ISSN 2007-9737

Identificación del perfil de autores en redes sociales usando nuevos esquemas de pesado que enfatizan información de tipo personal

Rosa María Ortega Mendoza[1,2], Anilú Franco Arcega[1], Manuel Montes y Gómez[2]

[1] Universidad Autónoma del Estado de Hidalgo,
Mineral de la Reforma, Hidalgo,
México

[2] Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla,
México

rmortega@inaoep.mx, afranco@uaeh.edu.mx, mmontesg@inaoep.mx

Resumen. Este artículo resume la tesis Ïdentificación del perfil de autores en redes sociales usando nuevos esquemas de pesado que enfatizan información de tipo personalçuya idea principal indica que los términos localizados en frases que exponen información personal son altamente valiosos para la tarea conocida como identificación del perfil de autores. Primero, se presenta un estudio sobre la relevancia de este tipo de frases en la tarea. Posteriormente, se propone un enfoque que enfatiza el valor de este tipo de información mediante dos novedosas propuestas: un método de selección de características y un esquema de pesado de términos, ambos basados en una nueva medida llamada ”índice de expresión personal”, la cual mide la cantidad de información personal revelada por un término. El enfoque fue evaluado en diferentes redes sociales prediciendo la edad y el género de usuarios. Los resultados señalan mejoras promedio con respecto a los mejores resultados del estado del arte: 7.34 % y 5.76 % para la identificación de edad y género respectivamente. Por lo tanto, se concluye que la información personal juega un rol importante en la tarea.

Palabras clave. Identificación del perfil de autores, información personal, esquemas de pesado, PEI, DPP, EXPEI.

Author Profiling on Social Media using New Weighting Schemes that Emphasize Personal Information

Abstract. This paper summarizes the thesis: Ïdentificación del perfil de autores en redes sociales usando nuevos esquemas de pesado que enfatizan información de tipo personal”whose main idea indicates that terms located in phrases exposing personal information are highly valuable for the AP task. Firstly, it is presented an study on the relevance of this information to this task. Secondly, it is proposed a novel approach, which aims to emphasize the value of this type of terms by two proposals: a feature selection method and a term weighting scheme; both of them are based on a novel measure called *personal expression intensity*, which estimates the quantity of personal information revealed by each term. The approach was evaluated in age and gender prediction on different social media. The results are encouraging, with average improvements about 7.34 % and 5.76 % for age and gender identification respectively in comparison with the best results from the state of the art. These results allow concluding that personal information play an important role in the task.

Keywords. Author profiling, term weighting schemes, personal information, PEI, DPP, EXPEI.

1. Introducción

La caracterización del perfil de los autores (en inglés *Author Profiling*, AP) es una tarea dirigida a estudiar el uso del lenguaje para distinguir grupos de autores que comparten una característica demográfica común (e.g. edad y género).

Existe una diversidad de rasgos socio-demográficos que conforman el perfil de los autores y que la literatura ha explorado mediante

enfoques de AP. Por ejemplo, se han propuesto métodos para detectar personalidad [11, 20], orientación política [15], idioma nativo [23], ocupación [8] y edad/género [2, 19]. Actualmente, esta tarea ha ganado gran relevancia para la comunidad científica debido a sus aplicaciones en diversas disciplinas. Por ejemplo, en lingüística forense puede generar evidencia adicional para identificar características de autores de mensajes de acoso. En Mercadotecnia, puede ayudar a generar publicidad dirigida de acuerdo con el perfil de las personas que gustan o disgustan de un producto. Inclusive, el impacto de la tarea en redes sociales ha motivado la creación de foros internacionales de evaluación para métodos referentes a AP [18].

Tradicionalmente, AP ha sido una tarea abordada como un problema de clasificación supervisado textos [21]. La mayoría de las contribuciones se concentran en la búsqueda de un conjunto de atributos que modele el perfil lingüístico de los autores. Dos tipos de características han sobresalido: temáticas (e.g. sustantivos, verbos y adjetivos) y características de estilo (e.g., palabras de función, signos de puntuación y etiquetas de partes de la oración). Combinaciones de ambos tipos han resultado exitosas en evaluaciones del PAN [18]. Recientemente, enfoques más sofisticados han sido considerados: n-gramas de caracteres, palabras o sintácticos [4, 17], representaciones basadas en tópicos [2], atributos de segundo orden [10] y representaciones continuas de palabras (*word embeddings representations*) [1, 5], así como técnicas de aprendizaje profundo [22, 24].

En contraste a la búsqueda común de atributos, esta investigación se enriquece de hallazgos psicológicos para seleccionar y ponderar la información más relevante para AP en redes sociales. De acuerdo con la perspectiva psicológica, el uso de pronombres personales en primera persona del singular (PP) se asocia con características que definen el perfil de las personas [16]. Por lo tanto, esta investigación sustenta que las frases que contienen PP[1], denominadas frases personales (FP), reflejan información personal como sentimientos, preferencias, intereses y hábitos entre otras características que pueden revelar el perfil del autor.

