Machine learning para predecir el sexo y pais del autor de un tweets

Autor Ricardo Cancar

RESUMEN.

Las redes sociales son una fuente de información muy valiosa, con el uso de herramientas Big Data y técnicas de machine learning se puede analizar la forma de escritura de un usuario de una red social y obtener conclusiones interesantes como el país de origen el sexo sus intereses económicos, políticos etc.. a gran velocidad. Aunque para la cantidad de datos que se maneja en este estudio no requiere de una solución big data si se aplican técnicas estadísticas y de machine learning para poder hacer la predicción de género, y nacionalidad entre países de habla hispana mediante un modelo que analice los tweets de estos usuarios, se desea hacer una predicción de género y nacionalidad de los usuarios de twitter de habla hispana según su forma de escribir, un caso de uso real del profiling en la redes sociales usando modelos de machine learning, es obtener un patrón de conducta según el tipo de usuario para realizar marketing dirigido es decir si es español y habla mucho de fútbol se le pueden enviar a dicho usuario publicidad para que compre entradas al partido de fútbol o un balón de fútbol o tenis deportivos etc...

INTRODUCCIÓN

En esta memoria se pretende describir el proceso que derivó en la creación de un modelo capaz de predecir tanto como el género “femenino o masculino” como el país (chile, Argentina, Mexico, Colombia, España, Venezuela y Peru) del autor del twitter analizado.

Para ello se dispone de un dataset para el training de 2800 autores donde cada país tiene 400 y de los cuales la mita es del género femenino y la otra mitad del generó masculino.

Como dato de test tendremos 1400 autores de los cuales hay 200 autores por cada país y de estos 100 son femeninos y 100 masculinos.

Mas adelante se vera en detalle los modelos aplicados para realizar estas predicciones, pero cabe destacar que el objetivo es realizar predicciones por clases, por lo que se usaron modelos de clasificación de los cuales el modelo que mejor nos dio resultados fue un randomforest, con un 74% de precisión en la clasificación por género y un 93% de precisión en la clasificación de variedad.

CARACTERÍSTICAS DEL DATASET.

El dataset con el que se trabajo cuenta con la id del autor del tweet, el tweet en texto, su sexo y el país al cual pertenece el autor, el tratar estos datos tienen su grado de complejidad, porque muchos tweets tienen errores ortográficos, símbolos pueden contener caracteres repetidos y emojis lo que dificulta el pre-procesado de datos.

Los Datos utilizados para este proceso están almacenados en archivos “.xml” los cuales tienen la identificación de usuario en twitter, 100 tweets por cada usuario los cuales son etiquetados con el sexo y país de origen del autor.

Para formar los datos de entrenamiento se tuvo que unir todos los tweets de cada uno de los autores con su id en twitter, etiquetados con su sexo y país de origen en un solo dataframe.

Para forma el conjunto de datos de pruebas se unieron los tweets de cada uno de los autores unido con su identificación de usuario de twitter.

Examinando la frecuencia de palabras utilizada por los hombres y la frecuencia de palabras utilizada por las mujeres y se realizo un df idf con respecto estos términos luego se creo una grafica que mostrara las palabras que usan con mas frecuencia los hombres y que las mujeres no utilizan tanto y lo mismo a la inversa.

Se observo que los hombres tienden hablar mas de juegos, fútbol, algo de política, y las mujeres de sentimientos, relaciones, familia.

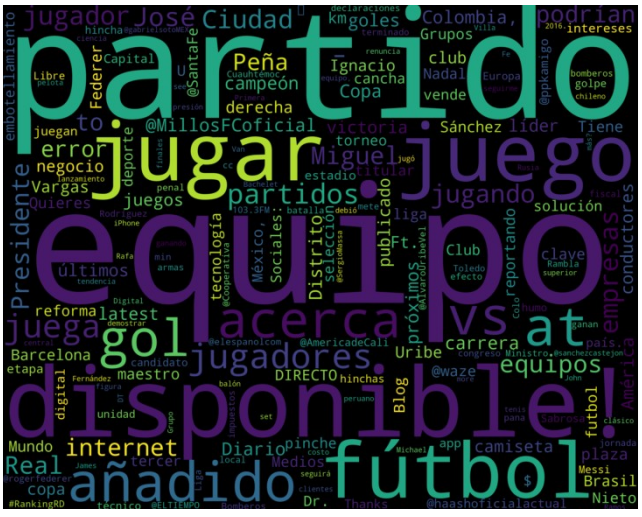

Fig1: Palabras mas usadas por hombres.

Fig 2:Palabras usadas por mujeres.

Con esto es puede hacer una bolsa de palabra que permita mejorar la precisión del modelo.

