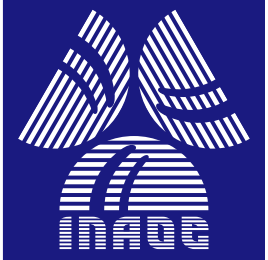

Determinación del perfil de autores en redes sociales con información multimodal

Miguel Ángel Álvarez Carmona, Luis Villaseñor Pineda

Laboratorio de Tecnologías del Lenguaje,
Coordinación de Ciencias Computacionales,
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), México

Reporte Técnico No. CCC-16-007
Julio de 2016

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Índice

Abstract

Hoy en día, las redes sociales han pasado a formar parte de nuestra rutina diaria. Éstas se han convertido en un medio interactivo y masivo de comunicación permitiendo el intercambio de información entre personas con distintos rasgos demográficos como ubicación geográfica, género, edad, nivel socio-económico, etc. En los últimos años los usuarios de las redes sociales han generado una cantidad muy importante de textos, no obstante, la gran mayoría de las veces solo se conoce el nombre de los autores de estos textos (no necesariamente el nombre real) y algunos datos extras que no nos dicen nada acerca de sus rasgos demográficos. Existen diversas aplicaciones donde es importante conocer datos relevantes de usuarios en redes sociales, por ejemplo: mercadotecnia, interacción humano computadora, ciencia forense, entre otras. Por este motivo ha surgido la necesidad de determinar los rasgos demográficos de los usuarios a través de los contenidos en sus cuentas de redes sociales. A esta tarea se le conoce como determinación del perfil de autores. Típicamente este problema se ha enfrentado diseñando la representación de la información textual de los usuarios, aunque trabajos recientes han empezado a utilizar diferentes modalidades de información como la red de contactos del usuario, su actividad en la plataforma o la información visual que comparte. En este trabajo doctoral atacaremos la tarea de determinación de perfiles de autores en Twitter representando a los usuarios con información multimodal. Para esto nos plantemos utilizar dos tipos de modalidades: i) información textual e ii) información de las imágenes compartidas por el usuario.

***Palabras clave**: detección de perfiles de autores, información multimodal, clasificación textual, análisis de imágenes, mezcla de espacios de atributos*

1. Introducción

Internet se ha consolidado como un medio interactivo y masivo de comunicación permitiendo el intercambio de información entre personas de distinta área geográfica, género, edad, nivel socio-económico, etc. Recientemente, este medio de comunicación ha ganado una importante

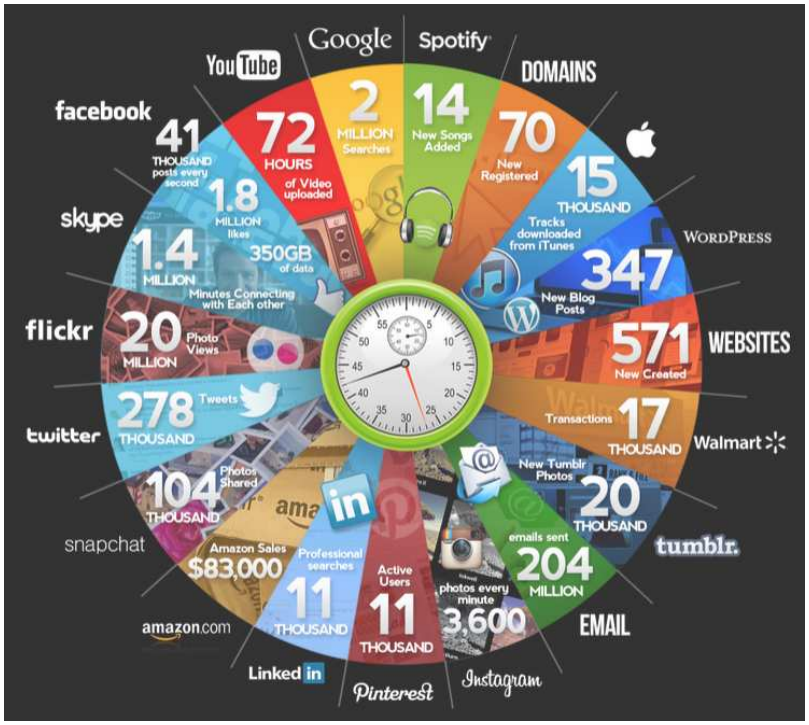

Figura 1. Lo que ocurre cada 60 segundos en la web según el sitio Qmee

popularidad gracias a algunos servicios que invitan a compartir fácilmente información como son: redes sociales, mensajería, *chats*, *blogs*, entre otros.

Este impacto se ha hecho evidente en los últimos años. En la figura 1 se muestra un gráfico donde, según el sitio Qmee[1], se muestra la información que se transmite en la web cada minuto. Según estos datos, mientras usted esta leyendo este párrafo, se genera más de 350 GB de datos en *facebook*[2], se escriben más de 278 mil *tweets*[3], hay más de 11 mil usuarios subiendo fotos en *pinterest*[4] y hay más de 347 nuevos post en *WordPress*[5].

Estas cifras demuestran que por minuto existe un número significativo de nuevos textos e imágenes compartidos por autores de los que, la gran mayoría de las veces, solo se conoce el nombre (no necesariamente el nombre real) y algunos datos extras que no nos dicen nada acerca

[1]http://blog.qmee.com/qmee-online-in-60-seconds/ visitado el 6 de julio de 2016
[2]https://www.facebook.com/
[3]https://twitter.com/
[4]https://pinterest.com/
[5]https://wordpress.com/

de la persona en sí.

Existen diversas razones por las que es importante conocer algunos datos relevantes de los autores de textos en redes sociales a pesar de que se mantengan anónimos. Por ejemplo, desde el punto de vista de mercadotecnia, existe el interés por conocer la identidad y los rasgos demográficos de los diversos usuarios de la red con la intención de dirigir la publicidad que se muestra en las diferentes plataformas en la web y que de esta manera la publicidad se distribuya y se aproveche de mejor manera (Bentolila *et al.*, 2015).

En el área de interacción humano computadora es importante conocer ciertas características de las personas para poder mostrar una interfaz acorde con las características y personalidad de cada individuo (De Andrés *et al.*, 2015).

También, aunado a ese impacto popular y sobre todo, a la facilidad de intercambiar información ocultando el perfil de las personas, la *web* ha sido usada para realizar actos ilícitos o engañosos como por ejemplo acoso sexual y extorsiones (Hall & Hall, 2007; Escalante *et al.*, 2015). En un esfuerzo por detectar y/o prevenir este tipo de actos ilícitos, la disciplina conocida como lingüística forense hace uso del conocimiento lingüístico para estudiar textos que evidencien este tipo de mal comportamiento.

Ya sea para dirigir publicidad, para mejorar una interfaz a partir de las características que definen a un usuario o para prevenir delitos provocados por engaños y extorsiones en la web, ha surgido la necesidad de determinar el perfil de los usuarios en redes sociales. Dado que, realizar manualmente un análisis de este tipo sobre las redes sociales es impensable, surge la necesidad de realizar este análisis de forma automática utilizando tecnologías computacionales.

En procesamiento de lenguaje natural, la tarea encargada de estudiar aspectos relacionados con el autor de un texto se le conoce como análisis de autoría (Indurkhya & Damerau, 2010). El análisis de autoría es el proceso de examinar las características de un texto con la intención de obtener conclusiones de su autor (El & Kassou, 2014). Diversos estudios dividen a la tarea de análisis de autoría en dos principales áreas (Zheng *et al.*, 2003; Abbasi & Chen, 2005; Zheng *et al.*, 2006):

1. Atribución de autoría.

2. Determinación del perfil de autores.

La atribución de autoría consiste en determinar la probabilidad de que un texto pertenezca a un autor dado (Stamatatos, 2009). Por otro lado, la determinación del perfil de autores consiste en extraer la mayor cantidad de información posible del autor a través de lo que escribe (Argamon *et al.*, 2009). La hipótesis detrás de esta área es que la forma en la que escribimos delata nuestra conducta y personalidad. Es en esta área donde se centra esta propuesta.

Generalmente, la tarea de la determinación del perfil de autores (DPA) consiste en extraer aspectos demográficos de una persona a partir de sus textos. Ejemplos de estos aspectos pueden ser: género, edad, nivel socio-económico, lugar de origen o lengua materna[6](Corney *et al.*, 2002; Koppel *et al.*, 2005; Schler *et al.*, 2006). También se han hecho esfuerzos por determinar otros aspectos como el nivel de bienestar (Schwartz *et al.*, 2013b), rasgos de personalidad tales como extraversión o neuroticismo (Argamon *et al.*, 2005; Mairesse & Walker, 2006) así como ideología política (Koppel *et al.*, 2009), afinidad por algunos productos (Argamon *et al.*, 2005), entre otros.

En los inicios de la tarea de DPA se analizaban textos formales como libros, periódicos o revistas para determinar los rasgos de sus autores (Argamon *et al.*, 2003). Sin embargo, determinar el perfil de una persona a través de sus cuentas en redes sociales es una tarea que ha tomado mucha fuerza en los últimos años (Rangel *et al.*, 2015; Stamatatos *et al.*, 2015).

A partir de la disponibilidad de volúmenes inmensos de información en la web, se reconoce cada día más el rol de la tarea de DPA como una herramienta fundamental para hacer un uso adecuado y ventajoso de esta información, lo que incluso ha llevado al incremento de *Workshops* y Competencias Internacionales específicas de esta tarea tales como el PAN[7], RepLab[8], el *workshop Authorship Attribution*[9], *Forensic Authorship Identification* [10], etc. (Rangel *et al.*,

[6]Si el texto está escrito en un idioma diferente a la lengua materna
[7]http://pan.webis.de/
[8]http://nlp.uned.es/replab2014/
[9]https://www.brooklaw.edu/intellectuallife/centerforlawlanguageandcognition/authorattributionsite
[10]http://grantome.com/grant/NSF/SES-1160828

2015; Stamatatos *et al.*, 2015).

Tradicionalmente, existen dos tipos de enfoques que han demostrado ser eficientes para atacar el problema de la DPA en redes sociales: los enfoques basados en estilo y los enfoques basados en contenido (Argamon *et al.*, 2003). Los enfoques basados en estilo se refieren al hecho de analizar cómo el autor se expresa al escribir, por otro lado, en los enfoques basados en contenido se analiza la temática del texto. El principal aporte de diversos trabajos con estos enfoques se basan en la selección de atributos que pueden medir el estilo y el contenido del autor (Schler *et al.*, 2006; Mairesse & Walker, 2006; Rangel *et al.*, 2015).

Más allá de la relevancia y ventajas que pueden tener estos enfoques en este tipo de tareas, también comienzan a identificarse ciertos problemas y aspectos desafiantes que requieren enfoques y técnicas más elaborados que los que se han usado hasta el momento.

Entre estos enfoques más avanzados, se puede mencionar aquellos que incorporan información de otra modalidad disponible más allá de la que se puede derivar del estilo y el contenido del texto del documento. Esta información puede ser de tipo visual, como por ejemplo la provista por los usuarios de redes sociales en las imágenes que comparten, información sobre sus redes de contactos, comportamiento de interacción en las redes sociales, etc. A esta información que utiliza distintas modalidades se le conoce como información multimodal. En este contexto, se puede observar que la mayoría de los trabajos recientes en DPA en el ámbito de las redes sociales se han enfocado principalmente en la definición de atributos temáticos y estilo-métricos apropiados para esta tarea; sin embargo hay muy pocos avances hacia la definición de representaciones multimodales que, por ejemplo, integren diversos tipos de información o que por la naturaleza de las redes sociales se incorporen también información de las imágenes compartidas por los usuarios o de su entorno social.