Específicamente, en la primer parte de la investigación se estudió la relevancia de las FP en AP. El estudio indicó que las FP conforman la esencia de los documentos para la tarea. En la segunda parte, se desarrolló un enfoque para AP que se compone de un nuevo método de selección y un novedoso esquema de pesado de términos, los cuales, además de cuantificar la frecuencia de los términos, consideran una calificación para la cantidad de información personal revelada por cada término. El enfoque fue evaluado para clasificar edad y género de autores; se obtuvieron mejoras promedio de exactitud de 7.34 % y 5.76 % respectivamente con respecto a resultados del estado del arte. Los resultados señalan que es posible mejorar el desempeño cuando se enfatiza el valor de los términos asociados a las FP. De esta manera se confirma la relevancia de las FP en la tarea.

El resto del manuscrito resume la investigación. Más detalles se presentan en la tesis [12] y las publicaciones representativas derivadas [14, 13].

2. El rol de las frases personales en la identificación del perfil de autores

La idea base de la investigación considera que las personas expresan sus intereses y estilo de escritura cuando hablan acerca de ellas mismas (mediante las FP). Por ejemplo, en la Tabla **??** se muestran algunas FP describiendo actividades que las personas comúnmente realizan al despertar en la mañana. Como se observa, cada persona usa su propio estilo de escritura y su vez, refleja sus intereses temáticos. Sin embargo, es posible encontrar patrones discriminativos de perfiles (e.g., hombres vs mujeres, o jóvenes vs adultos).

Por ejemplo, los hombres hablan más de comida que las mujeres, mientras las mujeres se expresan más acerca del cuidado personal, específicamente de su cabello. También puede ser notado que los

[1] Específicamente: *I, me, mine, my, myself* y la cadena *im* (común en textos de redes sociales).

Tabla 1. Ejemplos de frases personales extraídas de blogs provenientes de la colección de Schler [19] H=Hombre; M=Mujer

Fragmento de texto	Género	Edad
"And then I woke up at 11:00 & took a *shower* & got *dressed*. Then I was gonna fix my *hair* & put on my *makeup* & *mom* said there was no use in goin because it was late anyway.. So "	M	15
"I woke up Sunday morning and *cleaned* up the *house*. I have decided not to run away, just yet. Once the *house* was *cleaned* I took a long *bath* and *washed* my *hair* and gave it an intensive *conditioning treatment*."	M	41
"Wow what a day! I woke up about 11:30 to a great *breakfast* of *tacos!! Beef, egg, cheese* and *salsa sauce* to be precise, *yummmm!*"	H	15
"I woke up this morning feeling great. I went to the *kitchen*, *fried* me a *hamburger patty*, and some *eggs*. "	H	44

jóvenes tienden a mencionar a sus padres o a escribir informalmente (“..”, “&” o “!!”).

La destacable riqueza de la información observada en las FP motivó el desarrollo de un análisis más profundo que responde, principalmente, a las siguientes preguntas de investigación: a) ¿Es toda la información en un documento igualmente relevante para AP? Específicamente, ¿son las frases personales más discriminativas que las otras frases? y b)¿Son las frases personales igualmente relevantes en diferentes medios sociales?. El análisis se realizó de acuerdo con la siguiente configuración experimental.

2.1. Configuración experimental

La investigación adoptó un marco estándar de clasificación para AP: una combinación de características de contenido, estilo y sintácticas. Particularmente, se usaron los 1000 términos con mayor ganancia de información. Estos términos incluyeron palabras de contenido, signos de puntuación y palabras coloquiales (*slang words*). También se consideraron ocurrencias de palabras de función (FW) y unigramas de etiquetas POS. Usando tales términos, se construyó una bolsa de palabras cuyos pesos corresponden a su frecuencia normalizada (TF). En la fase de clasificación se aplicó SVM mediante un esquema de validación estratificada de 10 capas reportando exactitud.

Conjuntos de datos. Corresponden a textos en escritos en inglés provenientes de redes sociales y etiquetados automáticamente con la edad y género de los autores. Específicamente, se usaron dos conjuntos: blogs de Schler [19] y el corpus PAN-AP-2014[2]. En particular, los blogs de Schler[3] fueron concentrados en tres categorías según su edad [19]: 10s (desde 13 hasta 17 años), 20s (desde 23 hasta 27 años) y 30s (desde 33 hasta 47 años). Por su parte, el corpus PAN-AP-2014[4] se compone de cuatro dominios: blogs, revisiones de hoteles de TripAdvisor (de aquí en adelante, denotadas como Reviews), documentos de redes sociales (generalizados como Social Media) y publicaciones de Twitter (denotadas como Twitter). En este corpus, la etiqueta edad tiene 5 categorías: 18-24, 25-34, 35-49, 50-64 y $\geq$65.

