Inferencia de Sexo Basada en Características de un Texto

Sofía Abrevaya *(PSI-UBA),* Germán Romano *(FCEN-UBA),* Mercedes Terragno *(PSI-UBA),* Valeria Tiffenberg *(FCEN-UBA)*

***Abstract*—Gran parte del análisis de la conducta se compone por el estudio de las producciones humanas (verbales y no verbales). En particular, el lenguaje es característico de nuestra especie, y asimismo permite intercambiar información estructurada. Se ha discutido la posibilidad de analizar las características de un discurso para determinar rasgos del emisor. El objetivo del presente trabajo es lograr determinar qué características lingüísticas podrían distinguir de manera precisa el género del autor de un texto.**

Utilizando una muestra de 7126 libros en inglés del Project Gutenberg, y a través de diferentes características estructurales lingüísticas del texto (tales como cantidad de frases, cantidad de referencias a relaciones afectivas, etc.), se desarrolló un clasificador de libros según género con la técnica de Machine Learning *Random Forest Clasifier* en Python. Se destacaron tres características como indicadores principales para determinar el género del autor: la proporción de pronombres femeninos sobre masculinos, y la preponderancia de referencias a niños y a colores.

Agosto, 2016

***Index Terms*—Neurociencia, Inferencia de Sexo, Inteligencia Artificial, Machine Learning.**

I. Introducción

La sobrepoblación de datos que han generado las computadoras y que ha multiplicado Internet, ha llevado a preguntarse si esa información puede ser utilizada con otros fines más allá de su utilidad original. Teniendo en consideración que parte del análisis de la conducta se compone por el estudio de las producciones humanas (verbales y no verbales) se ha determinado entonces que la producción de datos es un modo sistemático posible de analizar ciertas características humanas.

El lenguaje en particular es característico de nuestra especie, y asimismo poseedor de la habilidad de intercambiar información estructurada. Si bien puede parecer evidente que las diferentes palabras aparecen con frecuencia variable en diferentes secciones de un texto, sólo recientemente ha sido señalado que los patrones de variabilidad de la frecuencia se relacionan con la función lingüística de las palabras[1].

En relación a ciertas particularidades del lenguaje se ha discutido la posibilidad de analizar ciertas características en un discurso para determinar rasgos del emisor, como ser: la autoría, el género, la edad, etc. La predicción exacta de atributos demográficos de medios de comunicación social y otros contenidos es información valiosa para la comercialización, personalización, y la investigación legal[2]. En este artículo se discute y experimenta con aquellas relativas a la predicción del género.

La predicción acertada del género en medios virtuales podría permitir la generación de contenido y marketing específico, y la caracterización adecuada del usuario de una aplicación. Por otro lado, la identificación del género en el texto de un libro podría ayudar a determinar la autoría o autenticidad del mismo, o la reinterpretación de roles de género preestablecidos socialmente.

Diversas publicaciones han demostrado poder inferir el género de una fuente escrita. Actualmente, la mayor parte de los trabajos relacionados se centran en las fuentes online y no en fuentes escritas tradicionales. Por tanto, el objetivo de este trabajo es explorar la posibilidad de clasificar de manera automática libros de texto (de ficción y no ficción) según el género del autor.

II. Materiales y Métodos

El set de datos utilizado fue obtenido de forma automática de la base de datos del sitio del Project Gutenberg[3]. Éste almacena una gran variedad de textos, mayoritariamente de ficción, que no se ven afectados por leyes de derechos de autor en Estados Unidos. Habitualmente, esto se debe a que los derechos de autor han expirado, lo que en Estados Unidos sucede entre 50 y 100 años luego del fallecimiento del autor, según el estado donde estuviera protegido.

La descarga de los textos se hizo eligiendo el idioma inglés, por ser la lengua mejor cubierta por el Project Gutenberg y con las herramientas mejor desarrolladas para análisis de texto. A su vez, se optó por textos en formato txt, para una mayor simplicidad al momento de realizar el análisis automático. Dentro de esta selección, no fue descargada la totalidad de los volúmenes disponibles por cuestiones de tamaño y tiempo, sino que el criterio de elección fue el de descarga automatizada en Gutenberg, quien fija una forma de descarga legítima y provee los títulos por orden dentro de su catálogo[4].

