Text Mining en Social Media. Master Big Data

Alejandro Iborra Casanova
iborra.alex@gmail.com

Abstract

El presente estudio Social Media intentara encontrar las claves para la determinación de un modelo capaz de predecir el género y la variedad lingüística del autor de un comentario de una red social. En concreto el estudio se centrara en un corpus de comentarios en Twiter de 2800 autores.
El corpus proporcionado se encuentra totalmente balanceado por lo que existirán muestras equivalentes tanto en género como en variedad lingüística, lo que nos permitirá generar un corpus de training óptimo para el entrenamiento de los diferentes modelos.
Una vez realizado el análisis podremos determinar si características como la longitud de un tweet, el número de emoticonos o el número de preposiciones se ve influenciado por el género de autor. De la misma forma, seremos capaces de afirmar si existen bolsas de palabras con una mayor frecuencia en función de la variedad lingüística usada por el autor, en el marco estudiado.
Una vez construido un dataset de training, utilizaremos el accuracy de modelos como SVM (Super Vector Machine), Arboles de Decisión, Random Forest o Redes Neuronales para finalmente seleccionar aquel que nos proporcione mejores resultados. Además del accuracy deberemos tener en cuenta el tiempo de ejecución del modelo antes de decantarnos por uno de los mencionados, ya que el objeto del estudio es buscar el equilibrio entre accuracy y tiempo de respuesta.

1. Introducción

El presente documento detallará el proceso que se ha llevado a cabo para generar un análisis de Text Mining en Social Media mediante modelos de clasificación capaces de determinar tanto el sexo como la variedad lingüística de los autores de un corpus de twiter. A partir de la información proporcionada se generara un dataset con características que iremos obteniendo del corpus inicial.

La información para el presente estudio se nos presenta en dos carpetas que contienen los tweets de cada uno de los autores que componen el corpus de estudio.

La carpeta de training contiene 2800 ficheros (uno por autor) con 100 tweets cada uno. El nombre de cada uno de estos ficheros se corresponde con un id que permite obtener el género y la variación lingüística de un autor en el fichero truth.

Por otro lado en la carpeta de test únicamente se encuentran los 2800 ficheros referentes a los distintos autores. En este caso, dado que serán los datos que se utilizaran para la evaluación del modelo no se dispone del fichero con las etiquetas.

Para la detección del género se han creado bolsas de palabras a partir del corpus del training. Adémas se han utilizado otras características como el número de emoticonos, la longitud de los tweets o el género de los adjetivos utilizados.

No obstante en los próximos apartados se detallará como se ha ido generando el dataset a partir de la información de training y cuales han sido los criterios que han influido en la decisión de uno de los diferentes modelos probados.

2. Dataset

En este apartado se profundizara en la creación del dataset y la obtención de las diferentes características que se usaran para el entrenamiento del modelo. En primer lugar se genera una bolsa de

palabras para cada género a partir del corpus de training, cuya diferencia se utilizara más adelante como característica en el dataset final.
A continuación se genera un primer dataset que contiene todos los tweets de la carpeta de training etiquetados en género. Para esto se recorren todos los fichero de la carpeta en cuestión obteniendo el tweet y el id de autor (nombre del fichero). El id del autor nos permitirá etiquetar en genero con la información del truth.

Una vez generado este primer dataset se irán incluyendo nuevas características obtenidas mediante funciones que se han creado para este fin y que se definen en las próximas líneas.
A partir de todos los tweets de un autor se obtiene el número de medio palabras por tweet. Esta característica será la utilizada para calcular los valores relativos de otras características.

Se obtiene ahora el número de emojis utilizados y dividiéndolo por el número medio de palabras se obtiene una nueva característica. Para poder tratar los emojis ha sido necesario incluir una librería especifica.

A partir de una lista de preposiciones se ha generado una característica nueva contando las que existen por autor y dividiéndolas por el número medio de palabras.

Mediante una bolsa de adjetivos y una función que los separa por género se ha obtenido el número medio de adjetivos masculinos y femeninos. Estos dos valores se han dividido por el número medio de palabras, generando así dos nuevos valores relativos (características).

Para la generación de una última característica se ha realizado una operación lógica entre el número de adjetivos masculinos y femeninos, guardándose el resultado como entero. Dicha operación lógica consiste en comparar si el número de adjetivos masculinos es mayor al de adjetivos femeninos, devolviendo 1 en caso afirmativo y 0 en caso negativo.

Finalmente se genera un dataset con todas las características detalladas anteriormente en un archivo de texto para a continuación poder aplicar los modelos.

3. Propuesta del alumno

En primer lugar comentar que todas las características detalladas en los apartados anteriores han sido fruto de una lluvia de ideas de todos los componentes del equipo. Comentado esto, algunas propuestas que podrían mejorar los resultados, serían las que a continuación se detallan.

Si observamos las bolsas de palabras, comprobamos que existen algunos números. Entiendo que estos números pueden ser valores puntuales y por ello descartándolos de las diferentes bolsas se podría obtener unas bolsas más fiables.

Por otro lado, se podría dar un peso distinto a cada una de las palabras. El peso de cada una de las palabras podría coincidir con la frecuencia (TF), quedándonos de esta forma con una bolsa de palabras más reducida pero más fiable.

