Text Mining en Social Media.
Máster Big Data Analytics

Alexander Marco Carpena
marcaralexander@gmail.com

Abstract

En este paper se describen las técnicas usadas para solventar la problemática del **Author Profiling** a partir de los datos del *PAN-AP-17. La metodología planteada se basa en aplicar una serie de técnicas de text mining sobre el dataset proporcionado con el objetivo de crear modelos que permitan predecir, ante nuevo texto, las características del mismo. A lo largo de la tarea, se ponen de relevancia el uso de técnicas para distinguir y filtrar palabras particulares frecuentes en cada una de las clases. Para esto, será necesario previamente realizar un buen estudio de los datos y preparlos adecuadamente, y finalmente, crear modelos con los que predecir. En este aspecto, el modelo con el que se ha obtenido mejor resultado, se basa en un modelo con* ***Random Forest*** *que tiene de entrada datos a los que se les ha aplicado la técnica de* ***tf-idf****.*

1. Introducción

Uno de los problemas actuales que han tenido un gran auge con el desarrollo del big data, es el análisis de autoría o Author Profiling, el cual aborda la clasificación de los textos en función de las elecciones estilísticas de sus autores. Más allá de las tareas de identificación y verificación de un cierto autor, donde se estudia el estilo particular del mismo, el perfil de autores distingue clases y estudia el aspecto sociolectal, es decir, cómo el lenguaje es compartido por las personas. Esto ayuda a identificar aspectos de creación de perfiles como el sexo, la edad, el idioma nativo o el tipo de personalidad.

En este trabajo, se aborda un pequeño aspecto de este estudio con el objetivo de identificar el género del autor y su variedad lingüística, de un dataset de Twitter en español. Para esta tarea, se nos proporciona un corpus anotado con el género del autor y su variedad de idioma (Argentina, Chile, Colombia, México, Perú, España y Venezuela)

2. Dataset

Antes de cualquier análisis, es conveniente describir el dataset usado. En nuestro caso, los datos proporcionados vienen ya divididos en dos carpetas, training y set, con un fichero XML para cada autor, el cual tiene una serie de tweets. Además, se nos proporciona un archivo con las etiquetas del training asociados al identificador de cada archivo XML, con lo que podemos etiquetar cada autor en particular.

En cuanto a las dimensiones del dataset, el conjunto de training contiene un total de 2800 muestras, 400 de cada variedad, de las cuales, 200 son de cada género, es decir, que del total de las 2800 muestras, también existe equidad y hay 1400 muestras de cada género. Esto también ocurre en el corpus de test, en este caso de un total de 1400 muestras, hay 200 de cada variedad de idioma y del total, 700 son hombres y 700 mujeres.

Este aspecto del equilibrio de muestras es de gran impacto a la hora de realizar el análisis, ya que nos evita problemas de desbalanceo y normalización de datos.

3. Propuesta del alumno

En un primer instante, decidimos enfocar el problema de forma distinguida, aunque finalmente calculamos la precisión del modelo tanto para los dos problemas y poder así sacar conclusiones complementarias.

En primer lugar nos enfocamos a identificar el género del autor, para esto, sobre el script base de R, probamos a modificar parámetros y usar diferentes algoritmos de clasificación. En concreto, probamos cómo afecta el uso de las stopwords, el tipo de clasificador, el número de palabras más

usadas que crearán la bolsa de palabras y probamos con pequeñas bolsas de palabras propias.

En este primer caso, la bolsa de palabras (bw) creada es muy simple:

- **Hombres**: 'enemigo', 'libertad', 'ganar', 'perder', 'batalla'
- **Mujeres**: 'maravilloso', 'feliz', 'cumpleaños', 'nerviosa', 'hija', 'bebé', 'agradecida'

Otra de las pruebas se basa en distinguir el género y la variedad según la **longitud de los tweets**, para lo cual ya nos pasamos a Pyhton. Esta prueba viene dada por la hipótesis de que las mujeres suelen escribir más que los hombres, y por tanto los tweets más largos deben ser de ellas.

Para validar esta hipótesis, calculamos la media, mediana, desviación típica y la asimetría de los tweets y creamos un modelo RandomForest con 500 árboles tanto para el género y la variedad. Con este modelo, los resultados obtenidos son incluso peores que los obtenidos con el baseline, por lo que descartamos este modelo.

El siguiente modelo que probamos trata de generar una **bag of words** con las palabras más relevantes usadas por hombres o mujeres. Para obtener estas palabras, primero calculamos la frecuencia de todas las usadas por hombres y mujeres. De estas, guardamos aquellas que son 3,2 veces más frecuentes hombres que en mujeres y viceversa.

