Identificación de perfiles de usuario

P. Espinoza, D. Vilariño, D. Pinto, M. Tovar y B. Beltrán

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Puebla, México

patricia.efong@gmail.mx
darnes,dpinto,mtovar,bbeltran@cs.buap.mx
http://nlp.cs.buap.mx

Resumen. En la presente investigación se propone un modelo para la identificación de perfiles de usuario. El modelo propuesto utiliza un conjunto de características extraídas de los textos. El modelo fue validado con 4 corpus en inglés: de Blogs, de Redes sociales, de Criticas y de Twitter y con 2 corpus en español: de Blogs y de Criticas. Se compara el desempeño de tres de los algoritmos más usados para clasificación: Naïve Bayes, Máquinas de Soporte Vectorial (SVM) y K Vecinos más cercanos (IBk).

Palabras clave: modelo de clasificación, perfil de usuario, patrones semánticos.

1. Introducción

Internet en la actualidad ofrece diversas herramientas para que los usuarios puedan expresarse libremente sin importar la edad, el sexo, el tema que traten y a que se dedican. La cantidad de conversaciones en línea (foros, salas de chats, redes sociales y blogs, entre otros medios) ha aumentado considerablemente. Dada esta situación es prácticamente imposible analizar manualmente una conversación y detectar el perfil del autor que la ha escrito.

La detección del perfil de un autor que puede ser edad, sexo, lenguaje nativo o tipo de personalidad es un problema que ha ganado importancia en aplicaciones forenses, de seguridad y de mercadotecnia. Hoy en día la comunidad de procesamiento de lenguaje natural desea estudiar la forma en que se comunican los diferentes grupos de edades y sexo, tratando de detectar los patrones de escritura comunes y diferentes entre estos grupos.

En la presente investigación se desea, dado un texto detectar la edad y el sexo de la persona que lo ha escrito. Dicho texto es un documento obtenido de los corpus de la Conferencia Internacional PAN 2014[1]. Lo que se pretende es encontrar patrones, características léxicas, sintácticas y semánticas propias de cada grupo de edad y género, para el desarrollo de modelos de aprendizaje que nos permitan clasificar adecuadamente dichas conversaciones.

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

La estructura del artículo es la siguiente. En la sección 2 se presentan los trabajos desarrollados en la literatura con respecto a la identificación de perfiles de usuario. La sección 3 presenta la descripción de las características seleccionadas para desarrollar el modelo de clasificación. La discusión acerca de los resultados obtenidos se presenta en la sección 4. Finalmente la conclusión del presente trabajo de investigación se realiza en la sección 5.

2. Trabajo relacionado

Se realizó un estudio sobre los trabajos desarrollados en esta área, enfatizando sus avances, alcance, enfoques, ventajas y desventajas, así como sus aportaciones científicas, encontrando el siguiente panorama general:

En la propuesta presentada en [1] se desarrollan dos modelos uno para el idioma español y otro para el idioma inglés, ambos totalmente diferentes. Para el idioma inglés se extrajeron características léxicas y sintácticas, sin embargo para el idioma español, se realizó una representación mediante grafos de las conversaciones y se extrajeron los patrones de cada clase utilizando la herramienta SUBDUE[2]. Se reporta que los resultados para el idioma inglés superaron considerablemente los resultados obtenidos para el idioma español.

En la investigación desarrollada en [3] se proponen 2 tipos de características que pueden ser usadas para esta tarea. Características basadas en el contexto y características basadas en el estilo. Las características basadas en el estilo incluyen características léxicas y sintácticas utilizando Pos-tagger como etiquetador. Para las características relacionadas al contexto se extraen las 1000 palabras individuales con mayor frecuencia de un corpus que incluye 19 320 post extraídos de blogs escritos en inglés. Aplican además Información Mutua para detectar los pares de palabras que son colocaciones. Los resultados que obtuvieron muestran que las características estilográficas que más ayudan a discriminar el género son las preposiciones para el caso de los hombres y los pronombres para el caso de las mujeres. Y con respecto al contexto, los hombres utilizan palabras relacionadas con la tecnología y las mujeres utilizan más palabras relacionadas a la vida personal y a las relaciones.