Este trabajo se enmarca en la tarea de determinación de perfiles de autores en redes sociales con información multimodal. En las siguientes páginas se detallará el resto de esta propuesta doctoral. El documento se encuentra organizado de la siguiente forma: en la Sección 2 se revisa el trabajo relacionado con esta investigación. Posteriormente en la Sección 3 se describirá la problemática. La Sección 4 presenta las preguntas de investigación, los objetivos y las princi-

pales contribuciones de esta propuesta. La Sección 5 muestra la metodología a seguir en esta investigación. Luego, en la Sección 6 mostramos de forma general el plan de trabajo para los siguientes tres años. En la Sección 7 delineamos el trabajo realizado y los resultados alcanzados hasta el momento. Finalmente, en la Sección 8 se mencionan las conclusiones generales de esta propuesta doctoral.

2. Trabajo relacionado

La determinación del perfil de autores (DPA) es una tarea del área del análisis de autoría que difiere de otras tareas como la atribución y verificación del autor donde se examina el estilo individual de los autores (Indurkhya & Damerau, 2010). En la DPA se distingue entre clases de autores, es decir, se identifican las características y patrones que son compartidos por un grupo de personas, como su edad, género, lenguaje nativo, nivel de educación, origen geográfico, ocupación, tipo de personalidad entre otros (Argamon *et al.*, 2003; Koppel & Schler, 2004; López-Monroy *et al.*, 2015).

La DPA es un área de creciente interés por su aplicabilidad en la ciencia forense, la seguridad, el *marketing* y en investigación psicológica y sociológica. Por ejemplo, la DPA puede asistir al *marketing* inteligente donde la información sobre clientes potenciales es de suma importancia para dirigir de mejor manera inversiones publicitarias(Pham *et al.*, 2009). También es claro su potencial en la prevención de delitos que utilizan la *web* como herramienta (Tam & Martell, 2009) detectando de forma automática rasgos de personalidad, género y edad (Mairesse *et al.*, 2007; Schwartz *et al.*, 2013a) para determinar si algún usuario está ocultando su verdadera identidad para facilitar el proceso delictivo. Existen trabajos donde, a través de la DPA se ha determinado ciertos rasgos particulares como del bienestar de distintas poblaciones (Schwartz *et al.*, 2013b), preferencia política (Rao *et al.*, 2010) u ocupación del usuario (Pham *et al.*, 2009). También existen trabajos que determinan el perfil psicológico de los pacientes a través de sus textos para tener un apoyo el diagnostico del experto (Mairesse & Walker, 2006; Nowson & Oberlander, 2006; Fink *et al.*, 2012).

La DPA es una tarea que tradicionalmente se ha atacado a través de enfoques supervisados para clasificar a cada usuario en una colección (Nowson & Oberlander, 2006; Goswami *et al.*, 2009; Rangel & Rosso, 2015; Tang *et al.*, 2015). Al igual que en otros enfoques de clasificación supervisada, un aspecto importante a definir es el modelo de representación, es decir, el tipo de atributos que se utilizará para representar a los usuarios de redes sociales a clasificar. Normalmente, los trabajos se han enfocado en determinar el perfil del autor únicamente a través del texto, sin embargo en tiempos recientes se ha intentado aprovechar otro tipo de información siendo los más exitosos los que aprovechan información de tipo visual. En las siguientes secciones se analizarán los enfoques de representación con los que se ha atacado el problema de DPA.

2.1. Enfoques basados en texto

En general, la mayoría de los trabajos en DPA se han enfocado en el uso de atributos temáticos (o de contenido) y estilísticos (Argamon *et al.*, 2005; Stamatatos, 2009; Mukherjee & Liu, 2010; Najib *et al.*, 2015; Álvarez-Carmona *et al.*, 2015; López-Monroy *et al.*, 2015).

Entre los atributos de contenido, los más usados son la bolsa de palabras (Wagner *et al.*, 2013; Ardehaly & Culotta, 2015) y de n-gramas de palabras (Rao *et al.*, 2010; De Silva & Riloff, 2014).

Por su parte, entre los enfoques basados en estilo se han usado atributos estilo-métricos como las longitudes de las oraciones y las palabras (Adali & Golbeck, 2012); también se han considerado las frecuencias de uso de varios tipos de elementos como signos de puntuación, mayúsculas, palabras fuera del diccionario, información sobre errores ortográficos y uso de *emoticones* (Yatam & Reddy, 2014; Mechti *et al.*, 2015; Simaki *et al.*, 2015). Otros métodos basados en n-gramas también han tenido buenos resultados como los n-gramas de POS (*Part Of Speech*) y los n-gramas de caracteres (González-Gallardo *et al.*, 2015).

En los últimos años, los principales aportes en esta tarea se han dado a través la selección y combinación de algunos de las representaciones descritas anteriormente. Sin embargo recientes trabajos han propuesto algunas representaciones diferentes a las tradicionales que han dado

buenos resultados. López-Monroy *et al.* (2015) proponen una representación concisa basada en encontrar subperfiles de usuarios y determinar la probabilidad de que un texto pertenezca a cada subperfil. Por otro lado (Rangel & Rosso, 2015) propone un método basado en la polaridad de los textos de los autores demostrando que este tipo de información puede ser valiosa para la tarea de DPA. Estos nuevos métodos dan evidencia de que es posible y que vale la pena diseñar nuevas representaciones que sean capaces de superar las representaciones tradicionales.

2.2. Enfoques basados en información no textual

Es importante observar que, a diferencia de las tareas de clasificación clásicas, el tipo de datos que se disponen para realizar la tarea de clasificación de usuarios en el contexto de las redes sociales es muy rico y heterogéneo. Esto se debe a que no sólo se tiene acceso al contenido textual de los mensajes sino también a imágenes y elementos lexicográficos propios del dominio (tales como *hashtags*, menciones, *links*, emoticones, etc.) (Wolf, 2000; Fink *et al.*, 2012), y diversos tipos de interacciones "sociales" del usuario con otros usuarios y con los contenidos que comenta o hace disponible (Pennacchiotti & Popescu, 2011; Staiano *et al.*, 2012). Existen diversas tareas donde se ha utilizado información multimodal con éxito como la búsqueda de imágenes en la web donde se utiliza información social (Bergamo & Torresani, 2010; Escalante *et al.*, 2012; Cui *et al.*, 2014), análisis de sentimientos donde se toma en cuenta imágenes, texto y características de la plataforma social (Maynard *et al.*, 2013) así como en la tarea de etiquetado automático de imágenes (Srivastava & Salakhutdinov, 2012; Eltaher & Lee, 2015; Merler *et al.*, 2015). Es por estas razones que la idea de utilizar un enfoque multimodal para construir un clasificador en la tarea de DPA surge de manera natural.

Por ejemplo, en Wagner *et al.* (2013) los resultados muestran que los atributos derivados de las palabras, conceptos y temas eran más útiles para predecir la profesión de los usuarios, mientras que los atributos derivados de aspectos sociales tales como la lista de contactos y los temas de los que hablan los amigos más cercanos eran más útiles para predecir aspectos de personalidad como introversión y apertura. También en (Culotta *et al.*, 2015) se observa que al hacer uso de ciertos elementos derivados del grafo social, tales como la información de los

usuarios a los que la persona sigue, se pueden predecir un conjunto de datos demográficos tales como la edad, género, etnia, etc.

Es importante remarcar que la principal ventaja en el uso de atributos multimodales radica en que diferentes tipos de características aportan diferentes niveles predictivos dependiendo de cada criterio a clasificar. De esta forma, al complementar la información que cada tipo de atributo aporta se obtiene un aumento en la precisión del clasificador. Así, cuando un tipo de característica no sea adecuado para predecir ciertos aspectos, se puede utilizar la información provista por algún otro tipo de característica para mejorar la predicción. Esta tendencia al uso de información multimodal comienza a observarse claramente en trabajos recientes como el de Rao *et al.* (2010) sobre DPA en *Twitter* donde, información de tipo socio-lingüística, como *emoticones* y *hashtags* es complementada con información obtenida de su red social como el número de seguidores y del comportamiento de comunicación. Pennacchiotti & Popescu (2011) por su parte, también incluyen atributos que describen el comportamiento y nivel de conectividad de los usuarios de las redes sociales, mientras que Montero *et al.* (2014) utilizan información sobre el uso de palabras con una particular carga emocional, y más recientemente se consideran las imágenes compartidas junto con sus etiquetas (You *et al.*, 2014).

Una tendencia similar puede observarse en estudios psicológicos que concluyen que las fotos que publican los usuarios en redes sociales (como *Facebook*) se relacionan con su género, edad y rasgos de personalidad (Hum *et al.*, 2011; Eftekhar *et al.*, 2014; Wu *et al.*, 2014).

Los trabajos con mejores resultados utilizando información multimodal han sido los que aprovechan información de tipo visual. Esto se debe en gran parte a que a diferencia de otro tipo de información como el texto, las imágenes son independientes del lenguaje (Ciot *et al.*, 2013; Nguyen *et al.*, 2014) lo que ha provocado que los esfuerzos se hayan dirigido a analizar las imágenes que comparten los usuarios.

El uso de información multimodal aún es un problema abierto en el área de DPA y abre alternativas para investigaciones futuras.

3. Problemática

La gran mayoría de los trabajos que han tratado de resolver la tarea de DPA se basan únicamente en la información textual que comparten los usuarios en redes sociales. Esto genera que mucha de la información disponible por la misma naturaleza de las redes sociales se desperdicie. Imágenes, vídeos, lista de contactos, horarios de actividad y otro tipo de información no son aprovechados por la mayoría de enfoques. Por esta razón no sabemos cuál de estas diferentes modalidades de información es más valiosa para la tarea de DPA ni qué rasgos demográficos mejoran sus resultados de clasificación con alguna de estas modalidades. Es por esto que es importante hacer un análisis de cómo impacta la información multimodal en la tarea de DPA.

Otro aspecto a resaltar es que los trabajos en DPA han dado evidencia de la importancia del contenido de los textos. El enfoque más común que se ha usado es de la bolsa de palabras. El problema de este enfoque cuando se trabaja sobre redes sociales es la poca información con la que se cuenta porque regularmente se analizan textos cortos, además de que no son textos formales lo que provoca que existan palabras fuera del diccionario y faltas de ortografía.

Un enfoque que no ha sido profundizado lo suficiente para representar el contenido de los textos es el de la extracción de tópicos (*Topic Modeling*). La extracción de tópicos es una tarea que consiste en descubrir patrones abstractos llamados tópicos que ocurren en un corpus (Blei, 2012), donde un tópico es un conjunto de palabras relacionadas entre sí de alguna forma.

Existen enfoques basados en tópicos, que se han usado en otras tareas como en recuperación de la información (Landauer *et al.*, 2013) o desambiguación de palabras Gabrilovich & Markovitch (2007), que podrían ser útiles para la tarea de DPA. Existen dos principales enfoques para la representación de tópicos: i) de forma implícita (Wagner *et al.*, 2013) y ii) de forma explicita, la cual consiste en definir una lista de tópicos a partir de una fuente externa (Gabrilovich & Markovitch, 2007).

La representación de tópicos de forma implícita consiste en obtener tópicos a través del mismo corpus. Existen dos grupos de algoritmos para extraer y representar los tópicos de esta manera. El primero consiste en obtener los tópicos con técnicas basadas en álgebra lineal

(De Lathauwer *et al.*, 1994). Mientras que el segundo grupo utiliza técnicas probabilísticas (Landauer *et al.*, 2013).

