ISSN 1870-4069

Identificación y pesado de términos para la detección de depresión en Twitter

Monserrat Vázquez-Hernández, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez

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

{mvazquez, villasen, mmontesg}@inaoep.mx

Resumen. Las redes sociales son parte de la vida diaria de las personas, a través de éstas los usuarios comparten información acerca de cualquier tema incluyendo aspectos emocionales, de ahí el interés de aprovecharlas para la detección de usuarios que sufren ciertos trastornos mentales. En este trabajo se propone un método para la identificación del léxico específico asociado a la depresión en Twitter. Este método identifica los términos más discriminativos y posteriormente construye una nueva representación de los textos al calcular el peso del resto de los términos con respecto al léxico identificado. Para validar el esquema de pesado propuesto se implementó una red neuronal de tipo CNN-LSTM para la detección de tuits relacionados con depresión. Los resultados obtenidos indican que este nuevo esquema de pesado mejora la detección de mensajes de usuarios con depresión en *Twitter*.

Palabras clave: Detección de depresión, Twitter, pesado de términos.

Term Identification and Weighing for Depression Detection in Twitter

Abstract. Social networks are part of people's daily lives, through these, users share information about any topic including emotional aspects, hence the interest in taking advantage of them to detect users suffering from certain mental disorders. This work proposes a method for identifying the specific lexicon associated with depression on Twitter. This method identifies the most discriminating terms and subsequently constructs a new representation of the texts by calculating the weight of the rest of the terms with respect to the identified lexicon. To validate the proposed weighing scheme, a CNN-LSTM neural network was implemented to detect tweets related to depression. The results obtained indicate this new weighing scheme improves the detection of messages from depressed users on *Twitter*.

Keywords: Depression detection, twitter, term weighing.

1. Introducción

Más de 300 millones de personas sufren depresión a nivel mundial siendo ésta la principal causa de ausentismo laboral en todo el mundo. Según la Organización Mundial de la Salud (OMS), en el mundo una de cada trece personas sufre depresión o algún trastorno mental relacionado[1].

Actualmente, las redes sociales son una parte integral en nuestras vidas diarias, a través de ellas los usuarios comparten información sobre sus intereses, opiniones, planes futuros e incluso sobre su estado emocional, por ejemplo, sus alegrías y tristezas. Es por ello que estas plataformas brindan una oportunidad para transformar los servicios de detección temprana de diferentes trastornos mentales como la depresión. Diversos trabajos de investigación se han enfocado en la detección de depresión a partir del contenido de redes sociales siguiendo, en su mayoría, un enfoque supervisado de clasificación de textos [1].

Twitter es un medio dinámico en el cual sus usuarios comparten mensajes frecuentemente durante el transcurso de sus actividades diarias. Esta característica ha convertido a este medio en uno de los más usados para la detección de personas que sufren depresión [2]. Cabe mencionar que la tarea no es sencilla, en primer lugar, los tuits son mensajes muy cortos que normalmente no proveen información de contexto y en segundo lugar, los indicadores de depresión suelen manifestarse de manera muy sutil, por lo tanto, su detección no es obvia para el lector. A pesar de estas características, trabajos recientes han reportando resultados muy alentadores en la detección de usuarios que sufren depresión y de otros trastornos mentales [3].

La mayoría de los métodos previos, orientados a la detección automática de depresión en redes sociales, han seguido un enfoque supervisado que los hace depender de datos etiquetados para su construcción. Estos métodos han explorado distintas representaciones y técnicas de clasificación, desde técnicas tradicionales hasta basadas en aprendizaje profundo [11,13,14,15]. En contraste, en este trabajo proponemos un esquema de pesado de términos orientado al dominio de depresión. Éste se basa en la identificación de los términos más discriminativos relacionados con depresión a partir de un conjunto etiquetado de referencia y posteriormente, el pesado del resto del vocabulario con respecto a su cercanía semántica con los términos identificados. Este nuevo enfoque de pesado, al igual que los tradicionales TF y TF-IDF, puede usarse en conjunto con distintos clasificadores permitiéndoles ponderar de mejor manera la información del dominio de aplicación, depresión en nuestro caso, y con ello su identificación.