El modelo propuesto en [8] se basa en el desarrollo de una variación del algoritmo *Exponential Gradient (EG)*, que permite detectar el género de un autor. Se propone una representación vectorial del conjunto de características que estudian y en cada paso bajo ciertos criterios de eliminación van reduciendo el espacio de representación, quitando aquellas características que aportan poco a la detección del género. Concluyen que las características más representativas son las palabras y las etiquetas de los textos.

En el trabajo desarrollado en [13] se estudia el comportamiento de hombres y mujeres blogueros, y mencionan que las características que mejores resultados ofrecieron son las palabras representativas de cada grupo, los hiperenlaces y palabras comúnmente usadas en los blogs (lol, haha, ur, entre otras).

[2] https://ailab.wsu.edu/subdue/

Los resultados que obtuvieron con estas características fueron del 80 % para el género y del 76 % para la edad. Se llegó a la conclusión de que las mujeres usan más pronombres y los hombres más preposiciones, también mencionan que se encontró que los blogs escritos por adolescentes son en su mayoría mujeres, que las mujeres hablan más sobre su vida privada y familia, mientras que los hombres hablan más sobre tecnología y política.

En otros trabajos precedentes para abordar esta tarea se puede observar, que las características más comúnmente utilizadas son:

- N gramas de palabras, [4],[9],[10] y [11].
- N gramas de caracteres, [4] y [11].
- Longitud de palabras, [5], [9] y [14].
- Longitud de oraciones, [6], [7] y [14].

En [10] y [9] se utiliza la herramienta *Linguistic Inquiry and Word Count (LIWC)*, la cual calcula el grado en que las personas usan diferentes categorías entre un conjunto de documentos, también se puede determinar el grado en el que un texto utiliza emociones positivas o negativas entre otras cosas. Además de las características mencionadas anteriormente, en el trabajo propuesto en [9] también se cuenta la frecuencia de uso de palabras en mayúscula, la frecuencia de uso de intensificadores y la longitud de las oraciones. En [12] las características que se usan son las mencionadas anteriormente y se agregan el uso de signos de puntuación, el uso de emoticones y el uso de las categorías gramaticales POS.

En el trabajo propuesto en [6] se utiliza la frecuencia de las clases a las que pertenecen las palabras. La clasificación de las palabras se realiza con la herramienta *RiTaWordNet* la cual establece la relación de una palabra con su clase mediante sinónimos e hiperónimos. Posteriormente se clasifican las palabras en positivas o negativas usando *SentiWordNet 3.0*, se cuenta los signos de puntuación usados, la frecuencia de las palabras cerradas, frecuencia de uso de pronombres, se reemplazan los emoticones por su palabra equivalente y se cuantifica una lista de palabras foráneas (meee, yesss, thy, u, urs, entre otras).

El modelo propuesto en [5] utiliza algunas de las siguientes características, la frecuencia de uso de palabras escritas en formato *CamelCase* y la frecuencia de uso de etiquetas POS. También comentan que las personas jóvenes utilizan más los pronombres en primera persona y que las personas que no son originarias de los Estados Unidos usan más las abreviaciones "u" y "ur". Mencionan que al igual que en [3] se aplica Información Mutua para detectar los pares de palabras que son colocaciones.

Otro trabajo que es importante destacar es el presentado en [5], donde se mencionan algunas características interesantes como son los determinantes (*a, the, that, these*) y cuantificadores (*one, two, more, some*), que sirven como indicadores para identificar a un hombre y una vez más se menciona que los pronombres (*I, you, she, her, their, myself, yourself, herself*) son indicadores para identificar a una mujer, ya que según los autores las mujeres tienden a personalizar más los textos que escriben, mientras que los hombres los generalizan.

