Manipulación de información y detección de bots en redes sociales

...

Tesis de Licenciatura en
Ciencias de la Computación

Rodríguez Fernández, Marcelo Patricio

Director: Diego C. Martínez

Introducción

Introducción

En el mundo virtual, predominan las redes sociales.

Plataformas como Facebook, Twitter, Instagram, YouTube son los principales medios para muchas actividades, como:

- publicidad
- comunicaciones personales
- transmisiones de noticias
- anuncios políticos
- defensa de causas sociales

Introducción

Un ejemplo destacable sobre la influencia de las redes sociales, es la elección de Trump en el año 2016.

Se destaca como una de las razones del éxito de Trump el papel importante que cumplieron las redes sociales, especialmente Twitter y Facebook.

Introducción

Elecciones Trump 2016 - Twitter

- Billón de tweets (Estados Unidos)
- Principal herramienta de comunicación de Trump (influencia sobre votantes).

Elecciones Trump 2016 - Facebook

La campaña de Trump la utilizó con tres objetivos principales:

- recaudar fondos a través de pequeñas donaciones
- difundir mensajes a públicos prioritarios
- diseminar noticias

El aspecto más controvertido fue el uso de esta red para la difusión de noticias falsas a favor de Trump o en contra de Clinton, las cuales fueron muy difundidas por Facebook.

Introducción

Tres noticias falsas que se diseminaron por Facebook durante la campaña de Trump:

Introducción

FBI Agent Suspected in Hillary Email Leaks Found Dead in Apparent Murder-Suicide

An FBI agent believed to be responsible for the latest email leaks was found dead in an apparent murder-suicide early Saturday morning, according to police.

DENVERGUARDIAN.COM

Introducción

THE POLITICAL INSIDER

Subscribe Login

WikiLeaks CONFIRMS Hillary Sold Weapons to ISIS... Then Drops Another BOMBSHELL!

Kosar
Featured Contributor

Julian Assange : Wikileaks Have The Email That Proves Hillary Sold Weapons to ISIS In Syria

WikiLeaks announcing that Hillary Clinton and her State Department were actively arming Islamic jihadists, which includes the ISIS in Syria.

Clinton has repeatedly denied these claims, including during multiple statements while under oath in front of the United States Senate.

WikiLeaks is about to prove Hillary Clinton deserves to be arrested.

Introducción

El manejo de las redes sociales puede ser impulsado por bots, un bot es una cuenta que produce contenido automatizado, se plantean algunas preguntas:

- ¿Somos víctimas de información errónea manipulada?
- ¿Qué pasa si las redes sociales se llenan de bots?
- ¿Se puede desarrollar una detección efectiva de bots antes de que sea demasiado tarde?

El objetivo de esta tesis es detectar bots automáticamente

Marco teórico

Marco teórico

¿Qué es un bot?

Es un algoritmo computacional que produce contenido automatizado e interactúa con humanos en las redes sociales, tratando de emular el comportamiento humano.

Marco teórico

Propósitos de los bots

Intenciones benignas o útiles

Por ejemplo:

Bots de servicio al cliente, donde un bot puede ayudar a encontrar información sobre pedidos o reservas de viajes de manera automática. Esto es muy útil y eficiente para las empresas, más aún en tiempos de distanciamiento social.

Marco teórico - Propósitos de los bots

Intenciones maliciosas: Bots que engañan, explotan y manipulan las redes sociales con rumores, spam, malware, información errónea o mentiras.

Ejemplos:

- Bots que pueden inflar el apoyo a un candidato político durante elecciones.
- Bots que pueden influir en la estabilidad de los mercados ya sea con un efecto positivo o negativo.
  - ⚠ Bots que aumentan la “charla” en las redes sociales sobre alguna compañía.
  - ⚠ En 2013, un ejército de bots publicaron un rumor falso de un ataque terrorista en la Casa Blanca, lo que provocó una caída en sus bolsas.
- Bots que explotan el delito cibernético logrando captar información privada en redes sociales.

Marco teórico

Evolución de los bots

Bots en su comienzo

Solo publicaban contenido automáticamente y eran fáciles de detectar.

Ejemplo: En 2011, un equipo de la Universidad de Texas, implementó una idea simple: crearon bots de Twitter que publicaban tweets sin sentido y observaron los seguidores captados (bots).

Marco teórico

Evolución de los bots

Bots actualmente

Ahora, el comportamiento de los bots es más difuso. Los bots pueden:

- sumarse a conversaciones
- comentar publicaciones
- responder preguntas
- adquirir visibilidad infiltrándose en discusiones populares

Marco teórico - Problemática y efecto bot

Los bots con intenciones maliciosas pueden causar problemas, pueden:

- dar la falsa impresión de que alguna información, independientemente su precisión, es muy popular y respaldada por muchos.
- alterar la percepción de la influencia en redes sociales
  - ampliando artificialmente la audiencia de algunas personas
  - pueden arruinar la reputación de una empresa con fines comerciales o políticos.
- un peligro potencial es que los bots manipulen la percepción de la realidad influyendo en el ánimo.

Marco teórico

¿Podemos confiar en la información de las redes sociales?

En algunos casos, los usuarios eligen comprar su base de fans, una estrategia que forma parte de lo que se llama fraude en las redes sociales.

El fraude en las redes sociales es el proceso de crear likes, seguidores, vistas o cualquier otra acción en redes como Facebook, Twitter, YouTube e Instagram para aumentar artificialmente la base de amigos/seguidores de una cuenta.

Marco teórico - ¿Podemos confiar en la información de las redes sociales?

Los servicios de fraude se basan en una de dos estrategias para obtener el control sobre las cuentas que utilizan para aumentar artificialmente la popularidad:

- Comprometer cuentas reales existentes
- Crear cuentas nuevas y falsas
  - granjas de clicks
  - botnets

Uno de los mayores desafíos que enfrentan los prestadores de servicio de fraude en las redes sociales es evadir las barreras de registro planteadas por las redes sociales.

Marco teórico

El precio del fraude en las redes sociales

Se puede acceder a cientos de sitios web especializados en ofrecer estos servicios fraudulentos mediante la búsqueda de "comprar likes" o "comprar seguidores" en los motores de búsqueda populares.

Algunos servicios son más caros que otros, por ejemplo, comprar un comentario es más costoso que comprar un like. Algunas redes sociales y el tamaño del servicio (número de me gusta) también afectan los precios.

Marco teórico - El precio del fraude en las redes sociales

	$USD / 1.000 seguidores	$USD / 1.000 likes
Facebook	$29	$20
Instagram	$13	$14
Twitter	$12	$15
YouTube	$51	$50

Marco teórico

El precio del fraude en las redes sociales

Los servicios de fraude merecen atención por parte de la comunidad investigadora por principalmente por dos razones:

- ★ Costos importantes para las empresas de redes sociales
- ★ Crea datos sesgados y aumenta artificialmente la popularidad de ciertas cuentas
  - ⇒ Desinformación
  - ⇒ Disminución de confianza en las redes sociales

Marco teórico - Manipulación y fake news

La manipulación en las redes sociales toma muchas formas, una de ellas son las noticias falsas o *fake news*.

El término *fake news* hace alusión a piezas de información que se difunden intencionalmente, mientras se sabe que son falsas.

Marco teórico - Manipulación y fake news

Algunos ejemplos que se pueden mencionar:

- Las manipulaciones de las redes sociales atribuidas a Rusia
- Cambridge Analytica

Marco teórico - Manipulación y fake news

Las manipulaciones de las redes sociales atribuidas a Rusia

Hay sospechas muy fuertes sobre la participación de los servicios rusos en la conducción de la campaña presidencial de Estados Unidos de 2016, que resultó en la elección de Donald Trump.

Una acción sospechosa pasa por la distribución de anuncios en las redes sociales, y especialmente en Facebook a través de memes. El Partido Demócrata hizo pública una breve lista de tales anuncios, y sugiere que los anuncios fueron pagados por “empresas cercanas al Kremlin”.

También aparecieron rastros menos convencionales, como una lista de cuentas de Twitter, baneadas pero etiquetadas como cuentas falsas manejadas desde Rusia.

