Asignatura Text Mining en Social Media. Master Big Data

Antonio Avia Antúnez
toni.avia@gmail.com

Abstract

Sirva el presente documento como parte del entregable a realizar de la asignatura Text Mining en Social Media en el marco del Master en Big Data Analytics por la Universidad Politécnica de Valencia.

El documento pretende reflejar el estudio realizado sobre un conjunto de tweets, a través del cuál, mediante diferentes técnicas de exploración, transformación y cálculo, ser capaces de identificar el género y la variedad de idioma de la persona que lo ha escrito. En el documento se detallarán esas técnicas y criterios, presentando diferentes resultados a nivel comparativo.

La presentación de los resultados irá en función de la tasa de acierto obtenida, presentando 2 valores:

- Tasa de acierto en identificación de género por tweet
- Tasa de acierto en identificación de la variedad del idioma por tweet

1 Introducción

Tal y como se ha comentado a modo preliminar, el objetivo del estudio es identificar el género y la variedad del idioma de un conjunto de tweets, por lo tanto tenemos como punto de partida los siguientes datos:

- Conjunto de tweets de entrenamiento. Almacenados en la carpeta training del sistema de ficheros. Se dispone de un total de 2800 ficheros en formato xml. Dentro de cada fichero xml a su vez se dispone del orden de 100, 120 tweets

- Conjunto de identificadores de cada tweet de entrenamiento con la identificación del género y variedad del idioma en el fichero truth.txt dentro de la carpeta training

- Conjunto de tweets de test. Almacenados en la carpeta test del sistema de ficheros. Se dispone de un total de 1400 de ficheros en formato xml, donde como en el caso anterior, cada fichero puede contener del orden de 100, 120 tweets

- Conjunto de identificadores de cada tweet de test con la identificación del género y variedad del idioma en el fichero truth.txt dentro de la carpeta test

Se introduce el concepto de bolsa de palabras. Una bolsa de palabras es la frecuencia con la que aparece determinada palabra a partir de un vocabulario y un conjunto de datos, tweets en este caso.

El número de frecuencias de palabras que compondrán la bolsa o el número de palabras que formarán el vocabulario es configurable en el pre-proceso y sus valores pueden oscilar entre 10, 50, 100, 500, 1000, 5000 y 10000, siendo su valor más alto el más preciso para el estudio, pero también el más costoso computacionalmente.

Dado que el estudio se diversifica en identificar el género y la variedad del idioma, la bolsa de palabras será diferente.

Una vez creada la bolsa de palabras la idea es aplicar diferentes técnicas de machine learning a los datos de entrenamiento para obtener como resultado la identificación del género y la variedad del idoma por tweet que reside en el fichero training/truth.txt. Con el algoritmo de machine learning aprendido, aplicando esa misma técnica al conjunto de datos de test podremos saber cuál es la tasa de acierto en la obtención del

género y la variedad de idioma por tweet ya que compararemos el resultado del algoritmo con el fichero test/truth.txt con los identificadores reales del género y variedad del idioma.

2 Dataset

Tenemos aproximadamente para realizar el estudio:

- 280.000 336.000 de tweets de entrenamiento
- 140.000 168.000 de tweets de test

Haciendo un análisis más profundo se podrían presentar valores estadísticos tales como la media, moda, mediana, desviación típica, máximo, mínimo, longitud, coeficiente de kurtosis, cuartiles, etc.

Como paso previo a transformar un vocabulario en una bolsa de palabras (frecuencia de aparición de palabras en valor numérico), se eliminan las mayúsculas, los retornos de carro y se aplica un tokenizador estándar del API de Tweeter, esto permitirá un mejor procesamiento de los datos.

A continuación se hará un cálculo de características que servirá para aportar información adicional en la técnica de machine learning no supervisada a aplicar, se detalla en el siguiente punto.

