Análisis de sentimientos en comentarios de Among Us

Mauricio Alvarado

4 de marzo de 2023

1. Motivación

Among Us es un videojuego multijugador en línea desarrollado por la compañía estadounidense *InnerSloth*. Fue lanzado en 2018 con poco recibimiento. No obstante, el juego alcanzó una enorme popularidad en 2020 durante la pandemia de COVID-19, convirtiéndose en uno de los juegos más jugados del año.
La historia de *Among Us* se desarrolla en una nave espacial, donde los jugadores asumen el papel de miembros de la tripulación o impostores que intentan sabotear la misión. Los jugadores trabajan juntos para mantener la nave en funcionamiento mientras completan tareas, pero algunos jugadores son impostores cuyo objetivo es sabotear y matar a los demás jugadores sin ser descubiertos. A medida que los jugadores completan tareas, también investigan los asesinatos y acusan a otros jugadores de ser los impostores. Estos deben utilizar la estrategia y la comunicación para determinar quién es el impostor y votar para expulsarlo de la nave.
Pese a que el juego aún cuenta con mucha popularidad y recibió muchas mejoras de parte de los desarrolladores, durante los primeros meses de la pandemia el juego contaba con muchos errores y caídas en sus servidores debido a la alta demanda. Esto llevó a muchas personas a calificar el juego de manera negativa. Actualmente cuenta con una calificación de 3.8 en *Google Play Store*, lo cual lo posiciona muy por debajo de muchos juegos de igual o menor popularidad. Por ello, lo que motiva a este proyecto es poder clasificar los comentarios de un juego muy popular y de mis favoritos del 2020, que actualmente ha solucionado gran parte de sus problemas, pero que aún cuenta con una calificación muy baja.
Se tomará en cuenta los cinco mil comentarios más actuales en castellano y en Perú. La última actualización es considerando los comentarios hasta la fecha de 22 de febrero del 2023.

2. Metodología

2.1. Modelos

Se consideraron un algoritmo de *machine learning* (*ML*) y dos arquitecturas neuronales de *deep learning* (*DL*). Cada uno corresponde a un modelo, por lo que se estimó 3 modelos diferentes.
El algoritmo considerado para *ML* es el de Regresión Logística. Fue elegido por ser uno de los algoritmos más usados para propuestas simples de clasificación. De hecho, es uno de los primeros algoritmos que se aprenden en diversos cursos a lo largo del mundo. Se estimó la mejor versión de este basado en la selección de los hiperpárametros que maximicen el ajuste con los datos de

entrenamiento. Se usó el método *GridSearchCV* que recoge las diferentes combinaciones de hiperparámetros definidos con anterioridad y estima el modelo con 5-Folds mediante *cross-validation*[1]. Entre las diferentes alternativas se consideró como parámetros de regularización a 0.001, 0.01, 0.1, 1, 10 y 100.
Para el caso los algoritmos de *DL*, se propuso dos arquitecturas neuronales que cuentan con capas de *embedding*, *convolucionales*, *dropout*, *pooling* y *densas*. La primera estructura será llamada "Básica" y la segunda será llamada "Intermedia". La primera se divide por las siguientes capas:

- Embedding: 15 000 dimensiones en el vocabulario
- Convolucional: 32 núcleos
- Pooling: Máximo global
- Dropout: 80 %
- Densa: 16 neuronas
- Densa: 1 neurona

Mientras que la segunda por lo siguiente:

- Embedding: 15 000 dimensiones en el vocabulario
- Convolucional: 64 núcleos
- Pooling: Máximo global
- Dropout: 30 %
- Densa: 128 neuronas
- Densa: 64 neuronas
- Densa: 1 neurona

En el entrenamiento se tomará en cuenta que todas las observaciones deben ser tokenizadas y no debe existir *stopwords*. En particular, para el modelo de Regresión Logística se aplicará una lematización al dataset previo a la vectorización vía TF-IDF. En cambio, para el caso de los modelos de redes neuronales se vectorizará directamente mediante el método *Tokenizer* de *keras*. Esto debido a que las observaciones no deben ser lematizadas en las redes neuronales si se cuenta con la capa Embedding, debido a que esta buscará relaciones entre las palabras. Este proceso puede alterarse y perder valor si se lematiza.
Como comentario final, que se retomará en la subsección 3.2, todos estos modelos se estimarán con el dataset traducido del Corpus Stanford Sentiment Treebank (SST-2). El mejor modelo será el seleccionado para clasificar los comentarios de *Among Us* en la *Google Play Store*.

