ISSN 1870-4069

Identificación del perfil de usuario en Twitter utilizando recursos semánticos

J. Víctor Carrera-Trejo, Miguel Á. Álvarez-Carmona, Luis Villaseñor-Pineda

Instituto Nacional de Astrofísica, Óptica y Electrónica,
Laboratorio de Tecnologías del Lenguaje,
México

jvcarrera@{ipn, inaoep}.mx, {miguelangel.alvarezcarmona, villasen}@inaoep.mx

Resumen. Los *hashtags*, las menciones de usuario o las direcciones *url* compartidas en Twitter son características de esta red social que pueden ser útiles para observar los intereses de un usuario. El presente trabajo evalúa la posibilidad de usar este tipo de características para identificar el perfil del usuario. No obstante, dada la variabilidad y especificidad de dichas características no es posible usarlas directamente, por lo que es necesario determinar el concepto asociado a través de recursos semánticos. En el caso particular del trabajo mostrado en este artículo se comprobó la utilidad del grafo de conocimiento de *Google*. Dicho grafo se construye a partir de los documentos públicos en Internet, reuniendo y asociando todo tipo de información de manera dinámica. La evaluación del método propuesto se realizó usando el corpus en inglés del PAN 2014. Los resultados alcanzados evidencian que la información de estas características puede aprovecharse en el perfilado de autores.

Palabras clave: clasificación no temática, perfilado de autor, Google, Twitter, grafo de conocimiento, hashtags, PAN 2014.

Author Profiling in Twitter using Semantic Resources

Abstract. Hashtags, user mentions or url addresses shared on Twitter are features of this social network that can be useful to observe the interests of a particular user. The present work evaluates the possibility of using this kind of characteristics to identify the user's profile. However, given the variability and specificity of these characteristics, it is not possible to use them directly, so it is necessary to determine the associated concept or meaning through semantic resources. In the particular case of the work presented in this article, the use of the Google's knowledge graph was verified. This kind of graph is built using public documents found on the internet, gathering and associating dynamically all kind of information. The evaluation of the proposed method was carried out

using the english corpus of the PAN 2014. The results obtained show that the information of these characteristics can be used in the author profiling.

Keywords: Twitter, knowledge graph, hashtags, classification, Google, author profiling, PAN 2014, concepts.

1. Introducción

El uso de internet ha generado nuevas formas de comunicarse, en las cuales se hace uso de diferentes aplicaciones para compartir información, entre estas aplicaciones se pueden encontrar blogs, microblogs, foros, etc. En ellas el principal objetivo es compartir información relacionada con diferentes temáticas o tópicos, ejemplos de estos son noticias, opiniones, revisiones de productos o servicios, etc. En algunas ocasiones la información se comparte de forma anónima, por lo que se desconocen datos acerca de quien realiza la publicación. Datos que podrían ser de interés por distintas razones; por ejemplo, para mercadotecnia o seguridad.

Dentro de las diferentes áreas de investigación de procesamiento del lenguaje natural existen diferentes tareas enfocadas a descubrir información relacionada con el autor de un texto [1], la primera de ellas, denominada "*author profiling*" o perfil de autor, se enfoca en identificar rasgos del autor de un documento como su edad, su género, su profesión, entre otros. Otra tarea se denomina "*author identification*" o identificación de autor, en la cual se intenta identificar al autor de un texto anónimo de entre un conjunto de autores dado.

Como se puede intuir, ambas tareas se pueden resolver con un enfoque de clasificación supervisado extrayendo características de textos con autores conocidos y a partir del análisis de estas características resolver la tarea que se requiera.

Existen diferentes tipos de características que pueden ser extraídas de los textos utilizados. Dentro de estas características [2,23] se pueden encontrar combinaciones léxicas, aquellas basadas en el estilo, como pueden ser signos de puntuación, el uso de mayúsculas, la longitud de las frases, etc., características semánticas [3,4] obtenidas mediante el uso de algoritmos de semántica latente o aquellas extraídas a partir del uso de etiquetadores de partes de la oración (*part-of-speech taggers*, POS), que permiten conocer la categoría gramatical a la que pertenece una palabra dentro de un texto [24].