El algoritmo LSA (*Latent Semantic Analysis*) ha sido el algoritmo más popular que utiliza técnicas de álgebra lineal, en particular, LSA utiliza una técnica llamada descomposición en valores singulares (Landauer *et al.*, 2013). LSA ha mostrado tener resultados competitivos con los demás algoritmo para extracción de tópicos en diversas tareas.

Por otro lado, los algoritmos basados en técnicas probabilísticas se basan en asumir alguna distribución de probabilidad de los datos para construir los tópicos. Existen diversos algoritmos basados en técnicas probabilísticas. En un intento por mejorar el algoritmo LSA se propuso pLSA (Hofmann, 1999) el cual es una versión probabilística y discreta de LSA. LDA (*Latent Dirichlet Allocation*) (Blei *et al.*, 2003) es un algoritmo que también intenta extraer tópicos, para esto, el algoritmo asume una distribución de Dirichlet en un intento de obtener la mejor agrupación de palabras que describan un tópico. Otro enfoque propuesto es el de *Cluster-based Retrieval* (Liu & Croft, 2004), el cual inicia asumiendo que cada palabra es un tópico para después mezclar las palabras maximizando la probabilidad de que un grupo de palabras pertenezcan a la mismo tópico.

En diversos trabajos se han hecho comparaciones entre estos tipos de algoritmos y los resultados apuntan a que el algoritmo más robusto y con mejores resultados en distintos dominios y tareas ha sido LDA (Wei & Croft, 2006; Seroussi *et al.*, 2011, 2014).

La extracción de tópicos al ser poco explorada dentro de la tarea de DPA se desconoce cuál de sus enfoques tiene el mejor rendimiento.

Por otro lado, existen pocos trabajos que han aprovechado la información extraída a partir de las imágenes que comparten los usuarios a pesar de que diversos trabajos en psicología han concluido que las imágenes que se comparten en redes sociales pueden decir bastante de las personas (Hum *et al.*, 2011; Grimshaw, 2013; Eftekhar *et al.*, 2014; Wu *et al.*, 2014; Kharroub & Bas, 2015). Lo más común es obtener el histograma de color de las imágenes para determinan el género de los usuarios pero no se han hecho estudios para otros rasgos de los autores. Otros trabajos han convertido las imágenes a textos a partir de etiquetadores

automáticos de imágenes que determinan asignar una lista de etiquetas de entre un conjunto previamente establecido para después ser procesados con el enfoque de bolsa de palabras. Esto nos lleva a la misma problemática de los enfoques basados en textos, es decir, el enfoque basado en tópicos no ha sido explorado por lo cuál no se conoce la importancia de estos enfoques para la representación de imágenes en la tarea de DPA.

4. Preguntas, objetivos y contribuciones

En la presente propuesta se pretende contestar las siguientes preguntas de investigación:

1. ¿Qué variante del enfoque basado en la extracción de tópicos de los textos de usuarios en *Twitter* puede ser aplicado para la tarea de DPA de tal modo que obtenga mejores resultados que la representación de bolsa de palabras?

2. ¿De qué forma se pueden modelar los tópicos de las imágenes compartidas por usuarios en *Twitter* de tal modo que sea posible determinar su perfil de autor?

3. ¿De qué manera se puede combinar la información textual y de las imágenes para tomar ventaja de ambos enfoques en la tarea de determinación del perfil de autores?

4.1. Objetivo general

Desarrollar un método para la determinación de perfiles de usuarios en *twitter* utilizando información textual y de imágenes para clasificar los rasgos de género y edad en los idiomas de inglés y español, de tal modo que obtenga mejores resultados que los obtenidos por la representación basada únicamente en información textual.

4.2. Objetivos particulares

1. Diseñar un conjunto de representaciones para usuarios en *Twitter* basadas en tópicos a partir de la información textual para su clasificación en la tarea de determinación de perfil de autores.

2. Diseñar un conjunto de métodos que capturen los tópicos de los textos extraídos de las imágenes compartidas por los usuarios de *Twitter* y representarlos para poder clasificar a los usuarios en la tarea de determinación de perfil de autores.

3. Diseñar e implementar un método para la determinación de perfiles de autores, que tome ventaja de la información textual y de imágenes generadas a partir de los *tweets* de los usuarios para los rasgos de género y edad.

4.3. Contribuciones esperadas

A través de esta investigación doctoral se espera obtener las siguientes contribuciones:

- Un método para generar representaciones del texto del autor a partir de sus tópicos de interés en *Twitter* que sean de utilidad para determinar su perfil de usuario. Además de un mejor entendimiento de los enfoques basado en la extracción de tópicos dentro de la tarea de DPA.

- Un método que sea capaz de capturar información visual compartida por los usuarios de *Twitter* para su clasificación en la tarea de determinación de perfil de autores. También se espera explicar cómo funciona el enfoque basado en tópicos para la representación de imágenes y cuál es su utilidad dentro de la tarea de DPA.

- Un enfoque multimodal para determinar el perfil de un autor en *Twitter* que logre aprovechar la información conjunta del texto y las imágenes compartidas por el usuario para la determinación del perfil del autor.

5. Metodología

En esta sección se explica en detalle la metodología propuesta para alcanzar los objetivos planteados. La metodología planteada consta de los siguientes pasos:

1. **Identificación y obtención de los conjuntos de datos**.

a) **Identificar y construir conjuntos de datos etiquetados de usuarios en *twitter*.**

Existen diversos trabajos que han compartido conjuntos de datos para la tarea de DPA (Schler *et al.*, 2006; Rangel *et al.*, 2014, 2015). Estos conjuntos son interesantes para evaluar trabajos con enfoques basados en texto, el problema es que no contienen información acerca de imágenes compartidas o del comportamientos de los usuarios en la red. Por esta razón es necesario extender algunas colecciones existentes y/o construir un conjunto de datos que contenga este tipo de información para poder hacer una correcta evaluación de los enfoques que se propongan en este trabajo doctoral.

Para construir este conjunto de datos se plantea seguir los siguientes pasos:

1) Desarrollar una herramienta capaz de obtener la información de un perfil dado en *twitter*. La entrada para esta herramienta es el identificador de alguna cuenta. El resultado es toda la información disponible en la cuenta de forma estructurada (texto, imágenes, comportamiento, y usuarios en su red de amigos). Después, el sistema empieza una búsqueda en profundidad de usuarios que son contactos de la cuenta original para obtener del mismo modo su información. El sistema se detiene después de obtener la información de un número de cuentas determinado.
2) Con esta herramienta es posible encontrar los perfiles existentes y ya etiquetados de algunas colecciones y descargar su información visual. De esta forma obtendríamos una colección extendida y apropiada para la evaluación del trabajo propuesto en este documento.
3) También es interesante obtener información únicamente de usuarios mexicanos. Para llevar a cabo el etiquetado se propone utilizar etiquetadores humanos que se encarguen de corroborar los datos de los usuarios en twitter y de esta manera recaudar perfiles con sus respectivas etiquetas de género y edad para poder obtener su información tanto visual como textual.

2. **Analizar y desarrollar métodos basados en características textuales**. En este paso nos plantemos aprovechar la información textual de los usuarios en *Twitter*. El principal objetivo de este trabajo es representar a los usuarios a partir del contenido de sus textos extrayendo sus tópicos de interés. Para esto exploraremos dos enfoques para determinar estos tópicos: los que se extraen de forma implícita y los que se definen de forma explícita. La hipótesis detrás de usar alguno de estos enfoques es que los usuarios que pertenecen a los mismos grupos (género, edad, región de origen, etc) tienden a escribir de los mismo temas.

   También es posible mejorar el resultado de la clasificación en la tarea DPA si se representa el estilo del autor.

   a) **Clasificar a los usuarios a partir de sus tópicos de interés de forma implícita**. Existen algoritmos que son capaces de representar textos por sus tópicos y encontrar relaciones entre palabras aunque éstas no co-ocurran directamente en el documento de manera implícita (Landauer *et al.*, 2013). Para esto se propone utilizar los algoritmos LSA (Landauer *et al.*, 2013) y LDA (Wagner *et al.*, 2013) los cuales han tenido buenos resultados en la tarea de extraer tópicos de interés.

   b) **Clasificar a los usuarios a partir de sus tópicos de interés de forma explícita**. Este enfoque consiste en determinar desde una fuente externa un conjunto de tópicos para después medir su intersección o similitud con los textos de los autores. Existen dos formas de obtener este conjunto de tópicos: i) se definen de forma manual o ii) se extraen automáticamente de un corpus.

   De los trabajos que han determinado manualmente este conjunto de tópicos se propone utilizar ESA (Gabrilovich & Markovitch, 2007) el cual define cada página de wikipedia como un tópico, LIWC (Pennebaker *et al.*, 2001; Tausczik & Pennebaker, 2010) que es un conjunto de colecciones de tópicos definidos por psicólogos para determinar la personalidad de autores y además de estos trabajos también se propone utilizar el árbol de WordNet y utilizar cada uno de sus nodos como un tópico.

Para obtener tópicos de forma automática se propone extraer las palabras más discriminantes para cada rasgo demográficos de los autores a partir de otras colecciones.

c) **Representación del estilo del autor**. Un aspecto muy importante en la tarea de DPA es el de representar el estilo del autor. Se ha demostrado que usuarios con los mismos rasgos demográficos tienden a tener un estilo para escribir similar (Koppel & Schler, 2004). Existen diversos trabajos que han propuesto un conjunto de medidas para capturar el estilo del autor que pueden ser útiles para esta tarea (Koppel *et al.*, 2005; Adali & Golbeck, 2012; González-Gallardo *et al.*, 2015).

3. **Capturar información de las imágenes**. Existen diversos trabajos que han demostrado que el hecho de utilizar información visual ha conseguido mejorar el resultado de clasificación en tareas de DPA (You *et al.*, 2014; Merler *et al.*, 2015). Para capturar esta información proponemos utilizar una representación análoga a la representación textual, es decir, basada en estilo y contenido.

   a) **Representación basada en el estilo de las imágenes**. Para esto se plantea utilizar diferentes tipos de información que capture el estilo de los usuarios con el que comparten imágenes.

      1) **Determinar la frecuencia con la que los usuarios comparten imágenes**. Algunos trabajos de psicología han llegado a la conclusión de que la cantidad de información visual que se comparte dice mucho de los autores en una red social (Tominaga & Hijikata, 2015). Esto provoca que aunque los usuarios no compartan imágenes, de alguna forma eso dice algo de su personalidad. Se plantea capturar la frecuencia de imágenes compartidas en el tiempo para observar si esta información es relevante para los rasgos de género y edad.

      2) **Capturar la información a través de descriptores**. Este punto se aborda bajo la hipótesis de que los colores y la textura en las imágenes que se comparten pueden decir algo con respecto a los rasgos demográficos del usuario (You *et al.*,

Figura 2. Ejemplo de una imagen compartida en twitter

2014). Se plantea observar el rendimiento de descriptores de información general como DCD, CSD, CLD, HTD, TBD, EHD, etc.

b) **Representación basada en el contenido de las imágenes**. En este paso proponemos utilizar etiquetadores automáticos de imágenes para poder obtener información del contenido de las mismas y conocer los objetos que aparecen en ellas. Esta información puede ser útil para la tarea de DPA porque nos permitiría saber los intereses de los usuarios.

1) **Etiquetado no supervisado de las imágenes**. Existen métodos que dada una imagen devuelven un etiquetado de los objetos que se reconocieron en dicha imagen. Ejemplos de estos métodos pueden ser (Pellegrin *et al.*, 2015), (Rashtchian *et al.*, 2010) y (Wang *et al.*, 2015).

