PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

DETECCIÓN AUTOMÁTICA DE BOTS EN TWITTER

TEODORO SANDOVAL PRADO

PROFESOR GUIA: **WENCESLAO PALMA MUÑOZ**

INFORME DE AVANCE DE PROYECTO DE TITULO
PARA OPTAR AL TÍTULO PROFESIONAL DE
INGENIERO DE EJECUCIÓN EN INFORMÁTICA

NOVIEMBRE, 2017

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

DETECCIÓN AUTOMÁTICA DE BOTS EN TWITTER

TEODORO SANDOVAL PRADO

PROFESOR GUIA: **WENCESLAO PALMA MUÑOZ**
PROFESOR CORREFERENTE: **HÉCTOR ALLENDE CID**

INFORME DE AVANCE DE PROYECTO DE TITULO
PARA OPTAR AL TÍTULO PROFESIONAL DE
INGENIERO DE EJECUCIÓN EN INFORMÁTICA

NOVIEMBRE, 2017

Resumen

En la plataforma *Twitter* es posible controlar usuarios por medio de programas automatizados que se definen como *bots*. Estas cuentas *bot* pueden causar problemas a otros usuarios cuando generan *Spam* o información confusa. En la comunidad científica también constituyen un problema puesto que generan información incoherente que constituye ruido y caos. Se pueden reconocer cuentas *bots* en *Twitter* con algoritmos de aprendizaje y así pueden ser reportadas o filtradas. Esto puede beneficiar tanto a la comunidad científica como a los usuarios comunes. En este trabajo se propone la creación de un clasificador de cuentas para ser implementado como modulo en otra investigación sobre *Twitter*.

Palabras Clave: twitter, bots, machine learning, clasificador, caracteristicas, botometer.

Índice

Lista de Figuras

Lista de Tablas

1. Introducción

Twitter es una red social que se caracteriza por conectar a sus usuarios de forma unilateral y por tener un límite de 140 caracteres por publicación (*Tweet*). Estas cualidades hacen de *Twitter* una plataforma de comunicación directa y concisa que hoy cuenta con cerca de 319 millones de usuarios cada mes y genera alrededor de 600 millones de *Tweets* por día. La masiva cantidad de información generada y distribuida por *Twitter* tiene muchos usos prácticos, como seguimiento de tendencias, promulgación de información, análisis demográfico, etc. Es posible automatizar tareas en cuentas de *Twitter* por medio de *bots*, de manera que se puede publicar noticias o publicidad de forma regular y perpetua. Sin embargo es bastante común el uso indebido de *bots*, propagándose por la red de usuarios y divulgando contenido no deseado (*spam*) o información falsa. Además, las cuentas *bots* son un obstáculo cuando se intenta hacer estudios sobre la plataforma, puesto que producen información confusa e incoherente con el comportamiento común de un usuario humano.

Un grupo de académicos de la Pontificia Universidad Católica de Valparaíso (PUCV) actualmente está realizando un estudio con el cual esperan poder predecir las elecciones presidenciales a través del análisis de *Tweets*, por lo que se necesita identificar y descartar los *twitters* de cuentas *bots* que puedan ensuciar las muestras. Los algoritmos de aprendizaje de maquina ofrecen una solución ideal, pero primero se deben recolectar datos característicos de cada usuario y *datasets* de entrenamiento.

2. Objetivos

2.1. Objetivo general

- Clasificar cuentas de *Twitter* como *Bot* o Humano a partir de la actividad y atributos del usuario.

2.2. Objetivos específicos

- recolectar las características necesarias para cada usuario de *twitter* con la API para *python*.
- entrenar algoritmos de aprendizaje con *datasets* de usuarios *bot* disponibles en la comunidad científica.
- identificar a los usuarios *bot* de la base de datos de la investigación asociada.
- mejorar resultados probando distintas configuraciones de características.
- probar técnicas de reducción de ruido.

