Propuesta recibida el mayo 2016 y aceptada para publicacin el julio 2016.

Perfilado de autor multilingüe en redes sociales a partir de n-gramas de caracteres y de etiquetas gramaticales

Social Network Multilingual Author Profiling using character and POS n-grams

Carlos-Emiliano González-Gallardo
LIA-Université d'Avignon,
GIL-Instituto de Ingeniería UNAM
carlos.gonzalez-gallardo@alumni.univ-avignon.fr

Juan-Manuel Torres-Moreno
LIA-Université d'Avignon,
École Polytechnique de Montréal
juan-manuel.torres@univ-avignon.fr

Azucena Montes Rendón
CENIDET
amr@cenidet.edu.mx

Gerardo Sierra
GIL-Instituto de Ingeniería UNAM
gsierram@iingen.unam.mx

Resumen

En este artículo presentamos un algoritmo que combina las características estilísticas representadas por los n-gramas de caracteres y los n-gramas de etiquetas gramaticales (POS) para clasificar documentos multilengua de redes sociales. En ambos grupos de n-gramas se aplicó una normalización dinámica dependiente del contexto para extraer la mayor cantidad de información estilística posible codificada en los documentos (emoticonos, inundamiento de caracteres, uso de letras mayúsculas, referencias a usuarios, ligas a sitios externos, *hashtags*, etc.). El algoritmo fue aplicado sobre dos corpus diferentes: los *tweets* del corpus de entrenamiento de la tarea *Author Profiling* de PAN-CLEF 2015 (Rangel et al., 2015) y el corpus de "Comentarios de la Ciudad de México en el tiempo" (CCDMX). Los resultados presentan una exactitud muy alta, cercana al 90 %.

Palabras clave

Minería de textos, Aprendizaje automático, Clasificación textual, n-gramas, Blogs, Tweets, Redes sociales

Abstract

In this paper we present an algorithm that combines the stylistic features represented by characters and POS n-grams to classify social network multilingual documents. In both n-gram groups a dynamic normalization by context was applied to extract all the possible stylistic information encoded in the documents (emoticons, character flooding, capital letters, references to other users, hyperlinks, hashtags, etc.). The algorithm was applied to two different corpus; *Author Profiling* of PAN-CLEF 2015 training tweets (Rangel et al., 2015) and the corpus of "Comments of Mexico City in time" (CCDMX). Results shows up to 90 % of accuracy.

Keywords

Text Mining, Machine Learning, Text Classification, n-grams, Blogs, Tweets, Social Networks

1 Introducción

La clasificación automática de texto se encarga de predecir de forma automática a cuál de las clases existentes pertenece un texto. Este modelo es creado a partir de un corpus etiquetado que contenga ejemplos de esas clases (Koppel et al., 2002).

A diferencia de la identificación de autor, que tiene como objetivo predecir si un texto pertenece o no a un autor específico, el perfilado de autor tiene como objetivo predecir si un texto pertenece o no a un grupo de autores que comparten ciertas características; como el género, la edad, el nivel educativo, la región geográfica, etc.

El interés por el perfilado de autor a partir de textos procedentes de Internet ha ido creciendo en los últimos años. Esto es debido a la gran cantidad de información que se produce continuamente en las redes sociales y los blogs. En marzo de 2016, Facebook reportó tener aproximadamente 1 090 millones de usuarios activos al día[1]; mientras que Twitter[2] 320 millones de usuario activos al mes.

Los documentos textuales producidos por los usuarios de estas redes, tienen características que los hacen difícilmente comparables con los textos literarios, documentales o ensayos en donde tradicionalmente el perfilado de autor es aplicado (Argamon et al., 2003, 2009); evitando así que

[1] http://www.facebook.com
[2] http://www.twitter.com

puedan ser analizados de forma similar (Peersman et al., 2011).

Dentro de las características que poseen los textos procedentes de Twitter y redes sociales, se encuentra su longitud, que es notablemente más corta (Peersman et al., 2011), el uso no estandarizado de mayúsculas y signos de puntuación, el gran número de errores ortográficos, etc.