[1]Consiste en separar las observaciones de entrenamiento en cinco grupos, estimar con cuatro de estos y reservar el quinto como un *testing* para estimar el ajuste. Tras realizar este ejercicio con las cinco combinaciones posibles, se obtiene cinco métricas de ajuste diferentes que serán promediadas. El objetivo de este ejercicio es corroborar la inexistencia de *overfitting*, lo cual se debería reflejar en métricas de ajuste en los datos de *test* muy similares al de los de entrenamiento.

2.2. Comparativa entre modelos

2.2.1. *Log-Loss*

Medida de la calidad de predicción de un modelo de clasificación, donde se evalúa la distancia entre las probabilidades de predicción y las etiquetas verdaderas. También se le conoce como "función de pérdida", por su traducción al castellano. Se define por lo siguiente:

$$Log - Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i * log(p) + (1 - y_i) * log(1 - p) \tag{1}$$

donde p es la probabilidad de clasificar como 1. Los valores de $Log - Loss$ varían entre 0 y ∞, donde un valor de 0 indica una predicción perfecta. En ese sentido, en la medida que un algoritmo asigne una menor función de pérdida, será preferido ante otro.

2.2.2. F1-score

Medida de combina la precisión (*precision*) y la exhaustividad (*recall*) de un modelo de clasificación. Dado que la precisión se define por:

$$precision = \frac{Verdadero\ positivo}{Verdadero\ positivo + Falso\ positivo} \tag{2}$$

y la exhaustividad por:

$$recall = \frac{Verdadero\ positivo}{Verdadero\ positivo + Falso\ negativo} \tag{3}$$

El F1-score es la media armónica de ambos:

$$F1 - score = 2\frac{precision * recall}{precision + recall} \tag{4}$$

El $F1 - score$ se sitúa entre 0 y 1, donde un valor de 1 indica una predicción perfecta.

2.2.3. Curva *Receiver Operating Characteristic* (ROC)

Métrica para evaluar la calidad de un modelo de clasificación binaria. La curva ROC representa la tasa de verdaderos positivos frente a la tasa de falsos positivos para diferentes umbrales de probabilidad. Dada la cantidad de aciertos del total de verdaderos, la sensibilidad:

$$sensibilidad = \frac{Verdadero\ positivo}{Verdadero\ positivo + Falso\ negativo} \tag{5}$$

y la cantidad de aciertos del total de falsos, la especificidad:

$$especificidad = \frac{Verdadero\ negativo}{Verdadero\ negativo + Falso\ positivo} \tag{6}$$

La curva ROC se grafica con 1-Especificidad en el eje de las abscisas y la Sensibilidad en el eje de las ordenadas. En la medida que el área debajo de la curva sea mayor, el modelo cuenta con mayor potencia para distinguir entre los verdaderos y falsos bajo diferentes umbrales. Es importante notar que a medida que se buscar acertar más los verdaderos o falsos, se pierde la potencia en alguno de los dos, por lo tanto es un trade-off para cada umbral. Aquel modelo que cuente con mayor potencia es el preferido.

3. Resultados

3.1. Datos

Dado que no se cuenta con una categoría para los comentarios del juego *Among Us* en la *Google Play Store*; es decir, que la clasificación se realizará a "ciegas" sin saber si alguno de los algoritmos ha asignado los comentarios de manera correcta o no; es necesario primero entrenarlo con datos en los que sí se cuente con esta división.
Por ello, los tres algoritmos fueron estimados con el *dataset* traducido del Corpus Stanford Sentiment Treebank (SST-2)[2]. Este Corpus está basado en el trabajo de Pang y Lee (2005), en el que se recolectó 11.855 oraciones individuales de *reviews* de películas. En Stanford se analizó y categorizó más de 215 mil frases únicas entre comentarios positivos y negativos. Para este proyecto se usará la versión traducida de dicho trabajo.
Tras haber entrenado y comparado los resultados entre los modelos, se usará este para clasificar los comentarios del juego *Among Us* de la *Google Play Store* extraídos hasta la fecha de 22 de febrero del 2023.

