Towards bots detection by analyzing the behavior of user data on Twitter

Francisco Moo-Mena[1], Sofía Robles-Sandoval[2], Karina González-Magaña[3], and Oliver Rodríguez-Adame[4]

[1] Facultad de Matemáticas, Universidad Autónoma de Yucatán
Mérida, Yuc., Mexico

[2] Universidad de Guadalajara
Guadalajara, Jal., Mexico

[3] Instituto Tecnológico de Ciudad Guzmán
Ciudad Guzmán, Jal., Mexico

[4] Instituto Tecnológico de Tepic
Tepic, Nay., Mexico

Abstract

Currently, social networks play an important role as a means of communication about various topics. In this way, this medium represents a very important source of data to know the opinions of its users on very diverse topics. However, the opinions expressed in this medium are exposed to the influence of specialized programs called bots. These bots are activated with the idea of influencing positively or negatively towards some point of view of the issues under discussion. When implemented through computer platforms accessible from any medium with Internet access, it is possible to access such content automatically through its APIs. Prior to an analysis of the opinions expressed in the social network, it would be highly recommended, as part of the process of debugging the data, some reliable bot detection mechanism. While there is still no optimized method for this task, this paper proposes a series of directives that can be considered in order to carry it out. As a case study, these directives are implemented on messages retrieved from Twitter, related to opinions about the candidates of the presidential election of Mexico in 2018.

Keywords: *Bots detection, social network, Twitter, Presidential elections.*

1. Introduction

Social networks have acquired great relevance in the dissemination of information and ideas, which have made them one more of the dissemination tools used by individuals and corporations. With the increase of information technologies and the rise of social networks, people spend more time on these platforms. Just in July 2018, the average number of daily tweets was 92,006 in Mexico. [1]. This means of communication also represents an excellent option to know the reactions of society to events of any kind. Considering this and taking into account its availability for most sectors of society, social networks have become easy targets for those who seek to manipulate or influence public opinion. Introducing, in this way, points of view and fictitious ideas not expressed by real people or institutions.

This intrusion is often done using the so-called bots, which are programs that publish on social networks in an automated way. To the best of our knowledge, there is still no optimized technique for automatic bot detection.

Among the most popular social networks are Twitter and Facebook. Through well-defined APIs, it is possible to access user data in these social networks. Given the enormous amount of data generated daily in these social networks, it is interesting to recover and analyze them to know trends in the opinions of millions of users on specific topics. However, to obtain reliable information, it is necessary to detect and eliminate data generated by bots in the data debugging stage.

In this work, after review of the literature, a series of measures related to the identification of bots are described. Being the Twitter API one of the most flexible to use, tweets are retrieved from thousands of users and the selected measures are calculated to try to determine which of the analyzed tweets were published by a bot and which were not. It should be noted that this first stage does not seek to define whether the bot is malicious or not. That is, if the bot's goal is to disorient or manipulate public opinion, since there are bots whose purpose is publicly known, such as those that publish news automatically. This will be used as a preliminary reference framework, or as a first filter of the tweets with a higher probability of having been published by bots.

The selected subject for the recovery of tweets was the opinion of users regarding the candidates of the 2018 presidential election in Mexico. The tweets were collected in the days close to the election in order to guarantee a broad production of them, considering the topic in question as a trend topic.

The rest of the paper is organized as follows: section 2 describes some important preliminary concepts to understand our proposal. Section 3 presents our proposal for the calculation of the measures associated with the detection of bots. Section 4 describes the experimentation carried out and the results obtained. Finally, we present the conclusions of this work, as well as the description of future works to extend it.

2. Preliminary Concepts

2.1 Twitter social network

Twitter is a micro-blogging service that allows users to connect with friends and other people by publishing short messages called tweets. Each tweet can contain up to 140 characters and text [2].

Some main concepts associated with this social network are the following [2]:

- **Retweet:** When a user decides to share another person's tweet with their group of followers.
- **Hashtag:** The symbol # is used in front of a word to categorize the message, and when clicked, you can read all the messages related to that topic.
- **Mentions:** Allows users to mention another user in the tweet using @ followed by the user's name. Users are identified by unique user names in the format @username. The user can also reply with @username and send messages. The user can respond to other users whether or not they are on their friends list. @username can be written anywhere in the tweet. Spammers also misuse this feature to send spam to other users. Therefore, according to twitter policies, if the message contains a large number of mentions and response labels, the user is considered a spammer [3][5].