2.2. Relevancia de las FP para AP

Para responder a las preguntas de investigación, se evaluó el rol de las FP en AP. Para ello, primero, se filtraron las FP de cada documento en la colección de textos, creando un corpus filtrado. Posteriormente, se comparó el desempeño de clasificación del corpus original (que contiene todas las frases de los documentos) y el corpus filtrado. Los resultados se muestran en la Tabla 2 y confirman valores de exactitud similares para el corpus filtrado y el original; sin embargo, es notable que las colecciones filtradas representan un pequeño subconjunto (de 15 % a 48 %) del corpus original.

[2] Este es el corpus de entrenamiento para la tarea AP en la competencia del PAN 2014

[3] Este corpus contiene 19320 documentos

[4] Contiene 147, 306, 4160 y 7746 documentos en blogs, twitter, reviews y social media, respectivamente

ISSN 2007-9737

Tabla 2. Resultados usando sólo FP en las colecciones de Schler y PAN 2014

Colección	Corpus	Exactitud		% en el corpus filtrado
		Edad	Género	
Schler	Original	77.49	80.07	48.12 % (de
	Filtrado	76.09	79.63	9,155,301)
Blogs	Original	36.56	68.42	24.20 % (de
	Filtrado	43.92	62.14	22,944 frases)
Twitter	Original	35.33	71.33	15.54 % (de
	Filtrado	37.49	59.55	318,691 frases
Reviews	Original	30.84	67.24	36.43 % (de
	Filtrado	29.21	65.21	52,833 frases)
Social Media	Original	34.84	53.64	22.97 % (de
	Filtrado	33.99	52.68	3,207,509)

Las pruebas de significancia estadística[5] indicaron que los resultados para la predicción de edad fueron comparables a través de todos los dominios considerados excepto para Blogs de Schler, mientras para la clasificación de género, se encontró una diferencia estadísticamente significante sólo para Twitter y Blogs. En general, estos resultados soportan la relevancia de las frases personales reafirmando su rol como la esencia de los documentos para AP.

Adicionalmente, se encontró que los resultados usando el corpus filtrado son significativamente mejores que esos correspondientes al corpus complemento, aunque hay menos información en el primero. Por ejemplo, en el caso de Blogs de Schler, el corpus complemento presentó una exactitud de 69.98 % para edad y 72.59 % para género respectivamente. Esto indica que la información personal de los autores es, en efecto, más importante que la información no personal.

Cabe señalar que, el análisis del rol de las frases personales en AP también incluyó el estudio de frases conteniendo pronombres personales en primera persona del plural, encontrando que éstas no tienen una relevancia especial para AP. Por otro lado, también se estudió la naturaleza del tipo de información contenida en las FP, concluyendo que la informacion del estilo de redaccion de los autores podría ser igualmente capturada tanto en FP como en FNP, pues ambos tipos de frases son escritas por el mismo autor. Sin embargo, los intereses temáticos de los autores son mejor capturados en las FP. Más detalles se muestran en [13].

[5] Para evaluar la diferencia estadística se utilizó la prueba pareada t aplicada sobre las 10 capas con un nivel de significancia de 0.05.

3. Enfatizando el valor de las frases personales

Los hallazgos anteriores motivaron el desarrollo de un nuevo enfoque de AP, el cual considera todos los términos de los documentes (presentes en FP o FNP) pero enfatiza el valor de aquellos contenidos en FP. Este enfoque inicia con la cuantificación de la cantidad de información personal del autor revelada por cada término. Para ello, se diseñaron tres nuevas medidas: precisión personal, cobertura personal y el índice de expresión personal.

Estas medidas son definidas considerando la siguiente notación: un documento d_j está formado por un conjunto de frases S_j, el cual a su vez está compuesto por los subconjuntos P_j y N_j, que representan los subconjuntos de frases personales y no personales, respectivamente. Por lo tanto, un término t_i puede aparecer en el subconjunto P_j y/o en N_j.

