Análisis de la influencia de los sentimientos en el proceso de detección de tópicos en redes sociales

Karel Gutiérrez-Batista
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad de Granada
Granada, España
karel@decsai.ugr.es

Jesús R. Campaña
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad de Granada
Granada, España
jesuscg@decsai.ugr.es

Maria-Amparo Vila
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad de Granada
Granada, España
vila@decsai.ugr.es

Maria J. Martin-Bautista
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad de Granada
Granada, España
mbautis@decsai.ugr.es

***Resumen*—En el presente trabajo se propone realizar un estudio para analizar la influencia que tienen los términos que expresan sentimientos en la detección automática de tópicos en redes sociales. Esta propuesta utiliza una metodología basada en una ontología, a la cual se le incorpora la capacidad de identificar y eliminar aquellos términos que presenten una orientación sentimental en textos de redes sociales, los cuales pueden influir de forma negativa en la detección de tópicos. Para ello se han utilizado dos recursos orientados al análisis de sentimientos con el fin de detectar dichos términos. El sistema propuesto ha sido evaluado con conjuntos de datos reales de las redes sociales Twitter y DreamChatchers en inglés y español respectivamente, demostrando en ambos casos la influencia de los términos con orientación sentimental en la detección de tópicos en textos de redes sociales.**

***Palabras Claves*—Detección tópicos, Agrupamiento jerárquico, Etiquetado grupos, Términos sentimientos, Multilingüe**

I. Introducción

Hoydía es un hecho reconocido el crecimiento y la popularidad alcanzada por las redes sociales, y como consecuencia de ello, el aumento del número de usuarios interactuando en dichas redes, lo que provoca la acumulación de grandes volúmenes de datos textuales no estructurados. Por tal motivo, las redes sociales constituyen una fuente de información de gran importancia, por lo que es de esperar que organizaciones, investigadores, etc., empleen tiempo y recursos en el estudio de estas. Sin embargo, el gran cúmulo y la falta de estructura de los textos, hace que sea prácticamente imposible su procesamiento y análisis automático de forma masiva, motivo por el cual resulta conveniente tener los textos previamente organizados teniendo en cuenta la temática abordada.

La detección de tópicos a partir de textos no estructurados, permite organizar dichos textos por temáticas, lo cual facilita su posterior análisis integrado con datos convencionales. En [1] se ha propuesto una metodología multilingüe para la detección automática de tópicos en datos textuales. Mediante la experimentación, se demostró la viabilidad de la propuesta, aunque se debe resaltar que los resultados no son lo suficientemente buenos cuando los textos provienen de redes sociales.

Esto se debe a que la detección de tópicos en textos más elaborados (librerías digitales, sitios web de noticias, etc.) es diferente cuando los textos pertenecen a redes sociales. En dichos sistemas los usuarios expresan ideas, hechos y sentimientos sobre cualquier tema utilizando un lenguaje coloquial, por lo que es de esperar que en los textos aparezcan con alta frecuencia términos que permiten expresar sentimientos relacionados con determinados productos, servicios, etc.

Motivado por la problemática anterior, en este artículo se realiza un estudio para analizar la influencia de los términos que expresan sentimientos en la detección automática de tópicos en redes sociales. Para ello se propone un nuevo enfoque para mejorar la metodología para la detección automática de los principales tópicos presentes en datos textuales propuesta en [1], la cual utiliza técnicas de minería de datos, recursos relacionados con el análisis de sentimientos, y una base de conocimiento multilingüe. La nueva propuesta permite identificar y eliminar los términos con orientación sentimental, con el objetivo de mejorar los resultados del sistema sobre textos de redes sociales.

La idea básica consiste en realizar un filtrado que elimine las palabras que expresen sentimientos durante la etapa de preprocesamiento semántico presente en la metodología. Para ello, se utilizan los recursos léxicos SentiWordNet [2] y WordNet Affect [3] por separado y juntos, para luego comparar los resultados obtenidos. Para la experimentación se han utilizado cuatro conjuntos de datos, los cuales pertenecen a las redes sociales Twitter y Dreamcatchers, en inglés y español

respectivamente.

