Atribución de Autoría y Determinación de la Orientación Política en Documentos Periodísticos

Viviana Mercado[1], Andrea Villagra[2], Guillermo Leguizamón[3], Marcelo Errecalde[4]

[1–2]Laboratorio de Tecnologías Emergentes (LabTEm)-Unidad Académica Caleta Olivia
Universidad Nacional de la Patagonia Austral, Ruta N° 3 (Acceso Norte)- Santa Cruz - Argentina.
[3–4]Laboratorio de Investigación y Desarrollo en Inteligencia Computacional -Dpto de Informática
Universidad Nacional de San Luis, Ejército de los Andes 950 - San Luis - Argentina.
{[1]vmercado,[2]avillagra}@uaco.unpa.edu.ar,{[3]legui,[4]merreca}@unsl.edu.ar

Resumen

Este artículo describe, brevemente, las tareas de investigación y desarrollo que se están llevando a cabo en forma conjunta en el área de análisis de autoría de documentos entre el LIDIC de la UNSL y el LabTEm de la UNPA. En particular, se ha tomado como caso de estudio primario los documentos de periodistas con diversas orientaciones políticas (oficialista vs opositor) con el objetivo de realizar con los mismos el Análisis de Autoría y la Determinación/Caracterización del perfil del autor. Ambos tipos de tareas, han ganado creciente interés en la comunidad científica internacional y en empresas dedicadas al análisis de la información en la Web, por lo que la línea de investigación propuesta permitiría la formación de recursos humanos en temáticas relevantes a corto y mediano plazo tanto en el ámbito académico/científico como en la industria.

Palabras clave: Minería de Textos, Análisis de Autoría, Determinación del perfil del Autor, Orientación Política, Artículos Periodísticos.

Contexto

Esta línea de trabajo se enmarca en los trabajos conjuntos que desde hace varios años llevan a cabo investigadores del LabTEm de la UNPA y el LIDIC de la UNSL. En particular, las tareas de investigación desarrolladas tienden a consolidar trabajos previos conjuntos relacionados a la Minería de Textos y la Web [16], y complementarlos con los desarrollos que en el LIDIC se están llevando a cabo en las áreas específicas de análisis de autoría y determinación del perfil del autor [8, 17]. En este contexto, ambos laboratorios no sólo disponen de financión obtenida de proyectos de investigación consolidados, sino que además se mantienen relaciones fluidas de investigación con centros de excelencia mundial especializados en estos temas como el Laboratorio de Tecnologías del Lenguaje del INAOE (Puebla, México) y el Artificial Intelligence Laboratory - DICSE de la University of the Aegean (Karlovassi, Grecia). En particular, una integrante del LabTEm desarrollará su trabajo de Maestría en esta temática, mientras que en el LIDIC un becario de doctorado y uno post-doctoral de CONICET trabajarán en la temática específica de determinación del perfil del autor, y co-

laborarán en aquellos temas que se solapen con la presente investigación.

Introducción

El *análisis de autoría* (AA) [15] es un área de investigación que ha ganado interés creciente en los últimos años principalmente por sus potenciales (y actuales) aplicaciones en problemas de seguridad nacional e inteligencia, lingüística forense, análisis de mercados e identificación de rasgos de personalidad, entre otros. El AA se enfoca en la clasificación automática de textos basándose fundamentalmente en las elecciones estilísticas de los autores de los documentos, e incluye distintas tareas de análisis como por ejemplo la *atribución de autoría*, la *verificación de autor*, la *detección de plagios*, la *determinación del perfil del autor* y la *detección de inconsistencias estilísticas*. Los enfoques predominantes en este área están basados en el aprendizaje automático/de máquina supervisado. En pocas palabras, estos enfoques derivan, a partir de un conjunto de datos etiquetados (conjunto de entrenamiento) y un proceso inductivo de aprendizaje/entrenamiento, un clasificador que puede generalizar sus predicciones a otros datos no observados previamente. La representación clásica de los textos/documentos en estos casos, incluye tanto atributos basados en el contenido (palabras) como en el estilo de escritura de los autores.

A partir de la disponibilidad de volúmenes inmensos de información en la Web, se reconoce cada día más el rol de la AA como una herramienta fundamental para hacer un uso adecuado y ventajoso de esta información, lo que ha quedado plasmado en un incremento de Workshops y Competencias Internacionales específicos de esta temática. En particular, un área que comienza a ganar creciente interés es la *determinación del perfil del autor*, es decir, aquella que identifica patrones compartidos por un grupo de gente y que aborda problemas de clasificación de acuerdo a la edad y género [13, 14, 2], nacionalidad, personalidad [4, 10], orientación política [1, 5, 11], etc.

