Aproximación a la tarea de *Author Profiling*: Identificación de género y variedad del lenguaje en Twitter

Lorena Ramírez Mondragón
loramon@masters.upv.es

Abstract

Esta aproximación presenta la tarea de identificación de género y variedad del lenguaje a partir de un dataset recolectado de Twitter, para el caso particular del español. La tarea es abordada mediante la especialización de bolsas de palabras.

1. Introducción

El reto de author profiling ha existido a lo largo de las últimas décadas, forma parte de un conjunto de desafíos que busca identificar o bien la identidad o bien rasgos específicos de los autores de determinados textos a través de ciertas características en el contenido y estilo de escritura. Los avances tecnológicos de los últimos años han permitido la evolución de los métodos empleados para dichos fines, haciéndolos cada vez más complejos y robustos, para lograr así un mayor alcance. Un mayor alcance en términos de capacidad y velocidad de procesamiento de textos, y de diversidad de técnicas y variables a incluir en el estudio. Hoy por hoy es posible incluir análisis de patrones lingüísticos mediante la observación de las características morfológicas, sintácticas y semánticas de los textos.

El perfilamiento de usuarios puede ser útil y aplicable en diversos sectores de la economía. En el campo financiero, una *Venture Capital* podría ver el impacto de una *start-up* en el mercado y determinar así invertir o no en esta; en el sector comercial, una empresa podría comprobar las preferencias por un producto u otro para enfocar así sus esfuerzos en la producción, y adicionalmente segmentar el mercado y sus respectivas campañas de una manera más eficiente; en los sectores político y social, el gobierno podría detectar el impacto y repercusión de sus políticas e iniciativas sociales.

Este artículo está enfocado en perfilar los autores de un conjunto de tweets según la variedad del lenguaje español empleado en los tweets y el género de los autores. El conjunto de tweets usado para este fin es en realidad un subconjunto del dataset PAN-AP'17 que incluye 4 lenguajes diferentes, inglés, español, portugués y árabe con sus respectivas variedades según su origen.

El resto del artículo está organizado de la siguiente forma. En la sección 2 presenta las características del dataset usado para este estudio, la sección 3 describe el enfoque y la propuesta de solución al problema planteado, la sección 4 muestra los resultados y análisis obtenidos, la sección 5 concluye sobre el trabajo realizado y presenta alternativas para mejorar los resultados en un futuro.

2. Dataset

El dataset PAN-AP'17 ha sido recolectado de Twitter (Rangel et al.), y reducido para este estudio al subconjunto de tweets escritos en español de siete diferentes países: Argentina, Chile, Colombia, España, México, Perú y Venezuela. El conjunto de datos de entrenamiento consta de 279,999 tweets correspondientes a 2800 autores, y está distribuido homogéneamente tanto como en la clase género como en la variedad del lenguaje. Asimismo el *dataset* de prueba consta de tweets correspondientes a 1400 autores dferentes y la distribución de los mismos sigue el mismo patrón que el conjunto de entrenamiento.

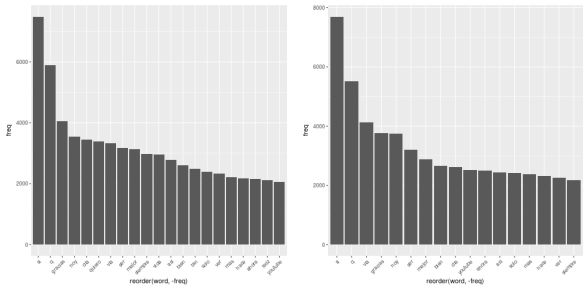
Para analizar el *corpus* de entrenamiento empezamos observando las palabras más frecuentes en cada clase mediante mapas y conteos de palabras. Para el caso del género, los mapas de palabras presentados en la Figura 1 permiten detectar que tanto hombres como mujeres tienden a usar más las mismas palabras, por ejemplo, si, hoy, gracias saltan a la vista rápidamente. Para complementar esta visualización, analizamos el número de veces que se usan las palabras más frecuentes (Ver Figura 2); en este punto observamos dos aspectos importantes: el primero tiene que ver con hecho de que unas

pocas palabras concentran la mayor parte de la frecuencia, mientras que el segundo está relacionado con el mito de que las mujeres hablan más que los hombres, a simple vista se observa que la Figura 2a correspondiente a las mujeres tiene una mayor cantidad de palabras que la Figura 2b para el caso de los hombres. En el caso de la variedad de lenguaje, observamos que por regla general el país o su capital hacen parte de las palabras más mencionadas. La Figura 3 muestra que entre las palabras más frecuentes en España están Madrid y España, así como en la Figura 4 se observa que Colombia tiene una de las mayores frecuencias en su propia variedad de lenguaje.

