Una metodología para la detección del perfil de un autor

Yuridiana Alemán, Darnes Vilariño, David Pinto
yuridiana.aleman@gmail.com, darnes@cs.buap.mx, and dpinto@cs.buap.mx

Facultad de Ciencias de la Computación-BUAP
Av. San Claudio y 14 sur, CP: 72570, Puebla, Mexico
http://www.cs.buap.mx/

Resumen En este artículo se presenta una metodología para la detección del perfil de un autor usando los siguientes dos elementos: género y edad. Para esto se parte de un conjunto de diálogos, escritos en dos idiomas: inglés y español, proporcionado para la competencia Author Profiling dentro del foro de evaluación "Uncovering Plagiarism, Authorship, and Social Software Misuse" (PAN2013). Se utiliza conteos de características léxicas, semánticas y sintácticas para generar un sistema de clasificación de dos fases, el cual clasifica primero el género y posteriormente la edad. Los resultados obtenidos muestran que, con la cantidad de datos disponibles, es posible caracterizar tanto la edad como el género de un autor con una exactitud mayor al 50 %. Sin embargo, estos valores podrían mejorarse al tener una mayor evidencia de información en los datos de entrenamiento.

Palabras clave: Clasificación supervisada, PAN 2013, Género, Edad, Bosque aleatorio

1. Introducción

La detección del perfil de un autor es un problema de creciente importancia en varios campos del conocimiento como medicina forense, seguridad y marketing. Por ejemplo, desde la perspectiva de la lingüística forense, sería importante conocer el perfil lingüístico del autor de un mensaje de acoso. Del mismo modo, desde el punto de vista del marketing, las empresas pueden estar interesadas en saber, mediante el análisis de los blogs y reseñas de productos en línea, que tipo de personas comentan sobre sus productos, y de esta manera encaminar sus campañas publicitarias hacia determinado género o rango de edad.

En este artículo se propone una metodología para detectar dos aspectos del perfil de los autores en los chats o blogs: género y edad. Esta metodología ha permitido crear un sistema que, a partir de un conjunto de diálogos escritos por diferentes personas (cada diálogo contiene el género y grupo de edad de la persona que lo escribió), catalogue un nuevo conjunto de diálogos (conjunto de evaluación) y determine los dos aspectos del perfil antes mencionados. Los textos utilizados se presentan en los idiomas inglés y español.

El artículo está estructurado de la siguiente manera. En la sección II se detalla el estado del arte referente a este tema de investigación. La sección III muestra la metodología planteada y el preprocesamiento dado a los conjuntos de datos. La sección IV muestra los resultados obtenidos en cada corpus. Finalmente, la sección V muestra las conclusiones obtenidas y el trabajo futuro para esta investigación.

2. Estado del arte

Gran parte de las investigaciones sobre blogs y redes sociales van encaminadas a descubrir al autor del texto, sin tomar en cuenta el perfil de dicho autor. En la rama de detección de perfil, específicamente género y edad, se han logrado algunos avances. En [1] se aplican técnicas de aprendizaje automático para la categorización de textos. Se utilizan dos tipos básicos de categorías: basadas en contenido y basadas en estilo, para determinar dos aspectos del perfil del autor (género y edad). Se realizaron tres tipos de experimentos, con los conjuntos de características tratados de forma separada y por último la unión de ambos conjuntos, utilizando taxonomías para determinar la relación de las palabras con la parte del discurso en cada texto escrito.

Esta misma metodología se utiliza en [2] donde el corpus analizado contiene 4 aspectos a determinar: género, edad, personalidad y lengua nativa. Los mejores resultados en género y edad fueron obtenidos utilizando la unión de las características de estilo y contenido; la lengua materna obtiene mejores resultados con el conjunto de características de contenido, mientras que la personalidad obtiene mejores resultados con las categorías de estilo. Las evaluaciones se realizaron con validación cruzada de 10 pliegues. En [3], se presenta un estudio donde se utilizan técnicas de categorización de texto para la predicción de género y edad en un corpus de mensajes escrito en el idioma holandés. Para los experimentos se utiliza el análisis de X^2, n-gramas (palabras y caracteres) y bolsa de palabras. Se realizan varios experimentos con diferentes números de instancias y características para clasificar dos tipos de categorías: adolescentes y adultos, utilizando el clasificador conocido como máquinas de soporte vectorial. Los mejores resultados se obtuvieron usando unigramas de palabras como esquema de representación de los textos.

