Asignatura Text Mining en Social Media. Master Big Data

José Antonio Miras Ferrando
pemife16@gmail.com

Abstract

En mi opinión, para abordar un problema de Author Profiling, debemos de tomar dos caminos diferentes, pero que finalmente convergerán en unos resultados más acertados. Un camino es el del enriquecimiento del modelo, estamos de acuerdo con que el dataset principal serán los propios tweets tokenizados, pero nuestra labor será incorporar al modelo nuevas variables que puedan hacer que clasifique de un manera mejor. Aquí es donde entra en juego la imaginación y aunque por motivos de tiempo en clase no se pudieron implementar, si que se nos ocurrieron un buen número de ideas que consideramos enriquecedoras. Algunos serán mejor para ayudar a determinar el genero y otras para determinar la región. Por ejemplo, para clasificar la región puede ser positivo recuperar las url mencionadas y recuperar el país de los dominios (.es, .co...). Por otro lado la creación de variables para determinar el género, como por ejemplo la longitud media de los tweets o una idea que surgió de una broma que fue a la de un diccionario que etiquetas con colores complejos (rosa palo) y colores simples (rojo), basándonos en la premisa no confirmada de que las mujeres son más precisas a la hora de nombrar colores. Aunque por temas de tiempo estas ideas no han sido implementadas. El otro camino a seguir es la elección del modelo y su ajuste de parámetros, en este caso el tratamiento sería de hacer pruebas empíricas y comparar resultados para elegir el modelo con un accuracy más acertado. Combinando un buen enriquecimiento del conjunto de datos y el modelo que hayamos encontrado más acertado es como creemos que podríamos tener una tasa de acierto mayor.

1. Introducción

Podríamos afirmar que el Author Profiling trata de obtener información de quien ha escrito un mensaje, ya sea su edad, su sexo, país u otras características. Concretamente el mismo corpus para este trabajo presenta dos problemas, en primer lugar, etiquetar el género del autor del tweet y por otro lado determinar la variedad de español del autor de cada tweet.

2. Dataset

Para la aplicación de un módelo a Author Profiling disponemos de un dataset de twitter de varios autores, donde de cada autor disponemos de 100 tweets, los cuales están etiquetados tanto a nivel de género como de variedad.

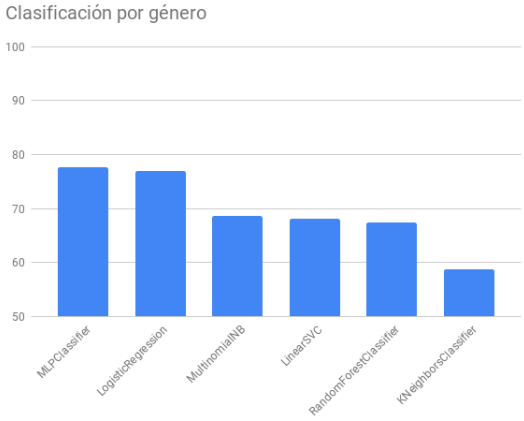
3. Propuesta del alumno

Despues de lo mencionado en el abstract creemos que la propuesta de valor que podemos hacer para este tipo de problemas es la aportación de ideas creeativas que enriquezcan el módelo y puedan aportar para una clasificación mejor.

4. Resultados experimentales

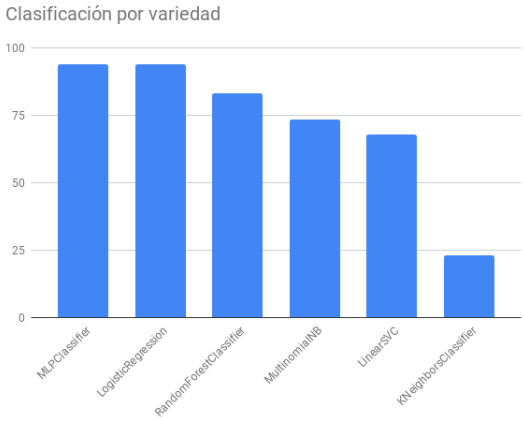
Procedemos a presentar dos tablas con el porcentaje de aciertos de los diferentes módelos probados.

Aquí tenemos los resultados para el problema de clasificación por género.



MLPClassifier	77,57
LogisticRegression	76,92
MultinomialNB	68,64
LinearSVC	68,14
RandomForestClassifier	67,35
KNeighborsClassifier	58,71

Aquí tenemos los resultados para el problema de clasificación por variedad.



MLPClassifier	93,78
LogisticRegression	93,64
RandomForestClassifier	83,14
MultinomialNB	73,21
LinearSVC	67,64
KNeighborsClassifier	23

Como análisis de los resultados resulta curisoso como un problema más sencillo como el de clasificar por género, ya que se trata de una clasificación binaria, devuelve unas tasas de acierto peores que la clasificación por variedad, que se trata de categorizar entre 7 variedades. Esta curiosidad, la podemos argumentar con el tipo de problema, probablemente en la sociedad actual no existen fuertes diferencias en la comunicación escrita entre hombres y mujeres y si que existan más palabras diferenciadoras entre los diferentes paśes de habla hispana.

5. Conclusiones y trabajo futuro

Breve presentación de las conclusiones sobre el trabajo realizado e ideas de futuro para mejorar los resultados.

Como hemos explicado a lo largo del documento, creemos que el factor diferenciador en un accuracy mejor, serán las variables adicionales que podamos crear, ya que sin añadir variables el problema simplente se resume en probar los difrentes módelos y quedarnos con el que mejor resultado devuelva.

Como trabajo futurio quedaría la implementación de algunas variables que hemos mencionado pero no implementado.

Otra posible tarea pendiente que podría ser interesante sería la de aplicar PCA para saber que variables son las más determinantes a la hora de determinar el género o la variedad.

References

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