Las redes sociales como Twitter tienen sus propias reglas y características que los usuarios explotan para expresarse y comunicarse entre sí. Estas reglas pueden ser aprovechadas para extraer una mayor cantidad de información estilística. (Gimpel et al., 2011) introducen esta idea para crear un etiquetador gramatical para Twitter. En nuestro caso, optamos por realizar una normalización dinámica dependiente del contexto. Esta normalización permite agrupar aquellos elementos que tengan la capacidad de proveer información estilística sin importar su variabilidad léxica. Esta fase ayuda al sistema de clasifición a mejorar su rendimiento.

El artículo está organizado de la siguiente manera: en la sección 2 hacemos una breve presentación del uso de n-gramas y etiquetas POS. En la sección 3 detallamos la metodología empleada en la normalización dinámica dependiente del contexto. La sección 4 presenta los corpus utilizados en el estudio. El modelo de aprendizaje es detallado en sección 5. Los diversos experimentos realizados y los resultados obtenidos son presentados en la sección 6. Para finalizar, en la sección 7 exponemos las conclusiones y algunas perspectivas de trabajo futuro.

2 N-gramas de caracteres y etiquetas gramaticales (POS)

Los n-gramas son un recurso de gran utilidad en el Procesamiento del Lenguaje Natural (PLN), ya que permiten la extracción de características de contenido y estilísticas a partir de los textos, que pueden ser utilizadas en tareas como resumen automático, traducción automática y clasificación textual.

Los n-gramas son secuencias de elementos de la unidad de información textual seleccionada (Manning & Schütze, 1999). Esta información cambia en función de la tarea a realizar y del tipo de información que se desea extraer. Por ejemplo, en traducción y resumen automático es común utilizar n-gramas de palabras y n-gramas de oraciones (Torres-Moreno, 2014; Giannakopoulos et al., 2008; Koehn, 2010). Dentro de la clasificación de texto, para la detección de plagio e identificación y perfilado de autor, los n-gramas de caracteres, palabras y etiquetas POS (*Part-of-Speech*) son utilizados (Doyle & Kešelj, 2005; Stamatatos et al., 2015; Oberreuter & Velásquez, 2013).

La unidades de información seleccionadas en este trabajo son caracteres y etiquetas POS. Con los n-gramas de caracteres se pretende extraer la mayor cantidad de elementos estilísticos posible: frecuencia de caracteres, uso de sufijos (género, número, tiempos verbales, diminutivos, superlativos, etc.), uso de signos de puntuación (frecuencia de uso, repetición), uso de emoticonos, etc. (Stamatatos, 2006, 2009).

Los n-gramas POS proporcionan información referente a la forma en que está estructurado el texto: la frecuencia de elementos gramaticales, la diversidad de estructuras gramaticales empleadas y la inteacción entre elementos gramaticales. Las etiquetas POS fueron obtenidas usando el etiquetador gramatical de Freeling[3]. Para controlar completamente el proceso de normalización y hacerlo independiente de un detector de nombres propios, preferimos realizar una normalización específica para estos corpus, en lugar de utilizar las funciones de Freeling (Padró & Stanilovsky, 2012).

Una etiqueta POS cuenta con varios niveles de detalle que permiten conocer los diferentes atributos de una categoría gramatical. En nuestro caso únicamente utilizamos el primer nivel de detalle que hace referencia a la categoría en sí misma (ver el cuadro 1).

Palabra: *versión*			
Atributo	**Código**	**Valor**	**Etiqueta**
Categoría	N	Nombre	
Tipo	C	Común	
Género	F	Femenino	
Número	S	Singular	*N*
Caso	0	-	
Género semántico	0	-	
Grado	0	-	

Cuadro 1: Etiquetado gramatical de la palabra *versión*.

3 Normalización dinámica dependiente del contexto

El léxico utilizado en las redes sociales es muy variado debido a la libertad que existe para codificar los mensajes. Para contrarrestar este he-

[3]Freeling está disponible en: `http://nlp.lsi.upc.edu/freeling/node/1`

cho, es necesario normalizar aquellos elementos que tengan la capacidad de proveer información estilística sin importar su variabilidad léxica: referencias a usuarios, ligas a sitios externos y *hashtags*. Este proceso denominado Normalización dinámica dependiente del contexto se separa en dos partes: Normalización del texto y Re-etiquetado POS:

- Normalización del texto

  Es común observar en redes sociales como Twitter las referencias a otros usuarios pertenecientes a la red. Esta referencia está determinada de la siguiente forma:

  `@nombre_de_usuario`

  La cantidad de posibles valores que se le pueden asignar a la etiqueta `nombre_de_usuario` es potencialmente infinita (dependiendo de la cantidad de usuarios de la que disponga la red social). Para evitar tanta variabilidad, decidimos normalizar este elemento con el fin de resaltar la intención de realizar una referencia a un usuario.

