Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 151-154 recibido 29-11-2016 revisado 12-01-2017 aceptado 09-02-2017

Author Profiling en Social Media: Identificación de Edad, Sexo y Variedad del Lenguaje*

Author Profiling in Social Media: Age, Gender and Language Variety Identification

Francisco Manuel Rangel Pardo
Universitat Politècnica de València
Camino de Vera s/n. 46022 Valencia (Spain)
kico.rangel@gmail.com

Resumen: Tesis doctoral escrita por Francisco Manuel Rangel Pardo en la Universitat Politècnica de València, bajo la dirección del PhD. Paolo Rosso. La tesis fue defendida el 3 de junio de 2016 en la misma universidad, ante el tribunal compuesto por los doctores Núria Bel de la Universitat Pompeu Fabra, Raquel Martínez Unanue de la Universidad Nacional de Educación a Distancia (UNED) y Rafael Berlanga Llavorí de la Universitat Jaume I. La tesis fue calificada con la puntuación de Sobresaliente *Cum Laude*.
Palabras clave: Author profiling, identificación edad, identificación sexo, identificación variedad lenguaje, social media, big data, emograph, representación de baja dimensionalidad

Abstract: PhD thesis written by Francisco Manuel Rangel Pardo at the Universitat Politècnica de València, under the supervision of PhD. Paolo Rosso. The thesis was defended on June 3rd 2016, with the commitee formed by the doctors Núria Bel from Universitat Pompeu Fabra, Raquel Martínez Unanue from Universidad Nacional de Educación a Distancia (UNED) and Rafael Berlanga Llavorí from Universitat Jaume I. The thesis was graded with Excellent *Cum Laude*.
Keywords: Author profiling, age identification, gender identification, language variety identification, social media, big data, emograph, low-dimensionality representation

1 Introducción

La posibilidad de conocer rasgos de una persona a partir únicamente de los textos que escribe se ha convertido en un área de gran interés denominada *author profiling* (LaVanguardia, 17/10/2016). Ser capaz de inferir de un usuario su sexo, edad, idioma nativo o los rasgos de su personalidad, simplemente analizando sus textos, abre todo un abanico de posibilidades desde el punto de vista forense, de la seguridad o del marketing.

Además, la proliferación de los medios sociales, que favorece nuevos modelos de comunicación y relación humana, potencia este abanico de posibilidades hasta cotas nunca antes vistas. La idiosincrasia inherente a estos medios sociales hace de ellos un entorno de comunicación especial, donde la libertad de expresión, la informalidad y la generación espontánea de temáticas y tendencias propician el acercamiento a la realidad diaria de las personas en su uso de la lengua. Sin embargo, esa misma idiosincrasia hace que en muchas ocasiones la aplicación de técnicas lingüísticas de análisis no sea posible, o sea extremadamente costoso.

La mayoría de aproximaciones propuestas por los investigadores para abordar las diversas tareas de *author profiling*, se basan en la frecuencia de uso de determinadas características (e.g. categorías gramaticales, palabras vacías o signos de puntuación), o en modelos que emplean n-gramas. Nuestra hipótesis, especialmente cuando hablamos de medios sociales donde no hay censura y prima la libertad de expresión, es que los usuarios expresan sus emociones de manera diferente dependiendo de ciertos rasgos de su persona. Nuestro objetivo es profundizar en el modo en que los usuarios expresan dichas emociones en el marco de su discurso, no sólo tomando en consideración su frecuencia relativa de

* Este trabajo ha sido parcialmente financiado por Autoritas Consulting SA (http://www.autoritas.net)

ISSN 1135-5948

aparición, sino también su posición con y en relación con el resto de elementos del discurso, y analizar cómo puede esto ayudar a determinar su edad y sexo, independientemente del medio social y del idioma. Para hacerlo, hemos propuesto EmoGraph, una representación basada en grafos, debido a su capacidad para modelar y analizar estructuras complejas como el lenguaje, que debido a la idiosincrasia propia de los medios sociales, hace compleja la aplicación de técnicas elaboradas de análisis sintáctico.

Además, hemos querido investigar si la expresión de emociones permitiría diferenciar entre hablantes de diferentes variedades de una misma lengua, por ejemplo españoles, mexicanos o argentinos, o portugueses y brasileños. Nuestra hipótesis es que la variación entre lenguas se basa más en aspectos léxicos, y así lo hemos corroborado tras comparar EmoGraph con representaciones basadas en patrones, representaciones distribuidas y una representación que toma en consideración el vocabulario completo, pero reduciendo su dimensionalidad a únicamente 6 características por clase y que se erige idónea para su aplicación en entornos *big data* como los medios sociales.