Marco teórico - Manipulación y fake news

Las manipulaciones de las redes sociales atribuidas a Rusia

Marco teórico - Manipulación y fake news

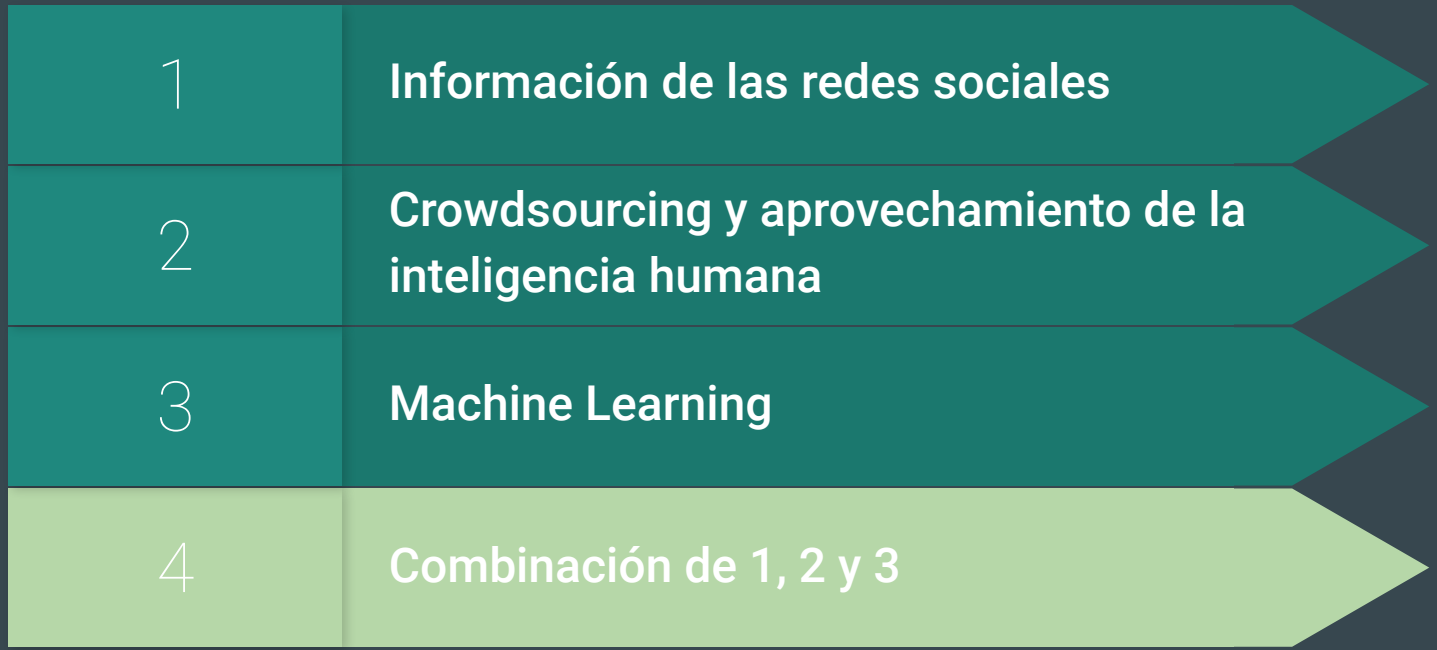
Cambridge Analytica

La compañía Cambridge Analytica afirmó (sin pruebas) haber inclinado la votación para "Leave" en el referéndum para Brexit, en junio de 2016. Entre sus herramientas aparece la distribución masiva de contenido en las redes sociales para imponer su narrativa al público.

Marco teórico - Manipulación y fake news

Cambridge Analytica

Marco teórico
Estado del arte: Sistemas de detección de bots



Marco teórico
Estado del arte: Sistemas de detección de bots

1 **Información de las redes sociales**

Detección de bots basada en grafos

Una entidad puede controlar múltiples bots para hacerse pasar por diferentes bots y lanzar un ataque (ataque sybil). Algunas estrategias para detectar cuentas sybil se basan en examinar la estructura de un grafo social. SibylRank, por ejemplo, explota una característica para identificar grupos densamente interconectados, y emplea el paradigma de inocentes por asociación.

Marco teórico
Estado del arte: Sistemas de detección de bots

2 Crowdsourcing y aprovechamiento de la inteligencia humana

Se exploró la posibilidad de detección humana, sugiriendo el crowdsourcing de detección de bots a grupos de trabajadores. Se probó la eficacia de los humanos para detectar cuentas de bots simplemente a partir de la información en sus perfiles, y se observó que la tasa de detección por humanos disminuye con el tiempo, pero es una muy buena técnica para un protocolo de votación por mayoría.

Tres inconvenientes de este enfoque:

- No es rentable en redes con una gran base de datos preexistente de usuarios.
- El costo de mantener trabajadores “expertos” para detectar con precisión las cuentas falsas, dado que el trabajador “promedio” no se desempeña bien individualmente.
- Plantea problemas de privacidad.

Marco teórico
Estado del arte: Sistemas de detección de bots

3 **Machine Learning**

Detección de bots basada en características

La ventaja de centrarse en los patrones de comportamiento es que pueden codificarse en características y adaptarse a técnicas de *machine learning*. Esto permite clasificar las cuentas según su comportamiento observado, y comúnmente se utilizan diferentes clases de características para capturar el comportamiento de los usuarios:

- Red
- Usuario
- Amigos
- Timing
- Contenido
- Sentimiento

Marco teórico
Estado del arte: Sistemas de detección de bots

3 **Machine Learning**

Ejemplo:
Botometer entra en esta categoría de sistema de detección de bots, fue lanzado en 2014 y proporciona el servicio para la detección de bots en Twitter, emplea algoritmos de aprendizaje supervisados entrenados con ejemplos de comportamientos tanto humanos como de bots.
Este sistema llega a una precisión de detección superior al 95%.

Marco teórico
Estado del arte: Sistemas de detección de bots

4 **Combinación de 1, 2 y 3**

Hay una necesidad de adoptar técnicas para detectar efectivamente los ataques sybil en las redes sociales.
Ejemplo:
Renren Sybil, un sistema que explora múltiples dimensiones de los comportamientos de los usuarios. Por ejemplo, examina los datos del flujo de clicks. Al identificar también características altamente predictivas, como la frecuencia de invitaciones, las solicitudes salientes aceptadas y el coeficiente de agrupación de red, Renren puede clasificar las cuentas en perfiles prototípicos de bots o humanos.
Las cuentas Sybil en Renren tienden a trabajar juntas para difundir contenido similar. El enfoque Renren combina ideas de los tres enfoques mencionados, y se logran buenos resultados incluso teniendo en cuenta sólo los últimos 100 clicks para cada usuario.

Introducción a Machine Learning

Introducción a Machine Learning

"[Machine Learning es el] campo de estudio que brinda a las computadoras la capacidad de aprender sin ser programadas explícitamente."

- Arthur Samuel, 1959

Introducción a Machine Learning

"Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna tarea T y alguna medida de performance P, si su desempeño en T, medido por P, mejora con la experiencia E."

- Tom Mitchell, 1997

Introducción a Machine Learning

¿Por qué utilizar machine learning?

El aprendizaje automático es ideal para:

- Problemas para los cuales las soluciones existentes requieren muchos ajustes manuales o largas listas de reglas: un algoritmo de aprendizaje automático a menudo puede simplificar el código y funcionar mejor.
  - Ejemplo: Un filtro de spam basado en técnicas de aprendizaje automático aprende automáticamente qué palabras y frases son buenos predictores de spam.
- Problemas complejos para los que no existe una buena solución utilizando un enfoque tradicional
  - Ejemplo: Reconocimiento de voz
- Obtener conocimientos sobre problemas complejos y grandes cantidades de datos.

Introducción a Machine Learning
Tipos de sistemas de aprendizaje automático

Hay diferentes tipos de sistemas de aprendizaje automático que se pueden clasificar en categorías según:

- Si están entrenados o no con supervisión humana
  - supervisado
  - no supervisado
  - semisupervisado
  - reforzado
- Si pueden aprender o no de forma incremental sobre la marcha
  - en línea
  - por lotes
- ...

Introducción a Machine Learning
Tipos de sistemas de aprendizaje automático

- ...
- Ya sea que funcionen simplemente comparando nuevos puntos de datos con puntos de datos conocidos o, en su lugar, detecten patrones en los datos de entrenamiento y creen un modelo predictivo
  - basado en instancias
  - basado en modelos

Introducción a Machine Learning

Aprendizaje supervisado

Los datos de entrenamiento que alimentan al algoritmo incluyen las soluciones deseadas, llamadas etiquetas.

Introducción a Machine Learning
Aprendizaje supervisado

Una tarea típica de aprendizaje supervisado es la *clasificación*.

Introducción a Machine Learning
Aprendizaje supervisado

Otra tarea típica es predecir un valor numérico como objetivo, como el precio de un auto, dado un conjunto de características (kilometraje, antigüedad, marca, etc.) llamadas predictores. Este tipo de tarea se llama *regresión*. Para entrenar el sistema, debe darle muchos ejemplos de automóviles, incluyendo tanto sus predictores como sus etiquetas (es decir, sus precios).