Por ejemplo, si le aplicáramos el sistema de Pellegrin *et al.* (2015) a la imagen 2, las 10 etiquetas más relevantes que obtendríamos se uestran en la tabla 1

Como podemos ver, los etiquetadores automáticos no son perfectos ya que aparecen etiquetas que son erróneas como *waterfall*. Sin embargo las etiquetas per-

Cuadro 1. Las 10 etiquetas más importantes según un etiquetados automático no supervisado para la imagen 2

cliff	beach
shore	boat
mountain	waterfall
lake	river
cave	pants

tenecen a un mismo campo semántico por lo que es posible que el ruido que se podría introducir no afecte de forma importante el resultado.

Las etiquetas obtenidas pueden ayudar a determinar algunos intereses del usuario. Se propone manipular a estas etiquetas como si fueran un texto y determinar sus tópicos como se describe en el paso 2 de esta metodología.

2) **Etiquetado supervisado de las imágenes**. Otros trabajos se enfocan en determinar si hay un objeto específico en la imagen o no. Si se seleccionan algunos objetos característicos de algunos rasgos demográficos y después se construyen clasificadores para determinar si estos objetos aparecen o no en las imágenes compartidas por el usuario entonces estas etiquetas pueden ser útiles para la tarea de DPA. De esta forma se pueden aplicar los métodos basados en tópicos de forma explicita para determinar los objetos importantes para detectar en las imágenes para la tarea de DPA

   Por ejemplo, Merler *et al.* (2015) proponen algunos "objetos"de importancia para la tarea de DPA que pueden aparecer en las imágenes de los usuarios. Las categorías que propone se muestran en la tabla 2. Para determinar si cada uno de estos n objetos se encuentra en alguna imagen se entrenan n clasificadores binarios donde el clasificador n_i determina si el objeto i se encuentra o no en dicha imagen.

   Por ejemplo, si utilizamos los clasificadores entrenados de la tabla 2 para la imagen 2 las clases *adult, beach, female adult, human portrait, view, human* y *nature* tendrían resultado positivo mientras que las demás categorías obtendrían

Cuadro 2. 25 clases importantes para la tarea DPA según Merler *et al.* (2015)

adult	animal	baby	beach	boy
brand	logo	building	CGI	car
cat	child	dog	elderly man	elderly person
elderly woman	female adult	girl	horse	human portrait
view	human	icon	male Adult	nature

un resultado negativo. De esta forma se puede aplicar este proceso a todo el historial de imágenes del usuario para obtener la frecuencia de cada categoría y con este vector se puede entrenar un clasificador final para obtener la decisión del rasgo demográfico que se esté determinando.

Para este trabajo se propone utilizar los clasificadores pre-entrenados de ImageNet Rastegari *et al.* (2016)

4. **Desarrollar e implementar un método para llevar a cabo la clasificación usuarios en *twitter* que integre la información textual y de imágenes**. Este último paso involucra el desarrollo de un método para la clasificación de usuarios en *twitter* que pueda tomar en cuenta distintos tipos de atributos para la clasificación. Por ejemplo, un algoritmo de ensamble que pueda sacar ventaja de distintos atributos como, textuales y de imágenes. El conjunto de ensambles podrían ser combinados a través de técnicas de fusión de información como:

   - **Fusión tardía**: cada conjunto de atributos, propiamente representados en un vector, es tomado para entrenar un clasificador del ensamble, o bien lanzar una consulta, ponderando y mezclando los resultados obtenidos (Kuncheva, 2004; Snoek *et al.*, 2005).
   - **Fusión temprana**: toda la información es tomada como un solo vector para entrenar un clasificador o lanzar una consulta (Kuncheva, 2004; Snoek *et al.*, 2005).
   - **Aprendizaje de múltiples *kernels* para clasificación**: cada conjunto de atributos propiamente representados en un vector, es tomado para entrenar una máquina

de vectores de soporte, posteriormente los *kernels* son combinados en uno mismo (e.g., generalmente a través de operaciones lineales) (Gönen & Alpaydın, 2011).

6. Plan de trabajo

A continuación se presenta de forma general un plan de trabajo para los siguientes tres años para algunas de las tareas más relevantes que se tienen planeadas hasta el momento. Este plan se puede ver en el cuadro 1.

7. Trabajo realizado y resultados preliminares

En este documento se describe el trabajo que se ha realizado en el periodo de Enero-Diciembre 2015. El trabajo realizado hasta hoy consiste en lo siguiente:

1. **Clasificar usuarios a partir de tópicos de forma implícita** (**Segundo paso de la metodología inciso a)**). Para llevar a cabo este paso se seleccionó el algoritmo LSA. Con este enfoque el equipo del INAOE participó en la competencia internacional del PAN del 2015 donde se obtuvieron los mejores resultados para la tarea de DPA.

2. **Clasificar usuarios a partir de tópicos de forma explícita automáticamente** (**Segundo paso de la metodología inciso b)**). Para este paso se extrajeron las palabras más discriminantes de una colección de *blogs* para DPA y se utilizó la representación de palabras de word2vec y la representación de documentos de doc2vec. La idea de utilizar estos enfoques era la de utilizar una extensión de las palabras aprovechando información de su contexto para convertir una palabra discriminante en un tópico.

3. **Clasificar usuarios a partir de sus tópicos de interés de forma explícita manualmente** (**Segundo paso de la metodología inciso b)**). A diferencia del experimento del punto anterior, en este paso nos plantemos una elección de tópicos definidos manualmente para la tarea de DPA. Para esto utilizamos un enfoque llamado LIWC el cual esta basado en una teoría psicológica para determinar la personalidad de autores.

Cuadro 3. Cronograma de actividades

Actividad	2015												2016												2017												2018											
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Revisión del estado del arte																																															
Identificación, extensión y construcción de los conjuntos de datos																																															
Implementación y desarrollo de métodos basados en tópicos de forma implícita																																															
Implementación y desarrollo de métodos basados en tópicos de forma explícita																																															
Análisis de los resultados																																															
Elaboración de la propuesta																																															
Defensa de la propuesta													x																																			
Escritura de un artículo																																															
Presentación de avances de segundo año																									x																						
Determinar la importancia de la frecuencia con la que los usuarios comparten imágenes																																															
Determinar la importancia del color y textura en las imágenes para DPA																																															
Analizar el texto extraído de un etiquetado automático no supervisado de imágenes																																															
Analizar el texto extraído de un etiquetado automático supervisado de imágenes																																															
Analizar los resultados y determinar la importancia de las imágenes para DPA																																															
Presentación de avances de tercer año																																					x										
Mezclar la información textual, de comportamiento y de imágenes a través de una fusión temprana																																															
Mezclar la información textual, de comportamiento y de imágenes a través de una fusión tardía																																															
Mezclar la información textual, de comportamiento y de imágenes a través de múltiples *kernels*																																															
Determinar la importancia de cada tipo de información para la tarea de DPA																																															
Escribir la tesis																																															
Escribir un artículo																																															
Entrega de la tesis																																															
Revisión y corrección del documento de tesis																																															
Defensa de la tesis																																															

Actividades
Avances
x Entregas

7.1. Clasificar usuarios a partir de tópicos de forma implícita

Se ha demostrado a lo largo de los últimos años que el contenido de los textos son importantes para determinar los rasgos demográficos de una persona (Schler *et al.*, 2006; Koppel *et al.*, 2009). La mayoría de trabajos han utilizado la representación de bolsa de palabras para capturar el contenido de los textos. Sin embargo con este enfoque los trabajos son incapaces de capturar relaciones entre palabras y extraer los tópicos de interés. Para capturar este tipo de información proponemos utilizar ideas que originalmente se utilizaron para tareas de recuperación de la información explotando la representación del algoritmo LSA (*Latent Semantic Analysis*)(Wiemer-Hastings *et al.*, 2004). LSA representa términos y documentos dentro de un espacio semántico. Los tópicos son resaltados bajo la representación de LSA puesto que por sus características, este algoritmo remueve la mayor cantidad de ruido en una colección de documentos dada para enfatizar patrones entre palabras.

La hipótesis detrás de esta idea es que las personas de los mismos grupos tienden a hablar de los mismos temas y con LSA los autores que hablan de los mismo temas tendrían una representación similar.

7.2. Clasificar usuarios a partir de tópicos de forma explícita automáticamente

Una forma de determinar las palabras discriminantes es obteniendo la ganancia de información de su representación de bolsa de palabras en alguna colección. Cada una de estas palabras aunque no representen en sí un tópico de alguna manera son valiosas para la tarea. Se espera que las palabras que pertenezcan a un mismo tópico se usen en un contexto similar por lo que capturar el contexto de las palabras más discriminantes para la tarea de DPA de alguna forma podría representar a los tópicos más discriminantes.

Existen representaciones que son capaces de capturar el contexto de las palabras. Uno de estos enfoques es word2vec (Mikolov *et al.*, 2013). Word2vec es un modelo no supervisado basado en una red neuronal recurrente que aprende representaciones de vectores por palabras a partir de un corpus de entrada que puede ser muy grande. Se ha demostrado que los vectores

aprendidos explícitamente codifican patrones y regularidades lingüísticas a través del contexto de las palabras. Así por ejemplo si se realiza la siguiente operación entre vectores es posible percatarnos de estas relaciones : vec ("king") – vec("man") + vec("woman") $\simeq$ vec("queen") . Los creadores de este método también demostraron la existencia de las regularidades lingüísticas mediante pruebas por analogía y similitud de palabras. Este enfoque ha tomado mucha fuerza en diferentes tareas como traducción automática, similitud textual, etiquetado de partes de la oración, encontrar sinónimos, agrupamiento de palabras, etc (Wang, 2014), y en este trabajo proponemos utilizarlo para la tarea de DPA.

Una extensión de word2vec es doc2vec, que a diferencia del enfoque original consiste en tener una representación vectorial por documento en lugar de tenerla por palabra (Campr & Ježek, 2015). Si obtenemos la representación de los documentos de cada autor mediante doc2vec podemos encontrar la distancia que existe entre cada autor y la representación vectorial de alguna palabra en específico. De esta forma podemos representar a cada autor a partir de las distancias que hay entre su vector proveniente de doc2vec y un conjunto de palabras importantes para discriminar entre rasgos demográficos para que se conviertan en las características que finalmente serán proporcionadas a un algoritmo de aprendizaje automático.

7.2.1. Selección de palabras discriminantes

Schler *et al.* (2006) realizaron un estudio sobre una colección importante de *blogs* para determinar las palabras más discriminantes para clasificar género y edad de los autores en *blogs*. Para esto, se buscaron las palabras con mayor ganancia de información. Los autores publicaron una lista de las 69 palabras más dicrimiantes. Estas palabras se pueden observar en la tabla 4. Después hicieron experimentos con las diez mil palabras con mayor ganancia de información. Para nuestro trabajo tomamos estas palabras y experimentamos con su representación vectorial.

En este trabajo utilizaremos la representación vectorial de las mismas palabras que se descubrieron en el trabajo de Schler *et al.* (2006).