El resto de este artículo está estructurado de la siguiente forma. Se presenta una revisión de los trabajos previos relacionados con el presente tema en la Sección 2. La Sección 3 describe brevemente la propuesta seleccionada para la detección automática de tópicos. La Sección 4, brinda una descripción detallada del proceso principal que permite analizar la influencia de los términos que expresan sentimientos en la detección automática de tópicos en datos textuales de redes sociales. Seguidamente la Sección 5 presenta los resultados experimentales. Finalmente se presentan las conclusiones, así como los trabajos futuros derivados de la presente investigación en la Sección 6.

II. Antecedentes

La detección de tópicos a partir de grandes volúmenes de textos, ha sido un tema ampliamente analizado en la literatura desde varios puntos vistas. Entre ellos destaca el uso de métodos tales como algoritmos de clasificación, Latent Dirichlet Allocation (LDA) y algoritmos de agrupamiento, entre otros. Para el caso de los algoritmos de clasificación es necesario contar con un conjunto de datos de entrenamiento que permita entrenar el clasificador, mientras tanto LDA y los algoritmos de agrupamiento no resulta necesario contar con un corpus previamente clasificado.

Son muchos los trabajos que podemos encontrar relacionados con la detección de tópicos mediante el uso de algoritmos de agrupamiento jerárquico supervisados y semi-supervisados, no así para los no supervisados. Tales son los casos propuestos en [4]–[6], donde los autores proponen enfoques basados en el uso de información experta, para de esta forma mejorar los resultados en la detección de los principales tópicos.

Desde el punto de vista no supervisado, en [1] se presenta una propuesta para la detección automática de tópicos en datos textuales basada en ontologías. Se utilizó el recurso léxico WordNet Domains [7] con el fin de homogeneizar la representación sintáctica de los conceptos presentes en los textos y así reducir considerablemente la dimensionalidad del problema. Para la experimentación se empleó el conjunto de datos Reuters-21578 el cual contiene textos relacionados con publicaciones de noticias. Los resultados muestran la viabilidad de la propuesta, donde los valores del Coeficiente de Silueta son mejores cuando se aplica la metodología propuesta.

Se debe destacar que aunque los textos de Reuters-21578 constituyen datos reales, son textos largos, están bien elaborados y además pertenecen a un dominio restringido. Si tenemos en cuenta que los textos presentes en redes sociales son textos cortos y los usuarios principalmente expresan sus sentimientos sobre un tema determinado, dichos textos deberían ser preprocesados de forma diferente con el fin de extraer los tópicos presentes.

En redes sociales, la detección de tópicos ha sido extensamente utilizada para el análisis de datos textuales. Muchas han sido los soluciones que han aparecido para el análisis textual en redes sociales, tales como el análisis de sentimientos [8], el filtrado de contenidos [9], [10], el modelado de los intereses del usuario [11], así como el seguimiento de eventos de interés [4], [12]. En [13] se realiza una comparación entre el contenido de los textos de Twitter con un medio de comunicación tradicional, el New York Times. Para ello se utiliza el modelado de tópicos sin supervisión utilizando el modelo Twitter-LDA, para descubrir dichos tópicos en mensajes cortos.

Por otra parte, son varios los trabajos que presentan un modelo donde se fusionan la detección y análisis de tópicos con el análisis de sentimientos [14]–[18]. En todos ellos la detección de tópicos se lleva a cabo sin tener en cuenta la influencia que tienen los términos con determinada orientación sentimental en dicha tarea.

En el presente trabajo se analiza la influencia de los términos con orientación sentimental en la detección de tópicos en redes sociales. Para ello se aplica un filtro durante el preprocesamiento semántico con el objetivo de eliminar los términos que expresan sentimientos, los cuales pueden introducir ruido en la detección de tópicos. Esta propuesta es totalmente novedosa ya que a diferencia de los trabajos mencionados donde se fusiona la detección de tópicos con el análisis de sentimientos, en este caso lo que se hace es descartar los términos de sentimientos, permaneciendo sólo los términos que aportan información útil para la detección automática de los principales tópicos.

