Aprendizaje automático y análisis de autoría. Investigaciones con textos periodísticos argentinos

Mercado V[1]., Villagra A.[1], Errecalde M.[1,2]

[1]Laboratorio de Tecnologías Emergentes (LabTEm)
Instituto de Tecnología Aplicada (ITA) - Unidad Académica Caleta Olivia
Universidad Nacional de la Patagonia Austral

[2]Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC)
Departamento de Informática - Universidad Nacional de San Luis

{vmercado, avillagra}@uaco.unpa.edu.ar, merreca@unsl.edu.ar

RESUMEN

El *análisis de autoría* (AA) 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: a) la *atribución de autoría*, b) la *verificación de autor*, c) la *detección de plagios*, d) la *determinación del perfil del autor* y e) la *detección de inconsistencias estilísticas*.

En este contexto, la línea de trabajo correspondiente a este trabajo de tesis de postgrado se propone el abordaje de dos tareas de AA, como lo son la *identificación del sesgo político* y *la atribución de autoría* en documentos periodísticos argentinos. Estas tareas de AA, serán complementadas con enfoques no-supervisados de aprendizaje automático como son el *agrupamiento de textos* (clustering) y la *modelización de tópicos*.

Palabras clave: Análisis Automático de Textos. Análisis de autoría. Determinación del perfil del autor. Identificación del sesgo político. Atribución de autoría. Análisis automático de documentos periodísticos.

CONTEXTO

Esta línea de trabajo se lleva a cabo en el Laboratorio de Tecnologías Emergentes (LabTEm), Instituto de Tecnología Aplicada (ITA) de la Unidad Académica Caleta Olivia de la Universidad Nacional de la Patagonia Austral, en el marco del Proyecto de Investigación 29/225 "Soluciones inteligentes para el desarrollo urbano sostenible" y particularmente en el marco del trabajo de tesis en la Maestría de Informática y Sistemas (MIS) de la Ing. Viviana Mercado.

1. INTRODUCCION

A partir de la disponibilidad de volúmenes inmensos de información en la Web, se reconoce cada día más el rol del *Aprendizaje Automático* y la *Extracción de Conocimiento a partir de los Datos* como herramientas fundamentales para hacer un uso adecuado y ventajoso de esta información. Esta tendencia crece día a día y se plantean nuevos escenarios relevantes como es el caso de *Big Data*, donde el contexto donde deben ser aplicados estos métodos es sumamente desafiante. En particular, un escenario que recibe creciente atención es el del *Análisis (automático) de Textos* (AAT, text analytics) o *aprendizaje automático con textos* (machine learning from text), lo cual resulta razonable si consideramos que gran parte de la información disponible en repositorios digitales, archivos de empresas y

organizaciones y la Web en general es de tipo textual.

El AAT surge de la interacción de 3 áreas principales: el Procesamiento del Lenguaje Natural, la Recuperación de la Información y el Aprendizaje Automático. El AAT incluye diversas tareas como el análisis de sentimiento/minería de opiniones, la extracción de información, la generación de resúmenes y el *análisis de autoría* (AA). El AA a su vez, comprende otras tareas más específicas como la *determinación del perfil del autor* (DPA) y la *atribución de autoría* (ATA). DPA es el área que identifica patrones compartidos por un grupo de gente y que aborda problemas de clasificación de acuerdo a la edad y género [Peersman et al., 2011, Schler et al., 2006, Argamon et al., 2009], nacionalidad, personalidad [Celli et al., 2014, Mairesse et al., 2007], orientación política [Abooraig et al.,2014, Conover et al.,2011, Malouf & Mullen, 2007], tendencias *suicidas* o *depresivas*, *rasgos de personalidad*, *perfiles de consumo* y *adicciones*, de *acosadores sexuales*, etc. La DPA, es un tema muy importante de investigación 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. La ATA, por su parte, consiste en la atribución de un texto de autoría desconocida a uno de un conjunto de autores potenciales. Desde el punto de vista científico, esta tarea plantea varios desafíos interesantes ya que, a diferencia del análisis basado en los temas o tópicos de los textos, aquí se debe capturar información representativa de los *estilos de escritura* de los autores.

En este contexto, y como parte de un proyecto de tesis de postgrado, se ha propuesto el abordaje de dos tareas de AA, una de ATA y otra de DPA como lo son *la atribución de autoría* y la *identificación del sesgo político* en *documentos periodísticos argentinos*. Estas tareas de AA, serán complementadas con enfoques no-supervisados de aprendizaje automático como son el *agrupamiento* de textos (clustering) y la *modelización de tópicos*.

El resto de este artículo describe estas líneas de trabajo en la Sección 2 y los resultados esperados/obtenidos en la Sección 3.

2. LINEAS DE INVESTIGACION y DESARROLLO

En esta sección se describen las 3 líneas de investigación que se llevan a cabo como parte de esta tesis de postgrado, y que identificaremos de ahora en más como 1) *atribución de autoría en documentos periodísticos*, 2) *identificación del sesgo político en documentos periodísticos* y 3) *Análisis no supervisado de documentos periodísticos*.

