivo era emplear este tipo de características para alimentar un clasificador supervisado que diferenciase entre oraciones con y sin opinión. En nuestro caso, adaptamos y enriquecemos este concepto al ámbito de la clasificación temática. Dada una tripleta de dependencias original (w_i, arc_{ij}, w_j), donde w_i y w_j son palabras, una tripleta sintáctica generalizada es aquella de

la forma $(g(w_i, A), arc_{ij}, g(w_j, B))$ donde, g es una *función de generalización*, y A y B el tipo de generalización deseada. Nuestro sistema soporta las siguientes generalizaciones: la propia palabra, su forma lematizada, su etiqueta morfológica y sus categorías psicométricas (en caso de tener varias, se devuelven todas sus combinaciones), e incluso la eliminación completa del término. El objetivo es capturar relaciones sintácticas entre pares de palabras que aporten una mayor información estructural que el uso de bigramas, donde la relación entre términos refleja únicamente una contigüidad física entre ellos. El uso de la generalización pretende disminuir los problemas de dispersión que pueden presentar este tipo de características, sin perder información semántica relevante que permita identificar temáticas.

3.3. Selección de tópicos relacionados

Cada tuit es evaluado por cada uno de los n clasificadores binarios de manera independiente, uno por temática, de tal modo que cada uno de ellos indica la pertenencia o no del tuit a la temática en cuestión. De esta manera, se obtiene un conjunto de tópicos predichos y otro de descartados, constituyendo el primer grupo el total de las temáticas relacionadas en un micro-texto.

4. Experimentos

Los experimentos realizados pretenden dar respuesta a las siguientes cuestiones relacionadas con la clasificación temática:

- Determinar si es o no recomendable aplicar técnicas de selección de características para entrenar un clasificador temático.
- Estudiar cómo influye la utilización de conocimiento morfológico, psicométrico y semántico al combinar dicha información.
- Analizar si el uso de información contextual ayuda a mejorar el rendimiento.

El diseño experimental se describe a continuación.

4.1. Descripción del corpus

El corpus TASS 2013 es una colección de tuits escritos en español por distintas personalidades públicas, incluyendo políticos, deportistas, intelectuales y periodistas. Dispone de un conjunto de entrenamiento y otro de test formados por 7 219 y 60 798 tuits, respectivamente. Cada tuit está anotado con las temáticas que en él se tratan. El conjunto de entrenamiento fue anotado manualmente mientras que la colección de test fue etiquetada semi-automáticamente: se llevó a cabo un *pooling* de los sistemas participantes [1] y a continuación la organización del TASS realizó una corrección manual para los casos conflictivos. Los 10 temas considerados son: *cine, fútbol, economía, entretenimiento, literatura, música, política, deportes* (salvo fútbol), *tecnología* y *otros*. La tabla 1 resume la distribución de los temas, tanto en el conjunto de entrenamiento como en el de test. Dado que un tuit puede estar asignado a más de una categoría la suma de la totalidad de temas es mayor que el total de tuits.

Tabla 1. Distribución de tópicos en el corpus TASS 2013. Un tuit puede estar asignado a más de una temática.

Tema	**TASS**$_{entrenamiento}$	**TASS**$_{test}$
cine	245	596
fútbol	252	823
economía	942	2 549
entretenimiento	1 678	5 421
literatura	103	93
música	566	1 498
política	3 120	30 067
deportes	113	135
tecnología	217	287
otros	2 337	28 191
Total	9 573	69 660

4.2. Métricas de evaluación

Las métricas utilizadas para evaluar nuestra propuesta son las estándar, en lo referido a la clasificación multi-etiqueta:

$$Hamming\ loss = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \triangle Z_i|}{|L|} \tag{1}$$

$$Label\text{–}based\ accuracy = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{2}$$

$$Exact\ match = \frac{\#instancias\ correctamente\ clasificadas}{\#instancias} \tag{3}$$

donde:

- L es el conjunto de todas las etiquetas.
- D es el conjunto de instancias de la colección.
- Y_i es el conjunto de etiquetas esperadas para una instancia $i \in D$.
- Z_i es el conjunto de etiquetas predichas para una instancia $i \in D$.
- $\triangle$ es el símbolo que representa la diferencia simétrica entre conjuntos.

4.3. Resultados experimentales

La tabla 2 ilustra el rendimiento para los modelos de características iniciales, donde se observa que los modelos basados en n-gramas obtienen los mejores resultados. En concreto, los unigramas de palabras obtienen el mejor rendimiento para las métricas *Hamming loss* y *label-based accuracy*. Respecto a la métrica *exact match*, el modelo basado en bigramas de lemas es el que obtiene el mejor rendimiento. Ello sugiere que la captura de contexto por medio de bigramas es útil para discriminar mejor los temas no relacionados con un tuit. Además, aplicar filtros previos para seleccionar aquellas características que aportan ganancia

Tabla 2. Rendimiento para los modelos de características iniciales.

