Compilación de un lexicón de redes sociales para la identificación de perfiles de autor

Helena Gómez-Adorno, Ilia Markov, Grigori Sidorov,
Juan-Pablo Posadas-Duran, Carolina Fócil-Arias

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

helena.adorno@gmail.com, markovilya@yahoo.com, sidorov@cic.ipn.mx,
jpposadas@gmail.com, focil.carolina@gmail.com

Resumen. En este trabajo presentamos un recurso léxico para el preprocesamiento de textos publicados en redes sociales desarrollado para los idiomas: inglés, español, holandés e italiano. El recurso se compone de diccionarios de palabras *slang*, abreviaturas, contracciones y emoticones utilizados comúnmente en redes sociales. Los diccionarios fueron utilizados en el preprocesamiento de *tweets* obtenidos del corpus de la competencia de identificación de perfiles de autor del PAN 2015 y los resultados demuestran que el uso de los diccionarios ayuda a mejorar la eficiencia de los clasificadores para la tarea de identificación de perfiles de autor.

Palabras clave: Lexicón, redes sociales, perfil de autor, clasificación de textos.

Compiling a Lexicon of Social Media for the Author Profiling task

Abstract. In this paper, we present a lexical resource for preprocessing of texts published in social networks. It is developed for the following languages: English, Spanish, Dutch, and Italian. The resource contains dictionaries of slang words, abbreviations, contractions, and emoticons commonly used in social networks. The dictionaries were used for preprocessing of tweets obtained from the corpus for the task of author profiling (PAN 2015). The results show that the use of the proposed dictionaries helps to improve the efficiency of classifiers for the author profiling task.

Keywords: Lexicon, social networks, author profiling, text classification.

1. Introducción

El uso de las redes sociales está en incremento constante a nivel mundial. Cientos de usuarios se inscriben a diario en las diferentes plataformas existentes, por lo tanto, el contenido extraído de las redes sociales es fundamental para tareas como análisis de sentimiento [11], detección de perfiles de autores [13], identificación de autores [9,14], minería de opiniones [4], detección de plagio [17], cálculo de similitud entre textos [18,20] y para desarrollar sistemas robustos que ayuden a la toma de decisiones en áreas relacionadas como la política, la educación, la economía, entre otras.

El procesamiento de los mensajes publicados en redes sociales no es una tarea sencilla de resolver [12,2]. Los mensajes publicados en éstas plataformas son generalmente cortos (cientos de palabras) y no siguen las reglas convencionales del idioma, por ejemplo, para componer los textos se utilizan con frecuencia palabras *slang*, abreviaturas y emoticones [8]. Las palabras consideradas como *slang* y las abreviaturas son específicas para cada idioma, y por lo tanto, los sistemas que realizan procesos sobre mensajes de redes sociales necesitan diccionarios específicos.

El objetivo de este trabajo es compilar diccionarios de palabras *slang*, abreviaturas, contracciones y emoticones para ayudar al preprocesamiento de textos publicados en redes sociales. Con el uso de estos diccionarios se pretende mejorar los resultados de las tareas relacionadas con datos obtenidos de dichas plataformas. Por lo tanto, evaluamos nuestra hipótesis en la tarea de identificación de perfiles de autor *(author profiling)*. El objetivo de esta tarea es obtener información respecto al autor de un texto, específicamente su edad y género, analizando mensajes publicados por el autor en Twitter [13,16].

Este trabajo está dividido de la siguiente forma: La sección 2 presenta los detalles de las investigaciones realizadas en el área Procesamiento de Lenguaje Natural y preprocesamiento de textos. La sección 3 describe el procedimiento para la compilación de los diccionarios y la estructura de los mismos. La sección 4 presenta la evaluación de la tarea de identificación de perfiles de autor utilizando los diccionarios desarrollados. Las conclusiones y el trabajo a futuro son presentados en la sección 5.

2. Trabajo relacionado

En esta sección, se presentan algunos de los principales trabajos que demuestran la importancia de la fase de preprocesamiento de datos en diferentes tareas de procesamiento automático de textos. Un correcto preprocesamiento conlleva a un análisis adecuado y ayuda a incrementar la precisión y la eficiencia de los procesos de análisis de textos. Algunos de los retos encontrados a la hora de realizar preprocesamiento de textos de redes sociales son presentados a detalle en el trabajo de Baldwin [3].