En investigaciones más recientes como en [4], se estudia la relación entre la edad y el uso del lenguaje en la red social de Twitter en tres diferentes tipos de experimentos. La clasificación de los usuarios por categorías de edades, por etapas de la vida y la predicción de su edad exacta. Se concluye con que un sistema automático logra mejor rendimiento que los seres humanos en estas tareas, y que los seres humanos y los sistemas automáticos tienen dificultades en la predicción de la edad de las personas mayores. Por otra parte, se presenta un análisis detallado de las variables que cambian con la edad. En particular, se encuentran fuertes patrones de cambio mayoritariamente en personas más jóvenes. En [5] se presenta una propuesta para la detección automática del género y de la edad utilizando ciertas características estilísticas, tales como:

emoticonos, signos de puntuación y frecuencia de palabras. Los experimentos fueron realizados utilizando máquinas de soporte vectorial obteniendo alrededor de 0.57 de exactitud en la detección de género y cerca de 0.65 para la detección de la edad. La conclusión obtenida a través de estos resultados es que el estilo de escritura está correlacionado mayormente con la edad de las personas y menos con el género. Es importante mencionar que aunque se han logrado avances en la detección del perfil del autor, estos resultados no han sido validados en otras colecciones de datos, y por tanto, se trata de valores que se encuentran ajustados a ciertos tipos de corpora. En este sentido, en el año 2013 se propuso una competencia internacional (PAN 2013), con la finalidad de proponer un corpus estándar y balanceado, así como evaluar el rendimiento de diversos sistemas a nivel mundial.

3. Metodología

La metodología propuesta se basa en la extracción de diversos conjuntos de características, ésta se muestra en la figura 1.

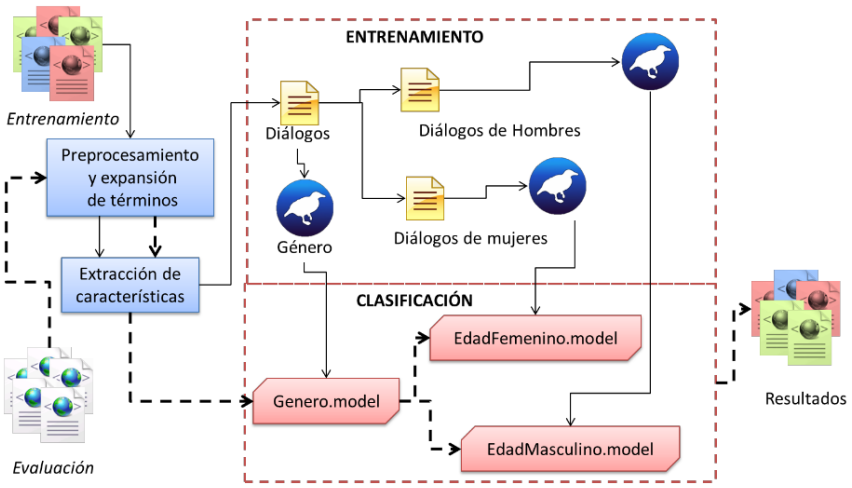

Fig. 1. Metodología propuesta para la investigación

El conjunto de entrenamiento consiste en documentos de tipo XML, que contienen conversaciones sobre diferentes temas agrupados por autor y etiquetados con su idioma (inglés o español), género (masculino y femenino) y grupo de edad. Existen tres grupos de edad:

1. 10s: Personas de 13 a 17 años.
2. 20s: Personas de 23 a 27 años.

3. 30s: Personas de 33 a 47 años.

El corpus del inglés contiene 236,600 autores, mientras que el corpus en español contiene 75,900. En la tabla 1 se muestran algunas estadísticas de los diálogos, separados por idioma y género.