Modelo	IG	HL	LBA	EM
Bigramas de lemas (BL)	Sí	0.077	0.626	**0.530**
Palabras (W)	Sí	**0.073**	**0.658**	0.527
Bigramas de palabras (BW)	Sí	0.080	0.613	0.524
Palabras (W)	No	0.079	0.634	0.498
Lemas (L)	Sí	0.078	0.640	0.493
Lemas (L)	No	0.085	0.611	0.460
Información morfológica (FT)	Sí	0.289	0.262	0.032
Propiedades psicométricas (P)	Sí	0.301	0.250	0.026

de información en el conjunto de entrenamiento parece ser beneficioso. Ello queda reflejado en la misma tabla, donde se muestran los resultados con y sin ganancia de información (columna IG), para las aproximaciones basadas en palabras y lemas. La información morfológica no parece ser de utilidad por sí misma. Lo mismo ocurre para el modelo entrenado con propiedades psicométricas, a pesar de que los lexicones del LIWC son capaces de asociar palabras con temáticas muy concretas como *televisión*, *deportes*, *dinero* o *trabajo*. Ello refuerza la hipótesis de la baja cobertura de este tipo de recursos, problema comentado previamente.

En la tabla 3 se ilustra el rendimiento que es posible obtener cuando los unigramas de palabras y bigramas de lemas, los mejores modelos iniciales para alguna de las métricas estándar, son combinados con otros conjuntos de caracerísticas. Vemos que la información morfológica sigue sin ser de utilidad incluso cuando se utiliza como conjunto de características complementario. El conocimiento psicométrico tampoco logra mejoras significativas. Por otro lado, combinar los dos mejores modelos iniciales según las métricas estándar sí mejora el rendimiento. Ello refuerza la hipótesis de que combinar conocimiento léxico con información contextual permite obtener modelos más precisos.

La tabla 4 ilustra el rendimiento cuando al modelo de bolsa de palabras, se le incorpora información contextual mediante tripletas de dependencias, en lugar de ngramas estándar. Las limitaciones de espacio nos impiden mostrar todos los resultados para las distintas tripletas consideradas. El modelo que agrega la tripleta sintáctica no generalizada mejora ligeramente el rendimiento de su correspondiente versión léxica. Las tripletas generalizadas también mejoran el rendimiento del modelo base. El modelo constituido por palabras y tripletas de lemas donde el término padre es eliminado, mejora a su homólogo léxico formado por palabras y lemas. Nuestra hipótesis es que palabras marcadas con funciones sintácticas importantes, como *atributo* o *complemento directo*, pueden ser relevantes para identificar los núcleos del mensaje, y por tanto sus tópicos.

Por último, la tabla 5 compara el modelo sintáctico con los sistemas que participaron en la tarea de clasificación temática de TASS 2013. Nuestra propuesta obtiene el mejor rendimiento en las tres métricas estándar, mejorando significativamente el estado del arte, con la excepción del grupo FHC25-IMDEA, que

Tabla 3. Rendimiento al combinar conjuntos de características iniciales: *bigramas de lemas* (BL), *bigramas de palabras* (BW), *propiedades psicométricas (P)*, *palabras* (W), *lemas* (L), *etiquetas morfológicas* (FT)

Modelo	HL	LBA	EM
W+BL	**0.068**	**0.671**	**0.573**
BL+P	0.076	0.632	0.539
BL	0.077	0.626	0.530
W+BW+P	0.078	0.647	0.530
W+BW	0.074	0.646	0.529
W+P+FT	0.073	0.656	0.528
W	0.073	0.658	0.527
W+P	0.073	0.655	0.526
W+L	0.073	0.656	0.525
BL+P+FT	0.082	0.612	0.495

Tabla 4. Rendimiento al incorporar características sintácticas sobre el modelo de bolsa de palabras: *palabras* (W), *lemas* (L), *tipo de dependencia* (DT) y *propiedades psicométricas* (P)

Características	HL	LBA	EM
W	0.073	0.658	0.527
W+(_,DT,L)	0.071	0.66	0.542
W+(L,DT,P)	0.071	0.661	0.551
W+(L,DT,L)	**0.067**	**0.674**	**0.579**

obtiene un rendimiento muy similar. Sin embargo, esta aproximación no sigue un enfoque multietiqueta, sino monoetiqueta. Siguiendo el mismo enfoque, la propuesta enviada por nuestro grupo al TASS ya hubiera obtenido una *exact match* de 0.589. Pero no consideramos este enfoque valioso ya que no aborda la verdadera naturaleza del problema.

