



Resumen Ejecutivo

Prueba técnica ML Engineer J & G

Aspirante: Jairo Andrés Hernández Mosquera

CLASIFICADOR DE NOTICIAS:

Descarga aquí el dataset de titulares de noticias (y sus descripciones cortas) publicadas entre 2012 y 2018 en el HuffPost y resuelve los siguientes retos. Puedes utilizar librerías y modelos pre-entrenados de acceso público (referenciar).

Retos:

1. ¿Se pueden catalogar las noticias con la descripción y los titulares? Compara tu clasificación con las categorías incluidas en el set de datos.
2. ¿Existen estilos de escritura asociados a cada categoría?
3. ¿Qué se puede decir de los autores a partir de los datos?
4. Ahora, utilizando técnicas de aprendizaje no supervisado, trata de identificar temas, “protagonistas” u otras entidades de las noticias.
5. Basándote en el texto de la descripción corta, caracteriza este dataset.
6. ¿Qué otra información útil se puede extraer de los datos?

Entregables:

1. Link de acceso al proyecto en GitHub (u otro repositorio)
2. Presentación (diapositivas) que comunique de la mejor manera posible y para cualquier público los resultados encontrados.

Solución propuesta

1. Se obtuvo la clasificación de las categorías mediante la aplicación de algoritmos supervisados, además de que este paradigma permite la comparación para la medición de rendimiento del mismo.

Para ello se exploran tres enfoques convenientes para este fin. Todo los datos pasan por la misma etapa de procesamiento, considerando en primera instancia una unión de los atributos *headline* y *short_description* en la adición de un nuevo atributo *text* [1], posteriormente una tokenización y un filtrado bajo el criterio texto inferior a 5 palabras, un label encoding de la categoría, Glove Embedding de para la representación de palabras y one hot encoding para la tranformación del label a datos interpretables por los modelos, En este punto, se configura los modelos para cada enfoque diferenciado

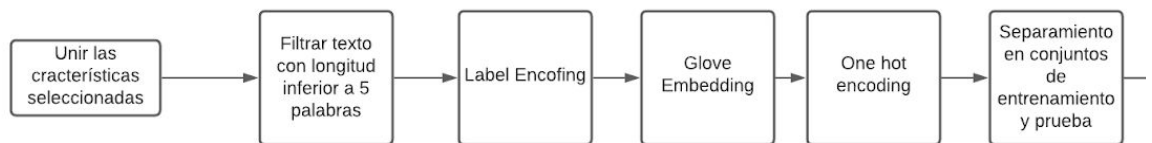


Figura 2. Etapa de procesamiento propuesta

Posteriormente, se configuran tres clasificadores plenamente diferenciados, El primero Es un Clasificador de texto simple de única capa, El segundo un claisifcador de capas LSTM bidireccional + capa convolutiva, siendo especialmente útil para encontrar la información semántica y contextual alrededor de cada token y finalmente un enfoque LSTM con capa de ateción para establecer comportamientos de ateción selectiva

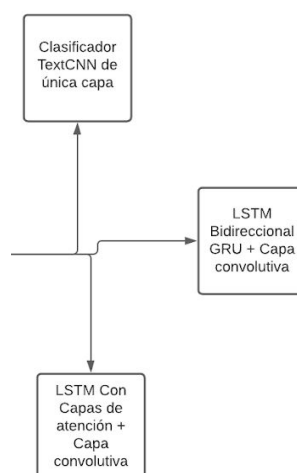


Figura 3. Modelos propuestos

Resultados Solución Propuesta (Comparación)

Estableciendo una comparación entre las categorías predichas y las reales, se tiene, los siguientes resultados:

```
local/lib/python3.6/site-packages/keras/callbacks.py:120: UserWarning: Method on_batch_end() is slow c
ompared to the batch update (0.536588). Check your callbacks.
159931/159931 [=====] - 662s 4ms/step - loss: 0.8761 - acc: 0.7236 - val_loss
: 1.1847 - val_acc: 0.6635
Epoch 19/20
159931/159931 [=====] - 653s 4ms/step - loss: 0.8661 - acc: 0.7271 - val_loss
: 1.1973 - val_acc: 0.6628
Epoch 20/20
159931/159931 [=====] - 653s 4ms/step - loss: 0.8576 - acc: 0.7297 - val_loss
: 1.2016 - val_acc: 0.6640
model TextCNN accuracy: 0.605182
model Bidirectional GRU + Conv: 0.641348
model LSTM with Attention: 0.663982
0.6729610084285822
```

Figura 1: Recorte de la verbosidad y métricas para tres enfoques establecidos

Para los tres enfoque se obtienen medidas de precisión entre la franja del 60 y 70 %, mostrando especial eficacia para el modelo LSTM con capas de atención con un 66%, hallazgos en notebooks publicados en repositorios públicos como Kaggle o GitHub muestran porcentajes de precisión no mayores al 75 %, usando modelos ligeros como DistilBERT o TF-IDF con alto grado de requisito computacional; En el estado de arte se logró una precisión del 88.72% usando una red CNN1-RNN 200 notablemente mas grande y tambien computacionalmente costosa (Experimentacion empiricas propia usando GPU y TFU de Colab sugieren enfoques usando TPU alojadas en servicios serverless).