  Las ligas a sitios de Internet tienen un comportamiento similar; la cantidad de ligas a estos sitios también es potencialmente infinita. Lo importante y rescatable es el hecho de utilizar un enlace a un sitio externo, por lo que todas las cadenas de texto que cumplen con el patrón:

  `http[s]://liga_sitio_externo`

  también fueron normalizadas.

- Re-etiquetado POS

  Estos elementos también proveen información gramatical importante que es necesario conservar, pero los etiquetadores gramaticales convencionales son incapaces de detectar. Por ello, en nuestro trabajo las referencias a usuarios, las ligas a sitios Internet y los *hashtags* son re-etiquetados de tal forma que se mantenga la interacción de estos elementos con el resto de los elementos gramaticales (ver un ejemplo en el Anexo, cuadro 17).

Una arquitectura general del sistema es mostrada en la figura 1.

4 Conjunto de datos

Con la finalidad de realizar pruebas pertenecientes a diversos contextos, hemos utilizado córpora provenientes de dos redes sociales: Twitter y Facebook. El corpus multilingüe de entrenamiento PAN-CLEF 2015 (Twitter) se encuentra etiquetado por género, edad y rasgos de personalidad. El corpus de "Comentarios de la Ciudad de México en el tiempo" (CCDMX) (comentarios de Facebook) dispone únicamente de etiquetas de género en español.

4.1 Corpus PAN-CLEF (train) 2015

El corpus PAN-CLEF (train) 2015[4] (Rangel et al., 2015) está conformado por un total de 324 muestras distribuidas en cuatro idiomas: español, inglés, italiano y holandés. Cada una de las muestras se compone de aproximadamente 96 *tweets* (Nowson et al., 2015).

Con respecto al género, la distribución del corpus está equilibrada en los cuatro idiomas (50 % como "Mujeres" y 50 % como "Hombres").

	Muestras		**Total**
	Mujeres	**Hombres**	
Español	50	50	100
Inglés	76	76	152
Italiano	19	19	38
Holandés	17	17	34

Cuadro 2: Corpus PAN-CLEF (train) 2015, Distribución de muestras por género.

En el caso de español e inglés las muestras también se encuentran etiquetadas por grupos de edad: 18-24, 25-34, 35-49 y >50 años. En este caso el corpus no está equilibrado, siendo el grupo "25-34" el más numeroso, y el grupo ">50" el que cuenta con el menor número de muestras, en ambos idiomas. Ver cuadro 3.

Grupo		**Español**	**Inglés**
18-24	muestras	22	58
	porcentaje	22 %	38 %
25-34	muestras	46	60
	porcentaje	46 %	40 %
35-49	muestras	22	22
	porcentaje	22 %	14 %
>50	muestras	10	12
	porcentaje	10 %	8 %
	Total muestras	100	152

Cuadro 3: Corpus PAN-CLEF (train) 2015, Distribución de muestras por edad.

Para los cuatro idiomas se cuentan con etiquetas de clases pertenecientes a cinco rasgos de per-

[4]Sitio web del PAN: `http://pan.webis.de/`

sonalidad: extraversión, inestabilidad emocional, amabilidad, responsabilidad y apertura al cambio.

Cada rasgo fue anotado con un valor discreto comprendido entre [-0.5, +0.5] (ver Anexo, cuadro 18).

4.2 Corpus de Comentarios de la Ciudad de México en el tiempo (CCDMX)

El corpus CCDMX está compuesto por 5 979 comentarios en español mexicano, procedentes de la página de Facebook "La Ciudad de México en el tiempo"[5]. La longitud promedio de los comentarios es de 110 caracteres. El corpus CCDMX fue anotado manualmente en el Grupo de Ingeniería Lingüística (GIL) de la UNAM en 2014[6].