El resto del artículo está organizado de la siguiente manera: la sección 2 presenta información acerca de trabajos relacionados donde previamente se ha abordado la tarea de detección de depresión en *Twitter*, en la sección 3 se describe el modelo propuesto para la identificación de los términos más frecuentes y con ellos la definición de pesos del vocabulario, en la sección 4 se evalúa el modelo mediante la implementación de diferentes esquemas de clasificación incluyendo

[1] https://www.who.int/es/news-room/fact-sheets/detail/depression

una red neuronal CNN-LSTM para la tarea de predicción, por último, en la sección 5 se presentan conclusiones y posibles direcciones futuras de trabajo.

2. Trabajo relacionado

La depresión es una enfermedad mental que afecta el equilibrio emocional de las personas. Su detección está dada por diferentes patrones de comportamiento de los individuos que la padecen [4]. En los últimos años, con el surgimiento y expansión del uso de redes sociales, diversas investigaciones se han llevado a cabo para la detección de este trastorno abordando el tema como un problema de *clasificación de textos*.

Estos trabajos han mostrado que el uso de ciertas palabras por parte de un usuario pueden dar pistas importantes sobre su condición psicosocial. Estudios en psicología sugieren que el uso de palabras específicas en nuestro vocabulario revelan indicadores sobre nuestro estado emocional [5]. Diferentes trabajos relacionados con la detección y análisis de depresión revelan que las palabras que usan las personas deprimidas permiten distinguirlas de entre el resto de las personas [3].

Siguiendo esta misma línea de trabajo, Tsugawa et al. así como Nadeem [6,7] realizaron un análisis respecto al vocabulario usado en tuits de usuarios con y sin depresión. Concluyen que observar la frecuencia de ciertas palabras es de gran utilidad para distinguir el comportamiento de una persona con este trastorno. De manera similar, Cavazos-Rehg et al. [8] realizaron un estudio exploratorio de tuits de varios usuarios observando que de acuerdo a las palabras más frecuentes, los usuarios pueden ser asociados al trastorno de estudio.

Por su parte, Stankevich et al. [9] combinaron un enfoque de bolsa de palabras con características estilométricas y morfológicas para la detección de depresión. Usaron representaciones distribuidas (*embeddings*) de palabras seleccionando las n palabras más informativas para cada usuario y obteniendo su representación a partir del promedio ponderado de los vectores de esas palabras. En sus experimentos observaron que el enfoque basado en *embeddings* aportó resultados inferiores a los obtenidos con el enfoque de bolsa de palabras, lo que les permitió reafirmar la utilidad de la frecuencia del uso de palabras para la predicción de depresión a partir de textos de redes sociales.

Basados en la evidencia previa de la utilidad de recursos léxicos, Losada y Gamallo [10] propusieron varios modelos para representar los textos a través de léxicos predefinidos y explotar las ventajas que éstos ofrecen. Después de hacer una experimentación con diferentes léxicos concluyeron que, hacer énfasis en la presencia de adjetivos ayuda a la detección de expresiones relacionadas con la depresión.

En resumen, varias de las investigaciones previas coinciden en que los textos de usuarios que padecen depresión muestran una existencia frecuente de emociones negativas y un constante uso de palabras referidas a síntomas y medicamentos relacionados con la enfermedad, por lo tanto, utilizan un vocabulario muy similar al momento de expresarse.

Esto indica que el análisis de frecuencia en el uso de ciertas palabras es un elemento distintivo para detectar la depresión. Nuestro enfoque se soporta en esta observación para proponer un nuevo esquema de pesado de palabras particularmente orientado a la tarea de detección de depresión.

