ISSN 1870-4069

Identificando signos de anorexia y depresión en usuarios de redes sociales

Alejandro Rosales-Martínez[1], Pablo Sotres-Castrejon[1], Griselda Velázquez[1], Esaú Villatoro-Tello[1,2], Gabriela Ramírez-de-la-Rosa[1,2]

[1] Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa, Maestría en Diseño, Información y Comunicación,, México

[2] Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa, Departamento de Tecnologías de la Información, México

{alesito500,p.sotres.c}@gmail.com, grisvillar@yahoo.com.mx, {evillatoro,gramirez}@correo.cua.uam.mx

Resumen. El perfilado de autores (PA) se ha convertido en una tarea muy relevante para la comunidad de Procesamiento del Lenguaje Natural. El objetivo principal del PA es determinar de forma automática características demográficas del autor, por ejemplo: género y edad. El PA tiene múltiples aplicaciones en áreas como la mercadotecnia y la lingüística forense, recientemente se investiga su utilidad en la identificación de trastornos, por ejemplo, detectar la depresión o anorexia. En este sentido, dentro de este trabajo presentamos una propuesta para resolver el problema de identificación de usuarios que padecen algún desorden mental; específicamente evaluamos la pertinencia de recursos léxicos que han sido generados desde el área de psicología. Para nuestros experimentos empleamos datos proporcionados por el foro eRisk. Nuestros resultados muestran que es posible identificar estos padecimientos por medio de emplear un conjunto reducido de términos para la construcción de la representación de los textos.

Palabras clave: procesamiento de lenguaje natural, perfilado de autores, representación de información, aprendizaje automático, clasificación no-temática de textos.

Identifying Signs of Anorexia and Depression in Social Media

Abstract. Author profiling (AP) has become an important task within the Natural Language Processing (NLP) field. The main goal of AP is to automatically determine demographics aspects from authors, for instance, age and gender. Despite the main applications of AP in the marketing and forensic fields, recently has been showed the utility of

AP techniques in preventing user's risk, such as detecting cues of mental illness. In this paper we describe a method for identifying signs of depression and anorexia in users' posts. Specifically, we evaluate the pertinence of psychological theory in the AP task. Our performed experiments were done using the data provided by the eRisk forum. Our results indicate that using a small number of features is possible to obtain comparable results against those obtained by traditional approaches.

Keywords: natural language processing, author profiling, knowledge representation, machine learning, non-thematic text classification.

1. Introducción

En la actualidad, Internet ha logrado tener un impacto importante en el mundo laboral, el ocio, y el conocimiento a nivel mundial. Gracias a Internet, millones de personas tienen acceso fácil e inmediato a una cantidad extensa y diversa de información en línea. Contrariamente a los medios de comunicación tradicionales, Internet ha permitido una descentralización repentina y extrema de la información; trayendo como consecuencia que una gran variedad de usuarios puedan gozar de estos beneficios.

Es claro que Internet, al volverse una parte importante de la vida cotidiana, permite a usuarios obtener cantidades significativas de información; así mismo, les permite mantener interactividad constante con otros usuarios a través de los servicios de mensajería instantánea o redes sociales tales como Facebook, Twitter, Instagram, Snapchat, etc. Estos servicios ofrecen atractivas ventajas, por ejemplo, permiten una fácil comunicación entre personas que pueden estar localizadas en distintos puntos geográficos; son muy sencillas de utilizar; no representan un costo para el usuario; además de ser medios virtuales y privados por naturaleza [11]; razones por la cuales su popularidad no se ha hecho esperar desde su aparición. De acuerdo al sitio flimper[3], durante el 2017, el número de usuarios activos de Facebook, la red social con mayor número de usuarios, es de aproximadamente de 1900 millones de personas; por otra parte, Twitter tiene 320 millones de usuarios activos generando un promedio de 500 millones de tuits al día.

A partir de la información que es producida por los usuarios de estas redes, áreas de investigación como lo son el Procesamiento de Lenguaje Natural (PLN), han centrado su atención en la diversidad de información vertida en la red, por ejemplo: vídeos, fotografías, opiniones, revisiones de productos, etc. Ejemplos de problemas que se han abordado en años recientes utilizando esta información son: identificación el estado de ánimo de las personas [8], predecir las fluctuaciones en la bolsa de valores [4], identificar a pedófilos en sitios web de conversaciones [7], así como la obtención de información general sobre el perfil de los usuarios [18], entre muchos otros.

