Determinación del género de autores de textos cortos a través de n-gramas

Francisco Antonio Castillo Velásquez, María Del Consuelo Patricia Torres Falcón, Ely Karina Anaya Rivera, Iván Peredo Valderrama, Jonny Paul Zavala de Paz

Universidad Politécnica de Querétaro,
Querétaro, México

{francisco.castillo, consuelo.torres, karina.anaya, ivan.peredo}@upq.mx,
jonny.zavala@upq.edu.mx

Resumen. En la actualidad, la posibilidad de comunicarse o de expresarse por un medio electrónico es muy amplia: correo electrónico, redes sociales, chats y otras herramientas son usadas por la mayoría de los usuarios de computadoras y dispositivos móviles. Uno de los problemas que se ha presentado con esta forma de comunicación es el exceso, como el plagio, falsa identidad, notas intimidatorias, etc. La atribución de autoría de textos (AAT) se encarga de responder a la cuestión de quién es el autor de un texto, dando algunos ejemplos previos de ese autor (conjunto de entrenamiento). Un proceso útil dentro de la AAT es la identificación de género o sexo (hombre, mujer) y que ha sido estudiado por varios autores pero principalmente para el inglés. El presente trabajo propone un modelo computacional basado en características léxicas (n-gramas) para la identificación del género para textos cortos en español. Se hicieron pruebas con un corpus de textos de mensajes en redes sociales y blogs, obteniendo resultados prometedores.

Palabras clave: Identificación de género, aprendizaje automático, n-gramas, clasificación, autoría

Gender Determination of Authors of Short Texts using N-grams

Abstract. Nowadays, the possibilities for communicating or expressing through an electronic way are very wide: e-mail, social networks, chats and other ways are used by the majority of computer and mobile device users. One of the problemas that is presented in this communication way is excess, such as plagiarism, identitiy falsification, blackmailing, etc. Text authorship attribution (TAA) is in charge of answering authoring issues by providing previous examples from said author (training set). A useful process within TAA is sex or gender identification (male, female), which has been studied by many authors for its use in English mostly. The present work proposes a computational model

based on lexical characteristics (n-grams) for gender identification in short texts in Spanish. Tests were carried out with a corpus from social network and blog text messages, producing promising results.

Keywords: Gender identification, machine learning, n-grams, classification, authorship.

1. Introducción

Las diferencias en la manera de expresarse verbalmente entre hombres y mujeres provienen de múltiples factores. Por una parte se encuentran elementos extrínsecos como la cultura, las exigencias sociales o la educación; y por otro lado, existen factores intrínsecos como las capacidades personales, el entrenamiento y la propia personalidad. De acuerdo con algunos estudios la manera de comunicarse en gran parte está determinada por las "diferencias en el funcionamiento y la estructura cerebral entre hombres y mujeres" [5]. Estas diferencias se manifiestan tanto en la comunicación oral como en la escritura.

La información textual que proporcionan los usuarios en redes sociales, sistemas de correo electrónico o blogs ofrece un potencial mercadológico y de seguridad, principalmente. Pero es indudable que parte de esa información no es del todo confiable. Muchos usuarios mienten sobre su edad, género, afiliación o gustos, apoyándose de fenómenos lingüísticos como el sarcasmo o la ironía; en otros casos, simplemente no reportan dicha información. Conocer el perfil demográfico o psicológico de tales usuarios es una oportunidad para las organizaciones y empresas y un reto para las tecnologías de Procesamiento del Lenguaje Natural (PLN).

Conocer el género del autor de un texto puede ser útil en tareas de la lingüística forense, como la identificación de escritos intimidatorios, detección de plagio y atribución de autoría.

Muchos estudios hechos están dirigidos a resolver el problema del género, pero la mayoría de estas investigaciones están limitadas al inglés y a los medios tradicionales. Este trabajo propone un modelo dirigido a texto en español (y con facilidad de aplicarse a muchos lenguajes) de fuentes como redes sociales y blogs.

Esta investigación presenta un modelo simple de identificación de género basado en la técnica de n-gramas, los cuales son secuencias de n elementos, que para el caso de la comunicación escrita pueden ser caracteres o palabras, por mencionar algunos. El principal objetivo es proporcionar un enfoque simple de identificación de género que pueda utilizarse para el español u otros lenguajes y con un grado de confiabilidad por lo menos similar a otros estudios hechos con otros enfoques diferentes.