PRE-PROCESADO.

Esta es la tarea que mas tiempo consume en cuando se trata de un proyecto de datascience, aunque para el análisis de texto existe muchas librerías que nos ahorran mucho trabajo estas suelen ser mas efectivas en ingles. En nuestro caso se decidió utilizar un tokenizador sencillo que eliminara caracteres repetidos puesto que muchas de las palabras usadas en twitter tienden a repetir caracteres debido a que no hay reglas de escrituras, transformar todas las letras en minúsculas para facilitar la comparación de las palabras.

También nos centramos encontrar las palabras que aportan más información con tfidf para crear una bolsa de palabras y entrenar el modelo con ello. Es decir palabras que es más probable que sean empleadas por hombres que mujeres y viceversa.

Eliminar palabras que no aporten mucha información al modelo es muy importante, porque reduce los tiempo de entrenamiento del modelo porque se reduce la cantidad de datos y ademas contribuye a mejorar la precisión del

modelo es por ello que se decidió eliminar todas aquellas palabras.

Otro enfoque es medir la longitud de los tweets para diferenciarlo por sexo bajo la hipótesis de que los tweets de las mujeres tienden a ser mas largos que el de los hombres.

MÉTODOS.

Baseline.
Para tener un modelo de referencia para la clasificación de tweets se uso el modelo baseline el cual hace un pre-procesado de los tweets eliminando signos de puntuación, números, stopwords, convierte a minúsculas todas la letras y elimina los espacios en blancos ademas crea una bolsa de palabras con los N términos más frecuentes donde N es el numero de términos que se desee añadir. A este pre-proceso se entreno con el modelo de suport vector machine con 10 términos se obtuvo una precisión de 56% en detección género y un 21% de precisión en variedad, al aumentar el número de termino mejora la precisión hasta llegar a un 70% de precisión en detección de género y 89% de precisión en variedad.

Random Forest con bag of words.
En este método se generó un vocabulario mediante un pre-procesado de los tweets similar al de baseline, pero con la diferencia que para crear el bag of words se uso un tfidf, es decir se tomo en cuenta las palabras que eran mas frecuentes en hombres que mujeres y viceversa. El mejor resultado se obtuvo con frecuencia de palabras usadas tres veces más en un género con respecto a otro. Con una precisión del 68% este método fue aplicado solo a la detección de género.

Random Forest entrenando con la longitud media, mediana, desviación estándar y la simetría en la longitud de los tweets por usuario para detectar género y variedad. Con una precisión de 54% en la detección de genero y 18% en la detección de variedad.

Random Forest con tfidf. Usando un pre-procesado de datos donde se redujo la longitud de aquellas palabras que repetían caracteres más de tres veces, se transformó el texto a minúsculas y se eliminaron los espacios en blanco, se usaron stopwords de la librería stop_words de python y las 20000 palabras con mayor frecuencia. Creando las matrices tfidf, usadas para entrenar el modelo en el cual se obtuvo un resultado de 74% de precisión en la detección de género y un 94% de precisión en la detección de variedad.

RESULTADOS.

El mejor resultado salio de transformar los tweets de los usuario en una matriz tfidf con los 20000 términos mas frecuentes en donde se pudo hacer aplicando esto al modelo de Random Forest con 500 iteraciones este resultado de 74% de precisión para detección de genero y un 94% de precisión para la detección de variedad mejora el modelo baseline con 5000 términos el cual tiene una precisión de 70% y 89% respectivamente.

Random Forest con vectores tfidf: género

	precision	recall	f1-score
female	0.75	0.71	0.73
male	0.72	0.76	0.74
avg	0.74	0.73	0.73

Random Forest con vectores tfidf: género

	precision	recall	f1-score
argentina	0.93	0.96	0.94
chile	0.97	0.96	0.97
colombia	0.92	0.94	0.93
mexico	0.89	0.93	0.91
peru	0.98	0.88	0.93
spain	0.89	0.95	0.92
venezuela	0.97	0.93	0.95
avg	0.94	0.94	0.94

CONCLUSIÓN.

Para finalizar hemos creado un modelo que es capaz de clasificar el género y el país de origen de un usuario de twitter de habla hispana, procesando sus tweets publicados. Los resultados obtenidos fueron de gran precisión sobre nuestro conjunto de prueba. El mejor resultado se obtuvo aplicando un random forest con un tfidf. Se observo los modelos llegaron a mejorar cuando disponían de un gran numero de términos ya que los mejores resultados se obtuvo con el baseline una bolsa de palabras con los 5000 términos mas frecuentes y con el random forest con los 20000 términos mas frecuentes.