Precisión personal (ρ) estima la concentración de información personal revelada en el contexto de un término. Es definida como el porcentaje de frases personales conteniendo el término t_i dentro del documento d_j:

$$\rho(t_i, d_j) = \frac{\#(t_i, P_j)}{\#(t_i, S_j)}. \quad (1)$$

Cobertura personal (τ) cuantifica la porción de frases personales de un documento (i.e., la porción de su "esencia") cubierta por el término t_i. Puede ser interpretada como la probabilidad condicional de la ocurrencia de un término dado el conjunto de frases personales:

$$\tau(t_i, d_j) = \frac{\#(t_i, P_j)}{|P_j|}. \quad (2)$$

Aunque ρ y τ son medidas cuyo valor incrementa cuando el número de ocurrencias en las frases personales es más grande, su comportamiento, es, de algún modo, opuesto. Por ejemplo, un término apareciendo una sola vez en un documento y particularmente en una frase personal, obtendría un valor muy alto de precisión personal (ρ), pero no necesariamente alta cobertura (τ), principalmente, porque el documento puede estar formado por varias frases personales. Por el contrario, un término apareciendo en la única frase personal de un documento conseguiría el más alto valor para ρ, independientemente de sus ocurrencias en frases no personales. De ahí que, para medir el balance entre ρ y τ se propone la siguiente medida.

Índice de expresión personal (PEI por sus siglas en inglés, *Personal Expressión Index*) es una combinación de ρ y τ e indica que entre más frecuente es la ocurrencia de un término en frases personales y menos frecuente en las frases no personales, el término revela más información del perfil del autor:

$$PEI(t_i, d_j) = 2\frac{\rho(t_i, d_j)\,\tau(t_i, d_j)}{\rho(t_i, d_j) + \tau(t_i, d_j)}. \qquad (3)$$

PEI establece que los términos más valiosos son aquellos con alta precisión personal así como alta cobertura.

3.1. El enfoque propuesto DPP-EXPEI

Desde la perspectiva de la clasificación supervisada, la construcción de un clasificador que asigne categorías predifinidas (categorías de autores) $C = \{c_1, .., c_{|C|}\}$ a una colección de documentos $D = \{d_1, ..., d_{|D|}\}$ involucra la transformación de documentos en una representación adecuada para los algoritmos de clasificación supervisada. Este proceso involucra dos etapas principales: selección y pesado de términos. El enfoque propuesto se enfoca en estas dos etapas. Específicamente, se considera la medida PEI para crear una nueva técnica de selección llamada pureza personal discriminativa (*discriminative personal purity*, DPP, por sus siglas en inglés) y un nuevo esquema de pesado, llamado recompensa exponencial de información personal (*exponential rewarding of personal information*, EXPEI).

3.1.1. Selección de términos: pureza personal discriminativa

La técnica de selección propuesta, DPP, permite elegir términos relacionados al perfil de los usuarios por medio de la medida PEI, tal como se muestra en la fórmula 4. Básicamente, DPP tiene dos componentes: un factor descriptivo definido como el máximo valor de la función PP_k (Eq. 5), que captura la capacidad de un término para describir información personal de autores pertenecientes a la categoría (c_k); y un factor discriminativo, basado en el coeficiente *gini* [7], que califica la habilidad de un término para discriminar ente categorías de autores (perfiles). Enseguida se describen ambos componentes:

$$DPP(t_i) = \max_{k=1}^{|C|} \{PP_k(t_i)\} \cdot gini(t_i). \qquad (4)$$

Pureza personal categórica como factor descriptivo.

La pureza personal categórica de un término t_i en un categoría c_k, definida como $PP_k(t_i)$, evalúa la información personal capturada por el término en los documentos pertenecientes a esa categoría. Formalmente, PP_k está representada por la ecuación 5:

$$PP_k(t_i) = \log_2\left(2 + \frac{1}{2}\sum_{d_j \in c_k} \frac{PEI(t_i, d_j) + 1}{NEI(t_i, d_j) + 1}\right), \qquad (5)$$

donde NEI[6], un concepto opuesto a PEI, captura el nivel de asociación de cada término a la información no personal.

Por lo tanto, PP_k es calculada como el cociente acumulativo de PEI entre NEI de todos los términos pertenecientes a los documentos de la categoría c_k. De esta manera, un término con valores de PEI mayores que NEI será premiado.

[6]Non-personal expression intensity (NEI) considera las ocurrencias de los términos en el subconjunto de las frases no personales.

Gini coeficiente como factor descriptivo. El factor discriminativo denotado como $gini(t_i)$ estima la capacidad de un término para discriminar documentos de las diferentes categorías de autores. Este segundo factor es determinado a través del coeficiente Gini, una medida que captura, en un solo valor, el nivel de concentración o desigualdad de cualquier distribución; en este caso, la distribución de los términos en todas las categorías. Por ejemplo, la presencia concentrada de un término en sólo una de las categorías señala su pertinencia para lograr la discriminación. Por el contrario, las ocurrencias de un término igualmente distribuidas en todas las categorías indica un bajo nivel de discriminación. Para estimar el coeficiente Gini, se aplicó la fórmula mostrada en [7], cuyo rango de valores va desde 0 hasta 1, indicando completa igualdad o desigualdad respectivamente.