[3] https://www.flimper.com/blog/es/2017-estadisticas-de-redes-sociales-facebook-instagram-linkedin-twitter-whatsapp

Específicamente, el perfilado de autor, una sub-disciplina del PLN, busca resolver el problema de identificar, a través de analizar el texto que escribe un usuario, características demográficas del autor de ese texto, por ejemplo: genero, edad, lenguaje nativo, preferencias políticas o religiosas, etc. Sin embargo, existen otros aspectos demográficos que son de interés no solo a la comunidad de computación, sino también a áreas de las Ciencias Sociales, particularmente a la Psicología; por ejemplo: la identificación de rasgos de personalidad, depresión, anorexia, etc., aspectos que se consideran como una dimensión más al problema de perfilado de autor [10].

En la actualidad, la depresión y la anorexia son trastornos que afectan a un gran número de personas en todo el mundo. Es un problema vigente con aproximadamente 350 millones de individuos que sufren este padecimiento [13]. Como se menciona en el estudio realizado por Goodwin y Jamison [9], la depresión es la principal causa de suicidio entre el 15 % y 20 % de pacientes que la padecen. Por otro lado, los datos referentes al crecimiento de pacientes con anorexia tampoco son alentadores, como lo indica la *National Eating Disorder Association*[4]: 70 millones de personas, tanto hombres como mujeres, sufren de problemas relacionados a desordenes alimenticios.

Este tipo de problemáticas pone en evidencia la necesidad de contar con herramientas computacionales que apoyen en la detección temprana de estos trastornos. Alertar a los individuos sobre la posibilidad de estar reflejando síntomas de un padecimiento de este tipo permitirá a los usuarios buscar un diagnóstico oportuno. Además, este tipo de herramientas se prevé servirán como sistemas de apoyo a la toma de decisiones, así como ayudar a disminuir la presencia de estos padecimientos en etapas avanzadas.

En este trabajo se propone un método automático para la identificación de depresión, y anorexia en usuarios de redes sociales. El método propuesto utiliza técnicas tradicionales de aprendizaje supervisado en combinación con estrategias de procesamiento de lenguaje natural. Nuestra hipótesis plantea que el sistema automático será más eficiente en la identificación de estos padecimientos al representar los documentos por medio de un conjunto cerrado de categorías de palabras, específicamente, palabras con funciones cognitivas y comunicativas muy particulares.

El resto del documento se organiza de la siguiente manera. La sección 2 se describe el trabajo relacionado más reciente; la sección 3 describe las características de los datos empleados para nuestros experimentos; la sección 4 muestra el método propuesto, en la sección 5 se describen los experimentos y los resultados obtenidos. Finalmente, la sección 6 plantea las conclusiones alcanzadas y proponen líneas de trabajo futuro.

2. Trabajo relacionado

El perfilado de autor es uno de los retos recientes que ha llamado la atención de la comunidad científica, en particular de áreas como el procesamiento de

[4] `https://www.eatingdisorderhope.com/blog/eating-disorders-world-overview`

lenguaje natural, ciencias forenses, estrategias de marketing y seguridad en internet. El objetivo principal del perfilado de autor (PA) es distinguir, a partir de un texto, entre clases de autores y no identificar a un autor en particular, siendo este último el escenario del problema conocido como atribución de autoría [16]. Así entonces, la tarea de PA busca modelar a través de atributos sociolingüísticos más generales a grupos de autores, dichos atributos son además indicadores de cómo los distintos grupos de autores emplean el lenguaje dependiendo de su género, edad y/o lenguaje nativo [2].

En el año 2017 se propone por primera vez una tarea de perfilado de autores donde las dimensiones que se desean identificar son condiciones mentales específicas, en concreto la identificación de usuarios con depresión [10]. Desde entonces, el foro de evaluación eRisk[5] convoca a los grupos interesados en este tipo de retos a presentar modelos computacionales que sean capaces de identificar anticipadamente usuarios con síntomas de depresión y anorexia.

Gran variedad métodos fueron propuestos en la edición 2017 de eRisk. Hubo propuestas de métodos que utilizaban sólo atributos léxicos, estadísticos, o atributos basados en emociones, representaciones basados en tópicos (LSA y LDA), métodos que empleaban representaciones basadas en grafos, y métodos que combinaban técnicas de recuperación de información en combinación con estrategias de aprendizaje supervisado. A pesar de la gran variedad de técnicas, el método que tuvo mejor desempeño en el 2017 fue el trabajo descrito en [6]. Este trabajo propone una representación semántica de los documentos que considera de manera explícita la información parcial de cada porción de texto que se va volviendo disponible. El enfoque temporal es complementado con técnicas tradicionales de categorización. Los resultados alcanzados por este método son de $F = 0.59$.