III. Descripción de la propuesta para la detección de tópicos

Como se mencionó anteriormente, en este artículo se analiza la influencia de los términos con orientación sentimental en la detección automática de tópicos en redes sociales. Teniendo en cuenta que los datos textuales de redes sociales se encuentran escritos de una manera más coloquial, y los usuarios tienden a expresar sus sentimientos y opiniones sobre determinados productos, servicios, entidades, atributos de estos, etc., resulta útil detectar y eliminar aquellos términos con una determinada orientación sentimental, ya que dichos términos no aportan información útil para la detección de tópicos.

A continuación, se presenta un resumen de cada una de las fases del sistema propuesto, para más detalles ver [1]. Para el caso específico de la fase de Preprocesamiento semántico, se ha resaltado el filtro que permite identificar y descartar los términos de sentimientos, el cual constituye el principal aporte del presente trabajo, y será explicado detalladamente en la Sección IV.

Preprocesamiento sintáctico

Una de las fases fundamentales en la detección de tópicos es el preprocesamiento sintáctico, el cual consiste en una limpieza sintáctica donde se aplican filtros a los datos textuales para facilitar su procesamiento automático. Primero, son ejecutados los procesos de etiquetado de categoría gramatical y de reconocimiento de entidades, dichos procesos se realizan con las herramientas Stanford POS [19] y Stanford NER respectivamente [20]. Luego se aplica el filtro de tokenización y los filtros necesarios para eliminar los términos que pertenecen al conjunto de palabras vacías, los que no son identificados

como sustantivos por el etiquetador gramatical, los que son identificados como sustantivos propios por el identificador de entidades, así como los aquellos términos que no se encuentren en la base de conocimiento Multilingual Central Repository (MCR) [21], ya que todos ellos no aportan información útil para la detección de los tópicos.

Preprocesamiento semántico

Una vez que los textos han sido preprocesados sintácticamente, se procede con el análisis semántico. En nuestro caso, el objetivo del preprocesamiento semántico, es el de homogeneizar la representación sintáctica de los conceptos presentes en el texto. Lo que se hace es sustituir las etiquetas de WordNet Domains [7] con las que han sido etiquetados los sentidos de WordNet por los términos presentes en los textos originales. Como ya se ha mencionado, en este trabajo se desea analizar la influencia de los términos con orientación sentimental en detección de tópicos, por tal motivo en la Sección IV, se explica en profundidad el proceso relacionado con la identificación y eliminación de los términos con orientación sentimental.

Agrupamiento Jerárquico

Una vez homogeneizados los textos, se procede a realizar el agrupamiento jerárquico de los textos a partir de las etiquetas de WordNet Domain. Para representar las características se ha utilizado el enfoque propuesto en [1].En este artículo sólo se analizará el algoritmo de agrupamiento jerárquico Complete Link utilizando como medida de similitud la distancia del coseno.

Etiquetado de grupos

Cuando termina la fase de agrupamiento, se realiza el proceso de selección de etiquetas de los grupos, la cual constituye una tarea de gran importancia, sobre todo en aplicaciones relacionadas con el análisis de datos, donde el usuario final necesita conocer de qué trata determinado grupo [22]. En el presente trabajo se ha utilizado la Media Aritmética para determinar las etiquetas más relevantes de cada grupo de textos.

IV. Influencia de los sentimientos en la detección de tópicos

En esta sección se describe el principal aporte del presente trabajo, que consiste en detectar y descartar los términos con orientación sentimental (positiva o negativa) con el fin de mejorar la detección de tópicos.