En principio, la tarea de "*author profiling*" se aplicó a textos formales, como noticias o libros; sin embargo, en los últimos años se ha tratado de determinar características de usuarios en redes sociales a través de los textos que ellos mismos comparten.

Una de las plataformas más utilizadas y estudiadas actualmente es Twitter, la cual es una red de servicio de microblogging que cuenta con más de 330 millones de usuarios, los cuales publican más de 500 millones de mensajes diariamente [7], también conocidos como tweets, los cuales tienen una longitud limitada de caracteres, en los que se pueden incluir enlaces (*url*) a sitios webs externos, imágenes o videos que puedan ser vistos por otros usuarios que tengan acceso al microblog. Esta red social ofrece una gran fuente de información y ha sido motivo de muy diversas investigaciones, entre las que se pueden encontrar: minería de opinión, análisis de

sentimientos, predicción de resultados electorales, estudios de mercado, análisis de desastres, etcétera.

Esta red social presenta un tipo propio de características, conocidas como interactivas [6], ya que éstas ofrecen un medio de interacción con otros usuarios, compartiendo menciones de otros usuarios, contenido mediante direcciones url o hashtags. Haciendo uso de los hashtags los usuarios comparten información temática, la cual es representada por el texto que el mismo usuario define. Sin embargo, muchas veces la interpretación del hashtag depende del contexto, del usuario o del dominio en que se comparte. De ahí que a pesar de ser "etiquetas" sobre un tema o un concepto, se necesite de un recurso para su interpretación. Por este motivo es necesario contar con alguna herramienta o recurso semántico a partir del cual extraer la información que un hashtag representa. Una herramienta que contiene dicha información es el grafo de conocimiento de Google [8,9]. Este grafo brinda información en general y puede ser usado para recuperar el concepto descrito por un hashtag.

El objetivo de este trabajo es analizar si el uso de características propias de Twitter, principalmente las url y hashtags compartidos, en conjunto con el grafo de conocimiento de Google [8,9], son útiles en la tarea de perfil de autor.

El resto de este artículo se estructura de la siguiente manera: En la sección 2 se presenta el trabajo relacionado; la sección 3 describe el corpus y el método propuesto; en la sección 4 se muestran los resultados obtenidos y, finalmente, en la sección 5 se analizan los resultados ofreciendo algunas conclusiones y comentarios acerca de posibles trabajos futuros.

2. Trabajo relacionado

Antes de abordar la descripción de los trabajos relacionados se describen algunos conceptos centrales para el trabajo propuesto.

2.1. Caracterización de la información

Existen diferentes formas de representación de los textos en tareas de clasificación, siendo la más común de ellas el modelo espacio vectorial [13,14,15], en lo cual el documento es representado como un vector, donde los valores de los componentes del vector representan los valores de las características extraídas del documento, por lo que el tamaño del vector corresponde con el número total de características. Estas características son usualmente palabras o algunas de sus invariantes morfológicas, como pueden ser sus lemas, entre otras.

El valor de cada característica se calcula de acuerdo con un tipo de pesado en particular [2,16], entre los que se encuentran principalmente binario, *tf*, *idf* y su combinación. La forma de caracterización más simple es la caracterización binaria, en la cual se debe indicar mediante un valor de "1" si una palabra en particular del vector de características se encuentra dentro de un texto en particular a caracterizar o bien se indica mediante un valor de "0" cuando dicha palabra no se encuentra.