3.2. Selección de modelo

Tras la estimación de los modelos, se seleccionó el modelo de Regresión Logística con un parámetro de regularización de 10. Hacia adelante, este modelo será el comparado con las métricas de las dos redes neuronales. Para los tres modelos, se usó la predicción con los datos de testeo y con sus verdaderas asignaciones (etiquetas).
El Cuadro 1 muestra los resultados de las métricas de $Log - Loss$ y $F1 - score$. Sorpresivamente, el modelo de Regresión Logística es aquel que presenta el mejor valor en la función de pérdida y también la que muestra mayor ajuste mediante la métrica de $F1 - score$. Es sorpresivo debido a ser el modelo más simple entre las tres alternativas. Este resultado se podría explicar a varios factores, entre ellos al criterio diferenciado que se consideró en los modelos, donde a la regresión logística se aplicó lematización y no a las redes neuronales. No obstante, también puede deberse las redes neuronales contaron con *overfitting* durante el proceso de estimación, lo cuál pese a varios experimentos no se logró reducir. Las Figuras se pueden encontrar al final del documento.

Cuadro 1: Métricas para cada modelo

Modelo	$Log - Loss$	$F1 - score$
Reg. Logística	**7.367**	**0.800**
CNN Intermedia	9.387	0.707
CNN Básica	10.378	0.655

La Figura 1 muestra las curvas ROC para diferentes umbrales. Se encuentra un resultado consistente con las anteriores métricas, en el que la Regresión Logística nuevamente es superior a las dos propuestas de arquitecturas neuronales. En este sentido, el modelo de simple de ML, con el hiperparámetro de 10 en la regularización resulta ser el mejor no solo entre sus diferentes alternativas,

[2]Disponible en: https://huggingface.co/datasets/mrm8488/sst2-es-mt.

sino también entre los diferentes modelos propuestos para el entrenamiento-testeo. Este modelo es el seleccionado con el que se clasificará los comentarios de *Among Us*.

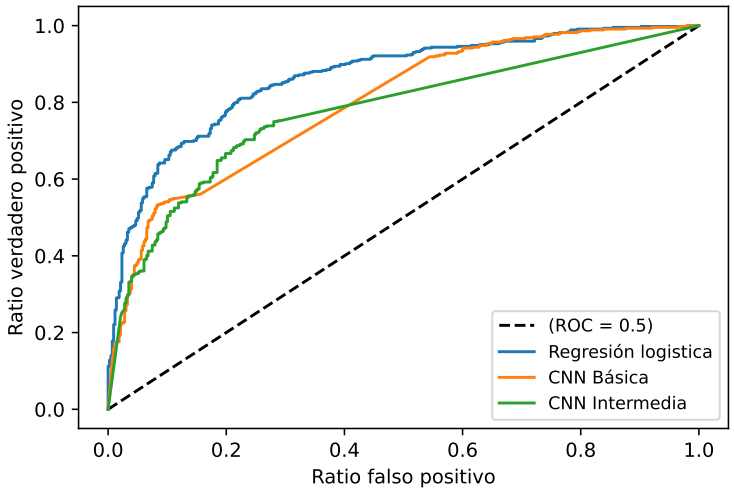

Figura 1: Curvas ROC

3.3. Hechos estilizados

En esta subsección se hará una revisión de las estadísticas de la base de datos de los comentarios de Among Us. Las principales estadísticas son las siguientes:

- Total de observaciones: 5 000
- Total de palabras: 38 227
- Total de observaciones por *score*: 1 (1 085), 2 (240), 3 (389), 4 (511), 5 (2 775)
- Total de palabras por *score*: 1 (9 472), 2 (2 701), 3 (4 808), 4 (5 737), 5 (15 509)
- Cantidad de verbos: 9 921
- Cantidad de adjetivos: 8 582

La mayor cantidad de comentarios, como es de esperarse, se encuentra en los extremos. Finalmente, no se pudo identificar correctamente cuáles son las personas u organizaciones más mencionadas en los comentarios: varias palabras que son asignadas como estas no lo son realmente. Los que efectivamente sí son palabras u organizaciones son:

- Among Us: 54
- Google: 7
- Messi: 4

3.4. Clasificación de comentarios

Con el modelo de clasificación elegido, la Regresión Logística, se procedió a predecir qué comentarios son positivos o negativos.
Como no se puede comparar el etiquetado con alguna referencia, debido a que este no existe, se comparó el etiquetado con el *score*. Es razonable esperar que un comentario positivo esté acompañado de una calificación alta, y por el contrario, un comentario negativo lo esté con una calificación baja. De esta manera, el análisis debería enfocarse principalmente en los comentarios con los *scores* que se encuentran en los extremos.
Según la Figura 2, el modelo parece ser capaz de predecir con mucha certeza los comentarios positivos. Los comentarios con *score* entre 2-4 los predice casi a la mitad de manera positiva y a la otra de manera negativa. Esto puede deberse a que estos cuenten con una fragmento del comentario que es positiva y otra que es negativa, propio de un comentario que no es calificado de manera categórica como un 1 o un 5. No obstante, para el caso extremo de negativo, el 1, el modelo no parece ser tan contundente al clasificar los comentarios como negativos. Se etiquetó a 656 comentarios como negativos y 429 como positivos, lo cual genera mucha incertidumbre.

