Importancia del lenguaje coloquial y de los símbolos de puntuación en el perfilado de autores

Diana M. Sepúlveda-Barrera[1], Daniel Martínez-Espino[1],
Esaú Villatoro-Tello[2], Gabriela Ramírez-de-la-Rosa[2]

[1] Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa,
Maestría en Diseño, Información y Comunicación (MADIC),
División de Ciencias de la Comunicación y Diseño,
México

[2] Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa,
Departamento de Tecnologías de la Información,
México

{dcg.disb,damaes83}@gmail.com, {evillatoro,gramirez}@correo.cua.uam.mx

Resumen. En años recientes el perfilado de autores (PA) se ha convertido en una tarea muy relevante para la comunidad de Procesamiento del Lenguaje Natural (PLN). El objetivo principal del PA es determinar de forma automática características demográficas del autor, por ejemplo género y edad. En este trabajo presentamos una propuesta para resolver el problema de PA; en particular nos interesó determinar el rol que juega el lenguaje coloquial así como el significado de los distintos símbolos de puntuación. Contrario a trabajos previos, nuestra propuesta considera cada símbolo de puntuación de manera independiente y no como un solo atributo que abarca todos los símbolos de puntuación. Nuestra hipótesis plantea que el uso de determinados símbolos de puntuación en conjunto con lenguaje coloquial aportan información relevante a un método de clasificación automática. Como contribución adicional de este trabajo, nos dimos a la tarea de compilar un diccionario de lenguaje coloquial, el cual consideramos un recurso valioso para la comunidad de PLN haciendo investigación en áreas afines. Los resultados obtenidos muestran que los atributos propuestos permiten enriquecer positivamente esquemas tradicionales de representación de textos.

Palabras clave: Perfilado de autores, atributos estilísticos, representación de textos, procesamiento de lenguaje natural, aprendizaje supervisado.

Importance of Colloquial Language and Punctuation for Author Profiling

Abstract. In recent years, author profiling (AP) has become a very relevant task for natural language processing (NLP). The main goal of

AP is automatically determine demographic aspects from an author, for example, genre and age. In this paper we present a method for author profiling; particularly, we are interested in determine the rol of colloquial language and the meaning of diverse punctuation marks. Contrary to previous works, our proposal considers each punctuation mark independently and not as a single attribute that covers all marks. Our hypothesis states that the use of certain punctuation marks together with the use of colloquial language can provide relevant information to a automatic classification method. As an aditional contribution, we compiled and made available a dictionary with the colloquial words we use in this paper. The obteined results show that the proposed features allow enhance traditional text representation schemas.

Keywords: Author profiling, stylistic features, text representation, natural language processing, supervised learning.

1. Introducción

El perfilado de autor es uno de los retos recientes que ha llamado la atención de la comunidad científica, en particular de áreas como el procesamiento de lenguaje natural, ciencias forenses, estrategias de marketing y seguridad en internet. El objetivo principal del perfilado de autor (PA) es distinguir, a partir de un texto, entre clases de autores y no identificar a un autor en particular, siendo este último el escenario del problema conocido como atribución de autoría [14]. Así entonces, la tarea de PA busca modelar a través de atributos sociolingüísticos más generales a grupos de autores, dichos atributos son además indicadores de cómo los distintos grupos de autores emplean el lenguaje dependiendo de su género, edad y/o lenguaje nativo [1].

Uno de los primeros trabajos que enfrentaron el problema de PA fueron los propuestos en [1,6], donde se mostró, a través de técnicas estadísticas, que el análisis sobre el uso de las palabras en distintos documentos permite determinar el género, edad, idioma nativo e incluso la personalidad del autor. A partir de entonces, muchos trabajos se han propuesto resolver el problema de PA, ejemplos de estas investigaciones son [12,3,10,9,8]. En muchos de estos trabajos se ha enfatizado el uso y análisis de representaciones textuales, las cuales han mostrado ser bastante eficientes cuando los documentos que se quieren clasificar son escritos formales (*e.g.,* artículos de noticias, libros, etc.). Sin embargo, cuando se trata de textos informales (*e.g.,* blogs, chats, tuits), las representaciones tradicionales tienen problemas determinando el perfil de los autores. Esto se debe en gran parte a la dificultad que representa analizar textos informales, los cuales contienen frecuentemente muchos errores ortográficos, variado uso de *jerga* específica de los medios sociales, así como el uso excesivo de emoticonos. A partir de esto, surge la idea de utilizar atributos estilísticos en combinación con representaciones tradicionales para resolver el problema de PA en medios sociales [12].