En el trabajo desarrollado por [2] se presenta un metodología para detectar el perfil de un autor, en particular se considera edad y género. Las características

que utilizan son: categorías gramaticales, palabras cerradas, sufijos y signos. Los autores logran detectar solamente en un 55 % el género y a lo sumo un 45 % la edad. En este trabajo solo se presentan los resultados para el corpus de blogs en el idioma inglés y el idioma español ofrecido en la conferencia PAN 2013.

Entre las técnicas de clasificación más usadas se encuentran: Naive Bayes, que ha sido reportada en los trabajos desarrollados en [4],[14] y [7] y las máquinas de soporte vectorial (SMV) que han sido utilizadas en las investigaciones desarrolladas en [11], [4], [14] y [12].

Las características usadas fundamentalmente han sido léxicas, sintácticas y conteos de las frecuencias de uso de algunos elementos. En la presente investigación se pretende proponer características que de alguna manera permitan detectar el sentido del texto que se está estudiando y con ello analizar si es posible descubrir el perfil del autor.

Es importante destacar que es más simple detectar el género, que la edad, pues los hombres y las mujeres escriben o se interesan por temas diferentes independientemente de la edad que tienen. Un aspecto importante a estudiar es la técnica de clasificación que se debe usar.

3. Descripción del enfoque propuesto

Se ha desarrollado un modelo supervisado, donde se extraen un conjunto de características del corpus de entrenamiento considerando cuantas veces aparecen en este o su probabilidad de aparición. A continuación se detallan las 3 fases que permiten construir este modelo.

3.1. Preprocesamiento del corpus

Debido a que el corpus con el que se trabaja es descargado directamente de la página del PAN, es necesario hacer varias operaciones antes de poder trabajar con él, algunas de ellas son:

1. Separar el corpus por autor.
2. Separar el corpus por género.
3. Sustituir los símbolos HTML que pueda contener el texto, por su equivalente en utf8.

Para el último punto se desarrolló un diccionario de símbolos HTML.

3.2. Características seleccionadas

En el enfoque propuesto se emplea un modelo supervisado basado en máquinas de aprendizaje, para el cual se construye un modelo de clasificación usando los siguientes conjuntos de características, obtenidas de los documentos de cada autor del corpus de entrenamiento proporcionado para esta tarea:

1. Número de slangs.
2. Número de contracciones.
3. Número de prefijos.
4. Número de signos.
5. Número de links
6. Número de imágenes
7. Número de palabras mal escritas.
8. Longitud de la oración.
9. Número de números.
10. Número de palabras que empiezan con mayúscula.
11. Número de palabras escritas en mayúscula.
12. Longitud de la palabra más larga.
13. Número de palabras de longitud 1, 2,10 y 15.
14. 39 categorías gramaticales.
15. Conteo de las 200 palabras más frecuentes.
16. Probabilidad de cada palabra del vocabulario (unigrama).
17. Bolsa de palabras.

El primero conjunto está conformado por las primeras 13 características que se muestran en la lista, se realiza un conteo para determinar el número de veces que aparece cada característica en un documento. Se decide cuantificar las palabras de longitud 1, 2, 10 y 15 con el objetivo de detectar algún patrón que permitiera separar por grupos de edad y género, ya que son extremos, es decir, palabras muy cortas y muy largas. Se desarrollaron diferentes recursos léxicos para realizar estos conteos como son: un diccionario de *slangs*, un diccionario de signos, diccionario de contracciones, diccionario de prefijos y un diccionario que nos permite detectar si la palabra ha sido mal escrita.

Para el segundo conjunto de características, se creó un diccionario de categorías gramaticales las cuales se extrajeron del corpus después de ser etiquetado con la herramienta de Clips llamada pattern.en[3]. Posteriormente se utilizó el diccionario para contar el número de veces que aparece cada categoría gramatical en el documento.