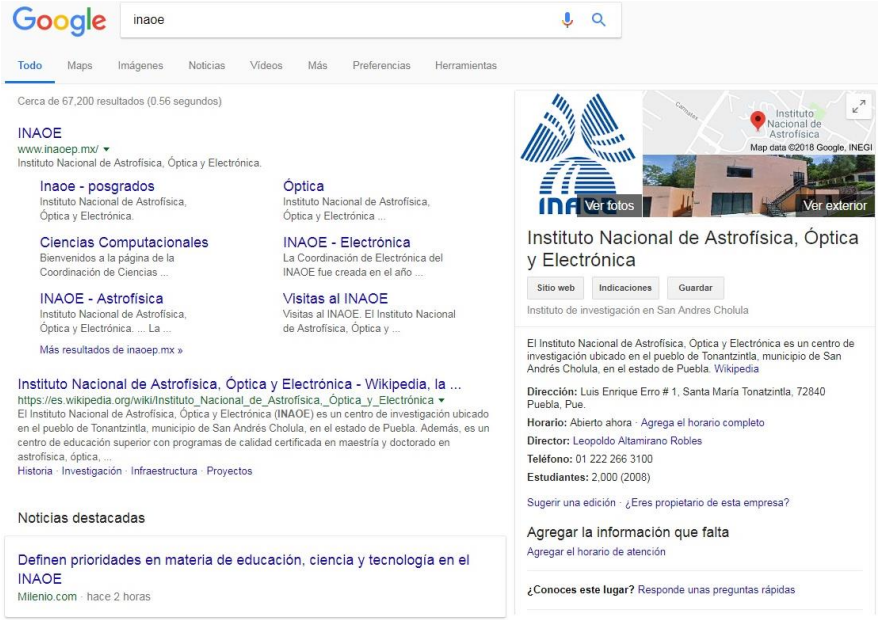

Fig. 1. Grafo de conocimiento para el término "INAOE".

2.2. Clasificación de la información

La clasificación de documentos busca determinar si un documento pertenece a una o varias categorías, también conocidas como clases o etiquetas. La clasificación de textos está basada en técnicas de *machine learning* [17,18], entre estas técnicas se pueden mencionar los clasificadores Naïve Bayes, máquinas de soporte vectorial con diferentes kernels, árboles de decisión, redes neuronales, etcétera.

2.3. El grafo de conocimiento de Google

El grafo de conocimiento de Google [8] es una base de conocimiento creada y utilizada por los servicios de Google. Para su creación se utilizan diferentes fuentes de información; por ejemplo, Freebase, Wikipedia, CIA World Facebook, entre otras. El grafo cuenta actualmente con más de 500 millones de objetos y 3.5 billones de hechos y relaciones entre estos objetos, y gracias a este grafo es posible obtener información relacionada con personas, eventos, lugares etcétera.

Esta herramienta puede ser consultada utilizando el motor de búsqueda de Google de dos formas, utilizando un web browser o un API de programación. En la figura 1 se observa la consulta de la palabra "INAOE", donde la información devuelta por el grafo de conocimiento se encuentra dentro del rectángulo del lado izquierdo. En dicho rectángulo se muestra la información relacionada con el concepto principal al que se asocia el término buscado; por ejemplo, de la consulta realizada, el término "INAOE"

se asocia con un centro de investigación del cual se puede conocer su dirección, horario, nombre de su director, entre otras. Por otro lado, al realizar la consulta utilizando un API de programación [10] es posible obtener el conjunto de conceptos asociados al término, como son, entre otros, una descripción del objeto y su tipo.

Por ejemplo, al realizar consulta de la palabra "INAOE" utilizando el API de Google, se obtienen, entre otros, los siguientes términos relacionados:

1. "name": "14674 INAOE", "description": "Asteroid", "@type": ["Thing"].
2. "name": "National Institute of Astrophysics, Optics and Electronics", "description": "Research institution in San Andres Cholula, Mexico", "@type": ["Thing, "Organization, Place, EducationalOrganization, CollegeOrUniversity"].
3. "name": "Héctor Manuel Moya Cessa", "description": "Physicist", "@type": ["Person, Thing"].
4. "name": "Guillermo Haro", "description": "Mexican astronomer", "@type": ["Person, Thing"].
5. "name": "Atacama Cosmology Telescope", "description": "Observatory in Chile", "@type": ["Place, Thing"].

2.4. Trabajos relacionados

Diversos trabajos se enfocan en buscar el mejor conjunto de características que permitan resolver la tarea de *autor profiling*; por ejemplo, en [12] se presentan diferentes trabajos que hacen uso de diferentes tipos de características: de estilo (signos de puntuación, tamaños de las sentencias, número caracteres, etc.), léxicas (*n*-gramas o bolsa de palabras), temáticas (utilizando recursos como LIWC) o bien, características basadas en representaciones distribucionales identificando relaciones entre términos, documentos, perfiles y sub-perfiles [25].