El tamaño total de esta muestra fue de aproximadamente 13.000 libros, unos 6GB de datos. El total no pudo ser usado para el análisis clasificatorio por haberse hallado dificultades para identificar de manera automatizada el género del autor. El tamaño de la muestra que fue catalogada con éxito entre autores hombre o mujer fue de 7126 libros, unos 3.2GB, de los cuales 1459 fueron libros escritos por mujeres, y 5667 por hombres.

El método automático utilizado para la catalogación del género real fue basado en el nombre del autor. Los libros descargados de Gutenberg llevan una encabezado estándar en todos los textos que identifica al autor con una línea que

comienza con “Author: ”. El nombre de pila indicado en esta línea fue buscado en dos listas: una con nombres de hombre y otra con nombres de mujer[5]. Si el nombre se hallaba en ambas listas, o en ninguna, ese libro era descartado. De lo contrario, el libro era etiquetado con el sexo que el nombre indicaba. Por último, se agregó a la automatización las formas corrientes de referencia por género, títulos nobiliarios y cargos militares, como ser ‘Mr’, ‘Miss’, ‘Mrs’, ‘Duque’, ‘Duchess’, ‘Count’, ‘Dame’, ‘Colonel’, ‘General’, etc.

La clasificación automatizada de los textos se hizo a través de la técnica de Machine Learning llamada *Random Forest Classifier* en su implementación dada por *sklearn*, una biblioteca (o *library*) para Python. Para llevarla a cabo, fue necesaria la extracción de características o ‘features’ cuantitativas de cada uno de los textos, que se realizó usando NLTK, una biblioteca de análisis del lenguaje para Python.

Las features elegidas fueron:

0) Cantidad de frases [6]
1) Cantidad de palabras [6]
2) Longitud promedio de palabras [6] [7]
3) Cantidad de palabras con letras repetidas [6]
4) Proporción de palabras con letras repetidas [6]
5) Proporción de preguntas [8]
6) Proporción de exclamaciones [8]
7) Cantidad promedio de palabras por frase
8) Cantidad promedio de signos de puntuación por frase [7] [9]
9) Cantidad de pronombres femeninos [10] [11]
10) Cantidad de pronombres masculinos [10] [11]
11) Relación entre pronombres femeninos y masculinos [10] [11]
12) Proporción de menciones a niños (en particular las palabras “child”, “children”, “baby”, “babies”, “son”, “daughter”)
13) Proporción de conceptos relacionados a la felicidad (en particular “happy”, “happiness”, “joy”, “exultant”, “exaltation”, “ecstatic”)
14) Proporción de conceptos relacionados a la tristeza (en particular “sad”, “sadness”, “depression”, “depressed”, “cry”, “crying”, “tears”, “cried”)
15) Proporción de conceptos relacionados a las relaciones amorosas (en particular las palabras “love”, “lovely”, “marriage”, “enamoured”, “fancy”, “care”, “seduce”, “girlfriend”, “boyfriend”, “fiance”, “fiancee”, “engaged”)
16) Proporción de menciones a colores (en particular “blue”, “brown”, “yellow”, “green”, “red”, “purple”, “violet”, “orange”, “black”, “white”, “lilac”, “turquoise”, “gray”, “pink”)
17) Proporción de conceptos relacionados a la demostración de cariño (en particular “hug”, “hugging”, “kiss”, “kissing”, “kissed”, “hugged”, “embrace”, “embracing”, “embraced”, “caress”, “caressing”, “caressed”)
18) Proporción de conceptos relacionados a sueños (en particular ”dream”, ”fantasy”, ”dreamed”, ”nightmare”, ”fantasize”, ”imagination”, ”imagine”, ”dreams”)
19) Proporción de adjetivos [7] [12] [13]
20) Proporción de adverbios [7] [12] [13]
21) Proporción de sustantivos [7] [12] [13]
22) Proporción de palabras referidas a números (cardinales) [7] [12] [13]

La mayoría de las features no requieren mayor detalle. Se hicieron todas por búsqueda por igualdad de la palabra o signo involucrado. La única excepción es el rol sintáctico de la palabra para caracterizar por adjetivos, adverbios, etc. Este taggeo se hizo con un POStagger (*Part-Of-Speech Tagger*) desarrollado por un grupo de la universidad de Stanford que, además de ser uno de los más veloces, tiene una precisión del 96.97%[14]. La elección de features fue en parte basada en la bibliografía leída, y en parte pensada especialmente para los experimentos de este trabajo (como el estudio de referencias a temáticas particulares).

