Twitter author profiling

Alejandro Fuster Baggetto
alejandrofuster1@gmail.com

Abstract

Se han fusionado bolsas de palabras y bigramas con características propias del estilo, con el objetivo de predecir el sexo y la variante lingüística de un autor dados sus tweets. Concretamente, para el estilo se han definido características como la cantidad de emoticonos y hashtags utilizados o la longitud de los tweets. Con estos añadidos se pretende complemetar la información obtenida a partir de las bolsas de palabras para subir la precisión del modelo especialmente en el caso del sexo, donde el estilo es mucho ms diferenciador que el vocabulario.

1 Introduccin

Millones de posts, tweets, publicaciones, etc son emitidos cada da en las redes sociales. En estos medios hay tal cantidad de información que un análisis de su contenido puede conducirnos a todo tipo de conclusiones sobre las tendecias sociales, los temas más actuales y la opinin pública sobre acontecimientos o personas. Siendo un problema tan importante, existe hoy en da una gran cantidad de software que analiza en tiempo real estos contenidos para obtener información de tipo valiosa a nivel comercial, de negocio, social, política, etc. Sin embargo, no solo es interesante analizar el contenido de las publicaciones, sino tambn el perfil de su autor, por ejemplo para agrupar autores con intereses u opiniones similares. Este otro problema, llamado author profiling, no es ni mucho menos trivial y se complica conforme vamos queriendo averiguar más sobre el autor. En este paper, únicamente vamos a predecir el sexo y la variante lingüística del autor en base a sus tweets, aunque muchas otras características como la edad, o la ideología política podrían ser predichas de forma semejante.

Todo el código desarrollado para la elaboración de este paper se encuentra en el este repositorio.

2 Dataset

Para este paper se ha utilizado un dataset de con un total de 4200 autores de Twitter de habla hispana, de los cuales 2800 han sido utilizados como conjunto de entrenamiento y 1400 como conjunto de vaalidación. El dataset incluye 100 tweets por autor, que son todos los datos de los que dispondremos para hacer las predicciones. No obstante, en un problema real, podrían tenerse en cuenta muchas otras características: Por ejemplo, el nombre de usuario nos aportaría mucha información a la hora de determinar el sexo, mientras que la localizacióon (u otros metadatos) de los tweets emitidos por un usuario y sus amigos podrían confirmarnos claramente la procedencia y por lo tanto la variante lingüística del mismo. Dicho esto, este paper se centra únicamente en el análisis del contenido de los tweets.

3 Propuesta del alumno

Una de las primeras hipótesis que hemos hecho ha sido la de que las palabras son mucho ms informativas para predecir la variante, mientras que el estilo es lo que más podría marcar la diferencia entre sexos. Para comprobar si esto es cierto o no, hemos eliminado las stopwords (palabras muy comunes que aportan poca información, como pueden ser las preposiciones o las conjunciones) de los tweets y hemos creado un wordcloud para hombres y otro para mujeres. Este tipo de gráficas representan las palabras más comunes de un texto junto con su frecuencia, que se representa mediante el tamao. En este caso, hemos concatenado todos los tweets de autores de cada sexo en un texto distinto. El resultado ha sido el esperado. Ambos sexos utilizan vocabularios similares, lo cual nos indica que necesitaremos extraer más características de los tweets si queremos obtener bue-

nas predicciones para el sexo. Hemos probado los wordclouds agrupando los tweets por variante y hemos podido confirmar que los vocabularios de cada variante se diferencian lo suficiente entre si como para que la bolsa de palabras nos de muy buenas predicciones. Un ejemplo de esto es que en los wordclouds se puede ver que una de las palabras más utilizadas en cada país es el propio nombre del país (En España se dice mucho España y en Mexico se dice mucho Mexico).