En otros trabajos, como en [6], los autores se enfocan a analizar características que surgen a partir de los textos de twitter, denominándolas "*social behavioral biometrics*", como son los hashtags, las menciones a usuario o los url compartidos para así poder inferir datos de los usuarios como patrones de comportamiento, de comunicación, entre otros. En [19] los autores realizan la clasificación temática de los hashtags de un corpus de twitter, utilizando una máquina de soporte vectorial como clasificador, utilizando un etiquetador temático y Wikipedia.

Finalmente, en [9] se presenta el concepto de baúl del conocimiento *(knowledge vault, KV)*, el cual se refiere a construir un gran repositorio de información a partir de la consulta online de diferentes bases de conocimiento estructuradas, como son, entre otras, Wikipedia, Freebase, YAGO, Satori de Microsoft, incluyendo el grafo de conocimiento de Google. En el KV se almacena información adicional de un concepto a partir de su búsqueda en las diferentes bases de conocimiento, esta información es almacenada en forma de una tripleta (sujeto, predicado, objeto). Por ejemplo, para los términos "Barack Obama", el sistema almacena la tripleta ("Barack Obama", "Place of birth", "Honolulu"), entre muchas otras que describen el concepto "Barack Obama".

Como se observa en párrafos anteriores, existen diferentes trabajos que proponen el uso de diversos tipos de características para identificar los temas de interés para un autor y con ellos, en conjunto de información estilística, intuir el perfil de dicho autor.

En el caso particular de Twitter, los hashtags son etiquetas propuestas por los usuarios para nombrar un tema. Desafortunadamente, el hashtag no puede interpretarse directamente, de ahí que para obtener las características temáticas sea necesario el uso

Tabla 1. Conceptos extraídos a partir del grafo de conocimiento.

Hashtag	Conceptos
#facebook	video_sharing_company technology_company social_network_company drama_series book_by_david_kirkpatrick song_by_rhett_and_link broadcasting_television_network
#graffiti	visual_art_type 1973_film studio_album_by_chris_brown studio_album_by_led_zeppelin video_game_series 1979_film 1990_film american_singer-songwriter comic_magazine_series company
#beirut	capital_of_lebanon, band, university_in_beirut_lebanon, country_in_the_middle_east city_in_oregon
#shazam	fictional_superhero television_program fictional_character 2019_film company rock_band american_television_show comic_series studio_album_by_the_move
#concacaf	league football_competition tournament competition sports_association soccer_team
#japan	country_in_east_asia soccer_team music_company state capital_of_japan war internet_services_company
#youtube	video_sharing_company orchestra swedish_comedian award_ceremony television_company event court_case american_sitcom public_university_in_milton_keynes_england

recursos adicionales. El presente trabajo analiza el uso de un recurso semántico: el grafo de conocimiento de Google, para evidenciar el o los conceptos detrás de un hashtag. A través de esta transformación se espera impactar el proceso de clasificación en la tarea de perfil de autor.

3. Metodología

Para este trabajo se utilizó el corpus del PAN-2014, el cual es descrito en la sección 3.1. En los apartados subsecuentes se describen los pasos utilizados en la metodología presentada en este trabajo: en la sección 3.2 se describe el proceso de tokenización y la obtención de los conceptos a partir de la consulta del grafo de conocimiento de Google; finalmente, en la sección 3.3 se indica el proceso para construir los diferentes conjuntos de características mencionados en la sección 3.2 y su clasificación, calculando los valores de las medidas precisión, cobertura (o recall) y *f1* para su posterior análisis y comparación.

3.1. Corpus PAN 2014

Una de las tareas del PAN-2014 [11,12] es la de *author profiling*, en la que el objetivo es: dado un documento identificar el género y la edad de su autor. Para llevar a cabo la tarea se construyó un corpus que incluye textos de blogs, revisiones de hoteles y twitter, en dos idiomas: inglés y español. Con base en este corpus se extrajo la parte relacionada con twitter en idioma inglés, y se obtuvieron 306 archivos.