3.1.2. Pesado de términos: recompensa exponencial de información personal

El esquema de pesado propuesto EXPEI, el cual se representa en la fórmula 6 considera todos los términos de los documentos, es decir, aquellos provenientes de FP así como de FNP, pero enfatiza el valor de la información personal.

$$w_{ij} = \left(\sqrt{TF(t_i, d_j)}\right)^{1-PEI(t_i,d_j)}, \qquad (6)$$

donde $TF(t_i, d_j)$ representa la frecuencia normalizada del término t_i en el documento d_j calculada como $\frac{\#(t_i,d_j)}{len(d_j)}$.

El esquema está basado en los valores de TF asignados a los términos. Por ejemplo, los términos con PEI = 1, obtendrán pesos iguales a 1 (EXPEI = 1, su máximo valor posible), independientemente de su frecuencia. Por otro lado, los términos con $0 <$ PEI < 1 serán proporcionalmente premiados; este premio es más importante para aquellos términos con baja frecuencia. Finalmente, los pesos de los términos con PEI = 0 serán suavizados por EXPEI (haciéndolos un poco más grandes que sus valores de TF). De esta manera se permite que los términos con baja frecuencia, pero relacionados con información personal, tengan la oportunidad de contribuir en la descripción del documento.

El enfoque DPP-EXPEI fue evaluado analizando su desempeño para predecir la edad y género de los autores bajo un marco experimental similar al presentado en la sección 2.1; naturalmente, se usó DPP como técnica de selección y EXPEI como esquema de pesado de términos. Los resultados fueron comparados con enfoques del estado del arte presentados en: [2], donde se explotan representaciones basadas en tópicos usando análisis semántico latente **(LSA)** y conteos de palabras **(LIWC)**; [10], donde se describe una representación basada subperfiles de usuario (**SSR**), es el método de referencia principal porque ha obtenido los mejores resultados en las colecciones del PAN 2013-2016; [9], donde se exploran atributos de segundo orden **(SOA)**; [25], donde se presenta un método basado en ideas de recuperación de información; [6], donde se usan atributos a nivel de grupo **(GLA)** mediante un análisis de tópicos; [3], donde se presenta un análisis de más de 140 millones de palabras en inglés **(MW)** obtenidas de blogs; [19], donde un conjunto de características estilísticas así como de contenido **(SC)** es usado para encontrar diferencias de género y edad.

Los resultados y comparaciones se muestran en la Tabla 3. Se observa que DPP-EXPEI supera a los enfoques de referencia en cada colección del PAN 2014 (únicamente en Social Media, SSR reportó una mejor exactitud). Además, el enfoque obtuvo mejores resultados que SSR en tres colecciones en el caso edad. Mientras, en el caso género, se mejoran los resultados en las cuatro colecciones. También existen ganancias importantes, por ejemplo, en el corpus Blogs para el caso edad existe una diferencia aproximada de 22 %; mientras las pérdidas obtenidas fueron muy pequeñas, la desventaja más representativa corresponde a 4.15 % en el caso edad de la colección Social Media.

Por otro lado, la prueba del rango con signo de Wilcoxon usando un nivel 0.05 de significancia indica que DPP-EXPEI es significativamente mejor que SSR (se comparó estadísticamente SSR y DPP-EXPEI sobre los diez conjuntos de datos).

Tabla 3. Resultados usando DPP para seleccionar términos y EXPEI para pesarlos

	Enfoque	Colecciones PAN 2014				Schler corpus
		Reviews	Twitter	Blogs	Social Media	
Edad	DPP-EXPEI	**44.83**	**61.44**	**75.34**	33.91	75.9
	LSA	34	39	48	36	-
	LIWC	29	47	42	34	-
	SSR	36.9	49.01	53.06	**38.06**	**77.68**
	SOA	33.92	47.97	48.07	37	-
	IRF	37.62	52.61	45.58	42.51	-
	GLA	-	-	-	-	72.83
	MW	-	-	-	-	77.4
	SC	-	-	-	-	76.01
Género	DPP-EXPEI	**76.42**	**81.5**	**84.25**	**58.57**	79.43
	LSA	65	66	70	52	-
	LIWC	62	71	60	50	-
	SSR	69.27	71.69	80.95	55.39	**82.01**
	SOA	68.05	71.92	77.96	55.36	-
	IRF	71.03	78.76	82.99	57.04	-
	GLA	-	-	-	-	75.04
	MW	-	-	-	-	80.5
	SC	-	-	-	-	80.01

Los resultados permiten concluir que DPP-EXPEI enfoca la atención en los términos más relevantes (intereses personales) cuando se tiene menos información (i.e., colecciones pequeñas). Sin embargo, cuando hay más información textual disponible (colecciones más grandes) DPP-EXPEI tiene menos impacto, principalmente, porque los términos frecuentes tienden a exponer directamente tales intereses personales.