A pesar de los avances obtenidos, el problema de identificación usuarios con síntomas de depresión y de anorexia aún no está resuelto. Motivados por esta problemática, nuestro trabajo propone evaluar la pertinencia de la información psicolingüística contenida en los mensajes de los usuarios. Para esto, realizamos un análisis exhaustivo en busca del tipo de dimensiones y categorías psicológicas presentes en los textos de los usuarios. A diferencia del trabajo previo, nos interesa evaluar la pertinencia de un conjunto cerrado de categorías psicolingüísticas para hacer la representación de los documentos.

3. Datos

Para la realización de los experimentos se trabajó con los datos proporcionados por el foro eRisk, foro de evaluación que se realiza en conjunto con la conferencia CLEF[6]. Durante su primera edición en 2017, los organizadores del eRisk proponen la tarea de detección anticipada de depresión [10], mientras que para el 2018 se propuso también la detección anticipada de anorexia.

[5] http://erisk.irlab.org/
[6] http://clef2018.clef-initiative.eu/

La Tabla 1 muestra algunas estadísticas básicas de los datos con los que se trabajó durante la realización de nuestros experimentos.

Tabla 1. Estadísticas de la partición de entrenamiento de los datos de eRisk 2018.

Estadísticas	**Depresión**		**Anorexia**	
	positivo	*negativo*	*positivo*	*negativo*
Num. de usuarios	135	752	20	132
Num. posts	4,956	48,184	745	7,738
Num. tokens	186,928	1,197,350	46,771	191,770
Vocabulario	16,581	63,840	6,111	20,657
Promedio tokens/post	37.71	24.85	62.79	24.79
Promedio tokens/usuario	1384.65	1592.22	2338.56	1452.80
Promedio riqueza léxica	0.089	0.053	0.130	0.108

Un aspecto importante a resaltar en los datos es el desbalance de clases. Observe que la clase positiva para ambos problemas es la clase minoritaria. Como consecuencia de este desbalance, el número de textos totales de la clase positiva es mucho menor al de los negativos, por ejemplo, 745 contra 7,738 posts para el problema de anorexia. Sin embargo, es importante resaltar que en promedio, la longitud de los posts producidos por las clases positivas es mayor que los posts de los usuarios de la clase negativa, esto significa que en este corpus, los sujetos que tienen presente el padecimiento tienden a escribir textos más extensos.

Finalmente, es conveniente mencionar que debido a que eRisk plantea el problema de detección de depresión y de anorexia como problemas de clasificación temprana, los datos mostrados en la Tabla 1 son proporcionados a través de 10 porciones (chunks) ordenados cronológicamente. De esta forma, el primer chunk corresponde a los textos más antiguos producidos por los usuarios, mientras que el chunk 10 contiene los mensajes más recientes. Para la realización de nuestros experimentos se conservó esta forma de organización de los datos.

4. Método propuesto

Para resolver el problema de identificación de usuarios con depresión y anorexia se utilizó un esquema de clasificación de textos. La clasificación de textos es la tarea de asignar un documento a una o más categorías predefinidas con base en su contenido [15]. El primer paso obligado es el *indexado* de los documentos, en este caso los textos P. El indexado denota la actividad de hacer el mapeo del conjunto de textos de cada usuario i (*i.e.*, p_i) en una forma compacta de su contenido. La representación más comúnmente utilizada para representar textos es a través de un vector con términos ponderados como entradas, concepto tomado del modelo de espacio vectorial usado en recuperación de información. Esta representación permite que cada texto p_i sea representado como el vector $\overrightarrow{p}_i = \langle w_{ki}, \ldots, w_{|\tau|i} \rangle$, donde τ es el *vocabulario*, *i.e.*, el conjunto de términos que

ocurren al menos una vez en algún elemento de P, mientras que w_{ki} representa la importancia del término t_k dentro del contenido del documento p_i. Este método de representación, también conocido como bolsa de palabras (BoW), propone varios esquemas para definir w_{ki}, los más comunes son un ponderado booleano, ponderado por frecuencia (tf), y ponderado por frecuencia relativa (tf-idf) [3].