El resto del artículo está organizado como sigue: en la sección 2 mencionamos un breve estado del arte de los estudios de la atribución o categorización del género. En la sección 3, presentamos los datos usados en este trabajo y definimos el problema de predicción del género. Las técnicas que explotamos para resolver el problema planteado se presentan en la sección 3. La sección 4 muestra los resultados de los experimentos llevados a cabo para evaluar la viabilidad de la predicción de género. Finalizamos discutiendo algunas conclusiones del efecto del género sobre el estilo de la escritura.

2. Estado del arte

Los estudios de autoría en la literatura pueden dividirse en tres categorías: atribución de autoría, detección de similitud y caracterización de la autoría. La atribución es la tarea de encontrar o validar el autor de un documento. Algunos ejemplos bien conocidos sobre atribución son la revisión de los trabajos de Shakespeare [4, 5] y la identificación de los autores de los disputados Documentos Federales (Federalist papers) [6, 7, 8]. La detección de similitud intenta encontrar la variación entre los escritos de un autor o diferenciar entre segmentos de texto escritos por diferentes autores, mayormente con propósitos de detección de plagio.

La caracterización de la autoría puede definirse como la tarea de asignar los escritos de un autor a un conjunto de categorías de acuerdo a un perfil sociolingüístico. Algunos atributos analizados previamente en la literatura son el género, nivel de educación, idioma y antecedentes culturales. En [11], se examina el género y el idioma usando técnicas de aprendizaje automático. En [12] se clasifican documentos en inglés de acuerdo al género del autor y al género del documento.

El análisis estilométrico también proporciona resultados interesantes, En general, las mujeres tienden a preferir usar palabras más grandes y con significado claro. A diferencia de los hombres, prefieren organizar oraciones más cortas y a omitir stopwords y signos de puntuación. El uso de caritas (*smiles*) y palabras que conlleven emociones es más común en las mujeres. Los mensajes largos de chat y el uso de palabras cortas son las características estilísticas más representativas de los hombres [7].

Procesando grandes cantidades de texto, usando un enfoque más orientado a palabras funcionales que en las de contenido y apoyándose de un software de análisis, Newman [10] intentó dar respuesta empírica a las cuestiones de cómo o por qué los hombres y las mujeres usan el lenguaje de forma diferente.

Yan & Yan [18] usaron la clasificación con Naïve Bayes para identificar el género de autores de blogs. Además de usar las características tradicionales de categorización, usaron algunas específicas, como los colores de fondo, fuente del texto y emoticones. Realizaron experimentos también con el enfoque de unigramas de palabras. Su corpus fue de 75000 entradas de 3000 bloggers. Sus experimentos más prometedores alcanzaron una precisión del 70%.

Chao-Yue [4] también desarrolló un clasificador Naïve Bayes y lo entrenó con frecuencias de palabras como principal característica. Sus resultados tuvieron una leve mejora con la adición de bigramas, trigramas y etiquetas PoS frecuentes. Su trabajo tomó en cuenta el efecto positivo que tienen las frases orientadas a la relación y las palabras orientadas al tópico sobre la precisión en los resultados, por lo que también realizó experimentos en donde las excluía.

Otro modelo de inferencia de género para redes sociales fue propuesto por Kokkos & Tzouramanis [6]. Ellos idearon una estrategia que explotaba tanto características basadas en contenido (características psicolingüísticas como los sentimientos no placenteros –ira, depresión, confusión, miedo) como características tradicionales de estilo (características basadas en carácter, en palabras y en sintaxis –total de palabras, cantidad de letras mayúsculas, cantidad de signos de interrogación, total de pronombres). La implementación del modelo incluyó un módulo de minería de texto

que combinó un etiquetador PoS y un clasificador SVM. Ellos reportan una precisión superior al 90% para pruebas con textos tomados de las redes Twitter y LinkedIn.

Nuestro trabajo de investigación está basado en un enfoque de n-gramas. De un estudio previo de nuestra parte surgió el concepto de n-gramas sintácticos. Estos son n-gramas definidos mediante caminos de un árbol sintáctico de dependencias o de constituyentes en lugar de la estructura lineal del texto [10]. Por ejemplo, la oración "las noticias económicas tienen poco efecto sobre los mercados financieros" puede ser transformada a n-gramas sintácticos siguiendo la estructura de sus relaciones de dependencia: *tienen-noticias*, *efecto-poco*, *tienen-sobre-mercados-los*.