En este trabajo proponemos una representación enriquecida que contempla dos aspectos: *i)* el uso del lenguaje coloquial (*i.e.,* jerga) en los textos; y *ii)* el uso y significado de distintos símbolos de puntuación. Nuestra hipótesis plantea que la combinación del uso de lenguaje coloquial junto con el uso de distintos símbolos de puntuación, considerando la frecuencia de uso de ambos, podría ser un factor importante al momento de distinguir el perfil del autor. Agregado a esto, se hizo la compilación de un diccionario coloquial, el cual es el resultado de identificar la *jerga* más comúnmente empleada en distintos y variados medios sociales. Este diccionario es un recurso valioso para la comunidad de PLN haciendo investigación en el área de PA.

El resto de este documento se encuentra organizado de la siguiente manera. En la sección 2 se describen algunos de los trabajos relacionados al problema de perfilado de autor. En la sección 3 se describe en detalle el método empleado en el perfilado de autor, así como los atributos estilísticos propuestos. En la sección 4 se describe la metodología experimental y los resultados obtenidos con el método propuesto. Finalmente, en la sección 5 se plantean las conclusiones obtenidas y se describen algunas lineas de trabajo futuro.

2. Trabajo relacionado

El problema del perfilado de autor ha sido atractivo para diferentes áreas de conocimiento, por ejemplo la psicología, lingüística, socio-lingüística y el procesamiento del lenguaje natural [6,12]. Tradicionalmente, como se describe en [8], el perfilado de autor involucra: *i)* la identificación y extracción de atributos textuales, *ii)* la construcción de una representación apropiada, por ejemplo bolsa de palabras (BOW[3]), y *iii)* la construcción de un modelo de clasificación, el cual es entrenado para identificar los perfiles de interés (*e.g.,* género o edad).

Dado lo anterior, uno de los trabajos que busca hacer detección de edad en conversaciones en línea, con el objetivo de identificar depredadores sexuales es el descrito en [15]. La hipótesis de los autores plantea que si un sistema de PA es capaz de detectar cuando un usuario adulto se quiere hacer pasar por un niño o adolescente, el problema del acoso sexual en internet se podría ver disminuido. En dicho trabajo se utilizó la biblioteca de NLTK [7] junto con el corpus NPS Chat Corpus, datos que reúnen textos provenientes de diferentes servicios de conversaciones en línea. Para probar su método, los autores formaron seis diferentes clases: adolescentes (13-19), adultos (20-59), 20s, 30s, 40s y 50s. Los atributos empleados fueron el uso de n-gramas de palabras, emoticones, tokens de puntuación (signos de puntuación), longitud promedio de oración y la cantidad de palabras promedio por documento. En los resultados obtenidos, las pruebas con SVM fueron siempre mejores, principalmente en el problema de identificar entre adolescentes *vs.* adultos (*i.e.* problema binario).

Por otro lado, en [5] se realiza un análisis estadístico en blogs para identificar variaciones del lenguaje dependiendo de la edad y del género. Los resultados

[3] Por sus siglas en Inglés: *Bag-of-Words*

obtenidos se basaron en dos atributos independientes; el primero, es el uso de lenguaje coloquial, y el segundo, en la longitud promedio de las oraciones según los grupos edad y género. Las pruebas se realizaron con un conjunto de 20,000 blogs como corpus, según los resultados reportados, la detección de género fue más acertada que la de edad. En general, el trabajo descrito en [5] se desprende de muchas de las ideas planteadas en [1,6].

En el trabajo descrito en [16] se utilizan 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. Otro trabajo que se aproxima al método propuesto en este artículo es el descrito en [11], donde los autores crearon un diccionario de emoticones. En sus experimentos, los autores muestran que es posible distinguir el genero de los autores a través de contar el número de emoticones empleados.