3. Definiciones

Para la mejor comprensión de este documento se explicaran y definirán algunos conceptos utilizados:

- **Twitter**: es una plataforma web con función de red social y *microbloggin*, se caracteriza por tener relaciones asimétricas entre usuarios y por limitar la cantidad de caracteres a 140 por post sin incluir Entidades. Internamente *Twitter* funciona con cuatro objetos principales: *Tweets*, Usuarios, Entidades y Lugares. *Twitter* ofrece abiertamente una API para acceder a todas sus funcionalidades por medio de programación, para utilizarla se debe registrar e identificar un usuario.
- **Usuario**: Cuenta de *Twitter* que puede tener un perfil personalizado o uno por defecto, y su actividad puede estar controlada por un humano o un software, puede seguir (Follow), ser seguido (Followed) o ser referenciado con una Mención.
- **Tweet**: También llamado *Status*, es la unidad de información principal en *Twitter*, puede contener texto plano o Entidades. puede ser Re-Tweeteado, que consiste en retransmitir el mismo mensaje entre usuarios y es la acción básica para propagar información a través de *Twitter*.
- **Entidad**: Son campos de texto que se reconocen y se extraen del contenido del *Tweet*, estas pueden ser *Hashtags*, Media, *URLs* o Menciones de usuario.
- **Follow**: Es la acción de conectarse con un usuario para recibir sus publicaciones en la portada de la cuenta (*Screen*), la persona a la que se realiza *Follow* no recibe información directa del usuario que lo realizó, pero si puede ver la cantidad total de *Follows* a su cuenta.
- **Screen**: Es la portada de la cuenta de usuario, en donde se reciben los *Tweets* realizados por todos los usuarios a los que se ha realizado *Follow*. La portada puede ser vista por cualquier usuario.
- **Part Of Speech**: Abreviado como POS, traducible como Partes del habla o Categoría gramatical, son identificadores de uso de las palabras, existen varias formas de clasificar las palabras según su uso y pueden varias entre idiomas. Para este documento se consideraran nueve: verbos, sustantivos, adjetivos, verbos auxiliares, predeterminantes, interjecciones, adverbios, preguntas informativas y pronombres.
- **Características**: son datos que pueden ser extraídos de forma directa o por medio de un proceso de datos, determinan cualidades cuantificables y distintivas de donde son extraídas.
- **Atributo**: se refiere a datos crudos contenidos en estructuras de datos.

4. Descripción del problema

En colaboración con un proyecto de predicción de elecciones presidenciales por medio de *Twitter* realizado por académicos de la PUCV, se necesita separar las cuentas *bots* y sus *Twitters* para poder realizar una lectura adecuada de los datos y mejorar la probabilidad de acierto. Para poder reconocer de forma automática cuentas *bot* en *Twitter* se deben reconocer las características que las distinguen de cuentas humanas. *Twitter* funciona con 4 estructuras que interactúan y mantienen información sobre los usuarios y *Tweets*: *Tweets* (o *Status*), usuarios (*User*), entidades (*Entities*) y lugares (Places), cada una con campos que definen características propias. Para poder reconocer las cuentas con algoritmos de aprendizaje se deben extraer ciertas características a partir de los datos crudos que contienen, para este propósito se utiliza como referencia la investigación de Botometer[1], que sugiere las características mostradas en el anexo, sin embargo, para el propósito de la investigación asociada se puede prescindir de algunas de dichas características.

4.1. Características

En el *paper* de referencia, se sugieren una serie de atributos ordenados en 6 grupos, que tras recolectar sus datos estadísticos (mínima, máxima, mediana, media, desviación estándar, asimetría, curtosis y entropía) reúnen un total de 1150 características(ver en anexo). Sin embargo, para reducir la envergadura de este trabajo se descartaron los atributos de "Sentimiento".

El proceso de extracción de características se realiza con la *API* de *Twitter*, específicamente con el *wrapper Tweepy* para *python*. Si bien la API permite acceder a todas las funcionalidades y datos públicos con los que cuenta *Twitter*, esta cuenta con ciertas limitaciones de escala de uso, por ejemplo, solo se puede solicitar los seguidores o *retweets* de un determinado usuario de a 100 por vez. También se cuenta con limites por ventanas de 15 minutos para cada tipo de solicitud. Estas limitaciones son importantes de considerar cuando se intentan obtener características que necesitan un amplio barrido de la red de usuarios, como las características de Amigos y de Red.

A grandes rázagos, las características agrupadas consisten en:

- **Meta-datos de usuario:** Estas características se obtienen de las propiedades de la cuenta del usuario, como datos de identificación, configuración de la cuenta, cantidad de amigos, seguidores favoritos, etc.

