Asignatura Text Mining II
Máster/Diploma Big Data Analytics

Pedro Henrique Mano Figueiredo Fernandes
pedromorfeu@gmail.com

Abstract

La aproximación al problema de *Text Mining* para *Author Profiling* ha tenido como base la técnica conocida por *bag of words*. Esta técnica es un modelo que aprende el vocabulário de un conjunto de documentos, a través del que se construye una representación de los datos en forma de matriz, adecuada para aplicar *machine learning*. A la matriz de vocabulario se han añadido otras características, como contadores de polaridad de sentimientos y determinadas estadísticas del texto de los documentos. El *metadata* de los tweets se ha usado también para reforzar las características de la matriz.
El modelo usado para el aprendizaje del vocabulário se basa en contadores de palabras, calculando un coheficiente de tipo TF-IDF. Se ha configurado el modelo con un máximo de 2000 características, que se traduce en el cálculo de las 2000 palabras ms significativas en el corpus de entrenamiento. Para la división de los documentos en trozos (*tokens*) se han aplicado *tokenizers* especializados en texto de Twitter.
Una vez construída la matriz con todas las características, se ha elegido un clasificador de tipo RandomForest para entrenar un modelo matemático. Este clasificador es de tipo *ensemble*, aplicando varias iteraciones de predicción sobre conjuntos aleatórios de los datos, lo que garantiza una mejor generalización del modelo.

1 Introducción

La exploración de datos de lenguage natural permite descubrir caracterśticas de los autores basadas en sus patrones de escrita. El objectivo de este ejercício es explorar la información de un dataset constituído por tweets de varios usuarios de distintos países de habla hispánica, con el intuito de crear un modelo clasificador de perfiles de sexo y país.

Determinados patrones de escrita pueden mostrar indícios de perfiles. Por ejemplo, los hombres tienen tendencia a usar más determinantes y adjectivos que las mujeres; y las mujeres suelen usar más pronombres y la negaci'on. En general, las mujeres tienen tendencia a demostrar más carga emocional en sus frases que los hombres. Métricas como el tamao de las frases pueden eventualmente ser significativas también. Además, hay que tener en cuenta los usuarios corporativos, cuya escrita será diferente de la de otros perfiles, probablemente más formal.

De la misma forma, los perfiles de países obedecen a patrones de escrita, con elementos significativos como el uso de modismos y otras variaciones lenguísticas.

El corpus de texto se usa para aprender el vocabulário significativo y luego aplicar técnicas de *machine learning*, que se encargarán de encontrar patrones y correlaciones para generar un modelo matemtico. Los datos sirven como matéria prima, a la que se aplican herramientas para crear nuevo conocimiento.

2 Dataset

El dataset de este ejercício está constituído por cientos de tweets de usuarios de 7 países de habla hispánica. Cada país tiene tweets de 650 usuarios y cada usuario tiene entre 600 y 1000 tweets.

El conjunto de entrenamiento cuenta con datos de 2730 usuarios previamente clasificados en sexo y país. En total hay 2.616.338 tweets, que constituye el corpus para calcular

el vocabulario. El conjunto de test tiene 1820 usuarios y respectiva clasificación, con un total de tweets de 1.741.060. El volumen de datos es bastante grande y de cómputo exigente: en una máquina con 4GB de RAM y un procesador i3, el cálculo del vocabulario para solamente 2000 palabras tarda más de 10 minutos.

Se ha realizado un breve análisis estadístico sobre la matriz de *bag of words* con R, cuyos resultados se presentan a continuación. El gráfico siguiente representa las palabras y símbolos con mayor coeficiente TF-IDF. Se verifica que las palabras más significativas son también las más usuales - este dato será relevante en el desarrollo del estudio propuesto.

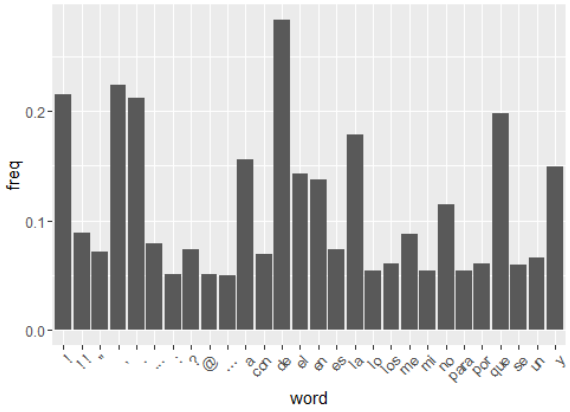

Figure 1: Palabras con mayor coeficiente TD-IDF.