El corpus CCDMX se encuentra únicamente etiquetado por género, siendo ligeramente mayor la cantidad de comentarios pertenecientes a la clase "Hombres" (ver cuadro 4).

	Comentarios	**%**
Mujeres	2573	43 %
Hombres	3406	57 %
Total de muestras	5 979	100 %

Cuadro 4: Corpus CCDMX, Distribución de muestras por género.

5 Modelo de aprendizaje

Para los experimentos utilizamos un modelo clásico de aprendizaje supervisado usando *Support Vector Machines* (SVM) (Vapnik, 1998), que ha mostrado ser robusto y eficaz en diversas tareas de PLN.

En particular, para realizar los experimentos empleamos el paquete Python *SciKit Learn*[7], usando un *kernel* lineal LinearSVC (Pedregosa et al., 2011), que produjo empíricamente los mejores resultados.

5.1 Características utilizadas

Las ventanas de *n*-gramas de caracteres y etiquetas POS contempladas fueron generadas con una longitud de 1 a 3 unidades. De esta forma, por ejemplo, la palabra *"versión"* está representada por los siguientes *n*-gramas de caracteres:

{*v, e, r, s, i, ó, n, _v, ve, er, rs, si, ió, ón, n_, _ve, ver, ers, rsi, sió, ión, ón_*}

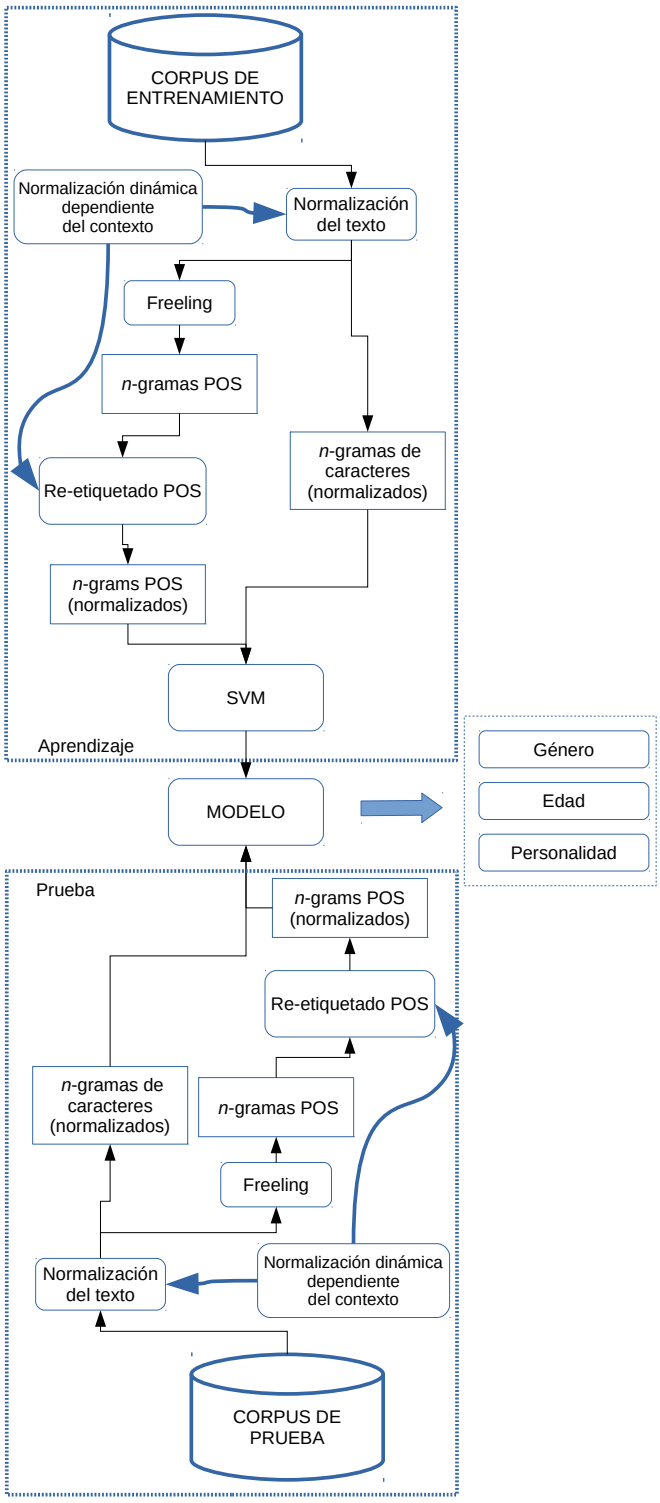

Figura 1: Arquitectura general del sistema de clasificación.