Teniendo en cuenta los resultados de los wordclouds, hemos intentado extraer más características (en este caso de estilo) que si que puedan ser significativas en el caso del sexo. Hemos extraido la cantidad de emojis, hashtags, menciones, URLs, retweets y la longitud. Sin duda, de entre las mencionadas, la característica más diferenciadora entre sexos es el número de emojis. Las mujeres utilizan una media de 2.75 emojis por tweet, mientras que los hombres utilizan una media de 12.24. Es decir, que las mujeres escriben en Twitter más del doble de emojis que los hombres. De forma mucho menos concluyente, las estadísticas nos muestran que los hombres hacen un 30% más de retweets. El resto de características extraidas varían muy poco entre géneros.

Como se ha indicado antes, se ha utilizado una bolsa de palabras. En este caso, se ha generado con TfxIDF, lo cual nos aporta mayor información sobre la importancia de las palabras que símplemente la frecuencia. Para esta bolsa, se han tenido en cuenta tanto los unigramas como los bigramas. Adems, para el caso de la variante lingüística se ha hecho una bolsa más grande (con el doble de características), debido a que el vocabulario es muy significativo para ese problema en concreto (como ya se ha comprobado). A estas características de la bolsa de palabras les hemos añadido las características de estilo que hemos extraido (cantidad de emojis, retweets, etc) normalizadas entre 0 y 1.

En cuanto a los clasificadores, se han probado tres de ellos: El más rápido y sencillo es el Naive Bayes multinomial. Este tipo de clasificadores intentan estimar la probabilidad aposteriori de cada clase dada una muestra, para lo cual asumen una determinada distribución de las variables. A pesar de ser tremendamente simples, estos clasificadores no suelen funcionar mal para lenguaje natural, ya que la frecuencia de cada palabra en un texto puede aproximarse a una distribución multinomial.

El segundo clasificador que se ha probado es la máquina de vectores de soporte lineal. Las SVMs obtienen el clasificador de margen máximo dada una función kernel. En este caso, por simplicidad, no se ha utilizado kernel (o lo que es lo mismo, se ha utilizado el lineal).

El tercer y último modelo que se ha probado es el random forest, que genera de forma semialeatoria una determinada cantidad de árboles de decisión y clasifica mediante una votación entre dichos árboles. Los random forest son ensembles, y como tal, a pesar de no ser conceptualmente complejos, son modelos relativamente pesados (Mucho más lentos de entrenar que cualquier Naive Bayes o SVM lineal).

4 Resultados experimentales

Los resultados obtenidos han sido bastante satisfactorios, sobre todo teniendo en cuenta que únicamente teníamos acceso a los tweets de cada autor y no a su cuanta, su lista de amigos, los metadatos de los tweets, etc. La máquina de vectores de soporte lineal es el modelo que mayor accuracy nos ha dado: 0.76 para el sexo, 0.93 para la variante y 0.7 para la combinada. El random forest, siendo un clasificador mucho más potente que la SVM lineal, nos ha dado resultados muy similares (incluso algo peores) que esta. Por otro lado, el Naive Bayes, nos ha dado resultados significativamente peores (0.68 para el sexo, 0.79 para la variante y 0.53 para la combinada), pero no muy malos teniendo en cuenta la simplicidad del modelo.

5 Conclusiones y trabajo futuro

Definitivamente, el problema de detectar la variante es mucho más sencillo que el del género, al menos con los datos de los que se partía para este paper.

Como ya se ha dicho, lo que más mejoraría estos resultados, sería enriquecer el dataset y no limitarnos únicamente al contenido de los tweets.

También se podrían mejorar los resultados (Concretamente los de detección de género) extrayendo más características de estilo como la cantidad de pronombres utilizados, que podrían ser o no relevantes para el problema.

Otra forma de aumentar ligeramente la precisión sería explorar más modelos y/o realizar

una buena y exhaustiva optimización de hiperparámetros.

Por último, sería interesante explorar el contenido de los hashtags y las URLs, pues podría ser de utilidad para el problema. Por ejemplo, una URL que contenga una noticia acontecida en España puede ayudarnos a deducir que el autor del tweet es español.

References

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