ISSN 1870-4069

Detección de bots en redes sociales usando técnicas procesamiento de lenguaje natural

Daniel Yacob-Espinosa[1], Helena Gómez-Adorno[2], Grigori Sidorov[1]

[1] Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico

[2] Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Mexico

espinosagonzalezdaniel@gmail.com, sidorov@cic.ipn.mx, helena.gomez@iimas.unam.mx

Resumen El uso de las redes sociales ha inundado nuestras vidas, no sólo para pasar el tiempo libre y la conexión entre las personas, sino también como una fuente de comunicación y difusión vía internet. Debido al aumento de usuarios y empresas que interactúan dentro de redes sociales, se han introducido ciertos programas para interactuar con los usuarios, mejor conocidos como: Bots. Los bots son programas los cuales pretenden tener un comportamiento humano en las redes sociales. En los últimos años, el uso de bots ha aumentado exponencialmente ya que pueden resolver ciertos problemas de manera rápida, pero también se han utilizado para con otros fines como: Alteración de información, creación de conflictos para usuarios humanos, difusión de noticias falsas, manipulación de opiniones, entre otros. Este trabajo de investigación muestra un método de clasificación entre usuarios humanos y bots en redes sociales principalmente Twitter. Los tweets de los usuarios son analizados y utilizados en diferentes estructuras de Ngramas donde se usan para entrenar un modelo de clasificación. El método logró una precisión superior al 90 % para los idiomas inglés y español.

Keywords: Detección de bots, redes sociales, procesamiento de lenguaje natural, usuarios humanos, tweets, clasificación, precisión.

Bot Detection on Social Media using Natural Language Processing Techniques

Abstract. Nowadays, social media is an essential component of our lives, not only for leisure time and connecting with people, but also as a source of communication and dissemination via the internet. Due to the increase of users and companies that interact within social networks, certain programs have been

introduced to interact with users, better known as bots. Bots are programs that pretend to have human behavior on social media. In recent years, the use of bots has increased exponentially since they can solve certain problems quickly, but they have also been used for other purposes such as: alteration of information, creation of conflicts for human users, dissemination of fake news, manipulation of opinions, among others. This research work shows a method of classification between human users and bots on social networks, mainly Twitter. Users' tweets are analyzed and used in different N-gram structures, which are used to train a classification model. The method achieved an accuracy of over 90% for the English and Spanish languages.

Keywords: Bot detection, social media, natural language processing, human users, tweets, classification, accuracy.

1. Introducción

En los últimos años hemos visto como las redes sociales han cambiado y se están convirtiendo en un medio de comunicación masivo donde cada día interactúan más usuarios entre ellas. Ahora que vivimos en una pandemia, la interacción en estos medios ha aumentado un 23 % según los reportes de Twitter [8].

En consecuencia a esto, las interacciones en redes sociales aumenta y a su vez, aumenta la difusión de información por estos medios. Estos aumentos en interacciones se deben a que las redes sociales son principalmente usadas por la inmediatez, facilidad y difusión de información.

Debido al uso cotidiano y en aumento en el uso de redes sociales; las empresas y marketing se comienzan a migrar e involucrarse más en estas tecnologías, haciendo uso estos medios para dar un mejor servicio como mostrar anuncios publicitarios, enfocar usuarios objetivos, etc.

Podemos ver que esto tiene sentido para dar un mejor servicio y controlar mejor al publico con el que se relacionan, pero también empiezan a usar diferentes implementaciones que les ayudan a mejorar su rol de negocios. Una implementación que ha tomado relevancia en los últimos tiempos son: Los bots.

Los bots en redes sociales son programas los cuales tratan de emular el comportamiento humano, haciendo parecer que los usuarios están interactuando con una persona como ellos a través de la red social, sin embargo es un programa de computadora.

Este tipo de programas actualmente suelen ser muy utilizados ya que las peticiones de los usuarios llegan a ser colosales y la mejor manera de distribuirlas y atenderlas es con el uso de estos programas.