Podemos resumir los objetivos de la tesis en los siguientes:

1. Proponer una representación que permita:
   a) modelar la estructura del discurso y la expresión de las emociones en el mismo, tomando en consideración no sólo su frecuencia de aparición sino su posición con y en relación al resto de elementos del discurso;
   b) verificar la hipótesis de que el modo en que el usuario articula su discurso y expresa en él sus emociones, sirve para determinar su edad y sexo;
   c) comprobar la independencia de la hipótesis con respecto al medio social; y
   d) con respecto al idioma.
2. Investigar si la expresión de las emociones es un rasgo diferenciador entre lenguas similares o variedades de una misma lengua, o si por el contrario, variaciones léxicas aportan más información a la tarea;
   a) comprobar la adecuación de las representaciones distribuidas a la tarea;
   b) proponer una representación que reduzca la dimensionalidad frente a las comunmente utilizadas basadas en n-gramas.
3. Crear los recursos necesarios para investigar las cuestiones planteadas y un marco de evaluación común que permita comparar las propuestas de diferentes investigadores construyendo así un estado del arte homogéneo, comparable y reproducible.

2 Estructura

La tesis se ha organizado en torno a 8 capítulos y 2 apéndices. En ellos se trata de responder a las preguntas motivadas en el apartado anterior. Concretamente, la estructura de capítulos y lo que en ellos se trata, es la siguiente:

1. Introducción. En este capítulo introducimos la oportunidad que brindan los nuevos medios sociales y la necesidad de ser capaces de obtener información sobre las personas que participan en ellos, introduciendo así el concepto de *author profiling*.

2. Identificación de edad y sexo. En este capítulo efectuamos una revisión exhaustiva al estado del arte en identificación de edad y sexo, describiendo las representaciones propuestas, los corpus disponibles, las medidas de evaluación utilizadas y los resultados alcanzados. Además, sobre la base de las teorías psicolingüísticas actuales, realizamos un estudio estadístico relativo al uso de las categorías gramaticales por medio social y por sexo.

3. Author profiling en el PAN. En este capítulo realizamos una descripción detallada de los tres años en los que organizamos la tarea de identificación de edad y sexo en el PAN[1]. La organización del PAN ha propiciado la creación de recursos como corpus etiquetados con edad y sexo en idiomas diferentes al inglés, la definición de un marco común de evaluación y la generación de un estado del arte consistente, reproducible y útil para la comparación (Rangel et al., 2013; Rangel et al., 2014; Rangel et al., 2015).

[1] http://www.pan.webis.de

4. Identificación de emociones en medios sociales. En este capítulo abordamos la tarea de identificación de emociones en medios sociales e investigamos su relación con el *author profiling*. Nuestra hipótesis central es que la expresión de las emociones tiene una fuerte correlación con nuestro sexo y edad, algo que en el estado del arte no ha sido abordado. Comenzamos así con una revisión del estado del arte en procesamiento afectivo, desde la perspectiva de la generación de recursos y desde la perspectiva de la identificación automática de emociones en texto, para posteriormente analizar la utilidad de la expresión de las emociones para la identificación de tendencias, o su relación con la ironía, la edad y el sexo. En este capítulo presentamos nuestra investigación en identificación de edad y sexo a partir del mismo conjunto de características que utilizamos para la identificación de las emociones, sentando las bases de la hipótesis central de esta tesis.

5. EmoGraph: Una aproximación basada en grafos. En este capítulo investigamos con mayor profundidad cómo el modo en que los usuarios expresan las emociones sirve para conocer su edad y su sexo. Para ello, tratamos de modelar cómo los usuarios estructuran su discurso y cómo las emociones se enmarcan en el mismo, utilizando grafos debido a su potencia para representar y analizar estructuras complejas, como en este caso el lenguaje. Tras una revisión del estado del arte en uso de grafos para diversas tareas de procesamiento del lenguaje natural, decidimos aprovechar los grafos para extraer el conocimiento relacional entre las partes del discurso y las emociones, y obtener un esquema de asignación de pesos para el aprendizaje automático de los modelos (Rangel y Rosso, 2016). Para finalizar el capítulo investigamos en la robustez del método ante diversos medios sociales e idiomas.

6. Identificación del lenguaje nativo y de las variedades del lenguaje. En este capítulo se proporciona una visión detallada del estado del arte relativo a dos tareas relacionadas: la identificación del idioma nativo de un usuario que escribe en una segunda lengua, y la discriminación entre variedades de una misma lengua. El objetivo del capítulo es presentar el transfondo de una tarea que tiene la doble vertiente de la clasificación de textos y el *author profiling*, y sobre la que deseamos contrastar una segunda hipótesis: la variación entre lenguas similares o variedades de una misma lengua, se debe más a cambios léxicos que al modo en que sus usuarios expresan las emociones.

7. Aproximaciones para la identificación de variedades del lenguaje. En este capítulo investigamos la adecuación de la representación propuesta para identificar la edad y el sexo a partir del modo en que los usuarios articulan su discurso y expresan las emociones a la tarea de discriminar entre usuarios que hablan variedades de una misma lengua, o lenguas muy similares. Así mismo proponemos representaciones alternativas (Rangel, Rosso, y Franco-Salvador, 2016) que permiten contrastar nuestra hipótesis de que en este caso, el léxico tiene un componente discriminativo mayor.

