STEAM:
ANALISIS Y PREDICCION DE SENTIMIENTO EN RESEÑAS DE VIDEOJUEGOS

Vallecillo, Heidel, Lopez Ballent

¿Qué es SteamLabs?

SteamLabs es un grupo dentro de Steam que busca a través de distintos experimentos internos de la plataforma, generar más visibilidad de juegos y contenido para los usuarios.

Dentro de sus esfuerzos más notables, se cuenta la iniciativa de Micro Avances, que genera “trailers” de 6 segundos de un juego

¿Qué proponemos?

Hasta el día de la fecha, todas las iniciativas de Steam Labs fueron desarrolladas únicamente con el ecosistema de Steam en mente. Por lo tanto, el objetivo de este proyecto es generar una herramienta que nos permita empezar a integrar la data obtenida en las reviews de los videojuegos con la del mundo exterior, permitiéndonos entender sí reviews u opiniones realizadas en otras plataformas como Twitter, Reddit, etc; sobre estos juegos son positivas o negativas y acoplarlas a las valoraciones que ya posee la plataforma.

¿Quiénes son nuestros consumidores?

STEAM LABS

Social Media

Usuarios de redes sociales que puedan no estar en contacto con el sistema de reviews de Steam

Non Users

Usuarios sin engagement con Steam que estén interesados en jugar algún juego y necesitan del voto popular para tomar una decisión

Usuarios

Usuarios de steam buscan una forma más resumida de apreciar si un juego es bueno o malo, sin necesidad de leer cada review.

¿Cómo lo logramos?

Generamos un modelo de Machine Learning capaz de clasificar con la mayor exactitud posible, si una review tiene un sentimiento positivo o negativo.

Para eso entrenamos el modelo con 100.000 reviews positivas y negativas utilizando conceptos de NLP básicos.

Con el modelo entrenado, el proceso se divide en 3 pasos:

1. Búsqueda de reviews realizadas por la gente en diversas plataformas (Reddit, Facebook, Twitter) sobre un determinado juego.
2. Predicción de sentimiento de cada review con nuestro modelo.
3. Valoración del juego de acuerdo a lo que usuarios de otras plataformas manifestaron.

¿Qué datos usamos?

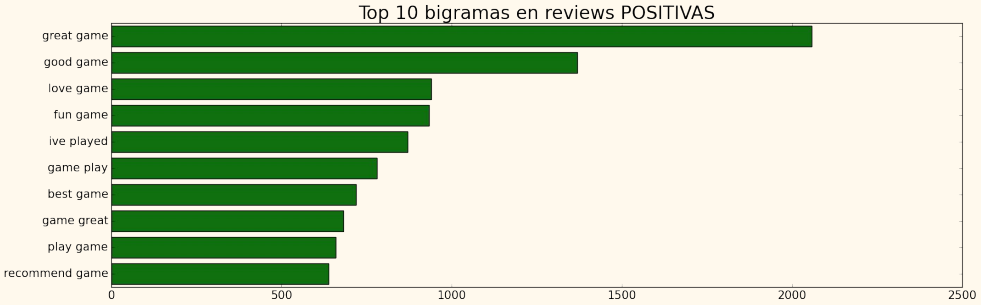
Generamos nuestro Dataset uniendo dos fuentes de datos:

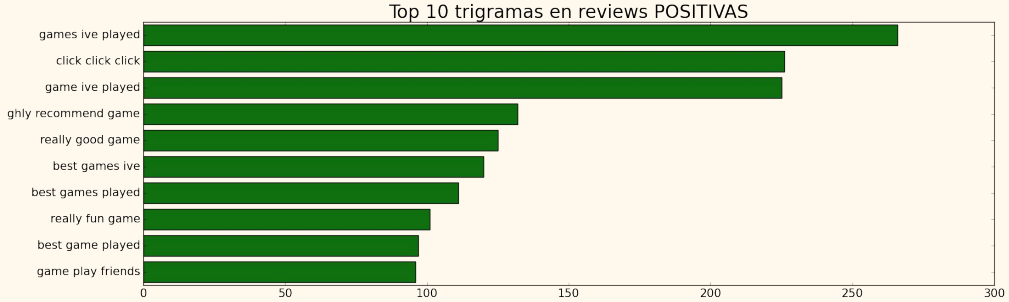
1. Steam Reviews
    - **Contenido original:** 6.4 Millones de Reviews en inglés de diferentes juegos publicados en la plataforma Steam de Valve con su respectivo sentimiento (Positiva o negativa).
2. Steam Store Games:
    - **Contenido original:** Data combinada de más de 27.000 juegos scrapeados de la plataforma Steam y de SteamSpy APIs.

