Entrega 1
APRENDIZAJE AUTOMÁTICO

Detección Automática de Bots en Plataformas de Streaming en Vivo

INTRODUCCIÓN

Últimamente las plataformas de streaming en vivo están experimentando un crecimiento exponencial en los últimos años, siendo los principales medios de entretenimiento digital. Kick, una plataforma emergente lanzada en 2022, ha ganado popularidad rápidamente al posicionarse como una alternativa a Twitch, ofreciendo mejores condiciones económicas para los creadores de contenido.
Sin embargo, este crecimiento acelerado ha traído consigo desafíos significativos relacionados con la autenticidad de la audiencia y la presencia de bots automatizados.

Siendo los bots una amenaza para las plataformas de streaming, que no solo afectan a la experiencia de los usuarios reales, sino también la integridad económica de los streamers. Los bots pueden inflar artificialmente métricas de audiencia, pueden hacer spam masivo, ejecutar ataques coordinados de acoso o manipular el engagement para atraer inversiones publicitarias fraudulentas.
Este proyecto aborda el desarrollo de un sistema de clasificación automática capaz de distinguir entre usuarios humanos reales y bots automatizados en plataformas de streaming, utilizando únicamente patrones de comportamiento observables en el chat.

CONTEXTO DEL PROBLEMA

El streaming en vivo es una industria valorada en más de $10 mil millones a nivel global, con millones de creadores de contenido que dependen de métricas auténticas para generar ingresos a través de publicidad, patrocinios y suscripciones. La presencia de bots distorsiona estas métricas, creando varios problemas:

Problemas Económicos:

- **Fraude publicitario**: Los anunciantes pagan por impresiones que son vistas por bots, no por usuarios reales.
- **Competencia desleal**: Streamers que usan bots artificialmente inflan su audiencia, obteniendo ventajas sobre competidores legítimos.
- **Pérdida de confianza**: Las marcas se muestran reticentes a invertir en plataformas donde no pueden verificar la autenticidad de la audiencia.

Problemas Técnicos:

- **Sobrecarga de servidores**: Los bots consumen recursos de red y procesamiento.
- **Dificultad de moderación**: Los moderadores humanos no pueden identificar y bloquear bots a la velocidad que se crean.
- **Evolución constante**: Los bots se actualizan continuamente para evadir sistemas de detección.

Relevancia del Problema

La detección automática de bots es crucial por varias razones:

1. **Escalabilidad**: La moderación manual es impráctica en canales con miles de mensajes por minuto
2. **Velocidad**: Los bots pueden causar daño significativo en segundos; la detección debe ser casi instantánea
3. **Adaptabilidad**: Los sistemas de ML pueden aprender nuevos patrones a medida que los bots evolucionan
4. **Precisión**: Los algoritmos pueden identificar patrones sutiles que los humanos podrían pasar por alto

Plataforma Kick como Caso de Estudio

Kick presenta características únicas que hacen relevante su estudio:

- **Plataforma emergente**: Menos sistemas anti-bot establecidos que Twitch
- **Crecimiento rápido**: Mayor vulnerabilidad a ataques automatizados
- **Comunidad activa**: Alta interacción en chats, facilitando la recolección de datos
- **Datos accesibles**: API pública permite la investigación académica
- **Originalidad**: Pocos estudios académicos sobre detección de bots en Kick específicamente

OBJETIVOS

Objetivo General

Desarrollar un modelo de aprendizaje automático capaz de clasificar usuarios como bots automatizados o usuarios humanos reales, basándose en características extraíbles de sus patrones de comportamiento en el chat, con el fin de facilitar sistemas de moderación automática y mejorar la integridad de las métricas de audiencia.

Objetivos Específicos

1. **Recolectar y construir un dataset etiquetado** de usuarios de Kick que incluya patrones de comportamiento característicos de bots y humanos, con al menos 1,000 instancias balanceadas entre ambas clases.

2. **Identificar y extraer características discriminativas** que permitan distinguir efectivamente entre usuarios automatizados y humanos, incluyendo frecuencia de mensajes, patrones de repetición, uso de URLs y características lingüísticas.