Cuadro 4. Las 69 palabras mas discriminantes para género y edad según Schler *et al.* (2006)

linux	gaming	server	software	gb	programming
google	data	graphics	india	nations	democracy
users	economic	shopping	mom	cried	freaked
pink	cute	gosh	kisses	yummy	mommy
boyfriend	skirt	adorable	husband	hubby	maths
homework	bored	sis	boring	awesome	mum
crappy	mad	dumb	semester	apartment	drunk
beer	student	album	college	someday	dating
bar	marriage	development	campaign	tax	local
democratic	son	systems	provide	workers	money
job	sports	tv	sleep	eating	sex
family	friends	emotions			

7.3. Clasificar usuarios a partir de tópicos de forma explícita manualmente

Un enfoque basado en tópicos de forma explicita manualmente que podría ser útil para este fin es LIWC (Pennebaker *et al.*, 2001; Tausczik & Pennebaker, 2010). Este método fue definido originalmente por psicólogos para determinar la personalidad de algún autor. LIWC es un conjunto de colecciones de tópicos y categorías estilo-métricas que se construyen a partir de diccionarios previamente definidos. De esta forma los atributos de cada usuarios ya no son las frecuencias de las palabras en el texto sino las apariciones de tópicos y estilo.

LIWC en su versión más reciente[11] contiene 41 categorías temáticas y 25 estilo-métricas. En la tabla 5 se muestran las 41 categorías temáticas. Cada una de estas categorías cuentan con una lista de palabras relacionadas temáticamente.

El uso de LIWC en la tarea de DPA ha ganado popularidad en los últimos años. Cada vez son más los trabajos que han utilizado este enfoque como parte de los atributos (Mukherjee & Liu, 2010; Nguyen *et al.*, 2011; Fink *et al.*, 2012; Schwartz *et al.*, 2013a,b; Kiprov *et al.*, 2015; Bayot *et al.*, 2015). Esto supone que alguna mejora a este método podría tener un impacto importante en la tarea de DPA.

La propuesta es representar a los tópicos definidos en LIWC a través de dos formas: i)

[11] http://liwc.wpengine.com/

Cuadro 5. Categorías temáticas de LIWC

relativity	feel	money	causation	insight
humans	discrepancy	sad	anger	see
affect	home	work	sexual	negative emotion
death	family	tentative	religion	verbs
quant	achievement	health	body	perception
assent	positive emotion	time	leisure	inhibition
hear	friends	anxiety	cognitive	certainty
space	motion	swear	social	biological
ingestion				

Cuadro 6. Descripción de las distribuciones de las clases para género en el corpus del PAN 2014

Clase	**Usuarios**	**Porcentaje**
Masculino	71	49.30 %
Femenino	73	50.69 %

tomando el promedio de los vectores de word2vec de las palabras que pertenecen a cada tópico y ii) tomando únicamente el vector de word2vec de los nombres de las categorías de LIWC de la tabla 5 y utilizar las distancias que hay entre cada una de ellas con los vectores de los autores en el corpus como atributos siguiendo la misma mecánica que el experimento anterior.

7.4. Córpora

7.4.1. Corpus de *Blogs* del PAN 2014 en inglés

Para estos experimentos se utilizó el corpus de *Blogs* en Inglés del PAN 2014. Este corpus está etiquetado para género (masculino y femenino) y para edad (18-24, 25-34, 35-49, 50-64 y 65 o más). En las tablas 6 y 7 se muestran las distribuciones de las clases en el corpus para los rasgos de género y edad respectivamente.

7.4.2. Corpus de *Tweets* del PAN 2015 en inglés

Se utilizó un subconjunto del corpus de *author profiling* del PAN 2015. Este subconjunto está compuesto por usuarios de *twitter* en inglés con 152 usuarios donde cada usuario etiquetado

Cuadro 7. Descripción de las distribuciones de las clases para edad en el corpus del PAN 2014

Clase	Usuarios	Porcentaje
18-24	6	4.61 %
25-34	59	37.48 %
35-49	53	39.66 %
50-64	23	17.12 %
65-xx	3	2.80 %

Cuadro 8. Descripción de las distribuciones de las clases para género en el corpus del PAN 2015

Clase	Usuarios	Porcentaje
Masculino	76	50.00 %
Femenino	76	50.00 %

para género (masculino y femenino) y edad (18-24, 25-34, 35-49 y 50 o más) [12]. En las tablas 8 y 9 se muestran las distribuciones de las clases en el corpus para los rasgos de género y edad respectivamente.

7.5. Configuración experimental

Para cada experimento utilizamos la siguiente configuración: i) consideramos los términos con al menos cinco apariciones en cada corpus, ii) para la representación de las palabras con word2vec se utilizó un modelo pre-entrenado con toda la wikipedia en inglés, estos vectores están representados con 200 dimensiones, iii) para la representación de doc2vec se unieron

[12]También para rasgos de personalidad pero para este experimento no se tomó en cuenta

Cuadro 9. Descripción de las distribuciones de las clases para edad en el corpus del PAN 2015

Clase	Usuarios	Porcentaje
18-24	58	38.15 %
25-34	60	39.47 %
35-49	22	14.47 %
50-xx	12	7.89 %

Cuadro 10. Resultados de la exactitud obtenida para género

Enfoque	*Blogs* PAN 2014	*Twitter* PAN 2015
BoW	73.87	74.00
LSA	**78.91**	**74.34**

Cuadro 11. Resultados de exactitud obtenida para edad

Enfoque	*Blogs* PAN 2014	*Twitter* PAN 2015
BoW	45.57	74.83
LSA	**51.70**	**78.94**

todas las entradas de un autor y se trató como un solo documento, posteriormente se tomó el promedio de los vectores de las palabras del documento de cada autor, iv) la distancia entre un documento y una palabra se obtendrá a través de la distancia coseno de sus vectores (Foote, 1997) v) Utilizamos validación cruzada a 10 pliegues con el clasificador *LibLINEAR* (Fan *et al.*, 2008).

7.6. Resultados experimentales

7.6.1. Resultados del algoritmo LSA para DPA

El objetivo de este experimento es analizar el rendimiento de LSA y compararlo con la representación de la bolsa de palabras para DPA. Los resultados de este experimento para el rasgo de género se muestran en la tabla 10. Para los dos conjuntos de datos el algoritmo de LSA supera los resultados obtenidos por la representación de bolsa de palabras aunque no parece que exista diferencia significativa en la colección del PAN 2015.

En la tabla 11 se muestran los resultados experimentales de la predicción del método propuesto para el rasgo de edad. Estos resultados se comportan de forma similar a los resultados obtenidos para el rasgo de género, es decir, en ambas colecciones el resultado obtenido por el algoritmo de LSA supera a los resultados obtenidos por la representación de bolsa de palabras.

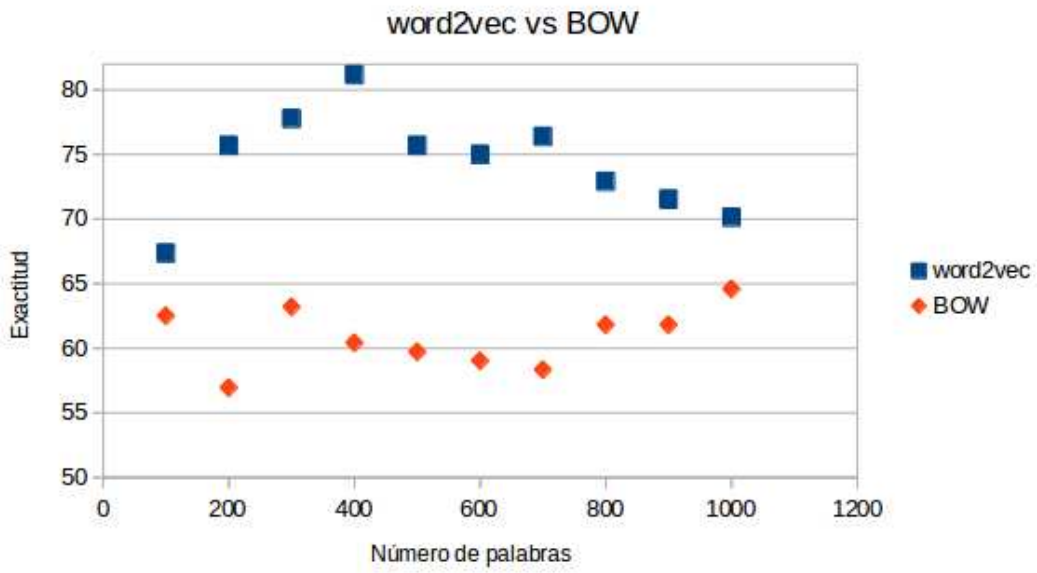

Figura 3. Resultados de exactitud para género con las primeras n palabras con mayor ganancia de información en el corpus del PAN 2014

7.6.2. Resultados con word2vec y doc2vec para DPA

Para mostrar los resultados del enfoque propuesto se hicieron 10 corridas tomando las primero n palabras con mayor ganancia de información según el estudio de Schler *et al.* (2006) donde n va desde cien hasta mil con intervalos de cien palabras y se compararon contra la bolsa de palabras. En la figura 3 se muestra la gráfica de exactitud de este enfoque contra la bolsa de palabras con las primeras n palabras con mayor ganancia de información para el género de los autores, por otro lado en la figura 4 se muestra la misma mecánica pero en esta ocasión para el rasgo de edad de los usuarios; ambas figuras muestran los resultados para el corpus del PAN 2014. En la figura 5 se muestran los resultados obtenidos para el rasgo de género mientras que en la figura 6 se muestran los resultados para el rasgo de edad; en esta ocasión ambas sobre la colección del pan 2015.

En las gráficas de género se ve como la representación basada en word2vec supera a la representación basada en bolsa de palabras en todos los casos. El mejor resultado se obtiene cuando se usan las 400 palabras más discriminativas para la colección del PAN 2014 y 300 para

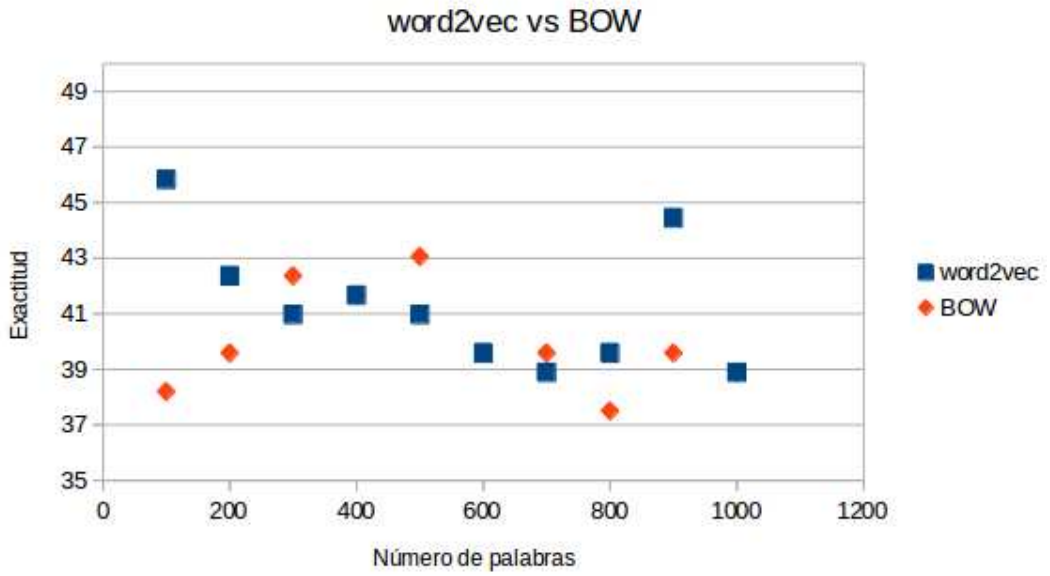

Figura 4. Resultados de exactitud para edad con las primeras n palabras con mayor ganancia de información en el corpus del PAN 2014