Como se mencionó en la introducción, nuestra hipótesis establece que para identificar adecuadamente a los sujetos que presentan algún trastorno, basta con representar los textos con un conjunto cerrado de categorías de palabras, en específico palabras con funciones cognitivas y comunicativas, las cuales tienen un significado dentro de la teoría psicológica. Para lograr esto, empleamos como recurso base el diccionario psicolingüístico LIWC [17].

LIWC (Linguistic Inquiry and Word Count) es un recurso léxico que está conformado por un total de 5,690 palabras, las cuales están asociadas a cuatro grandes dimensiones: procesos estándar, procesos psicológicos, aspectos personales, y actos del habla. En total, estas cuatro dimensiones contemplan 64 categorías de palabras. Para conocer en más detalle la conformación y el proceso de construcción de este recurso refiérase a [18, 15]. Algunos artículos de investigación recientes que han empleado LIWC como parte de su método para la identificación del perfil de autor, sobre todo en identificación de genero y edad, son [1, 14].

Motivados por el trabajo previo, nuestro método propone utilizar como términos del vocabulario (τ) solo aquellas palabras que pertenecen a las categorías de LIWC más representativas para cada una de las tareas. En otras palabras, se definió un vocabulario específico para depresión (τ_D), y uno para anorexia (τ_A). Note que tanto τ_D como τ_A son subconjuntos de LIWC.

Para identificar el lenguaje más representativo se hizo un análisis que permitiera detectar aquellas categorías de palabras que son claramente utilizadas en proporciones diferentes entre los usuarios de la clase positiva y los de la clase negativa. Este análisis se hizo para ambos problemas de clasificación, es decir, depresión y anorexia. En la figura 1 y figura 2 se muestra a través de una gráfica de barras el grado de importancia de cada una de las 64 categorías de LIWC en los problemas de depresión y anorexia respectivamente. Para esto, se contabilizó la frecuencia de aparición de los términos de cada una de las categorías de LIWC tanto en la clase *positiva* como en la *negativa*. Las frecuencias obtenidas se normalizan por el tamaño de los documentos de su respectiva clase. Finalmente, la diferencia entre las frecuencias obtenidas es lo que nos permite identificar las categorías LIWC más representativas para cada problema.

Observe que para depresión (figura 1) solo 7 categorías tienen un porcentaje de uso distinto mayor al 40 %. Estas categorías son: *i, article, family, friend, anx, sad, health*. De este análisis es importante resaltar la presencia de la categoría *'i'*, misma que refiere al uso de pronombres personales. Este hallazgo ha sido discutido previamente en [14], donde se menciona que las personas con depresión usan en mayor cantidad de palabras como *I*, *Me* y *My*, debido a que cuando las personas se deprimen tienden a enfocarse más en ellos mismos, prestando menos atención al mundo a su alrededor. La presencia de las categorías *sad*