Podría parecer que el uso de bots en redes sociales llega a ser benéfico en muchos aspectos pero la realidad muestra que estos programas se han visto muy involucrados en muchas controversias, una de ellas es el referéndum por la independencia catalana de octubre del 2017 [7], el cual se vio gravemente afectado por el uso de bots, ya que uno cada tres usuarios que daba su opinión era bot, donde principalmente eran usados para crear conflictos con los usuarios humanos.

Tabla 1. Combinaciones de Ngramas.

Español	Inglés
2 Caracteres-Ngramas	1 Caracteres-Ngramas
3 Caracteres-Ngramas	2 Caracteres-Ngramas
5 Caracteres-Ngramas	3 Caracteres-Ngramas
1 Palabras-Ngramas	2 Palabras-Ngramas
3 Palabras-Ngramas	3 Palabras-Ngramas

Este tipo de investigaciones muestra como este tipo de programas se ven relacionados en problemas tan sensibles para la sociedad.

Como podemos observar los bots pueden tener tareas sencillas como la de promocionar un producto o servicio, sin embargo también pueden tener tareas con otros fines como: Crear controversias de opiniones hacia un cierto interés [3].

En relación a este tipo de conflictos existen ciertas aplicaciones para ayudar a realizar la clasificación, uno de ellos es BorOrNot [2], el cual es una aplicación que nos indica la probabilidad de que un usuario en Twitter pueda ser un bot.

Esta aplicación es una gran herramienta donde no sólo usa el texto de los tweets para realizar la tarea, también emplea el uso de los metadatos de la cuenta para realizar la clasificación, como pueden ser: Cambios de foto de perfil, número de seguidores, si la cuenta ha sido verificada, entre otros.

Es una aplicación que relaciona todos estas características para determinar su tarea, sin embargo las tecnologías implementadas para el uso de bots también esta mejorando y puede complicarse si se deja de lado este problema.

Debido a este problema decidimos adentrarnos y realizar esta investigación proponiendo una solución utilizando técnicas de lenguaje natural sobre los tweets de los usuarios para realizar una clasificación e identificar los usuarios humanos y los bots.

Anteriormente participamos en PAN 2019 [5] para CLEF 2019 [1], con la tarea "Bots and Gender Profiling" donde mostramos una solución formada con estructuras de Ngramas, en particular usamos bigramas de carácter; de manera que para esta tarea decidimos usar la misma metodología; utilizando una estructura de Ngramas mucho más amplia [9].

2. Metodología

2.1. Conjunto de datos

Para realizar la investigación usamos el corpus de PAN 2019 [5], el cual consta de 3000 usuarios; los cuales; 1500 usuarios son bots y 1500 usuarios son humanos, esto para el idioma español. Para el idioma inglés usamos contamos con 4000 usuarios de los cuales 2000 son bots y los otros 2000 son humanos. Cada usuario es representado por un archivo XML el cual contiene 100 tweets de máximo 140 caracteres. Este corpus lo usamos únicamente para entrenar el modelo.

Tabla 2. Evaluación de clasificación entre bots y humanos en inglés en la tarea de PAN 2019 [5] con el corpus de entrenamiento.

Método de clasificación	1-grama	2-grama	3-grama
J48	63.44	65.54	71.29
NaiveBayes	66.48	69.25	69.25
RandomForest	83.21	85.55	83.24
RandomForest	86.79	83.24	89.31
SVM	90.42	**91.86**	89.41

2.2. Pasos del pre-procesamiento

Consideramos indispensable crear una capa de pre-procesamiento para cualquier investigación relacionada en el lenguaje, ya que en nuestro caso esta capa ayudó a tener mejor precisión para la clasificación. Los pasos del pre-procesamiento son los siguientes:

- Se cambió todo el texto a minúsculas para todos los conjutos de datos.
- Los signos de puntuación fueron removidos.
- Los dígitos fueron removidos.
- Los links y url fueron removidos directamente.
- Las menciones también fueron removidas sin ser etiquetadas.
- Los emoticones son removidos sin ser etiquetados.