2.1. *Atribución de autoría en documentos periodísticos*

La *atribución de autoría (ATA)* es una de las tareas principales dentro del *análisis de autoría* (AA) [Stamatatos, 2009]. 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: a) la *atribución de autoría*, b) la *verificación de autor*, c) la *detección de plagios*, d) la *determinación del perfil del autor* y e) la *detección de inconsistencias estilísticas*. Los enfoques predominantes en esta 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.

Para el caso particular de llevar a cabo una tarea de ATA con documentos

periodísticos, el primer paso es la construcción de un "corpus" (colección de documentos) con los documentos de distintos periodistas. En nuestro caso, el foco fue en la recopilación de documentos de periodistas de reconocida adhesión a las políticas del gobierno Nacional en Argentina, en el período finalizado el 10 de Diciembre de 2015 y de documentos de periodistas que clara y abiertamente eran opositores a dichas políticas. Estos documentos, están 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.

Así, la primera tarea fue identificar fuentes de información donde obtener textos periodísticos de Argentina, con una clara orientación política (oficialista vs opositor), como así también de los autores de los mismos. Posteriormente se procedió a recopilar toda esta información, compatibilizando los distintos formatos de los documentos (html, pdf, etc.) y generando un corpus con los textos "planos" de los mismos.

Para la ATA en documentos periodísticos 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 cuáles son las "features" (estilográficas o de contenido) que son más relevantes para discriminar los distintos autores que pertenecen al mismo (o diferente) espectro político.

Como hipótesis de trabajo, se plantea que cuando la atribución de autoría se realice sobre la totalidad de los periodistas, tanto las características de contenido (palabras) como estilográficas serán relevantes. Sin embargo, cuando la atribución se restrinja a periodistas de la misma orientación política, el uso común de ciertas palabras y expresiones hará que los aspectos estilográficos tomen mayor relevancia para la identificación de los mismos.

2.2. *Identificación del sesgo político en documentos periodísticos*

La *identificación del sesgo político* (ISP) en un texto o documento es una forma de *determinación del perfil del autor* (DPA), otra de las tareas principales del análisis de autoría (AA). La ISP, al igual que otras tareas de DPA como la detección de personas depresivas o con diversos rasgos de personalidad, pedófilos y suicidas es una tarea desafiante dentro del análisis automático de textos ya que involucra, en general, el uso de representaciones de los textos que capturen aspectos de contenido y estilográficos de sus autores.

En este contexto, un área particular dentro de la ISP es la que se orienta al estudio de la orientación política en textos escritos por periodistas, y que referiremos de ahora en más como textos periodísticos (TP). Como ya se explicó previamente, consideraremos como textos periodísticos a aquella información que un periodista publica en diversos medios como puede ser un blog personal, un artículo escrito en un medio masivo como un diario o bien el contenido expresado en un libro de su autoría.

La ISP ha sido realizada con textos generados por usuarios comunes de medios sociales como Twitter [Cohen & Ruths, 2013], aunque más recientemente se ha realizado con los documentos producidos por periodistas [Lazaridou & Krestel, 2016]. Sin embargo, estos textos han sido escritos mayoritariamente en inglés o en otros idiomas, no existiendo, de acuerdo a nuestro conocimiento, estudios sobre ISP de documentos periodísticos en español.

En este trabajo realizaremos una primera aproximación a la ISP en textos periodísticos en español, en particular, de textos generados por periodistas argentinos como ya fue explicado en la sección previa. La tarea en este caso, será agrupar todos los documentos de periodistas "oficialistas" por un lado y "opositores" por el otro y convertir el problema de clasificación en uno de clasificación binaria ("oficialista" vs "opositores").

Nuestro objetivo aquí será responder las siguientes preguntas de investigación:

1) ¿Cuáles son las formas de representación de documentos más apropiadas para esta tarea?

2) ¿Cuáles son los algoritmos de aprendizaje más efectivos para usar con esas representaciones?

3) ¿Cuál es el impacto de los enfoques de reducción de dimensionalidad en las representaciones de los documentos?

4) ¿Cuánto se asemejan los resultados obtenidos con estudios similares con textos periodísticos escritos en otros idiomas?

2.2. *Análisis no supervisado de documentos periodísticos*

Las dos tareas de AA planteadas previamente son básicamente tareas de aprendizaje automático *supervisado,* es decir se basan en la información de *las clases* (periodistas en un caso, "oficialista" u "opositor" en el otro) que los documentos tienen asociadas. Sin embargo, existen otras formas de análisis automático de textos no supervisadas que suelen obtener información complementaria valiosa sobre las colecciones de documentos. Estos métodos, que incluyen el agrupamiento (*clustering*) no supervisado de documentos, la reducción no supervisada de dimensionalidad (por ejemplo, mediante *análisis de componentes principales*) o la *modelización de tópicos* se centran en analizar las propiedades de las "features" utilizadas para representar los documentos, sin hacer uso de la información de la clase asignada al documento.