Análisis Exploratorio de los Datos

Palabras y frases más utilizadas

Reviews positivas

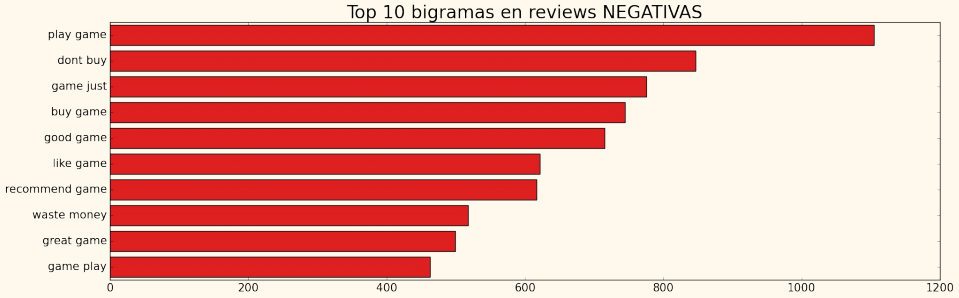


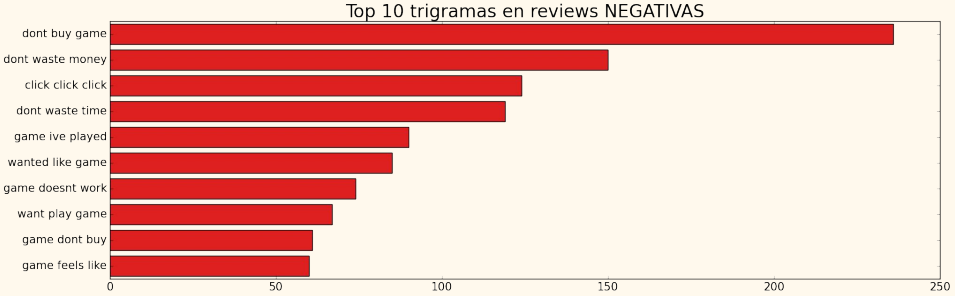


Análisis Exploratorio de los Datos

Palabras y frases más utilizadas

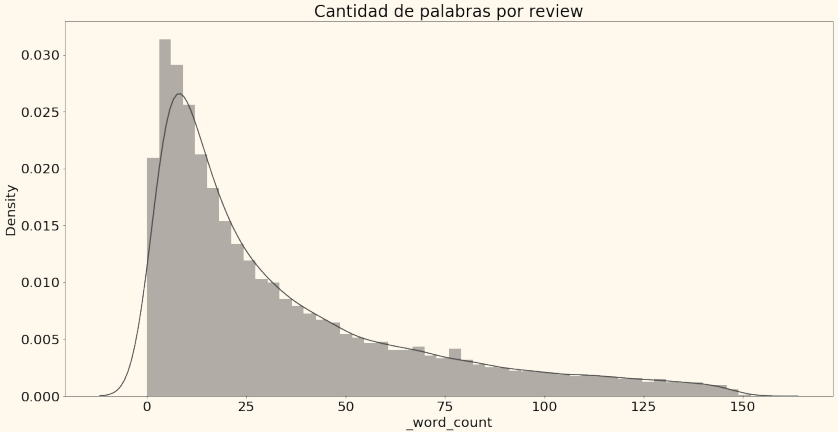
Reviews negativas

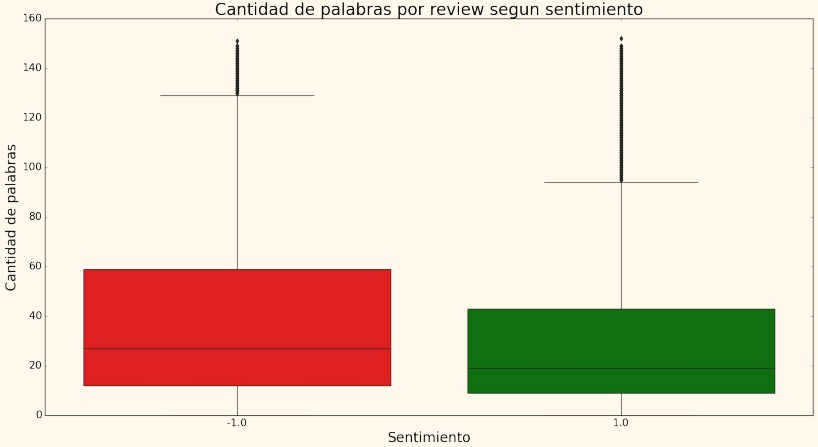




Análisis Exploratorio de los Datos

Distribución de cantidad de palabras por review





Desarrollo del modelo de ML

Precisión obtenida

Modelo	Accuracy TEST	Accuracy TRAIN
Random Forest	0.7847	0.9947
LightGBM	0.8017	0.8665
XGBoost	0.7839	0.8117

Optimización del modelo

Randomized Search CV

Este método genera un espacio de hiperparámetros de los cuales el algoritmo selecciona aleatoriamente para entrenar el modelo.

Se especificaron 10 iteraciones del modelo para generar el espacio de hiperparametros para que el modelo entrene y valide internamente a través de un algoritmo de Cross Validation interno.

Se optimizaron los modelos de XGBoost y LightGBM, sin optimizar Random Forest debido al tiempo de procesamiento requerido y la falta de recursos.