En el estudio desarrollado por Clark y Araki [7] se discuten los problemas relacionados con el procesamiento de mensajes obtenidos a partir de redes

sociales. Los autores mejoraron el rendimiento de un corrector ortográfico de código abierto sobre datos de Twitter, mediante el desarrollo de un sistema de preprocesamiento automático para la normalización de dichos datos. Los resultados reportados indican que el sistema es capaz de disminuir el promedio de error por mensaje de 15 % a 5 %.

El trabajo realizado por Hemalatha *et al.* [11] presenta algunos de los pasos de preprocesamiento de textos que deben ser tomados en cuenta para mejorar la calidad de los mensajes obtenidos a través de Twitter. Entre las técnicas mencionadas se encuentran: remover URLs, caracteres especiales, letras repetidas de una palabra y palabras de preguntas (qué, cuándo, como, etc.). Este estudio demostró que al realizar los pasos mencionados anteriormente, el resultado de la tarea de análisis de sentimiento mejora considerablemente.

En la investigación realizada por Haddi *et al.* [10] se utilizó una combinación de diferentes técnicas de preprocesamiento tales como limpieza de etiquetas HTML, expansión de abreviaturas, manejo de palabras de negación, eliminación de palabras auxiliares (*stop words*) y uso de métodos para reducir una palabra a su raíz. El objetivo de este trabajo es el de analizar los sentimientos sobre opiniones relacionadas con películas. Los autores reportaron que un preprocesamiento de textos apropiado puede mejorar el desempeño del clasificador y aumentar los resultados considerablemente en la tarea de análisis de sentimientos.

En [15] se propone la corrección ortográfica de los mensajes encontrados en redes sociales. Esto incluye letras repetidas, vocales omitidas, sustitución de letras con números (típicamente sílabas), uso de ortografía fonética, uso de abreviaturas y siglas. En un enfoque dirigido por datos *data-driven approach* [5] se aplica un filtro de URL combinándolo con técnicas estándar de preprocesamiento de textos.

Como puede observarse, existen diversas investigaciones relacionadas con el preprocesamiento de textos publicados en redes sociales. En este trabajo, se presenta un recurso léxico y se demuestra su importancia para la tarea de identificación de perfiles de autor. En la siguiente sección se describe el procedimiento utilizado para la compilación de los diccionarios y se muestran ejemplos del contenido de los mismos.

3. Creación del lexicón de redes sociales

Este trabajo de investigación comprende el análisis y la recopilación de vocabulario abreviado (utilizado en redes sociales) para la creación de diccionarios en varios idiomas como el inglés, español, holandés e italiano. Los diccionarios fueron recopilados para estos cuatro idiomas ya que son necesarios para el preprocesamiento de *tweets* para la tarea de identificación de perfil del autor del PAN 2015 [16]. El PAN es un laboratorio de evaluación sobre descubrimiento de plagio, autoría y uso indebido de software social, que se celebra en el marco de la conferencia CLEF[1].

[1] *Conference and Labs of the Evaluation Forum:* `http://www.clef-initiative.eu/`

El tipo de vocabulario acortado que generalmente se utiliza en las redes sociales se pueden dividir en tres categorías: palabras *slang*, abreviaturas y contracciones. A continuación se describe brevemente cada categoría:

Palabras *Slang* vocabulario estructurado en una lengua dada, que generalmente se utiliza entre personas del mismo grupo social. Es un metalenguaje que se usa para enriquecer las expresiones, y las palabras tienen una representación fonológica intacta. Algunos ejemplos de palabras *slang* encontrados en el idioma español son bb (bebé), xq (porque), dnd (dónde), tb (también), tqm (te quiero mucho) y xfa (por favor).

Abreviaciones son representaciones ortográficas de una palabra o frase. También se incluyen en esta categoría los acrónimos, los cuales se forman a partir de las letras iniciales de un nombre o partes de palabras o frases. Dentro de esta categoría podemos encontrar los siguientes ejemplos: Arq. (Arquitecto), Sr. (Señor), NY (Nueva York), kg. (kilogramo), Av. (Avenida), entre otras.

Contracciones ocurren cuando dos palabras se reducen en una sola y un apóstrofe toma el lugar de la letra que falta. Hay muchas reglas entre las lenguas para crear contracciones. Sin embargo, esta investigación no tendrá en cuenta ninguna de ellas. Ejemplos de contracciones son: al (a el) y del (de el).