Más recientemente, el trabajo descrito en [8] se enfoca en proponer una representación vectorial comprimida (*i.e.*, pocas dimensiones) para eliminar el problema de la alta dimensionalidad que significa el trabajar con representaciones tipo BOW. Para esto, los autores proponen la construcción de sub-perfiles, donde la idea principal es capturar los elementos más discriminativos a nivel de intra-perfiles. Al final esta nueva representación es empleada en el esquema general de PA para entrenar un clasificador que permita identificar género y edad.

A diferencia del trabajo descrito previamente, en esta investigación nos interesa evaluar la importancia que tiene el significado de los diferentes símbolos de puntuación empleados por los autores. En trabajos previos los símbolos de puntuación son considerados indistintamente, es decir como un sólo atributo, así entonces una coma (,) es tratada igual que un punto y coma (;) o que dos puntos (:), etc. Nuestra intuición es que el uso y significado que les dan los autores a estos símbolos en conjunto con el uso de lenguaje coloquial (*jerga*), pueden ser factores relevantes al momento de entrenar un clasificador para identificar el género y el rango de edad de los autores.

3. Método propuesto

3.1. Representación de los documentos

En este trabajo se aborda la problemática de la identificación del perfil de autor aplicando el paradigma de clasificación de textos (CT)[4]. Bajo este paradigma el primer paso consiste en el *indexado* de los documentos de entrenamiento (Tr), esta actividad consiste en mapear un documento d_j a una forma compacta de su contenido. La representación más comúnmente utilizada para representar cada documento es un vector con términos ponderados como entradas, concepto tomado del modelo de espacio vectorial usado en recuperación de información [2]. Es decir, un texto d_j es representado como el vector $\overrightarrow{d}_j = \langle w_{kj}, \ldots, w_{|\tau|j} \rangle$, donde

[4] La Clasificación de Textos es la tarea de asociar automáticamente categorías predefinidas con documentos a partir del análisis de su contenido [13].

τ es el *diccionario*, *i.e.*, el conjunto de términos que ocurren al menos una vez en algún documento de Tr, mientras que w_{kj} representa la importancia del término t_k dentro del contenido del documento d_j.

El peso w_{kj} se puede determinar de distintas maneras, entre las más usadas en la comunidad científica están el ponderado booleano y el ponderado por frecuencia relativa de términos. Una breve descripción es dada a continuación:

- *Ponderado Booleano:* Consiste en asignar el peso con valor de 1 si la palabra ocurre en el documento y 0 en otro caso:

$$w_{kj} = \begin{cases} 1, \text{ si } t_k \in d_j \\ 0, \text{ en otro caso.} \end{cases} \tag{1}$$

- *Ponderado por frecuencia relativa (TF-IDF):* Este tipo de ponderado es una variación del tipo anterior y se calcula de la siguiente forma:

$$w_{kj} = TF(t_k) \times IDF(t_k), \tag{2}$$

donde $TF(t_k)$ es la frecuencia del término t_k en el documento d_j. IDF es conocido como la "frecuencia inversa" del término t_k dentro del documento d_j. El valor de IDF es una manera de medir la "rareza" del término t_k. Para calcular el valor de IDF se utiliza la siguiente ecuación:

$$IDF(t_k) = \log \frac{|D|}{|\{d_j \in D : t_k \in d_j\}|}, \tag{3}$$

donde D representa la colección de documentos que está siendo indexada.

3.2. Atributos de estilo

Como estrategia para enriquecer la representación BOW se decidió considerar la presencia de 8 atributos estilísticos. A continuación se listan los atributos de estilo contemplados:

- LINK - indica la frecuencia de uso de URLs en los documentos en revisión
- EMOTICON - indica la frecuencia del uso de emoticonos, para esto se utilizó un diccionario de los emoticonos más comunes (Sección 3.3).
- IMG - representa la presencia de imágenes en los documentos.
- PUNTUATRESP - indica la frecuencia de uso de puntos suspensivos (...)
- PUNTUAP - refiere a la aparición de puntos separadores de oraciones y/o párrafos (.).
- PUNTUAC - indica la frecuencia de uso de comas (,).
- PUNTUADOSP - representa la aparición del símbolo de dos puntos (:).
- PUNTUAPC - indica la presencia del símbolo de punto y coma (;).

3.3. Diccionario de lenguaje coloquial

Como elemento estilístico adicional se definió el atributo INFORMALEXXX, el cual es un atributo que refleja la cantidad de lenguaje informal/coloquial (*jerga*) empleado por el autor. Para poder calcular este atributo fue necesaria la compilación de un diccionario de lenguaje informal.