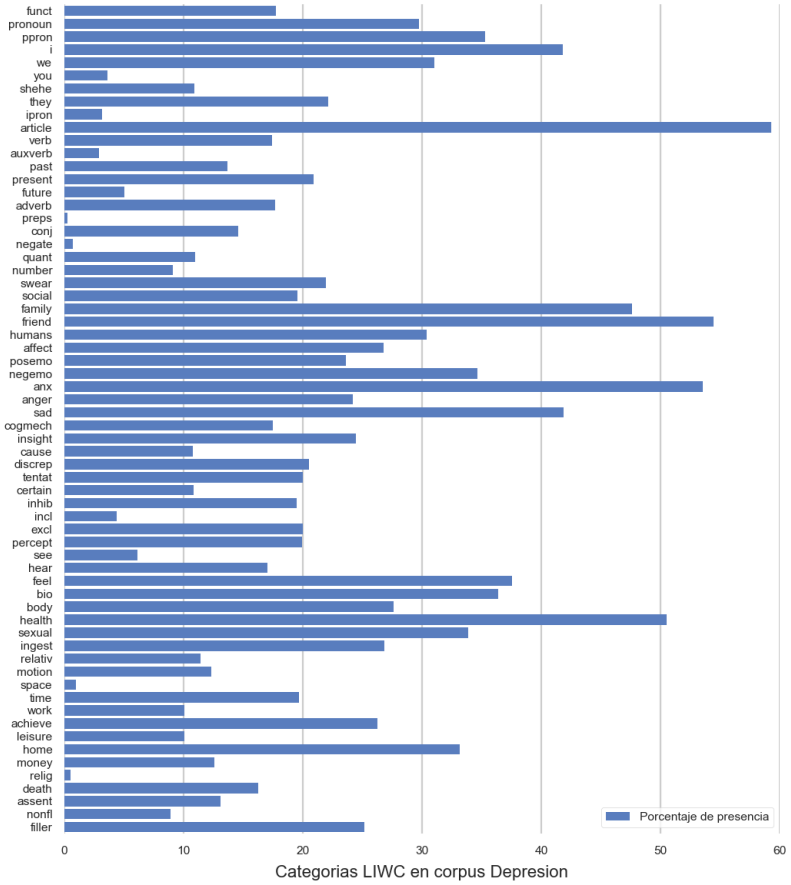

Fig. 1. Porcentaje de presencia de las categorías de LIWC en el corpus de usuarios con depresión.

(tristeza) y *anx* (ansiedad), son conjuntos de palabras que pertenecen a una familia de palabras relacionadas a procesos afectivos, ejemplos de palabras que caen en estas categorías son *nervous, afraid, tense, grief, cry, sad.* Finalmente, las categorías *family* y *friend* son conjuntos de palabras que aluden a procesos sociales, los cuales se ven afectados en personas con depresión.

Respecto al corpus de anorexia (figura 2) el análisis arrojó que existen 13 categorías con un porcentaje de uso distinto mayor al 40 %, *i, you, they, article,*

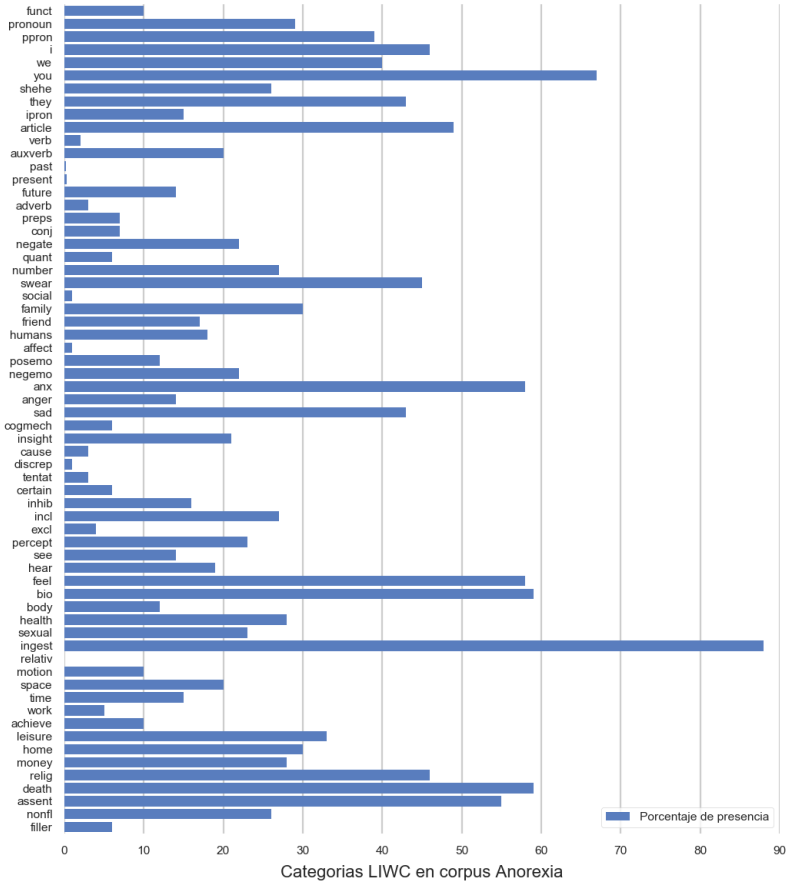

Fig. 2. Porcentaje de presencia de las categorías de LIWC en el corpus de usuarios con anorexia.

swear, anx, sad, feel, bio, ingest, relig, death, assent. Entre las más relevantes es la categoría de *ingest*, la cual es una familia de palabras que refieren al consumo de alimentos y en general a procesos biológicos. Otro aspecto relevante es el uso de las categorías *you* y *they*, es decir palabras que refieren al uso de pronombres personales en 2a y 3a persona. Este aspecto es importante, pues nos hace suponer que los usuarios anoréxicos, contrario a los usuarios con depresión, son más consientes de otras personas.

Finalmente, otro aspecto que llamó nuestra atención es como las palabras ofensivas (*swear*) tienen una presencia importante.