- **Amigos:** Estas características se obtienen de los usuarios a los que se sigue. Los usuarios pueden interactuar siguiendo, *retweeteando* y mencionando explícitamente a otros usuarios. Estas características se extraen separando 4 grupos de usuarios amigos: *retweeteando*, mencionando, *retweeteado* y mencionado.

- **Red:** Estas características se obtienen del análisis de estructuras creadas a partir de la comunicación entre usuarios. Se deben reconstruir tres tipos de redes: de *retweets*, menciones, y co-ocurrencia de *hashtags*. Las redes de *retweets* y menciones usan a los usuarios como nodos y se constyruyen siguiendo la direccion de propagación. la red de co-ocurrencia de *hashtags* usa a los *hashtags* como nodos y se arman tomando la aparición simultanea de estos en un mismo *tweet*. a cada red se calcula una serie de características en torno a los pesos de los nodos.

- **Contenido:** Estas características se obtienen del texto contenido en los *tweets* del usuario. Se utilizan etiquetas de partes del habla, o *Part-of-Speech* en ingles (POS), para identificar componentes de lenguaje natural.

- **Tiempos:** Estas características se obtienen de los tiempos entre ocurrencias de *tweets*, *retweets* y menciones realizadas por el usuario a lo largo de toda su historia. de estos tiempos de extraen características de distribución.

Meta-datos de usuario	Largo del nombre en pantalla (*Screen name*) Numero de dígitos del nombre en pantalla Largo de nombre de usuario Desplazamiento de tiempo (*UTC*) (seg.) Perfil por defecto (binario) Imagen por defecto (binario) Edad de la cuenta (días) Número de descripciones de perfil únicas (*) Largos de las descripciones de perfil (*) Distribución del numero de amigos (*) Distribución del numero de seguidores (*) Distribución del numero de favoritos Número de amigos (razón de señal-ruido y cambio relativo) Número de seguidores (razón de señal-ruido y cambio relativo) Número de favoritos (razón de señal-ruido y cambio relativo) Número de *tweets* (por hora y total) Número de *retweets* (por hora y total) Número de menciones (por hora y total) Número de respuestas (por hora y total) Número de *retweeteado* (por hora y total)	**Amigos (†)** Número de idiomas distintos Entropía del uso de idiomas (*) Distribución de edad de la cuenta (*) Distribución del desplazamiento de tiempo (*) Distribución del número de amigos (*) Distribución del número de seguidores (*) Distribución del número de *tweets* (*) Distribución del largo de la descripción Fracción de usuarios con perfil e imagen por defecto
Red (‡)	Número de nodos Número de bordes (también para recíprocos) (*) Distribución de fuerza (*) Distribución de fuerza interna (*) Distribución de fuerza externa Densidad de la red (también para recíprocos) (*) Coeficiente de *clustering* (también para recíprocos)	**Contenido** (*,**) Frecuencia de etiquetas POS en el *tweet* (*,**) Proporción de etiquetas POS en el *tweet* (*) Número de palabras en el *tweet* (*) Entropía de palabras en el *tweet*
Tiempos	(*) Tiempo entre dos *tweets* consecutivos (*) Tiempo entre dos *retweets* consecutivos (*) Tiempo entre dos menciones consecutivas	

† Se consideran cuatro tipos de usuarios conectados: *retweeteando*, mencionando, *retweeteados*, mencionados.
‡ Se consideran tres tipos de redes: de *retweet*, mención y co-ocurrencia de *hashtags*.
* Tipos de distribución. Por cada distribución, se computan las siguientes ocho estadísticas y se usan como características individuales: mínima, máxima, mediana, media, desviación estándar, asimetría, curtosis y entropía.
** Etiquetas de partes del habla (*Part-Of-Speech, POS*). Hay nueve etiquetas *POS* : verbos, sustantivos, adjetivos, verbos auxiliares, predeterminantes, interjecciones, adverbios, preguntas informativas y pronombres.

Tabla 1: Lista de características basadas en la investigación de *Botometer[1]*

Table 1: List of 1150 features extracted by our framework.