5. Conclusiones y trabajo futuro

En este trabajo hemos evaluado cómo diferentes aproximaciones supervisadas basadas en conocimiento lingüístico son capaces de identificar los temas tratados en mensajes de Twitter. El problema se ha abordado desde un punto de vista de clasificación multi-etiqueta, dado que es frecuente que los usuarios relacionen en un mismo mensaje más de una temática. Los resultados muestran que la inclusión de información estructural, ya sea en forma de tripletas sintácticas generalizadas o de bigramas de lemas, ayuda a mejorar el rendimiento respecto a sistemas clásicos basados en bolsas de palabras. Por contra, la inclusión de información morfológica y psicométrica no ha permitido obtener mejoras respecto

Tabla 5. Comparación del mejor modelo sintáctico con respecto a los modelos participantes en el TASS 2013. Los sistemas han sido ordenados según la *label-based accuracy*.

Modelo	HL	LBA	EM
Modelo sintáctico	**0.068**	**0.674**	**0.579**
FHC25-IMDEA [5]	0.072	0.637	0.573
Modelo inicial enviado al TASS 2013 [19]	0.086	0.614	0.456
UPV [2]	0.084	0.608	0.468
UNED-JRM [9]	0.124	0.417	0.358
ETH-ZURICH [20]	0.098	0.370	0.291
UNED-LSI [7]	0.185	0.197	0.070
SINAI-CESA [8]	0.182	0.126	0.093

a los modelos base. La aproximación propuesta también mejora los resultados de sistemas previos que han evaluado el mismo corpus bajo las mismas condiciones.

Respecto al trabajo futuro, pretendemos incorporar el uso de meta información. Nuestro enfoque sólo tiene en cuenta la información proporcionada por el propio tuit, sin embargo muchos tuits contienen enlaces a webs externas, cuyo análisis también puede ser de utilidad. Además, considerar información que pueda extraerse acerca del usuario puede ser útil de cara a detectar la temática.

Agradecimientos

Trabajo parcialmente financiado por el Ministerio de Economía y Competitividad y FEDER (TIN2010-18552-C03-02) y por la Xunta de Galicia (CN2012/008).

Referencias

1. Villena-Román, J., García-Morera, J.: TASS 2013 — workshop on sentiment analysis at SEPLN 2013: An overview. [21] 112–125
2. Pla, F., Hurtado, L.F.: ELiRF-UPV en TASS-2013: Análisis de sentimientos en Twitter. [21] 220–227
3. Platt, J.C.: Advances in kernel methods. MIT Press, Cambridge, MA, USA (1999) 185–208
4. Chang, C., Lin, C.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems Technology **2**(3) (April 2011) Article 27
5. Cordobés, H., Anta, A.F., Núñez, L.F., Pérez, F., Redondo, T., Santos, A.: Técnicas basadas en grafos para la categorización de tweets por tema. [21] 160–166
6. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proc. of the Seventh International Conference on World Wide Web, Brisbane, Australia (1998) 107–117
7. Castellano González, A., Cigarrán Recuero, J., García Serrano, A.: UNED LSI @ TASS 2013: Considerations about textual representation for IR based tweet classification. [21] 213–219

8. Montejo-Ráez, A., Díaz Galiano, M.C., García-Vega, M.: LSA based approach to TASS 2013. [21] 195–199
9. Rufo Mendo, F.J.: Are really different topic classification and sentiment analysis? [21] 206–212
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explorations **11**(1) (November 2009) 10–18
11. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: 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. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 42–47
12. Taulé, M., Martí, M.A., Recasens, M.: AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Tapias, D., eds.: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco (2008)
13. Brill, E.: A simple rule-based part of speech tagger. In: Proceedings of the workshop on Speech and Natural Language. HLT'91, Stroudsburg, PA, USA, Association for Computational Linguistics (1992) 112–116
14. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E.: Maltparser: A language-independent system for data-driven dependency parsing. Natural Language Engineering **13**(2) (2007) 95–135
15. Pak, A., Paroubek, P.: 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) (May 2010)
16. Pennebaker, J., Francis, M., Booth, R.: Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates (2001) 71
17. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B.: Combining lexicon-based and learning-based methods for Twitter sentiment analysis. Technical Report HPL-2011-89, HP Laboratories, Palo Alto, CA (2011)
18. Joshi, M., Penstein-Rosé, C.: Generalizing dependency features for opinion mining. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. ACLShort '09, Suntec, Singapore, Association for Computational Linguistics (2009) 313–316
19. Vilares, D., Alonso, M., Gómez-Rodríguez, C.: LyS at TASS 2013: Analysing Spanish tweets by means of dependency parsing, semantic-oriented lexicons and psychometric word-properties. In: Proc. of the TASS workshop at SEPLN 2013. IV Congreso Español de Informática. (2013) 179–186
20. García, D., Thelwall, M.: Political aligment and emotional expressions in Spanish tweets. [21] 151–159
21. Díaz Esteban, A., Alegría Loinaz, I., Villena Román, J., eds.: XXIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis at SEPLN 2013, Madrid, Spain, SEPLN (September 2013)