Así entonces, el diccionario fue creado haciendo una recopilación de diversas listas de jerga digital en español e inglés (fail, noob, gamer, etc.), símbolos representativos de emoticones (n_n, :D, x_x, etc.), y siglas de uso común (lol, brb, afk, etc.), obtenidos de fuentes como Wikipedia, foros de uso popular como Taringa!, blogs latinoamericanos de análisis de tendencias digitales y de uso personal. El diccionario quedó conformado por 1,178 palabras y símbolos[5].

En la Tabla 1 se puede observar la incidencia en porcentaje de palabras contenidas en el diccionario coloquial digital que aparecen en los textos del corpus utilizado para nuestros experimentos. Como es posible observar, el porcentaje de uso de lenguaje coloquial es notoriamente diferente entre distintos rangos de edades, por lo cual se consideró para hacer las distinciones.

Tabla 1. Porcentaje de incidencia de palabras coloquiales en el corpus

Clase	Porcentaje de aparición
Hombres	2.83 %
Mujeres	2.02 %
10's (11-19)	9.48 %
20's (20-29)	4.55 %
30's (30-39)	2.01 %

3.4. Métodos de clasificación

Dado que nuestra propuesta para identificar perfiles de autores no depende en particular de ningún algoritmo de aprendizaje, podemos emplear prácticamente cualquier clasificador para enfrentar el problema. Para los experimentos realizados seleccionamos dos diferentes algoritmos de aprendizaje, los cuales son algoritmos representativos dentro de la gran variedad de algoritmos de aprendizaje disponibles actualmente en el campo de aprendizaje computacional. Específicamente, consideramos los siguientes:

- **Naïve Bayes(NB).** Método probabilístico que asume la independencia de los atributos entre las diferentes clases del conjunto de entrenamiento.
- **Arboles de decisión (J48)**. Un algoritmo que permite generar un árbol de decisión, el cual selecciona los atributos más discriminativos basándose en su medida de entropía.

[5] El recurso está disponible en: `http://ccd.cua.uam.mx/~evillatoro/Resources/Slang_Dictionary_RCS_2016.txt`

En nuestros experimentos se empleó la implementación de Weka de cada uno de estos algoritmos, donde los parámetros empleados fueron los entregados por defecto por Weka [4]. Es importante mencionar que para todos los experimentos se aplicó como estrategia de validación la técnica de validación cruzada a diez pliegues.

3.5. Evaluación

Para evaluar un sistema de clasificación se utilizan las medidas de *Precisión* y *Recuerdo*, que son medidas comunes en el área de recuperación de información. La precisión (P) es la proporción de documentos clasificados correctamente en una clase c_i con respecto a la cantidad de documentos clasificados en esa misma clase. El recuerdo (R), la proporción de documentos clasificados correctamente en una clase c_i con respecto a la cantidad de documentos que realmente pertenecen a esa clase. Así, la precisión se puede ver como una medida de la corrección del sistema, mientras que el recuerdo da una medida de cobertura o completitud.

Adicionalmente, es común emplear la medida-F para describir el comportamiento de la clasificación, la cual se define como:

$$medida - F = \frac{(1+\beta^2)Precision * Recuerdo}{\beta^2 Precision + Recuerdo}, \tag{4}$$

donde con $\beta = 1$ representa la media armónica entre la precisión y el recuerdo. La función de β es la de controlar la importancia relativa entre las medidas de precisión y recuerdo. Es común asignar un valor de 1 indicando igual importancia a ambas medidas.

4. Resultados experimentales

4.1. Conjunto de datos

Para la realización de nuestros experimentos trabajamos con datos extraídos del corpus PAN-2013[6], el cual está compuesto por documentos en inglés y en español. Nuestro trabajo se enfocó a realizar las pruebas con los textos en español. El corpus contiene documentos de 75,900 autores diferentes, ambos géneros y distintos segmentos de edad (37,950 documentos para cada género; 2,500 textos en el segmento de 11 a 19 años de edad, que denominaremos 10's; 42,600 textos en el segmento de 21 a 29 años de edad, que denominaremos 20's; 30,800 documentos en el segmento de 31 a 39 años de edad, que llamaremos 30's).