Introducción a Machine Learning

Aprendizaje no supervisado

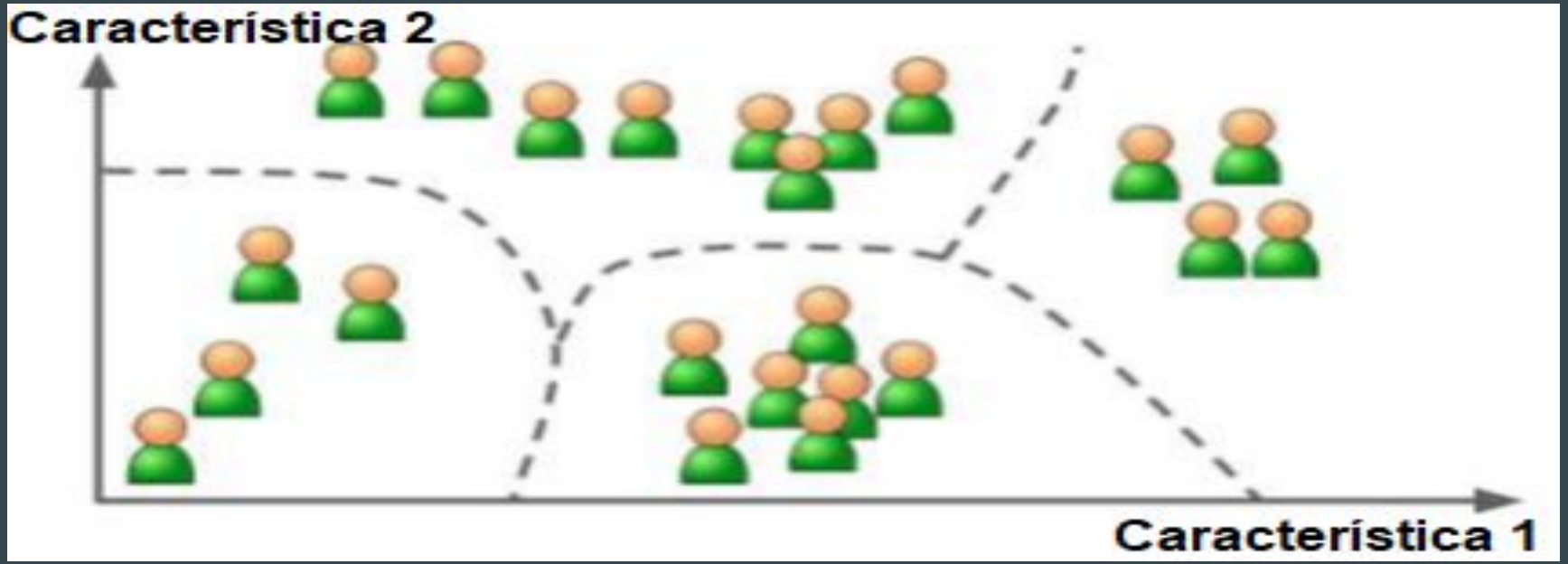
En el aprendizaje no supervisado, los datos de entrenamiento no están etiquetados.

Los algoritmos de aprendizaje no supervisados más importantes entran en alguna de estas categorías:

- Agrupamiento (clustering)
- Detección de anomalías
- Visualización y reducción de dimensionalidad
- Aprendizaje de reglas de asociación

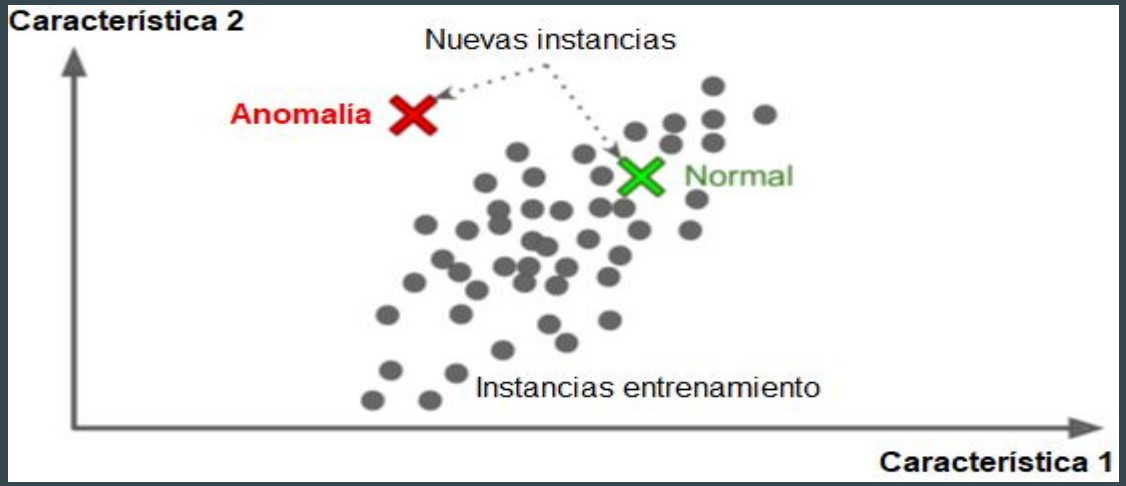
Introducción a Machine Learning
Aprendizaje no supervisado - Agrupación

Consiste en la agrupación automática de datos.



Introducción a Machine Learning
Aprendizaje no supervisado - Detección de anomalías

Los algoritmos de detección de anomalías buscan detectar valores anormales, por ejemplo, detectar transacciones inusuales de tarjetas de crédito para evitar fraudes, detectar defectos de fabricación o eliminar automáticamente los valores atípicos de un conjunto de datos antes de enviarlos a otro algoritmo de aprendizaje.



Introducción a Machine Learning
Aprendizaje no supervisado - Visualización

Los algoritmos de visualización también son buenos ejemplos de algoritmos de aprendizaje no supervisados: se alimenta con una gran cantidad de datos complejos y sin etiquetar, y generan una representación 2D o 3D de los datos.

Introducción a Machine Learning
Aprendizaje no supervisado - Reducción de dimensionalidad

El objetivo es simplificar los datos sin perder demasiada información.

Una forma de hacer esto es fusionar varias características relacionadas en una, a esto se le llama extracción de características. Por ejemplo, desgaste de auto (kilometraje, antigüedad).

Introducción a Machine Learning
Aprendizaje no supervisado - Reglas de asociación

El objetivo del aprendizaje de reglas de asociación es profundizar en grandes cantidades de datos y descubrir relaciones interesantes entre los atributos o características.

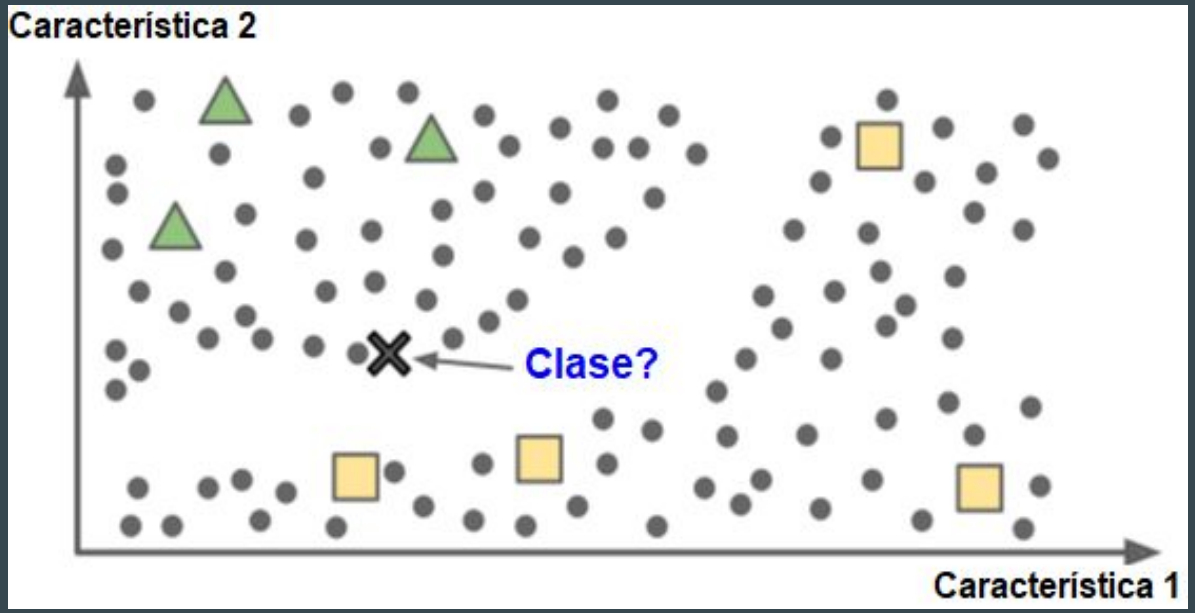
Por ejemplo, en un supermercado la ejecución de una regla de asociación en sus registros de ventas puede revelar que las personas que compran jamón y queso también tienden a comprar mayonesa, por lo tanto, es posible que desee colocar esos productos cerca unos de otros.

Introducción a Machine Learning

Aprendizaje semisupervisado

Algunos algoritmos pueden tratar con datos de entrenamiento parcialmente etiquetados, generalmente muchos datos sin etiquetar y un poco de datos etiquetados, esto se llama aprendizaje semisupervisado.

