Esquema de paper. Asignatura Text Mining en Social Media. Master Big Data

Javier Montagud Gómez
javiganval@gmail.com

Abstract

El siguiente artículo trata de predecir tanto el género como la variedad del autor de un tuit.
Para ello partimos de un dataset con un conjunto de tuits de diferentes autores. Partiendo de esto se consigue entrenar un modelo que se utilizará para predecir ambas variables, gnero y variedad.
Dado que se trata de dos problemas diferentes, se han trabajado de forma separada. Por un lado, para conseguir predecir el Género he intentado plantearme preguntas relacionadas con el cómo?. De ahí me han surgido diferentes técnicas que se explican con detalle en los siguientes apartados.
Por otro lado para predecir la Variedad he intentado plantearme preguntas relacionadas con el qué?. De ahí me han surgido diferentes técnicas que se explican con detalle en los siguientes apartados.
He realizado diferentes pruebas con diferentes modelos para tratar de obtener los mejores resultados para los dos problemas a tratar. En los siguientes apartados se comentan los resultados.
Por último se pueden consultar tanto las concluisiones obtenidas como los posibles problemas pendientes de tratar.

1 Introducción

El presente documento trata de explicar cómo predecir con mayor exactitud la siguiente información:

- El sexo del autor de un tuit, si es hombre o mujer.
- La variedad del autor de un tuit.

Para ello, partimos de un conjunto de 100 tuits por autor y un total de 2800 autores. Como punto de partida tenemos unos resultados de acierto que muestro a continuación y que trataré de mejorar:

- Género: 66.43%
- Variedad:77.021%

Para conseguir ese resultado inicial se parte de la siguiente información:

- Un vocabulario con las 1000 palabras más frecuentes de todos los tuits del Corpus de entrenamiento. Previamente se procesa toda la información con el fin de eliminar acentos, números, palabras vacías,
- Una bolsa de palabras con las frecuencias relativas de las palabras del vocabulario en los tuits de cada autor.
- Un modelo entrenado con los datos de train que consisten en un conjunto de 100 tuits por autor con un total de 2800 autores. Tras evaluar el modelo con ciertos datos de test se consiguen los resultados mencionados anteriormente.

A continuación paso a describir el dataset y las diferentes técnicas utilizadas para conseguir mejorar el resultado inicial. Los mejores resultados obtenidos después de aplicar dichas técnicas son considerablemente más alto:

- Género: 74%
- Variedad: 86%

2 Dataset

El proceso de construcción del dataset ha sido el siguiente:

- Se recuperan tuits enmarcados en una región geográfica. Estos son longitud, latitud, radio

- Se preseleccionan los usuarios únicos que han emitido tuits (filtrados por idioma del perfil)
- Se recuperan los timelines de los usuarios únicos.
- Se seleccionan los autores con más de 100 tuits (que no sean retuits) en el idioma correspondiente y con la localización geográfica esperada en su perfil.
- Se revisan manualmente los perfiles para asegurar el sexo.
- Se seleccionan 100 tuits por autor para la construcción del dataset final.

Las características del dataset son las siguientes:

- Se obtiene de Twitter
- Se busca una colección de miles de autores.
- Se recuperan cientos de tuits de autor.
- Se recupera información de gran variedad de temas.
- Aproximadamente ocupa unos 54Mb descomprimido.

El formato de los ficheros descomprimidos es la siguiente:

- Un par de ficheros de verdad: training.txt y test.txt. El formato es: id:::sexo:::variedad
- Un fichero .json por autor:

Lo que nos interesa explorar es lo siguiente:

- Número de autores por clase (sexo y variedad del lenguaje).
- Número de tuits por autor.
- Número de tuits por clase.
- Número de palabras por documento / autor / clase.
- Distribución de palabras/documentos/autores por documento/autor/clase
- Longitud media de tuits, palabras, documentos...por clase.
- Distribución temporal de los tuits, tuit más antiguo, más nuevo, media,
- Palabras extrañas, frecuentes, comunes

3 Propuesta del alumno

Como punto de partida tenemos 2 resultados, uno para el **género** y otro para la **variedad**. Es por ello que he tratado tanto el **género** como la **variedad** como 2 problemas diferentes.