El agrupamiento (clustering) de documentos, por ejemplo, busca generar *grupos* de documentos *relacionados*. La idea es que documentos dentro del mismo grupo sean similares (cercanos) entre sí (alta cohesión) y sean diferentes (alta separación) a los documentos de otros grupos.

Las técnicas de *reducción de dimensionalidad* como el *análisis de componentes principales* (PCA por sus siglas en inglés) permite transformar colecciones de datos de alta dimensionalidad a dos o tres dimensiones facilitando de esa manera su visualización.

Las técnicas de *modelización de tópicos* por su parte, como la *Latent Dirichlet Allocation* (LDA), intenta encontrar grupos de palabras (los *tópicos*) que aparecen en conjunto frecuentemente. Así, la LDA requiere que cada documento pueda ser entendido como una "mixtura" de un subconjunto de los tópicos.

En el contexto del análisis de documentos periodísticos, las tareas no supervisadas en este caso serán diversas:

1) En el contexto del agrupamiento, el objetivo será identificar cuáles son las representaciones y algoritmos de clustering que detectan en forma más efectiva los grupos oficialistas/opositores en un caso y los distintos periodistas en el otro.

2) Los métodos de reducción de dimensionalidad permitirán visualizar cuales son las clases que, a priori, lucen como más sencillas o complicadas para abordar por los enfoques supervisados. También se analizará el impacto de estas técnicas en la efectividad de los distintos clasificadores, considerando distintas dimensiones para los datos a analizar.

3) La modelización de tópicos por otra parte, permitirá identificar cuáles son los tópicos comunes a ambos tipos de periodistas (oficialistas y opositores) y cuales son propios de cada categoría.

3. RESULTADOS OBTENIDOS/ ESPERADOS

Como resultado de esta tesis se espera lograr un *sistema integrado* de *atribución de autoría* con periodistas de la Argentina y de *determinación de la orientación política en documentos periodísticos*, que también soporte otras tareas de aprendizaje no supervisado como el *descubrimiento de tópicos* y el *agrupamiento* de documentos similares

Al tiempo presente, ya se ha generado un corpus de 200 documentos de 10 periodistas (5 oficialistas y 5 opositores) representándose los documentos con un enfoque de *bolsa de términos (bag of terms)*

variándose en cada caso el tipo de término utilizado (*n*-gramas de palabras, *n*-gramas de caracteres, términos "lematizados", términos "truncados", etc.). En cada caso se han utilizado distintos enfoques para el pesado de los términos como la ponderación *binaría*, ponderación *tf-idf*, etc. También se han representado los documentos utilizando todas las características del sistema *Language Inquiry Word Count* (LIWC). Los resultados preliminares en las tareas de atribución de autoría e identificación del sesgo político son muy prometedores y comparables a los obtenidos en tareas similares en otros idiomas como el inglés. En las tareas de análisis no supervisado ya se han obtenido 100 tópicos representativos de las temáticas abordadas por los distintos tipos de periodistas. En este momento se están comenzando los estudios con las restantes técnicas no supervisadas. Es importante observar, que si bien nuestro estudio está actualmente restringido al uso de técnicas "clásicas" en la representación de documentos, se prevé en un futuro considerar nuevos enfoques de *aprendizaje de representaciones*, que incluyan el aprendizaje de "*embeddings*" tales como **Word2vec**, **GloVe**, **fastText**, **StarSpace**, **ULM-FiT**, **ELMo** y **BERT**.

4. FORMACION DE RECURSOS HUMANOS

En esta línea de trabajo actualmente una integrante está desarrollando su tesis de Maestría correspondiente a la Maestría en Informática y Sistemas de la UNPA.

5. BIBLIOGRAFIA

Abooraig R., Alwajeeh A., Al- Ayyoub M., and Hmeidi I. **(2014)**. 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)*.

Argamon S., Koppel M., Pennebaker J. W., and Schler J. **(2009)** *Automatically profiling the author of an anonymous text*. Communications of the ACM, 52:119123.

Celli F., Lepri B., Biel J.-I., Gatica- Perez D., Riccardi G., and Pianesi F.**(2014)**. *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. ACM.

Cohen R., and Ruths D. **(2013)** Classifying Political Orientation on Twitter: It's Not Easy!. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.

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

Lazaridou K., and Krestel R. **(2016)** Identifying Political Bias in News Articles. TCDL Bulletin, vol. 12, 2016.

Mairesse F., Walker M. A., Mehl M. R. and Moore R. K. (**2007**). *Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text*. In JAIR, 30, 457-500.

Malouf R. and Mullen T.**(2007)** *Graph-based user classification for informal online political discourse*.

Peersman C., Daelemans, W. and Van Vaerenbergh L. **(2011)**. *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. ACM.

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

Stamatatos E. (**2009**). *A Survey of Modern Authorship Atrribution Methods*. Journal of the American Society for Information science and technology. 60(3). 538-556.