[5]Sitio web del blog: http://www.facebook.com/laciudaddemexicoeneltiempo

[6]Este corpus puede ser solicitado en el sitio web del GIL, en http://corpus.unam.mx

[7]Disponible en el sitio: http://scikit-learn.org

Y la secuencia de etiquetas POS

```
REF@USERNAME V C D N P V REF#LINK
```

está representada por los siguientes n-gramas POS:

```
{REF@USERNAME, V, C, D, N, P, V,
REF#LINK, REF@USERNAME V, V C, C D, D
N, N P, P V, V REF#LINK, REF@USERNAME
V C, V C D, C D N, D N P, N P V, P V
REF#LINK}
```

Una escala lineal de frecuencia es utilizada en todos los casos con excepción de los n-gramas POS para los textos en español, en donde se aplica una función logarítmica del tipo:

$$\log_2(1 + frecuencia) \quad (1)$$

que permite evitar una desviación en los cálculos debido a las grandes frecuencias.

5.2 Protocolo experimental

Cuatro experimentos fueron realizados con el corpus PAN-CLEF (train) 2015, uno por cada idioma. El 70 % de las muestras fue utilizado para entrenar el modelo de aprendizaje y el 30 % durante su evaluación.

Por otro lado, tres experimentos fueron realizados con el corpus CCDMX.

- En primer lugar, el 100 % de los comentarios fueron utilizados como muestras de prueba, utilizando el modelo de aprendizaje generado con las muestras en español de entrenamiento del corpus PAN-CLEF (train) 2015.
- Para el segundo experimento, se crearon muestras de 50 comentarios, juntando así 121 muestras que fueron probadas utilizando el mismo modelo de aprendizaje que el primer experimento.
- Finalmente, el tercer experimento consitió en tomar el 70 % de las 121 muestras para entrenar el modelo de aprendizaje y el 30 % para probar su desempeño.

6 Resultados

Para evaluar el desempeño del sistema en ambos corpus, varias medidas clásicas fueron implementadas:

La exactitud (Ex), precisión (Pr), cobertura (Co) y valor-F (F_1) (Manning & Schütze, 1999) fueron medidos en el corpus CCDMX para evaluar la predicción de género.

En el corpus PAN-CLEF (train) 2015, las mismas medidas fueron utilizadas para evaluar la predicción de género (español, inglés, italiano y holandés) y la edad (español e inglés).

Finalmente, para la evaluación de los rasgos de personalidad en el corpus PAN-CLEF (train) 2015, la medida RMSE (Rangel et al., 2015) fue utilizada.

6.1 Resultados sobre el corpus PAN-CLEF (train) 2015

Los cuadros 5 a 12 presentan los resultados multilingües obtenidos sobre el corpus PAN-CLEF (train) 2015.

Los casos para el experimento en italiano (tabla 9) y para el experimento en holandés (tabla 11) ameritan ser explicado. Las medidas de evaluación reportan 1 en prácticamente todos los casos; esto es debido a que la cantidad de muestras existentes eran muy pocas para italiano y holandés.

Pensamos que valdría la pena probar con una mayor cantidad de datos para validar los resultados en estos dos idiomas.

Español
Las pruebas se realizaron sobre 30 muestras

	Pr	**Co**	F_1	**Ex**
Hombres	0.929	0.867	0.897	0.900
Mujeres	0.875	0.93	0.902	
18-24	0.750	1	0.857	0.800
25-34	0.750	0.875	0.807	
35-49	1	0.667	0.800	
>50	1	0.500	0.667	

Cuadro 5: Corpus PAN-CLEF (train) 2015, Resultados género y edad (español).

Rasgo	**RMSE**
Extraversión	0.106
Inestabilidad emocional	0.128
Amabilidad	0.158
Responsabilidad	0.164
Apertura al cambio	0.138
Promedio	0.139

Cuadro 6: Corpus PAN-CLEF (train) 2015, Resultados rasgos de personalidad (español).