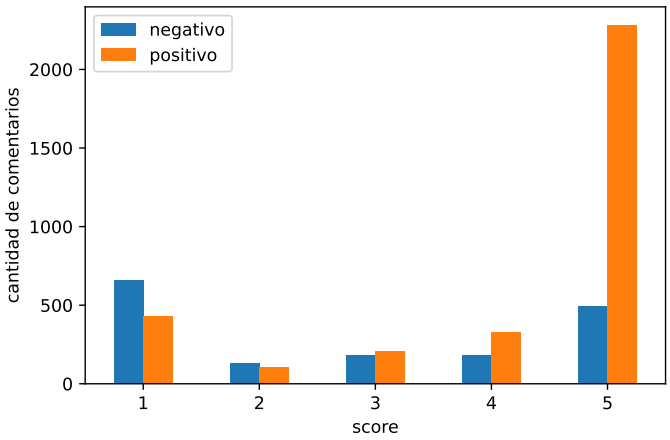

Figura 2: Distribución de clasificación dado el *score*

4. Conclusiones

A modo de conclusión se encontró que un modelo simple de Regresión Logística superó a las dos propuestas de redes neuronales en el proceso de entrenamiento con el *dataset* traducido del Corpus Stanford Sentiment Treebank (SST-2). El resultado es consistente en los tres criterios de comparación: *Log-Loss*, *F1-score* y la curva ROC.
Luego, en la clasificación de los *Among Us* de la *Google Play Store*, el modelo es capaz de clasificar los comentarios positivos de manera muy acertada. Esto tomando en consideración que estos comentarios positivos deberían ser congruentes con un *score* alto. Por el contrario, el modelo parece ser inexacto para clasificar comentarios negativos.

Queda en agenda aumentar más algoritmos de clasificación como el Decision Tree, Random Forest, XGBoost y SVM. Además, se debe mejorar las arquitecturas de redes neuronales de tal manera que se encuentre una que no genere *overfitting*. Tomando esto en consideración, se espera reducir la incertidumbre en la clasificación de comentarios muy negativos y positivos.

Referencias

[1] Pang, B., y Lee, L. (2005). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. *arXiv preprint* **cs/0506075**. doi: https://doi.org/10.3115/1219840.1219855.

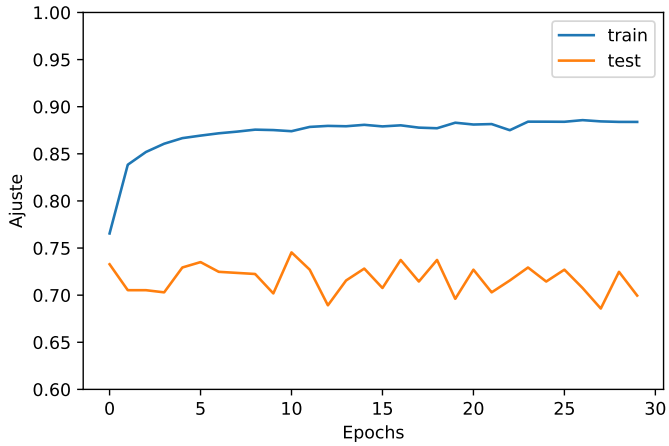

Figura 3: Ajuste en la Red Neuronal Convolucional Básica

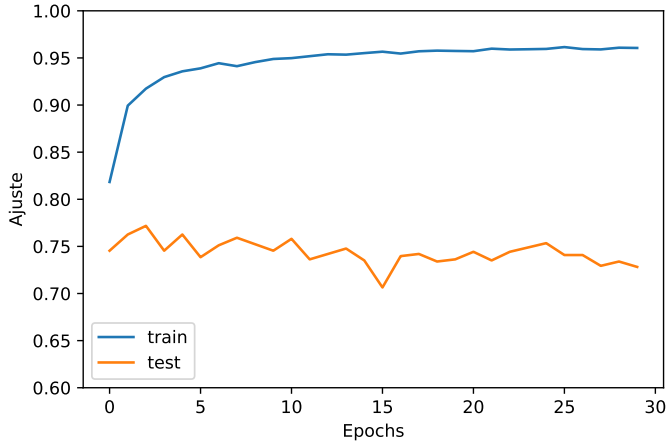

Figura 4: Ajuste en la Red Neuronal Convolucional Intermedia