Criterios para etiquetar como BOT (is_bot=1):

- Nombres de usuario genéricos o secuenciales (user####, bot####, test#####)
- Frecuencia de mensajes superior a 30 por hora de forma sostenida
- Ratio de repetición superior a 0.6 (más del 60% de mensajes idénticos)
- Ratio de URLs superior a 0.5 (más del 50% de mensajes contienen enlaces)
- Presencia de enlaces acortados sospechosos (bit.ly, tinyurl, etc.)
- Comportamiento mecánico (mensajes a intervalos exactos)
- Participación simultánea con mensajes idénticos en múltiples canales

Criterios para etiquetar como HUMANO (is_bot=0):

- Nombres de usuario personalizados y únicos
- Conversaciones contextuales relacionadas con eventos del stream
- Frecuencia de mensajes dentro de rangos humanos (2-30 por hora)
- Mensajes variados en contenido, longitud y estilo
- Uso natural del lenguaje (errores de tipeo, jerga, emojis)
- Interacción bidireccional con otros usuarios
- Respuestas a preguntas del streamer o de la comunidad

DEFINICIÓN DEL PROBLEMA

Tipo de Problema

Este proyecto constituye un **problema de clasificación binaria supervisada**.

Justificación:

- **Variable objetivo**: Categórica con dos clases (bot=1, humano=0)
- **Aprendizaje supervisado**: Disponemos de instancias etiquetadas para entrenamiento
- **Clasificación binaria**: Sólo dos clases mutuamente excluyentes (un usuario es bot O humano, no ambos)
- **Predicción discreta**: El modelo predice una categoría, no un valor numérico continuo

Hipótesis del Modelo

Existen patrones de comportamiento distintivos y medibles en la forma en que los bots interactúan en el chat, que permiten clasificarlos automáticamente con un alto grado de precisión (>85%).

Hipótesis secundarias:

- Los bots tendrán frecuencias de mensajes significativamente mayores que los humanos
- Los bots presentarán ratios de repetición más altos debido a mensajes automatizados
- La presencia de URLs, especialmente links acortados, será más común en bots
- Los nombres de usuario genéricos con comportamiento automatizado

MODELOS DE APRENDIZAJE AUTOMÁTICO A UTILIZAR

Basándonos en la naturaleza del problema (clasificación binaria con variables mixtas numéricas y categóricas), se evaluarán los siguientes algoritmos:

1. Regresión Logística

Justificación:

- Modelo baseline interpretable para clasificación binaria
- Proporciona probabilidades de pertenencia a cada clase
- Coeficientes permiten entender la importancia relativa de cada característica
- Computacionalmente eficiente

Ventajas esperadas:

- Rápido entrenamiento e inferencia
- No requiere muchos datos
- Resultados fácilmente interpretables para stakeholders

Posibles limitaciones:

- Asume relaciones lineales entre features y log-odds
- Puede no capturar interacciones complejas entre variables

2. K-Nearest Neighbors (k-NN)

Justificación:

- No asume distribución específica de los datos
- Efectivo cuando existen "clusters" de comportamiento similar
- Bots similares probablemente exhiben patrones parecidos

Ventajas esperadas:

- Captura patrones locales de comportamiento
- No requiere fase de entrenamiento explícita
- Adaptable a nuevos patrones de bots

Posibles limitaciones:

- Sensible a la escala de las variables (requiere normalización)
- Computacionalmente costoso en predicción con datasets grandes
- Requiere selección cuidadosa de k

3. Árboles de Decisión

Justificación:

- Altamente interpretables (reglas if-then)
- Manejan naturalmente variables categóricas y numéricas
- No requieren normalización de datos
- Capturan interacciones entre variables automáticamente

Ventajas esperadas:

- Fácil visualización y explicación del modelo
- Identificación clara de reglas de clasificación
- Útil para entender qué características son más discriminatorias

Posibles limitaciones:

- Propensos a overfitting sin poda adecuada
- Pueden ser inestables ante pequeños cambios en datos
- Sesgo hacia variables con más valores únicos