3. Esquema de pesado propuesto

El propósito de este trabajo es diseñar un nuevo esquema de pesado de términos orientado a la detección de depresión. Este proceso parte de un corpus de textos etiquetados como deprimidos y no-deprimidos. A partir de ellos, en una primera etapa se identifican los términos más discriminativos de la depresión, es decir, se crea un léxico de depresión. Posteriormente, en una segunda etapa, se calcula un peso para cada término del vocabulario indicando su cercanía semántica con las palabras del léxico extraído. Las siguientes subsecciones explican los dos pasos de este proceso y la siguiente sección muestra su aplicación en la identificación de tuits relacionados con la depresión.

3.1. Identificación de un léxico de depresión

Para el análisis del lenguaje se utilizan publicaciones de *Twitter* etiquetadas como deprimidos y no deprimidos. El principal objetivo en esta etapa es detectar aquellas palabras más representativas de los tuits etiquetados como deprimidos.

Para la extracción de este léxico, y para los experimentos posteriormente reportados, se utilizó un conjunto de datos en inglés que consta de 11,595 tuits de los cuales 8,000 son de la clase no-depresivos y 3,595 de la clase depresiva. [2]. Para los experimentos en este trabajo, los datos se dividieron en conjuntos de entrenamiento y prueba de manera aleatoria: el 80 % (9,276 tuits) se utilizó para la extracción del léxico de depresión y el cálculo del peso de los términos, además fueron usados para entrenar los distintos clasificadores; el restante 20 % (2,319 tuits) se utilizó para evaluar el desempeño de los clasificadores.

Los pasos que corresponden al proceso de extracción del léxico de depresión son los siguientes:

1. Preprocesamiento de tuits. El conjunto de datos (entrenamiento y prueba) fue preprocesado de la siguiente manera:
   - Se convierten a minúsculas todos los tuits.
   - Se realiza tokenización de los tuits.
   - Se eliminan espacios redundantes y referencias a URLs.
   - Se eliminan las referencias a usuarios '@' de *Twitter*.
   - Se eliminan signos de puntuación y números.

[2] Se utiliza el conjunto de datos propuesto en el trabajo https://github.com/viritaromero/Detecting-Depression-in-Tweets el cual contiene 10,314 tuits (8,000 no-depresivos y 2,314 depresivos), adicionalmente se agregaron 1,281 tuits de la clase depresiva extraídos de Twitter con la herramienta Twint https://github.com/twintproject/twint

- Se eliminan palabras vacías haciendo uso de las palabras definidas en la librería NLTK de Python para el idioma inglés.
- Se aplica un proceso de lematización.

Extracción de las palabras más frecuentes tanto para el subconjunto de tuits depresivos como para no-depresivos. Algunas de estas palabras se muestran en la Tabla 1.

Tabla 1. Palabras más frecuentes en tuits deprimidos y no deprimidos.

Deprimidos	**No deprimidos**
1. depression	1. love
2. care	2. thank
3. sad	3. day
4. hate	4. good
5. hurt	5. think
6. stress	6. hope
7. problem	7. great
8. suicide	8. well
9. pain	9. fun
10. heart	10. happie

2. Cálculo de la frecuencia normalizada de cada una de las palabras extraídas. Esta frecuencia se calcula por separado para los vocabularios de los tuits depresivos y no-depresivos de acuerdo con la fórmula 1, donde $freq(t, C)$ es la frecuencia del término t en los tuits de la clase C (depresivo o no-depresivo) y $freq_{max}(C)$ es la frecuencia correspondiente al término más común de la clase C:

$$freq_{norm}(t) = freq(t, C)/freq_{max}(C). \quad (1)$$

   El puntaje resultante, expresado por la frecuencia normalizada se obtiene de cada término en cada una de las clases, por lo que, los términos presentes en ambas tendrán dos puntajes. En la Tabla 2 se muestran algunos de los puntajes obtenidos.
3. Selección del conjunto de términos que conforman el léxico de depresión denotado como $\mathcal{L}_{dep}$. Para ello se consideran todos los términos cuya frecuencia normalizada sea mayor a un umbral β predefinido (el umbral β puede tomar valores de 0.01 a 0.06). En caso de que la palabra ocurra en los vocabularios de ambas clases, se mantienen solo aquellas con frecuencia normalizada mayor en la clase depresión que en la de no-depresión.