Introducción a Machine Learning
Aprendizaje semisupervisado

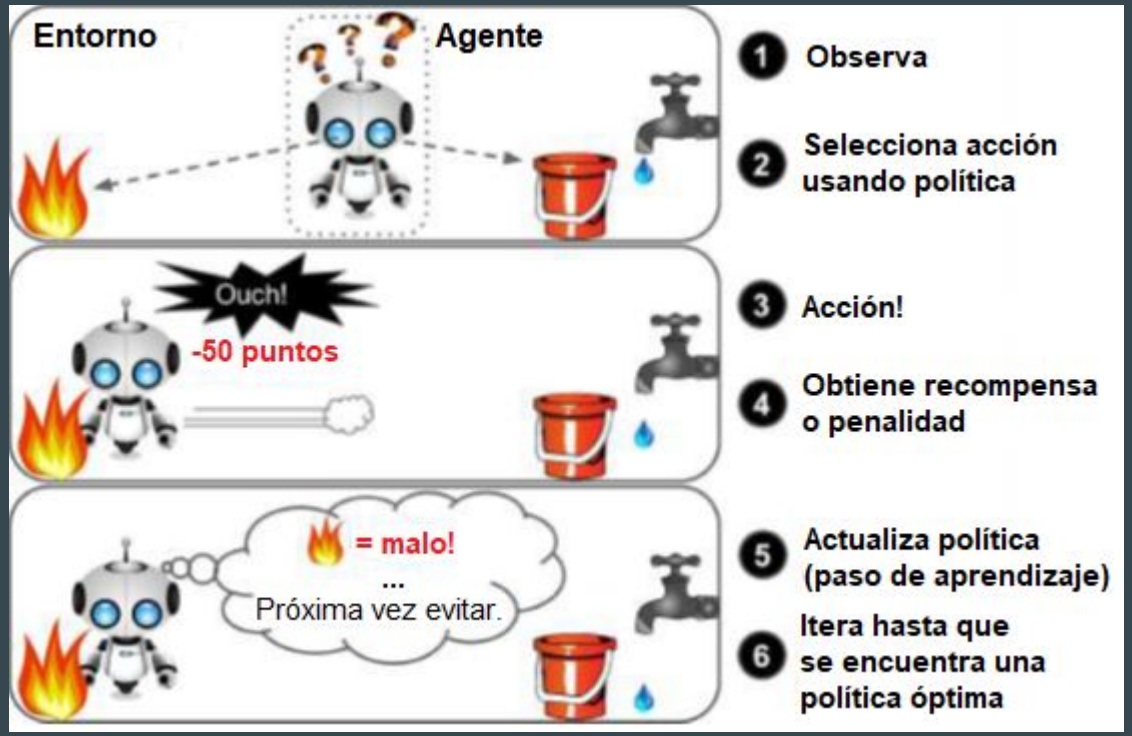


Introducción a Machine Learning

Aprendizaje reforzado

En el aprendizaje reforzado, el sistema de aprendizaje, llamado agente en este contexto, puede observar el entorno, seleccionar y realizar acciones y obtener recompensas a cambio (o penalizaciones en forma de recompensas negativas). Luego, debe aprender por sí mismo cuál es la mejor estrategia, llamada política, para obtener la mayor recompensa a lo largo del tiempo.

Introducción a Machine Learning
Aprendizaje reforzado



Introducción a Machine Learning

Aprendizaje por lotes

En el aprendizaje por lotes, el sistema es incapaz de aprender de forma incremental: debe entrenarse utilizando todos los datos disponibles.

Introducción a Machine Learning

Aprendizaje en línea

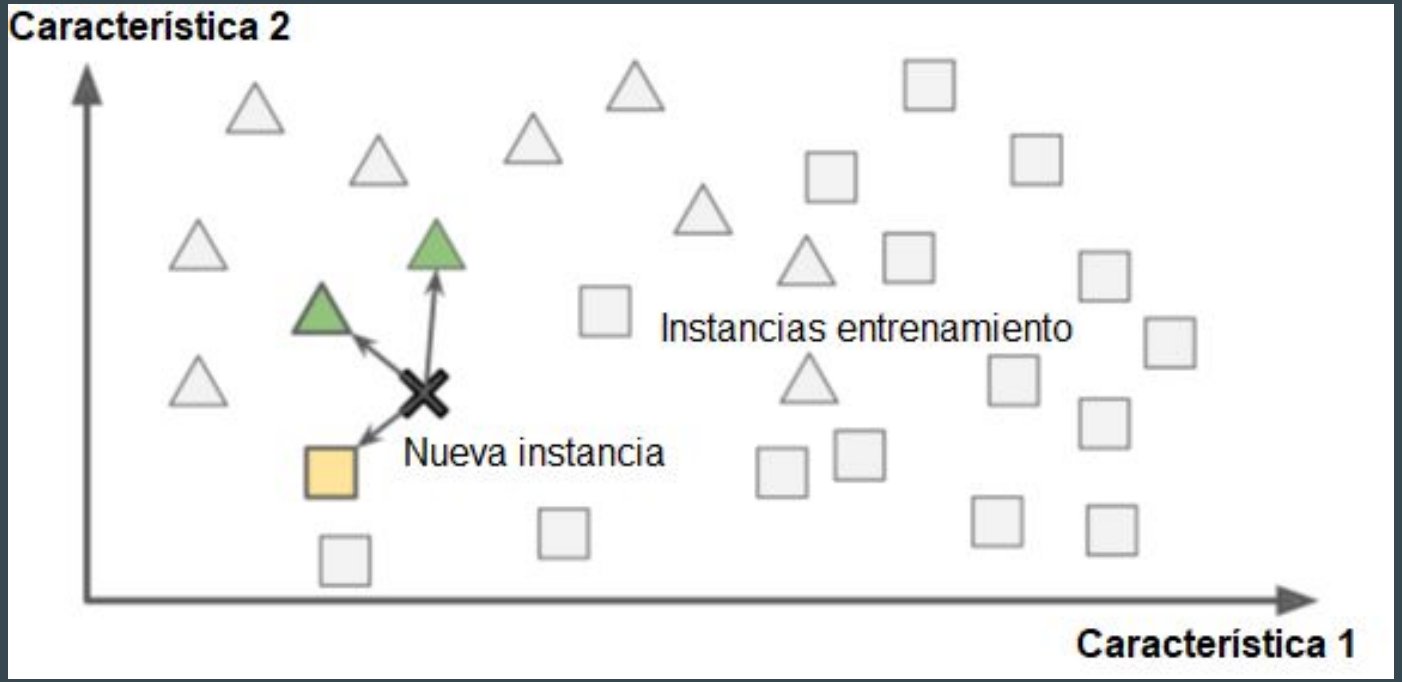
En el aprendizaje en línea, se entrena al sistema de manera incremental al alimentarlo con instancias de datos de manera secuencial, ya sea individualmente o por grupos pequeños llamados mini lotes. Cada paso de aprendizaje es rápido y económico, por lo que el sistema puede aprender sobre nuevos datos sobre la marcha, a medida que llegan.

Introducción a Machine Learning

Aprendizaje basado en instancias

En el aprendizaje basado en instancias, el sistema “memoriza” los ejemplos del conjunto de entrenamiento y su clase, luego para generalizar, lo realiza en base a una medida de similitud de la instancia a clasificar con los ejemplos memorizados.