Un análisis detallado sobre el desempeño de cada uno de los componentes del enfoque DPP-EXPEI es mostrado en [14].

3.2. Discusión

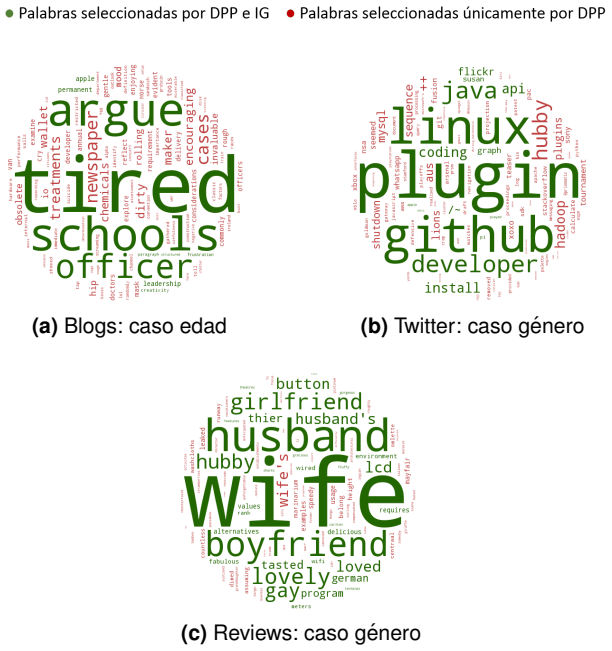
A pesar de la heterogenidad de los esquemas de selección y pesado de términos, la mayoría están soportados en inferencias estadísticas sobre las ocurrencias de los términos en los documentos sin considerar características cualitativas de esas ocurrencias. En contraste, las técnicas DPP y EXPEI están basadas en la idea de que no toda la información en un documento es igualmente relevante; por lo tanto, analizan el contexto de procedencia de cada término tomando ventaja de su ocurrencia en las FP. Por ejemplo, en los experimentos, la mayoría de los términos más relevantes seleccionados por DPP fueron también elegidos por IG, como se muestra en la Figura 1.

Si bien, varios términos seleccionados únicamente por DPP no son frecuentes, ellos parecen ser términos intuitivos conocidos en la literatura de AP. En específico, para identificar adultos hay términos valiosos como: *newspaper*, *doctors* y *treatments*. A su vez, para la identificación de género las palabras *mysql, haddop* o *plugins* son muy cercanas a tópicos de tecnología, los cuales han sido asociados con el género masculino. Finalmente, palabras como *xoxo, aws, hubby* han mostrado ser de gran ayuda para la identificación de mujeres. De esta manera, DPP enriquece la selección incluyendo varios términos relacionados con expresiones personales que IG califica como no informativos.

Particularmente, también se estudiaron correlaciones entre el esquema EXPEI y el tradicional TF, como se muestra en la Figura 2. Se encontró que ambos esquemas están menos correlacionados en los documentos con pocas publicaciones, esto indica que EXPEI extrae información relevante para AP incluso cuando es poco frecuente, lo cual se traduce en una ventaja importante cuando hay poca información. Por otro lado, cuando existen más publicaciones, las correlaciones tienden a incrementar, sugiriendo que en documentos largos la frecuencia es suficiente para detectar información discriminativa de perfiles de autores.

Finalmente, también se estudió la influencia de las características de las colecciones con el desempeño del enfoque tomando como referencia el método SSR. Mediante correlaciones extraídas usando el coeficiente de Spearman se encontró que el enfoque es apropiado para AP cuando existan pocos ejemplos con alta densidad léxica y colecciones de entrenamiento desbalanceadas. Todas ellas son condiciones desafiantes en AP. Por el contrario, se recomienda evitar el uso del enfoque propuesto cuando exista un gran número de ejemplos; en este caso, se sugiere el uso de otros métodos como SSR o tradicionales.