En el caso de género, el corpus se encuentra balanceado ya que el corpus cuenta con 153 archivos para género femenino y 153 para género masculino. Sin embargo, en el

Tabla 2. Características.

Característica	Descripción
normales	Tokens extraídos del corpus, en minúsculas, se incluyen las stopwords.
lematizadas	Tokens que pertenecen a la característica "normales", se lematizan todos ellos excepto las características de twitter.
urls	Urls extraídas del corpus.
dominios	Dominios de las urls extraídas del corpus, se sustituyen cada una de las urls por estos dominios.
hashtags	Hashtags extraídos del corpus.
grafo	Conceptos indicados por el grafo del conocimiento para los hashtags, se sustituyen cada uno de los hashtags por estos conceptos.
usuarios	Usuarios mencionados en los tweets.

caso de edad el corpus no se encuentra balanceado, ya que se tienen 5 clases con diferentes números de archivos: se tiene 20 archivos que corresponden a autores cuyo rango de edad se encuentra entre 18 y 24 años; 88 corresponden a aquellos entre 25 y 34; 130 se relacionan a personas entre 35 y 49; 60 a aquellos entre 50 y 64 y, finalmente, solamente 8 corresponden a autores con más de 65 años.

3.2. Caracterización del corpus y grafo del conocimiento

Para realizar la caracterización del corpus se tomó cada uno de los 306 archivos y utilizando la herramienta de software *Ark Tokenizer* [5], se separó cada archivo en tokens, incluyendo cada uno de ellos en cada uno de los 5 conjuntos de características definido por la herramienta.

Se lematizaron cada uno de los tokens obtenidos, excepto aquellos que pertenecen al conjunto denominado "*twitter/online specific*" ya que contiene características específicas de twitter, utilizando para ello la herramienta *CoreNLP* de Stanford [20], obteniendo así conjuntos de características lematizadas y normales.

Utilizando el conjunto de características de twitter se separaron las direcciones url y a partir de cada una de ellas se separó el dominio de la url; por ejemplo, de la siguiente url compartida "*http://bit.ly/j5I178*" el dominio extraído es "*bit.ly*", con esto se obtienen dos conjuntos, uno de ellos normales y el otro denominado dominios.

Para extraer la información relacionada con cada uno de los hashtags aplicando el grafo del conocimiento, se desarrolló una aplicación en el lenguaje de programación Python [21], donde se realizó la consulta de cada hashtag en el grafo. De la información devuelta para cada consulta se extrajeron los datos mostradas en el campo "*description*". Algunos conceptos extraídos se muestran en la tabla 1, en esta tabla se observa, por ejemplo, que para el hashtag #facebook, el grafo de conocimiento de Google lo identifica como una red social o una compañía en la que se comparten videos, aunque también lo identifica en menor medida como una canción o un libro, otro ejemplo son los conceptos relacionados con #japan, éste es relacionado con un país o un equipo de soccer.

Finalizado el proceso de caracterización del corpus, se construyeron los diferentes conjuntos de características, las cuales se describen en la tabla 2.

3.3. Construcción y clasificación de los conjuntos de características

Considerando las características mostradas en la tabla 2 y la descripción del corpus que se realizó en la sección 0, se construyeron los diferentes archivos para la tarea de clasificación. En el caso de la clasificación para género se consideró un esquema *10-fold cross* validation, construyendo un archivo de entrenamiento y uno de validación por *fold.* Por otro lado, para el caso de edad se utilizó un esquema *5-fold cross validation*, ya que en este caso las 5 clases no se encuentra balanceadas y se buscó incluir principalmente archivos de todas las clases tanto en el archivo de entrenamiento como en el de prueba.

Por otro lado, los vectores de caracterización de cada uno de los archivos se construyeron utilizando un esquema de pesado binario y considerando los tokens que se encuentren en dos o más archivos.