2.2 Metadata on Twitter

Beyond the content of messages posted by users, Twitter carries a significant number of metadata that can be retrieved and analyzed. By combining such metadata one could obtain interesting relationships related to the detection of bots. Some of them are described below:

- **Semantics analysis:** To analyze the content of the tweets from a semantic aspect. This implies the sentiment analysis of the tweet, if it is a positive, negative or neutral opinion on a subject, the degree of contradiction in the tweet or if there is a polarity of feelings on the subject.
- **Syntax analysis:** To analyze the syntax of a tweet you can use the number of hashtags it contains, the number of mentions, the links or urls and the special characters.
- **User behavior:** In user behavior is taken into account the frequency of user posting. If the user repeats the same tweet on several occasions. If the user has changes in his feelings on a topic, the average number of tweets the user has. If the user has the geolocation activated, the number of followers that counts, the number of followings, the number of mentions, the date of creation of the account and the time of the tweets. Based on these attributes through formulas we can obtain other data such as: the user reputation, the age of the account and the followers ratio.
- **Followers ratio:** It is the relationship between the number of followers and the number of followings and is expressed as follows.

$$Followers\ ratio = \frac{\#\ of\ followings}{\#\ of\ followers} \quad (1)$$

- **Reputation:** It is the relationship between the followers with the followers plus the followings, expressed as follows.

$$Reputation = \frac{Followers}{Followers + Followings} \quad (2)$$

- **Average daily tweets:** It is obtained by dividing the number of tweets (statuses) by the number of days since the creation of the account. According to [4] spammers post tweets in a robotic way using the twitter API or a web interface, at regular intervals, and research shows that spammers are active at a specific time of the day. Moreover, the frequency of tweets is greater than that of a genuine Twitter user. The basic idea when including this feature is to detect automated behavior of spammers, while normal users show random behavior.

Before identifying the different ways in which we could detect bots, we must define what a bot is. Bots are programs that pass themselves off as humans and carry out publications in an automated way. On Twitter there are several bots: they can be malicious, such as those that carry out scams or spam, or they can be broadcast like those used to publish news. In many cases these malicious bots seek to influence the opinion of other users of the network. It is known that social networks are a good way to know the opinion of a group of people on a given topic. That is why detecting bots within a social network is important, as they can sometimes influence and create favorable or unfavorable trends on a topic. In the particular case of elections you can find bots whose objective is to discredit or favor a candidate, sometimes spreading false information.

3. Methodology

In order to make measurements aimed at detecting the presence of bots, a source from which we can obtain sufficient data is required in the first instance. For this work it was decided to gather information from the social network Twitter for the facilities it provides for researchers and programmers. As a case study, it was decided to analyze the tweets related to the accounts of the candidates for the Mexican presidency. This chosen theme guaranteed us an abundance of data, given that the samples were taken in the days close to the presidential election.

First, it was necessary to store tweets in a database. Being unstructured the nature of the data, we created a database using MongoDB. 1,349,442 tweets were obtained from the twitter API during some days before and after the elections in Mexico. For the collection of these tweets the code available at http://pythondata.com/collecting-storing-tweets-with-python-and-mongodb/, was used with some minor changes. After this, several publications focused on the identification of bots on twitter were reviewed, in order to provide the variables that are commonly considered as indicators of automated behavior by a user.

In the end, we established the following measurement variables as the most representative to link tweets with the activity of a bot:

1. Mentions,
2. Language,
3. Popularity,
4. Tweets average,
5. Tweets with URLs,
6. Tweets with hashtags,
7. Platform,
8. Followers ratio,
9. Account´s age,
10. Human capabilities versus bot activity.

Once these variables were obtained, it was decided to separate into different collections within the database, one for each candidate, in order to then graph the results and be able to perform an analysis.

To refer to the candidates, their full names were not always used, but rather the terms used by the press and society on a daily basis for each of them during the campaigns and elections were used. Meade to mention José Antonio Meade Kuribreña, AMLO for Andrés Manuel López Obrador, Anaya for Ricardo Anaya Cortés and Bronco or El Bronco to allude to Jaime Rodríguez Calderón.

4. Experimental Results

In this section we present the results obtained from the measured variables.

4.1 Mentions for each candidate

To begin with the analysis we decided to graph the percentage of tweets that mention each candidate and we obtained the following proportion (fig. 1).

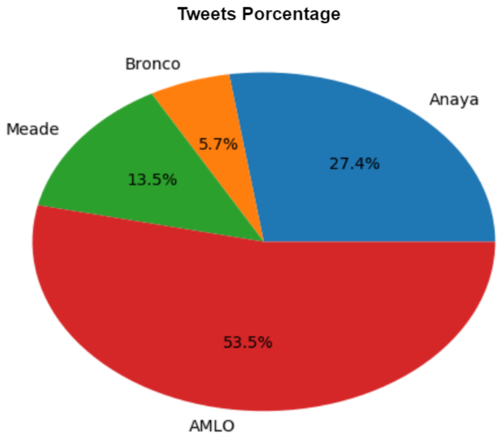

Fig. 1. Percentage of tweets that mention each candidate.