Introducción a Machine Learning
Aprendizaje basado en instancias

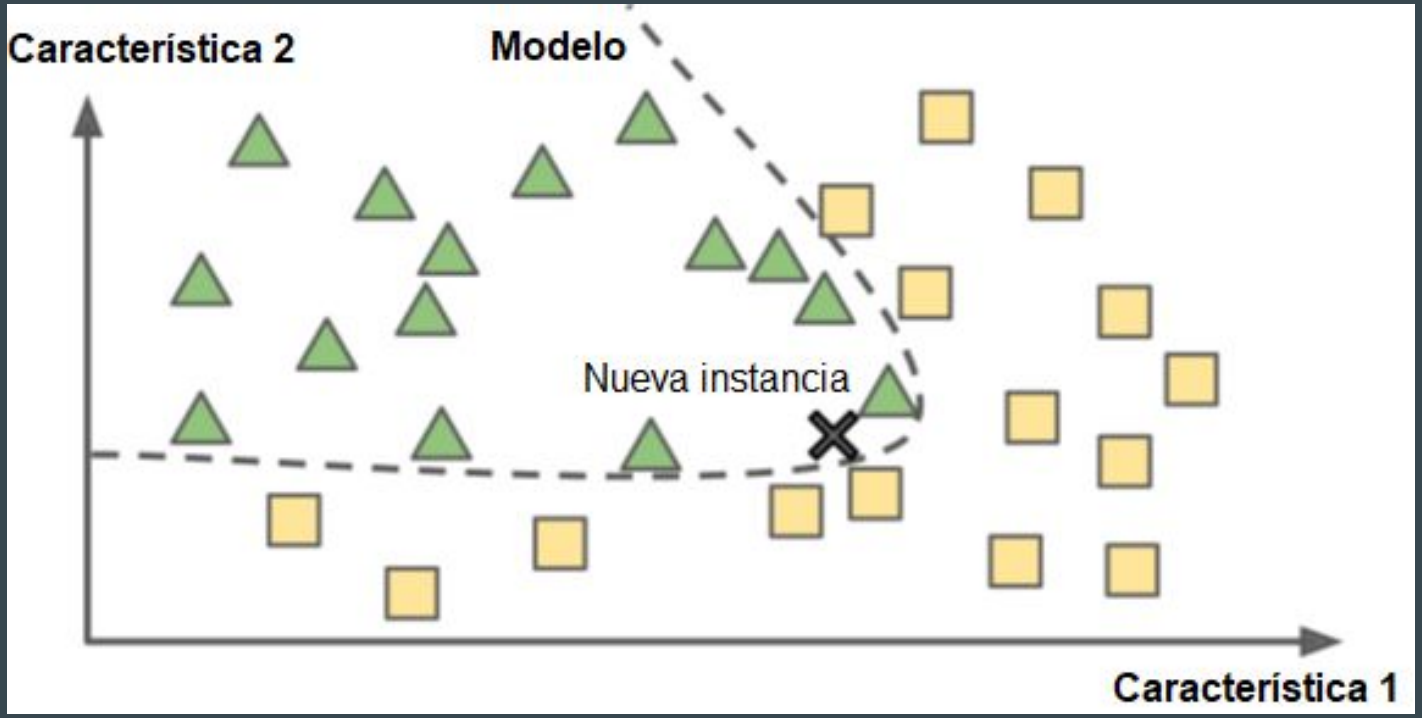


Introducción a Machine Learning

Aprendizaje basado en modelos

Otra forma de generalizar a partir de un conjunto de ejemplos es construir un modelo de estos ejemplos y luego usar ese modelo para hacer predicciones, esto se llama aprendizaje basado en modelos.

Introducción a Machine Learning
Aprendizaje basado en modelos



Introducción a Machine Learning
Proceso del aprendizaje automático

Introducción a Machine Learning
Proceso del aprendizaje automático

01 Recolectar los datos

Recolectar los datos consiste en juntar, agrupar o conseguir los diferentes datos a utilizar, podemos recolectar los datos desde diversas fuentes, por ejemplo, datos de un sitio web, utilizando una API o desde una base de datos. Podemos también utilizar otros dispositivos que recolectan los datos por nosotros; o utilizar datos que son de dominio público. El número de opciones que tenemos para recolectar datos es sumamente diverso. Esta parte del proceso parece obvia, pero es uno de los más complicados y conlleva mucho tiempo.

Introducción a Machine Learning
Proceso del aprendizaje automático

02 Prepocesar los datos

Preprocesar los datos, consiste en que una vez que tenemos los datos, los datos obtenidos tengan el formato correcto para alimentar el algoritmo de aprendizaje. Es prácticamente inevitable tener que realizar varias tareas de preprocesamiento antes de utilizar los datos.

Introducción a Machine Learning
Proceso del aprendizaje automático

03 Entrenar el algoritmo

Finalizados los pasos anteriores, podemos realizar un preanálisis para corregir los casos de valores faltantes o intentar encontrar a simple vista algún patrón en los mismos que nos facilite la construcción del modelo. En este punto podemos detectar valores atípicos que debamos descartar; o encontrar las características que poseen mayor influencia para realizar una predicción.

Entrenar el algoritmo requiere utilizar las técnicas de aprendizaje automático, en esta etapa alimentamos o introducimos al o los algoritmos de aprendizaje los datos que hemos procesado en las etapas anteriores. Los algoritmos deben ser capaces de extraer información útil de los datos preprocesados, para luego realizar predicciones de forma eficiente.

Introducción a Machine Learning
Proceso del aprendizaje automático

04 Evaluar el algoritmo

Evaluar el algoritmo consiste en poner a prueba la información o conocimiento que el algoritmo obtuvo del entrenamiento del paso anterior. Evaluamos qué tan preciso es el algoritmo en sus predicciones y si no se obtiene el rendimiento esperado, podemos volver a la etapa anterior y continuar entrenando el algoritmo cambiando algunos parámetros hasta lograr un rendimiento aceptable.

Introducción a Machine Learning
Proceso del aprendizaje automático

05 Utilizar el modelo

Utilizar el modelo consiste en poner a prueba el modelo seleccionado, utilizando los nuevos datos, con el fin de etiquetarlos de forma correcta. Se evalúa el rendimiento del modelo y en caso de no obtener el resultado esperado se regresa a revisar todos los pasos anteriores, hasta obtener buenos resultados.

Marco metodológico

Marco metodológico

Ya se introdujo el concepto de bot y una introducción al Machine Learning, decidí por desarrollar una aplicación web que sea capaz de detectar si una cuenta pública de Twitter es un bot o no, utilizando técnicas de Machine Learning.

Los usuarios reciben muchos tweets en los que algunos de ellos son de bots. La detección de bots es necesaria para identificar a los usuarios falsos y proteger a los usuarios genuinos de información errónea y de intenciones maliciosas.

Marco metodológico

El bot de Twitter es un software que envía tweets automáticamente a los usuarios. Las intenciones maliciosas de los bots de Twitter son:

1. Difundir rumores y noticias falsas.
2. Difamar a alguna persona o institución.
3. Se crean comunicaciones falsas para robar credenciales.
4. Dirigir a usuarios genuinos hacia sitios web falsos.
5. Cambiar pensamientos sobre un individuo o grupo, influyendo en la popularidad.

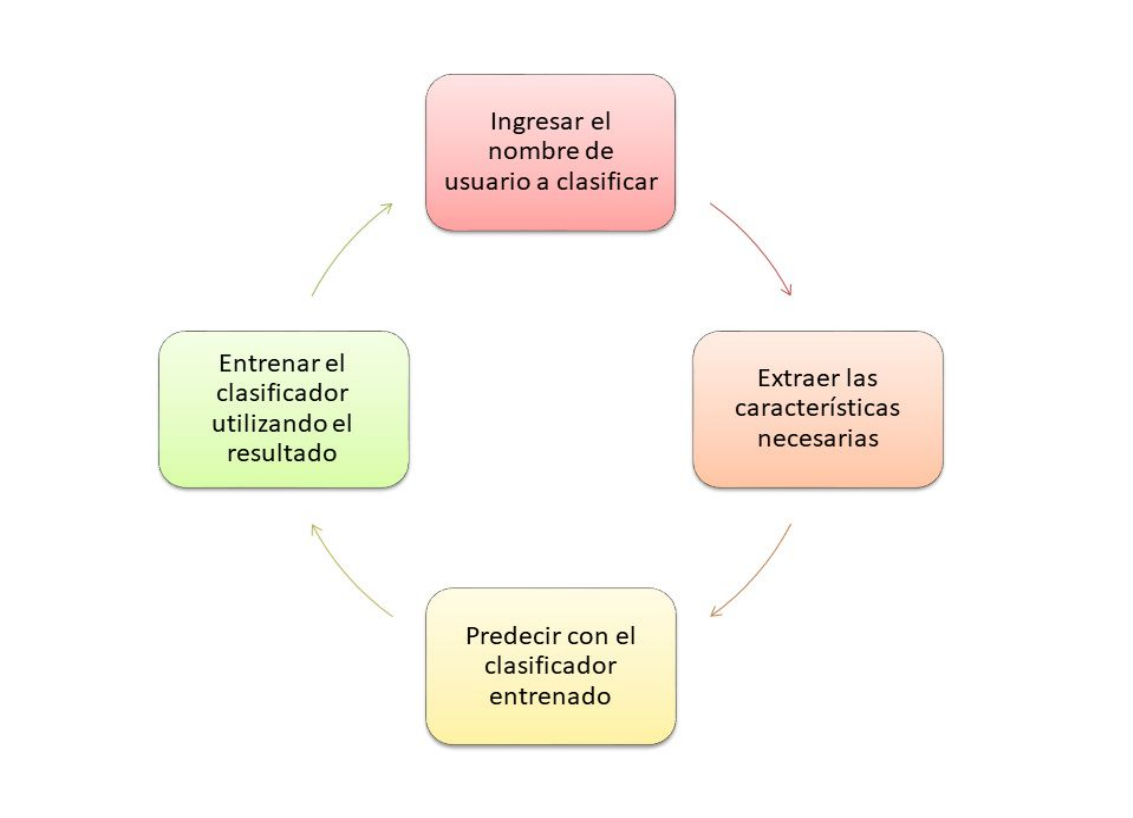
Marco metodológico

Según Twitter, la manipulación de su plataforma ocurre cuando visualizan alguna de estas conductas:

1. El uso malicioso de la automatización para perjudicar e interrumpir la conversación pública, como tratar de conseguir que algo se convierta en tendencia.
2. La amplificación artificial de las conversaciones en Twitter, incluso mediante la creación de cuentas múltiples.
3. Generar, solicitar o comprar falso compromiso.
4. Tuitear, participar en la conversación o “seguir” de manera masiva o agresiva.
5. El uso de hashtags en forma de spam, incluyendo el uso de hashtags no relacionados en un tweet.