Con los vectores de features de cada libro almacenados en un archivo txt, se procedió a utilizar las herramientas de la biblioteca *sklearn* para: separar los datos de validación de los de entrenamiento (se separó un 10% de los libros para validación) con una seed fija para utilizar siempre los mismos libros. Luego se procedió a separar los libros restantes en un conjunto de training (80% de los restantes) y otro de testing, para ser utilizados en las pruebas a realizar.

III. Resultados

A partir de las herramientas para Python, la biblioteca sklearn en particular, se seleccionó un tipo de clasificador: Random Forest Classifier. Se eligió el que daba mejores resultados consistentemente para 10 seeds. Es decir, con el mismo fraccionamiento entre libros de training y testing, el Random Forest Classifier dio el mejor, o muy cercano al mejor resultado que los otros dos clasificadores probados. Por otro lado, se hizo una prueba con un parámetro del clasificador hasta llegar a usar 20 estimadores. No fue una prueba exhaustiva ni exacta si no que se probaron con números por encima y por debajo del default (15), y se perdía precisión muy rápido a mayor número de estimadores por encima de 20 (los porcentajes de precisión bajaban más de 5%). Este clasificador realizó el entrenamiento y se obtuvieron los resultados que se muestran a continuación.

La precisión del clasificador sobre los datos de verificación fue del 94% (clasificaciones correctas sobre total de clasificaciones). Se obtuvo también la matriz de confusión, que se observa en la figura 1. La cantidad de falsos positivos, es decir, libros escritos por hombre pero clasificados como mujer, es un porcentaje ínfimo. Los falsos negativos, en cambio, fueron porcentualmente más altos. A partir de esto, se podría decir que se requiere un análisis más exhaustivo para extraer la identidad de los libros escritos por mujeres que para aquellos escritos por hombres. Esto podría deberse a dos causas: en primer lugar, que el tamaño de la muestra no haya sido lo suficientemente grande, especialmente porque conseguimos pocos libros escritos por mujeres y el ; y en segundo lugar, a que exista una mayor variabilidad de tópicos en los libros escritos por mujeres, lo que los hace más difíciles de caracterizar.

	Clasif. como mujer	Clasif. como hombre
Autor mujer	113	30
Autor hombre	12	522

Fig. 1. *Confusion Matrix*

Para comparar con una medida aleatoria, se realizó un "clasificador" extra que utilizó los mismos datos de entrenamiento para extraer el porcentaje de novelas escritas por mujeres (alrededor de un 20%) y se corrió mil veces sobre los datos de verificación clasificando entre hombre y mujer. Para cada libro, la clasificación se hizo eligiendo un número aleatorio entre 0 y 1, para los menores a 0.2, se determinó mujer, y para los mayores hombre. El promedio de las corridas resultó en una exactitud del 67%. Esto permite determinar que el clasificador armado en base a los features fue significativamente mejor que el azar.

	Ratio de clasificaciones correctas
Clasificador con features	0.93796
Azar	0.67112

Fig. 2. Clasificador en base a features versus clasificador aleatorio

Para obtener conclusiones sobre la pregunta inicial de "qué features ayudarían a distinguir el género del autor", se utilizaron las herramientas brindadas por el clasificador que permiten saber la importancia (o peso) de cada feature dentro del entrenamiento del clasificador. Con los datos ingresados para este trabajo se destacaron 3 features como indicadores principales para determinar el género, mientras que el resto no aportaron datos significativos: la proporción de pronombres femeninos sobre masculinos, las menciones a niños y la preponderancia de menciones a colores. Y entre esos 3, la proporción de pronombres femeninos fue el más significativo con una distancia significativa de los otros dos.