MODELOS ANALIZADOS

- Random Forest
- LightGBM
- XGBoost
- LightGBM optimizado
- XGBoost optimizado

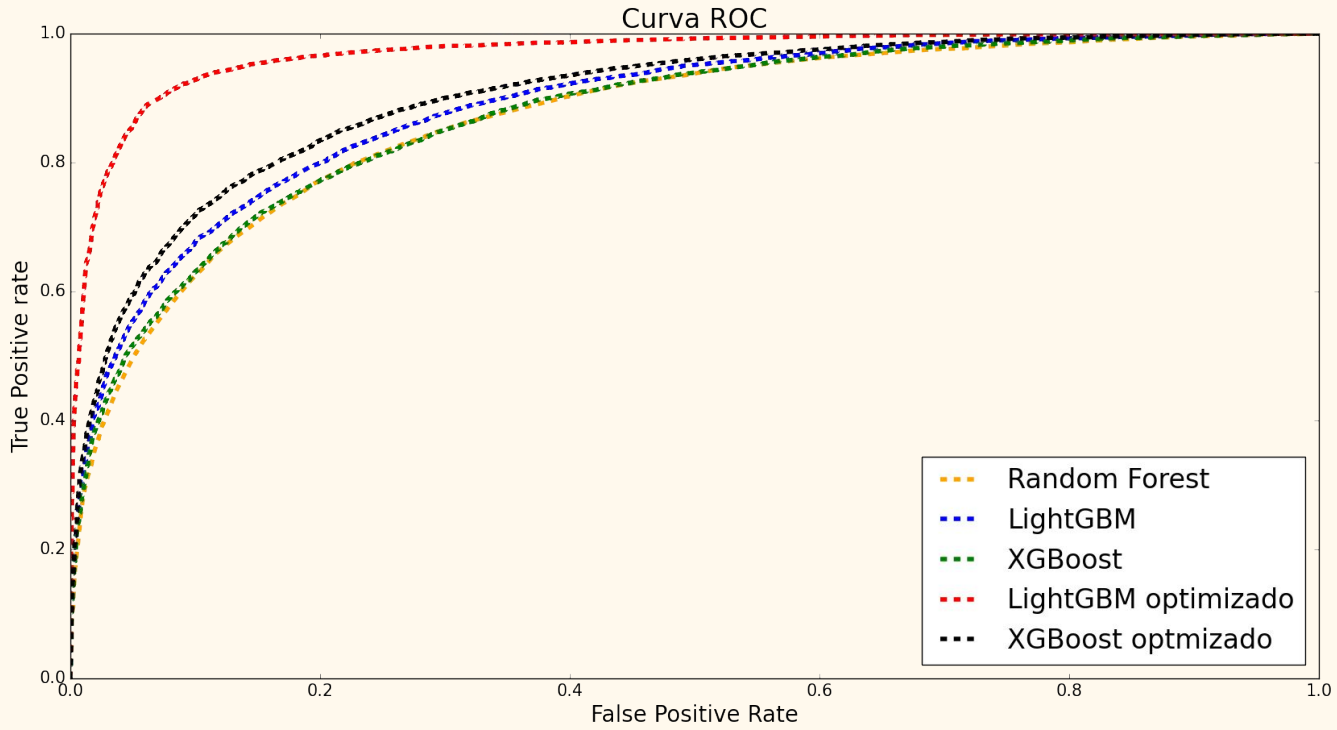
Comparación de modelos

Comparación de métricas de los 5 modelos

Modelo	Accuracy TEST	Accuracy TRAIN	Precisión	Recall
Random Forest	0.784734	**0.994774**	0.783862	0.828479
LightGBM	0.801742	0.866506	0.810151	0.824957
XGBoost	0.783965	0.811769	0.790003	0.815344