IV-A. Análisis de los términos de sentimientos en la detección de tópicos

Para detectar los términos con orientación sentimental, se han utilizado los recursos SentiWordNet [2] y WordNet Affect [3], los cuales están basados en WordNet y permiten determinar si un término en un contexto determinado expresa algún tipo de sentimiento. El primer paso sería desambiguar el término, así de esta forma se conoce el verdadero significado del término en cuestión y finalmente determinar si tiene o no orientación sentimental.

Se debe resaltar que la idea del filtro aplicado para detectar los términos que expresan sentimientos, es totalmente novedosa, ya que permite separar los términos con información relevante para la detección de tópicos, de aquellos términos vinculados con algún tipo de sentimiento.

SentiWordNet : Es un recurso léxico creado especialmente para tareas relacionadas con la clasificación de sentimientos, así como en aplicaciones basadas en la minería de opinión [2]. Constituye una versión mejorada de SentiWordNet 1.0 [23] y se encuentra disponible públicamente para propósitos de investigación. SentiWordNet es el resultado de asignar a todos los sentidos de WordNet tres valores numéricos que indican el valor de polaridad (positivo, negativo y neutro), y dichos valores están en el rango [0,1] [2].

En nuestro caso, una vez desambiguado cada término, determinamos el valor positivo y negativo asignado en SentiWordNet al sentido correspondiente para cada uno de los términos analizados. En caso de que el sentido presente un valor positivo mayor que cero o un valor negativo mayor que cero, este término queda totalmente descartado y no se tiene en cuenta para el posterior análisis mediante el cual se detectan los principales tópicos abordados en los textos.

WordNet Affect: Es un recurso lingüístico para la representación léxica del conocimiento afectivo. Al igual que SentiWordNet, está basado en WordNet. Fue creado mediante la selección y etiquetado de sentidos de WordNet, que representan conceptos afectivos, y luego fue extendido mediante el uso de las reacciones entre términos y conceptos presentes en WordNet [3].

A diferencia de SentiWordNet, WordNet Affect no presenta la polaridad de los distintos conceptos, en su lugar etiqueta los sentidos con un conjunto de 1,903 categorías que constituyen estados afectivos. Actualmente consta de 2,874 sentidos y 4,787 términos [3]. La forma en la que ha sido utilizado es similar al anterior, cuando los términos son desambiguados, son descartados si su significado en el contexto actual ha sido etiquetado con algún estado mental en WordNet Affect.

De forma general, cuando los términos que expresan sentimientos son descartados, independientemente del recurso léxico utilizado, los términos más frecuentes que pertenecen a un tópico determinado son más afines al mismo. Además se debe señalar que son muchos los textos que tras aplicar el filtrado de sentimientos no pueden ser procesados, ya que son descartados todos sus términos. Esto se debe también en gran medida a que la gran mayoría de los textos de redes sociales solamente expresan opiniones sobre cierta temática.

V. Experimentos

A continuación demostraremos de forma experimental la validez de nuestra propuesta. Como se ha mencionado, no contamos con información previa de los tópicos presentes en los textos(categorías o etiquetas), por tal motivo debemos utilizar una medida no supervisada. En este caso se ha seleccionado el

Coeficiente de Silueta [24], que permite determinar la cantidad de grupos para la que el algoritmo brinda un mejor resultado.

Los conjuntos de datos para evaluar el sistema pertenecen a dos redes sociales Twitter y Dreamcatchers. Lo primera es una de las redes sociales más populares y de las más utilizadas en investigaciones relacionadas con el tema. La segunda ha sido desarrollada bajo un enfoque colaborativo entre sus miembros, y se cuenta con la base de datos que le da soporte. Los datos seleccionados de Twitter y Dreamcatchers se encuentran en inglés y español respectivamente, demostrando que la propuesta es independiente del idioma.

V-A. Conjuntos de datos

Se han seleccionado cuatro conjuntos de datos Tabla I, pertenecientes a Twitter y a la Red Social Dreamcatchers. Los datos de Twitter se obtuvieron del conjunto de entrenamiento de Sentiment140[1], se encuentran en formato CSV y consta de seis campos, entre ellos el texto de los *tweets* el cual será utilizado en el presente trabajo. Se debe mencionar que se seleccionaron estos datos por estar orientados al Análisis de Sentimientos, y constituyen una fuente de gran importancia para la experimentación.