2.3. Caracterísiticas

Ya que tenemos los datos preprocesados, pasamos a formar las características que nos ayuden a crear un modelo donde podamos clasificar dichos usuarios. Como parte de la investigación, decidimos entrar al concurso de PAN 2019, el concurso consistía en realizar 2 clasificaciones: Dado el conjunto de datos, se debía clasificar si un usuario era bot o humano, después ya identificados los humanos, la próxima tarea era clasificar por el género.

Dentro del concurso nos fueron asignados algunos recursos tecnológicos muy limitados por lo cual no podíamos sobrepasarlos ya que si lo hacíamos, los procesos daban por terminada la tarea y salimos sin resultados del concurso; por lo tanto; teníamos que crear un modelo que fuera limitado pero eficaz para dicha clasificación.

En nuestro caso usamos únicamente bigramas de caracteres. Utilizando los bigramas de caracteres para ambos lenguajes arrojaron 84.13 % de precisión para la clasificación entre bots y humanos evaluados por el comité del concurso.

Debido al desempeño que mostraron los bigramas, decidimos incrementar las características con diferentes combinación de Ngramas. La configuración dependiendo del lenguaje la podemos ver en la Tabla 1.

Estas características fueron seleccionadas de esta manera debido a que mientras realizábamos los experimentos e íbamos agregando o modificando la combinación de Ngramas, notábamos que bajaba la puntuación de la clasificación; como por ejemplo;

Tabla 3. Resultados de precisión evaluados en corpus de entrenamiento por PAN 2019 [5].

Clasificador	Español	Inglés
J48	77.98	83.90
NaiveBayes B	84.83	87.22
RandomForest	89.98	92.31
SVM	**94.13**	**96.04**

cuando agregábamos Ngramas de tamaño 4, la precisión bajaba un 12 %, de esta manera nos dimos cuenta que mientras hacíamos más grandes los Ngramas; tanto de caracteres como de palabras; era peor la precisión de la clasificación.

Después de obtener las combinaciones de Ngramas donde obtuvimos los mejores resultados, estos son contados y se obtiene la frecuencia de aparición dentro de texto entre todos los tweets de los usuarios. Con estas frecuencias agrupadas por cada usuario podemos crear vectores los cuales si juntamos a todos los usuarios del mismo idioma crearemos un modelo de espacio vectorial donde cada dimensión corresponde con un usuario en particular.

2.4. Modelo de espacio vectorial

De esta manera representamos los tweets en un módelo de espacio vectorial [6], donde cada columna es representada por una formación de Ngramas de los tweets, llenando así la matriz con todas las formaciones de Ngramas seleccionadas. Cada usuario es representado por una nueva dimensión en la matriz.

Si existe un Ngrama que no este dentro la matriz de un usuario, esta característica será incluida para todas las matrices y llenada con 0, esto es importante debido a que pueden ayudarnos a diferenciar los datos dentro del algoritmo clasificador.

3. Experimentos

En la parte experimental usamos distintos algoritmos los cuales puedan identificar similitudes entre los dos tipos de usuarios que tratamos de clasificar. Como habíamos comentado; anteriormente participamos en el concurso de PAN 2019 y seleccionamos la estructura de bigramas de caracteres debido a su buena precisión con el clasificador Support-vector machine.

Podemos ver en el Cuadro 4, los experimentos con las diferentes configuraciones, todas estas pruebas fueron evaluadas con K-fold Cross Validation con un K=10. En ese momento usamos este verificador debido a que no teníamos acceso al conjunto de datos de prueba. Los resultados para el idioma español se muestran en el Cuadro 2 y para el idioma inglés se encuentran en el Cuadro 3.

Con respecto al concurso [5], el ganado fue Juan Pizzarro [4], donde seleccionó una arquitectura de Ngramas de caracteres y de palabras. Su resultado para la clasificación de bots en español fue 93.33 % y en inglés fue 93.60 %, estos resultados fueron evaluados con por el cómite de PAN [5].

