Text Mining en Social Media. Master Big Data Analytics

Jose Carlos Sola Verdú
josover1@alumni.upv.es

Abstract

El **objetivo** de la práctica consiste en analizar una serie de tuits para **conocer** la **variación geográfica** donde se encuentra la persona que lo escribe y su **género**.

Para ello, se recogen en un *dataset* todas las palabras incluidas en los tuits, posteriormente se realizará un preprocesado al *dataset* eliminando números, espacios en blanco, letras sueltas, signos de puntuación, preposiciones, conjunciones, etc., así como pasar todos los tuits a minúsculas para facilitar su análisis.

Una vez han sido eliminados todos los elementos no reseñables del *dataset*, se generará una bolsa de palabras (*Bag of Words* en inglés) con los 1.000 términos más frecuentes para ser tratados durante el entrenamiento (training) del modelo. También se utilizará esta bolsa de palabras para generar las representaciones tanto del corpus de *training* y de *test*.

Se representará la bolsa de palabras como la frecuencia relativa en la que se repite cada palabra de la bolsa en el *dataset*, donde una representación más elaborada consistiría en tener en cuenta los pesos *tf / idf* de las palabras, así como tener en cuenta la presencia de cada palabra en el *dataset*.

A partir de la representación del *training*, se probarán diferentes herramientas de aprendizaje para evaluar la representación del *test*, y calcular el *accuracy* para el género y la variedad geográfica. También se tendrá en cuenta el cálculo de medias como la precisión, el *kappa* (cuando más cercano es a 0, significa que es un resultado más próximo al azar) y el *recall,* para complementar el análisis.

1 Introducción

Para llevar a cabo la práctica, se ha proporcionado un *dataset* que contiene un gran número de tuits de la comunidad hispano-hablante. Lo que se pretende es, al analizar dichos tuits, conocer si el tuit ha sido escrito por un hombre o por una mujer, y la variedad geográfica de dicho escritor/a.

En primer lugar, se tratará el *dataset*, es decir, se eliminarán elementos que no son relevantes para el análisis, como números, signos de puntuación, espacios en blanco, etc.

En segundo lugar, se generará una bolsa de palabras que se utilizará para realizar el *training* y el *test*. Tras aplicar varias herramientas de aprendizaje, se valorará cual es la que mejor *accuracy* obtiene, de esta forma, cuánto más accuracy mejor será la precisión para determinar el género y la variedad geográfica del usuario.

2 Dataset

El *dataset* con el que se realizará el análisis contiene los tuits de **300 autores** de la comunidad hispano-hablante. Además, contiene **100 tuits por** cada **autor**.

Tras una **limpieza** básica como eliminación de signos de puntuación, números, espacios en blanco, convertir todas las letras a minúsculas… se **genera** la **bolsa de palabras** (*Bag of Words*), la cual se tendrá en cuenta para realizar el *training*.

Una vez generado el vocabulario a partir de la bolsa, se emplea ***Support Vector Lineal Machine*** (SVM) como modelo de aprendizaje, obteniendo un ***accuracy*** para el **género** de **66.43%** y un **77.21%** para la **variedad** geográfica, siendo estos los resultados a batir.

3 Propuesta del alumno

Para conseguir batir los resultados anteriores, en la limpieza de datos se han eliminado *stopwords*, letras que se encuentran sueltas, dos o más espacios en blanco, así como las preposiciones y conjunciones, ya que son palabras que se repiten mucho pero carecen de peso para efectuar el análisis.

Una vez se ha efectuado la limpieza de datos, se optó por evaluar el modelo con SVM como en el caso original, para comprobar que la limpieza realizada era efectiva.

Además, se han probado otros modelos de entrenamiento, como **redes neuronales**, árboles de decisión (***Random Forest***) y ***Naive Bayes***.

4 Resultados Experimentales

A continuación, se detallan los resultados para cada uno de los modelos testeados:

- **SVM**: Tras probar este modelo con la limpieza de datos que generamos, conseguimos mejorar el *accuracy* de la **variedad**, siendo del **77.29%**, pero, sin embargo, en el **género** descendió, **66.36%.**

- **Red Neuronal:** Finalmente fue descartado porque requería demasiado tiempo para su ejecución, no siendo viable para este dataset.

- **Random Forest**: Tras probar con 10 árboles de decisión, n = 10, la ejecución se llevó a cabo de forma satisfactoria, pero no se consiguió mejorar el resultado obtenido en SVM, así que se decidió realizar la ejecución con n = 100 para 1.000 palabras, obteniendo unos resultados altamente satisfactorios. Para el **género** se consiguió un *accuracy* del **72.21%** y para la **variedad** un **88.71%.**

- **Naive Bayes**: Realmente este modelo fue un poco frustrante, ya que después de estar casi unas 4h ejecutándose, finalmente saltó un error y no se pudo llevar a cabo la correcta ejecución de este modelo.

5 Conclusiones y trabajo futuro

Tras realizar todo el tratamiento del dataset, el equipo quedó muy contento con el resultado obtenido con el modelo ***Random Forest***, ya que permitió obtener unos ***accuracys*** bastante ***altos***.

Como **trabajo futuro**, se barajan diferentes propuestas para mejorar el *accuracy* en la **variedad,** como **mantener** sólo la **raíz** de la **palabra** y probar con las **palabras** más **específicas** de cada **país**. En cuanto a mejorar el **género**, se tendría en cuenta la **última letra** de la **palabra** y la longitud de éstas.

En el código fuente del proyecto se encuentran todos los bloques comentados para facilitar la ejecución del script.

References

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