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 algortimos 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

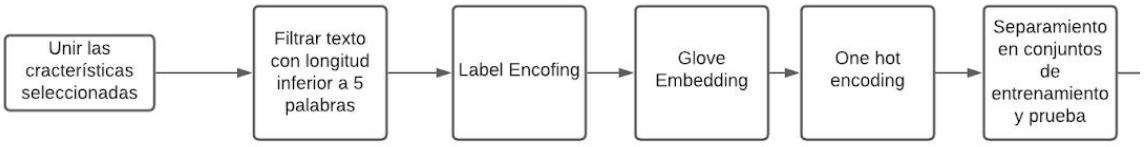

Figura 2. Etapa de procesamiento propuesta

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 bidreccional + 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

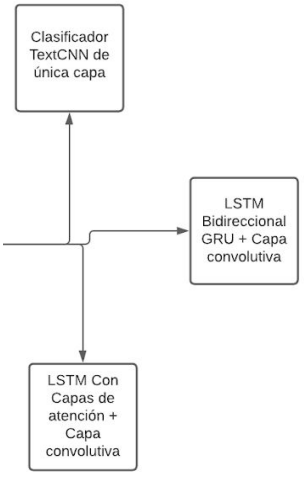

Figura 3. Modelos propuestos

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 trés 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 ateción con un 66%, hallazgos en notebooks publicados en repositorios públicos como Kaggle o GitHub muestrán porcentajes de precisón 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 empíricas 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.

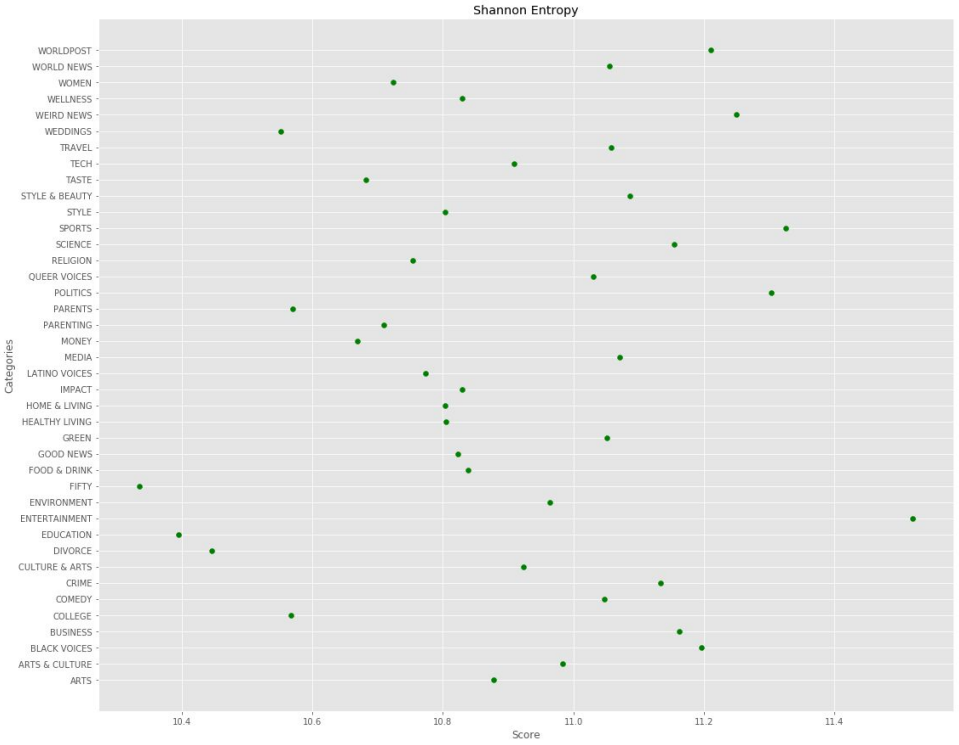

Figura 2: Esquema Shannon Entropy

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

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

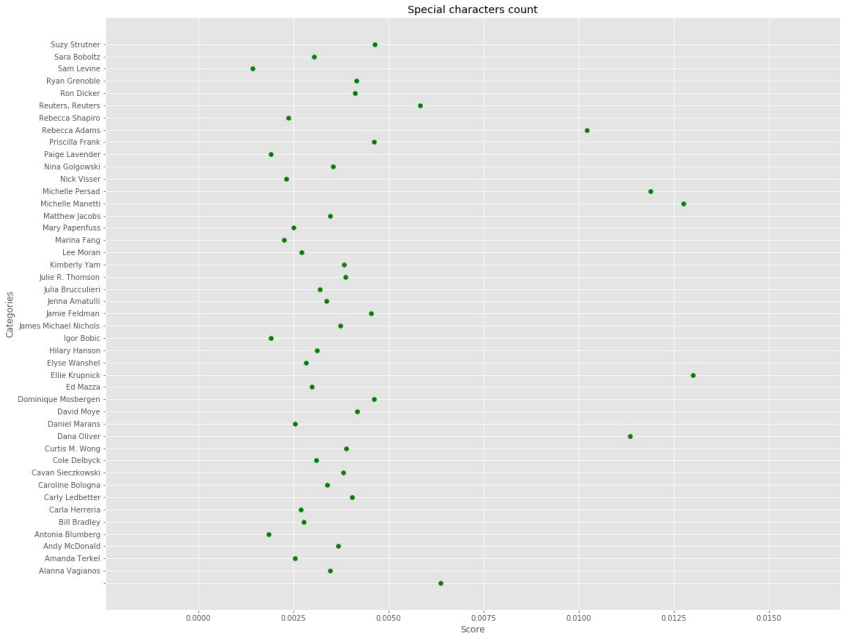

Figura 3: Esquema Special characters

La representación gráfica de estas medidas permiten establecer características propias de ciertores 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 critererios 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 categorias 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 extración 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

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

[1] Classification of News Dataset, Olga Fuks, Stanford University, ofuks@stanford.edu