We found that Andrés Manuel López Obrador (AMLO) was mentioned in 397,983 tweets, Ricardo Anaya Cortés in 203,905, José Antonio Meade Kuribreña in 100,284 and Jaime Rodríguez Calderón "El Bronco" in 42,168. The remaining tweets are related to the elections, however, they do not mention explicitly any specific candidate. We found interesting the proportion of tweets obtained since it is very close to the proportion of votes obtained in the elections for the presidency, which was as follows:

53.19% for Andrés Manuel López Obrador
22.27% for Ricardo Anaya.
16.40% for José Antonio Meade and
05.23% for Jaime Rodríguez Calderón.

This data was retrieved in August 1, 2018 from https://centralelectoral.ine.mx/2018/07/06/da-conocer-ine-resultados-del-computo-de-la-eleccion-presidencial-2018/

4.2 Language

In the case of the language, we obtained the field of the profile of each user and we decided to classify them in 2 categories, Spanish and other. Tweets that come from a user with a language other than Spanish may be more likely to be a bot. The following graphs show the tweets in Spanish in orange and the tweets in other languages in blue. Four graphs are shown, the first (left) dedicated to the tweets of "López Obrador", the second those of "Anaya", the third to those of "Meade" and finally to those of the "Bronco" (fig. 2).

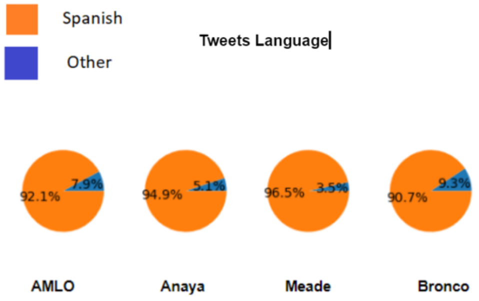

Fig. 2. Percentage of tweets in Spanish and other languages.

In the four cases, the proportion of tweets in other languages is small, the fact that they are in another language may be due to the news accounts of other countries that spoke about the elections in Mexico.

The following is an example tweet within the database that contains tweets that mention Meade:

"@aurelionuno Cínicazo!!!\n\nCORRUPTOS!!!\n#SomosPRI @PRI_Nacional \n#YoMero\n 🐀 🐀🐀🐀🐀\n#Fobaproa\n#Pemexgate\n#Monexgate \n#Odebrecht\n#EstafaMaestra\n#OHL\n#Chihuahua \n#LaCasaBlanca\n#LaEstafaBursátil\n#EtilenoXXI\n#PetroMansión"

Analyzing this tweet, we found that the location of the account is in Brazil and the determined language is Portuguese. The user is LavaJatoNews and seeing its description, we realize that it is a news bot in Brazil.

This is an example where we can identify a bot using the language tag.

4.3 Popularity

For popularity we use the reputation formula noted above, using the number of followers and followings.

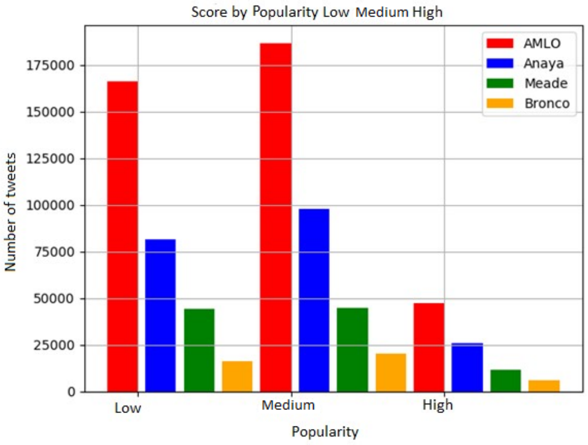

Fig. 3. Levels of popularity among candidates.

For the popularity we found that the values are between 0 and 1, for this reason we decided to group them in levels of popularity:

low (0-0.33),
medium (0.33-0.66) and
high (0.66-1).

Whereas when the popularity tends to 1 is more likely to be a real person, since usually the bots have a smaller number of followers (fig. 3).

An example is the account @Roberto73202922 that published the following tweet:

"Última encuesta Reforma:\n\nAMLO 51%\n\nAnaya 27%\n\nMeade 19%\n\nBronco 3%\n\nÚltima encuesta parametría:\n\nAMLO 53%\n\nAnaya 22%\n\nMeade 18%\n\nBronco 2%\n\nÚltima encuesta El Financiero:\n\nAMLO 54%\n\nMeade 22%\n\nAnaya 21%\n\nBronco 3%\n\n#AMLO imparable para ser el próximo presidente #JuegaMéxicoMX! https://t.co/ma9skEc8JR"

It has a popularity of 0.27, therefore it is in the low popularity category. Analyzing the account we can see that the probability of being a bot is greater, since the name of the account is long and with many numbers, which could be generated in an automated way. In addition, has a high tweets average since it generated 12,842 tweets in 56 days.