Otro tipo de elemento que aparece con frecuencia en los mensajes de redes sociales son los emoticones. Los emoticones son visualizaciones tipográficas que permiten representar las expresiones faciales de las emociones, es decir, es una manera de darle una carga emotiva a un texto. Se incluyeron dos estilos de emoticones conocidos como occidental y oriental. El estilo occidental se utiliza comúnmente en los Estados Unidos y Europa, los emoticones de este estilo se escriben de izquierda a derecha, como si una cara se gira 90 grados hacia la derecha. Los emoticones mostrados a continuación pertenecen a este estilo: :-) (cara sonriente), :-/ (cara dudosa) y :-o (cara sorprendida). Por el otro lado, se tienen a los emoticones de tipo oriental que son populares en el este de Asia y a diferencia del estilo occidental, los emoticones orientales no se rotan. En este estilo, los ojos son a menudo vistos como una característica importante de la expresión. Algunos ejemplos de este estilo son (^v^) (cara sonriente), ((+_+)) (cara dudosa) y (o.o) cara sorprendida.

En este trabajo realizamos la recopilación de vocabulario abreviado y emoticones que se utilizan generalmente en redes sociales. A continuación se describe el proceso de compilación de los diccionarios:

1. Búsqueda e identificación de sitios web que se utilizan como fuente para la extracción de las listas de palabras *slang*, abreviaturas y contracciones en los cuatro idiomas (inglés, español, italiano y holandés).
2. Extracción manual o semi-automática de todas las palabras *slang*, abreviaturas y contracciones junto con sus respectivos significados de cada sitio web en los diferentes idiomas.
3. Identificación y fusión de todos los archivos de la misma categoría. Limpieza, formateo y estandarización de cada archivo, eliminando duplicados. Verificación manual de significados de cada entrada de los diccionarios.

Mediante el proceso descrito anteriormente se crearon doce diccionarios, divididos en cuatro idiomas, uno para cada categoría (palabras *slang*, abreviaturas y contracciones). Los diccionarios están disponibles de manera gratuita en nuestro sitio web[2], donde además se presenta una breve descripción de los diccionarios, una lista de sitios web utilizados para la recolección de las tres categorías de vocabulario para los cuatro idiomas, y la lista sitios web usados para obtener los emoticones. En el caso del diccionario de palabras *slang* en español también se incluyeron entradas del trabajo [6], en el que se realizó una extracción manual de palabras *slang* de una colección de mensajes de Twitter.

Cada diccionario se ha almacenado en un archivo diferente, los elementos se encuentran ordenados de manera alfabética y la información se codifica usando dos columnas separadas por una tabulación. La primera columna corresponde a una entrada de palabra *slang*, abreviatura o contracción, según sea la naturaleza del diccionario, y la segunda columna corresponde al significado de la entrada correspondiente.

La Tabla 1 presenta las estadísticas de cada diccionario, donde se puede observar que existe un número significativo de palabras *slang* disponibles para inglés y español, mientras que para el caso del holandés e italiano el número de entradas es menor. Por otro lado, se puede observar que hay un gran número de abreviaturas en el idioma holandés. El número total de entradas en nuestro lexicón de redes sociales es de 7,212.

Tabla 1. Número de entradas en cada diccionario

Tipo de diccionario	Holandés	Italiano	Inglés	Español
Abreviaturas	1,237	107	1,346	527
Slangs	250	362	1,249	939
Contracciones	15	56	131	11
Emoticones	-	-	482	482
Totales	1,520	525	3,208	1,959

4. Caso de estudio: Identificación de perfiles de autor

La tarea de identificación de perfiles de autor consiste en la identificación de algunos aspectos de una persona como su edad, sexo, o algunos rasgos de comportamiento basados en el análisis de muestras de texto. El perfil de un autor puede ser utilizado en muchas áreas, por ejemplo, en las ciencias forenses para obtener la descripción de un sospechoso mediante el análisis de los mensajes publicados en redes sociales, y en las empresas para personalizar los anuncios que aparecen en las redes sociales o enviados por medio de correo electrónico[1].

[2] `http://www.cic.ipn.mx/~sidorov`

En los últimos años, se han propuesto diferentes métodos para abordar la tarea de identificación de perfiles de autor, la mayoría de ellos utilizan técnicas de aprendizaje automático, minería de datos y procesamiento del lenguaje natural. Desde un punto de vista de aprendizaje automático, la tarea identificación de perfiles de autor puede ser considerada como un problema de clasificación multi-clase y multi-etiqueta, donde cada elemento S_i de un conjunto de muestras de texto $\mathbf{S} = \{S_1, S_2, \ldots, S_i\}$ se le asignan múltiples etiquetas $(l_1, l_2, \ldots, l_k)$, cada una de ellas representando un aspecto del autor (género, edad, rasgos de comportamiento) y el valor asignado en cada etiqueta representa una categoría dentro del aspecto correspondiente. El problema se traduce a la construcción de un clasificador M que asigna varias etiquetas a los textos no etiquetados.

