Manuel Sánchez Gomis
nelosan@gmail.com

Abstract

A lo largo del siguiente documento el lector podrá conocer el estudio realizado para clasificar un conjunto de *tweets* procedentes de la red social Twitter y cómo a través de ellos se ha generado un modelo que permita clasificar futuros *tweets*. Este modelo podrá identificar el género y el país de origen del autor.

En el documento se va a describir cómo se ha obtenido el *dataset* que conforma el estudio, y cómo se ha clasificado para su posterior estudio. El lector podrá conocer las técnicas que permite generar el *dataset* completo y su manera de almacenarlo.

Posteriormente, se mostrará cómo se han tratado estos datos para generar un modelo que permita clasificar los futuros *tweets*. Se partirá de la generación de una bolsa de palabras de un total de mil palabras, la cual se irá ajustando al aplicar algunas modificaciones, tales como eliminar acentos o no tener en cuenta algunos *stopwrords*, o palabras vacías.

Una vez generada la bolsa de palabras se pasará a utilizar un modelo de clasificación basado en *Suport Vector Machine* para posteriormente pasar a otro modelo de tipo *Random Forest*. Se presentarán las técnicas que se han utilizado para ajustar este nuevo modelo, tales como incrementar el número de árboles o incrementar el número de palabras de la bolsa.

Finalmente, se podrán observar los resultados generados por estas técnicas así como las conclusiones alcanzadas por el autor y las futuras pruebas a realizar para mejorar los resultados obtenidos.

Introducción

Todos conocemos la red social Twitter, una plataforma donde los usuarios publican diariamente todos aquellos mensajes que quieran compartir con el mundo. Esto proporciona un mecanismo de estudio social muy interesante para todo aquel que esté *observando* los mensajes de esta red social.

A lo largo de este artículo se va a mostrar un estudio realizado para obtener información de los *tweets* generados por un conjunto de usuarios de habla hispana. Mediante estos *tweets*, se va a intentar generar un modelo que permita identificar el género del autor y el país del que procede este autor.

Dataset

Como se ha comentado en el apartado anterior, el objetivo de este artículo es realizar un estudio sobre *tweets* obtenidos de la red social Twitter de personas de habla hispana.

Para ello, se ha realizado la descarga de un conjunto de tweets que se han almacenado en función del autor, generando varios documentos. Estos documentos tienen como nombre el identificador del usuario y como contenido todos los *tweets* de ese usuario.

A la hora de tener en cuenta los *tweets* a descargar, se han seguido los siguientes criterios:

- Se recuperan los *tweets* pertenecientes a las regiones geográficas de habla hispana.
- Se seleccionan los tweets de autores con más de 100 *tweets* que no sean *retweets*.
- Se revisan manualmente los perfiles para asegurar el sexo y localización.
- Se seleccionan 100 *tweets* por autor para la construcción final del *dataset*.

Con toda esta información, se ha generado un dataset para el *training* del modelo compuesto por *tweets* de 2801 autores.

Además de estos datos, se han obtenido otros *tweets* pertenecientes a 1401 autores para realizar el test del modelo.

Propuesta

Una vez descrito el problema y los datos con los que se cuenta para realizar el estudio, vamos a mostrar todas las operaciones realizadas así como los resultados obtenidos.

En primer lugar se generó un corpus que contenía todas las palabras pertenecientes a los *tweets* de *training* con su frecuencia. Una vez generado el corpus, se obtienen las mil palabras más frecuentes, aplicando distintos tipos de preprocesado, como pasar a minúscula, eliminar número, eliminar palabras vacías, etc.
Por último, con el vocabulario y los datos pertenecientes a los *tweets*, se obtiene una representación en bolsa de palabras por frecuencias relativas.

Con la bolsa de palabras, se pasó a aprender y evaluar el modelo. El primer modelo con el que se probó fue un *SVN*[1] con método lineal. Este modelo nos daba un *accuracy* del **0,6643** para género y del **0,7721** para la variedad.