El análisis exploratorio nos permitió identificar que youtube es sin duda alguna uno de los *trending topics* para todos los países lo que confirma una vez más la transformación digital que estamos viviendo. Por otra parte, observamos también una tendencia en la preferencia por usar Twitter para expresar opiniones políticas, en Colombia, por ejemplo, aparecen juanmansantos y enriquepenalosa, haciendo referencia a Juan Manuel Santos y Enrique Peñalosa, presidente de Colombia y alcalde de Bogotá en el momento de la recolección de datos, respectivamente; en Venezuela, el nombre del país se lleva la mayor parte de la frecuencia y está acompañada por otras palabras como gobierno, nicolasmaduro, elcooperante, las dos primeras son una clara referencia a la situación política del país, mientras que la última es una mención a uno de los diarios digitales del país; en el caso de México, se destacan trump y epn, haciendo referencia al jefe de Gobierno de su país vecino al norte y a su presidente del momento Enrique Peña Nieto. Un aspecto que sería interesante contrastar en este análisis es la edad de los autores de los tweets para corroborar tendencias polítcas segmentadas por generaciones.

(a) Female **(b)** Male

Figura 1: Palabras más frecuentes por género. *Dataset* sin procesar



(a) Female **(b)** Male

Figura 2: Frecuencia de palabras por género. *Dataset* sin procesar

Figura 3: Palabras más frecuentes en España. *Dataset* procesado

3. Propuesta del alumno

En esta sección se describe el punto de referencia, el enfoque utilizado para resolver el problema, las características utilizadas, y las razones que las sustentan.

3.1. *Baseline*

El punto de referencia está basado en una bolsa de palabras (*BoW*) por clase construida a partir de un vocabulario de n términos en el que todas las palabras están en minúscula, no hay signos

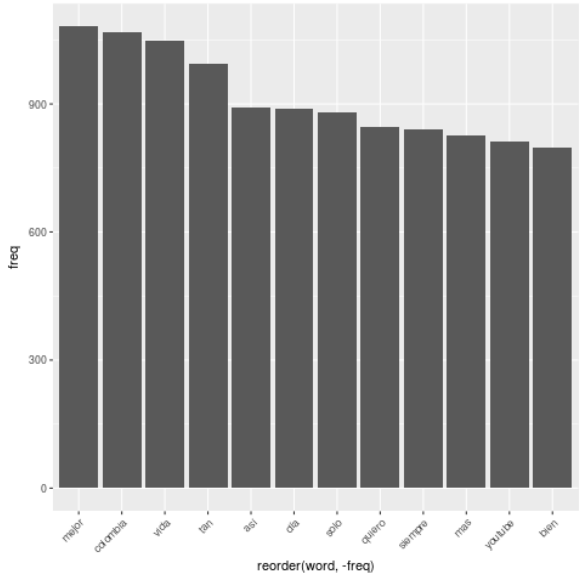


Figura 4: Frecuencia de palabras en Colombia. *Dataset* procesado

de puntuación ni números y se han eliminado las *stopwords* del lenguaje español. Esta bolsa de palabras fue empleada para entrenar una SVM con *cross validation* de diez *folds* y una repetición.

3.2. Propuesta

Nuestra propuesta consiste en una especialización del *BoW* original. Nos enfocamos en pulir la construcción del vocabulario a usar para la conformación de las bolsas de palabras con base en las observaciones realizadas durante la fase de exploración de datos. El primer paso fue eliminar una lista de palabras con aquellos términos frecuentes comunes a las diferentes clases: ”si”,”q”,”d”,”x”,”hoy”,”gracias”,”vía”,”ser”. Está elección está basada en que estas palabras estaban en el top 10 de palabras más frecuentes generando picos muy altos de frecuencia y de alguna forma podrían generar ruido en los modelos. Incluir esta lista fue de utilidad para poder ver diferencias más significativas *a priori* y elegir así otros caminos para seguir.