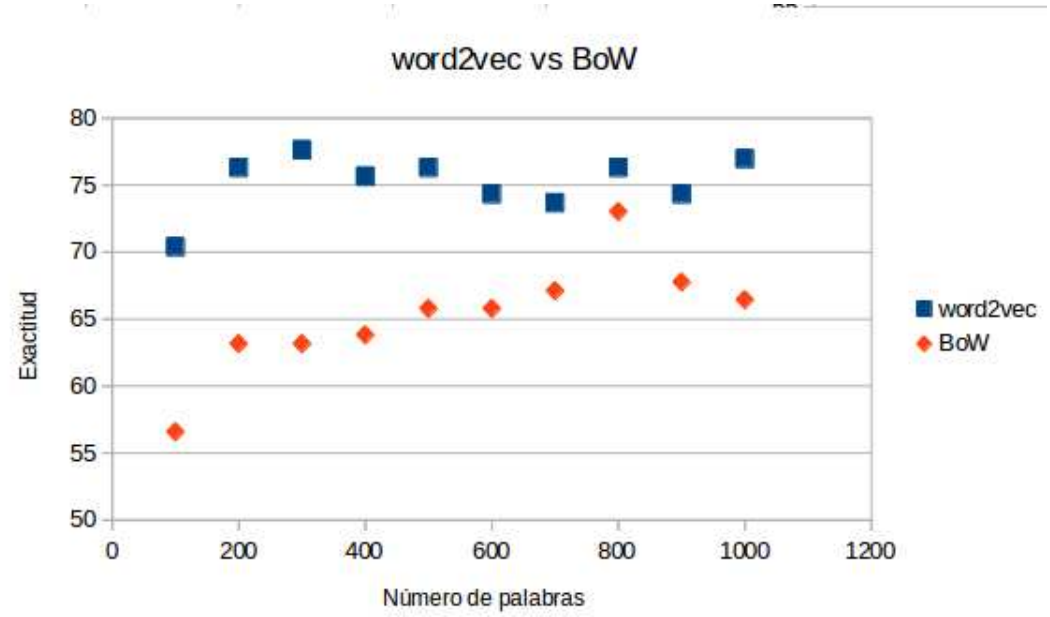

Figura 5. Resultados de exactitud para género con las primeras n palabras con mayor ganancia de información en el corpus del PAN 2015

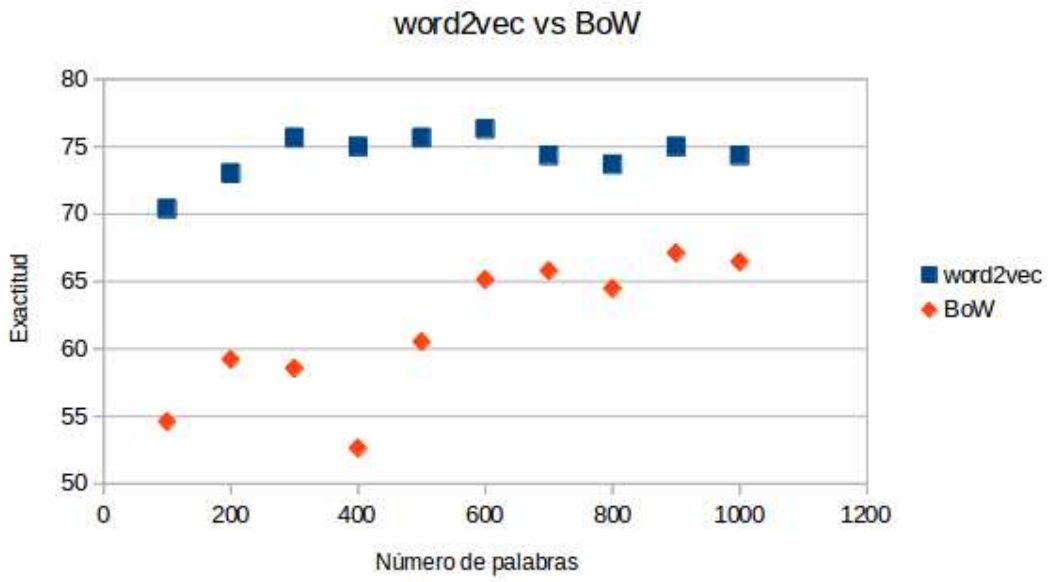

Figura 6. Resultados de exactitud para edad con las primeras n palabras con mayor ganancia de información en el corpus del PAN 2015

Cuadro 12. Resultados de exactitud para género

Enfoque	*Blogs* PAN 2014	*Twitter* PAN 2015
69 palabras discriminantes	60.41	68.14
69 vectores	70.13	74.34
BOW (mejor n)	64.5	73.02
word2vec (mejor n)	**81.16**	**77.63**

el corpus del PAN 2015 mientras que la bolsa de palabras obtiene su mejor resultado con mil palabras y 900 respectivamente. En la tabla 12 se muestra una comparación entre el enfoque basado en representación vectorial y la bolsa de palabras. Estos resultados dan evidencia de que para género la representación basada en word2vec funciona mejor que la bolsa de palabras.

Por otro lado, en la gráfica de edad en la colección del PAN 2014 se ve que mientras se incrementa el número de palabras se va introduciendo ruido ya que el mejor resultado se obtiene con las primeras cien palabras. Incluso hay corridas donde la bolsa de palabras supera al enfoque vectorial. Sin embargo en los resultados obtenidos en la colección del PAN 2015 se observa un comportamiento similar al obtenido en el rasgo de género, es decir, la representación

Cuadro 13. Resultados de exactitud para edad

Enfoque	*Blogs* PAN 2014	*Twitter* PAN 2015
69 palabras discriminantes	39.58	59.30
69 vectores	43.05	60.45
BOW (mejor n)	43.00	67.10
word2vec (mejor n)	**45.80**	**75.65**

Cuadro 14. Resultados de exactitud para género

Enfoque	*Blogs* PAN 2014	*Twitter* PAN 2015
LIWC	58.33	62.5
LIWC -representación vectorial	**74.30**	**71.05**
LIWC -promedio vectorial	64.58	63.15

vectorial supera en todos los casos a la representación de bolsa de palabras. En la tabla 13 se muestra una comparación entre el enfoque basado en representación vectorial y la bolsa de palabras.

7.6.3. Resultados con LIWC para DPA

Para comparar los resultados obtenidos se va a experimentar con el enfoque tradicional de LIWC, con la representación vectorial de las categorías y con el promedio vectorial de las palabras que componen cada categoría.

En la tabla 14 se pueden observar los resultados para determinar el género de los autores del método de LIWC original comparado con las variantes basadas en word2vec. En esta tabla se puede ver que en ambas colecciones el mejor resultado es obtenido por la representación vectorial de las categorías de LIWC. Estos resultados hacen pensar que al promediar todas las palabras de una categoría se esta introduciendo ruido ya que los resultados son menores que cuando solo se toma el vector que representa el nombre de las categorías. Ambos enfoques superan al método original.

En los resultados mostrados en la tabla 15 se describen los valores de exactitud obtenidos determinando la edad de los usuarios. En esta tabla se puede observar un patrón similar que en los resultados anteriores. Para este rasgo también se mejoraron los resultados obtenidos por LIWC para ambas colecciones. Con estos resultados se puede comprobar que es posible la

Cuadro 15. Resultados de exactitud para edad

Enfoque	*Blogs* PAN 2014	*Twitter* PAN 2015
LIWC	32.63	61.18
LIWC-representación vectorial	**47.22**	**67.76**
LIWC-promedio vectorial	45.83	66.44

Cuadro 16. Resultados generales de exactitud para género

Enfoque	*Blogs* PAN 2014	*Twitter* PAN 2015
BoW	73.87	74.00
LSA	78.91	74.34
word2vec (mejor n)	**81.16**	77.63
LIWC -representación vectorial	74.30	71.05
Estado del arte	80.95	**78.28**

introducción de ruido al promediar los vectores de las palabras por categoría. En este rasgo el mejor resultado se obtiene utilizando la representación vectorial del nombre de las categorías.

7.6.4. Comparaciones generales

En la tablas 16 y 17 se pueden observar los mejores resultados obtenidos de cada enfoque que se ha llevado a cabo como parte del trabajo de esta propuesta doctoral para los rasgos de género y edad. Estos resultados se comparan con la representación de la bolsa de palabras y con el mejor resultado en el estado del arte para cada colección. Para el caso de la colección del PAN 2014 el mejor resultado lo obtiene el trabajo de López-Monroy *et al.* (2015) mientras que el mejor resultado en el corpus del PAN del 2015 se obtiene con el trabajo de Álvarez-Carmona *et al.* (2015).

Tanto LSA como el enfoque de word2vec superan a la bolsa de palabras en ambas colecciones

Cuadro 17. Resultados generales de exactitud para edad

Enfoque	*Blogs* PAN 2014	*Twitter* PAN 2015
BoW	45.57	74.83
LSA	51.70	78.94
word2vec (mejor n)	45.80	75.65
LIWC-representación vectorial	47.22	67.76
Estado del arte	**53.06**	**79.60**

para los rasgos de género y edad. Incluso la representación basada en word2vec supera al mejor resultado en el estado del arte en la colección del PAN 2014 para el rasgo de género. Por otro lado el enfoque basado en LIWC no logra superar a la representación de bolsa de palabras en ambas colecciones para el rasgo de edad aunque sí obtiene mejores resultados clasificando género.

Es importante resaltar que estos enfoques están basados únicamente en contenido. Estos resultados mejorarán al incluir medidas estilo-métricas que el estado del arte sí considera.

7.7. Participación en la competencia del PAN 2015

Para participar en esta competencia se propuso utilizar una representación de tópicos a través del algoritmo LSA. La desventaja es que LSA no considera las etiquetas del conjunto de datos para construir su representación, es decir, no es supervisado. Para atacar de algún modo este problema se propone utilizar una representación concisa llamada CSA (Lopez-Monroy *et al.*, 2013) la cual consiste en un vector del tamaño del número de clases posibles en un conjunto de datos donde cada elemento en el vector representa una probabilidad de que un documento dado pertenezca a una clase.

De esta forma la representación de LSA y CSA se unen bajo un enfoque llamado representación temprana (Kuncheva, 2004) la cual consiste en pegar los dos espacios para que los vectores resultantes pasen por un algoritmo de clasificación.

El objetivo de este experimento es analizar el rendimiento de LSA, CSA y la representación de la bolsa de palabras para DPA en el corpus del PAN 2015. Experimentamos con LSA y CSA por separado y con su unión. Esto con la finalidad de observarla contribución de cada una de las representaciones.

7.7.1. Corpus

Se utilizó el corpus completo de prueba de *author profiling* del PAN 2015. El corpus del PAN 2015 esta compuesto por usuarios de *twitter* de cuatro idiomas: Español, Inglés, Italiano

Cuadro 18. Descripción del conjunto de datos

Idioma	Perfiles de autores
Inglés	152
Español	100
Italiano	38
Holandés	34

Cuadro 19. Información de los rasgos de personalidad por idioma

	Inglés		Español		Italiano		Holandés	
Rasgo	Rango	Clases	Rango	Clases	Rango	Clases	Rango	Clases
Extrovertido	[-0.3,0.5]	9	[-0.3,0.5]*	8	[0.0,0.5]*	5	[0.0,0.5]	6
Estable	[-0.3,0.5]	9	[-0.3,0.5]	9	[-0.1,0.5]	7	[-0.2,0.5]	8
Agradable	[-0.3,0.5]	9	[-0.2,0.5]	8	[-0.1,0.5]*	6	[-0.1,0.4]	6
Consiente	[-0.2,0.5]	8	[-0.2,0.5]*	7	[0.0,0.4]	5	[-0.1,0.4]	6
Apertura	[-0.1,0.5]	7	[-0.1,0.5]	7	[0.0,054]	6	[0.1,0.5]	5

y Holandés. Cada idioma esta etiquetado para género (masculino y femenino), edad [13] (18-24, 25-34, 35-49, 50 o más) y cinco rasgos de personalidad (extrovertido, estable, agradable, consiente, apertura). Los valores para los rasgos de personalidad están en un rango continuo entre -0.5 y 0.5. En la tabla 18 se muestra el número de perfiles por cada idioma.