Para la característica número 15 (conteo de las 200 palabras más frecuentes) se hizo un análisis del corpus de hombres y otro del corpus de mujeres para identificar cuáles son las 100 palabras más frecuentes de cada uno, descartando las palabras cerradas y las palabras que se repiten en ambos corpus. Algunas de las palabras que se extrajeron del corpus de blogs en inglés se pueden observar en la Tabla 1.

[3] www.clips.ua.ac.be/pages/pattern-en

Tabla 1. Palabras más frecuentes.

Corpus mujeres	Corpus hombres
Art	Banks
Design	Building
Diet	Development
Exercise	Economy
Food	Financial
Gallery	Goverment
Health	Information
Heart	Job
Home	Money
Personal	Payment

Para calcular la probabilidad de los unigramas, se eliminaron las palabras cerradas y símbolos y se calculó la probabilidad de aparición de cada palabra del corpus, las cuales son aproximadamente 20 mil palabras. Este conjunto de palabras se agregan al vector de la siguiente forma: si la palabra aparece en el documento, se pone la probabilidad calculada anteriormente para esa palabra, si la palabra no aparece en ese documento, el valor que se pone en el vector para esa palabra será cero.

Por último se agrega el texto de cada autor como una bolsa de palabras.

3.3. Representación de las características

Todas las características mencionadas en la sección 3.2 permiten construir un vector representativo de cada autor considerando ya sea la frecuencia o la probabilidad de aparición de cada una de las características seleccionadas.

Para la fase de entrenamiento, un ejemplo de dicho vector se muestra en la figura 1 donde el campo con el valor *Clase* al final del vector, es el atributo clasificador del documento que en el caso del género podría indicar si el documento pertenece a una mujer o a un hombre.

Fig. 1. Vector de entrenamiento.

Para la fase de prueba se utiliza un vector de características como se muestra en la en la figura 2, donde el atributo clasificador se sustituye por un signo de interrogación ya que se desconoce la clase a la que pertenece.

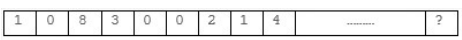

Fig. 2. Vector de prueba.

3.4. Proceso de clasificación

El modelo descrito anteriormente se puede observar en la figura 3, en ella se muestra el proceso que se sigue. Este se ha dividido en dos fases, en la primera fase se realiza el pre procesamiento descrito en la sección 3.1 y después se extraen las características descritas en la sección 3.2 en donde el atributo clasificador será el género del autor. Por último se envía a Weka el conjunto de vectores característicos que sirven para crear el *Modelo de clasificación por género*.

En la segunda fase se utiliza el mismo conjunto de vectores característicos que en la fase anterior, pero el atributo clasificador ahora será el rango de edad del autor. Aquí se crean dos modelos de clasificación diferentes, el *Modelo de clasificación de edadMujer* y el *Modelo de clasificación de edadHombre*. Como ya se sabe de antemano a que género pertenecen los documentos gracias a la fase anterior, se pueden discriminar los documentos para que a cada modelo solo entren vectores que correspondan a ese género.

Como paso final se evalúan los resultados de los modelos.

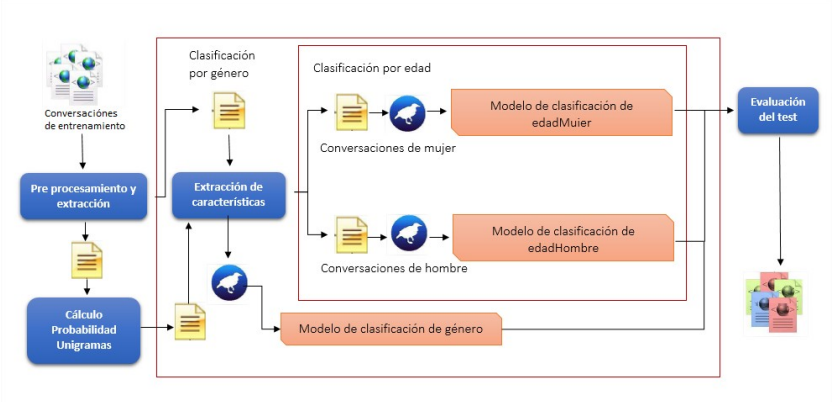

Fig. 3. Descripción del modelo.