Este tipo de n-gramas están destinados a reflejar la estructura sintáctica más fielmente que los n-gramas lineales, y tienen muchas de las mismas aplicaciones, especialmente como características en un Modelo de Espacio Vectorial. Los n-gramas sintácticos dan mejores resultados que el uso de n-gramas estándar para ciertas tareas, por ejemplo, para atribución de autoría [16]. En el presente trabajo no se usan n-gramas sintácticos, pero sí n-gramas simples de carácter, obteniendo resultados alentadores en la tarea de identificación de genero de autores.

3. Desarrollo

En muchas tareas del PLN los documentos son representados como vectores de características. Estos vectores pueden servir como entrada a varios algoritmos como de clusterización y clasificación de documentos. Las características más usadas son las léxicas y las de carácter, que consideran a un texto como una secuencia de palabras y de caracteres, respectivamente. La frecuencia de palabras, riqueza del vocabulario, n-gramas, frecuencias de letras, n-gramas de caracteres, etc., son ejemplos específicos. Una gran ventaja de estas características de bajo nivel es que son muy fáciles de extraer de forma automática [3].

Este trabajo también se vio motivado por la conclusión de Sarawgi [12], donde concluye que el enfoque más robusto está basado en modelos del lenguaje basados en caracter (que aprenden patrones morfológicos) más que en modelos basados en tokens (que aprenden patrones léxico-sintácticos).

En esta sección describiremos el modelo propuesto de trabajo, desde la compilación del corpus de trabajo, pasando por el proceso de obtención de los n-gramas y generación de estadísticas, hasta la tarea de clasificación (ver figura 1).

De forma general el proceso inicia con una recopilación de textos cortos que eventualmente formarán el corpus de trabajo. Estos textos son procesados para obtener n-gramas a nivel de carácter con un programa especial (text2ngram). Este programa no permite la ordenación por el campo de frecuencia por lo que se hizo necesario tener un proceso semi-automatizado para ordenar los n-gramas por frecuencias. El resultado de esta ordenación y de la generación del corpus son dos archivos de texto, los cuales son procesados para generar un archivo en formato reconocido por Weka, software que más adelante nos permitirá realizar la clasificación. Este archivo .arff contiene la información de las características representativas para la clasificación de género, que estarán definidas por los n-gramas más frecuentes. Se utilizaron varios algoritmos para el proceso de clasificación, entre ellos Naïve Bayes, máquina de soporte a vectores y

árboles de decisión. Finalmente, se obtienen las cifras de la clasificación. Estas etapas del modelo propuesto se detallan a continuación.

3.1 Compilación del corpus

La parte inicial del trabajo fue la generación de un corpus de mensajes de texto cortos en español (que no sobrepasaran de 300 caracteres) obtenidos de comentarios en diversas páginas que tenían añadido el plug-in de comentarios de Facebook. Este corpus estará disponible para la comunidad investigadora.

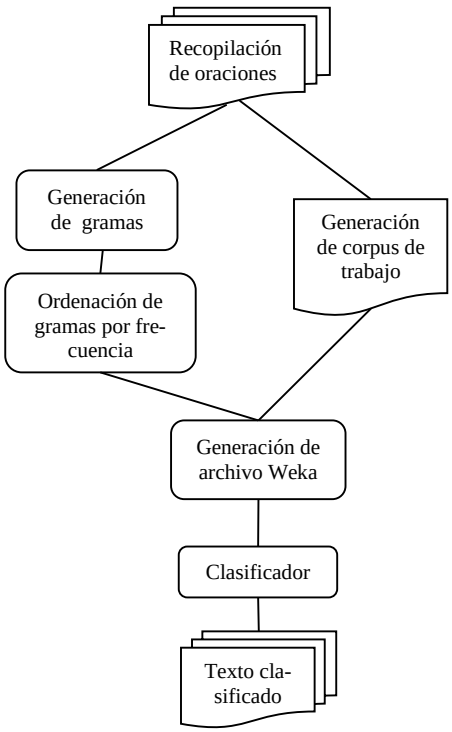

Fig. 1. Modelo propuesto para la identificación de género de textos cortos.

Se generó un corpus de 400 textos, la mitad de mujeres y la otra mitad de hombres. Como se mencionó anteriormente, cada uno de los textos no sobrepasa de 300 caracteres. A continuación se muestran algunos ejemplos de textos que forman parte del corpus. Se ha dejado la redacción original, inclusive con errores ortográficos o uso de *smyles*.

si lo sabes conservar sin caer en la monotonía sii es posible...!!
Ánimo compi no tengas meyo jeje pronto te iré a visitarte
No es gusto.. es por amor!
Extracto del corpus de textos de hombres

se pasa bien, estamos felices y mas unidos k nunk...