Continuando con la investigación, obtuvimos el conjunto de prueba de PAN 2019 [5]. Después seleccionamos esta metodología debido a que la combinación de Ngramas mostrada en el Cuadro 1, incrementaban la precisión de la clasificación.

Los resultado mostrados en el Cuadro 4 muestran diferentes clasificadores los cuales fueron evaluados únicamente con precisión en el concurso de PAN 2019 [5], resaltando que el mejor resultado obtenido fue utilizando Support-Vector Machine con la combinación de Ngramas del Cuadro 1. Con los procedimientos en esta investigación, logramos obtener mejores resultados del estado del arte, en este caso a el ganador [4] del concurso PAN de 2019 [5].

4. Conclusiones

Dados los resultados obtenidos en esta investigación, consideramos viable el uso de Ngramas para realizar una clasificación entre bot y usuarios humanos dentro de una red social. En este caso únicamente realizamos las pruebas con tweets debido al corpus que utilizamos. Planteamos realizar la misma investigación con otra red social ya que pensamos que se puede replicar esta misma metodología debido a que hacemos uso únicamente de texto y no de otro tipo de características particulares como los metadatos de los usuarios.

Algo sobresaliente que mostró esta investigación en el concurso de PAN 2019 [5] fue la precisión de la clasificación únicamente usando bigramas de caracteres para ambos lenguajes [9]. Podemos observar que ambos lenguajes guardan similitudes para este tipo de tareas y si se realiza investigaciones entre estos dos lenguajes se pueden usar metodologías muy similares para las clasificaciones.

Cabe resaltar que esta investigación fue usando únicamente tweets de Twitter, por lo cual nos parecería muy interesante replicar toda esta investigación para bots de otra red social, con esto podemos ver si existen similitudes y diferencias entre los bots de distintas redes sociales.

En la actualidad y debido a la era en la que vivimos consideramos importante continuar con mejoras de implementaciones y evolución de los bots, así mismo, llevar estas evoluciones de manera responsable ya que lo importante de este tipo de tecnologías es mejorar para ayudar y no obstaculizar o corromper a la sociedad misma.

Referencias

1. Cappellato, L., Ferro, N., Losada, D. E.: Working notes of conference and labs of the evaluation forum 2019 - conference and labs of the evaluation forum. In: Conference and Labs of the Evaluation Forum 2019 Working Notes. vol. 2380 (2019), ceur-ws.org/Vol-2380/
2. Davis, C. A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: BotOrNot: A system to evaluate social bots. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 273–274 (2016) doi: 10.1145/2872518.2889302
3. País, E.: Así fabrican los partidos políticos un trending topic, (2018)
4. Pizarro, J.: Using n-grams to detect bots on Twitter notebook for PAN at the conference and labs of the evaluation forum 2019. In: Conference and Labs of the Evaluation Forum 2019 Labs and Workshops, Notebook Papers (2019), ceur-ws.org/Vol-2380/

5. Rangel, F., Rosso, P.: Overview of the 7th author profiling task at pan 2019: Bots and gender profiling. In: Conference and Labs of the Evaluation Forum 2019 Labs and Workshops, Notebook Papers (2019), ceur-ws.org/Vol-2380/
6. Sidorov, G.: Syntactic n-grams in computational linguistics. Springer (2019)
7. Stella, M., Ferrara, E., De-Domenico, M.: Bots increase exposure to negative and inflammatory content in online social systems. In: Proceedings of the National Academy of Sciences. vol. 115, pp. 12435–12440 (2018)
8. Vives, J.: El coronavirus dispara el número de usuarios de twitter (2020)
9. Yacob-Espinosa, D., Gómez-Adorno, H., Sidorov, G.: Bots and gender profiling using character bigrams. In: Conference and Labs of the Evaluation Forum 2019 Labs and Workshops, Notebook Papers (2019), ceur-ws.org/Vol-2380/