Inglés
Las pruebas se realizaron sobre 46 muestras

	Pr	**Co**	F_1	**Ex**
Hombres	0.826	0.826	0.826	0.826
Mujeres	0.826	0.826	0.826	
18-24	0.895	0.944	0.919	0.848
25-34	0.789	0.833	0.810	
35-49	0.800	0.667	0.727	
>50	1	0.750	0.857	

Cuadro 7: Corpus PAN-CLEF (train) 2015, Resultados género y edad (inglés).

Rasgo	**RMSE**
Extraversión	0.182
Inestabilidad emocional	0.182
Amabilidad	0.150
Responsabilidad	0.123
Apertura al cambio	0.162
Promedio	0.160

Cuadro 8: Corpus PAN-CLEF (train) 2015, Resultados rasgos de personalidad (inglés).

Italiano
Las pruebas se realizaron sobre 12 muestras.

	Pr	**Co**	F_1	**Ex**
Hombres	1	1	1	1
Mujeres	1	1	1	

Cuadro 9: Corpus PAN-CLEF (train) 2015, Resultados género (italiano).

Rasgo	**RMSE**
Extraversión	0.065
Inestabilidad emocional	0.194
Amabilidad	0.091
Responsabilidad	0.100
Apertura al cambio	0.112
Promedio	0.112

Cuadro 10: Corpus PAN-CLEF (train) 2015, Resultados rasgos de personalidad (italiano).

Holandés
Las pruebas se realizaron sobre 10 muestras

	Pr	**Co**	F_1	**Ex**
Hombres	0.833	1	0.901	0.900
Mujeres	1	0.800	0.889	

Cuadro 11: Corpus PAN-CLEF (train) 2015, Resultados género (holandés).

Rasgo	**RMSE**
Extraversión	0.118
Inestabilidad emocional	0.161
Amabilidad	0.145
Responsabilidad	0.032
Apertura al cambio	0.118
Promedio	0.139

Cuadro 12: Corpus PAN-CLEF (train) 2015, Resultados rasgos de personalidad (holandés).

6.2 Laboratorio de evaluación PAN-CLEF 2015

En 2015 se llevó a cabo el treceavo laboratorio de evaluación organizado por PAN-CLEF[8]. La tarea de perfilado de autor consistió en predecir el género, la edad y 5 rasgos de personalidad de usuarios de Twitter a partir de los *tweets* emitidos.

El corpus de entrenamiento corresponde al corpus descrito en la sección 4.1, mientras que el corpus de prueba se encuentra constituido por 142 muestras en inglés, 88 en español, 36 en italiano y 32 en holandés (Rangel et al., 2015). Estos dos corpus constituyen el conjunto de datos oficial.

El método propuesto ne este artículo se posiciona en segundo lugar *(gonzalesgallardo15)* de la tabla general de resultados descrita en (Rangel et al., 2015). Un extracto de la misma se muestra en el cuadro 13.

Lugar	**Equipo**	**Global**
1	alvarezcarmona15	0.8404
2	gonzalesgallardo15	0.8346
3	grivas15	0.8078
4	kocher15	0.7875
5	sulea15	0.7755
...	...	...
19	bayot15	0.6178

Cuadro 13: Extracto de la tabla de resultados en (Rangel et al., 2015).

[8]Sitio web del PAN-CLEF 2015: `http://pan.webis.de/clef15/pan15-web/index.html`

6.3 Resultados sobre el corpus CCDMX

El primer experimento realizado con este corpus pretende descubrir qué tanto repercute la diferencia en el tamaño de las muestras de entrenamiento y prueba. La fase de entrenamiento fue realizada con el 70 % de las muestras en español del corpus PAN-CLEF (train) 2015. Hay que recordar que una muestra de este corpus está compuesta por aproximadamente 100 *tweets*.

Se probaron las 5 979 muestras disponibles del corpus CCDMX. Los resultados se muestran en el cuadro 14.

	Pr	**Co**	F_1	**Ex**
Hombres	0.598	0.631	0.614	0.549
Mujeres	0.474	0.439	0.456	

Cuadro 14: corpus CCDMX, Resultados experimento 1.