2. Para este apartado se optó por un análisis Estilométrico del corpus, el procesamiento va hasta el el labelencoding y se aplica con FeatureExtration() para obtener represetaciones cuatitativas que permitan identificar al escritor, para ello se tomaron caracteron representaicon cuantitativas por autor como la longitud media por palabra, longitud media por oración, Entropia de Shannon y otras características.

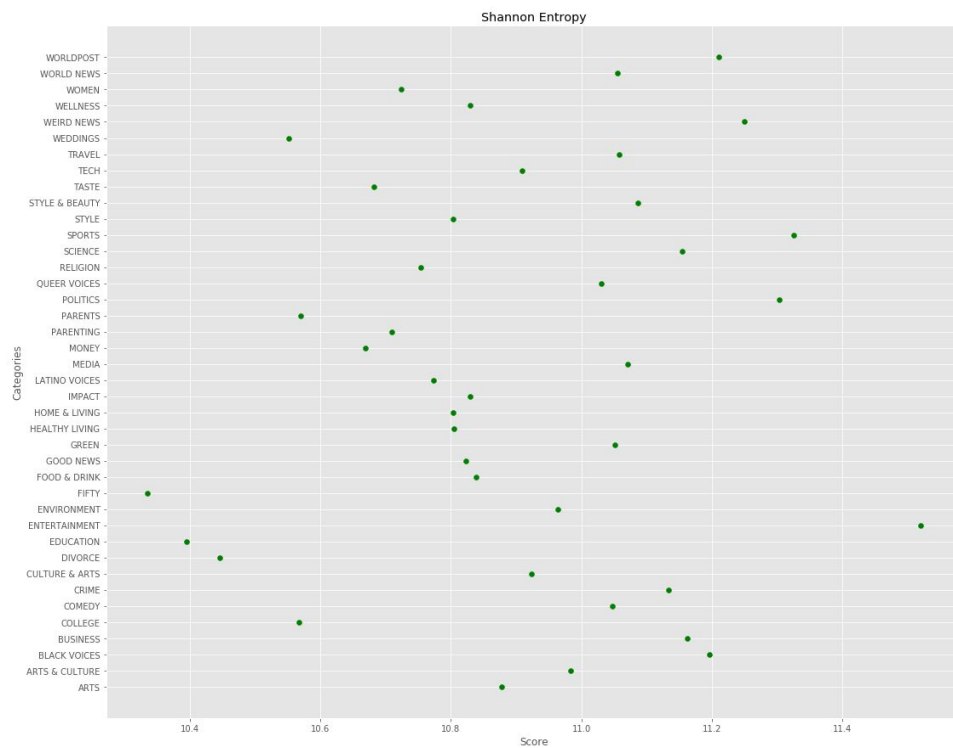


Figura 2: Esquema Shannon Entropy

En este caso a manera de ejemplo se puede establecer un alto contenido de lectura en la lengua analizada para la Categoría Entretenimiento

- Para este apartado se optó por un análisis Estilométrico del corpus, con las mismas características del apartado 2