Finalmente, utilizando la librería *Scikit-Learn* [22] de Python [21] se realizó la clasificación de los diferentes *folds.* Para medir la eficacia de la representación propuesta se aplicaron las medidas *precision y recall* para por clase, así como los promedios generales. Para el cálculo de la medida F1 se utilizó la ecuación mostrada en (1), tanto por clase, como de forma general:

$$F1 = 2 * \frac{precision * recall}{precision + recall}. \quad (1)$$

4. Experimentos

En esta sección se mostrarán los resultados obtenidos por la representación propuesta en este trabajo. Como clasificador se utilizó el algoritmo SVM de Scikit-Learn [22].

4.1. Bolsa de palabras

Para la bolsa de palabras, en el caso del género se obtuvieron los resultados que se muestran en la tabla 3. Como se puede observar, el mejor resultado se obtuvo al utilizar el conjunto de los lemas de las palabras del corpus y sustituyendo las url compartidas por los dominios de ellas, que se denomina *lematizadas_dominios.* En segundo término, se encuentran los tokens del conjunto de características normales, las url que se incluyen dentro de este conjunto de características que se encuentran como en el corpus, este conjunto es llamados normales_*urls.*

A continuación, se encuentra el conjunto de características normales donde se utilizaron los dominios de las url en lugar de todas ellas, dicho conjunto se identifica como *normales_dominios* y, finalmente, el conjunto llamado *lematizadas_urls* en el que se incluyen los lemas de los tokens y donde las url compartidas no sufren ninguna modificación, una descripción más detallada del conjunto de características se encuentra en la tabla 2. Finalmente, cabe hacer notar que los valores de F1 no varían demasiado entre los diferentes conjuntos.

Tabla 3. Resultados macro de clasificación para género.

	Precision	**Recall**	**F1**
lematizadas_dominios	0.81075	0.8	0.79745
normales_urls	0.7985	0.78333	0.78003
normales_dominios	0.80205	0.78333	0.77955
lematizadas_urls	0.78832	0.78	0.77772

Tabla 4. Resultados macro de clasificación para edad.

	Precision	**Recall**	**F1**
lematizadas_dominios	0.39064	0.45	0.40389
normales_urls	0.36542	0.44333	0.3928
lematizadas_urls	0.37096	0.43667	0.39162
normales_dominios	0.35124	0.43	0.38097

En el caso de la clasificación por edad, los resultados se muestran en la tabla 4, como se puede observar al igual que en el caso de género el conjunto de características mejor clasificado es el conocido como *lematizadas_dominios*.

De acuerdo con los valores obtenidos para género y edad, el conjunto denominado como *lematizadas_dominios* obtuvo los valores más altos de F1, por lo que estos serán utilizados como *baseline* en los experimentos posteriores.

4.2. Características de twitter extraídas del corpus

Utilizando las diferentes características propias de twitter se tienen los siguientes resultados de clasificación. Para el caso de género, estos se muestran en la tabla 5. Como se puede observar, el conjunto de características que obtiene el mejor valor de F1 son los dominios que se extraen de las url compartidas. Un valor similar, pero por debajo, es el que se obtiene al utilizar las menciones de usuarios, los hashtags y sus conceptos extraídos a partir del grafo. El peor resultado es el que se indica al utilizar las url sin modificarlas. Es importante mencionar que ninguno de estos resultados mejora el *baseline* propuesto.

Para el caso de la clasificación por edad, se obtuvieron los siguientes datos mostrados en la tabla 6. Se puede ver que nuevamente el uso de los dominios ofrece el mejor resultado de F1; sin embargo, los conceptos representados por el grafo obtienen un resultado muy similar al del *baseline.* Sin embargo, los resultados obtenidos por los hashtags, las url completas y las menciones de usuarios se encuentran por debajo de éste.

Finalmente, en las tablas 7 y 8 se pueden revisar los valores obtenidos de la medida F1 micro para cada una de las clases para género y edad, respectivamente, y las características utilizadas.

Para género, el mejor resultado para ambos sexos se obtiene utilizando el *baseline*, analizando las características de twitter, el uso de dominios obtiene buenos resultados tanto para el género femenino y masculino, aunque las menciones de usuarios obtienen el mejor resultado para el género femenino.

Tabla 5. Resultados macro para la clasificación de género.