Tabla 1. Conjunto de entrenamiento para la tarea *Author profiling*

	Inglés		Español	
	Femenino	**Masculino**	**Femenino**	**Masculino**
Total de diálogos	118,300	118,300	37,950	37,950
Diálogos de "10s"	8,600	8,600	1,250	1,250
Diálogos de "20s"	42,900	42,900	21,300	21,300
Diálogos de "30s"	66,800	66,800	15,400	15,400
Vocabulario	1,228,711	1,219,020	533,873	592,605
Promedio de palabras	798	705	249	273
Diálogo con más palabras	10,648	12,917	11,806	11,714

El vocabulario es muy extenso, sobre todo para el idioma inglés, esto es lógico debido al número de diálogos para este idioma. Para el caso del idioma español es más extenso el vocabulario en la categoría de masculino, mientras que para el inglés es más extenso el vocabulario en la categoría de femenino. Además, se observa que en todas las categorías de edad, el número de textos es el mismo para ambos géneros (en este sentido es un corpus balanceado por género para cada rango de edad).

Analizando los conjuntos de textos, se llegó a la conclusión de que éstos presentan muchas palabras mal escritas, truncadas, emoticones y vocabulario propio de blogs y chats. Por lo tanto, se realiza una expansión de los textos construyendo algunos recursos léxicos para ambos idiomas,como son: un diccionario de emoticones, un diccionario de abreviaturas, un diccionario de palabras comunes en SMS y de las contracciones mas utilizadas. Usando los recursos léxicos anteriormente mencionados, se fue capaz de sustituir en el corpus de entrenamiento cada ocurrencia de emoticones y contracciones por su correspondiente significado. Se eliminaron además los signos de puntuación y los caracteres no imprimibles. Después de aplicar este pre-procesamiento, el vocabulario en ambos idiomas se redujo drásticamente.

En la fase de extracción de características la mayoría de las investigaciones realizadas emplean todo el vocabulario, sin embargo, en este caso se observa que este tipo de enfoque consumiría demasiados recursos de espacio y tiempo de cómputo, e incluso ciertas herramientas de aprendizaje automático no serían capaces de soportar tal volumen de información, es por ello que se propone utilizar los siguientes conteos:

1. **Categorías Gramaticales:** Se obtuvo la categoría gramatical de cada palabra dentro de los textos, para posteriormente realizar el conteo de cada una de ellas. Para esto, se utilizó la herramienta *Tree-Tagger* [6] para el idioma español y el *Stanford POS-tagger* [7] para el inglés, obteniendo 102 características para el idioma español y 52 para el inglés.

2. **Palabras Cerradas:** Dentro de esta categoría se clasifican grupos de palabras como preposiciones, conjunciones y determinantes. Para los experimentos se realiza un conteo de palabras cerradas en todas las conversaciones; así, cada palabra cerrada representa una característica, y el valor de ésta en cada instancia está dado por las veces que aparece en dicha conversación. Se obtienen 195 palabras para el idioma inglés y 178 para el español.
3. **Sufijos:** Se tomaron como características los sufijos existentes para ambos idiomas. Al igual que en los conjuntos anteriores, cada sufijo representa una característica, y las veces que aparece en cada conversación es el valor para dicho atributo. En este conjunto se obtuvieron 131 características para el idioma inglés y 172 para el español.
4. **Signos:** Se contabilizan todos los signos de puntuación existentes.

Tal y como se ha mencionado anteriormente, en este artículo se utiliza el enfoque de aprendizaje automático para detectar el perfil de los autores. Este enfoque parte de la premisa de la existencia de un corpus supervisado que se usa para entrenar un modelo de clasificación, que es usado posteriormente para calcular la clase asociada a un texto de entrada cuya clase es desconocida. Utilizando las características antes mencionadas, se han generado los siguientes modelos de clasificación mediante el método de *Random Forest* [8]:

1. *Género*: Se usan todos los textos del conjunto de entrenamiento usando el género (masculino, femenino), como atributo clasificador.
2. *EdadMasculino*: Se agrupan todos los textos escritos por hombres usando el rango de edad (10s, 20s y 30s), como atributo clasificador.
3. *EdadFemenino*: Se agrupan todos los textos escritos por mujeres usando el rango de edad (10s, 20s y 30s), como atributo clasificador.