8. Conclusiones y trabajo futuro. En este capítulo presentamos las conclusiones al trabajo que hemos llevado a cabo en el marco de nuestro doctorado, subrayando los principales descrubrimientos que soportan nuestras hipótesis, las principales contribuciones al estado del arte y marcando las directrices de trabajos futuros que pueden derivarse del mismo.

Apéndice I. Author profiling en PAN 2014. En este apéndice se muestran las tablas de significación estadística en una comparación pareada de los sistemas participantes de la tarea del PAN 2014, además de las distancias entre la edad predecida y la edad real de los autores.

Apéndice II. Author profiling en PAN 2015. En este apéndice se muestran las tablas de significación estadística en una comparación pareada de los sistemas participantes en la tarea del PAN 2015, así como una comparativa de resultados entre idioma.

3 Contribuciones

Se pueden destacar las siguientes contribuciones:

1. Hemos propuesto la representación EmoGraph para modelar el estilo discursivo y la expresión de las emociones en textos, y la hemos aplicado a la identificación de edad y sexo. Además, hemos verificado su aplicabilidad y robustez a diferentes medios sociales e idiomas.

2. Hemos investigado la aplicabilidad de la representación EmoGraph en la tarea de identificación de variedades de una misma lengua, comprobando que la expresión de emociones y el estilo discursivo no varía de modo discriminativo. Para verificar nuestra hipótesis de que las variaciones se producen principalmente a nivel léxico, hemos analizado varias representaciones: una basada en patrones (IG-WP), dos basadas en representaciones distribuidas sobre el conocido modelo de Skip-gramas continuos, y la representación de baja dimensionalidad (LDR) que hemos propuesto y que permite trabajar de manera eficiente en entornos *big data*.

3. Hemos creado los recursos necesarios para llevar a cabo la investigación, concretamente:

   a) Con respecto a la identificación de edad y sexo, hemos colaborado en la organización y creación de un marco de evaluación en la tarea de identificación de edad y sexo del PAN en el CLEF, lo que ha permitido crear un conjunto de corpus recopilados de diferentes medios sociales (Twitter, blogs, revisiones online, redes sociales) y en diferentes idiomas (inglés, español, holandés e italiano), etiquetados con edad y sexo.

   b) Con respecto a la identificación de emociones en medios sociales, y concretamente en Twitter, hemos compilado el corpus Barcenas con tuits tratando un caso de corrupción política ocurrido en España entre el 9 de julio y el 2 de octubre de 2013, con un total de 4.397.023 tuits en español.

   c) Con respecto a la identificación de emociones en medios sociales y su relación con el sexo, así como con el uso de la ironía, hemos generado el corpus EmIroGeFB con comentarios de Facebook anotados con las seis emociones básicas de Ekman, la presencia/ausencia de ironía, y el sexo de los autores de los comentarios. El corpus se enmarca dentro de tres temáticas (política, fútbol, famosos) y consta de 1.200 comentarios en español.

   d) Por último, con respecto a la identificación de la variedad de lenguaje, hemos construido el corpus HispaBlogs con posts escritos en blogs personales en cinco variedades del español: Argentina, Chile, España, México y Panamá. El corpus consta de dos particiones de 2.400 y 1.000 autores por partición.

Bibliografía

LaVanguardia. 17/10/2016. Los textos escritos delatan el sexo. http://www.lavanguardia.com/cultura/20161017/411054860677/linguistica-estudio-determina-sexo-autor-articulo-upf.html.

Rangel, F. y P. Rosso. 2016. On the impact of emotions on author profiling. *Information processing & management*, 52(1):73–92.

Rangel, F., P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, y W. Daelemans. 2014. Overview of the 2nd author profiling task at pan 2014. En L. Cappellato N. Ferro M. Halvey, y W. Kraaij, editores, *CLEF 2014 labs and workshops, notebook papers. CEUR-WS.org*, volumen 1180, páginas 898–927.

Rangel, F., P. Rosso, y M. Franco-Salvador. 2016. A low dimensionality representation for language variety identification. En *17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing*. Springer-Verlag, LNCS.

Rangel, F., P. Rosso, M. Moshe Koppel, E. Stamatatos, y G. Inches. 2013. Overview of the author profiling task at pan 2013. En P. Forner R. Navigli, y D. Tufis, editores, *CLEF 2013 labs and workshops, notebook papers. CEUR-WS.org*, volumen 1179, páginas 352–365.

Rangel, F., P. Rosso, M. Potthast, B. Stein, y W. Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. En L. Cappellato N. Ferro G. Jones, y E. San Juan, editores, *CLEF 2015 labs and workshops, notebook papers. CEUR Workshop Proceedings. CEUR-WS.org*, volumen 1391.