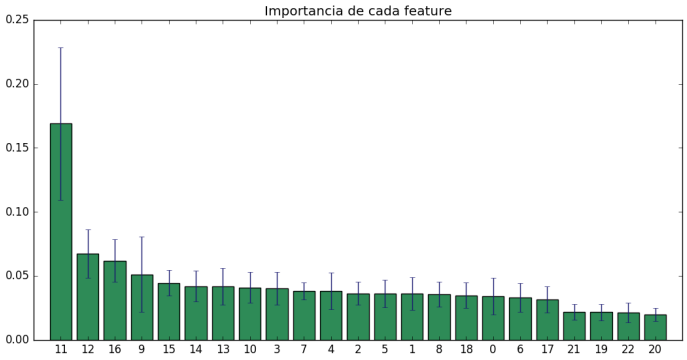

Fig. 3. Features ordenadas según importancia (la numeración se corresponde con el listado de la sección II)

Una vez obtenidos estos resultados, se decidió hacer una investigación un poco más exhaustiva sobre los features más significativos, para evaluar con más exactitud qué pasa con esos rasgos en particular.

En el caso de la comparación entre pronombres femeninos y masculinos, esta medida compara la cantidades de menciones a mujeres contra la cantidad de menciones a hombres, en base a pronombres. Los nombres propios no fueron evaluados, con lo cual, una mayor cantidad de pronombres de un género se puede atribuir a una mayor cantidad de personajes de ese género o a personajes de dicho género con mayor protagonismo y, por lo tanto, con mayor número de menciones. Como puede observarse en la figura 4, este feature no resulta distintivo porque las mujeres escriban sólo sobre mujeres y los hombres sólo sobre hombres, sino que, como lo muestran las curvas de frecuencia del gráfico con formas totalmente diferentes, se ve una clara curva descendiente entre los libros escritos por hombres que marca que la mayoría de los autores masculinos hablan mayoritariamente sobre hombres. El promedio de los libros escritos por hombres tienen una relación de aproximadamente 0.35 de pronombres femeninos sobre pronombres masculinos. Por otro lado, en los libros escritos por mujeres se ve una variabilidad mucho mayor: en la figura 4, que sólo refleja una tasa entre 0 y 1.6, se ve poca variación, con muchos libros con personajes masculinos y otros tantos con personajes femeninos. Pero el promedio está aproximadamente en 1.1, lo que quiere decir que las menciones están casi balanceadas entre ambos géneros con un leve favorecimiento por las mujeres. Esto podría deberse a los outliers de libros que tratan exclusivamente sobre mujeres.

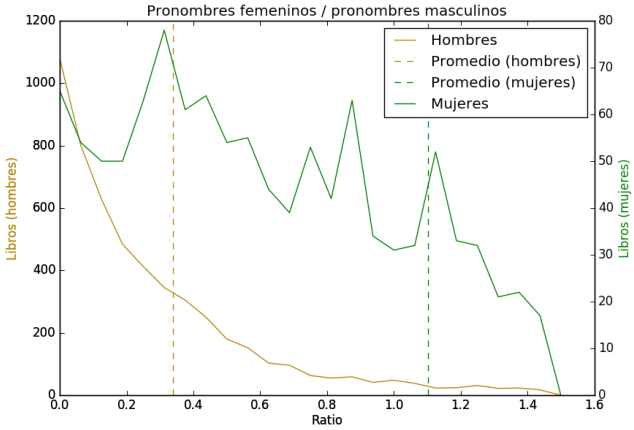

Fig. 4. Curvas de frecuencia y promedio del rate de pronombres femeninos sobre masculinos, según sexo del autor

Si se observa la mediana en lugar del promedio, la diferencia también es muy marcada y es aún más orientada a personajes masculinos. La mitad de los libros escritos por hombres tienen una relación menor a 0.2, mientras que en los escritos por mujeres la mitad tiene apenas menos de 0.8.