En el segundo experimento se optó por generar muestras de 50 comentarios, que representan un compromiso razonable entre el número de muestras y número de caracteres por muestra (aproximadamente 5 000 caracteres).

Un total de 121 muestras fueron probadas con un modelo de aprendizaje entrenado con el 70 % de las muestras en español del corpus PAN-CLEF (train) 2015. Los resultados son ligeramente mejores que en el experimento anterior, pero el cambio de dominio parece repercutir en gran medida el desempeño del sistema (ver cuadro 15).

	Pr	**Co**	F_1	**Ex**
Hombres	0.657	0.942	0.774	0.686
Mujeres	0.818	0.346	0.486	

Cuadro 15: Corpus CCDMX, Resultados experimento 2.

Por último, un tercer experimento fue realizado sobre este corpus. De las 121 muestras, el 70 % fue utilizado para entrenar el modelo de aprendizaje y el 30 % para medir su desempeño.

Estos últimos resultados obtenidos son mucho mejores que los anteriores, reafirmando la hipótesis de que el cambio de dominio afecta en gran medida el desempeño del sistema presentado (ver cuadro 16).

	Pr	**Co**	F_1	**Ex**
Hombres	0.950	0.900	0.924	0.920
Mujeres	0.880	0.940	0.909	

Cuadro 16: Corpus CCDMX, Resultados experimento 3.

7 Conclusiones y trabajo futuro

El uso de n-gramas de caracteres y n-gramas de etiquetas POS, como lo muestra los resultados, es una buena opción en textos densos debido a su capacidad de extracción de información.

En el caso de n-gramas de caracteres, fue posible extraer emoticonos, exageración de signos de puntuación (inundamiento de caracteres), uso de letras mayúsculas y todo tipo de información emocional codificada en los *tweets* y en los comentarios de Facebook.

Con los n-gramas de etiquetas POS, para el español y el inglés fue posible capturar los subconjuntoss más representativos de dos y tres elementos gramaticales. En el caso del italiano y el holandés se pudieron capturar los elementos gramaticales más frecuentes.

El algoritmo de clasificación presentado muestra ser bastante eficaz para detectar el genero, aunque un poco menos adecuado en las tareas de clasificación de la edad. Una idea interesante a desarrollar en un trabajo futuro podría ser la traducción de los emoticonos usados en las redes sociales en términos que puedan ser procesados con los mismos algoritmos de este artículo. Así la frase:

"Estoy muy feliz :) :)"

cuyas etiquetas gramaticales son:

"V R A F F F F"

sería procesada como:

```
V R A EMOT#H_SMILE EMOT#H_SMILE
```

Pensamos que esta estrategia podría mejorar aún más los resultados del sistema de clasificación.

Otro estudio en el corpus CCDMX podría consistir en agrupar el conjunto de comentarios en grupos de tamaños variables, por ejemplo: $1, 2, 4, 8, \ldots, n2^n$ comentarios y medir su impacto en el desempeño del algoritmo.

El enfoque multilingüe del algoritmo da la oportunidad de ser aplicado en tareas que involucren la detección de género o edad en opiniones dentro de redes sociales (Cossu et al., 2014, 2015).

Agradecimientos

Este trabajo fue parcialmente financiado por el proyecto CONACyT-México No. 215179 "Caracterización de huellas textuales para el análisis forense". Igualmente agradecemos el financiamiento del proyecto Europeo CHISTERA CALL - ANR: *Access Multilingual Information opinionS (AMIS)*, (Francia - Europa).

Referencias

Argamon, Shlomo, Moshe Koppel, Jonathan Fine & Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text* 23(3). 321–346.

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

Cossu, Jean-Valère, Eric SanJuan, Juan-Manuel Torres-Moreno & Marc El-Bèze. 2015. Multidimensional reputation modeling using microblog contents. En F. Esposito, O. Pivert, M.-S. Hacid, W. Z. Rás & S. Ferilli (eds.), *Foundations of Intelligent Systems: 22nd International Symposium, ISMIS 2015*, 452–457. Springer.

Cossu, Jean-Valère, Rocio Abascal-Mena, Alejandro Molina, Juan-Manuel Torres Moreno & Eric SanJuan. 2014. Bilingual and Cross Domain Politics Analysis. *Research in Computing Science* 1(85). 9–19.