El detector bot implementado es un sistema de aprendizaje automático supervisado, basado en modelos y en línea.

Marco metodológico - Descripción del modelo propuesto



Marco metodológico

Datos

Los algoritmos de aprendizaje automático supervisados requieren un dataset de características con una etiqueta que clasifique cada fila o resultado. Estas características pueden ser los atributos que se obtienen a través de APIs que describen una pieza de información sobre una cuenta de una plataforma de red social, como el número de amigos.

Marco metodológico

Datos

Construí un dataset llamado *Humanos*, obteniendo 75.000 tweets provenientes de 2.500 usuarios verificados que cuentan con ciertas características. Para obtener 2.500 usuarios se armó una lista con nombres de usuarios seguidos por la cuenta @verified, esos usuarios están verificados, de cada usuario verificado se obtienen 30 tweets, y de cada tweet se toman ciertas características

Marco metodológico

Datos

También se arma un dataset *Bots*, en este caso, se extraen nombres de usuario de cuentas bots que se encuentran en el dataset botwiki-2019. Se extraen los nombres de usuario de los bots, de los cuales se toman 670 y se obtienen 40 tweets de cada uno, llegando así a 26.800 tweets de los cuales se extraen las mismas características que para el dataset *Humanos*.

Marco metodológico

Datos

Para el dataset *Total* se toman los dos datasets anteriores y se extraen algunas características y se agregan otras que servirán para la predicción. Este dataset servirá como base para tomar algunas características con las que se entrenarán los clasificadores.

Marco metodológico
Datos

Dataset	Fuente	Descripción
Humanos	Twitter API (Tweepy): recolectados 2.500 usuarios verificados de los cuales se obtuvieron 75.000 tweets.	Dimensión: (75000, 19) Variables: 19
Bots	Twitter API (Tweepy): Para usuarios bots, se toman 670 usuarios del dataset botwiki-2019, de los cuales se obtuvieron 26.800 tweets.	Dimensión: (26800, 19) Variables: 19
Total	Se toman los datasets Humanos y Bots, se extraen algunas características y se agregan otras para la predicción.	Dimensión: (3188, 17) Características: 16 Target: 1 (bot)

Marco metodológico
Extracción y análisis de características

Marco metodológico
Extracción y análisis de características

Se extraen muchas características de Twitter, a continuación presento una lista de la información extraída de cada tweet junto a datos del usuario autor que lo publicó:

- ID del tweet
- Texto del tweet
- Fecha de creación del tweet
- Cantidad de favoritos del tweet
- Cantidad de RTs del tweet
- Cantidad de menciones en el tweet
- Cantidad de links en el tweet
- Cantidad de hashtags en el tweet
- Fuente del tweet
- ...

Marco metodológico
Extracción y análisis de características

- ...
- ID del usuario
- Nombre de pantalla del usuario
- Nombre del usuario
- Fecha de creación del usuario
- Biografía del usuario
- Longitud de la biografía del usuario
- Ubicación del usuario
- Cantidad de seguidos por el autor del tweet
- Cantidad de seguidores del autor del tweet
- Cantidad de tweets del usuario autor del tweet

Marco metodológico

Extracción y análisis de características

- $Promedio\ de\ favoritos = \frac{\text{Sumatoria de la cantidad de favoritos}}{30}$
- $Promedio\ de\ retweets = \frac{\text{Sumatoria de la cantidad de retweets}}{30}$
- $Promedio\ de\ menciones = \frac{\text{Sumatoria de la cantidad de menciones}}{30}$
- $Promedio\ de\ links = \frac{\text{Sumatoria de la cantidad de links}}{30}$
- $Promedio\ de\ hashtags = \frac{\text{Sumatoria de la cantidad de hashtagss}}{30}$
- $Reputación\ de\ usuario = \frac{\text{Promedio de retweets}}{\text{Cantidad de seguidores}}$
- $Antiguedad =$ fechaActual $-$ Fecha de creación del usuario
- $Tweet\ por\ día = \frac{\text{Cantidad de tweets}}{\text{Antiguedad}}$
- $Tweet\ por\ seguidor = \frac{\text{Cantidad de tweets}}{\text{Cantidad de seguidores}}$
- $Antiguedad\ por\ seguidos = \frac{\text{Antiguedad}}{\text{Cantidad de seguidos}}$
- $name_{binary} = \begin{cases} 1 \\ 0 \end{cases}$ dependiendo si el nombre contiene alguna palabra relacionada a bot.
- $user\_name_{binary} = \begin{cases} 1 \\ 0 \end{cases}$ dependiendo si el nombre de pantalla contiene alguna palabra relacionada a bot.
- $description_{binary} = \begin{cases} 1 \\ 0 \end{cases}$ dependiendo si la biografía contiene alguna palabra relacionada a bot.

Marco metodológico

Selección de características para la predicción

Las características que se tienen en cuenta para entrenar los clasificadores son las siguientes:

- Reputación de usuario.
- Promedios de hashtags, links, retweets, favoritos y menciones en los tweets.
- Antigüedad de la cuenta.
- Cantidad de usuarios seguidos.
- Cantidad de usuarios seguidores.
- Cantidad de tweets.
- Cantidad de tweets por día.
- Cantidad de tweets por seguidor.
- Cantidad de días de la cuenta (antigüedad) por usuarios seguidos.
- Valores binarios que indican si el nombre de usuario, el nombre o la descripción contienen alguna palabra relacionada a bot.

Marco metodológico

Clase objetivo

bot (atributo): toma el valor “bot” para el perfil de usuario de una cuenta bot, y el valor “noBot” para el perfil de usuario de una cuenta manejada por un humano.

Marco metodológico

Métricas de evaluación

Exactitud: Fracción de predicciones que se realizaron correctamente en un modelo de clasificación.

En la clasificación de clases múltiples, la exactitud se define de la siguiente manera:

$$Exactitud = \frac{Predicciones\ correctas}{Número\ total\ de\ ejemplos}$$

En la clasificación binaria, la exactitud se calcula con la siguiente fórmula:

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

Marco metodológico

Métricas de evaluación

Precisión: La precisión identifica la frecuencia con la que un modelo predijo correctamente la clase positiva, mide la calidad de la predicción.

Para calcular la precisión se hace uso de la siguiente fórmula:

$$Precisión = \frac{VP}{VP + FP}$$

Marco metodológico

Clasificación

Una tarea típica de aprendizaje supervisado es la clasificación.

La clasificación es una técnica de categorizar un objeto en una clase particular basada en el conjunto de datos de entrenamiento que se utilizó para entrenar al clasificador, con la particularidad que las muestras del conjunto de datos están etiquetadas, para tal fin, alimentamos los clasificadores con el conjunto de datos *Total*.

El detector bot implementa tres clasificadores: K vecinos más cercanos, Naïve Bayes Gaussiano y Bosque aleatorio.

Marco metodológico

K vecinos más cercanos

En la clasificación por K vecinos más cercanos, la salida es la pertenencia a una clase. Un objeto es clasificado como perteneciente a una clase si la mayoría de sus k vecinos pertenecen a esa clase.

Marco metodológico

K vecinos más cercanos

Funcionamiento:

Calcular la distancia entre el objeto a clasificar y el resto de los ítems del dataset de entrenamiento.