El enfoque basado en aprendizaje automático está dividido en dos etapas: entrenamiento y prueba. En la etapa de entrenamiento, se obtiene una representación vectorial de cada uno de los textos de ejemplo de cada categoría, es decir, $v^i = \{v_1, v_2, \ldots, v_j\}$ donde v^i es la representación vectorial del texto de ejemplo S_i.

Luego, un clasificador es entrenado utilizado la representación vectorial de las muestras etiquetadas. En este trabajo utilizamos un clasificador basado en máquinas de soporte vectorial (SVM) y generamos diferentes modelos de clasificación para cada uno de los aspectos del perfil de un autor, es decir, aprendemos un modelo para determinar la edad y otro modelo para determinar el género de un autor.

Las características utilizadas en este trabajo se basan en una representación vectorial de la frecuencia de ocurrencia de palabras usando el modelo estándar de bolsa de palabras (en inglés, *Bag of words (BOW)*), que ha demostrado ser efectivo en tareas relacionadas con la caracterización de autores en trabajos previos [19]. En este artículo se utilizan solamente la frecuencia de palabras que ocurren en el conjunto de textos de entrenamiento para construir el modelo de representación.

En la fase de prueba o evaluación, la representación vectorial de los textos no etiquetados es obtenida utilizando las mismas características extraídas en la etapa de entrenamiento. Luego, se utiliza el clasificador para asignar valores a las etiquetas de cada aspecto del perfil del autor de cada usuario del conjunto de prueba.

Con el objeto de evaluar la utilidad de nuestros diccionarios, utilizamos el corpus diseñado para la tarea identificación de perfiles de autor del PAN 2015. El corpus está compuesto de *tweets* en cuatro idiomas diferentes: inglés, español, italiano y holandés. Cada idioma tiene un conjunto de *tweets* etiquetados que corresponden a la edad y género del autor de dicho *tweet*. Los valores de las etiquetas de la clase género pueden ser: hombre o mujer. Los valores de las etiquetas de la clase edad pueden ser: 18-24, 25-34, 35-49, 50-xx.

El corpus de identificación de perfiles de autor del PAN-2015 está parcialmente disponible. Debido a la política de los organizadores, sólo el corpus de entrenamiento ha sido liberado. En este sentido, se realizaron los experimentos

Tabla 2. Resultados obtenidos para la clasificación de género

Language	**SVM Liblinear**	
	sin preprocesamiento	**con preprocesamiento**
Inglés	74.91	**76.33**
Español	80.00	**81.00**

Tabla 3. Resultados obtenidos para la clasificación de edad

Language	**SVM Liblinear**	
	sin preprocesamiento	**con preprocesamiento**
Inglés	75.14	**76.31**
Español	68.70	**69.11**

utilizando el corpus de entrenamiento y se realizó validación cruzada de 10 capas para evaluar nuestra propuesta.

Las tablas 2 y 3 presentan la exactitud obtenida para las clases género y edad respectivamente, con y sin preprocesamiento del corpus. Podemos concluir que para cada lenguaje, los mejores resultados fueron obtenidos cuando se realiza el preprocesamiento utilizando nuestros diccionarios.

La etapa de preprocesamiento consiste básicamente en la identificación dentro del corpus de palabras que se encuentren en nuestros diccionarios y reemplazarlas por sus respectivos significados. Cabe mencionar que para este trabajo no realizamos ningún proceso de desambiguación del sentido de las palabras y por tanto, solo se selecciona el primer significado disponible para cada término.

5. Conclusiones y trabajo futuro

En este trabajo presentamos un lexicón de redes sociales que contiene diccionarios de palabras *slang*, abreviaturas, contracciones y emoticones más populares en las redes sociales. El recurso contiene diccionarios en idioma inglés, español, holandés, e italiano. Además, describimos la metodología de la recopilación de datos, listamos las direcciones URL utilizadas como fuentes para la creación de cada diccionario, y explicamos el proceso de estandarización de los mismos. Luego, proporcionamos información relativa a la estructura de los diccionarios y una descripción de la longitud de cada uno de ellos.

Al momento de utilizar los diccionarios para preprocesamiento de textos nos dimos cuenta de que hay algunos términos que se usan comúnmente en las redes sociales que no están presentes en nuestras fuentes web, especialmente para los idiomas inglés, italiano y holandés. Por lo tanto, para un trabajo futuro, tenemos la intención de ampliar los diccionarios de palabras *slang* con entradas

recogidas manualmente para cada idioma, de la misma manera que se hizo para el diccionario de palabras *slang* en español.