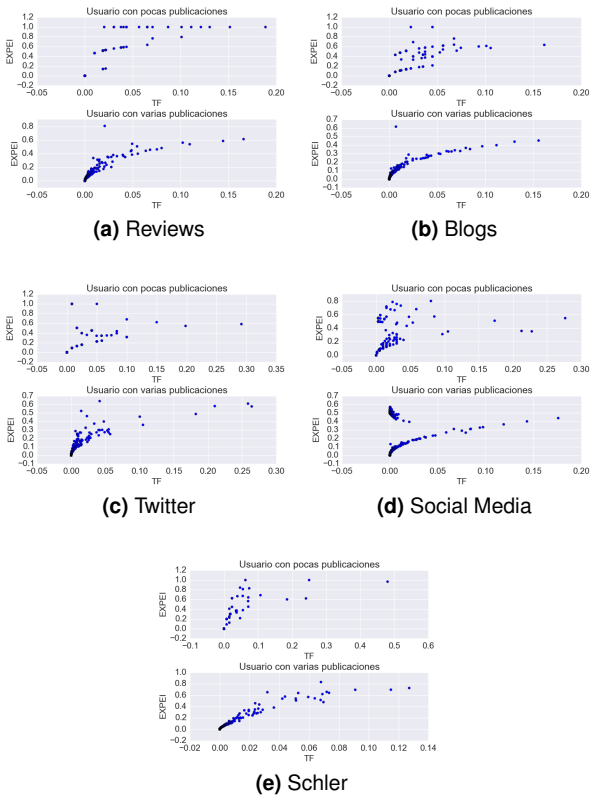
ISSN 2007-9737



(a) Blogs: caso edad **(b)** Twitter: caso género

(c) Reviews: caso género

Fig. 1. Las 100 palabras mejor calificadas por DPP. El tamaño de la fuente corresponde a la posición del término dentro de las 100 mejores. El color verde y orientación horizontal representan las palabras seleccionadas por ambos esquemas: IG y DPP; mientras el color rojo y orientación vertical identifica las palabras seleccionadas únicamente por DPP



(a) Reviews **(b)** Blogs

(c) Twitter **(d)** Social Media

(e) Schler

Fig. 2. Correlaciones de los esquemas EXPEI y TF entre las representaciones vectoriales de los documentos correspondientes a usuarios con pocas y varias publicaciones respectivamente. Ambas representaciones son menos correlacionadas cuando hay menos publicaciones y más correlacionadas en caso contrario

4. Conclusiones y trabajo futuro

En este trabajo de investigación se aborda la tarea AP en redes sociales. La hipótesis principal de esta investigación indica que las frases personales integran la esencia de los textos para la tarea AP. Principalmente, porque en este tipo de frases se exponen términos que revelan rasgos del perfil de las personas. Para comprobar la hipótesis se comparó el desempeño de un método del estado del arte [19] usando los documentos completos versus usando el subconjunto de FP. Los resultados fueron sorprendentes, se obtuvo una exactitud similar usando cualquiera de los dos conjuntos, aunque el subconjunto de frases personales representa sólo una pequeña porción de la colección original. Los resultados corroboran la hipótesis y sugieren que las FP conforman *.esencia"* de los documentos para AP.

El hallazgo mencionado inspiró el diseño de dos nuevos esquemas: DPP para selección de términos y EXPEI para pesado de términos; ambos basados en una medida llamada índice de expresión personal (PEI), la cual fue propuesta para cuantificar el grado de asociación de los términos a la información personal del autor. La combinación de tales esquemas (DPP-EXPEI) superó los resultados reportados en el estado del arte en la mayoría de las colecciones, mostrando mejoras promedio de 7.34 % y 5.76 % para edad y género respectivamente.

Los resultados de esta investigación han

motivado el interés de evaluar el enfoque propuesto: usando otros idiomas (e.g. español), prediciendo otras dimensiones del perfil de autores (e.g. personalidad), así como su aplicación en la caracterización de tipos de comportamientos sociales de los usuarios (e.g. identificación de acoso y detección de usuarios con depresión).

Agradecimientos

Esta investigación fue realizada con el apoyo de CONACyT bajo las subvenciones de los proyectos PDCPN2014-01-247870 y FC-2016-2410. El trabajo también fue apoyado por Red Temática en Tecnologías del Lenguaje (proyectos 260178 y 271622).