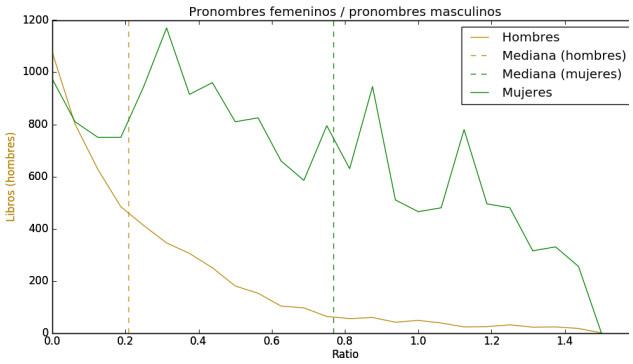

Fig. 5. Curvas de frecuencia y mediana del rate de pronombres femenino sobre masculinos, según sexo del autor

Las diferencias en otros features no resultaron tan marcadas aunque sí visibles. En el caso de las menciones a niños y bebés, la mediana de las menciones en libros escritos por mujeres y por hombres difiere notablemente, tal como puede observarse en la figura 6; sin embargo las curvas son relativamente parecidas, con el pico de la curva de las mujeres mostrando una leve prevalencia de menciones a niños, mientras que el pico de la curva de los autores masculinos es más cercano a ninguna mención.

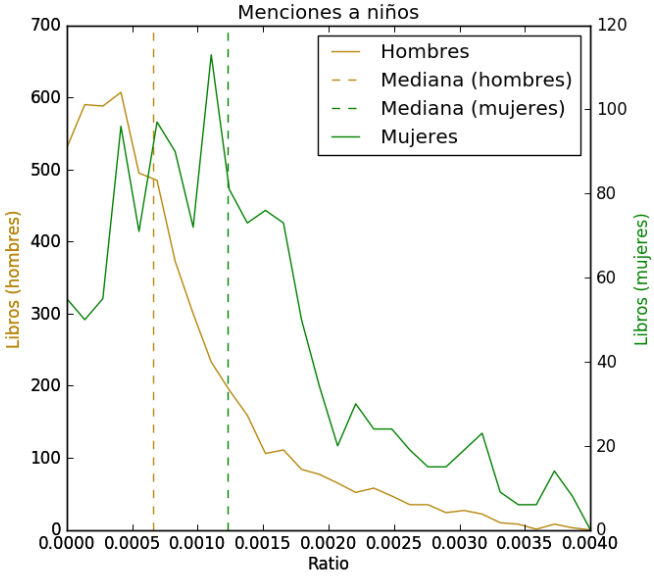

Fig. 6. Curvas de frecuencia y mediana del rate de menciones a niños, según sexo del autor

Un comportamiento similar se puede observar en el rate de menciones a colores que muestra la figura 7.

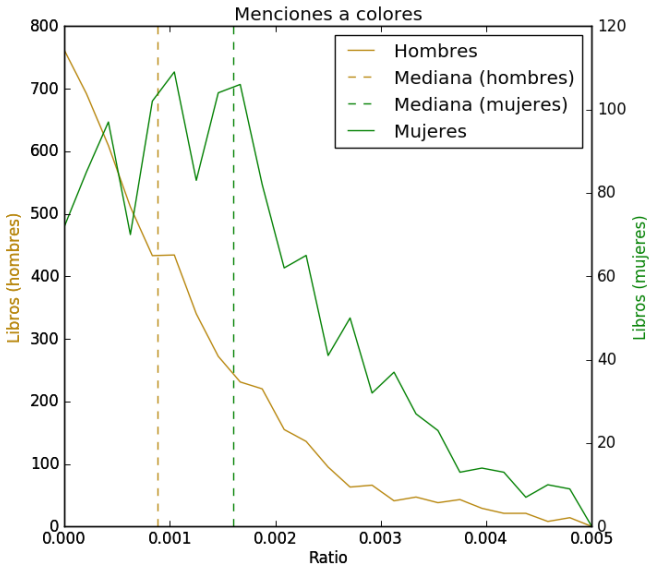

Fig. 7. Curvas de frecuencia y mediana del rate de menciones a colores, según sexo del autor

IV. Conclusiones

Los resultados observados no permiten vislumbrar diferencias en el estilo de escritura entre hombres y mujeres. En la bibliografía leída, y de la cual fueron obtenidos muchos de los features que se recolectaron para esta investigación, las diferencias de estilos eran de carácter más coloquial. Por ejemplo, buscar las palabras con letras repetidas pretendía encontrar escrituras especiales de algunas palabras que denotaran una emoción en particular, o que proveyeran algún efecto (tal como "looooooove", para expresar entusiasmo en el sentimiento). Pero en una novela, por ejemplo, este tipo de afectaciones del lenguaje no tienen lugar.