4. Resultados

A continuación se muestra una descripción de los corpus que se utilizaron para el entrenamiento del modelo propuesto y posteriormente se muestran los experimentos realizados, así como los resultados de cada experimento.

4.1. Conjunto de datos

Se trabajó con cuatro corpus en el idioma inglés, los cuales contienen documentos de fuentes como Blogs, Criticas, Redes Sociales y Twitter. También se

contó con dos corpus en idioma español con documentos de Blogs y de Redes Sociales. Todos los corpus fueron obtenidos del sitio web del PAN 2014. Se proporciona una breve descripción en los siguientes tablas en donde se muestra el número de instancias con las que se cuenta en las diferentes corpus. La columna Autores representa el número de autores, las columnas 18-24, 25-34, 35-49, 50-64 y 65+ representan los rangos de edad de estos autores.

Tabla 2. Número de instancias por corpus y por clase para el idioma inglés.

Género	Autores	18-24	25-34	35-49	50-64	65+
Hombre(blog)	74	3	30	27	12	2
Mujer(blog)	73	3	30	27	11	2
Hombre(review)	2,080	180	500	500	500	400
Mujer(review)	2,080	180	500	500	500	400
Hombre(socialmedia)	3,529	693	945	1035	851	5
Mujer(socialmedia)	3,503	699	944	1025	828	7
Hombre(twitter)	149	9	44	63	29	4
Mujer(twitter)	152	10	43	65	30	4

Tabla 3. Número de instancias por corpus y por clase para el idioma español.

Género	Autores	18-24	25-34	35-49	50-64	65+
Hombre(blog)	44	2	13	21	6	2
Mujer(blog)	44	2	13	21	6	2
Hombre(socialmedia)	636	165	213	162	80	16
Mujer(socialmedia)	636	165	213	162	80	16

4.2. Experimentos

En las siguientes tablas se muestra un resumen de los mejores resultados obtenidos por corpus y por lenguaje. Todos los experimentos se hicieron aplicando validación cruzada de 10 pliegues y se utilizaron los algoritmos de clasificación vecinos mas cercanos (IBk), máquinas de soporte vectorial (SVM) y Naïve Bayes sobre el conjunto de características escogidas, las cuales fueron descritas en la sección 3.2. Se muestra en **negritas** los mejores resultados para cada idioma. Se puede observar que el algoritmo de clasificación que mejor comportamiento mostró fue la máquina de soporte vectorial, con polikernel . Los resultados varían de acuerdo al corpus. Esto nos indica que la forma en que se escribe en cada uno de ellos es diferente, ya que se han utilizado las mismas características. La detección del género en los corpus de Blogs, Twitter y Criticas superó el 80 %, sin embargo en las Redes Sociales las características escogidas no permitieron detectar fácilmente si el texto fue escrito por un hombre o una mujer.

Se realizaron diversos experimentos con el objetivo de analizar las características que nos permiten diferenciar cada una de las clases; llegando a la conclusión que la probabilidad de cada palabra le permite al clasificador poder detectar el género de la persona que ha escrito el texto. La detección de la edad es más compleja, debido a que el corpus no está balanceado y se dispone de 5 clases. Esto nos condujo a obtener resultados que no superan el 70 % de precisión.

Tabla 4. Resultados para los corpus en inglés.