Doyle, Jonathan & Vlado Kešelj. 2005. Automatic Categorization of Author Gender via N-Gram Analysis. En *6th Symposium on Natural Language Processing, SNLP*, n/a.

Giannakopoulos, George, Vangelis Karkaletsis & George Vouros. 2008. Testing the use of n-gram graphs in summarization sub-tasks. En *Text Analysis Conference*, 158–167.

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan & Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, 42–47. ACL.

Koehn, Philipp. 2010. *Statistical machine translation*. New York, NY, USA: Cambridge University Press 1st edn.

Koppel, Moshe, Shlomo Argamon & Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17(4). 401–412.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge.

Nowson, Scott, Julien Perez, Caroline Brun, Shachar Mirkin & Claude Roux. 2015. XRCE Personal Language Analytics Engine for Multilingual Author Profiling—Notebook for PAN at CLEF 2015. En L. Cappellato, N. Ferro, G. Jones & E. SanJuan (eds.), *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*, vol. 1391 CEUR Workshop Proceedings, .

Oberreuter, Gabriel & Juan D. Velásquez. 2013. Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications* 40(9). 3756–3763.

Padró, Lluís & Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, ELRA.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & É. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Machine Learning Research* 12. 2825–2830.

Peersman, Claudia, Walter Daelemans & Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. En *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 37–44. ACM.

Rangel, F., P. Rosso, M. Potthast, B. Stein & W. Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. En Cappellato L., Ferro N., Gareth J. & San Juan E. (eds.), *CLEF 2015 Labs and Workshops, Notebook Papers*, online.

Stamatatos, Efstathios. 2006. Ensemble-based Author Identification Using Character N-grams. En *3rd International Workshop on Text-based Information Retrieval*, 41–46.

Stamatatos, Efstathios. 2009. A Survey of Modern Authorship Attribution Methods. *American Society for information Science and Technology* 60(3). 538–556.

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

Torres-Moreno, Juan-Manuel. 2014. *Automatic text summarization*. London: Wiley-Sons.

Vapnik, Vladimir N. 1998. *Statistical learning theory*. New York: Wiley-Interscience.

Anexo

En este anexo presentamos algunos ejemplos de normalización dinámica, y una distribución de muestras por rasgos de personalidad en el corpus PAN-CLEF 2015.

tweet original	@username creo que esta versión la supera... ...http://t.co/peOlOweM Lo va petar en la #feriaJaen2012
tweet normalizado	@us creo que esta versión la supera... ...htt Lo va petar en la #feriaJaen2012
tweet original (POS)	F N V C D N P V N N V V S D F N
tweet normalizado (POS)	REF@USERNAME V C D N P V... ...REF#LINK N V V S D REF#HASHTAG

Cuadro 17: Normalización dinámica dependiente del contexto.

		Rango								
Idioma	Rasgo	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5
Español	Extraversión	3		5	5	28	32	9	9	9
	Inestabilidad emocional	2	10	25	9	12	19	10	10	2
	Amabilidad		3	16	6	16	40	12	2	5
	Responsabilidad		2		21	7	20	12	21	17
	Apertura al cambio			7	10	37	15	9	14	8
Inglés	Extraversión	1	4	10	17	41	37	20	13	9
	Inestabilidad emocional	11	5	22	9	19	37	19	18	12
	Amabilidad	5	2	12	19	44	46	13	7	4
	Responsabilidad		1	4	30	38	27	33	12	7
	Apertura al cambio			2	1	47	39	23	19	21
Italiano	Extraversión				8	13	9		3	5
	Inestabilidad emocional		1	3	3	8	4	12	5	2
	Amabilidad			1	3	11	9	7		7
	Responsabilidad				3	18	6	5	6	
	Apertura al cambio				1	14	9	2	7	5
Holandés	Extraversión				3	5	11	7	6	2
	Inestabilidad emocional		1	5	3	3	4	6	8	4
	Amabilidad		2	1	5	10	10	2	4	
	Responsabilidad			2	4	15	6	5	2	
	Apertura al cambio					4	11	4	12	3

Cuadro 18: Corpus PAN-CLEF 2015, Distribución de muestras por rasgos de personalidad.