3 Propuesta del alumno

El cálculo de características sirve para ayudar a determinar un patrón de identificación del género y la variedad del idioma. Diferenciando entre género y variedad del idioma, los criterios son:

GÉNERO

- Longitud del tweet, pudiendo desglosarse por número de caracteres, número de palabras y por número de frases

- Número de menciones por el autor del tweet a otros usuarios. Es posible que dependiendo del género se hagan más o menos menciones

- Estudio del género de los influencers que siguen las menciones del autor. Se puede construir un dataset por género con los principales influencers

- Las relaciones con otros usuarios pueden ser relevantes para la determinación del género

- Número de hashtags, emojis y retweets por autor

- Contador de tipos de palabras: verbos, nombres, determinantes, adjetivos, signos de puntuación y admiración, valores numéricos

- Elaborar una bolsa de palabras con los temas más comunes por género, tales como deporte, moda, política, sexo, economía, empleo, etc

- Número de urls

- Analizar los dominios de esas urls que pueden referenciar a la bolsa de palabras con los temas más comunes del punto anterior

VARIEDAD

- Al igual que para el género, se considera que la longitud del tweet puede ser importante

- Localización del perfil de la persona con la que el autor se relaciona y de amigos del mencionado

- Se puede construir un dataset con los principales influencers del país

- Contador de tipos de palabras: verbos, nombres, determinantes, adjetivos, signos de puntuación y admiración, valores numéricos

- Comparar los hashtags con los trending topics actuales del país

- Tener una bolsa de palabras de los principales periódicos digitales de cada país ya que en los tweets se hacen muchas referencias a periódicos digitales

- Analizar los dominios de las urls del tweet para identificar el país origen

- Tener una bolsa de palabras o un dataset con las ciudades y/o municipios de un país. Comparable a información que nos pudiera aportar el INE en españa y el organismo correspondiente de los países de sudamérica.

4 Resultados experimentales

Los resultados para una bolsa de palabras de 10 y 50 en función del tiempo y tasa de acierto son:

Mining en Social Media/Entrega/images/tabla.pdf

	n=10				n=50			
Modelo	Género	Tiempo (s)	Variedad	Tiempo (s)	Género	Tiempo (s)	Variedad	Tiempo (s)
SVM Lineal	0.6814	4.46	0.5921	3.01	(*)		0.6764	14.36
Naive Bayes	0.6778	0.08	0.6864	0.11			0.7321	0.04
Random Forest	0.6735	0.71	0.6457	0.72			0.83	0.46
KNN	0.5871	6.88	0.5771	6.47			0.23	3.02
Regresión Logística	0.775	0.9	0.7692	0.61			0.9364	2.37
Red Neuronal	0.7742	69.2	0.7757	75.25			0.9378	23.47

- El método que mejor tasa de acierto en relación al tiempo de computación presenta es la Regresión Logística

- Su tasa de acierto es similar a la red neuronal, presentando esta última el tiempo de ejecución más alto de todos los modelos aplicados, por lo tanto no es el más recomendable por su alto coste de computación

- La explicación de esta situación puede ser debido a que existe una estrecha relación entre las frecuencias obtenidas de la bolsa de palabras y la propia concepción del algoritmo que se basa en probabilidades de variables categóricas

- En general el modelo que peores resultados obtiene es KNN, mostrando además en sus tiempos de ejecución valores elevados

- De los resultados también se extrae que conforme aumenta la bolsa de palabras la tasa de acierto de cada modelo va mejorando, excepto para el modelo KNN

(*) No he podido ejecutar en mi local la parte del Género para n=50

5 Conclusiones y trabajo futuro

Según los resultados, es muy importante enfocar un proyecto de este tipo tanto desde un punto de vista de qué modelo funciona y se adapta mejor a la hora de clasificar o predecir un conjunto de datos, como desde qué coste computacional se está dispuesto a asumir.

Como posibles mejoras, tal y como se ha mencionado en el estudio de las características, se pueden elaborar diferentes datasets por temáticas que permitirían aproximar mejor la identificación del género y la variedad del idioma:

- Influencers
- Páginas web de periódicos digitales
- Urls por temas: deportes, economía
- Información INE (españa) y resto de países
- Otros

References

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