Por otro lado, al momento de buscar features que tuvieran más que ver con la temática del texto, parte de la bibliografía refería a los textos escritos por mujeres como más centrados en "sensaciones" y a los escritos por hombres como más centrados en la "acción". En este trabajo, no se profundizó esta línea de investigación (ya que la técnica utilizada no es la indicada para buscar cercanías por campo semántico, debido a que el análisis de porcentaje de palabras asociadas es una medida débil y que cubre pocos casos) y por ende la misma no resultó en diferencias perceptibles. Pero la búsqueda de otra diferencia temática, como es la presencia de personajes femeninos e infantiles, sí resultó en diferencias notables. Es evidente en los resultados que, al menos en los libros cuyos derechos de autor ya vencieron, los hombres escriben mayormente sobre personajes masculinos, o por lo pronto, que sus protagonistas son mayoritariamente hombres. Mientras que dentro de los libros escritos por mujeres hay mucha mayor variabilidad: hay libros con mayoría de referencias masculinas y otros con mayoría de referencias femeninas, pero promediando, la relación entre uno y otros es muy pareja. Si se observa la mediana, igual se puede ver que la mitad de los libros escritos por mujeres tienen mayor cantidad de

referencias a personajes masculinos. Es decir, en la gran mayoría de los 7000 libros analizados, la acción se centra alrededor de personajes masculinos.

V. Discusión

Este trabajo puede servir para aportar información a una línea de investigación respecto del rol de las mujeres en la cultura y en la literatura en particular; sin embargo, quedan abiertos varios análisis por hacer. En primer lugar, se podría tomar una mayor cantidad de libros ampliando la selección dentro de Project Gutenberg o buscando textos de autores más contemporáneos, pero también se podría hacer un análisis más exhaustivo sobre la cantidad exacta de personajes que se reconocen por nombres en una novela, y con qué género se los identifica; investigar si dentro de aquellos escritos en primera persona se favorece a los personajes femeninos o masculinos; y evaluar la preponderancia de cada personaje que aparece en el texto, entre otros.

References

[1] Montemurro, M., & Zanette, D, 2009. Towards The Quantification Of The Semantic Information Encoded In Written Language. Advances in Complex Systems, 135-135
[2] John D. Burger, John C. Henderson, George Kim, and Guido Zarrella, 2011. Discriminating gender on Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing
[3] Michael S. Hart and Gregory B. Newby. Project Gutenberg. http://www.gutenberg.org/
[4] Michael S. Hart and Gregory B. Newby. Project Gutenberg: Information about robot access to out pages. https://www.gutenberg.org/wiki/Gutenberg:Information_About_Robot_Access_to_our_Pages
[5] Anónimo. Word Lists. http://www.outpost9.com/files/WordLists.html
[6] Christian S., 2012. Gender Prediction from Blog Posts
[7] Zhangand, C. and Zhang, P., 2010. Predicting gender from blog posts
[8] Deitrick, W., Miller, Z., et al. Author Gender Prediction in an Email Stream Using Neural Networks
[9] de Vel, et al., 2002. Language and gender author cohort analysis of e-mail for computer forensics. Proceedings Digital Forensics Research Workshop, Syracuse, NY, USA
[10] Argamon, S., Koppel, M., Fine, J., & Shimoni, A., 2003. Gender, Genre, And Writing Style In Formal Written Texts. Interdisciplinary Journal for the Study of Discourse, 321-346
[11] Koppel, M., Argamon, S. and Shimoni, 2004. Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17, 4
[12] Miller et al., 2012. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. International Journal of Intelligence Science, 2012, 2, 143-148
[13] Schler, et al. Effects of Age and Gender on Blogging
[14] Stanford NLP Group. Stanford Log-linear Part-Of-Speech Tagger. http://nlp.stanford.edu/software/tagger.shtml