Para identificación de personalidad se muestra en la tabla 19 la cantidad de valores de diferentes de cada rasgo. Si tomamos cada valor diferente como una clase, para cada idioma se muestra el número de clases para cada rasgo[14]

Cuadro 20. Resultados de la exactitud obtenida para género

Idioma	BOW	CSA	LSA	LSA+CSA
Inglés	74.00	70.86	74.34	**78.28**
Español	84.00	74.00	**91.00**	**91.00**
Italiano	76.31	73.68	**86.84**	**86.84**
Holandés	82.35	91.07	**91.17**	**91.17**

Cuadro 21. Resultados de la exactitud obtenida para edad

Idioma	BOW	CSA	LSA	LSA+CSA
Inglés	74.83	68.21	78.94	**79.60**
Español	80.00	74.00	81.00	**82.00**

7.7.2. Resultados en el corpus de entrenamiento

En la tabla 20 se muestran los resultados experimentales de la predicción del método propuesto para género. LSA obtiene los mejores resultados en todos los idiomas excepto en inglés donde mezclar este algoritmo con CSA sí mejora el resultado de predicción.

En la tabla 21 se muestran los resultados experimentales de la predicción del método propuesto para la edad. En este caso solo se muestra para Inglés y Español. Al igual que en el experimento anterior, LSA obtiene los mejores resultados individuales mientras que la combinación de LSA y CSA obtiene el mejor rendimiento en general para los dos idiomas. Es posible que el rendimiento de CSA mejoraría si las colecciones fuesen más grandes[15].

Finalmente, en la tabla 22 se muestra los resultados obtenidos por la representación de la bolsa de palabras (BW) y la combinación propuesta en este trabajo para cada rasgo de personalidad en los cuatro idiomas. Se puede ver que los resultados de la combinación de LSA (L) y CSA (C) superan a la bolsa de palabras. Aunque los resultados podrían prometer mucho se deben tomar con cautela por la poca información en los datos para los rasgos de personalidad ya que una instancia bien/mal clasificada cambia bruscamente el resultado de exactitud en la clasificación.

7.7.3. Resultados oficiales de la competencia PAN 2015

Para participar en la competencia del PAN 2015 era necesario subir el sistema entrenado con el corpus descrito en la sección anterior a una plataforma que los organizadores proporcionaban [16]. Una vez que el software estuviese montado se hacia la evaluación con un corpus de prueba

[13]Solo disponible en Español e Inglés

[14]Los rangos con asterisco indican que existen valores dentro del rango que no aparecen en el corpus. Por ejemplo, en Español(extrovertido y consiente) el valor -0.1 no aparece en ninguna instancia.

[15]Los mejores resultados de CSA se dan en el rasgo de edad según (Rangel *et al.*, 2013, 2014)

[16]www.tira.io/task/author-profiling/

Cuadro 22. Resultados de la exactitud obtenida para personalidad

	Inglés		Español		Italiano		Holandés	
Rasgo	BW	L+C	BW	L+C	BW	L+C	BW	L+C
Extrovertido	64	**87**	62	**87**	65	**94**	64	**91**
Estable	56	**85**	69	**91**	52	**94**	61	**94**
Agradable	60	**80**	62	**84**	71	**92**	61	**88**
Consciente	61	**78**	62	**86**	57	**94**	67	**91**
Apertura	65	**86**	62	**74**	55	**84**	64	**97**

(aún no liberado) publicando posteriormente los resultados finales.

Para cada idioma se calculó el error cuadrático entre la salida de cada sistema (f_{sal}) para personalidad y el resultado del *ground truth*(f_{gt}) de la siguiente manera:

$$RMSE = \sqrt{\frac{\sum_i^n (f_{gti} - f_{sali})^2}{n}}$$

Después se obtiene la exactitud conjunta para género y edad y se obtiene el resultado para cada idioma de la siguiente forma:

$$rank = \frac{(1 - RMSE) + jointAccuracy}{2}$$

Finalmente se obtiene un resultado global a partir del promedio aritmético de los cuatro idiomas. En la figura 7 se muestra la tabla publicada por los organizadores del PAN con los resultados globales de todos los competidores. En esta tabla se puede ver que el resultado del equipo del INAOE (alvarezcarmona15) obtiene los mejores resultados para los idiomas de Inglés, Español y Holandés haciendo que en el resultado global el INAOE obtenga el mejor promedio.

8. Conclusiones

En este documento se describe el trabajo que se ha realizado en el periodo de Enero-Diciembre 2015, y el trabajo que se planea llevar a cabo durante el programa de Doctorado. El principal interés de esta investigación se centra la representación de tópicos tanto textuales como en

Ranking	Team	Global	English	Spanish	Italian	Dutch
1	alvarezcarmona15	**0.8404**	**0.7906**	**0.8215**	0.8089	**0.9406**
2	gonzalesgallardo15	0.8346	0.7740	0.7745	**0.8658**	0.9242
3	grivas15	0.8078	0.7487	0.7471	0.8295	0.9058
4	kocher15	0.7875	0.7037	0.7735	0.8260	0.8469
5	sulea15	0.7755	0.7378	0.7496	0.7509	0.8637
6	miculicich15	0.7584	0.7115	0.7302	0.7442	0.8475
7	nowson15	0.7338	0.6039	0.6644	0.8270	0.8399
8	weren15	0.7223	0.6856	0.7449	0.7051	0.7536
9	poulston15	0.7130	0.6743	0.6918	0.8061	0.6796
10	maharjan15	0.7061	0.6623	0.6547	0.7411	0.7662
11	mccollister15	0.6960	0.6746	0.5727	0.7015	0.8353
12	arroju15	0.6875	0.6996	0.6535	0.7126	0.6843
13	gimenez15	0.6857	0.5917	0.6129	0.7590	0.7790
14	bartoli15	0.6809	0.6557	0.5867	0.6797	0.8016
15	ameer15	0.6685	0.6379	0.6044	0.7055	0.7260
16	cheema15	0.6495	0.6130	0.6353	0.6774	0.6723
17	teisseyre15	0.6401	0.7489	0.5049	0.6024	0.7042
18	mezaruiz15	0.6204	0.5217	0.6215	0.6682	0.6703
19	bayot15	0.6178	0.5253	0.5932	0.6644	0.6881
	ashraf15	-	0.5854	-	-	-
	kiprov15	-	0.7211	0.7889	-	-
	markov15	-	0.5890	0.5874	-	0.6798

Figura 7. Tabla de resultados finales según el PAN 2015 para la tarea de DPA (Rangel $et\ al.$, 2015)

imágenes y de cómo aprovechar la información multimodal de usuarios en *twitter* para la tarea de DPA.

De esta forma, el trabajo se enfocará en hacer clasificación de rasgos demográficos mejorando los métodos basados en tópicos mediante la introducción de información de las imágenes compartidas por los usuarios, además de proponer métodos basados en texto que sean comparables con los del estado del arte. A través de la utilización de imágenes se pretende capturar otro tipo de información, que actualmente no es ampliamente utilizada para la tarea de DPA. Dado que el contenido temático de los documento de los autores ha tenido buenos resultados en la tarea, se propone utilizar representaciones que puedan capturar tópicos de interés para hacer un análisis profundo de estos enfoques. Estos métodos podrían ser algoritmos conocidos como LSA o enfoques más recientes que han tenido buenos resultados en diversas tareas como word2vec y doc2vec.

Otro tipo de información importante para esta tarea podría ser la de las imágenes que comparten las personas. Trabajos recientes han dado evidencia que este tipo de información resulta útil para determinar algunos rasgos de los autores. En general nos propones extraer esta información a cinco niveles: i) a partir de la frecuencia con la que se comparten imágenes, ii) a partir de las etiquetas con las que los mismos usuarios describen a las imágenes. iii) a partir de histogramas de color o representando las imágenes con algún descriptor, iv) extrayendo los tópicos de las imágenes con etiquetadores automáticos no supervisados y iv) con etiquetadores supervisados para encontrar objetos importantes para determinar ciertos rasgos demográficos dentro de las imágenes. Finalmente nos planteamos tomar ventaja de las dos modalidades de información mezclándolas para obtener una representación final de los usuarios. Para ello, se propone trabajar con métodos de fusión tardía, fusión temprana o aprendizaje de múltiples *kernels*, que logren combinar de mejor manera la información de cada espacio de atributos.

9. Publicaciones

Algunos de los resultados preliminares contenidos en esta propuesta de investigación se encuentran publicados en:

- Miguel A. Álvarez-Carmona, A Pastor López-Monroy, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, & Hugo Jair Escalante. Inaoe's participation at pan'15: Author profiling task. *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, 1391, 2015.

Referencias

Ahmed Abbasi & Hsinchun Chen. Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, IEEE*, 20(5):67–75, 2005.

Sibel Adali & Jennifer Golbeck. Predicting personality with social behavior. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, páginas 302–309. IEEE Computer Society, 2012.

Miguel A Álvarez-Carmona, A Pastor López-Monroy, Manuel Montes-y Gómez, Luis Villaseñor-Pineda, & Hugo Jair Escalante. Inaoe's participation at pan'15: Author profiling task. *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, 1391, 2015.

Ehsan Mohammady Ardehaly & Aron Culotta. Inferring latent attributes of twitter users with label regularization. páginas 185–195, 2015.

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, & James W Pennebaker. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, & Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-*, 23(3):321–346, 2003.

Shlomo Argamon, Moshe Koppel, James W Pennebaker, & Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.

Roy Bayot, Teresa Gonçalves, & Paolo Quaresma. Author profiling of twitter users. 2015.

Isaac Bentolila, Yiming Zhou, Labeeb K Ismail, & Richard Humpleman. System, method, and software application for targeted advertising via behavioral model clustering, and preference programming based on behavioral model clusters, May 21 2015. US Patent 20,150,143,414.

Alessandro Bergamo & Lorenzo Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Advances in Neural Information Processing Systems*, páginas 181–189, 2010.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

David M Blei, Andrew Y Ng, & Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Michal Campr & Karel Ježek. Comparing semantic models for evaluating automatic document summarization. In *Text, Speech, and Dialogue*, páginas 252–260. Springer, 2015.

Morgane Ciot, Morgan Sonderegger, & Derek Ruths. Gender inference of twitter users in non-english contexts. In *EMNLP*, páginas 1136–1145, 2013.

Malcolm Corney, Olivier De Vel, Alison Anderson, & George Mohay. Gender-preferential text mining of e-mail discourse. In *Computer Security Applications Conference, 2002. Proceedings. 18th Annual*, páginas 282–289. IEEE, 2002.

Peng Cui, Shao-Wei Liu, Wen-Wu Zhu, Huan-Bo Luan, Tat-Seng Chua, & Shi-Qiang Yang. Social-sensed image search. *ACM Transactions on Information Systems (TOIS)*, 32(2):8, 2014.

Aron Culotta, Nirmal Kumar Ravi, & Jennifer Cutler. Predicting the demographics of twitter users from website traffic data. In *Proceedings of the International Conference on Web and Social Media (ICWSM), in press. Menlo Park, California: AAAI Press*, 2015.