En un breve análisis de sentimientos (positivo y negativo), se puede ver que la media es ligeramente superior en las mujeres.

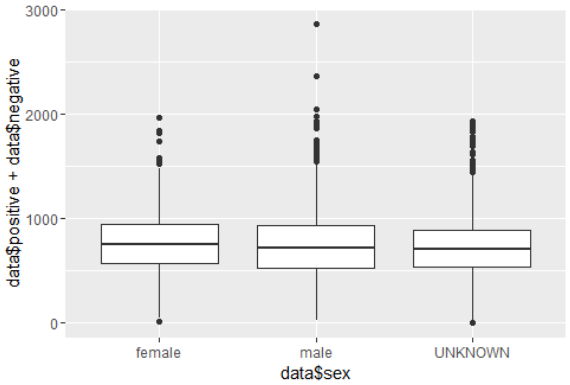

Figure 2: Sentimiento.

El tamaño medio de frases muestra que los hombres escriben frases más largas; interesante que el tamaño medio de las frases de mujeres y desconocido (*UNKNOWN*) es parecido:

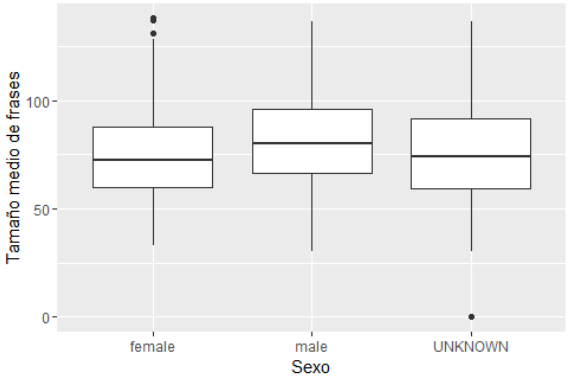

Figure 3: Tamaño medio de frases.

El *metadata* de los tweets puede contener datos significativos. Se han usado, por ejemplo, sobre todo los colores usados en el perfil de *Twitter*. Por ejmplo, en el elemento `profile_sidebar_fill_color`, se pueden identificar algunos patrones:

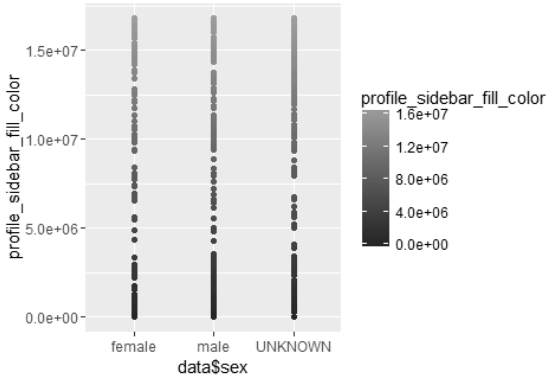

Figure 4: Colores del perfil de Twitter.

3 Propuesta del alumno

La técnica *bag of words* ha dado un *baseline* para acercar el problema. Esta técnica se basa sencillamente en contadores de palabras, en su concepto. Sin embargo, se puede customizar la forma de contar (acumulador, cuoficiente, binário), la forma de separar los componentes del texto, el número máximo de palabras del vocabulario o conjuntos de varias palabras. Las decisiones de configuración se describen a continuación.

Se ha usado un *tokenizer* especializado en texto de tweets (`TweetTokenizer`, del paquete

NLTK), que ofrece opciones de ignorar referencias a otros usuarios y reducir las repeticiones de texto. Se ha usado también la opción de tratamiento *case-sensititve*, por poder ser significativo en la distinción de género.

La transformación de los datos en características numéricas adecuadas para *machine learning* es un área conocido por *feature extraction*; en el caso de *bag of words* el proceso es de vectorización, generando un vector de características. Para esta tarea se ha usado la clase `TfidfVectorizer` de *scikit-learn*, que implementa la separación en *tokens* y el cálculo de frecuencias, as como la normalización y atribución de pesos de importancia a las palabras. Se ha configurado el modelo para generar una matriz de características de cuoficientes TF-IDF.

El aprendizaje del vocabulario es una tarea intensiva para CPU y memoria, directamente proporcional al numero de características/palabras del vector generado. Se ha configurado el modelo para un máximo de 2000 palabras, que presentaba rendimiento y resultados aceptables. Sin embargo, se han realizado pruebas y se ha verificado que los resultados de predicción son mejores con mayor numero de características, a coste de mayor tiempo de ejecución. Otra configuración importante es el rango de *ngrams*, que son conjuntos de palabras consecutivas. Se ha aplicado el rango de 1 a 2 palabras, con lo que se han mejorado los resultados.