User meta-data	Screen name length	**Sentiment**	(***) Happiness scores of aggregated tweets
	Number of digits in screen name		(***) Valence scores of aggregated tweets
	User name length		(***) Arousal scores of aggregated tweets
	Time offset (sec.)		(***) Dominance scores of single tweets
	Default profile (binary)		(*) Happiness score of single tweets
	Default picture (binary)		(*) Valence score of single tweets
	Account age (days)		(*) Arousal score of single tweets
	Number of unique profile descriptions		(*) Dominance score of single tweets
	(*) Profile description lengths		(*) Polarization score of single tweets
	(*) Number of friends distribution		(*) Entropy of polarization scores of single tweets
	(*) Number of followers distribution		(*) Positive emoticons entropy of single tweets
	(*) Number of favorites distribution		(*) Negative emoticons entropy of single tweets
	Number of friends (signal-noise ratio and rel. change)		(*) Emoticons entropy of single tweets
	Number of followers (signal-noise ratio and rel. change)		(*) Positive and negative score ratio of single tweets
	Number of favorites (signal-noise ratio and rel. change)		(*) Number of positive emoticons in single tweets
	Number of tweets (per hour and total)		(*) Number of negative emoticons in single tweets
	Number of retweets (per hour and total)		(*) Total number of emoticons in single tweets
	Number of mentions (per hour and total)		Ratio of tweets that contain emoticons
	Number of replies (per hour and total)		
	Number of retweeted (per hour and total)		
Friends (†)	Number of distinct languages	**Network (‡)**	Number of nodes
	Entropy of language use		Number of edges (also for reciprocal)
	(*) Account age distribution		(*) Strength distribution
	(*) Time offset distribution		(*) In-strength distribution
	(*) Number of friends distribution		(*) Out-strength distribution
	(*) Number of followers distribution		Network density (also for reciprocal)
	(*) Number of tweets distribution		(*) Clustering coeff. (also for reciprocal)
	(*) Description length distribution		
	Fraction of users with default profile and default picture		
Content	(*,**) Frequency of POS tags in a tweet	**Timing**	(*) Time between two consecutive tweets
	(*,**) Proportion of POS tags in a tweet		(*) Time between two consecutive retweets
	(*) Number of words in a tweet		(*) Time between two consecutive mentions
	(*) Entropy of words in a tweet		

† We consider four types of connected users: retweeting, mentioning, retweeted, and mentioned.
‡ We consider three types of network: retweet, mention, and hashtag co-occurrence networks.
* Distribution types. For each distribution, the following eight statistics are computed and used as individual features: min, max, median, mean, std. deviation, skewness, kurtosis, and entropy.
** Part-Of-Speech (POS) tag. There are nine POS tags: verbs, nuns, adjectives, modal auxiliaries, pre-determiners, interjections, adverbs, wh-, and pronouns.
*** For each feature, we compute mean and std. deviation of the weighted average across words in the lexicon.

Figura 1: Tabla de características del documento original[1]

5. Entrenamiento

para el entrenamiento se necesita armar una base de datos de usuarios y *tweets* dividida en bots y humanos, generar las características para cada usuario y entrenar un algoritmo *Random Forest* con esas características.

5.1. Base de datos

La base de datos utilizada para el entrenamiento fue creada a partir de datasets obtenidos de del sitio web de Botometer [1]. Los datos están separados en 4 grupos: humanos, bots falsos follow, bots sociales y bots tradicionales. cada grupo cuenta con una lista de datos de usuario y una lista de tweets asociados a los usuarios. La base de datos se creó siguiendo esta estructura, resultando 8 colecciones (tabla 2).

Set de colecciones	Nombre	Cantidad de registros
Human	UsersEN_Human	3211
	TweetsEN_Humans	2839361
Bot	UsersEN_Bot_traditional	2400
	UsersEN_Bot_follow	837
	UsersEN_Bot_social	4715
	TweetsEN_Bot_traditional	145094
	TweetsEN_Bot_follow	196027
	TweetsEN_Bot_social	3457133

Tabla 2: Colecciones en la base de datos de entrenamiento

5.2. Características

La base de datos no cuenta con los requerimientos para obtener todas las características, por lo que en el entrenamiento se utilizaron menos que las originalmente planteadas. Solamente se utilizaron las características de “Meta-datos de usuario”, “Tiempos” y “Contenido” acumulando un total de 165 métricas. Las características de .ᴬmigosz Red"se tuvieron que descartar temporalmente debido a que para calcularlas se necesita disponer de subconjuntos de usuarios que se relacionan con cada usuario. Es posible completar la base de datos utilizando la API de Twitter, sin embargo este proceso tomaría mas tiempo del que se dispone debido a las restricciones de tiempo en las consultas que impone Twitter.