Javier De Andrés, Beatriz Pariente, Martin Gonzalez-Rodriguez, & Daniel Fernandez Lanvin. Towards an automatic user profiling system for online information sites: Identifying demographic determining factors. *Online Information Review*, 39(1):61–80, 2015.

L De Lathauwer, B De Moor, J Vandewalle, & Blind Source Separation by Higher-Order. Singular value decomposition. In *Proc. EUSIPCO-94, Edinburgh, Scotland, UK*, volume 1, páginas 175–178, 1994.

Lalindra De Silva & Ellen Riloff. User type classification of tweets with implications for event recognition. *ACL 2014*, página 98, 2014.

Azar Eftekhar, Chris Fullwood, & Neil Morris. Capturing personality from facebook photos and photo-related activities: How much exposure do you need? *Computers in Human Behavior*, 37:162–170, 2014.

Sara El Manar El & Ismail Kassou. Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86(12), 2014.

Mohammed Eltaher & Jeongkyu Lee. User profiling of flickr: Integrating multiple types of features for gender classification. *Journal of Advances in Information Technology Vol*, 6(2), 2015.

Hugo Jair Escalante, Manuel Montes, & Enrique Sucar. Multimodal indexing based on semantic cohesion for image retrieval. *Information retrieval*, 15(1):1–32, 2012.

Hugo Jair Escalante, Manuel Montes-y Gómez, Luis Villaseñor-Pineda, & Marcelo Luis Errecalde. Early text classification: a naive solution. *arXiv preprint arXiv:1509.06053*, 2015.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, & Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

Clayton Fink, Jonathon Kopecky, & Maksym Morawski. Inferring gender from the content of tweets: A region specific example. In *ICWSM*, 2012.

Jonathan T Foote. Content-based retrieval of music and audio. In *Voice, Video, and Data Communications*, páginas 138–147. International Society for Optics and Photonics, 1997.

Evgeniy Gabrilovich & Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, páginas 1606–1611, 2007.

Mehmet Gönen & Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

Carlos E González-Gallardo, Azucena Montes, Gerardo Sierra, J Antonio Núñez-Juárez, Adolfo Jonathan Salinas-López, & Juan Ek. Tweets classification using corpus dependent tags, character and pos n-grams. 2015.

Sumit Goswami, Sudeshna Sarkar, & Mayur Rustagi. Stylometric analysis of bloggers' age and gender. In *Third International AAAI Conference on Weblogs and Social Media*, 2009.

Mark Grimshaw. *The Oxford handbook of virtuality*. Oxford University Press, 2013.

Ryan CW Hall & Richard CW Hall. A profile of pedophilia: definition, characteristics of offenders, recidivism, treatment outcomes, and forensic issues. In *Mayo Clinic Proceedings*, volume 82, páginas 457–471. Elsevier, 2007.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 50–57. ACM, 1999.

Noelle J Hum, Perrin E Chamberlin, Brittany L Hambright, Anne C Portwood, Amanda C Schat, & Jennifer L Bevan. A picture is worth a thousand words: A content analysis of facebook profile photographs. *Computers in Human Behavior*, 27(5):1828–1833, 2011.

Nitin Indurkhya & Fred J Damerau. *Handbook of natural language processing*, volume 2. CRC Press, 2010.

Tamara Kharroub & Ozen Bas. Social media and protests: An examination of twitter images of the 2011 egyptian revolution. *New Media & Society*, página 1461444815571914, 2015.

Yasen Kiprov, Momchil Hardalov, Preslav Nakov, & Ivan Koychev. Su@ pan'2015: Experiments in author profiling. 2015.

Moshe Koppel, Navot Akiva, Eli Alshech, & Kfir Bar. Automatically classifying documents by ideological and organizational affiliation. In *ISI*, páginas 176–178, 2009.

Moshe Koppel & Jonathan Schler. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, página 62. ACM, 2004.

Moshe Koppel, Jonathan Schler, & Kfir Zigdon. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, páginas 624–628. ACM, 2005.

Ludmila I Kuncheva. Combining pattern classifiers. *Methods and Algorithms. Wiley, Chichester*, 2004.

Thomas K Landauer, Danielle S McNamara, Simon Dennis, & Walter Kintsch. *Handbook of latent semantic analysis*. Psychology Press, 2013.

Xiaoyong Liu & W Bruce Croft. Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 186–193. ACM, 2004.

A Pastor López-Monroy, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, & Efstathios Stamatatos. Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems*, 2015.

Adrian Pastor Lopez-Monroy, Manuel Montes-y Gomez, Hugo Jair Escalante, Luis Villasenor-Pineda, & Esaú Villatoro-Tello. Inaoe's participation at pan'13: Author profiling task. In *CLEF 2013 Evaluation Labs and Workshop*, 2013.

François Mairesse & Marilyn Walker. Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, páginas 543–548, 2006.

François Mairesse, Marilyn A Walker, Matthias R Mehl, & Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, páginas 457–500, 2007.

Diana Maynard, David Dupplaw, & Jonathon Hare. Multimodal sentiment analysis of social media. 2013.

Seifeddine Mechti, Maher Jaoua, & Lamia Hadrich Belguith. Machine learning for classifying authors of anonymous tweets, blogs, reviews and social media, 2015.

Michele Merler, Liangliang Cao, & John R Smith. You are what you tweet... pic! gender prediction based on semantic analysis of social media images. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, páginas 1–6. IEEE, 2015.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, & Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, páginas 3111–3119, 2013.

Calkin Suero Montero, Myriam Munezero, & Tuomo Kakkonen. Investigating the role of emotion-based features in author gender classification of text. In *Computational Linguistics and Intelligent Text Processing*, páginas 98–114. Springer, 2014.

Arjun Mukherjee & Bing Liu. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, páginas 207–217. Association for Computational Linguistics, 2010.

Fahad Najib, Waqas Arshad Cheema, & Rao Muhammad Adeel Nawab. Author's traits prediction on twitter data using content based approach. 2015.

Dong Nguyen, Noah A Smith, & Carolyn P Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, páginas 115–123. Association for Computational Linguistics, 2011.

Dong-Phuong Nguyen, RB Trieschnigg, AS Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, & FMG de Jong. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. Association for Computational Linguistics, 2014.

Scott Nowson & Jon Oberlander. The identity of bloggers: Openness and gender in personal weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, páginas 163–167, 2006.

Luis Pellegrin, Jorge A Vanegas, John Arevalo, Viviana Beltrán, Hugo Jair Escalante, Manuel Montes-y Gómez, & Fabio A González. Inaoe-unal at imageclef 2015: Scalable concept image annotation. 2015.

Marco Pennacchiotti & Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 430–438. ACM, 2011.

James W Pennebaker, Martha E Francis, & Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

Dang Duc Pham, Giang Binh Tran, & Son Bao Pham. Author profiling for vietnamese blogs. In *Asian Language Processing, 2009. IALP'09. International Conference on*, páginas 190–194. IEEE, 2009.

Francisco Rangel, P Rosso, M Potthast, B Stein, & W Daelemans. Overview of the 3rd author profiling task at pan 2015. In *CLEF*, 2015.

Francisco Rangel & Paolo Rosso. On the multilingual and genre robustness of emographs for author profiling in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, páginas 274–280. Springer, 2015.

Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, & Walter Daelemans. Overview of the 2nd author profiling task at pan 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes)*, 2014.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, & Giacomo Inches. Overview of the author profiling task at pan 2013. *Notebook Papers of CLEF*, páginas 23–26, 2013.

Delip Rao, David Yarowsky, Abhishek Shreevats, & Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, páginas 37–44. ACM, 2010.

Cyrus Rashtchian, Peter Young, Micah Hodosh, & Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, páginas 139–147. Association for Computational Linguistics, 2010.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, & Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. *arXiv preprint arXiv:1603.05279*, 2016.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, & James W Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, páginas 199–205, 2006.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, *et al.* Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013a.

Hansen Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, *et al.* Characterizing geographic variation in well-being using tweets. In *ICWSM*, 2013b.

Yanir Seroussi, Ingrid Zukerman, & Fabian Bohnert. Authorship attribution with latent dirichlet allocation. In *Proceedings of the fifteenth conference on computational natural language learning*, páginas 181–189. Association for Computational Linguistics, 2011.

Yanir Seroussi, Ingrid Zukerman, & Fabian Bohnert. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310, 2014.

Vasiliki Simaki, Christina Aravantinou, Iosif Mporas, & Vasileios Megalooikonomou. Automatic estimation of web bloggers' age using regression models. In *Speech and Computer*, páginas 113–120. Springer, 2015.

Cees GM Snoek, Marcel Worring, & Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, páginas 399–402. ACM, 2005.

Nitish Srivastava & Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, páginas 2222–2230, 2012.

Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, & Alex Pentland. Friends don't lie: inferring personality traits from social network structure. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, páginas 321–330. ACM, 2012.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, & Benno Stein. Overview of the pan/clef 2015 evaluation lab. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, páginas 518–538. Springer, 2015.

Jenny Tam & Craig H Martell. Age detection in chat. In *Semantic Computing, 2009. ICSC'09. IEEE International Conference on*, páginas 33–39. IEEE, 2009.

Duyu Tang, Bing Qin, & Ting Liu. Learning semantic representations of users and products for document level sentiment classification. In *Proc. ACL*, 2015.

Yla R Tausczik & James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

Tomu Tominaga & Yoshinori Hijikata. Study on the relationship between profile images and user behaviors on twitter. In *Proceedings of the 24th International Conference on World Wide Web Companion*, páginas 825–828. International World Wide Web Conferences Steering Committee, 2015.

Christoph Wagner, Sitaram Asur, & Joshua Hailpern. Religious politicians and creative photographers: Automatic user categorization in twitter. In *Social Computing (SocialCom), 2013 International Conference on*, páginas 303–310. IEEE, 2013.

Huizhen Wang. Introduction to word2vec and its application to find predominant word senses. 2014.

Josiah Wang, Krystian Mikolajczyk, Alba G Seco de Herrera, Stefano Bromuri, M Ashraful Amin, Mahmood Kazi Mohammed, Burak Acar, Suzan Uskudarli, Neda B Marvasti, José F Aldana, *et al.* General overview of imageclef at the clef 2015 labs. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 6th International Conference of the CLEF Association, CLEF'15, Toulouse, France, September 8-11, 2015, Proceedings*, volume 9283, página 444. Springer, 2015.

Xing Wei & W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 178–185. ACM, 2006.

Peter Wiemer-Hastings, K Wiemer-Hastings, & A Graesser. Latent semantic analysis. In *Proceedings of the 16th international joint conference on Artificial intelligence*, páginas 1–14. Citeseer, 2004.

Alecia Wolf. Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior*, 3(5):827–833, 2000.

Yen-Chun Jim Wu, Wei-Hung Chang, & Chih-Hung Yuan. Do facebook profile pictures reflect user's personality? *Computers in Human Behavior*, 2014.

Satya Sri Yatam & T Raghunadha Reddy. Author profiling: Predicting gender and age from blogs, reviews & social media. In *International Journal of Engineering Research and Technology*, volume 3. ESRSA Publications, 2014.

Quanzeng You, Sumit Bhatia, Tong Sun, & Jiebo Luo. The eyes of the beholder: Gender prediction using images posted in online social networks. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, páginas 1026–1030. IEEE, 2014.

Rong Zheng, Jiexun Li, Hsinchun Chen, & Zan Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2006.

Rong Zheng, Yi Qin, Zan Huang, & Hsinchun Chen. Authorship analysis in cybercrime investigation. In *Intelligence and Security Informatics*, páginas 59–73. Springer, 2003.