Determinadas palabras son muy comunes y aparentemente aportan poco significado, como es el caso de artículos y preposiciones, como 'de', 'la', 'para'. Estas palabras se llaman *stop words*. La clase `TfidfVectorizer` ofrece la posibilidad de usar una lista de *stop words* y el paquete *NLTK* tiene listas pre-determinadas para varias lenguas. Se ha probado con esa configuración y no se han obtenido mejoras, probablemente porque son palabras significativas en la escrita de hombres y mujeres - por ejemplo, las mujeres usan más artículos que los hombres. Así que se ha decidido no usar *stop words*. Con el vector numérico generado y los datos de training etiquetados para sexo y país, ya se pueden aplicar algoritmos de aprendizaje supervisado. Para esta tarea, se ha usado un clasificador de tipo *Random Forest*, en particular `RandomForestClassifier` de *scikit-learn*. Este clasificador es una técnica de *ensemble*, que entrena varios árboles de decisión con sub-conjuntos del *dataset* para así obtener un mejor acierto y prevenir el sobreajuste.

4 Resultados experimentales

Resultados obtenidos con un vocabulario de 2000 caractersticas y *Random Forest* de 100 árboles:

	Sexo	País
Acierto	52.69%	93.46%

Los tiempos de ejecución en una máquina con 4GB de RAM y procesador i3 han sido:

Operación	Tiempo
Leer *tweets* de training	5 min.
Cálculo TF-IDF de training	10 min.
Características extra de training	4 min.
Leer *tweets* de test	4 min.
Cálculo TF-IDF de test	4 min.
Características extra de test	4 min.
Random Forest	1 min.
Total	32 min.

Los resultados obtenidos para la predicción de género de los usuarios han sido peores que para la predicción del país. Se verifica así que las variaciones de la lengua castellana en los distintos países son más significativas que las variaciones entre hombre y mujeres. La presencia de la etiqueta 'UNKNOWN' (instituciones, por ejemplo) puede dificultar esta tarea.

El vocabulario con *stop words* (palabras comunes) mejora un poco los resultados - como descrito anteriormente, puede ser significativo para distinguir la escrita de hombres y mujeres.

La inclusión de información extra (sentimiento y metadata) en el vector de características ha tenido efectos importantes. En la predicción de sexo, por ejemplo, la mejora fue de la orden del 10%.

El método de *Random Forest* ha demostrado mejores resultados respecto a otros clasificadores, como gausianas. El hecho de ser una técnica de *ensemble* contribuye bastante, al entrenar varios modelos con sub-conjuntos de datos. Sin embargo, esta técnica exige más tiempo de ejecución.

5 Conclusiones y trabajo futuro

Text Mining para *Author Profiling* es un área con muchas aplicaciones prácticas, que pueden ser, por ejemplo, la determinación del género o la detección de fraude. Es interesante verificar como las técnicas de *machine learning* permiten descubrir patrones en la forma de escribir, que pueden dar mucha información. De esa forma se consigue nuevo conocimiento, al que se puede aplicar inteligencia de decisión.

El procesamiento de lenguage natural puede ser una tarea compleja. Hay variaciones en la escrita de la misma persona, errores ortográficos o palabras con repeticiones para dar enfasis. Los sentimientos manifestados pueden ser ambíguos. Y, por otro lado, el uso de ironía es difícil de detectar.

La predicción de sexo no tiene tan buenos resultados como la de país. Probablemente el análisis de sentimientos podría ayudar en ese sentido, por lo que necesitaría una explotación más profunda. Igualmente, otros elementos de metadata podrían haber sido usados. El uso de mayor número de características de vocabulario podría ser una mejora futura, para lo que haría falta mayor capacidad de computo.

Hay otros clasificadores de *supervised learning* que podrían aportar mejoras en el ajuste a este tipo de datos. Los métodos de *Support Vector Machines* serían buenos candidatos, pues construyen uno o varios hiper-planos en espacios dimensionales grandes, para así dividir el plano y clasificar los datos.

References

Kaggle. *Part 1: For Beginners - Bag of Words* `https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-1-for-beginners-bag-of-words`

Scikit-learn. *1. Supervised learning* `http://scikit-learn.org/stable/supervised_learning.html#supervised-learning`

NLTK 3.0 documentation. *nltk.tokenize package* `http://www.nltk.org/api/nltk.tokenize.html`