En base a esto, se decidieron realizar las siguientes operaciones para ver si se mejoraba el resultado del modelo:

- Eliminación de los acentos de las palabras. Esto puede ser influyente ya que no todas las personas escriben correctamente ni acentúan sus palabras.
- Incrementar el corpus de *stopwords* o palabras vacías. Para ello, se realizó una búsqueda de un conjunto de palabras que pudieran ser añadidas a las palabras eliminadas anteriormente.
- Cambiar el modelo a *Random Forest*. El *Random Forest* es un modelo que para problemas de clasificación da muy buenos resultados, así que se realizó la prueba con este modelo.
- Incrementar el número de palabras de la bolsa de palabras. En vez de crear una bolsa de palabras con mil palabras, se decidió hacer la prueba con dos mil palabras.

[1] Support Vector Machine

Resultados experimentales

En esta sección se va a mostrar los resultados obtenidos de aplicar las técnicas indicadas en el apartado anterior. Vamos a partir de los resultados obtenidos para el *SVN* con método linear.

Gender	Variety
0,6643	0,7721

Resultados SVN método lineal

En primer lugar se eliminaron los acentos manteniendo el mismo modelo. Solo con este cambio los resultados mejoraron un poco, aunque no de manera muy notable.

Gender	Variety
0,665	0,780

Resultados SVN eliminación de acentos

La segunda prueba que se hizo fue cambiar el modelo a *RF*[2]. Se hicieron varias pruebas cambiando el número de árboles para ver cuál era la que daba un mejor resultado.

Se empezaron haciendo las pruebas con diez árboles y manteniendo la eliminación de los acentos, lo cual mejoró un poco los resultados.

Gender	Variety
0,675	0,790

Resultados RF con 10 árboles

Posteriormente, se incrementaron el número de árboles a cincuenta, lo cual ya mejoró un poco más el modelo, llegando a casi tres décimas en el caso de la variedad.

[2] Random Forest

Gender	Variety
0,690	0,819

Resultados RF con 50 árboles

Cabe destacar que el tiempo de entrenamiento del modelo también descendía con el *RF*.

Se realizaron tres pruebas más, para cien, ciento cincuenta y doscientos. Con estas pruebas, el tiempo de entrenamiento ya iba notándose, pero los resultados también iban mejorando.

Gender	Variety
0,712	0,823

Resultados RF con 100 árboles

Gender	Variety
0,720	0,830

Resultados RF con 150 árboles

Gender	Variety
0,722	0,836

Resultados RF con 200 árboles

Posterior a esto, se realizó un incremento del número de *stopwords* con un diccionario de palabras que contenía más de 600 palabras. Este diccionario permitió mejorar los resultados para la variedad, pero empeoraba los resultados con el género.

Gender	Variety
0,710	0,851

Resultados RF con 100 árboles y más stopwords

El último cálculo realizado fue incrementar el número de palabras de la bolsa, lo cual

incrementaba bastante el tiempo de cálculo.

Gender	Variety
0,735	0,863

Resultados RF con 100 árboles, más stopwords y 2000 palabras

Conclusiones y trabajos futuros

Llegados a este punto, podemos concluir con todos los experimentos realizados anteriormente, que para un problema de clasificación el modelo basado en *Random Forest* da muy buenos resultados sin emplear mucho tiempo. Comparándolo con el primer cálculo basado en *svn*, las mejoras son notables en un tiempo menor.

Incrementar el número de palabras de la bolsa ayuda a mejorar los resultados pero también el tiempo de procesamiento.

Otro factor que mejora los resultados en el cálculo de la variedad es incrementar el número de *stopwords*, pero en el cálculo del género empeora los resultados.

Así pues, como trabajos futuros, sería interesante llegar a configurar de manera eficiente el número de palabras para la bolsa.

Por otro lado, habría que comprobar ciertos estudios realizados por otros grupos, para ver si los resultados mejoran:

- Realizar un estudio de los emoticonos, ya que parece que son más utilizados por las mujeres que los hombres y esto podría identificar de manera más fácil el género.

- Realizar estudios de palabras más utilizadas por hombres y mujeres. Podríamos identificar estas palabras en publicaciones dirigidas a los distintos géneros.

- Realizar un estudio de palabras más utilizadas por países, lo cual nos vendría bien para identificar el país de procedencia del emisor del mensaje.

El autor considera que solo con estos estudios, los resultados ya mejorarían notablemente para el cálculo del género y la variedad en los *tweets*.