	Precision	**Recall**	**F1**
dominios	0.65509	0.64	0.63115
usuarios	0.63388	0.61667	0.60237
hashtags	0.61169	0.6	0.59026
grafo	0.56705	0.564	0.55648
urls	0.58497	0.52333	0.45354

Tabla 6. Resultados macro para la clasificación de edad.

	Precision	**Recall**	**F1**
dominios	0.41255	0.43333	0.41746
grafo	0.41038	0.42745	0.41143
hashtags	0.35619	0.36667	0.34743
urls	0.42057	0.42667	0.32523
usuarios	0.31029	0.32667	0.30802

Tabla 7. F1 micro para género.

Características	**Femenino**	**Masculino**
lematizadas_dominios (*baseline*)	0.809	0.787
grafo	0.562	0.551
hashtags	0.643	0.536
usuarios	0.674	0.53
urls	0.642	0.263
dominios	0.666	0.598

Tabla 8. F1 micro para edad.

Característica	**18-24**	**25-34**	**35-49**	**50-64**	**65+**
lematizadas_dominios (*baseline*)	0.0	0.438	0.572	0.162	0.0
grafo	0.18	0.4	0.524	0.256	0.0
hashtags	0.124	0.29	0.48	0.242	0.0
usuarios	0.058	0.262	0.444	0.19	0.0
urls	0.0	0.162	0.582	0.134	0.0
dominios	0.168	0.404	0.518	0.336	0.0

Para edad, se puede observar que el uso de los dominios o conceptos del grafo de conocimiento como característica ofrecen los mejores resultados en 4 de 5 clases, en comparación con el *baseline*, esta última característica ofrece buenos resultados en 3 de 5 clases, siendo la clase de 65 o más años donde ninguna característica obtiene resultados al clasificarse, pero en la 18 a 24 años el *baseline* no obtiene resultados, pero el uso de conceptos o dominios si los obtienen.

5. Conclusiones y trabajo futuro

A partir de los resultados obtenidos para la tarea de perfil de autor se puede observar que, para el caso de género, el conjunto de características que ofrece los mejores

resultados son aquellas que se basan en la utilización de todas las palabras, siendo el conjunto que se compone por los lemas de los tokens y los dominios de las url el que mejor resultado obtiene. Cabe hacer notar que no presenta una gran diferencia con los otros conjuntos de características basadas en bolsa de palabras. Por otro lado, con respecto a los resultados obtenidos utilizando las características propias de twitter no se observa ninguna mejora. Al analizar cada característica, en particular se observa que el uso de los dominios de las url mejora la clasificación en el caso del género masculino, lo que permite presuponer que los hombres comparten diferentes tipos de contenido pero que proviene de sitios similares. Mientras para el caso del género femenino el uso del dominio o de toda la url ofrece resultados similares, lo que puede interpretarse como que comparten información de sitios distintos. En el caso de las menciones de usuario, hashtags y conceptos identificados vía el grafo de conocimiento ofrecen resultados similares para el caso de los hombres, por lo que comparten contenidos similares y en el caso de las mujeres, la clasificación cae utilizando sólo los conceptos.

Para el rasgo de edad se observa que el uso de conceptos del grafo de conocimiento y el uso de dominios presenta una ligera mejora en la clasificación, de acuerdo con el conjunto de características basado en el uso de lemas y dominios, caso contrario para los hashtags, los usuarios y las url. A nivel micro se tiene que el uso de conceptos o dominios mejora la clasificación en edades de 35-49 y 50-64, no así en el resto de ellas, resultados que a su vez mejoran la clasificación final.

En general se identifica que el uso de conceptos extraídos a partir de hashtags y dominios de las url compartidas como características en la tarea de perfil de autor permite obtener resultados interesantes. Aunque es importante mencionar que esto depende de la presencia de hashtags en el corpus a analizar, por lo que como trabajo futuro es importante realizar un estudio estadístico de la cantidad de url y hashtags que contienen las clases que mejoraron.

Por otro lado, la identificación de conceptos utilizando herramientas semánticas como el grafo de conocimiento, permite conocer el significado de características temáticas, por lo que pueden ser útiles en otras tareas de minería de textos. Un trabajo a futuro es comparar los resultados que se obtienen con ellas con los obtenidas utilizando algoritmos de semántica latente.

