Resumen

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*. 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.

En este trabajo hemos propuesto EmoGraph, una representación basada en grafos con el objetivo de modelar el modo en que los usuarios expresan sus emociones, y el modo en que las articulan en el marco de su discurso, teniendo en consideración no sólo su frecuencia, sino también su posición y relación con y respecto a los elementos del mismo. Nuestra hipótesis de partida es que los usuarios se expresan y expresan sus emociones de manera diferente dependiendo de su edad y sexo, y además, pensamos que esto es así independientemente de su idioma y del medio donde escriban. Hemos colaborado en la creación de un marco común de evaluación en el laboratorio PAN del CLEF, generando recursos que nos han permitido verificar nuestra hipótesis y conseguir resultados comparables y competitivos con los mejores resultados obtenidos por los investigadores del área.

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.