En general, los documentos que conforman el corpus pertenecen a distintas tipologías de texto, como entradas de blogs, mensajes de foros, conversaciones de chat, anuncios, noticias, artículos, por mencionar algunos, en donde todos los autores tienen como enlace común el idioma (incluso pertenecen a distintas

[6] `http://pan.webis.de/clef13/pan13-web/author-profiling.html`

nacionalidades). Para realizar nuestros experimentos, fue necesario reducir la muestra de datos, esto principalmente debido a las limitaciones en cuanto a poder de cómputo se refiere. En la Tabla 2 podemos observar como quedó conformado el corpus final. En la tabla se muestra información referente al número de documentos por clase (Num. documentos), el tamaño promedio de cada documento (medido en tokens), tamaño promedio de tokens (medido en caracteres); tamaño promedio de oraciones (cantidad de tokens por oracion); y diversidad léxica (número de veces promedio que se utilizan una palabra en todo el documento).

Tabla 2. Estadísticas de los documentos en el corpus empleado

	Género		*Edad*		
	Hombres	Mujeres	10's	20's	30's
Num. documentos	300	300	200	200	200
Tamaño documentos	1984.05	1840.38	1309.09	2171.74	1677.87
Tamaño tokens	5.52	5.50	5.82	5.48	5.54
Tamaño oraciones	98.66	96.56	79.92	95.34	92.28
Diversidad léxica	1.35	1.35	1.28	1.40	1.35

4.2. Resultados del método base

Como se mencionó en secciones anteriores, como método base se empleó una forma de representación de bolsa de palabras. Bajo este esquema, la dimensionalidad del vector de atributos es de 19,275 atributos. La Tabla 3 muestra el desempeño obtenido para los problemas de identificación de género y edad respectivamente.

Tabla 3. Resultados de clasificación de género y edad empleando una bolsa de palabras

Pesado - Algoritmo	*Género*			*Edad*		
	Precisión	*Recuerdo*	*F-score*	*Precisión*	*Recuerdo*	*F-score*
BOOL-NB	**0.571**	**0.567**	**0.56**	**0.442**	0.415	0.39
BOOL-J48	0.549	0.548	0.54	0.419	**0.422**	**0.42**
TF-IDF-NB	0.554	0.553	0.55	0.412	0.410	0.40
TF-IDF-J48	0.534	0.533	0.53	0.381	0.383	0.38

Se puede observar que en general un esquema de pesado booleano permite a los clasificadores obtener un mejor desempeño en términos de *precisión*, lo cual significa que la presencia/ausencia de ciertos términos es un factor importante al resolver el problema de perfilado.

Al hacer un análisis sobre los atributos con mayor ganancia de información para el problema de identificación de género, observamos que los atributos más relevantes contienen una carga emocional, por ejemplo: *Sonrisa, reales, diez, ley, familias, tierras, inlove, dy, letras, lamento*. Por otro lado, para el caso de la identificación de edad los atributos con mayor ganancia de información fueron términos asociados a la jerga empleada en medio sociales, por ejemplo: *k, hora, fr, go, qe, qu, errores, contar, super, spero*. A partir de este análisis, se decidió realizar un experimento adicional, el cual consistió en aplicar como técnica de reducción de dimensionalidad la estrategia de ganancia de información. Tras este proceso, se realizaron nuevamente los experimentos empleando el método base más ganancia de información (IG). Los resultados obtenidos bajo este esquema aparecen en la Tabla 4

Tabla 4. Resultados de clasificación de género y edad empleando una bolsa de palabras con ganancia de información

Pesado - Algoritmo	*Género*			*Edad*		
	Precisión	*Recuerdo*	*F-score*	*Precisión*	*Recuerdo*	*F-score*
BOOL-NB	0.678	**0.645**	**0.63**	0.543	**0.498**	**0.48**
BOOL-J48	0.589	0.555	0.51	0.474	0.445	0.42
TF-IDF-NB	**0.729**	0.597	0.53	0.532	0.428	0.40
TF-IDF-J48	0.644	0.532	0.42	**0.562**	0.403	0.33

Como es posible observar, la reducción de dimensionalidad a través de IG permite al clasificador mejorar significativamente su desempeño en términos de la medida F. Es particularmente notorio el desempeño de un esquema de pesado booleano empleando un clasificador bayesiano (BOOL-NB).