Los modelos se crean de esta manera bajo la premisa de que en primera instancia, es más fácil detectar el género, que la edad en la forma de escribir de un autor. En la Figura 1, se puede observar el proceso de dos fases usado para determinar el perfil del autor. La clasificación de las instancias se realizó evaluando primeramente el modelo Género, y una vez obtenida la categoría (masculino o femenino), los textos etiquetados como "femenino" se evalúan con el modelo *EdadFemenino* para obtener la etiqueta de edad, mientras que los etiquetados como "masculino" son evaluados con el modelo *EdadMasculino*. Con la finalidad de medir objetivamente los resultados obtenidos, se han llevado a cabo dos tipos de evaluaciones:

1. **Validación cruzada a 10 pliegues:** Los datos se dividen en K subconjuntos, uno de los subconjuntos se utiliza como datos de prueba y el resto (K-*1*) como datos de entrenamiento. Este proceso se repite 10 veces.
2. ***Software submission*:** Se genera el modelo de clasificación usando la totalidad de los datos de entrenamiento para el conjunto de evaluación de la competencia "PAN 2013"[1]. Regularmente, los foros de evaluación proporcionan el conjunto de prueba para validar los modelos de clasificación, de

[1] http://pan.webis.de/

tal manera que se envían los resultados obtenidos, los cuales son evaluados y comparados contra otras aproximaciones. Sin embargo, en el marco del "PAN 2013", la metodología establecida fue que cada equipo debe enviar el software programado e instalarlo en una máquina virtual. Así, los organizadores son capaces de ejecutarlo usando un conjunto de datos prueba que no es accesible para los equipos que envían el software.

4. Resultados obtenidos

En esta sección se presentan los resultados obtenidos tanto para la evaluación por validación cruzada, como por *software submission*. Para el análisis de los resultados obtenidos se utilizan las métricas de precisión, recuerdo y la medida F_1, la cual se define como la media armónica entre la precisión y el recuerdo.

4.1. Validación cruzada a 10 pliegues

Como se mencionó anteriormente, en este tipo de evaluación se usa únicamente el conjunto de datos de entrenamiento. En la Tabla 2 se muestran los resultados obtenidos al clasificar solamente el género, mientras que en la Tabla 3 se muestran los valores obtenidos al clasificar únicamente la edad del autor. El valor de medida F_1 obtenido para el caso de la clasificación de género es de 0.54, para el idioma Español, mientras que es de 0.55 para el idioma Inglés. La clasificación de la edad arrojó un F_1 igual a 0.37 para el idioma Español y 0.44 para el idioma Inglés. Como puede observarse, sobre el conjunto de datos usado como entrenamiento, los valores promedio validan la hipótesis planteada, en la cual se indica que es más fácil detectar el género del autor, que su edad. El comportamiento observado fue similar para ambos idiomas (Español e Inglés).

Tabla 2. Evaluación por validación cruzada usando el género (masculino y femenino) como atributo clasificador

	Español			Inglés		
Categoría	**Precisión**	**Recuerdo**	F_1	**Precisión**	**Recuerdo**	F_1
Masculino	0.5350	0.6430	0.5840	0.5410	0.6480	0.5897
Femenino	0.5560	0.4440	0.4937	0.5640	0.4530	0.5024
Promedio	**0.5455**	**0.5435**	**0.5445**	**0.5525**	**0.5505**	**0.5515**

Sin embargo, este comportamiento podría estar afectado de alguna manera por el número de muestras para la clase de autores que se encuentran en el rango de los 10 años (10s), ya que al ser pocas pueden no representar correctamente la clase y generar un modelo sobreajustado, que no se comporte adecuadamente sobre el corpus de prueba. Esta clase es demasiado pequeña con relación a las otras dos (20s y 30s) y por tanto, el proceso de clasificación puede tender a favorecer a las clases con mayor número de muestras. Bajo este razonamiento, y considerando un proceso de clasificación basado en dos fases, se proponen dos

Tabla 3. Evaluación por validación cruzada usando la edad (10s, 20s y 30s) como atributo clasificador

	Español			Inglés		
Categoría	**Precisión**	**Recuerdo**	F_1	**Precisión**	**Recuerdo**	F_1
10s	0.0000	0.0000	0.0000	0.3000	0.0200	0.0375
20s	0.6100	0.7900	0.6900	0.5100	0.4700	0.4892
30s	0.5100	0.3300	0.4100	0.6400	0.7400	0.6864
Promedio	**0.3733**	**0.3733**	**0.3733**	**0.4833**	**0.4100**	**0.4437**

evaluaciones: primeramente clasificar el género del autor seguido de la clasificación de la edad (tal y como se propuso en la metodología inicial) y en una segunda evaluación se realiza el proceso inverso.