El software desarrollado esta pensado para obtener todas las características desde una base de datos, por lo que es posible trabajar con todas las características en caso de tener una base de datos mas completa.

5.3. Pruebas de entrenamiento

El algoritmo utilizado para la clasificación es *Random Forest*, que es un algoritmo utilizado para regresión y clasificación, nos entregara un valor entre 0 y 1 según la probabilidad que tiene de ser bot o no.

Se realizaron pruebas de entrenamiento del clasificador con una colección de 600 sets de caracteristicas, 300 de usuarios humanos y 300 de usuarios bot (100 de cada tipo de bot). Se utilizo

la técnica *cross validation* para realizar pruebas de precisión probando diferentes convinaciones de parametros de entrada para el clasificador, la tabla 3 muestra algunos de estos resultados.

numero de arboles	profundidad máxima	estado aleatorio	porcentaje de acierto
10	6	2	0.98755186722
50	5	2	0.98755186722
11	6	2	0.983402489627
90	5	2	0.983402489627
99	6	0	0.979253112033
24	9	6	0.979253112033
88	8	0	0.97510373444
14	8	2	0.97510373444
15	6	0	0.970954356846
49	8	7	0.970954356846
81	2	3	0.921161825726
11	2	0	0.917012448133
25	2	0	0.912863070539
67	2	0	0.908713692946
19	2	1	0.904564315353
12	2	3	0.821576763485
57	2	5	0.804979253112
14	2	3	0.800829875519
43	2	3	0.796680497925
22	2	3	0.788381742739
65	2	5	0.788381742739
34	2	3	0.771784232365
31	2	3	0.767634854772
30	2	3	0.763485477178
15	2	3	0.759336099585
17	2	3	0.755186721992
10	2	5	0.751037344398

Tabla 3: resultados preliminares del entrenamiento

La tabla 3 muestra resultados seleccionados de ente los mayores y los menores porcentajes de acierto, un análisis superficial sugiere que la profundidad máxima es el parámetro con mayor influencia en los resultados.

6. Clasificación

Para la clasificación se deben extraer las características especificadas previamente, pero esta vez de una base de datos con usuarios y *tweets* Chilenos, específicamente, relacionados con los candidatos presidenciales 2017. La adaptación de los algoritmos de recolección no es de mayor complejidad, sin embargo, esta base de datos tiene una estructura diferente y faltan algunos datos necesarios para la clasificación.

6.1. Características

Para la correcta generación de características fue necesario recopilar los datos de usuarios desde la *API* de *Twitter*, en este caso particular fue posible depender de la *API* puesto que la restricción de solicitudes de usuarios es mas baja (1500 solicitudes de usuarios cada 15 minutos), lo suficiente para recolectar una cantidad aprovechable de usuarios.En el caso de los *tweets* también ocurre una falta de datos que impide la recolección de características de usuario, en este caso no hay mas alternativa que descartar esas características. Para mantener la consistencia de los datos, se les dará un valor por defecto a las características faltantes.

6.2. Dependencias para los resultados

Lamentablemente, para poder analizar resultados concretos se necesita usuarios pre-clasificados entre manos y *bots*. Los datos actuales cuentan con una etiqueta sugerida por un equipo evaluador, sin embargo esta no es del todo confiable puesto que no se realizaron con la atención necesarias.

7. Conclusión

El carácter abierto de la plataforma *Twitter* y el apoyo de sus desarrolladores a la comunidad científica, ofrece innumerables posibilidades de investigación y desarrollo de tecnologías que ayuden en ámbitos sociales, comerciales, políticos, etc. Sin embargo esto es un arma de doble filo para el funcionamiento integro de la plataforma. Al ser de libre uso se puede utilizar de forma abusiva para influir en las masas y generar información falsa o confusa con consecuencias negativas, tanto para usuarios comunes como para desarrollo científico. La capacidad de discernir automáticamente si una cuenta es un *bot* o un humano ayuda a controlar el ruido en el contenido de la red de *Twitter*, lo que favorece a la comunidad científica y ayuda a mantener el correcto funcionamiento de *Twitter*.

Para la correcta implementación de algoritmos de aprendizaje, es importante tener a disposición un set de datos grande y consistente. Datasets incompletos o diferentes unos de otros llevan a importantes retrasos en la investigación y pueden dar resultados insatisfactorios, provocando que el algoritmo no pueda ser aplicado en la practica.

Referencias

[1] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization, 2017.