Author Profiling: Distinción entre género y procedencia geográfica

Javier Galarza Hernández[1]

***Abstract*— En el presente paper se presenta el proyecto final de la asignatura de Text Mining in Social Media del plan de estudios de Máster en Big Data Analytics por la Universidad Politécnica de Valencia. Éste consiste en aplicar técnicas de Machine Learning para distinguir rasgos específicos sobre los autores de un conjunto de tweets facilitados, más concretamente la distinción de sexo y procedencia geográfica. Se establece un baseline de 66 % de precisión en predicción de género y un 77 % de precisión en diversidad. A partir de éstos, se elaboran varios modelos y se escoge el mejor de ellos como resultado final, consiguiendo un 70 % de precisión en género y un 80,7 % en diversidad.**

I. INTRODUCCIÓN

Se define como Author Profiling la disciplina englobada dentro del Text Mining que se encarga de la extracción de características de un/a autor/a a partir de un texto escrito. De esta manera, es posible distinguir edad, género, personalidad, corrientes ideológicas o procedencia geográfica a partir del estilo de escritura, palabras más usadas, rasgos lingüísticos definidos y un largo etcétera.

Esta área está en auge en la actualidad debido a sus amplias utilidades en nuestra sociedad como puede ser seguridad anti-terrorista, detección de patrones culturales y sociológicos, detección de robo de identidad o detección de posibles delitos de odio con alevosía detectando ironía o ausencia de ésta.

En el proyecto propuesto en la asignatura, el problema se acota en la identificación de género del autor/a y en su procedencia geográfica a partir de un conjunto de tweets. Debido a que no existe correlación ninguna entre la procedencia geográfica y el género del autor en cuestión, se tratarán como dos problemas diferentes con diferentes aproximaciones y modelos. Se prestará especial atención a la repetición de palabras y a la limpieza previa de los tweets para mejorar los resultados de discriminación, así como a la selección de los algoritmos que conformarán el núcleo del modelo Machine Learning.

II. CONJUNTO DE DATOS

Se parte como base de un conjunto de datos facilitados por PAN [1], organización dedicada al análisis de textos y generadora de una comunidad internacional para estos fines. La fuente de información para la elaboración del conjunto de datos ha sido Twitter, recopilando tweets de miles de autores con cientos de tweets por autor cubriendo un gran espectro de temas. Los datos corresponden específicamente a la versión 2017, recopilando en concreto 50,378 autores de los cuales 24,429 son mujeres y 25,949 hombres. De esta manera hay una representación de un 51,50 % de mujeres y un 48,5 % de hombres, distribución que roza la equidad de género. En el mismo conjunto de datos contempla cuatro idiomas: Español, Portugués, Inglés y Árabe. En el caso de estudio propuesto, abarcará únicamente autores de procedencia geográfica hispanohablante, es decir, sólo se contarán con tweets escritos en lengua Española. Dentro de esta segmentación inicial, encontramos las siguientes posibles procedencias: Argentina, Chile, Colombia, Méjico, Perú, España y Venezuela.

[1]Máster en Big Data Analytics, Universidad Politécnica de Valencia, Valencia, España jagaher@teleco.upv.es

III. PRE-PROCESAMIENTO Y MODELOS

Tras una primera exploración del conjunto de datos, se observan varias casuísticas que podrían interferir en el buen comportamiento de los algoritmos Machine Learning a la hora de procesar texto. Por este motivo, se procede a realizar una limpieza de:

- Enlaces
- Acentos
- Palabras con vocales repetidas
- Emoticonos
- Signos de puntuación y exclamación

En el caso de los enlaces, signos de puntuación y exclamación y los emoticonos, se procede a la eliminación de los mismos, mientras que con acentos y vocales repetidas, se procede a la sustitución de la vocal sin ninguna de esas casuísticas [2].

Adicionalmente a esa limpieza, se realiza también una técnica de pre-procesamiento de texto denominada *Stemming*. Esta técnica consiste en quedarse únicamente con las raíces de las palabras, de manera que es más sencillo categorizar las familias de palabras que proceden de la misma raíz. Una vez se realiza este primer pre-procesamiento, se genera una *Bag of Words*, una lista de palabras más frecuentes para utilizar en la predicción de sexo y procedencia geográfica. En este proyecto específico, al realizar *Stemming*, lo que se obtendrá en la *Bag of Words* son las raíces de las palabras más frecuentes. Esta lista de raíces de palabras será en lo que se basen los algoritmos para predecir si un texto procede de un autor o una autora, o la procedencia geográfica concreta del autor/a dependiendo del problema que se esté abarcando.

A continuación, se escogen tres diferentes algoritmos de Machine Learning de los cuales se elegirá el mejor como resultado final.

III-A. Random Forest

Los Random Forest se definen como una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la

misma distribución para cada uno de estos. Es uno de los algoritmos de aprendizaje más certeros que hay disponible. Para un set de datos lo suficientemente grande produce un clasificador muy certero. Computa las proximidades entre los pares de casos que pueden usarse en los grupos, localizando valores atípicos, o (ascendiendo) dando vistas interesantes de los datos. No obstante, Si los datos contienen grupos de atributos correlacionados con similar relevancia para el rendimiento, entonces los grupos más pequeños están favorecidos sobre los grupos más grandes [3].

Se ha utilizado el método *CForest* del paquete *Party* para el entrenamiento y la predicción de este algoritmo en *R*.