En la Tabla 4 se presentan los valores de precisión, recuerdo y F_1 para cada modelo de clasificación aplicado en la segunda fase de la metodología inicial, es decir, tomando los modelos de género para determinar la edad.

Tabla 4. Resultados de los modelos que usan la edad (10s, 20s y 30s) como atributo clasificador en la segunda fase de clasificación.

		Español			Inglés		
Modelo	Categoría	**Precisión**	**Recuerdo**	F_1	**Precisión**	**Recuerdo**	F_1
Edad Masculino	10s	0.1430	0.0040	0.0078	0.3100	0.0280	0.0514
	20s	0.6110	0.7820	0.6860	0.5180	0.4750	0.4956
	30s	0.5190	0.3550	0.4216	0.6240	0.7330	0.6741
	Promedio	0.4243	0.3803	0.4011	0.4840	0.4120	0.4451
Edad Femenino	10s	0.0000	0.0000	0.0000	0.3060	0.0300	0.0546
	20s	0.6040	0.7680	0.6762	0.5170	0.4580	0.4857
	30s	0.5050	0.3530	0.4155	0.6440	0.7610	0.6976
	Promedio	0.3697	0.3737	0.3717	0.4890	0.4163	0.4497
Promedio General		**0.3970**	**0.3770**	**0.3864**	**0.4865**	**0.4142**	**0.4474**

Se observa un mejor comportamiento en el rango de edad es de 20 años, para los diálogos escritos en Español (independientemente del género). En el caso de los diálogos escritos en Inglés, el mejor resultado se obtiene en el rango de edad de los 30 años. Desde nuestro particular punto de vista, estos resultados están co-relacionados con la cantidad de diálogos que se tienen por rango de edad en el conjunto de datos de entrenamiento. El experimento arroja un F_1 de 0.39 para el idioma Español y de 0.45 para el Inglés. Tal y como se esperaba, estos valores de F_1 son menores que aquellos obtenidos en la primera fase de clasificación. Los diálogos mal clasificados en la fase uno, están destinados a provocar un margen de error cercano al 45 % en la segunda fase de clasificación.

Para el caso de la segunda evaluación, se creó un modelo de clasificación en donde se utiliza únicamente la edad como atributo clasificador. La salida de este modelo puede ser "10s", "20s" o "30s" (Resultados en la tabla 3). Una vez identificado el rango de edad del autor, se procede a identificar el género, de acuerdo a la edad identificada en la fase anterior. Así, se tienen ahora tres modelos de clasificación adicionales, uno que entrena sobre los diálogos escritos por personas en el rango de edad de los 10 años, otro sobre el rango de edad de los 20 años, y el último sobre el rango de edad de los 30 años.

Tabla 5. Resultados de los modelos que usan el género (masculino y femenino) como atributo clasificador en la segunda fase de clasificación

		Español			Inglés		
Modelo	Categoría	**Precisión**	**Recuerdo**	F_1	**Precisión**	**Recuerdo**	F_1
Genero10s	Masculino	0.0000	0.0000	0.0000	0.1912	0.2453	0.2149
	Femenino	0.0000	0.0000	0.0000	0.3069	0.2366	0.2672
Genero20s	Masculino	0.3218	0.3915	0.3532	0.3344	0.3927	0.3612
	Femenino	0.3405	0.2678	0.2998	0.2849	0.2240	0.2508
Genero30s	Masculino	0.2703	0.3372	0.3001	0.3193	0.3769	0.3457
	Femenino	0.2844	0.2128	0.2435	0.3656	0.3055	0.3329
Promedio General		**0.2028**	**0.2015**	**0.2022**	**0.3004**	**0.2968**	**0.2986**

La salida final indica el rango de edad y el género identificado al que pertenece un determinado autor. Si bien, este sistema tiene la misma entrada y la misma salida que el presentado anteriormente, los resultados que se observan en la Tabla 5 muestran el impacto de seleccionar una fase por encima de la otra. Se presentan de nuevo los valores de precisión, recuerdo y F_1 para cada modelo de clasificación aplicado en la segunda fase.

Al comparar los resultados obtenidos con respecto al esquema anterior, se observa una pérdida mayor de precisión en todos los rangos de edad, sin importar el idioma. En particular, para el caso del idioma Español, no se logró distinguir los diálogos escritos por los autores en el rango de edad de los 10 años. Este hecho sugiere de nuevo que los resultados están co-relacionados con la calidad del corpus de entrenamiento, de tal manera que los diálogos no son representativos de la clase y pueden haber generado un modelo sobreajustado, tal y como se mencionó anteriormente.