Corpus	Tipo	Clase	Características	Clasificador	Resultado
BLOGS		Género	254 + Probabilidades + Texto	SMO	87.07
	Female	Edad	254 + Probabilidades + Texto	SMO	43.83
	Male	Edad	254 + Texto	SMO	**54.05**
CRITICAS		Género	254 + Probabilidades + Texto	SMO	**99.87**
	Female	Edad	254 + Probabilidades	SMO	30.76
	Male	Edad	54	SMO	29.61
REDES		Género	254	SMO	52.38
	Female	Edad	254	SMO	37.91
	Male	Edad	54	SMO	36.78
TWITTER		Género	254 + Texto	SMO	83.05
	Female	Edad	54	SMO	**48.29**
	Male	Edad	254 + Texto	SMO	49.48

Tabla 5. Resultados para los corpus en español.

Corpus	Tipo	Clase	Características	Clasificador	Resultado
BLOGS		Género	254 + Texto	SMO	**79.54**
	Female	Edad	254 + Texto	SMO	**68.18**
	Male	Edad	254	SMO	**45.45**
REDES		Género	254	SMO	60.61
	Female	Edad	254 + Texto	SMO	40.72
	Male	Edad	254 + Texto	SMO	38.36

5. Conclusiones

Se desarrolló un modelo para la detección del perfil de un autor (género y edad). Se pudo observar que el comportamiento del modelo propuesto fue similar para ambos idiomas y que los resultados obtenidos para varios corpus ofrecidos en la Conferencia Internacional PAN 2014 fueron satisfactorios para la detección del género, sin embargo es necesario incluir características diferentes para la detección de la edad.

Como trabajo futuro se propone desarrollar una representación de los textos mediante grafos de co-ocurrencia y calcular las medidas de centralidad que posee la herramienta gephi[4], con el objetivo de obtener las palabras representativas de cada clase y que puedan ser incluídas en el modelo de clasificación.

[4] http://gephi.github.io/

Referencias

1. Aleman, Y., Loya, N., Vilariño, D.: Two methodologies applied to the author. PAN 2013 (2014)
2. Aleman, Y., Vilariño, D., Pinto, D.: Avances en la ingeniería del lenguaje y del conocimiento. Research in Computer Science 85, 93–103 (2014)
3. Argamon, S., Koppel, M., J., P., Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM - Inspiring Women in Computing 52(2), 119–123 (2009)
4. Burger, J.D., Henderson, J., Kim, G., G., Z.: Discriminating gender on twitter. In: EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1301–1309 (2011)
5. Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: Author profiling for english emails. In: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics. pp. 263–272 (2007)
6. Gopal, P., Banerjee, S., Das, D.: Automatic author profiling based on linguistic and stylistic features. In: Proceedings of the 9th PAN at CLEF Conference (2013)
7. Goswami, S., Sarkar, S., Rustagi, M., Meder, T.: Stylometric analysis of bloggers age and gender. In: Proceedings of the Third International ICWSM Conference (2013)
8. Koppel, M., Argamon, S., A., S.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
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 (2013)
10. Nguyen, D., Smith, N., Rosé, C.: Author age prediction from text using linear regression. In: LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 115–123 (2011)
11. Peersman, C., Daelemans, W., Vaerenbergh, L.V.: Predicting age and gender in online social networks. In: SMUC '11 Proceedings of the 3rd international workshop on Search and mining user-generated contents. pp. 37–44 (2011)
12. Rangel, F., Rosso, P.: Use of language and author profiling: Identification of gender and age. http://users.dsic.upv.es/ prosso/resources/RangelRosso_NLPCS13.pdf (2009)
13. Schler, J., Koppel, M., S., A., J., P.: Effects of age and gender on blogging. In: Proceedings of the AAAI Spring Symposium on Computational (2006)
14. Zhang, C., Zhang, P.: Predicting gender from blog posts. http://people.cs.umass.edu/ pyzhang/course/genderClassify.pdf (2010)