Por otra parte, se cuenta con la base de datos de Dreamcatchers, con un total de 61 tablas. La información recogida es toda la relacionada con los datos personales y de afiliación del usuario, así como, las interacciones que realiza en su perfil y con otros usuarios. En este caso utilizaremos los *Comentarios* de los usuarios.

Se debe mencionar que además del idioma, los textos de Twitter difieren con los de Dreamcatchers en los siguientes aspectos:

- Los textos de Dreamcatchers aunque abordan diversos tópicos, principalmente constituyen temas relacionados con el ámbito universitario, ya que los usuarios de dicha red social son estudiantes de la Universidad de Camagüey.
- La longitud de los textos de Twitter que hemos utilizado está restringida a 140 caracteres, mientras los textos de Dreamcatchers no presentan ninguna restricción.

V-B. Evaluación

En esta sección, se explica el procedimiento para evaluar el funcionamiento de la metodología para la detección automática de tópicos aplicando el filtro para eliminar los términos que expresan sentimientos con los distintos recursos y sin aplicarlo.

Para ello se ha utilizado como medida el Coeficiente de Silueta [24] 1. Esta medida permite analizar la calidad de los grupos creados por los algoritmos de agrupamiento jerárquico. Los valores de este coeficiente están en el intervalo de [-1;1], siendo 1 el valor ideal que deben alcanzar los distintos algoritmos de agrupamiento jerárquico.

$$S(i) = \frac{a(i) - b(i)}{max\left\{a(i), b(i)\right\}} \quad (1)$$

[1] http://www.sentiment140.com/

La experimentación se ha llevado a cabo utilizando el método Complete Link y realizando cortes para las siguientes cantidades de grupos (17, 25, 40, 60, 80, 100 y 120), y para cada caso se ha determinado el valor del Coeficiente de Silueta cuando no se aplica el filtro para eliminar los sentimientos, así como cuando es aplicado haciendo uso de los distintos recursos. En todos los casos se ha utilizado como medida de similitud la distancia del coseno.

V-C. Resultados y discusión

En las Tablas II-V se muestran los valores del Coeficiente de Silueta de cada conjunto experimental y las distintas cantidades de grupos, tanto cuando no se aplica el filtro para eliminar los términos que expresan sentimientos **(NF)** y cuando es aplicado utilizando SentiWordNet **(SWN)**, WordNet Affect **(WA)** and la combinación de ambos **(SWN-WA)**. Se puede observar a simple vista que cuando se aplica el filtrado de sentimientos, se obtienen mejores resultados que cuando no es aplicado (valores en negrita).

Con el fin de realizar un análisis más detallado de los resultados, se ha realizado un análisis estadístico para así poder determinar si existen diferencias significativas entre los valores obtenidos para los distintos número de grupos, los recursos utilizados y la fuente de los datos textuales. Las Figuras 1 y 2 muestran las gráficas del Coeficiente de Silueta con respecto a la cantidad de grupos y el recurso utilizado para detectar los términos con orientación sentimental respectivamente, utilizando en ambos casos la prueba de Kruskal-Wallis [25]. A continuación se resumen las conclusiones del análisis:

- A partir de 60 grupos en adelante los valores del Coeficiente de Silueta se estabilizan, mostrando diferencias significativas con las cantidades anteriores.
- Existe una notable diferencia entre los resultados obtenidos con los recursos **SWN** y **SWN-WA**, y los obtenidos con el recurso **WA** y cuando no se aplica ningún filtro para detectar los términos con orientación sentimental, aunque no existen diferencias significativas entre **SWN** y **SWN-WA** si tenemos en cuenta que la utilización de **SWN-WA** conlleva el uso de otro recurso léxico.