Al momento, no existe evidencia de que el proceso de clasificación se comporte diferente para cada idioma evaluado. Los resultados obtenidos se corresponden completamente con el número de muestras que tiene cada clase en el corpus de datos.

Finalmente, se decidió investigar el comportamiento del proceso de clasificación usando un modelo único, es decir, una sola fase en donde existen 6 clases distintas asociadas a las diferentes categorías que presenta el corpus: masculino en el rango de 10 años, masculino en el rango de 20 años, masculino en el rango de 30 años, femenino en el rango de 10 años, etc. Los resultados obtenidos y su comparación con los dos modelos presentados anteriormente (Género->Edad y Edad->Género) se muestran en la Figura 2.

Al utilizar las 6 clases se evita arrastrar errores de una fase a otra , sin embargo, existe el problema del incremento en el número de clases. El clasificador tiene mayor dificultad para discernir entre las diferentes posibilidades. Si a este problema, se le agrega el hecho de que las clases no están balanceadas, entonces se obtiene el comportamiento mostrado en la Figura 3. Básicamente, la mejor aproximación que se pudo tener en los experimentos presentados en este trabajo fue cuando se usó un proceso de clasificación en dos fases, identificando primero el género y después el rango de edad del autor. En función de estos resultados obtenidos, se ha decidido usar esta aproximación para la evaluación del corpus de prueba en la aproximación "software-submission".

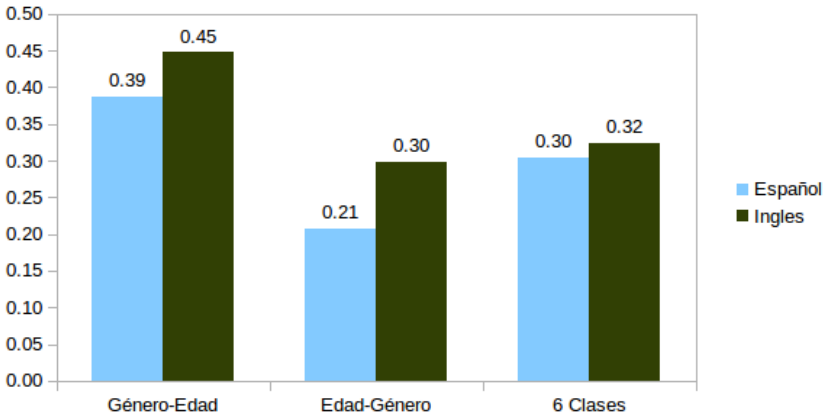

Fig. 2. Comparación entre las aproximaciones de dos fases presentadas y un sólo clasificador con 6 clases

4.2. Software submission

El sistema presentado para la competencia "PAN 2013"(Aleman) obtuvo el lugar 7 con una exactitud de 55.22 % para la detección del género, 59.23 % para la detección de la edad y una exactitud general de 32.92 %. Como se puede observar en la tabla 6.

Tabla 6. Resultados obtenidos por la aproximación de dos fases (Género-Edad) en la competencia PAN 2013 y comparados con otras aproximaciones

Team	Total	Gender	Age
Meina	0.3894	0.5921	0.6491
Pastor L.	0.3813	0.5690	0.6572
Seifeddine	0.3677	0.5816	0.5897
Santosh	0.3508	0.5652	0.6408
Yong Lim	0.3488	0.5671	0.6098
Ladra	0.3420	0.5608	0.6118
Aleman	**0.3292**	**0.5522**	**0.5923**
Gillam	0.3268	0.5410	0.6031
Kern	0.3115	0.5267	0.5690
Cruz	0.3114	0.5456	0.5966
Pavan	0.2843	0.5000	0.6055
Caurcel Diaz	0.2840	0.5000	0.5679
H. Farias	0.2816	0.5671	0.5061
Jankowska	0.2814	0.5381	0.4738
Flekova	0.2785	0.5343	0.5287
Weren	0.2564	0.5044	0.5099
Sapkota	0.2471	0.4781	0.5415
De-Arteaga	0.2450	0.4998	0.4885
Moreau	0.2395	0.4941	0.4824
baseline	0.1650	0.5000	0.3333
Gopal Patra	0.1574	0.5683	0.2895
Cagnina	0.0741	0.5040	0.1234

Estos resultados son bastante similares a los obtenidos en las pruebas con el conjunto de entrenamiento y aunque de manera global la exactitud no supera el 50 %, ningún equipo participante alcanzó este valor, quedando el primer lugar con 38.9 % de exactitud.