Guapaa!!! Ame tu falda
Ultimo día en mi trabajo :(pero que bonito detalle !!!!!!!
Extracto del corpus de textos de mujeres

Estos textos fueron compilados en un solo archivo plano con el campo del género y el texto correspondiente. Por ejemplo, los textos de los corpus anteriores quedarían compilados de la siguiente manera:

hombre,'si lo sabes conservar sin caer en la monotonía sii es posible...!!'
hombre,'Ánimo compi no tengas meyo jeje pronto te iré a visitarte'
hombre,'No es gusto.. es por amor!'
mujer,'se pasa bien, estamos felices y mas unidos k nunk...'
mujer,'Guapaa!!! Ame tu falda'
mujer,'Ultimo día en mi trabajo :(pero que bonito detalle !!!!!!!'

También se eliminaron frases orientadas a la relación, como aquellas que incluían "*mi novia*", "*mi esposo*" ya que esto podría causar ruido en los resultados de la clasificación posterior.

3.2 Generación de n-gramas

Tomando como punto de referencia el segundo texto del extracto del corpus de mujeres ("*Guapaa!!! Ame tu falda*") podemos generar los bigramas de caracteres *Gu,ua, ap, pa, aa, a!, ¡!(2), ¡_, _A, Am, me, e_, _t, tu, u_, _f, fa, al, ld* y *da*, (tomamos en cuenta símbolos de puntuación). También podemos generar los trigramas *Gua, uap, apa, paa, aa!, a!!, ¡!!, ¡!_, ¡_A, _Am, Ame, me_, e_t, _tu, tu_, u_f, _fa, fal, ald* y *lda*, . La idea de trabajar con gramas es muy simple y tiene la ventaja adicional que puede aplicarse prácticamente para cualquier idioma.

Nuestro modelo hace un análisis estadístico de las apariciones de los gramas en cada una de los textos. Se pretende obtener un conjunto de características definitorias del género de un autor basado en este fundamento léxico de caracteres. Un análisis de este nivel (superficial) no necesita de un procesamiento profundo de los textos, como lo hace un análisis sintáctico (tanto de dependencias como de constituyentes).

El término n-grama refiere a una serie de tokens secuenciales en un documento. La serie puede ser de longitud 1 (unigramas), longitud 2 (bigramas), etc., hasta llegar al n-grama correspondiente. Los tokens usados pueden ser palabras, letras o cualquier otra unidad de información presente en el documento. [15]

El uso de modelos de n-gramas en el PLN es una idea relativamente simple, pero se ha encontrado que es efectiva en muchas aplicaciones. Por ejemplo, modelos del lenguaje a nivel de caracter pueden ser aplicados a cualquier lenguaje, inclusive a otros tipos de secuencias como las del ADN y la música. Otras tareas en donde se ha aplicado esta idea es en la compresión de textos y la minería de datos. [15]

Cada n-grama se convertirá posteriormente en un atributo de tal forma que el algoritmo de aprendizaje que usemos intentará generar conocimiento sobre el uso de los n-gramas por parte de cada autor.

Cada atributo (n-grama) tendrá un valor real asociado que sale de la fórmula:

$$v_i = \frac{frec_{jd}}{Tfrec_j}, \quad (1)$$

donde V_j es el valor del atributo j, $frec_{jd}$ es la frecuencia del atributo j en el documento d y $Tfrec_j$ es la frecuencia del atributo j en la colección de textos.

La generación de estos gramas es automática y a través de la herramientas text2ngram. Un extracto de salida de este módulo es algo como se muestra a continuación (tomando como ejemplo la información mostrada en la sección 3.1):

si 4	!! 9
to 3	a_ 3
... ...	

Esta es solo una muestra de ejemplo de salida, ya que el proceso se hace al analizar todos los textos de un autor (hombre o mujer). El número a la derecha del grama representa el total de apariciones.

A estos archivos se les aplica un proceso de delimitación de fin de línea para que la herramienta text2ngram –que ayudará en la extracción de los n-gramas, pueda generar gramas "falsos" y así poderlos identificar para su eliminación.

Luego, ya estamos en condiciones para aplicar un subproceso parecido al que se llevó a cabo en la construcción del *baseline*, compilando frecuencias de n-gramas.