Por otra parte, la Figura 3 muestra la gráfica del Coeficiente de Silueta con respecto a la red social de la cual provienen los textos. Se realizó la prueba de Wilcoxon [26] y se puede concluir que existe una gran diferencia para las redes sociales utilizadas, pues Dreamcatchers brinda los mejores resultados. Esto se debe en gran medida a que gran parte de los usuarios de esta red pertenecen a un contexto universitario, por lo que el dominio de conversación es más restringido.

Tabla I
DESCRIPCIÓN DE LOS CONJUNTOS DE DATOS

Conjunto	Cantidad de documentos	Fuente	Idioma	Intervención	Cantidad de términos diferentes	Cantidad total de términos
Conjunto 1	5000	Twitter	Inglés	Tweet	3189	12915
Conjunto 2	10000	Twitter	Inglés	Tweet	4597	25634
Conjunto 3	5000	Dreamcatchers	Español	Comentario	1661	8851
Conjunto 4	10000	Dreamcatchers	Español	Comentario	2218	17141

Tabla II
COEFICIENTE DE SILUETA DEL CONJUNTO 1

Recurso/ Grupos	17	25	40	60	80	100	120
NF	0.0	0.09	0.19	0.23	0.24	0.24	0.22
SWN	0.03	0.1	0.2	0.26	**0.28**	0.27	0.27
WA	0.0	0.09	0.18	0.22	0.23	0.24	0.22
SWN-WA	0.02	0.09	0.21	0.26	0.27	0.26	0.27

Tabla III
COEFICIENTE DE SILUETA DEL CONJUNTO 2

Recurso/ Grupos	17	25	40	60	80	100	120
NF	0.04	0.09	0.14	0.21	0.22	0.23	0.22
SWN	0.06	0.09	0.15	0.25	0.26	**0.27**	0.26
WA	0.05	0.08	0.12	0.21	0.23	0.23	0.22
SWN-WA	0.06	0.08	0.15	0.24	**0.27**	**0.27**	0.25

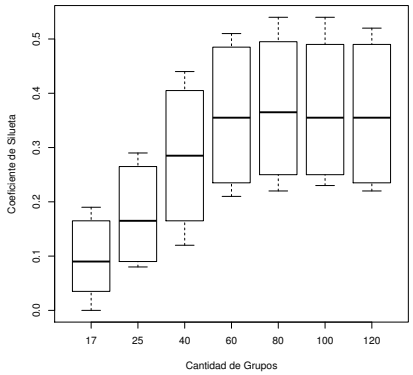

Figura 1. Gráfica entre el Coeficiente de Silueta y el número de grupos

Tabla IV
COEFICIENTE DE SILUETA DEL CONJUNTO 3

Recurso/ Grupos	17	25	40	60	80	100	120
NF	0.13	0.24	0.37	0.45	0.45	0.44	0.44
SWN	0.19	0.29	0.44	0.51	0.52	0.52	0.52
WA	0.12	0.23	0.37	0.45	0.46	0.45	0.45
SWN-WA	0.16	0.27	0.44	0.51	**0.53**	0.51	0.52

Tabla V
COEFICIENTE DE SILUETA DEL CONJUNTO 4

Recurso/ Grupos	17	25	40	60	80	100	120
NF	0.17	0.24	0.36	0.45	0.46	0.45	0.45
SWN	0.15	0.28	0.43	0.51	**0.54**	**0.54**	0.52
WA	0.18	0.26	0.38	0.46	0.47	0.47	0.46
SWN-WA	0.17	0.28	0.44	0.51	**0.54**	0.53	0.52