4.3. Resultados del método propuesto

En los siguientes experimentos se reportan los resultados obtenidos al hacer la extensión de la BOW por medio de incorporar los atributos estilísticos definidos en la sección 3, los cuales capturan las frecuencias de uso de emoticones, links, diferentes signos de puntuación y vocabulario coloquial digital.

Los resultados de los experimentos realizados en esta sección están descritos en la Tabla 5 y Tabla 6 para cuando no se aplica ganancia de información y cuando si se hace uso de IG respectivamente.

Como se puede observar, la representación propuesta, sin emplear ganancia de información, funcionó mejor para el caso de identificación de edad. Nótese que para el caso de género, los resultados son muy similares a los obtenidos con el método base (Tabla 3). Por el contrario, para el problema de identificación de edad, se pudo lograr una pequeña mejora en los resultados, de aproximadamente un 1 % relativo (Tabla 3 *vs.* Tabla 5).

Tabla 5. Resultados de clasificación de género y edad empleando una bolsa de palabras extendida con los atributos estilísticos propuestos

Pesado - Algoritmo	*Género*			*Edad*		
	Precisión	*Recuerdo*	*F-score*	*Precisión*	*Recuerdo*	*F-score*
BOOL-NB	**0.573**	**0.568**	**0.56**	**0.446**	0.418	0.39
BOOL-J48	0.533	0.533	0.53	0.422	0.423	0.42
TF-IDF-NB	0.540	0.540	0.53	0.436	**0.433**	**0.43**
TF-IDF-J48	0.543	0.543	0.54	0.399	0.402	0.39

Tabla 6. Resultados de clasificación de género y edad empleando una bolsa de palabras extendida con los atributos estilísticos propuestos y utilizando ganancia de información

Pesado - Algoritmo	*Género*			*Edad*		
	Precisión	*Recuerdo*	*F-score*	*Precisión*	*Recuerdo*	*F-score*
BOOL-NB	0.703	**0.698**	**0.70**	0.440	0.440	0.43
BOOL-J48	0.606	0.605	0.60	0.457	0.448	0.44
TF-IDF-NB	**0.773**	0.600	0.53	**0.557**	**0.463**	**0.44**
TF-IDF-J48	0.737	0.600	0.53	0.473	0.460	0.44

A pesar de la mejora mínima en el desempeño del clasificador para el caso de identificación de edad, un análisis sobre los atributos con mayor ganancia de información arrojó que entre los atributos con mayor relevancia están los siguientes: *k, hora, PUNTUAC, fr, go, qe, qu, PUNTUATRESP, INFORMALEXXX, errores, EMOTICON*. De manera similar, para el caso de identificación de género, los atributos con mayor ganancia de información fueron: *LINK, sonrisa, diez, reales, EMOTICON, ley, familias, dy, canto y lamento*. Este análisis indica, hasta cierto punto, que los atributos propuestos para enriquecer la bolsa de palabras son efectivamente discriminativos en ambos problemas. Dado que los atributos propuestos aparecieron entre los atributos con mayor IG, se replicaron los experimentos empleando la bolsa de palabras enriquecida aplicando IG como estrategia de reducción de dimensionalidad. Los resultados obtenidos se muestran en la Tabla 6.

Como es posible observar, para el caso de identificación de género se logró un 70 % en la medida F, lo cual representa una ganancia significativa en comparación con emplear la BOW enriquecida sin aplicar IG (Tabla 5). Para el caso de identificación de edad es posible obtener mejoras mínimas en general.

Como pruebas adicionales se realizó una serie de experimentos en los cuales no se hizo distinción entre los diferentes símbolos de puntuación, es decir, se consideraron los atributos descritos en la sección 3 como un solo atributo en el proceso de representación de los datos[7]. Así entonces, para la tarea de identifica-

[7] Esta representación corresponde a lo que normalmente se ha aplicado en trabajos previos (Ver sección 2)

ción de género, se replicó la mejor configuración de la Tabla 6 (*i.e.*, BOOL-NB), y los resultados obtenidos fueron: $F = 0.48$, $P = 0.53$ y $R = 0.50$. Note que el desempeño decrece en comparación con los resultados de la Tabla 6, lo cual muestra la pertinencia de hacer la distinción entre símbolos de puntuación para el problema de identificación de *género*. De manera similar, para el problema de identificación de edad se tomó la mejor configuración lograda de la Tabla 6 (*i.e.*, TF-IDF-NB), para este caso los resultados fueron: $F = 0.40$, $P = 0.41$ y $R = 0.40$. Nuevamente, el desempeño del clasificador empeora, evidenciando la efectividad de hacer una caracterización individualizada de los símbolos de puntuación.