   La Tabla 3 muestra ejemplos de los léxicos de depresión para dos umbrales distintos.

Tabla 2. Frecuencia normalizada de los términos en tuits depresivos y no depresivos.

Tuits con depresión		Tuits sin depresión	
Palabra	$freq_{norm}$	Palabra	$freq_{norm}$
depression	1.000	love	1.000
care	0.070	thank	0.803
sad	0.058	day	0.699
hate	0.056	good	0.696
hurt	0.056	think	0.445
stress	0.049	hope	0.413
problem	0.043	great	0.385
suicide	0.042	well	0.380
pain	0.041	fun	0.261
heart	0.040	happie	0.259

Tabla 3. Ejemplos de palabras de léxicos de depresión para umbrales β distintos.

$\mathcal{L}_{dep}(\beta = 0{,}02)$	$\mathcal{L}_{dep}(\beta = 0{,}03)$	$\mathcal{L}_{dep}(\beta = 0{,}04)$
depression	depression	depression
people	people	people
someone	life	life
care	care	care
hurt	hurt	hurt
stress	sad	sad
sad	fuck	stress
understand	hate	cause
problem	stress	understand
anxiety	problem	anxiety
pain	suicide	suicide
suicide	anxiety	sometimes

3.2. Pesado de términos orientado a la tarea

Una vez identificado el léxico representativo de usuarios depresivos ($\mathcal{L}_{dep}$), éste es utilizado para calcular el peso del resto de los términos del vocabulario. El peso se determina a partir de la *cercanía semántica* de cada término con respecto al léxico pre-establecido, para lo cual se utilizan representaciones distribuidas (*embeddings*) de los términos. Los *word embeddings* se definen como una técnica para modelar el lenguaje que representa cada palabra a través de un vector que es calculado observando los contextos de uso de cada palabra en grandes corpus de documentos [11]. En nuestro caso, se utilizaron las representaciones de *word embeddings* de GloVe (Global Vector for Word Representation) calculados sobre *Wikipedia* (6B tokens, 400K vocab, uncased 300d vectors) [12].

El cálculo de los pesos de los términos se realiza mediante los siguientes pasos:

1. Se determina el vector (*word embedding*) de cada una de las palabras del léxico de depresión $\mathcal{L}_{dep}$.
2. Se calcula un vector que represente a todo el léxico del lenguaje depresivo. Este vector único lo denominaremos *embedding depression vector* (EDV) y lo definimos como el vector promedio de los vectores de los términos en $\mathcal{L}_{dep}$. El objetivo de este vector es capturar la "semántica"del lenguaje depresivo.
3. Se determina el grado de asociación entre un término cualesquiera y lenguaje depresivo modelado en EDV, para ello se mide la similitud coseno entre el vector del término a ponderar y el vector EDV. De este modo, el peso de cada término corresponde a su similitud con el vector EDV.
   En la Tabla 4 se presenta un listado de algunas de las palabras fuertemente relacionadas, de acuerdo a la similitud coseno calculada, al léxico de depresión $\mathcal{L}_{dep}$, considerando un umbral β =0.01. En este listado se pueden observar términos asociados a emociones negativas además de algunos relacionados con síntomas y tratamientos de la depresión.