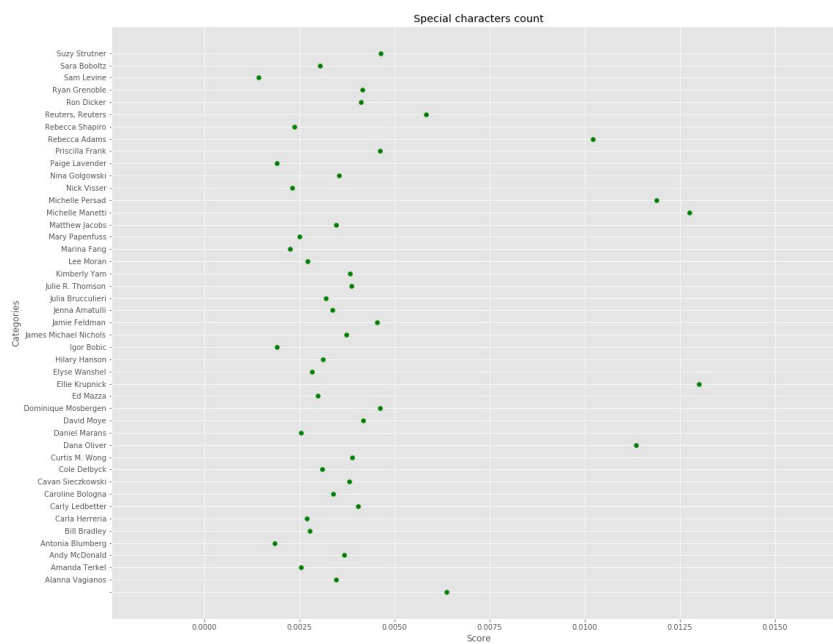


Figura 3: Esquema Special characters

La representación gráfica de estas medidas permiten establecer características propias de ciertos autores, como en el caso de analizar la cantidad de caracteres especiales permite encontrar una tendencia en el estilo de por ejemplo Michelle Persad y Michelle Manetti, de forma homóloga, se pueden establecer criterios descriptivos para agrupar y analizar la data en función de los autores, la aplicación de esta técnica en la misma demostro que La contribución promedio de las tres categorías más escritas para el top 100 de los autores con más articulos es de alrededor del 80 % configurando una alta correlación conveniente para tomar en cuenta

4. Para este apartado se empleó un enfoque NER para la extracción de algunas entidades de interés por noticia, los resultados se encuentran anexados en la carpeta correspondiente, de forma complementaria, se recurre a un Topic modeling para una serie de categorías plenamente representables

```
0 -0.38928 love person experience word story relationship share healthy words
1 -0.29589 music show night awards song hit rock credit singer
2 -0.35965 police gun death killed shooting violence dead man shot
3 -0.4754 make making ways lives important change time working find
4 -0.39932 long woman takes break leave days started face secret
5 -0.36943 time life family real friends talk death happy thanksgiving
6 -0.36792 media social event moment message today big news attention
7 -0.31224 star film movie series tv director actor hollywood talks
8 -0.44423 world problem part control stand free stop culture today
9 -0.2714 food eat eating healthy foods kitchen recipes diet taste
10 -0.31521 women school men college students sexual high education young
11 -0.30378 man divorce woman couple marriage video husband wife married
12 -0.57644 good feel people bad makes don matter hard fact
13 -0.28785 study sleep risk found shows health research heart science
14 -0.37822 black white million house people friday group online store
15 -0.29631 day summer hot morning make chocolate cake valentine ice
16 -0.27929 street air wall america flight st fire board car
17 -0.40653 state government bill law federal public report congress data
18 -0.41747 trump president donald obama house russia campaign speech white
19 -0.35426 change climate states united global deal future crisis world
20 -0.38512 found question perfect asked questions times read running dog
21 -0.26874 wedding art photos ideas color project design favorite planning
22 -0.20302 twitter check facebook huffpost post pinterest instagram style tumblr
23 -0.46571 year back day made end big coming set heart
24 -0.47143 people money don won call stop save ll big
25 -0.33161 season travel holiday christmas list summer top vacation road
26 -0.31822 photos world space visit place places natural light beautiful
27 -0.29163 gay court justice lgbt anti ban supreme decision july
28 -0.38963 kids children child parents baby mother mom son daughter
29 -0.50422 life business people live living work fear lives true
30 -0.27835 home run line front room video wanted idea hurricane
31 -0.22529 north south west china town kim green beach korea
32 -0.65425 ve make time don things work thing lot great
33 -0.22295 cancer drug treatment medical patients dr doctors doctor drugs
34 -0.35836 party clinton gop hillary election republican vote presidential campaign
35 -0.41103 years american america war national ago history security month
36 -0.34183 photos style photo red dress beauty fashion hair wear
37 -0.37386 video game news super team show night live watch
38 -0.24988 lost human weight rights loss lose san trans gender
39 -0.33211 week city york fashion days weekend times show spring
40 -0.32504 health care plan people mental tax americans pay working |
```

Figura 4: Temas agrupados resultantes del Topic modeling

5. A modo de consideraciones adicionales y conclusiones se tiene:
 - Implementar una imputación de missing values con Web Scraping puede ayudar a enriquecer el modelo
 - Hay una disminución en el grado de contenido entre el 2012 - 2018 del HuffPost
 - Implementar conjunción de categorías similares para mejorar clasificador
 - Introducir el autor como otra variable de entrada en el clasificador