4.4. Discusión

Como se mencionó en la sección anterior, los atributos estilísticos propuestos para enriquecer la representación BOW mostraron tener mayor pertinencia en el problema de identificación de edad. Técnicas de selección de atributos (*e.g.,* IG) identifican como características relevantes, *i.e.,* asignan un *rank* alto, a un subconjunto importante de los atributos propuestos. A continuación (Tabla 7) mostramos un ejemplo que refleja el uso distinto que diferentes autores, de diferentes rangos de edad, dan a los símbolos de puntuación en sus textos.

De estos ejemplos se puede notar una tendencia hacia el uso indiscriminado de símbolos de puntuación entre más joven es el autor (*e.g.* autores del rango 10's). Por el contrario entre mayor edad tiene el autor, tiende a hacer un uso más ordenado y correcto de dichos símbolos, además de que hace uso de diversos símbolos y no unos cuantos, que es el caso de los mas jóvenes.

En general, el análisis realizado indica que el hacer distinción entre los símbolos de puntuación empleados es un elemento relevante para sistemas de PA. Contrario a trabajos anteriores, nosotros consideramos que hacer una distinción explícita de los distintos símbolos de puntuación es importante debido a que distintos autores tienden a utilizarlos de formas distintas, tal y como se muestra en la Tabla 7. Agregado a esto, la secuencia en que aparecen los mismos podría ser otro factor atractivo a considerar como un atributo adicional en un sistema de perfilado de autor.

5. Conclusiones

En este trabajo hemos descrito nuestro método para enfrentar el problema del perfilado de autor. Nuestro objetivo principal fue determinar la pertinencia del uso del lenguaje coloquial *digital* en conjunto con distintos símbolos de puntuación para determinar el perfil de autor. Contrario al trabajo previo, nosotros hacemos una distinción entre símbolos de puntuación, bajo el supuesto de que varios autores los utilizan con un significado, forma y frecuencias diferentes.

Para la realización de nuestros experimentos se tomó una muestra aleatoria de datos extraídos de la competencia del PAN-2013. Los documentos de dicha competencia son en su mayoría datos obtenidos de medios sociales, lo cual

Tabla 7. Porciones de textos extraídos del corpus utilizado que reflejan la diversidad de uso que se les da a los distintos símbolos de puntación

Muestra de texto de documento perteneciente a clase 10's
niñas lo mejor que me ha pasado en la vida juntas por siempre nenas sois mi vida enteraa os adoroo encerioo no cambieis cada una de vosotras es unica de difernte maneraaa! oOOOooOOoOOooO PUNTUADOSP
Muestra de texto de documento perteneciente a clase 20's
Me convertí en adulto y hasta hoy cuando veo a una persona hablando demasiado PUNTUAC interrumpiendo la conversación de todos PUNTUAC siendo inoportuna o violenta PUNTUAC presumiendo de lo que tiene PUNTUAC sintiéndose prepotente y haciendo de menos a la gente PUNTUAC tengo la impresión de oír la voz de mi padre diciendo PUNTUADOSP "Cuanto más vacía la carreta PUNTUAC mayor es el ruido que hace"
Muestra de texto de documento perteneciente a clase 30's
Quizás sea la rama poética la que más siente el gallego-berciano PUNTUAP Pero también hay que comprender que los poetas utilizan el código escrito que más dominan PUNTUAC el castellano PUNTUAC porque es el que recibieron durante su formación educativa PUNTUAP Sólo los poetas más populares PUNTUAC en el sentido de cercanos a su tierra PUNTUAC se atreven con el gallego PUNTUAP Este es el caso paradigmático de Gilberto Núñez Ursinos PUNTUAC tan concienciado por la poesía social PUNTUAC "déixame tamén a min escuitar PUNTUAC oilo aturuxo os teus fillos que ben cedo emigraron" PUNTUAP Por su parte PUNTUAC el poeta Antonio Bouza Pol declara con firmeza que "el gallego es nuestra segunda lengua PUNTUAC y esta realidad antiquísima tiene que ser conocida PUNTUAC respetada y querida por toda la Comunidad de Castilla y PUNTUAC más aún PUNTUAC por León"

representa un reto importante para los sistemas automáticos de PA debido a la presciencia de errores ortográficos, el uso frecuente de jerga específica de medios sociales, así como el uso de emoticones. Los resultados obtenidos de nuestros experimentos muestran que el método propuesto es pertinente para el problema de identificación de edad.

