Asignatura Text Mining en Social Media. Master Big Data. Author Profiling en Twitter

Juan Falomir Orti
jfalomirorti@gmail.com

Abstract

En nuestra sociedad, regida por una democracia burguesa representativa del siglo XXI, es realmente importante controlar los canales de comunicación por parte de los poderes fácticos. Esto desemboca en un paradigma en el cual es fundamental conocer la capacidad de imbuir a la población en estos mensajes que buscan, muchas veces, ser mayoritarios. Tener herramientas para el análisis sociológico es una tarea capital y, las tradicionales encuestas teléfonicas, parecen cada vez más obsoletas.

La aparición de nuevas formas de comunicarse, como las redes sociales (Whatsapp, Instagram, Twitter, etc...), que han tenido una expansión exponencial y que muchas veces funcionan como termómetro de la población frente a los problemas sociales o cotidianos ha abierto la oportunidad de extraer información de estos contextos y cubrir la necesidad sociológica que expresábamos en el párrafo anterior.

La llegada de tecnologas conmunmente llamadas ”tecnologás de BIG DATA” ha permitido afrontar este tipo de problemas o necesidades que hace 10 años eran difícilmente asumibles o resolubles. Y en este caso que nos ocupa, es importante conseguir que los ordenadores puedan entender el lenguaje natural de los humanos.

La morfología, la sintaxis, la morfología y la pragmática son elementos vitales en el lenguaje escrito. El análisis de los mismos dependen del propio contexto en el que son utilizados y elementos como la ironía o el sarcasmo hacen que la parametrización de un modelo que sea efectivo en su estudio resulte tecnicamente imposible.

Sumando todos estos factores nos lleva a la problemática que este estudio quiere resolver, que es ser capaces de reconocer ciertas características de un usuario por su forma de comunicarse en Twitter. Las características elegidas son el género (masculino o femenino) y la variedad del castellano que usan (ñol, chileno, venezolano, mexicano, argentino, colombiano o peruano). De esta forma, es fácil analizar los problemas que pueden afectar a distintos países o como se posicionan los dos géneros frente a situaciones sociales.

Nuestra solución pasa por analizar estadísticamente características relativas al contexto comunicativo de Twitter, y entrenar un modelo de Machine Learning (ya sea de multinomial de NaiveBayes, randomForest o Support Vector Machine) basándonos en la técnica de Tf-idf para parametrizar la importancia de estos elementos y los del lenguaje natural en Twitter.

1 Introducción

Para conseguir crear el modelo que sea capaz de analizar el género y variedad linguistica en Twitter nos parece importante explicar y exponer ciertas características de la comunicación en este contexto, ya que no estamos en la pragmática de una novela, un ensayo o una conversación.

En Twitter, la característica más distintiva es que los mensajes no pueden superar los 280 carácteres. Además , hay otras diferencias entre esta vía de comunicación y cualquiera natural que son especialmente importantes para nuestra forma de abordar el problema:

- Menciones: son palabras precedidas por @

que indican la persona o usuario a la que va dirigido el tuit.

- Hashtags: son palabras o expresiones precedidas por # que suelen indicar el tema del que habla el twit.

- Emojis: son expresiones prediseñadas que muestran sensaciones o estados de ánimo como :-) (Alegría).

- Retuit: son tuits caracterizados por estar escritos por un tercero pero que se muestran como si fueran propios.

- URLs: son enlaces a páginas de terceros o a elementos multimedia.

Por último en twitter hay trending topics, son temas de los que se hablan mayoritariamente y que marcan mucho el vocabulario utilizado: deportes, política o actualidad son algunos ejemplos.

El estudio de estadísticas relativas al uso de estos elementos que son distintivos de twitter nos parece importante para resolver el problema de Author Profiling. Creemos que la metodología a aplicar no es la misma en un entorno como Twitter que en análisis de literatura clásica o más forma. Así que vamos a dividir el trabajo en tres partes: análisis del dataset desde este punto de vista, propuesta para resolver el problema, y resultados finales. Además daremos alguna clave de futuro.

2 Dataset

El dataset es una colección de tuits de 2800 usuarios de train y 1400 de test con 100 tuits cada uno. Cada uno de los usuarios tiene asociada un género y una variedad del lenguaje. Antes de comenzar con en análisis estadístico hemos hecho un primer proceso en el cual creamos una cadena con todos los tuits juntos, sin diferenciar entre ellos, ya que creemos que esto ayudará al estudio y al propio modelo.

Llegados a este punto tenemos 2800 cadenas de texto de train y 1400 cadenas de texto de test.

Hemos calculado el número de menciones, retuits, hashtags, emojis, URLs y menciones entre los ditintos textos y hemos calculado la media entre los que tienen el género o variedad. Además hemos calculado la longitud media normalizada de las cadenas de textos. Los resultados son los siguientes:

male :		female :	
Emojis	12.245714	Emojis	27.532857
hashtags	25.573571	hashtags	26.440357
Mentions	73.790357	Mentions	62.311071
URLs	41.506786	URLs	37.919286
RTs	1.118571	RTs	0.783571
longitud	0.666084	longitud	0.611633

chile :	
Emojis	18.753750
hashtags	28.263750
Mentions	72.977500
URLs	35.403750
RTs	0.768750
longitud	0.641618

venezuela :	
Emojis	4.757500
hashtags	27.551250
Mentions	67.102500
URLs	48.678750
RTs	2.792500
longitud	0.702906

peru :	
Emojis	24.821250
hashtags	23.731250
Mentions	65.630000
URLs	38.721250
RTs	0.458750
longitud	0.632717

spain :	
Emojis	21.398750
hashtags	31.010000
Mentions	79.577500
URLs	47.763750
RTs	0.760000
longitud	0.674259

colombia :	
Emojis	22.218750
hashtags	20.487500
Mentions	65.093750
URLs	32.512500
RTs	0.717500
longitud	0.617075

mexico :	
Emojis	25.612500
hashtags	34.196250
Mentions	71.040000
URLs	47.806250
RTs	0.570000
longitud	0.647087

argentina :	
Emojis	21.662500
hashtags	16.808750
Mentions	54.933750
URLs	27.105000
RTs	0.590000
longitud	0.556347