Tabla 4. Palabras fuertemente relacionadas al léxico de depresión, correspondientes a $\mathcal{L}_{dep}(\beta$ =0.01)

Término	Similitud con EDV	Término	Similitud con EDV
need	0.697	treatment	0.583
pain	0.648	patient	0.577
suffer	0.644	lose	0.561
fear	0.641	chronic	0.543
treat	0.636	hurt	0.534
rather	0.621	wrong	0.527
care	0.614	trauma	0.523
help	0.613	afraid	0.520
problem	0.612	sick	0.511
stress	0.610	depression	0.499
enough	0.603	cure	0.498
mental	0.585	effect	0.492
heart	0.584	respect	0.490

4. Evaluación del modelo de predicción

Diferentes trabajos se han propuesto para la predicción de depresión en redes sociales. Estos esfuerzos se han realizado en diferentes niveles, es decir, se puede considerar todo el historial de un usuario como un único documento a partir

del cual se estimará la predicción ó se puede dar una estimación para cada tuit individualmente. En este trabajo, la tarea de predicción se realiza a nivel de tuit.

Para concluir sobre el beneficio del esquema de pesado propuesto, éste se comparó contra un esquema de pesado tradicional (*tf-idf*) utilizando diferentes algoritmos de clasificación.

Clasificadores Para evaluar el esquema de pesados propuesto se utilizaron los siguientes algoritmos de clasificación: 1) Clasificador Naïve Bayes, 2) *k-nearest neighbors* y 3) un modelo SVM. En cada caso los textos fueron representados usando una bolsa de palabras (BoW) con (i) pesado *tf-idf* tradicional y (ii) usando el esquema de pesado propuesto (es decir, pesando cada palabra respecto al EDV depresivo). Adicionalmente, se implementó una arquitectura tradicional usada en aprendizaje profundo de tipo CNN-LSTM, los parámetros utilizados para la configuración de la arquitectura están basados en implementaciones anteriores para tareas de clasificación de textos: num. épocas=3, num. de filtros=64, tamaño de los filtros=5, dropout = 0.2 [3]. Usando esta red se realizaron dos experimentos: i) las palabras en los tuits se representan con sus correspondientes *embeddings* pre-entrenado de GloVe y ii) de igual forma, las palabras en los tuits se representan con sus correspondientes *embeddings* pre-entrenados de GloVe, seguido de la multiplicación de cada *embedding* por el peso propuesto en función de su relación con el vocabulario EDV depresivo, de esta forma se premió o castigó a las palabras.

Tabla 5. Comparación de resultados para la clase depresiva entre los esquemas de pesado tradicional (*tf-idf*) y el propuesto, con diferentes tipos de clasificadores. Se muestran diferentes umbrales para determinar el vocabulario del EDV depresivo.

	Precisión	Cobertura	F1-score
Naïve Bayes + tf-idf	0.91	0.87	0.88
Naïve Bayes + EDV $\mathcal{L}_{dep}(\beta = 0.03)$	**0.95**	0.83	**0.89**
Naïve Bayes + EDV $\mathcal{L}_{dep}(\beta = 0.04)$	0.95	0.82	0.88
K-NN (k = 5) + tf-idf	0.58	0.26	0.36
K-NN (k = 5) + EDV $\mathcal{L}_{dep}(\beta = 0.03)$	**0.66**	0.27	**0.38**
K-NN (k = 5) + EDV $\mathcal{L}_{dep}(\beta = 0.04)$	0.59	0.27	0.37
SVM + tf-idf	0.93	0.88	0.90
SVM + EDV $\mathcal{L}_{dep}(\beta = 0.03)$	0.94	0.90	0.92
SVM + EDV $\mathcal{L}_{dep}(\beta = 0.04)$	**0.97**	**0.90**	**0.93**
CNN-LSTM pretrained GloVe	0.69	0.50	0.58
CNN-LSTM pretrained GloVe * peso $\mathcal{L}_{dep}(\beta = 0.03)$	**0.68**	**0.64**	**0.66**
CNN-LSTM pretrained GloVe * peso $\mathcal{L}_{dep}(\beta = 0.04)$	0.68	0.63	0.65

Resultados En la Tabla 5 se muestran los resultados obtenidos en cada clasificador (precisión, cobertura y F1-score) para la clase de tuits depresivos. Como se puede observar, el esquema de pesado propuesto mejora el esquema tradicional *tf-idf*. La Tabla 5 muestra el efecto al considerar diferentes umbrales ($\beta = 0{,}03$ y $\beta = 0{,}04$) para determinar el vocabulario del EDV depresivo. En general, con un umbral más exigente, es decir, con un conjunto menor de palabras para formar el EDV se alcanzan mejores resultados.