Los resultados de este análisis indican que existen diferencias importantes en el uso del lenguaje entre los usuarios que tienen, y los que no tienen, los trastornos de depresión y de anorexia. Así entonces, para la conformación de los vocabularios τ_D y τ_A se tomó el vocabulario de las categorías que tuvieron un porcentaje de presencia mayor al 35 % respectivamente.

5. Configuración experimental

En esta sección se describe la configuración experimental. Comenzaremos describiendo el método base, las métricas de evaluación, y finalmente se discuten los resultados obtenidos.

5.1. Método base

Como método base se utilizó como forma de representación una bolsa de palabras (BoW) tradicional, es decir, se emplea todo el vocabulario de la colección P para calcular la representación. A esta configuración la denominamos como "ALL" en los experimentos realizados.

Además de lo anterior, dos variantes del método base fueron evaluadas. La modificación consistió en emplear los k términos más frecuentes para construir la representación. Esta variante se inspiró en algunos trabajos previos, los cuales han mostrado que solo empleando los términos más frecuentes de la colección es suficiente para representar la semántica de los documentos de las distintas clases [1,5]. De esta forma, se emplearon valores de $k = 1000$ y $k = 5000$.

Para la construcción de la representación de bolsa de palabras se empleo la implementación disponible en SciKitLearn[7]. Como esquemas de pesado se utilizó: booleano (BOOL), TF y TF-IDF.

5.2. Clasificador

El algoritmo de aprendizaje utilizado fue Naïve Bayes (NB). Este método de aprendizaje se considera como parte de los clasificadores probabilísticos, los cuales se basan en la suposición que las cantidades de interés se rigen por distribuciones de probabilidad, y que la decisión óptima puede tomarse por medio de razonar acerca de esas probabilidades junto con los datos observados [12]. Para los experimentos realizados utilizamos la implementación de bayes proporcionada por SciKitLearn[7] con sus parámetros por defecto.

[7] `http://scikit-learn.org/stable/`

5.3. Evaluación

Como métrica de evaluación principal utilizamos la medida F, la cual se define como se muestra en la ecuación (1):

$$medida - F = \frac{(1+\beta^2)P * R}{\beta^2 P + R}, \tag{1}$$

donde con $\beta = 1$ representa la media armónica entre la precisión y el recuerdo. La precisión (P) es la proporción de documentos clasificados correctamente en una clase c_i con respecto a la cantidad de documentos clasificados en esa misma clase. El recuerdo (R), la proporción de documentos clasificados correctamente en una clase c_i con respecto a la cantidad de documentos que realmente pertenecen a esa clase. Así, la precisión se puede ver como una medida de la corrección del sistema, mientras que el recuerdo da una medida de cobertura o completitud.

Como se mencionó en la sección 3, los datos están divididos en 10 *chunks*. Para la realización de los experimentos se entrenó y evaluó un modelo de clasificación por cada chunk empleando una estrategia de validación cruzada de 10 pliegues para cada experimento. Así entonces, los resultados mostrados en las tablas 2 y 3 representan el promedio del desempeño obtenido en los 10 chunks.

5.4. Resultados

Las tablas 2 y 3 muestran los resultados de los experimentos realizados. Los resultados se reportan en términos de la medida F sólo para la clase de interés, es decir, la clase positiva. Observe que el mejor resultado obtenido en los experimentos base (tabla 2) para el problema de depresión es cuando se utilizan los 5 mil términos más frecuentes con un esquema de pesado binario (BOOL). En forma similar, los resultados para detección de anorexia muestran que es conveniente emplear los cinco mil términos más frecuentes, pero contrario al problema de depresión, aquí se vuelve relevante el esquema de pesado, resultando TF-IDF como el mejor esquema de ponderación de términos.

Los resultados obtenidos por el método base indican que, mientras que para el problema de identificación de usuarios con depresión basta con la aparición (o no) de ciertos términos, para detectar a los usuarios con anorexia, es necesario considerar las frecuencias relativas de dichos términos.

La tabla 3 muestra los resultados de utilizar los diccionarios τ_D y τ_A para construir la representación de los datos de depresión y anorexia respectivamente (vea sección 4). Note que el número de atributos empleado para la representación de los documentos con el método propuesto es significativamente menor en comparación al método base (5000). En promedio, se requieren de 927 atributos para el corpus de depresión y 608 para el de anorexia.

A pesar de que no es posible superar al mejor resultado del método base, los resultados obtenidos con nuestro método son alentadores. Observe que para el caso de identificación de depresión se obtiene un $F = 0.456$ empleando un esquema de pesado de TF en comparación con un $F = 0.473$ que se obtuvo en el método base bajo la misma configuración. De manera similar, para el

Tabla 2. Resultados empleando una representación tradicional de bolsa de palabras. Como medida de evaluación se empleó la métrica F de la clase positiva.