En general, vemos que hay una diferencia notable a nivel de género en el uso de Emojis (más del doble el género femenino) menciones (casi un 20% más por parte de los hombres), a nivel de RT (un 40% más los hombres) y la longitud (un 9% más por parte de los hombres). Desde el punto de vista estadístico y si hicieramos intervalos de confianza, podríamos concluir que las evidencias son suficientes como para decir que son, realmente, grupos diferentes.

A nivel de variante llama la atención que en venezuela no se usan casi emojis y muchos retuits, la escasa longitud de los tuits en Argentina y pocos enlaces, los pocos hashtags en colombia o la cantidad ingente de menciones en España, pero lo que más llama la atención a nivel de lenguaje son las palabras utilizadas.

Además del estudio realizado antes, también hemos extraído la lista de palabras más utilizadas en los distintos países y géneros. Entre las personas de género masculino destacan palabras como ”gobierno”, ”pais”, ”equipo” o expresiones como ”jajaja” y en cambio en las personas de género femenino encontramos palabras como ”quiero”, ”amor”, ”vida” o ”siempre. Pero la diferencia principal la encontramos entre las variedades del idioma y llama la atención que entre las palabras utilizadas en cada país hay muchos gentilicios o nombres de sitios del país de de donde procede el hablante.

Así que tras analizar el dataset podemos llegar a dos conclusiones muy claras: la primera es que en Twitter abundan recursos que en un lenguaje más formal no se usan, o no está el lenguaje preparado, como en el caso de los emojis; y la segunda es que, haciendo un estudio de estos recursos ya encon-

tramos muchas diferencias entre los géneros en el uso del lenguaje y de estas herramientas. Ahora vamos a desarrollar, a partir de estos datos una solución al problema.

3 Propuesta del alumno

Tras analizar los datasets ya estamos en disposición de diseñar una solución para el problema. Además de las propios textos vamos a añadir la información del número de emojis, hashtags, retuits, URLs y menciones junto a la longitud media de los párrafos al dataset. Posteriormente normalizaremos estas medidas ya que en situaciones de Machine Learning se mejoran ligeramente los resultados. Tras este preprocesado, aplicamos Tf-idf a los dataset para extraer las palabras o expresiones más representativas del corpus ya que también queremos incluír bi-gramas (un n-grama es un conjunto de n palabras juntas). Y tras todo este proceso entrenamos tres tipos de modelos para quedarnos con el que mejor resultado obtenga, los modelos elegidos son:

- Multinomial de Naive-Bayes
- Support Vector Machine
- Random Forest con 100 nodos y una profundidad máxima de 25.

Tras ejecutar los tres modelos, hemos perfilado las distintas variables de entrada para obtener los mejores resultados:

```
vectorizers={
    'sexo':TfidfVectorizer(analyzer="word", stop_words=stopwords.words('spanish'),\
                           ngram_range=(1,2), max_df=0.9,max_features=10000),
    'pais':TfidfVectorizer(analyzer="word",\
                           ngram_range=(1,2),max_features=20000)
}
```

Esto es, usar bigramas, stopwords en caso de género, 10000 features en caso de género y 20000 en caso de variedad.

4 Resultados experimentales

Tras la propuesta perfilada los resultados son los siguientes:

```
MultinomialNB :
         sexo :  0.6871428571428572
         pais :  0.7971428571428572
         Joint:  0.5342857142857143
LinearSVC :
         sexo :  0.76
         pais :  0.9342857142857143
         Joint:  0.7085714285714285
RandomForestClassifier :
         sexo :  0.7257142857142858
         pais :  0.9292857142857143
         Joint:  0.6721428571428572
```

Y podemos observar como los mejores resultados los obtenemos con Support Vector Machines, con una tasa del 76% de acierto a nivel de género, un 93,4% a nivel de variedad y un 70% en los casos en los que hemos acertado ambos. Tras ver los resultados de otros años en el PAN vemos que es un resultado que se situaría en el Top 10 de ambas categorías (género en castellano y variedad de castellano).

5 Conclusiones y trabajo futuro

Tras los resultados obtenidos podemos concluír que es muy importante analizar un problema en su contexto adecuado y analizando los pequeños detalles. La solución aportada es muy efectiva en el entorno Twitter pero no es exportable a otras pragmáticas; así como tampoco sería tan efectivo usar algorítmos de profiling de literatura clásica a Twitter porque el contexto no es el mismo.

De cara a futuro trabajo y vías de mejora vemos que podría analizarse muchos más factores en el entorno Twitter que en el dataset no se reflejaban, como personas con las que interactua el usuario, si la ubicación (que sería una gran ayuda para la variedad) no está activada, la hora en la que se escribe el tuit (puede ayudar a diferenciar el dialecto Europeo de los sudamericanos) o tener triggers de uso de temas relativos a la sociedad en la que vive, como conciertos.

References

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