3.3 Generación de archivo para clasificación

El proceso de clasificación hace uso de un archivo en formato .arff (Weka). Este software contiene una colección de algoritmos de aprendizaje para tareas de minería de datos. La generación del archivo es un proceso semi-automático, que como se ve en la figura 1, necesita un listado de gramas y el archivo que representa al corpus de trabajo. A continuación se muestra un extracto de un ejemplo de archivo generado.

```
@relation generacionweka
@attribute ente numeric
@attribute !!!! numeric
@attribute ento numeric
@attribute ment numeric
... ...
@attribute ades numeric
@attribute genero {'mujer','hombre'}
@data
0,0,0,0,0,1,0,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,'hombre'
0,0,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,1,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,
0,1,0,0,1,'hombre'
... ...
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,'mujer'
0,0,0,0,0,0,2,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,'mujer'
```

4. Resultados

Los experimentos fueron desarrollados sobre los datos de un corpus para el problema de identificación de género. Dicho corpus está conformado por textos

obtenidos de redes sociales y blogs. Seleccionamos textos que no sobrepasaran los 300 caracteres y que fueran independientes de las relaciones. Se presentan los resultados de los experimentos para el corpus de 200 y 400 textos. Para la evaluación de los experimentos usamos el 60% de los datos para entrenamiento y el resto para clasificación.

En los resultados que se mostrarán a continuación, usamos el término "profile size" (tamaño del perfil) para representar los primeros n-gramas más frecuentes; por ejemplo, un tamaño del perfil de 40 significa que se usaron solo los primeros 40 n-gramas más frecuentes. Probamos varios umbrales para el perfil y seleccionamos 5 de ellos, como se muestra en todas las tablas de resultados.

Cuando alguna celda de la tabla contiene ND (no disponible) significa que nuestros datos fueron insuficientes para obtener el número correspondiente de n-gramas. Sucede solo con los bigramas, ya que en general hay menos bigramas que trigramas, etc. En estos casos el número total de todos los bigramas es menor que el tamaño del perfil.

Tabla 1. Resultados experimentales para el corpus de 200 oraciones (100 de mujeres y 100 de hombres).

tamaño del perfil	**clasificador**	**tamaño del n-grama**			
		3	**4**	**5**	**6**
40	SVM	48	51	**98**	58
	NB	55	55	**92**	53
	J48	49	50	**98**	50
80	SVM	49	52	**98**	**98**
	NB	55	54	**96**	95
	J48	54	56	97	**98**
120	SVM	52	95	**99**	98
	NB	56	84	**98**	95
	J48	52	97	97	**98**
160	SVM	52	94	**99**	98
	NB	58	82	**98**	95
	J48	53	97	97	**98**
200	SVM	53	95	**99**	98
	NB	56	94	**98**	95
	J48	50	97	97	**98**

La tarea de clasificación consiste en seleccionar características para construir el modelo de espacio de vectores, algoritmos supervisados de entrenamiento y clasificación; es decir, decidir a qué clase pertenece el fragmento de texto –en nuestro modelo de espacio de vectores. En este trabajo presentamos resultados para tres clasificadores: SVM (NormalizedPolyKernel de SMO), Naïve Bayes y J48.

En términos generales vemos los mejores resultados para los gramas más grandes (5 y 6) con una precisión que alcanza hasta un 99%. Esto se obtiene con ambos corpus

de trabajo (200 y 400 textos). Los experimentos se realizaron con un modelo de validación cruzada con 10 iteraciones (10-folds).

Es interesante notar el salto fuerte que hay entre cifras, ya que de pasar de los 50s llega bruscamente a los 80s o 90s, sin detenerse entre 60s y 70s, en ambas pruebas.

Los clasificadores fueron usados motivados por otros trabajos donde han dado resultados aceptables, como en [13].