Para el problema de la detección de la variedad del idioma, es importante el uso de técnicas como el **TFIDF**, el cual expresa cómo de importante es una palabra en un texto, dentro de una colección de textos. En este caso, esta técnica nos ayuda a distinguir las palabras más relevantes de cada variedad, creando una bolsa de palabras con más significado que las recogidas sin esta técnica.

Para la creación de los modelos, aunque, tal y como se ha visto anteriormente, el uso de las stopwords ayudan en la clasificación de la variedad, al usar el TFIDF, estas palabras no tienen mayor relevancia en el texto, por lo que decidimos quitarlas y así ayudar al sistema a ser más eficiente.

Para aplicar el TFIDF, usamos el TfdifVectorizer, el cual convierte un raw de documentos en una matriz de objetos TFIDF. Tras aplicar esta transformación, entrenamos un modelo RandomForest de 500 árboles tanto para género y variedad, aún sabiendo que este modelo debe ser más fuerte para la detección de la variedad y no tanto en la de género.

4. Resultados experimentales

De todas los ensayos realizados en la primera aproximación, es decir, variando los parámetros iniciales de la baseline, obtenemos los siguientes datos:

Preprocesado	**Gender**	**Variety**
Baseline (SVM n=10)	0,5635	0,2157
SVM n=10 with sw	0,6407	0,2492
RF n= 10	0,5721	0,2064
RF n=10 with sw	0,6514	0,2721
RF n=10 with sw & bw	0,6571	0,275
SVM n=10 with sw & bw	0,6507	0,2642
RF n=100 with sw & bw	0,6571	0,485

A la vista de estos resultados sacamos varias conclusiones:

- El uso de las stopwords ayuda a mejorar el modelo. Esto indica que existen diferencias entre hombres y mujeres en el uso de estas palabras, ya sea en la variedad o en la cantidad de las mismas.
- De los dos clasificadores probados, el Random Forest presenta mayor precisión para el mismo 'n'.
- El aumento de 'n' implica bastante mejora en la detección de la variedad, aunque no tanto en la de género. Esto es debido al margen de mejora existente y no tanto al preprocesado realizado.
- El uso de las bolsas de palabras, por muy simple que esta fuese, ha ayudado a mejorar la predicción, por lo que es un factor a tener en cuenta para ambos problemas.

El modelo creado con esta bolsa de palabras está orientado únicamente a la detección del género, por lo que lo validamos solamente en esta prueba. Creamos un modelo RandomForest con 500 árboles y aplicándolo sobre el test.

Preprocesado	**Gender**
Bag of words	0,66

Este resultado mejora por poco el obtenido en las primeras pruebas con bolsas de palabras más

simples. A pesar de esto, de este proceso obtenemos una gran información, y es que entre las palabras más frecuentes por mujeres y que no usan tanto los hombres se resaltan los emoticonos, lo que nos da una idea de los siguientes pasos a seguir.

Finalmente, nuestro mejor modelo se basa en usar TF-IDF, donde la precisión mejora considerablemente frente a los demás modelos:

Preprocesado	**Gender**	Variety
Tf-Idf	0,94	0,74

A la vista de estos resultados, concluimos que la variedad del idioma viene dada principalmente por la variedad de palabras de cada una. Sin embargo, también concluimos que en la diferencia de género este aspecto también es notable, tal vez, debido a que cada género tienen temas particulares de los que suelen hablar más. Aún así, la precisión obtenida no permite sacar este tipo de conclusiones.

5. Conclusiones y trabajo futuro

A la vista de estos resultados, concluimos que la variedad del idioma viene dada principalmente por la variedad de palabras de cada una. Sin embargo, también concluimos que en la diferencia de género este aspecto también es notable, tal vez, debido a que cada género tienen temas particulares de los que suelen hablar más. Aún así, la precisión obtenida no permite sacar este tipo de conclusiones.

El trabajo realizado hasta ahora, nos ha deslumbrado algunas claves en la distinción de variedad y género, sin embargo también se han quedado muchas pruebas por hacer.

Tal vez, una de las más prometedoras sería el análisis de los emoticonos, los cuales ya se ha visto que tienen una gran frecuencia de aparición en los textos femeninos y no tantos en los masculinos.

Además, se podrían usar otros métodos de machine learning para la realización de lo modelos, como las Redes Neuronales, o realizar un estudio sobre las URLs citadas en los tweets.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.