Otra vertiente a explorar es la combinación con técnicas de análisis de sentimientos. Ya que una vez que se obtiene el tema asociado a un hashtag también es posible identificar la aprobación o rechazo a ese tema.

Información que enriquecería la caracterización del mensaje y ayudaría a una mejor identificación de los rasgos del perfil del usuario.

Agradecimientos. Este trabajo fue desarrollado con el apoyo parcial del CONACYT bajo el programa de posdoctorados nacionales.

Referencias

1. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the reproducibility of PAN's shared tasks: Plagiarism detection, author identification, and author

profiling. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction, pp. 268–299 (2014)
2. Tellez, E.S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O.S., Villaseñor, E.A.: A case study of spanish text transformations for twitter sentiment analysis. In: Expert Systems with Applications, pp. 457–471. DOI: 10.1016/j.eswa.2017.03.071. 81 (2017)
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. In: Journal of the American society for information science, American Documentation Institute, 41(391) (1990)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. In: The Journal of Machine Learning Research, 3, pp. 993–1022 (2003)
5. 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, 2, pp. 42–47 (2011)
6. Sultana, M., Paul, P.P., Gavrilova, M.: Mining social behavioral biometrics in twitter. In: Proceeding (CW'14), Proceedings of the International Conference on Cyberworlds, IEEE Computer Society Washington (2014)
7. Statista Homepage: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of users/ (2014)
8. Google: https://googleblog.blogspot.mx/2012/05/introducing-knowledge-graph-things-not.html (2012)
9. Dong, X.L., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun S., Zhang, W.: Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In: The 20th (ACM SIGKDD) International Conference on Knowledge Discovery and Data Mining, (KDD'14), pp. 601–610 (2014)
10. Google Knowledge Graph API: https://developers.google.com/knowledge-graph/ (2018)
11. PAN 2014: http://pan.webis.de/clef14/pan14-web/author-profiling.html (2014)
12. Rangel, F., Chugur, P.R.I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at pan 2014. In: Proceedings of the Conference and Labs of the Evaluation Forum (2014)
13. Salton, G., McGill, M.: Introduction to Modern Information Retrieval, McGraw Hill (1983)
14. Salton, G.: Automatic Text Processing: The Transform, Analysis, and Retrieval of Information by Computer. Addison-Wesley Longman Publishing Co. Inc. (1989)
15. Sidorov, G.: Construcción no lineal de n-gramas en la lingüística computacional: n-gramas sintácticos, filtrados y generalizados. 166 p. (2013)
16. Lan, M., Tan, C.L., Low, H.B., Sung, S.: A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In: Special interest tracks and posters of the 14th international conference on World Wide Web (WWW'05). (ACM), pp. 1032–1033 (2005)
17. Cilibrasi, R.L., Vitányi, M.B.P.: The Google Similarity Distance. In: IEEE Transactions on knowledge and data engineering, 19(3) (2007)
18. Sebastiani, F.: Text Categorization. In: Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press (2005)
19. Ferragina, P., Piccinno, F., Santoro, R.: On Analyzing Hashtags in Twitter. In: International (AAAI) Conference on Web and Social Media (2015)
20. Manning, C.D., et. al.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)

21. Van Rossum, G.: Python tutorial. In: Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI) (1995)
22. Pedregosa et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pp. 2825–2830 (2011)
23. Villegas, M.P., Garciarena-Ucelay, M.J., Fernández, J.P., Álvarez-Carmona, M.A., Errecalde, M.L., Cagnina, L.: Vector-based word representations for sentiment analysis: a comparative study. In: XXII Congreso Argentino de Ciencias de la Computación, (CACIC), pp. 785–793 (2016)
24. Kocher, M., Jacques, S.: Distance measures in author profiling. In: Information Processing & Management, 53(5), pp. 1103–1119 (2017)
25. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y-Gómez, M., Villaseñor-Pineda, L., Jair-Escalante, H.: INAOE's Participation at PAN'15: Author Profiling task. In: Working Notes Papers of the CLEF (2015)