III-B. Regresión por mínimos cuadrados

Mínimos cuadrados es una técnica de análisis numérico enmarcada dentro de la optimización matemática, en la que, dados un conjunto de pares ordenados y una familia de funciones, se intenta encontrar la función continua, dentro de dicha familia, que mejor se aproxime a los datos, de acuerdo con el criterio de mínimo error cuadrático. Desde un punto de vista estadístico, un requisito implícito para que funcione el método de mínimos cuadrados es que los errores de cada medida estén distribuidos de forma aleatoria [4].

Se ha utilizado el método *PLS* del paquete de mismo nombre en *R* para la elaboración de este modelo.

III-C. Support Vector Machines

Se define las máquinas de vector soporte o Support Vector Machines (SVM) como un modelo que representa a los puntos de muestra en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama vector soporte. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase. En ese concepto de separación óptima es donde reside la característica fundamental de las SVM: este tipo de algoritmos buscan el hiperplano que tenga la máxima distancia con los puntos que estén más cerca de él mismo. Por eso también a veces se les conoce a las SVM como clasificadores de margen máximo. De esta forma, los puntos del vector que son etiquetados con una categoría estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado [5].

Se ha utilizado el método *svmLinearWeights* del paquete *e1071* para la elaboración de este modelo en *R*.

IV. RESULTADOS EXPERIMENTALES

Aplicando el pre-procesado explicado anteriormente y para todos los modelos propuestos, se ha superado el *baseline* propuesto por los docentes de la asignatura: 66,43 % en distinción de género y 77,21 % en distinción de procedencia geográfica.

IV-A. Procedencia geográfica

Abarcando la problemática sobre la distinción de la procedencia geográfica, únicamente se ha implementado un único modelo por restricciones de tiempo. El algoritmo implementado ha sido un *Random Forest*. Se ha escogido este algoritmo en concreto por ser uno de los que mayor eficiencia computacional aporta y por su sencillez. Se ha conseguido un 80,7 % de aciertos sobre total de casos.

IV-B. Género

Abarcando la problemática sobre la distinción de género del autor/a, se han implementado los tres algoritmos explicados en el apartado anterior: *Random Forest*, *Regresión por mínimos cuadrados* y *SVM*.

Se ha conseguido un total de 70,5 % de aciertos sobre total utilizando un *Random Forest*, destacando su rapidez de cómputo y su buen resultado a pesar de ser un algoritmo de remarcada sencillez. En el caso de la *Regresión por mínimos cuadrados*, se obtiene un resultado de 67,3 % de precisión, igualado en eficiencia de cómputo pero por detrás del *Random Forest* en sencillez y en resultado final. Por último, con la implementación de las *SVM*, se obtiene un 66,71 % de precisión. Este resultado, además de estar por detrás de los dos anteriores, se ve impactado de manera negativa como opción debido a su alto coste computacional por complejidad de cálculo.

TABLE I
RESULTADOS

	Random Forest	Mín. cuadrados	SVM
Género	70,5 %	67,3 %	66,71 %
Procedencia Geográfica	80,7 %	-	-

Observando los resultados obtenidos en la tabla I, se concluye que el mejor algoritmo que ha funcionado tanto para la problemática de identificación de género como de procedencia geográfica ha sido el *Random Forest*. El *Random Forest* además destaca también no sólo por su precisión de resultado sino también en términos de eficiencia computacional, por lo que lo sitúan como algoritmo escogido como resultado de este proyecto.

V. CONCLUSIONES Y TRABAJO FUTURO

En este proyecto se ha abarcado la problemática de *Author profiling* haciendo énfasis en la distinción de género y de procedencia geográfica de países hispanohablantes. En primer lugar se ha realizado un análisis exploratorio para tener conocimiento sobre las características del conjunto de datos con el que se iba a trabajar. Seguidamente, se propone un preprocesamiento básico para poder contar con un conjunto de datos sólido con el que los algoritmos de Machine Learning puedan operar y obtener los resultados óptimos. Se plantean tres algoritmos: *Random Forest*, *Regresión por mínimos cuadrados* y *SVM*. De éstos, el que mejor resultados provee y el que mejor opera en términos de eficiencia computacional resulta ser el *Random Forest* con diferencia.

V-A. Trabajo Futuro

Como trabajo futuro se propone la elaboración de otros algoritmos para la casuística de procedencia geográfica y comprobar si sigue la misma tendencia que la problemática de género. Adicionalmente, podrían mejorarse los resultados mediante agregación *Bootstrap* [6], meta-algoritmo que consiste en el promediado de diversos algoritmos como implementación de un único modelo, pudiendo ser capaz de proveer resultados más robustos y ajustados. Se plantea también la extracción de más información acerca del autor a partir de la elaboración de un *web crawler* [7] para la extracción de más datos como puede ser el color de fondo de perfil, descripción del usuario o nombre de usuario.

REFERENCES

[1] PAN website, Author Profiling: http://pan.webis.de/clef17/pan17-web/author-profiling.html

[2] Gurusamy, Vairaprakash and Kannan, Subbu on *Preprocessing Techniques for Text Mining, Oct 2014*

[3] Breiman, Leo. *Random forests. Machine learning*, 2001, vol. 45, no 1, p. 5-32.

[4] ABDI, Hervé. "Partial least square regression (PLS regression)". *Encyclopedia for research methods for the social sciences*, 2003, vol. 6, no 4, p. 792-795.

[5] Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features."*Machine learning: ECML-98* (1998): 137-142.

[6] Efron, Bradley, and Robert Tibshirani. "The bootstrap method for assessing statistical accuracy."*Behaviormetrika* 12.17 (1985): 1-35.

[7] Shkapenyuk, Vladislav, and Torsten Suel. "Design and implementation of a high-performance distributed web crawler."*Data Engineering, 2002. Proceedings. 18th International Conference on. IEEE, 2002*.