Un análisis realizado de ganancia de información sobre el vector de atributos reflejó que 5 de los 9 atributos estilísticos propuestos aparecen entre los atributos más relevantes para resolver el problema de PA. Experimentos realizados muestran que, tras hacer una reducción de dimensionalidad empleando la estrategia de IG, el desempeño del clasificador mejora significativamente. Estos resultados nos motivan a seguir trabajando sobre esta línea de investigación. Agregado a esto, como una contribución adicional de nuestro trabajo se conformó, a través

de distintas fuentes, un diccionario de lenguaje coloquial digital, el cual contiene una gran variedad de términos de la *jerga* más comúnmente empleada en medios sociales. Este recurso lingüístico lo consideramos de gran importancia para la comunidad científica de PLN trabajando en áreas afines.

Como trabajo futuro nos proponemos hacer pruebas con una muestra mayor, de forma que podamos validar los hallazgos hasta ahora encontrados. También planeamos hacer una representación que considere la secuencia de aparición de los símbolos de puntuación, por ejemplo n-gramas de símbolos de puntuación. Finalmente, una definición más fina de atributos asociados a símbolos de puntuación podría resultar en un beneficio mayor del sistema de clasificación.

Agradecimientos. Este trabajo fue parcialmente financiado por el CONACyT a través de las becas 708534 y 717783, el proyecto de investigación No. 258588 y programa del SNI. También se agradece el apoyo otorgado a través de la Coordinación de la Maestría en Diseño, Información y Comunicación (MADIC) de la UAM Cuajimalpa.

Referencias

1. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. TEXT 23, 321–346 (2003)
2. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
3. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1301–1309. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2145432.2145568
4. Garner, S.R.: Weka: The waikato environment for knowledge analysis. In: In Proc. of the New Zealand Computer Science Research Students Conference. pp. 57–64 (1995)
5. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers' age and gender. In: Third International AAAI Conference on Weblogs and Social Media (2009)
6. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002), http://llc.oxfordjournals.org/content/17/4/401.abstract
7. Loper, E., Bird, S.: Nltk: The natural language toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1. pp. 63–70. ETMTNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), http://dx.doi.org/10.3115/1118108.1118117
8. López-Monroy, A.P., y Gómez, M.M., Escalante, H.J., Villaseñor-Pineda, L., Stamatatos, E.: Discriminative subprofile-specific representations for author profiling in social media. Knowledge-Based Systems 89, 134 – 147 (2015), http://www.sciencedirect.com/science/article/pii/S0950705115002427

9. 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. AAAI Press (2013), reporting year: 2013
10. 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. pp. 37–44. SMUC '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2065023.2065035
11. Rangel, F.: Author profile in social media: Identifying information about gender, age, emotions and beyond. In Proceedings of the 5th BCS IRSG Symposium on Future Directions in Information Access pp. 58–60 (2013), http://ewic.bcs.org/upload/pdf/ewic_fdia13_paper14.pdf
12. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. vol. 6, pp. 199–205 (2006)
13. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. 34(1), 1–47 (Mar 2002), http://doi.acm.org/10.1145/505282.505283
14. Stamatatos, E.: A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol. 60(3), 538–556 (Mar 2009), http://dx.doi.org/10.1002/asi.v60:3
15. Tam, J., Martell, C.H.: Age detection in chat. In: Semantic Computing, 2009. ICSC '09. IEEE International Conference on. pp. 33–39 (Sept 2009)
16. Yuridiana Alemán, D.V., Pinto, D.: Una metodología para la detección del perfil de un autor. Avances en la Ingeniería del Lenguaje y del Conocimiento 93 (2014), http://www.sciencedirect.com/science/article/pii/S0950705115002427