Más allá de la relevancia y ventajas que pueden tener este tipo de tareas existe, actualmente, un desarrollo limitado en nuestro país de trabajos y grupos de investigación especializados en la problemática del AA. En este contexto, en la presente línea de investigación nos enfocaremos en dos áreas claves de la AA como lo son la determinación del perfil del autor (DPA), y la atribución de autoría (ATA).

Respecto a la DPA, también conocida como *caracterización del autor* y en inglés como *author profiling*, incluye actividades como la determinación automática de la *edad*, *género*, rasgos de personalidad y orientación política, entre otras. En nuestro caso, nos concentraremos en la orientación política (progobierno vs opositor) de documentos periodísticos de acceso público, como libros de investigación periodística, blogs periodísticos, artículos en revistas y diarios on-line, etc.

Respecto a la ATA, analizaremos las particularidades que surgen para la identificación automática de autores, en aquellos contextos en donde los mismos tienen igual o diferente orientación política. En estos casos, se analizará cuales son las "features" (estilográficas o de contenido) que son más relevantes para discriminar los distintos autores que pertenecen al mismo (o diferente) espectro político.

Líneas de Investigación y Desarrollo

Como se expresó en la sección previa, nuestro estudio vinculado al análisis de autoría en documentos periodísticos de contenido político se centrará en dos áreas: 1) *caracterización del autor* y 2) *atribución de autoría*. En este dominio en particular, estas áreas corresponden a la caracterización de periodistas de acuerdo a su orientación política y la atribución del periodista autor de un documento político. Ambas líneas se describen a continuación.

Caracterización de periodistas de acuerdo a su orientación política

Para el desarrollo de esta línea de investigación del proyecto, se comenzará con la construcción de un corpus de documentos de periodistas de reconocida adhesión a las políticas del gobierno y de documentos de periodistas que clara y abiertamente son opositores a dichas políticas. Estos documentos, estarán originados en la información textual que dichos periodistas han hecho disponible en redes sociales, blogs, artículos en periódicos on-line, libros de investigación periodística, etc.

Con los documentos de ambas orientaciones políticas, se construirán perfiles diferenciados que referenciaremos como *oficialistas* vs *opositor*. Para ello, se utilizarán representaciones de documentos que combinen atributos de contenido y estilo con esquemas de pesado que ponderan tanto la ocurrencia como la coocurrencia de estos elementos. Una alternativa a considerar en este caso, es el enfoque de representación de segundo orden con clustering [12] que obtuvo el primer lugar en la tarea de DPA de la competencia internacional PAN 2014 (pan.webis.de).

A partir de estas representaciones se construirá un sistema prototipo que se enfocará en la determinación del perfil ideológico de los periodistas. En este caso, se utilizarán técnicas de evaluación de aceptación generalizada en el área y se analizarán los atributos más relevantes en la caracterización del periodista. Estos resultados serán contrastados con experiencias similares realizadas en la determinación ideológica de usuarios en redes sociales en los Estados Unidos [5]. Asimismo, se hará una primera experiencia "cross-domain" analizándose en qué medida el modelo obtenido con documentos periodísticos "formales" puede ser aplicado en dominios similares pero, en los cuales, el contenido generado por los usuarios contiene alto nivel de ruido, o no corresponden a periodistas.

Atribución del periodista autor de un documento político

Un aspecto bien conocido de la atribución de autoría es su diferencia con la clasificación *temática* (o por *tópico*), en el sentido de los requerimientos que involucra cada una. En la clasificación temática los atributos de contenido (palabras) suelen ser fundamentales, mientras que en la clasificación por autor los atributos que cuantifican el *estilo* adquieren mayor importancia. Esto es comprensible si consideramos que un mismo autor puede hablar de diferentes temas por lo que más importante que capturar las temáticas abordadas es identificar patrones estilográficos de escritura como la longitud de sentencias, riqueza de vocabulario, uso de preposiciones y "palabras de paro" específicas, etc.

En este contexto, esta línea de investigación tiene el desafío de identificar autores de similar orientación política, los cuales seguramente tendrán patrones similares en el uso de determinados términos, por lo que la diferenciación estilográfica tendrá un rol fundamental. Algunas de las alternativas que se probarán en este caso, serán por ejemplo las basadas en n-gramas de caracteres, que han demostrado una efectividad considerable en problemas de atribución de autoría [3, 7]. También se considerará el uso de enfoques *basados en perfiles* [9, 6] con los cuales, investigadores del LIDIC ya han realizado algunas experiencias en tareas de DPA [8].

Resultados y Objetivos

Este trabajo es continuación de trabajos iniciales que se han desarrollado en el LIDIC en temáticas relacionadas a la DPA con textos formales en español [17] y a la determinación de género y edad en blogs en español mediante enfoques basados en perfil [8] y se dan en el marco de las tareas de Minería de Textos y la Web que el LabTEm y el LIDIC han comenzado a realizar en forma conjunta recientemente [16].