Ambos problemas los abordo con planteamientos diferentes para conseguir optimizar los resultados en cada uno de ellos.

Para el problema del **género** he intentado plantearme preguntas relacionadas con el cómo?. Por ejemplo, cómo escriben las mujeres y cómo escriben los hombres? cómo utilizan los verbos las mujeres y cómo los utilizan los hombres? Haciéndome este planteamiento, he encontrado diferencias semánticas entre hombres y mujeres. He tratado de trasladarlas al dataset.

Para este primer problema he creado el fichero Gender.R.
Sobre la propuesta inicial he realizado las siguientes modificaciones:
En la fase de generación de la bolsa de palabras, he añadido las siguientes columnas, que contienen las frecuencias de las siguientes palabras en los tweets de cada autor:

- Adjetivos: palabras localizadas en el fichero adjetivos.txt. En principio las mujeres utilizan más los adjetivos que los hombres.
- Pronombres: palabras localizadas en el fichero pronombres.txt. En principio las mujeres utilizan más los pronombres que los hombres.
- Palabras cariñosas: palabras localizadas en el fichero sentimientos.txt. En principio las mujeres utilizan más los adjetivos que los hombres
- Los horóscopos. Palabras localizadas en el fichero horoscopo.txt
- Las palabras más frecuentes de las mujeres.Palabras localizadas en el fichero mujeres.txt
- Las palabras más frecuentes de los hombres.Palabras localizadas en el fichero hombres.txt

- Las palabras de la revista Superpop.Palabras localizadas en el fichero superpop.txt

Estas frecuencias las he dividido por el número de palabras de los 100 tweets de cada autor.
No es lo mismo que la palabra casa aparezca 5 veces en un total de 100 palabras escritas por un autor que en un total de 200 palabras de otro autor. No tendría el mismo peso. Es por ello que lo dividimos por el total de palabras de cada autor.
Adicionalmente he añadido una columna más con el número de palabras de todos los tweets de cada autor.

En la fase de entrenamiento pruebo con los siguientes modelos:

- SVMLinear
- RandomForest
- Penalized Multinomial Regression

Para el problema de la **variedad** he intentado plantearme preguntas relacionadas con el qué? Por ejemplo, qué palabras son las más frecuentes en España? De esta forma consigo ver diferencias entre las diferentes variedades e intento trasladarlas al dataset.
Para este problema he creado el fichero Variety.R.
Sobre la propuesta inicial he realizado las siguientes modificaciones:
En la fase inicial, donde se preprocesan los datos, he quitado los acentos para conseguir que no se tuviesen en cuenta a la hora de entrenar el modelo.
En la fase posterior, donde se genera la bolsa de palabras, he añadido 8 nuevas columnas. 7 de las columnas hacen referencia a cada una de las 7 variedades posibles.
Los valores de cada una de estas 7 columnas son:

- Frecuencia de las 300 palabras más repetidas en cada **variedad** dividido por el numero de palabras de los 100 tweets de cada autor.La última columna hace referencia al número de palabras de todos los tweets de cada autor.

En la fase de entrenamiento pruebo con los siguientes modelos:

- SVMLinear
- RandomForest
- Penalized Multinomial Regression

4 Resultados experimentales

Para el problema del **género** se realizan diferentes pruebas con diferentes métodos que tras las pruebas se desestiman.
Los métodos probados han sido entre otros los siguientes:

- SVMLinear
- RandomForest
- Penalized Multinomial Regression

El método más favorable es el Penalized Multinomial Regression.
Sobre dicho método realizamos diferentes pruebas con diferentes combinaciones.
Los resultados obtenidos de dichas pruebas son los siguientes:

MÉTODO	ACCURACY	KAPPA
Multinomial (1)	0.7321	0.4643
Multinomial (2)	0.7336	0.4671
Multinomial (3)	0.7329	0.4657
Multinomial (4)	0.7336	0.4671
Multinomial (5)	0.7345	0.4612
Multinomial (6)	0.7379	0.4757
Multinomial (7)	0.7364	0.4729

(1) Sin tener en cuenta los archivos mujeres.txt y hombres.txt
(2) Con acentos
(3) Aadiendo Superpop2.txt
(4) Aadiendo el logaritmo del tamao
(5) Dividiendo el tamao por la variedad
(6) Aadiendo superpop.txt
(7) Poniendo deportes.txt

Para el problema de la **variedad** se realizan diferentes pruebas con los siguiente modelos:

- SVMLinear
- RandomForest
- Penalized Multinomial Regression

El mejor resultado obtenido es con el método RandomForest con un 82,92%.
Los resultados obtenidos son los siguientes:
(1) Palabras vocabulario: 1000, palabras variedad: 300
(2) Palabras vocabulario: 500, palabras variedad: 300

MÉTODO	ACCURACY	KAPPA
SVMLinear (1)	0.7829	0.7467
SVMLinear (2)	0.7364	0.6925
SVMLinear (3)	0.7029	0.6533
RandomForest (4)	0.8636	0.8408
RandomForest (5)	0.8292	0.8292
Multinomial (6)	0.7893	0.7542

(3) Palabras vocabulario: 0, palabras variedad: 300
(4) Palabras vocabulario: 1000, palabras variedad: 300
(5) Palabras vocabulario: 1000, palabras variedad: 500
(6) Palabras vocabulario: 1000, palabras variedad: 500

5 Conclusiones y trabajo futuro

Como conclusiones destaco las siguientes:

- Para solucionar el problema del **género** es más influyente buscar posibles patrones que den respuesta al cómo?, por ejemplo cómo se expresa la mujer?, mientras que para el problema de la **variedad** es más influyente dar respuesta a preguntas relacionadas con el qué?. Por ejemplo, qué palabras son las más frecuentes en Colombia?
- El hecho de que hayan menos opciones en una clase a predecir no implica que por ello se obtengan mejores resultados y de forma más fácil. Para este caso se obtienen mejores resultados y más fácilmente cuando se trata de solucionar el problema de la **variedad**, teniendo ésta muchas más posibles opciones que el problema del **género**, que sólo tiene dos posibles. Hombre o mujer.
- A nivel de resultados en la predicción del **género** cabe destacar que el método de Multinomial ha sido el mejor de todos los utilizados.
- A nivel de resultados en la predicción de la **variedad** cabe destacar que el método de RandomForest ha sido el mejor de todos los utilizados con diferencia.

Como trabajo futuro, lo desglosaría en los 2 problemas a resolver:
Para el problema del **género**:

- Detección de los emoticonos. Creemos que las mujeres los utilizan más y por tanto podría obtenerse mejoras.
- Etiquetar cada palabra con algún etiquetador. De esta forma sabríamos que es un adjetivo, un pronombre y sobre todo que es un verbo y si es presente o pasado. Los verbos en presente son más utilizados por las mujeres y esta sería una forma de detectarlos.

Para el caso de la **variedad**:

- Etiquetar cada palabra con algún etiquetador. De esta forma se podría tratar la información del dataset de forma separada para que el modelo consiguiera encontrar posibles patrones.

References

El presente documento trata de explicar cómo predecir con mayor exactitud la siguiente información:

Libro de Hadley Wickham Garrett Grolemund. R for Data Science.

Las siguientes URLS:
- https://stackoverflow.com/questions/tagged/r
- http://www.statmethods.net/r-tutorial/index.html
- https://stats.idre.ucla.edu/r/faq/

Las fuentes de datos utilizadas para obtener diccionarios de diferentes conjuntos de palabras son:
- http://www.superpop.es/
- http://www.hola.com/horoscopo/