Esquema de pesado	**Num. de atributos**	**Medida F**	
		Depresión	*Anorexia*
BOOL	1000	0.446	0.284
	5000	**0.488**	0.377
	ALL	0.012	0.070
TF	1000	0.459	**0.594**
	5000	0.473	0.574
	ALL	0.146	0.417
TF-IDF	1000	0.423	0.591
	5000	0.462	0.586
	ALL	0.351	0.350

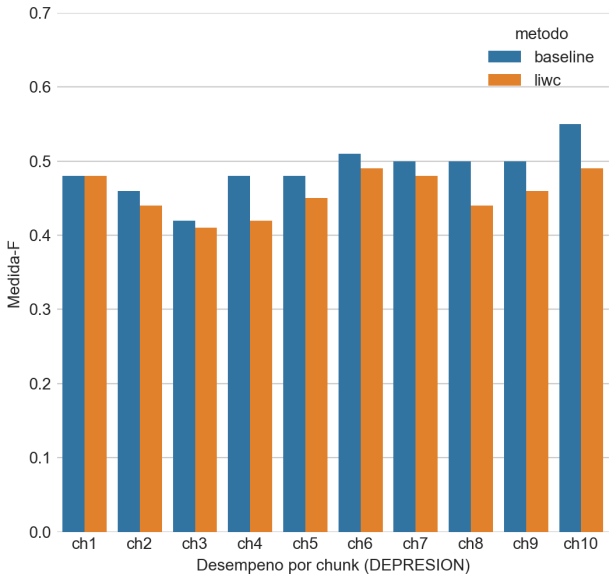

Fig. 3. Resultados por chunk para el problema de depresión.

problema de identificación de usuarios con anorexia se obtiene un $F = 0.494$ con nuestro método contra un $F = 0.594$ obtenido por el método base bajo la misma configuración.

Las figuras 3 y 4 muestran el desempeño tanto del mejor método base como del método propuesto para los problemas de identificación de depresión y

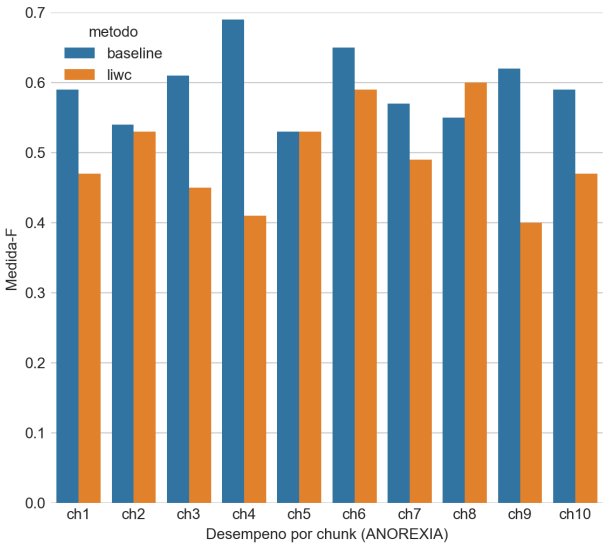

Fig. 4. Resultados por chunk para el problema de anorexia.

Tabla 3. Resultados empleando como representación las categorías psicolingüísticas de LIWC. Como medida de evaluación se empleó la métrica F de la clase positiva.

Problema	**Num. de atributos**	**Esquema de pesado**		
		BOOL	**TF**	**TF-IDF**
Depresión	927	0.224	**0.456**	0.426
Anorexia	608	0.358	**0.494**	0.474

anorexia respectivamente. Como se puede observar, el desempeño del método propuesto es muy cercano al mejor baseline para el problema de depresión (figura 3). Incluso se observa que en el primer chunk, nuestro método es capaz de igualar el desempeño del método base.

Por otro lado, para el caso de identificación de anorexia, el desempeño obtenido en cada chunk muestra que las diferencias entre el método base y el método propuesto son mayores. Sin embargo, el método propuesto es capaz de igualar al método base en el chunk 5 e incluso obtiene un mejor desempeño en el chunk 8.

