M.I. Jesús Roberto López Santillán,
Dr. Luis Carlos González Gurrola
Facultad de Ingeniería, Universidad Autónoma de Chihuahua,
FINGUACH Año 3, Núm. 10, diciembre-febrero 2016

Determinación de autoría de textos

En la era del *internet* la disponibilidad de información crece de manera geométrica, es decir, cada cierto tiempo la información disponible en línea al menos se duplica. Esto lleva al surgimiento paralelo de situaciones adversas como los crímenes informáticos, los cuales pueden ir desde plagios de trabajos académicos hasta el robo de identidad o ciberterrorismo. A causa de lo anterior, la determinación de la autoría de los textos disponibles digitalmente como tesis académicas, libros electrónicos, entradas en *blogs*, foros, mensajes electrónicos de texto, entre otros, se ha vuelto indispensable. Si bien es un problema sumamente complejo, las técnicas para aumentar la eficacia de las herramientas se mejora constantemente. En el presente artículo trataremos de exponer los diversos métodos históricos y en desarrollo para la determinación del autor de un texto electrónico.

Cuando hablamos de determinación de autoría de textos nos referimos al estudio sistemático de características encontradas en los textos que se pueden atribuir a un autor determinado con cierta probabilidad. Este tipo de análisis nos permite establecer con un alto grado de probabilidad el autor de una entrada de texto digital. El éxito de estas técnicas depende en gran medida del volumen de escritos con que se cuente de diversos autores. Como esto en la mayoría de las ocasiones es difícil de obtener, resulta necesario el desarrollo de nuevas técnicas que permitan incrementar la efectividad con menor cantidad de información disponible; debido a que un autor no necesariamente tiene producción abundante o porque los textos no se encuentran siempre disponibles.

La autoría de textos digitales se puede clasificar básicamente en tres vertientes:

a) La búsqueda sistemática de autor(es) de textos en función de las muestras disponibles.
b) La creación de un perfil del autor con base a las características encontradas en el texto.
c) La detección de similitudes en un mismo texto con el propósito de determinar si fue escrito por una misma persona, muy útil en la determinación de plagios.

En los últimos 15 años las técnicas que han entregado mejores resultados son aquellas que analizan la información desde los puntos de vista léxicos, sintácticos, semánticos y de contenido. Estas técnicas utilizan diversas maneras de representar las características que se buscan en un texto y que eventualmente se pueden atribuir a un determinado autor. El proceso normalmente consiste en recopilar las fuentes de información, extraer las características lingüísticas de ellas, crear un modelo que establezca la probabilidad de acierto y finalmente clasificar los textos y sus autores.

Las formas más populares de representar las características que puedan identificar autores pueden incluir algunas de las siguientes:

a) Características en forma de n-gramas. Los n-gramas son secuencias de elementos que aparecen en un texto, estos pueden tener grado 1, 2, 3...n. Los elementos a secuenciar pueden ser símbolos, palabras, enunciados; de esta manera representamos características de un autor y las podemos buscar en un texto.

b) Estilometría. El "estilo" de cada autor normalmente se puede determinar a través de ciertas características lingüísticas que éste imprime a sus escritos, desde el tipo de palabras seleccionadas hasta la frecuencia con que aparecen en sus textos.
c) Selección de la estructura del texto. Cada autor normalmente imprime un sello general a sus documentos, es decir, la estructura que escoge para transmitir su mensaje se puede considerar como una huella digital de tal forma que con cierto nivel de certidumbre se puede atribuir a un autor específico.

Una vez que se selecciona el estilo de representación de las características se puede optar por alguna de las técnicas de clasificación ya probadas y que normalmente entregan buenos resultados; entre estas podemos mencionar: a) Maquinas de Vectores de Soporte (SVM por sus siglas en inglés), b) Análisis de Componentes Principales (PCA por sus siglas en inglés), c) Estrategias de Aprendizaje Maquina *(Machine Learning)*.

El enfoque de *machine learning* como estrategia de Inteligencia Artificial (IA) cobró gran relevancia en la década de los 90 del siglo pasado, sin embargo no fue sino hasta hace un par de años que resurgió debido a la disponibilidad de tecnología que permite implementar redes neuronales más eficaces, las cuales son parte fundamental de esta vertiente. Las redes neuronales ahora se pueden implementar con más capas interiores lo que las hace más "profundas", dando origen al denominado *"Deep Learning* (DL)". Este tipo de arquitectura de aprendizaje máquina ha entregado muy buenos resultados en las áreas de reconocimiento de imágenes y voz dando lugar a implementaciones prácticas en estas áreas, las cuales podemos utilizar en gran cantidad de productos con los que cotidianamente interactuamos, como consolas de videojuegos, dispositivos de audio/video, dispositivos móviles, entre otros.

Entre las aplicaciones más recientes del DL se encuentra el Procesamiento de Lenguaje Natural (NLP por sus siglas en inglés). Debido a los resultados altamente efectivos entregados por esta arquitectura se esperan iguales resultados en esta área. Si bien es un área de aplicación relativamente nueva para DL es de esperarse que en los próximos años se obtengan clasificadores de autoría de textos basados en DL con una eficacia mayor a los obtenidos actualmente, al menos esta es la "esperanza".

Referencias

Shaukat tamboli, M., & Prasad, R. S. (2013). Authorship Analysis and Identification Techniques: A Review. *International Journal of Computer Applications, 77(16)*, 11-15. doi:10.5120/13566-1375

Tan, R. H., & Tsai, F. S. (2010). Authorship Identification for Online Text. 2010 International Conference on Cyberworlds. doi:10.1109/cw.2010.50

Nirkhi, S., & D. (2013). Comparative study of Authorship Identification Techniques for Cyber Forensics Analysis. *International Journal of Advanced Computer Science and Applications, 4(5)*. doi:10.14569/ijacsa.2013.040505