Agradecimientos. Este trabajo ha sido realizado gracias al apoyo de la "Red Temática en Tecnologías del Lenguaje - CONACYT" y Gobierno Mexicano (Proyecto CONACYT 240844, SNI, COFAA-IPN, SIP-IPN 20151406, 20161947).

Referencias

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM 52(2), 119–123 (2009)
2. Atkinson, J., Figueroa, A., Pérez, C.: A semantically-based lattice approach for assessing patterns in text mining tasks. Computación y Sistemas 17(4), 467–476 (2013)
3. Baldwin, T.: Social media: Friend or foe of natural language processing? In: 26th Pacific Asia Conference on Language, Information and Computation. pp. 58–59 (2012)
4. Ben-Ami, Z., Feldman, R., Rosenfeld, B.: Using Multi-View Learning to Improve Detection of Investor Sentiments on Twitter. Computación y Sistemas 18, 477–490 (2014)
5. Brigadir, I., Greene, D., Cunningham, P.: Adaptive Representations for Tracking Breaking News on Twitter. ArXiv e-prints (2014)
6. Camacho-Vázquez, V., Sidorov, G., Galicia-Haro, S.N.: Machine learning applied to a balanced and emotional corpus of tweets with many varieties of Spanish. submitted (2016)
7. Clark, E., Araki, K.: Text normalization in social media: Progress, problems and applications for a pre-processing system of casual English. Procedia - Social and Behavioral Sciences 27, 2–11 (2011)
8. Das, D., Bandyopadhyay, S.: Document Level Emotion Tagging: Machine Learning and Resource Based Approach. Computación y Sistemas 15, 221–234 (2011)
9. Gómez-Adorno, H., Sidorov, G., Pinto, D., Markov, I.: A graph based authorship identification approach: Notebook for PAN at CLEF 2015. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. (2015)
10. Haddi, E., Liu, X., Shi, Y.: The role of text pre-processing in sentiment analysis. Procedia Computer Science 17, 26–32 (2013), first International Conference on Information Technology and Quantitative Management
11. Hemalatha, I., Varma, D.G.P.S., Govardhan, D.A.: Preprocessing the informal text for efficient sentiment analysis. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) 1(2), 58–61 (2012)
12. Pinto, D., Vilariño-Ayala, D., Alemán, Y., Gómez-Adorno, H., Loya, N., Jiménez-Salazar, H.: The soundex phonetic algorithm revisited for sms-based information retrieval. In: II Spanish Conference on Information Retrieval CERI 2012 (2012)
13. Posadas-Durán, J.P., Gómez-Adorno, H., Markov, I., Sidorov, G., Batyrshin, I.Z., Gelbukh, A.F., Pichardo-Lagunas, O.: Syntactic n-grams as features for the author profiling task: Notebook for PAN at CLEF 2015. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. (2015), `http://ceur-ws.org/Vol-1391/136-CR.pdf`

14. Posadas-Duran, J.P., Sidorov, G., Batyrshin, I.: Complete syntactic n-grams as style markers for authorship attribution. In: Human-Inspired Computing and Its Applications, pp. 9–17. Springer (2014)
15. Rangarajan Sridhar, V.K.: Unsupervised text normalization using distributed representations of words and phrases. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. pp. 8–16. Association for Computational Linguistics, Denver, Colorado (2015), http://www.aclweb.org/anthology/W15-1502
16. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. CLEF (2015)
17. Sanchez-Perez, M.A., Gelbukh, A., Sidorov, G.: Adaptive algorithm for plagiarism detection: The best-performing approach at pan 2014 text alignment competition. In: Mothe, J., Savoy, J., Kamps, J., Pinel-Sauvagnat, K., Jones, J.G., SanJuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction: 6th International Conference of the CLEF Association, CLEF'15, Toulouse, France, September 8-11, 2015, Proceedings, pp. 402–413. Springer International Publishing, Cham (2015)
18. Sidorov, G., Gómez-Adorno, H., Markov, I., Pinto, D., Loya, N.: Computing text similarity using tree edit distance. In: Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC), 2015 Annual Conference of the North American, Redmond, WA, USA. pp. 1–4 (2015)
19. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2009)
20. Vilariño, D., Pinto, D., León, S., Alemán, Y., Gómez-Adorno, H.: Buap: N-gram based feature evaluation for the cross-lingual textual entailment task. Atlanta, Georgia, USA p. 124 (2013)