En este contexto, se busca cumplir con el objetivo general de formar profesionales

e investigadores especializados en la explotación de información disponible en la Web, un área que, como marca la tendencia en publicaciones científicas a nivel mundial y en el desarrollo de competencias internacionales como *PAN*[1] o *kaggle*[2] atraen un número cada vez mayor de investigadores y personal de la industria. Es importante remarcar que, considerando el poco desarrollo que han tenido en nuestro país las temáticas vinculada a la Minería de Textos y la Web, mediante estos trabajos iniciales y el desarrollo de profesionales idóneos en distintos Laboratorios y Centros de Investigación del país, se favorece a la formación de una masa crítica de recursos humanos con un impacto al corto y mediano plazo no sólo en la academia sino también en la industria.

Como objetivo específico además, se plantea el desarrollo de sistemas de DPA y ATA con capacidades comparables a las obtenidas por otros sistemas a nivel mundial que se están desarrollando en este momento. En este sentido, es interesante remarcar los trabajos conjuntos que en el LIDIC se están realizando en este momento con personal del INAOE de México, quienes han obtenido los mejores resultados del PAN en el área de DPA de las ediciones correspondientes a los años 2013 y 2014. Consideramos que como fruto de esa interacción, se lograrán importantes avances y experiencia en el área de DPA que podrá ser luego transferida en otros trabajos conjuntos con grupos de nuestro país como es el caso del LabTEm.

Formación de Recursos Humanos

La responsabilidad principal de este trabajo estará a cargo de una tesista de la "Maestría en Informática y Sistemas" de la Universidad Nacional de la Patagonia Austral que se desempeña en el LabTEm y que contará con el soporte de investigadores del LIDIC trabajando en temáticas afines. En particular, se prevé que la tesista tenga una activa interacción con un Doctorando del LIDIC trabajando en temáticas de DPA, un becario de Doctorado de CONICET trabajando en DPA con información multimodal y una becaria de post-Doctorado del CONICET enfocada en aspectos psicológicos relacionados a las tareas del DPA.

[1] pan.webis.de

[2] www.kaggle.com

Referencias

[1] R. Abooraig, A. Alwajeeh, M. Al-Ayyoub, and I. Hmeidi. On the automatic categorization of arabic articles based on their political orientation. In *Proc. of the Third International Conference on Informatics Engineering and Information Science (ICIEIS2014)*, 2014.

[2] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52:119–123, 2009.

[3] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[4] F. Celli, B. Lepri, J.-I. Biel, D. Gatica-Perez, G. Riccardi, and F. Pianesi. The workshop on computational personality recognition 2014. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 1245–1246, New York, NY, USA, 2014. ACM.

[5] M. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of twitter users. In *Proceedings of 3rd IEEE Conference on Social Computing (SocialCom)*, 2011.

[6] H. J. Escalante, M. M. y Gómez, and T. Solorio. A weighted profile intersection measure for profile-based authorship attribution. In *Proceedings of MICAI 2011*, volume 7094, pages 232–243, 2011.

[7] G. Frantzeskou, E. Stamatatos, S. Gritzalis, and S. Katsikas. Source code author identification based on n-gram author profiles. In *Artificial Intelligence Applications and Innovations*, volume 204 of *IFIP*, pages 508–515. Springer US, 2006.

[8] D. G. Funez, L. Cagnina, and M. L. Errecalde. Determinación de género y edad en blogs en español mediante enfoques basados en perfil. In *Anales del XIX Congreso Argentino de Ciencias de la Computación (CACIC 2013)*, pages 1003–1012, 2013.

[9] R. Layton, P. Watters, and R. Dazeley. Recentred local profiles for authorship attribution. *Natural Language Engineering*, 18:293–312, 2012.

[10] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500, 2007.

[11] R. Malouf and T. Mullen. Graph-based user classification for informal online political discourse, 2007.

[12] A. Pastor López-Monroy, M. Montes-y Gómez, H. J. Escalante, and L. Villaseñor-Pineda. Using intra-profile information for author profiling. In *Proceedings of the 11th evaluation lab on uncovering plagiarism, authorship, and social software misuse, PAN 2014*, 2014.

[13] C. Peersman, W. Daelemans, and L. Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pages 37–44, New York, NY, USA, 2011. ACM.

[14] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205, 2006.

[15] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society For Information Science and Technology*, 60(3):538–556, 2009.

[16] D. Taquias, A. Villagra, and M. L. Errecalde. Detección de plagios con adversarios. In *Anales del XVI Workshop de Investigadores en Ciencias de la Computación (WICC 2014)*, pages 233–237, 2014.

[17] M. P. Villegas, M. J. Garciarena Ucelay, M. L. Errecalde, and L. Cagnina. A spanish text corpus for the author profiling task. In *Anales del XX Congreso Argentino de Ciencias de la Computación (CACIC 2014)*, pages 621–630, 2014.