5. Conclusiones y trabajo futuro

En este trabajo se propuso un nuevo esquema de pesado de términos para la detección de signos de depresión en publicaciones de *Twitter*. El peso de los términos se estimó en función de su similitud respecto al vector único del léxico asociado al lenguaje depresivo. Para identificar este léxico se presentó un método basado en las palabras más frecuentes. Los resultados alcanzados indican que el método propuesto mejora la clasificación en comparación a las representaciones tradicionales.

Como trabajo futuro se busca implementar el esquema propuesto con representaciones de *embeddings* diferentes como FastText, además evaluar el esquema en otros conjuntos de datos para la tarea de detección de depresión, lo que nor permita comparar nuestro esquema contra otros métodos. Por otro lado, dada la generalidad del método, éste podría evaluarse en otro tipo de tareas de clasificación como perfilado de autor o detección de engaño. Adicionalmente, sería posible aplicarlo a colecciones de datos en otros idiomas, en específico para el idioma español.

Agradecimientos. El primer autor agradece el apoyo otorgado por el CONACYT a través de la beca No. 650291. Los autores también agradecen el apoyo recibido por el CONACYT a través del proyecto de CB-2015-01-257383 para la realización de esta investigación.

Referencias

1. Coello Guilarte, D.L.: Clasificación translingue para la detección de depresión en usuarios de Twitter. Tesis de Maestría en Ciencias Computacionales, INAOE (2019)
2. Tsugawa, S., Mogi, Y., Kikuchi, Y., Kishino, F., Fujita, K., Itoh, Y., Ohsaki, H.: On estimating depressive tendencies of twitter users utilizing their tweet data. In: 2013 IEEE Virtual Reality (VR), pp. 1–4 (2013)
3. Sandhya, G.: Using Word Embeddings to Explore the Language of Depression on Twitter. Master Thesis in Computer Science, The University of Vermont (2019)
4. Spitzer, R. L., Md, K. K., Williams, J. B. W.: Diagnostic and statistical manual of mental disorders (DSM-5 R). Naklada Slap, Jastrebarsko, Croatia (2013)
5. Pennebaker, J. W., Booth, R. J., Francis, M. E.: Liwc 2007: Linguistic inquiry and word count. Austin, Texas, liwc.net. (2007)

6. Tsugawa, S., Mogi, Y., Kikuchi, Y., Kishino, F., Fujita, K., Itoh, Y., Ohsaki, H.: On estimating depressive tendencies of twitter users utilizing their tweet data. In: 2013 IEEE Virtual Reality (VR), pp. 1–4 (2013)
7. Nadeem, M.: Identifying depression on twitter. (2016)
8. Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M., Bierut, L. J. : A content analysis of depression-related tuits. Computers in human behavior, 54, pp. 351–357 (2016)
9. Stankevich, M., Isakov, V., Devyatkin, D., Smirnov, I.: Feature engineering for depression detection in social media. In: ICPRAM, pp. 426–431 (2018)
10. Losada, D. E., Gamallo, P.: Evaluating and improving lexical resources for detecting signs of depression in text. Language Resources and Evaluation, pp. 1–24 (2018)
11. Wohlgenannt, G., Chernyak, E., Ilvovsky, D.: Extracting social networks from literary text with word embedding tools. In: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), pp. 18–25 (2016)
12. Pennington, J., Socher, R., Manning, C. D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543 (2014)
13. Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao: Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI conference on artificial intelligence (2015)
14. Chunting Zhou, Chonglin Sun, Zhiyuan Liu, Francis Lau: A C-LSTM neural network for text classification. (2015)
15. Mathieu C.: BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Canada (2017)