El segundo paso, exclusivo para la clase género, nace a raíz del vocabulario resultante del paso anterior, este nos permitió detectar que varias de las palabras más usadas seguían siendo comunes entre ellos y ellas, por cuanto no había una evidencia clara que permitiera diferencias las palabras más usadas por un género u otro. Así las cosas, decidimos construir un vocabulario con las palabras más frecuentes exclusivas de los hombres y de las mujeres a partir de un vocabulario de n palabras.

Una vez construidos los vocabularios, la siguiente tarea consistió en entrenar diferentes modelos para determinar el impacto de las estrategias elegidas.

4. Resultados experimentales

En esta sección se presentan los resultados obtenidos y se analiza el impacto y la contribución de la propuesta realizada.

4.1. Por variedad del lenguaje

La eliminación de las palabras más frecuentes comunes a las diferentes clases contribuyó para mejorar el baseline obteniendo una mayor precisión con un menor número de palabras. En el *baseline* por variedad, la mejor precisión, 0.7275, fue obtenida usando un vocabulario de 5000 palabras y requirió casi 1 hora de procesamiento; mientras que en nuestro caso para todos los modelos empleados usando 1000 palabra, los modelos SVM y C5.0 costaron menos de la tercera parte del tiempo mientras que *Random Forest* requirió aproximadamente 1 hora y arrojó la mejor precisión de los 3 modelos.

	TRAIN		***TEST***	
Modelo	*Accuracy*	*Kappa*	*Accuracy*	*Kappa*
SVM	0.7518	0.7104	0.7721	0.7342
C5.0	0.7589	0.71875	0.7621	0.7225
RF	0.8764	0.8558	0.8936	0.8758

Cuadro 1: Resultados para Variedad del Lenguaje. n = 1000

4.2. Por género

En este caso se confirmó la hipótesis de que las palabras exclusivas por género aportan un mejor resultado que el vocabulario original procesado. La construcción del vocabulario exclusivo arrojó 185 palabras a partir de un vocabulario de 500, el vocabulario resultante está compuesto por palabras que dicen los hombres que no dicen las mujeres y viceversa.

	TRAIN		***TEST***	
Modelo	*Accuracy*	*Kappa*	*Accuracy*	*Kappa*
SVM	0.7007	0.40143	0.7064	0.4129
C5.0	0.6525	0.305	0.6657	0.3314
RF	0.7039	0.4079	0.6964	0.3929

Cuadro 2: Resultados para Género con vocabulario exclusivo. n original = 500, n exclusivo = 185

La mejor precisión obtenida, 0.7064, fue alcanzada por SVM, y aunque supera la precisión del vocabulario procesado con n = 1000, no llega a superar las precisiones alcanzadas por el modelo de referencia en el que con un vocabulario de 100 palabras se logró una precisión de 0.7375 en 6 minutos.

	TRAIN		***TEST***	
Modelo	*Accuracy*	*Kappa*	*Accuracy*	*Kappa*
SVM	0.6646	0.3293	0.6586	0.3171
C5.0	0.6443	0.2886	0.6764	0.3529

Cuadro 3: Resultados para Género. n = 1000

5. Conclusiones y trabajo futuro

En este artículo abordamos la tarea de identificación de la variedad del lenguaje español y el género a partir de un dataset recolectado de Twitter. El enfoque principal consistió en usar bolsas de palabras construidas con base en vocabularios de palabras frecuentes usando diferentes tratamientos para la selección de términos.

El análisis realizado permite concluir que la tarea puede ser potenciada al utilizar enfoques diferentes para cada caso. En el primer caso, fue de utilidad hacer una sencilla modificación en la bolsa de palabras para mejorar la precisión del modelo, mientras que para el caso del género la contribución no fue la misma, por el contrario, puede decirse que la precisión se redujo. Esto nos lleva a pensar que lo que explica las diferencias entre la escritura de ellos y ellas no están asociadas con las palabras usadas sino con la forma de usarlas, escribiendo oraciones más o menos largas y conectándo las palabras de formas diferentes.

El trabajo futuro en esta materia puede enfocarse agregando características relacionadas con la longitud de los tweets y las frecuencias de las subsecuencias de palabras (n-gramas), y especializando el análisis de frecuencias con TFIDF.

References

Francisco Rangel, Paolo Rosso, Martin Potthast and Benno Stein. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In *Cappellato L., Ferro N., Goeuriot L., Mandl T. (Eds.) CLEF 2017 labs and workshops, notebook papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1866*, 2017.