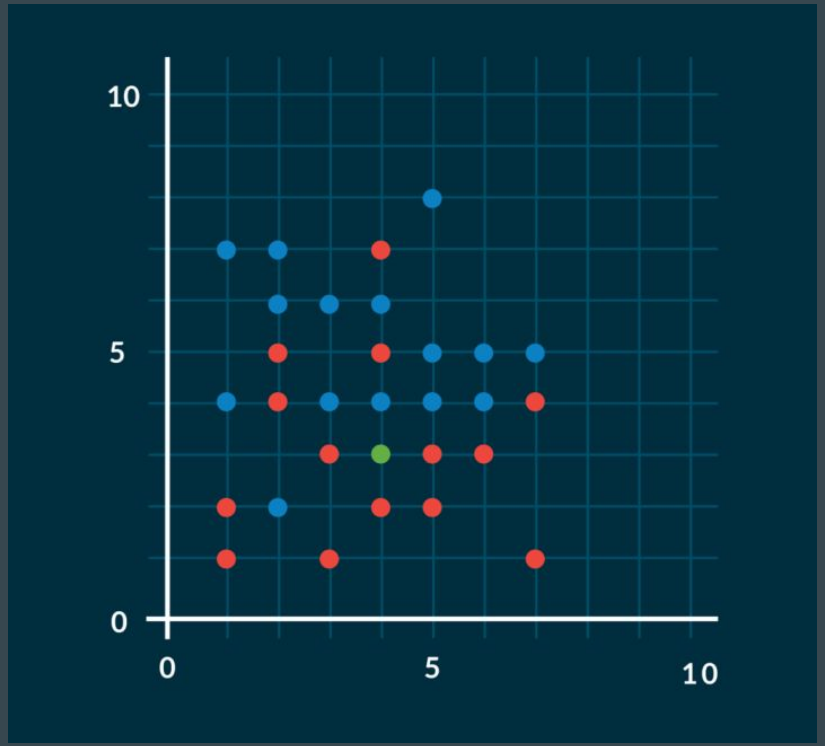
Seleccionar los “k” elementos más cercanos (con menor distancia, según alguna métrica de distancia como la distancia euclidiana).

Realizar una “votación de mayoría” entre los k puntos: los puntos de una clase que “dominen” decidirán la clasificación final.

Marco metodológico
K vecinos más cercanos

Ejemplo:

Se busca clasificar el punto (4, 3) para saber si es de color azul o rojo, con k=7 vecinos y como métrica la distancia euclidiana.



Marco metodológico
K vecinos más cercanos

Rojo	Distancia al (4,3)
(1,1)	3.605551275
(1,2)	3.16227766
(2,4)	2.236067977
(2,5)	2.828427125
(3,1)	2.236067977
(3,3)	1
(4,2)	1
(4,5)	2
(4,7)	4

Rojo	Distancia al (4,3)
(5,2)	1.414213562
(5,3)	1
(6,3)	2
(7,1)	3.605551275
(7,4)	3.16227766

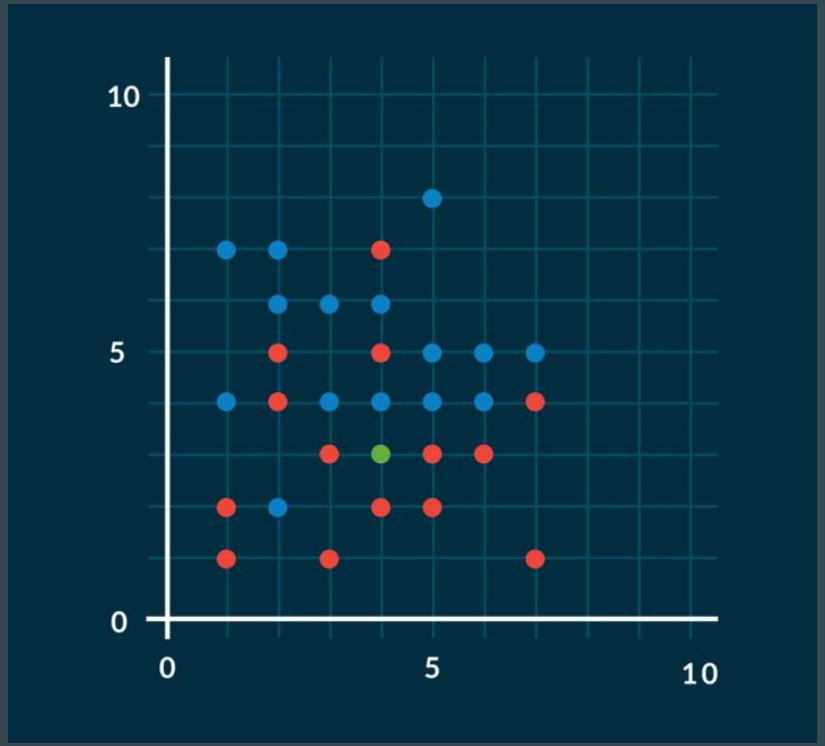
Azul	Distancia al (4,3)
(1,4)	3.16227766
(1,7)	5
(2,2)	2.236067977
(2,7)	4.472135955
(2,8)	5.385164807
(3,4)	1.414213562
(3,6)	3.16227766
(4,4)	1
(4,6)	3

Azul	Distancia al (4,3)
(5,4)	1.414213562
(5,5)	2.236067977
(5,8)	5.099019514
(6,4)	2.236067977
(6,5)	2.828427125
(7,5)	3.605551275

Marco metodológico
K vecinos más cercanos

Ejemplo:

Los puntos (3, 3), (3, 4), (4, 2), (4, 4), (5, 2), (5, 3) y (5, 4) son los puntos con menor distancia al punto (4, 3), de los cuales 4 son rojos y 3 son azules.
Por lo tanto, el punto (4, 3) se clasifica rojo.



Para el clasificador K vecinos más cercanos:

- k = 3
- distancia Manhattan

0.9545454545454545

Exactitud del clasificador K vecinos más cercanos

0.9403441629211 2

Precisión del clasificador K vecinos más cercanos

Marco metodológico

Naïve Bayes Gaussiano

Naïve Bayes es un método de clasificación basado en el Teorema de Bayes. Los clasificadores Naïve Bayes tienen alta precisión y velocidad en grandes conjuntos de datos.

El clasificador Naïve Bayes asume como hipótesis la independencia condicional de clase: el efecto de una característica particular en una clase es independiente de otras características.

Marco metodológico

Naïve Bayes Gaussiano

El Teorema de Bayes está expresado por la siguiente ecuación:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

- ★ *P(H)* es el conocimiento inicial que tenemos sobre que la hipótesis *H* sea la correcta, se le suele denominar la probabilidad a priori de *H*.
- ★ *P(D)* se define de forma similar, pero esta vez sobre los datos *D*.
- ★ *P(D|H)* denota la probabilidad de observar los datos *D* dado que tenemos la hipótesis *H*, se le suele denominar verosimilitud.
- ★ *P(H|D)* es la probabilidad a posteriori que la hipótesis *H* tiene, dados los datos observados *D*.

Marco metodológico
Naïve Bayes Gaussiano

Clima	Temperatura	Humedad	Ventoso	Tenis (Clase objetivo)
Lluvioso	Calor	Alta	Falso	No
Lluvioso	Calor	Alta	Verdadero	No
Nublado	Calor	Alta	Falso	Sí
Soleado	Templado	Alta	Falso	Sí
Soleado	Frío	Normal	Falso	Sí
Soleado	Frío	Normal	Verdadero	No
Nublado	Frío	Normal	Verdadero	Sí
Lluvioso	Templado	Alta	Falso	No
Lluvioso	Frío	Normal	Falso	Sí
Soleado	Templado	Normal	Falso	Sí

Clima	Temperatura	Humedad	Ventoso	Tenis (Clase objetivo)
Lluvioso	Templado	Normal	Verdadero	Sí
Nublado	Templado	Alta	Verdadero	Sí
Nublado	Calor	Normal	Falso	Sí
Soleado	Templado	Alta	Verdadero	No

Marco metodológico
Naïve Bayes Gaussiano

Ejemplo:
Sea el vector dependiente (clase objetivo) y, y el conjunto de características X donde contiene características como $X = \left(x_1, x_2, \ldots, x_n \right)$ se puede expresar el Teorema de Bayes como:

$$P(y|X) = \frac{P(X|y)\, P(y)}{P(X)}$$

Suponiendo el caso X = (Soleado, Calor, Normal, Falso) y se busca predecir y.

Marco metodológico
Naïve Bayes Gaussiano

Ejemplo:

$$P(Sí \mid hoy) = \frac{P(Clima\ Soleado \mid Sí)\,P(Temperatura\ Calor \mid Sí)\,P(Humedad\ Normal \mid Sí)\,P(Ventoso\ Falso \mid Sí)\,P(Sí)}{P(hoy)}$$

$$P(No \mid hoy) = \frac{P(Clima\ Soleado \mid No)\,P(Temperatura\ Calor \mid No)\,P(Humedad\ Normal \mid No)\,P(Ventoso\ Falso \mid No)\,P(No)}{P(hoy)}$$

$$P(Sí \mid hoy) = \frac{2}{9} \times \frac{2}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14} \approx 0.0141$$

$$P(No \mid hoy) = \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14} \approx 0.0068$$

X = (Soleado, Calor, Normal, Falso)

$$P(Sí \mid hoy) > P(No \mid hoy) \Rightarrow y = Sí$$

Marco metodológico
Naïve Bayes Gaussiano

Clima				
	Sí	No	P(Sí)	P(No)
Soleado	2	3	2/9	3/5
Nublado	4	0	4/9	0/5
Lluvioso	3	2	3/9	2/5
Total	**9**	**5**	**100%**	**100%**

Temperatura				
	Sí	No	P(Sí)	P(No)
Calor	2	2	2/9	2/5
Templado	4	2	4/9	2/5
Frío	3	1	3/9	1/5
Total	**9**	**5**	**100%**	**100%**