6. Conclusiones y trabajo a futuro

Este artículo describe la metodología propuesta para identificar perfiles psicológicos de los usuarios de redes sociales. En específico nos enfocamos en el pro-

blema de identificación de usuarios con depresión y anorexia. Nuestra hipótesis de trabajo plantea que es posible identificar a los usuarios que presentan dichos trastornos por medio de utilizar un conjunto muy reducido de palabras que tienen un significado dentro de la teoría psicolingüística.

Para comprobar la validez de nuestra hipótesis se utilizó como recurso el diccionario LIWC, el cual define cuatro grandes dimensiones psicológicas. Para la realización de nuestros experimentos se utilizó el corpus proporcionado por eRisk. Los resultados obtenidos son alentadores, se mostró que usando entre un 1.5 % y un 3 % de atributos es posible obtener un desempeño similar al de métodos que emplean todo el vocabulario para la construcción de la representación.

Como trabajo futuro queremos explorar técnicas de fusión de información para la construcción de la representación. Nos interesa evaluar tanto técnicas de fusión temprana (ealy-fusion) como fusión tardía (late-fusion) para la construcción de la representación. Además de esto, es de nuestro particular interés incorporar información de comportamiento. La hipótesis detrás de esta idea es que los usuarios con un padecimiento tendrán comportamientos diferentes al de usuarios que no presentan un perfil depresivo y/o anoréxico.

Agradecimientos. El trabajo de los primeros tres autores fue parcialmente financiado por el CONACyT a través de las becas de maestría 836519, 673283, 869688 respectivamente. El trabajo de los dos últimos autores fue financiado a través del proyecto CONACyT CB-2015 No. 258588. También se agradece el apoyo otorgado a través de la Coordinación de la Maestría en Diseño, Información y Comunicación (MADIC) de la UAM Cuajimalpa, así como al Departamento de Tecnologías de la Información de la UAM Cuajimalpa.

Referencias

1. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Meza, I.: Evaluating topic-based representations for author profiling in social media. In: Ibero-American Conference on Artificial Intelligence. pp. 151–162. Springer (2016)
2. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. TEXT 23, 321–346 (2003)
3. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
4. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), 1 – 8 (2011)
5. Chung, C., Pennebaker, J.W.: The psychological functions of function words. Social communication 1, 343–359 (2007)
6. Errecalde, M.L., Villegas, M.P., Funez, D.G., Ucelay, M.J.G., Cagnina, L.C.: Temporal variation of terms as concept space for early risk prediction. In: Proceedings Conference and Labs of the Evaluation Forum CLEF 2017 (2017)
7. Escalante, H.J., Villatoro-Tello, E., Juarez, A., Montes-y-Gomez, M., Villaseñor, L.: Sexual predator detection in chats with chained classifiers. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment

and Social Media Analysis. pp. 46–54. Association for Computational Linguistics, Atlanta, Georgia (2013)
8. Golder, S.A., Macy, M.W.: Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science 333(6051), 1878–1881 (2011)
9. Goodwin, F.K., Jamison, K.R.: Manic-depressive illness: bipolar disorders and recurrent depression, vol. 1. Oxford University Press (2007)
10. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental foundations. In: Proceedings Conference and Labs of the Evaluation Forum CLEF 2017. Dublin, Ireland (2017)
11. Miah, M.W.R., Yearwood, J., Kulkarni, S.: Detection of child exploiting chats from a mixed chat dataset as a text classification task. In: Proceedings of the Australasian Language Technology Association Workshop 2011. pp. 157–165 (2011)
12. Mitchell, T.M., et al.: Machine learning. 1997. Burr Ridge, IL: McGraw Hill 45(37), 870–877 (1997)
13. Organization, W.H.: The World Health Report 2001: Mental health: new understanding, new hope. World Health Organization (2001)
14. Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., Booth, R.J.: The Development and Psychometric Properties of LIWC2007. This article is published by LIWC Inc, Austin, Texas 78703 USA in conjunction with the LIWC2007 software program., `http://www.liwc.net/LIWC2007LanguageManual.pdf`
15. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1), 1–47 (2002)
16. Stamatatos, E.: A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol. 60(3), 538–556 (Mar 2009)
17. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of Language and Social Psychology 29, 24–54 (2010), `http://homepage.psy.utexas.edu/homepage/students/Tausczik/Yla/index.html`
18. Villatoro-Tello, E., Ramírez-de-la-Rosa, G., Sánchez-Sánchez, C., Jiménez-Salazar, H., Luna-Ramírez, W.A., Rodríguez-Lucatero, C.: UAMCLyR at RepLab 2014: Author profiling task. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 1547–1558 (2014)