Por último, podría ser interesante, analizar (aunque sea de una forma superficial) las referencias y links introducidos en los tweets. Para analizar esto se podría seguir un funcionamiento similar al utilizado hasta ahora, es decir, contar el número de menciones y de links y obtener su valor relativo dividiéndolo por el número medio de palabras.

Referente al análisis de variedad lingüística, además de las características obtenidas para el análisis de género se podrían incluir dos características más.

Por un lado, se podría evolucionar la técnica de las bolsas de palabras comparándolas con las demás bolsas. Es decir, una vez obtenidos los términos más frecuentes (TF), compararlos con los términos más frecuentes de las demás variedades lingüísticas. De esta forma obtendremos los términos más frecuentes de cada una de las variedades lingüísticas en concreto.

La ultima característica de basaría en el análisis de las extensiones de los links. Debido a que las direcciones en Twiter se muestran de una forma reducida, en primer lugar se debería obtener la dirección completa y después generar una bolsa de extensiones más comunes para cada una de las variedades lingüísticas. Además podríamos reducir las bolsas manteniendo únicamente las exten-

siones con mayor frecuencia incrementando así la fiabilidad de la técnica.

4. Resultados experimentales

A partir del dataset generado en el apartado anterior aplicaremos diferentes modelos y utilizaremos el accuracy y el tiempo de ejecución para intentar determinar cuál es el modelo más adecuado. Los modelos que aplicaremos son: SVM, decisión Tree, Random Forest y Neural Net.

SUPER VECTOR MACHINE: Observamos se sufre un incremento importante en el tiempo de ejecución conforme aumenta el tamaño de la bolsa de palabras. Por otro lado el modelo reduce su tiempo a partir de las 1000 palabras. Sin embargo el accuracy no se ve demasiado afectado por la variación del tamaño de la bolsa de palabras (incluso empeora a mayor número de palabras).

WORDS	ACCURACY	TIME
50	**0.68**	12.16
100	0.68	27.8
500	0.66	83.5
1000	0.63	10.08

Cuadro 1: Resultados SVM.

DECISION TREE: Se observa que a medida que crece la bolsa de palabras se incrementa levemente el tiempo, quedando siempre por debajo del segundo, mientras que el accuracy no supera el 0,6.

WORDS	ACCURACY	TIME
50	0.56	0.09
100	0.57	0.15
500	**0.60**	0.52
1000	0.60	0.92

Cuadro 2: Resultados DECISION TREE.

RANDOM FOREST: En este caso, se no se aprecia una variación importante ni del accuracy, ni en el tiempo de proceso, a medida que aumenta la bolsa de palabras.

NEURAL NET: En este modelo se observa un crecimiento importante del accuracy, mientras que el tiempo de procesado no sufre demasiadas variaciones (variando en algo más de medio segundo).

WORDS	ACCURACY	TIME
50	0.60	0.25
100	0.64	0.25
500	**0.64**	0.23
1000	0.64	0.26

Cuadro 3: Resultados RANDOM FOREST.

WORDS	ACCURACY	TIME
50	0.60	2.5
100	0.65	3.85
500	0.71	2.89
1000	**0.72**	3.19

Cuadro 4: Resultados NEURAL NET.

5. Conclusiones y trabajo futuro

Durante la realización de este trabajo hemos podido comprobar las particularidades de realizar modelos para la clasificación de textos escritos en redes sociales. Hemos comprobado cómo, en estos casos, Big Data se aplica más al enriquecimiento de los datos que a la cantidad de los mismos.

Centrándonos en el análisis de los resultados obtenidos, rechazaremos los modelos obtenidos mediante SVM y Decision Tree dado que tienen los accuracy más bajos de modelos utilizados. Sin embargo para decantarnos por uno de los otros dos modelos (Random Forest o Neural Net), deberíamos plantearnos si sacrificamos accuracy o tiempo de ejecución. Es decir, si en nuestro caso real, tenemos alguna limitación de tiempo, es posible que debamos sacrificar accuracy y seleccionar **Random Forest**. Si por el contrario no existe ninguna limitación de tiempo, los mejores resultados han sido los obtenidos por **Neural Net**.

Como ampliación del análisis o trabajo futuro se podrían añadir nuevas características para enriquecer la información proporcionada. Se proponen la creación de dos nuevas características que se definen a continuación.

Por un lado se podría crear una nueva bolsa de palabras pero en este caso que contenga diferentes temáticas. Al igual que se han realizado trabajos de forma manual para el ejemplo de AuthorProfiling, en este caso también debería ser, en principio, una bolsa manual. Por un lado se

crearía una lista de temáticas como por ejemplo ocio, viajes, corazón, deportes, etc y por otro de forma manual se asociarían los diferentes links a estas temáticas y se etiquetarían.
De esta forma se podría ver si hay una temática que es más propia de una variedad lingüística u otra o si hay temáticas más propias de un género u otro. Con toda esta información podríamos crear una característica para cada una de las temáticas en las que se indicaría el número de links de esa temática divido por el número medio de palabras por autor.

La siguiente técnica propuesta se basaría en el cálculo de n-gramas. La idea es crear bolsas de bígamas y trigramas más utilizados por género o por variedad lingüística. Las características que se derivarían de esta técnica son el número de bígamas (y trigramas) dividido por el número medio de palabras por autor.

References

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