Tabla 2. Resultados experimentales para el corpus de 400 oraciones (200 de mujeres y 200 de hombres).

tamaño del perfil	clasificador	tamaño del n-grama			
		3	**4**	**5**	**6**
40	SVM	53	**54**	53	49
	NB	**57**	53	52	50
	J48	52	**54**	50	50
80	SVM	55	51	57	**99**
	NB	57	54	56	**97**
	J48	43	54	50	**99**
120	SVM	53	56	**99**	**99**
	NB	56	56	**97**	**97**
	J48	50	55	**99**	**99**
160	SVM	53	98	**99**	98
	NB	55	86	**97**	**97**
	J48	48	98	**99**	**99**
200	SVM	51	97	**99**	**99**
	NB	56	85	**98**	97
	J48	50	98	98	**99**

5. Conclusiones y trabajo futuro

En este artículo se propuso un modelo computacional para la identificación del género de autores de textos cortos. El enfoque se basa en la técnica de n-gramas de caracteres, que entre otras ventajas, se puede aplicar a cualquier lenguaje. Se hicieron pruebas con un corpus de 200 y 400 escritos (sin ningún pre-procesamiento de corrección ortográfica o gramatical). La clasificación se aplicó con los algoritmos NaiveBayes, SVM (SMO) y J48, alcanzando cifras de hasta casi un 100% de clasificación correcta en algunos casos.

Como trabajo futuro se hace necesario probar el modelo con una mayor cantidad de textos cortos y añadiendo características de estilo (riqueza del vocabulario, frecuencia de palabras, por ejemplo) para verificar si esto aumenta los resultados en la precisión.

Referencias

1. Argamon, S., Koppel, M., Pennebaker, J., Schler, J.: Automatically profiling the Author of an Anonymous Text. Communications of the ACM - Inspiring Women in Computing, Vol. 52, No. 2, pp. 119–123 (2009)

2. Bamman, D., Eisenstein, J., Schnoebelen, T.: Gender identity and lexical variation in social media. Journal of Sociolinguistics, Vol. 18, No. 2, pp. 135–160 (2014)
3. Bogdanova, D.: Extraction of High-Level Semantically Rich Features from Natural Language Text. Conference Proceedings II of the 15th East-European Conference on Advances in Databases and Information Systems, pp. 262–271 (2011)
4. Chao-Yue, L. Author Gender Analysis. (retrieved link) (2010)
5. de Iceta, M.: Diferencias cerebrales en función del sexo. Revista web de psicoanálisis, aperturas psicoanalíticas, Vol. 15, (http://www.aperturas.org/) (2003)
6. Kokkos, A., Tzouramanis, T.: A Robust Gender Inference Model for Online Social Networks and its Application to Linkedin & Twitter. First-Monday peer–reviewed journals on the Internet, Vol. 19, No. 9 (2014)
7. Koppel, M., Argamon, S., Shimoni, A.: Automatically Categorizing Written Texts by Author Gender. Literary & Linguistic Computing, Vol. 17, No. 4, pp. 401–412 (2002)
8. Kucukyilmaz, T., Cambazoglu, B., Aykanat, C., Can, F.: Chat Mining for Gender Prediction. Lecture Notes in Computer Science (4243), pp. 274–283 (2006)
9. Muhammad, M., Wolfe, B.: Gender Classification of Mobile Application Reviews (2013)
10. Newman, M., Groom, C., Handelman, L., Pennebaker, J.: Gender differences in language use: an analysis of 14,000 text samples. Discourse Processes, Vol. 45, No. 3, pp. 211–236 (2008)
11. Rosso, P., Rangel, F.: On the identification of emotions and authors' gender in Facebook comments on the basis of their writing style. CEUR Workshop Proceedings 1096, pp. 34–46 (2013)
12. Sarawgi, R., Gajulapalli, K., Choi, Y.: Gender Attribution: Tracing Stylometric Evidence beyond Topic and Genre. Proc. CoNLL '11 Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 78–86 (2011)
13. Sidorov, G., Velásquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L.: Syntactic Dependency-based N-grams as Classification Features. LNAI 7630, pp. 1–11 (2012)
14. Sidorov, G., Velásquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L.: Syntactic N-grams as Machine Learning Features for Natural Language Processing. Expert Systems with Applications, Vol. 41, No. 3, pp. 853–860 (2014)
15. Doyle, J., Keselj, V.: Automatic Categorization of Author Gender via N-Gram Analysis. Proceedings of the 6th Symposium on Natural Language Processing, SNLP'2005 (2005)
16. Singh, S., Sarwan, M., Bharshiv, M., Sathe, A.: Gender and Age Classification on the Basis of Blogs.
17. Ugheoke, T., Saskatchewan, R.: Detecting the Gender of a Tweet Sender. M.Sc. Project Report. Department of Computer Science, University of Regina (2014)
18. Yan, X., Yan, L.: Gender classification of Weblog Authors. Proceedings of the AAAI Spring Symposia on Computational Approaches, pp. 228–230 (2006)