Referencias

1. **Alekseev, A. & Nikolenko, S. (2017).** Word Embeddings for User Profiling in Online Social Networks. *Computación y Sistemas*, Vol. 21, No. 2, pp. 203–226.
2. **Álvarez-Carmona, M. A., López-Monroy, A. P., Montes-y-Gómez, M., Villaseñor-Pineda, L., & Meza, I. (2016).** Evaluating topic-based representations for author profiling in social media. In **Montes y Gómez, M., Escalante, H. J., Segura, A., & Murillo, J.**, editors, *Advances in Artificial Intelligence - IBERAMIA 2016: 15th Ibero-American Conference on AI, San José, Costa Rica, November 23-25*. Springer International Publishing, Cham, pp. 151–162.
3. **Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007).** Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, Vol. 12, No. 9.
4. **Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., & Nissim, M. (2017).** N-gram: New groningen author-profiling model. *CoRR*, Vol. abs/1707.03764.
5. **Bayot, R. K. & Gonçalves, T. (2016).** Author Profiling Using SVMs and Word Embedding Averages—Notebook for PAN at CLEF 2016. **Balog, K., Cappellato, L., Ferro, N., & Macdonald, C.**, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*, CEUR-WS.org.
6. **Booker, L. B. (2008).** Finding identity group "fingerprints" in documents. In **Srihari, S. N. & Franke, K.**, editors, *Computational Forensics: Second International Workshop, IWCF 2008, Washington, DC, USA, August 7-8, Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 113–121.
7. **Dixon, P., Weiner, J., Mitchell-Olds, T., & Woodley, R. (1988).** Erratum to bootstrapping the gini coefficient of inequality. *Ecology*, Vol. 68, pp. 1307.
8. **Duong, D. T., Pham, S. B., & Tan, H. (2016).** *Using Content-Based Features for Author Profiling of Vietnamese Forum Posts*. Springer International Publishing, Cham, pp. 287–296.
9. **López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., & Villaseñor-Pineda, L. (2014).** Using intra-profile information for author profiling. **Cappellato, L., Ferro, N., Halvey, M., & Kraaij, W.**, editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*, volume 1180 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 1116–1120.
10. **López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., & Stamatatos, E. (2015).** Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems*, Vol. 89, pp. 134–147.
11. **Olivares, N., Vivanco, L. M., & Figueroa, A. (2018).** The Big Five: Discovering Linguistic Characteristics that Typify Distinct Personality Traits across Yahoo! Answers Members. *Computación y Sistemas*, Vol. 22, No. 3, pp. 795–807.
12. **Ortega-Mendoza, R. M. (2017).** *Identificación del perfil de autores en redes sociales usando nuevos esquemas de pesado que enfatizan información de tipo personal*. Ph.D. thesis, Universidad Autónoma del Estado de Hidalgo.
13. **Ortega-Mendoza, R. M., Franco-Arcega, A., López-Monroy, A. P., & Montes-y-Gómez, M. (2016).** I, me, mine: The role of personal phrases in author profiling. In **Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., & Ferro, N.**, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*. Springer International Publishing, Cham, pp. 110–122.

14. **Ortega-Mendoza, R. M., López-Monroy, A. P., Franco-Arcega, A., & Montes-y-Gómez, M. (2018).** Emphasizing personal information for author profiling: New approaches for term selection and weighting. *Knowledge-Based Systems*, Vol. 145, pp. 169–181.
15. **Pennacchiotti, M. & Popescu, A.-M. (2011).** Democrats, republicans and starbucks afficionados: User classification in twitter. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, ACM, New York, NY, USA, pp. 430–438.
16. **Pennebaker, J. (2011).** *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury USA.
17. **Posadas-Durán, J. P., Gómez-Adorno, H., Markov, I., Sidorov, G., Batyrshin, I. Z., Gelbukh, A. F., & Pichardo-Lagunas, O. (2015).** Syntactic n-grams as features for the author profiling task: Notebook for PAN at CLEF 2015. **Cappellato, L., Ferro, N., Jones, G., & San Juan, E.**, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*.
18. **Rangel, F. and Rosso, P. and Potthast, M. and Stein, B. (2017).** Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. **Cappellato, L., Ferro, N., Goeuriot, L., & Mandl, T.**, editors, *Working Notes Papers of the CLEF 2017 Evaluation Labs*, volume 1866 of *CEUR Workshop Proceedings*, CLEF and CEUR-WS.org.
19. **Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006).** Effects of age and gender on blogging. *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp. 199–205.
20. **Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., & Ungar, L. H. (2013).** Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, Vol. 8, No. 9, pp. e73791.
21. **Sebastiani, F. (2002).** Machine learning in automated text categorization. *ACM Computing Surveys*, Vol. 34, No. 1, pp. 1–47.
22. **Sierra, S., Montes-Y-Gómez, M., Solorio, T., & González, F. (2017).** Convolutional Neural Networks for Author Profiling in PAN 2017—Notebook for PAN at CLEF 2017. **Cappellato, L., Ferro, N., Goeuriot, L., & Mandl, T.**, editors, *CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland*, CEUR-WS.org.
23. **Tofighi, P., Köse, C., & Rouka, L. (2012).** Author's native language identification from web-based texts. *International Journal of Computer and Communication Engineering*, Vol. 1, No. 1, pp. 47–50.
24. **Tutubalina, E. & Nikolenko, S. (2017).** Demographic Prediction Based on User Reviews about Medications. *Computación y Sistemas*, Vol. 21, No. 2, pp. 227–241.
25. **Weren, E. R. D., Moreira, V. P., & P. M. de Oliveira, J. (2014).** Exploring Information Retrieval features for Author Profiling—Notebook for PAN at CLEF 2014. **Cappellato, L., Ferro, N., Halvey, M., & Kraaij, W.**, editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*, volume 1180 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 1164–1171.

Article received on 04/09/2018; accepted on 20/12/2018.
Corresponding author is Manuel Montes y Gómez.