Título

Análisis de sentimientos en reseñas de videojuegos con DistilBERT

Autores

Andreas Gehrts

1) Resumen

Este trabajo aplica modelos de lenguaje grandes (LLMs) al análisis de sentimientos en reseñas de videojuegos publicadas en la plataforma Steam. Se entrenó un modelo DistilBERT con fine-tuning sobre un dataset balanceado de más de 260.000 reseñas clasificadas como recomendadas o no. El modelo alcanzó una precisión del 85% y mostró resultados competitivos considerando el mínimo preprocesamiento realizado. Se discuten además ejemplos cualitativos y posibilidades de mejora.

2) Planteamiento del problema

El volumen creciente de reseñas en plataformas digitales como Steam genera la necesidad de clasificarlas automáticamente para mejorar la experiencia de usuario y la toma de decisiones de desarrolladores. El análisis de sentimientos permite identificar si una reseña refleja una experiencia positiva o negativa. En este proyecto se entrena un modelo de lenguaje preentrenado, distilbert-base-uncased, para clasificar reseñas como recomendadas o no recomendadas.

3) Corpus

El dataset fue extraído de Kaggle - Steam Reviews Dataset, compuesto por 434.891 reseñas con las siguientes columnas relevantes:

1. review: texto de la reseña
2. recommendation: etiqueta binaria ("Recommended" / "Not Recommended")
3. funny, helpful: métricas de interacción
4. hour_played: horas jugadas antes de la reseña
5. date_posted, title, is_early_access_review: metadatos

Para entrenamiento se utilizó un subconjunto balanceado:

1. Total de registros: 261.248 (130.624 positivos, 130.624 negativos)
2. Se descartaron columnas no textuales y se trabajó exclusivamente con review y recommendation.

4) Metodología

a) **Preprocesamiento**

i) Conversión de etiquetas: recommendation: 1 (Recommended) / 0 (Not Recommended)

ii) Eliminación de nulos

iii) Balanceo de clases

iv) Tokenización con DistilBertTokenizerFast, max_length=128, padding=True, truncation=True

5) Entrenamiento

i) **Modelo**: distilbert-base-uncased (Hugging Face Transformers)

ii) Fine-tuning con Trainer

iii) **Hiperparámetros:**

(1) **Epochs:** 5

(2) **Batch size:** 32

(3) **Learning rate:** 3e-5

(4) **Optimización:** AdamW

(5) **Aceleración:** uso de GPU (fp16) mediante accelerate

6) Evaluación

a) Accuracy

b) F1 score (ponderado)

c) Matriz de confusión

d) Predicciones de prueba

7) Resultados

Tras el entrenamiento con DistilBERT, se evaluó el modelo utilizando métricas estándar como **accuracy** y **F1-score**, obteniendo los siguientes resultados sobre el conjunto de validación:

a) Resultados finales:
   i) eval_loss: 0.4539
   ii) eval_accuracy: 0.8936
   iii) eval_f1: 0.8936
   iv) eval_runtime: 31.5706
   v) eval_samples_per_second: 1655.019
   vi) eval_steps_per_second: 25.878
   vii) epoch: 5.0

8) Análisis cualitativo

Ejemplos de clasificación

Texto	Etiqueta	Predicción	Obs.
"One of the worst games you can play right now...."	0	0	Acierto
"36$ so i can stare at failed connection screen?"	0	0	Acierto
"This game rocks but overrun with hackers just ..."	1	0	Ambigüedad

(0 = not recommended, 1 = recommended)

9) **Conclusiones**

a) DistilBERT logra buenos resultados con poco preprocesamiento, aprovechando su conocimiento del lenguaje general.

b) El uso de datos balanceados es crucial para evitar sesgos en la clasificación binaria.

c) El modelo es sensible a ambigüedades e ironía, comunes en el lenguaje informal de reseñas.

10) **Recomendaciones**

a) Incluir variables adicionales (helpful, hour_played) como input multimodal para mejorar el contexto.

b) Aplicar técnicas de detección de sarcasmo o modelos más grandes como BERT o RoBERTa.

c) Extender el análisis a múltiples idiomas o hacer clasificación multiclase (positivo, neutral, negativo).