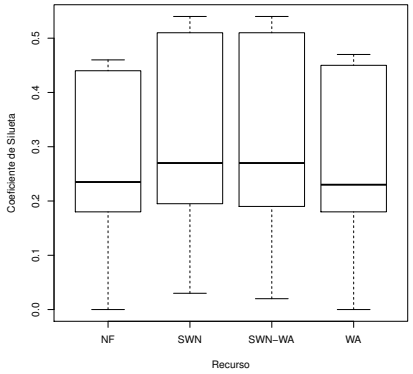

Figura 2. Gráfica entre el Coeficiente de Silueta y los recursos utilizados

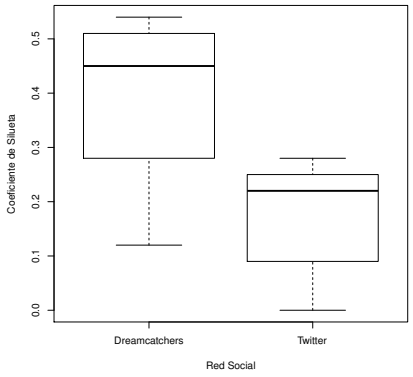

Figura 3. Gráfica entre el Coeficiente de Silueta y las redes sociales utilizadas

VI. Conclusiones y Trabajos Futuros

En este artículo se presenta una nueva propuesta para la detección automática de tópicos en textos de redes sociales. Para ello se ha incorporado un filtro durante el preprocesamiento semántico de los textos, permitiendo detectar y eliminar los términos que presenten una orientación sentimental, ya que estos no constituyen información relevante para la detección de tópicos. Con este fin, fueron utilizados los recursos SentiWordNet y WordNet Affect por separado y combinándolos.

Los experimentos realizados tanto con Twitter como con Dreamcatchers, permiten mostrar la viabilidad del sistema. Se ha experimentado sin aplicar el filtro para eliminar los sentimientos y aplicándolo con dos recursos léxicos relacionados

con el análisis de sentimientos. En cada caso se realizaron cortes en las cantidades de grupos ya mencionadas y se calculó el Coeficiente de Silueta. Los resultados alcanzados cuando se aplicó el filtro mejoran los resultados obtenidos cuando no se aplica el filtro. El recurso con el que se obtuvo un mejor rendimiento fue SentiWordNet , pues aunque la combinación de ambos recursos mejora a SentiWordNet en determinados casos, las diferencias no son significativas teniendo en cuenta que se incorporan todos los conceptos de WordNet Affect.

El uso de algoritmos de agrupamiento jerárquico, brinda la posibilidad de agrupar los documentos en tópicos o temas y crear una jerarquía de tópicos, la cual puede ser utilizada como jerarquía de una dimensión en un modelo multidimensional y de esta forma facilitar el análisis de los datos de las redes sociales. La metodología propuesta puede ser aplicada a distintas redes sociales independientemente del idioma y de los temas tratados en dichas redes, gracias a la base de conocimiento utilizada (MCR). Se debe mencionar que MCR está basado en el recurso WordNet, integra recursos como WordNet Domains, Base Concepts, Top Ontology y la ontología AdimenSUMO y además mediante un índice (Inter-Lingual-Index ILI) integra wordnets de seis idiomas diferentes: Inglés, Español, Catalán, Euskera, Gallego y Portugués [21].

Como se mencionó anteriormente el tema tratado en la presente investigación deja abierta la posibilidad para realizar extensiones tales como incluir la jerarquía de tópicos obtenida como una nueva dimensión en un modelo multidimensional, permitiendo un mejor análisis de los datos principalmente en redes sociales, y luego de tener creados los tópicos estudiar el proceso relacionado con la de detección automática de tópicos para un nuevo texto.

Referencias

[1] K. Gutiérrez-Batista, J. R. Campaña, M.-A. Vila, y M. J. Martin-Bautista, "An ontology-based framework for automatic topic detection in multilingual environments," *International Journal of Intelligent Systems*, vol. 33, no. 7, pp. 1459–1475, 2018. [Online]. Disponible: https://onlinelibrary.wiley.com/doi/abs/10.1002/int.21986

[2] S. Baccianella, A. Esuli, y F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." in *LREC*, vol. 10, 2010, pp. 2200–2204.

[3] R. Valitutti, "Wordnet-affect: an affective extension of wordnet," in *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004, pp. 1083–1086.

[4] L. Chung-Hong, "Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 338 – 13 356, 2012.