Esta misma aproximación fue evaluada en el marco de la competencia con la finalidad de detectar conversaciones en las que intervienen pederastras. Los resultados obtenidos nos colocan en el lugar 12, de 21 equipos participantes. De hecho, existen 8 equipos que detectan al 100 % el género de los depredadores, sin embargo, este resultado no es tan significativo dado que es bastante conocido que la mayoría de los depredadores sexuales son del sexo masculino, lo cual también se refleja en el conjunto de datos de la competencia. Nuestra aproximación obtuvo un 74 % de exactitud en la detección del género, lo cual consideramos es consistente con los resultados que se han reportado a lo largo de este trabajo. La conclusión es que se debe afinar el conjunto de características de acuerdo al tipo de atributo clasificador, ya que no es lo mismo clasificar género que edad. En el género influye el carácter de las personas y los hábitos de comunicación entre hombres y mujeres, pero en el caso de la edad, el tamaño del vocabulario, por ejemplo, debería ser significativo, puesto que las personas suelen incrementar y modificar el vocabulario a medida que crecen.

5. Conclusiones y trabajo futuro

En este artículo se presentó una metodología para la detección de género y edad en blogs y chats. Esta metodología usa conteos de características léxicas, sintácticas y semánticas para representar los diálogos de las personas con la finalidad de entrenar un modelo de clasificación supervisado que permita determinar el género y el rango de edad de un autor.

Los resultados obtenidos utilizando solamente el conjunto de entrenamiento muestran que existe una mejor clasificación en género que en edad, sin embargo, en ninguno de los dos casos se supera el 55 % de medida F_1, además, esta medida baja cuando se unen los dos clasificadores, llegando a un valor F_1 entre 40 % y 44 %.

Las métricas muestran valores más altos para el idioma inglés, que para el idioma español, lo cual parece estar directamente relacionado con el número de diálogos para cada idioma. En particular, el número de diálogos para los autores que se encuentran en el rango de edad de los 10 años es muy bajo y por tanto, existe el riesgo de que no sean representativos de la clase en la vida real. Consideramos así, que la calidad del corpus tuvo un efecto negativo en el proceso de clasificación, sobreajustando los datos de entrenamiento y generando un modelo que no puede reconocer adecuadamente los datos de prueba, especialmente para el conjunto de diálogos que fueron escritos por autores en el rango de edad de los 10 años.

Aunque los resultados de la competencia no sobrepasan el 50 %, a nivel mundial se logró una buena posición respecto al total de equipos participantes. Por lo tanto, aún falta mucho en investigaciones para lograr superar los resultados

aquí obtenidos. Con todos los experimentos realizados se puede concluir que es necesario construir un corpus representativo del género, pero principalmente de los rangos de edad, puesto que éstos se encuentran totalmente desbalanceados y, desde nuestro particular punto de vista, pocamente representados, principalmente para el caso de aquellos escritos por autores en el rango de edad de los 10 años.

Es preciso hacer un análisis lingüístico profundo para determinar las características representativas para cada modelo de clasificación. En este sentido, se considera firmemente que estas características serán diferentes cuando se busque identificar el género de una persona que cuando se intente determinar su rango de edad.

Referencias

1. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006. (2006) 199–205
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM **52**(2) (February 2009) 119–123
3. Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting age and gender in online social networks. In: Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents. SMUC '11, New York, NY, USA, ACM (2011) 37–44
4. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "how old do you think i am?": A study of language and age in twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. ICWSM 2013 (2013)
5. Rangel, F., Rosso, P.: Use of language and author profiling: Identification of gender and age. In: Proceedings of the 10th Workshop on Natural Language Processing and Cognitive Science (NLPCS-2013). (2013)
6. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK (1994)
7. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Human Language Technology Conference (HLT-NAACL 2003). (2003)
8. Breiman, L.: Random forests. Mach. Learn. **45**(1) (October 2001) 5–32