LightGBM Opt	**0.915574**	0.916983	**0.919651**	**0.923853**
XGBoost Opt	0.818443	0.878164	0.826609	0.838568

Comparación de modelos

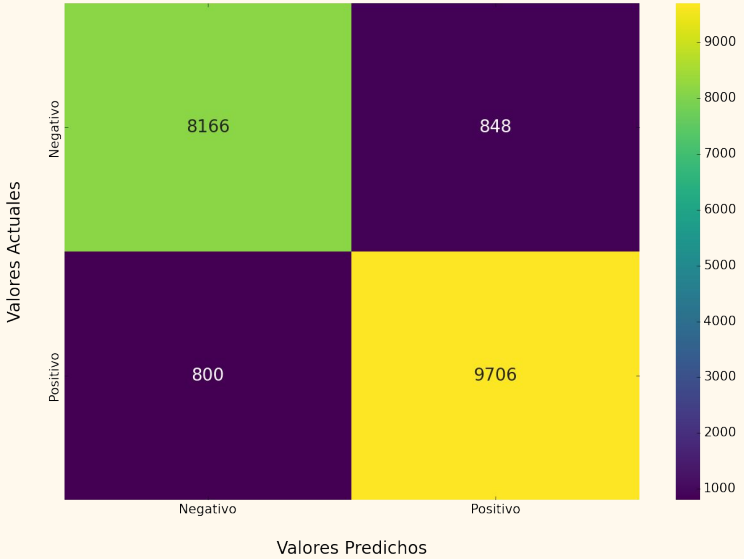
Curva ROC



Evaluación del modelo LightGBM Opt

Matriz de confusión

La matriz de confusión nos permite visualizar la performance general del modelo. Nos permite separar los resultados en 4 categorías de valor de verdad y determinar la efectividad del modelo desde estas categorías.