[5] A. G. Skarmeta, A. Bensaid, y N. Tazi, "Data mining for text categorization with semi-supervised agglomerative hierarchical clustering," *International Journal of Intelligent Systems*, vol. 15, no. 7, pp. 633–646, 2000. [Online]. Disponible: http://dx.doi.org/10.1002/(SICI)1098-111X(200007)15:7<633::AID-INT4>3.0.CO;2-8

[6] L. Zheng y T. Li, "Semi-supervised hierarchical clustering," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ser. ICDM '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 982–991. [Online]. Disponible: http://dx.doi.org/10.1109/ICDM.2011.130

[7] B. Magnini y G. Cavaglia, "Integrating subject field codes into wordnet." in *LREC*. European Language Resources Association, 2000.

[8] C. Lin y Y. He, "Joint sentiment/topic model for sentiment analysis," in *18th ACM Conference on Information and Knowledge Management 8CIKM09)*. New York, NY, USA: ACM, 2009, pp. 375–384.

[9] J. Duan y J. Zeng, "Web objectionable text content detection using topic modeling technique," *Expert Systems with Applications*, vol. 40, pp. 6094–6104., 2013.

[10] J. Martinez-Romo y L. Araujo, "Detecting malicious tweets in trending topics using a statistical analysis of language," *Expert Systems with Applications*, vol. 40, pp. 2992–3000, 2013.

[11] M. Pennacchiotti y S. Gurumurthy, "Investigating topic models for social media user recommendation," in *20th International Conference Companion on World Wide Web*. New York, NY, USA: ACM, 2011, pp. 101–102.

[12] J. Wu, W. Gao, B. Zhang, J. Liu, y C. Li, "Cluster based detection and analysis of internet topics," in *4th International Symposium on Computational Intelligence and Design, ISCID 2011*, vol. 2, 2011, pp. 371–374.

[13] W. X. Zhao, J. Weng, J. He, E.-P. Lim, y H. Yan, "Comparing twitter and traditional media using topic models," in *33rd European conference on advances in information retrieval (ECIR11)*. Berlin, Heidelberg: Springer-Verlag., 2011, pp. 338–349.

[14] C. Lin, Y. He, R. Everson, y S. Ruger, "Weakly supervised joint sentiment-topic detection from text," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 6, pp. 1134–1145, June 2012.

[15] J. Sowmiya y S. Chandrakala, "Joint sentiment/topic extraction from text," 2015, pp. 611–615, cited By 0.

[16] Y. Rao, Q. Li, X. Mao, y L. Wenyin, "Sentiment topic models for social emotion mining," *Information Sciences*, vol. 266, pp. 90 – 100, 2014. [Online]. Disponible: http://www.sciencedirect.com/science/article/pii/S002002551400019X

[17] K. Cai, S. Spangler, Y. Chen, y L. Zhang, "Leveraging sentiment analysis for topic detection," in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*, vol. 1, Dec 2008, pp. 265–271.

[18] X. Ding, L. Zhang, Y. Tian, X. Gong, y W. Wang, "Dynamic topic detection model by fusing sentiment polarity," vol. 159, 2015, pp. 65–71, cited By 0.

[19] K. Toutanova, D. Klein, C. D. Manning, y Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, ser. NAACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 173–180. [Online]. Disponible: http://dx.doi.org/10.3115/1073445.1073478

[20] J. R. Finkel, T. Grenager, y C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 363–370. [Online]. Disponible: http://dx.doi.org/10.3115/1219840.1219885

[21] A. Gonzalez-Agirre, E. Laparra, y G. Laparra, "Multilingual central repository version 3.0," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[22] C. D. Manning, P. Raghavan, y H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[23] A. Esuli y F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, 2006, pp. 417–422.

[24] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, Nov. 1987. [Online]. Disponible: http://dx.doi.org/10.1016/0377-0427(87)90125-7

[25] W. Kruskal y W. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, pp. 583–621, 1952.

[26] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.