Humedad				
	Sí	No	P(Sí)	P(No)
Alta	3	1	3/9	4/5
Normal	6	1	6/9	1/5
Total	**9**	**5**	**100%**	**100%**

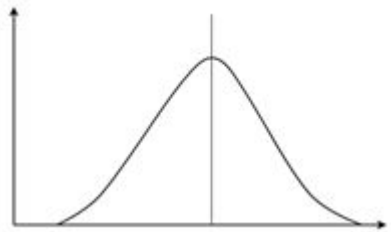
Ventoso				
	Sí	No	P(Sí)	P(No)
Falso	3	1	3/9	4/5
Verdadero	6	1	6/9	1/5
Total	**9**	**5**	**100%**	**100%**

Tenis		P(Sí)/P(No)
Sí	9	9/14
No	5	5/14
Total	**14**	**100%**

Marco metodológico

Naïve Bayes Gaussiano

En el ejemplo, los datos proporcionados fueron de tipo discreto. Si los datos son variables continuas, se puede aplicar Naïve Bayes Gaussiano. Los diferentes clasificadores ingenuos de Bayes se diferencian principalmente por las suposiciones que hacen con respecto a la distribución de $P(x_i \mid y)$



0.9764890282131 7

Exactitud del clasificador Naïve Bayes Gaussiano

0.9694524048527 3

Precisión del clasificador Naïve Bayes Gaussiano

Marco metodológico

Bosque aleatorio

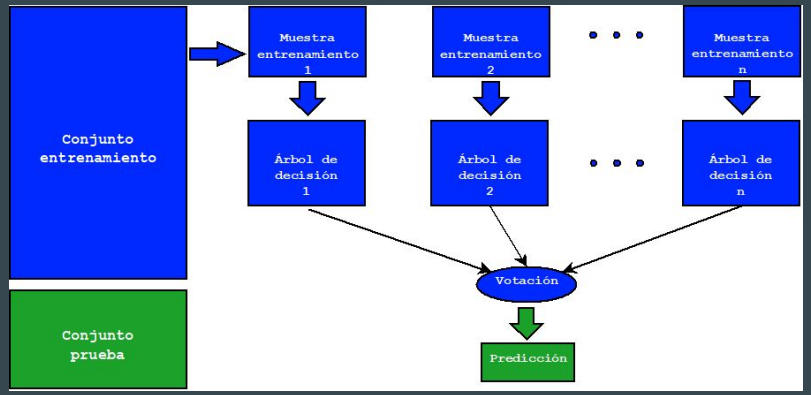
Los bosques aleatorios es un algoritmo de aprendizaje supervisado. Se puede utilizar tanto para clasificación como para regresión.

Los bosques aleatorios crean árboles de decisión con ejemplos seleccionados al azar, obtienen una predicción de cada árbol y seleccionan la mejor solución mediante votación.

Marco metodológico
Bosque aleatorio

Funcionamiento:

1. Selecciona muestras aleatorias de un dataset determinado.
2. Construye un árbol de decisión para cada muestra y obtiene un resultado de predicción de cada árbol de decisión.
3. Realiza una votación por cada resultado previsto.
4. Selecciona el resultado de la predicción con más votos como predicción final.



0.8338557993730 4

Exactitud del clasificador Bosque aleatorio

0.8260531626146 3

Precisión del clasificador Bosque aleatorio

Marco metodológico
Funcionalidades adicionales

Está la posibilidad de visualizar los tweets que sirvieron para extraer las características para la clasificación de una cuenta en bot o humano.

Se pueden obtener los puntajes de Botometer. Botometer toma un nombre de usuario de Twitter y da ciertas puntuaciones, el rango de esas puntuaciones va desde 0 hasta 5, mientras esa puntuación es más cercana a 0 esa cuenta de Twitter tiene una actividad similar a un humano, y si esa puntuación es cercana a 5 dicha cuenta de Twitter tiene una actividad más similar a un bot.

Marco metodológico
Funcionalidades adicionales

Los puntajes mostrados por el puntaje Botometer son los siguientes:

- General: Basado en una comparación de varios modelos entrenados en diferentes tipos de bots y en cuentas humanas, corresponde al modelo con mayor confianza.
- Astroturf: Bots políticos etiquetados manualmente y cuentas involucradas en trenes de seguimiento que eliminan contenido sistemáticamente.
- Fake follower: Bots comprados para aumentar el número de seguidores.
- Financial: Bots que publican utilizando cashtags.
- Other: Varios otros bots obtenidos a partir de anotaciones manuales, feedback de los usuarios, etc.
- Self declared: Bots de botwiki.org.
- Spammer: Cuentas etiquetadas como spambots de varios datasets.

Detector bot:
https://tesis-uns.herokuapp.com/

Demostración

...

Caso 1:
Cuenta con nombre de usuario inválido

Demostración

•••

Caso 2:
Cuenta con nombre de usuario inexistente

Demostración

...

Caso 3:
Cuenta sin tweets

Demostración

...

Caso 4:
Cuenta con los tweets protegidos

Demostración

...

Caso 5:
Cuenta válida con tweets públicos

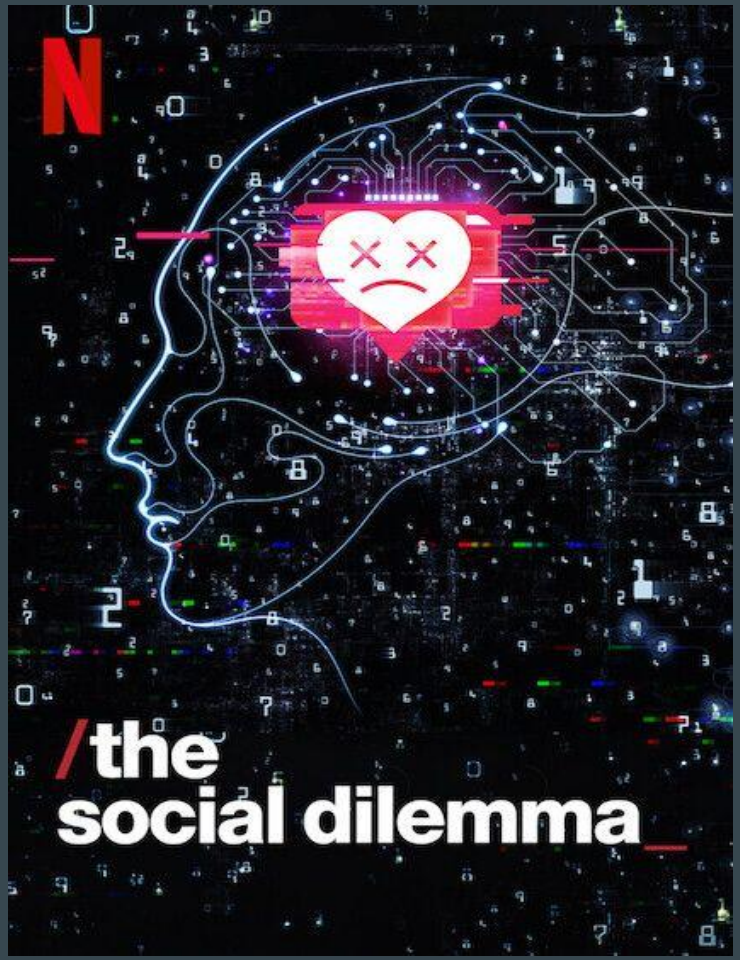
Conclusiones

Conclusiones

- Es necesario que los bots y los humanos puedan reconocerse entre sí.
- Necesidad de instancias de entrenamiento.
- La manipulación de datos de redes sociales crea desinformación. Es necesario tomar medidas para disminuir este fenómeno desde todos los ángulos:
  - eliminar sitios web fraudulentos.
  - desarrollar técnicas de detección más sofisticadas.
  - establecer fuertes barreras de registro.
  - contribuir a la conciencia de los potenciales compradores de servicios de fraude.

Conclusiones

- Con respecto al detector bot implementado:
  - La detección de bots no es una tarea fácil, se utilizan muchas características para tal fin, por lo que el resultado obtenido puede no ser certero.
  - Los modelos entrenados dieron buenos resultados, siendo el modelo de Bosques Aleatorios el modelo menos confiable
  - La aplicación web cuenta con limitaciones para el tamaño de la base de datos.
  - La aplicación web implementada parte de información de la cuenta pública que se puede ver en Twitter e identifica las características que la convierten en un bot; por ejemplo, el nombre de la cuenta, el número de tweets, la ubicación en la biografía, los hashtags utilizados, etc. Este enfoque es limitado.
- Un trabajo futuro, más allá de las características públicas, puede enfocarse en patrones de interacción y en el contenido de los mensajes.


/the
social dilemma


NETFLIX
OFFICIAL
TRAILER

¡Gracias!