→ Verdaderos positivos: 9706 (49,7%)

→ Verdaderos negativos: 8166 (41,8%)

→ Falsos positivos: 848 (4,3%)

→ Falsos negativos: 800 (4,1%)

Testing del modelo

Resultados de las predicciones realizadas por el modelo obtenido

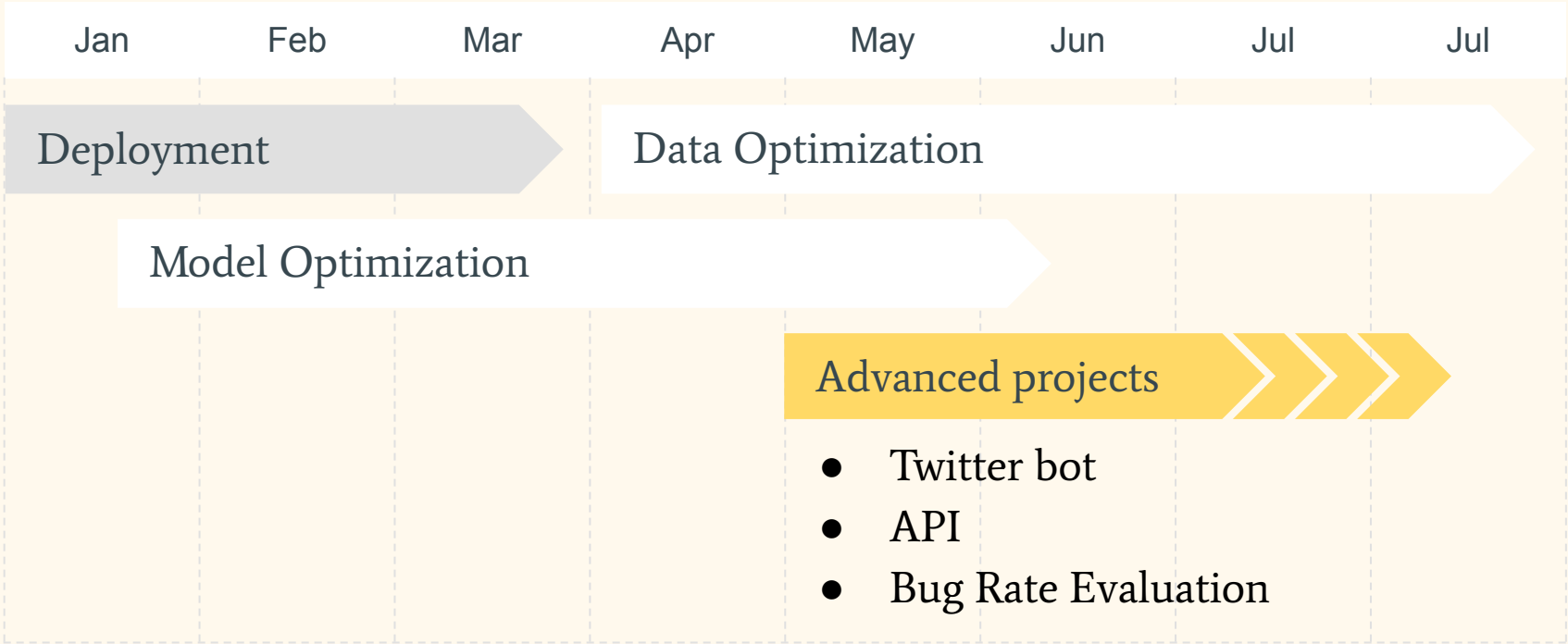
still feels like it has a lot of bugs
Score: -1
Probabilidad Negativa: 0.867944674681083
Probabilidad Positiva: 0.13205532531891698

not worth your time its one battle worthless voice acting no real plot one of the most pathetic things ive ever seen
Score: -1
Probabilidad Negativa: 0.9067791851970665
Probabilidad Positiva: 0.09322081480293354

hilariously glitchy and just plain fun
Score: 1
Probabilidad Negativa: 0.3844595676605228
Probabilidad Positiva: 0.6155404323394772

this is awesome but sometimes we got server problem but at least this greatest game i ever played
Score: 1
Probabilidad Negativa: 0.32067260462911584
Probabilidad Positiva: 0.6793273953708842

Futuras iniciativas



Equipo

Ignacio Vallecillo

Francisco Lopez Ballent

Gerardo Heidel