Obtención de información de usuarios a través de Twitter

Antonio López Dios

***Resumen*—A lo largo de este trabajo se pretende hacer una breve aproximación a la tarea de la obtención de la información a través de los textos producidos por usuarios anónimos. Para ello se utilizará un conjunto de mensajes previamente almacenados de la red Twitter. Se pretende informar paso a paso de los procesos más relevantes realizados para conseguir averiguar el género del usuario y su región, teniendo en cuenta de que todos los usuarios seleccionados son de habla hispana. El objetivo de este artículo es doble, por un lado informar del proceso elaborado para dar respuesta al problema y por otro dejar constancia de la cantidad de información que se puede conseguir de un usuario con un relativo poco esfuerzo**

***Keywords*—Twitter, analisis, R, Social Media, Text Mining.**

I. Introducción

I-A. ¿Qué es Twitter?

Twitter es una de las aplicaciones webs más conocidas a día de hoy. Presenta, entre otras características, todas las virtudes de los blogs, la mensajería instantánea y las redes sociales. Dicho de otro modo, Twitter es una plataforma de comunicación bidireccional con naturaleza de red social. Gracias a esta aplicación, sus usuarios pueden estar en contacto en tiempo real a través de micromensajes de texto, llamados tweets, los cuales constan de un tamaño máximo de 140 caracteres. Evidentemente también adolece de algunos de los problemas de las redes sociales, como puede ser el no saber nada de la persona que escribe. El obtener información de los usuarios a través de sus mensajes en twitter es una tarea de gran interés debido a la enorme cantidad de aplicaciones prácticas que de ella se pueden derivar, desde marketing, estudios de mercado hasta incluso seguridad. Es por ello que se ha decidido como primer acercamiento a la tarea, conseguir un programa que sea capaz de clasificar a los usuarios por género (Hombre-Mujer) según lo escrito en sus tweets y otro en el que lo que se distingue es la nacionalidad del usuario (siendo diferentes paises de habla hispana). Para conseguirlo se han gastado como herramientas únicamente el lenguaje de programación R en su entorno de programación R-Studio y un dataset proporcionado por el profesor de la asignatura, el cual se pasará a explicar a continuación.

II. Dataset

El dataset con el que se va a trabajar está caracterizado entre otros factores en que:

- Todos sus datos se obtienen de Twitter
- Es una colección de información sobre miles de usuarios de la aplicación.

- Hay una enorme variedad de temas tratados en sus tuits, no se filtra por tema.
- Ocupa un tamaño aproximado de 54 MB descomprimido.

Para la creación del dataset se ha seguido el siguiente proceso de construcción:

1. Se han recuperado una serie de tuits enmarcados en ciertas regiones concretas.
2. Se hace un filtrado por idioma de los tuits.
3. Se recuperan los timelines de los usuarios
4. Se recuperan los autores que reúnan como requisito que tengan más de 100 tuits en el idioma buscado y en la localización geográfica adecuada en su perfil.

El formato de los ficheros descomprimidos es la siguiente:

- Un par de ficheros: 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

III. Propuesta del alumnado

Como se ha comentado anteriormente se ha pretendido obtener información de los usuarios de Twitter a través de sus mensajes en dicha aplicación. El primer acercamiento ha sido intentar por una parte averiguar el género del usuario y en segundo lugar la región. Para ello se han utilizado diferentes métodos que a continuación se detallarán: En el caso del género primero se ha seguido los siguientes pasos:

1. Limpiar el corpus quitando las palabras más comunes que carecen de significado
2. Eliminar los acentos y valores numéricos del corpus

Se ha pensado de que manera sería posible distinguir entre hombres y mujeres, llevando a cabo una sucesión de experimentos a través de los estereotipos más comunes, como por ejemplo que las mujeres hablan más de sentimientos que los hombres o que utilizan más adjetivos. Para ello se han creado los siguientes diccionarios:

1. Adjetivos.txt: Es un archivo en el que se encuentran cientos de adjetivos
2. Pronombres.txt: En este archivo se encuentran diversos pronombres
3. Sentimientos.txt: En este caso se trata de un archivo con palabras relativas a los sentimientos
4. Horóscopos.txt: El vocabulario de este fichero está extraído del horóscopo del amor de la revista Cosmopolitan, después de haberlo utilizado se ha pensado que esta revista va dirigida a un público muy concreto, con lo que se paso a realizar otro diccionario de una revista más generalista.
5. Superpop.txt: Diccionario de palabras obtenidas

de diversos artículos de dicha revista.

6. En mujeres.txt y hombres.txt nos encontramos con diccionarios de las palabras más utilizadas por los hombres y las mujeres.

Se debe indicar que para el cálculo no se utilizaba las veces que un autor repetía una palabra (o grupo de palabras) sino que esa frecuencia era dividida por el número total de palabras de ese autor. No es lo mismo que un autor diga 5 veces la palabra amor en 100 palabras que en 100000. Se ha ido probando de diversas formas mezclando la utilización de los diccionarios y la preparación del corpus y hay que decir que la mejor combinación ha sido el dejar las palabras con acentos, sin quitar las stopwords y usando los diccionarios anteriormente relatados. Con todo esto se ha obtenido un resultado de 74 % de accuracy y un kappa de 0,48 mientras que el valor inicial era de 66 % de accuracy, luego la mejora ha sido considerable. Hay que destacar que se han utilizado diferentes modelos (SVM, Random forest…) y el que ha dado unos resultados mejores ha sido el "Penalized multinomial Regression" En el caso de la variedad se ha intentado abordar el problema centrándose en las palabras más frecuentes de cada variedad dialectal. En este caso lo primero que también se ha realizado ha sido quitar los acentos de las palabras. Posteriormente al dataset se le han añadido 7 columnas, las cuales hacen referencia a cada una de las variedades por región de las que trata el problema. En dichas columnas se reflejan la frecuencia en las que se utilizan las 300 palabras más características de cada región. En la fase de entrenamiento se prueba con los mismos métodos que los descritos en el problema del género. Hay que indicar que en este caso el método más efectivo es el Random Forest. El mejor resultado obtenido en este problema ha sido de 86,36 % lo cual es casi 10 puntos por encima del resultado inicial.

IV. Conclusion

De todo lo comentado anteriormente se puede extraer algunas conclusiones interesantes: En el caso de las diferencias por género, se puede decir que al menos inicialmente, es más sencillo distinguir a los usuarios a través del significado de las palabras más que por la repetición de las mismas. Es decir, se ha podido comprobar que en principio las mujeres hablan de más temas diferentes que los hombres. Esto no quiere decir que ellas hablen más sino que ellas hablan de todos los temas y los hombres hay temas que tratan menos, por eso mismo es más sencillo clasificarlas a ellas que a ellos. Además la forma en la que están escritas las palabras también es característica del género, de ahí que mejore el resultado cuando se dejan las palabras con acentos. En el caso de la variedad se puede decir que en todas las regiones las palabras se escriben de manera similar, aunque hay expresiones típicas en ciertas regiones que no son para nada utilizadas en otras.

V. Trabajo futuro

En primer lugar sería interesante continuar con estos problemas añadiendo por ejemplo una forma de contabilizar la frecuencia en la que se gastan los emoticonos. Sí también pudiese contabilizarse la frecuencia con la que se utilizan ciertos tiempos

verbales se cree que sería de mucha utilidad en el análisis. Por otra parte, sería interesante ampliar la actividad haciendo un problema en el que se predijera la edad del usuario por ejemplo o una combinación de las anteriores es decir que fuese capaz de predecir la nacionalidad y el género del usuario a la vez.

VI. Referencias

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.cosmopolitan.com/horoscopo/