4.4 Tweets average

In the next graphs we can see the distribution of tweets averages from the different accounts that wrote about the candidates. For the calculation of the average, we divided the total of tweets between the number of days, since the creation of the account and the publication of the tweet. Accounts with a higher average number of tweets per day are more likely to be a bot since a feature of these is that they generate many tweets in a short time (fig 4).

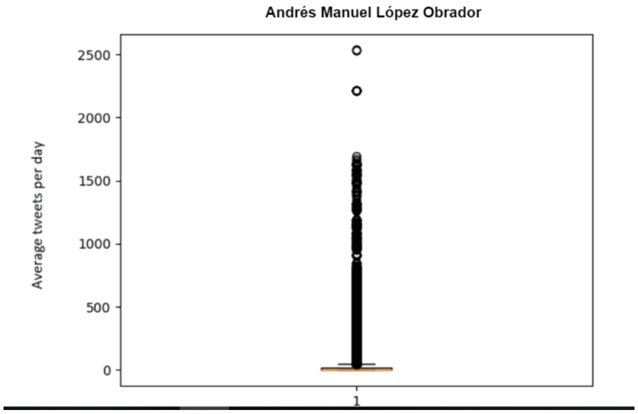

Fig. 4(a). Tweets average for AMLO.

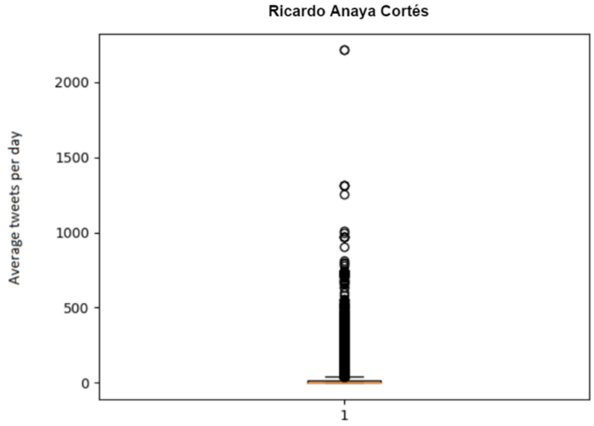

Fig. 4(b). Tweets average for Anaya.

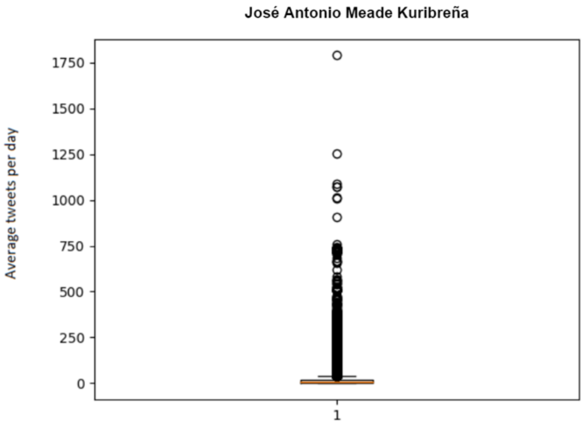

Fig. 4(c). Tweets average for Meade.

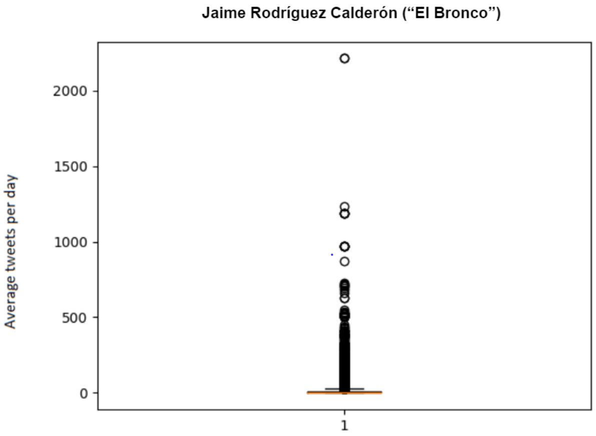

Fig. 1(d). Tweets average for El Bronco.

An example that we found was the @NewsBossIndia account with an average of 2,533 tweets per day. Being such a high number tells us that tweets are generated automatically. Analyzing the account we can see that it is a news bot. His tweet was the following:

> **"'AMLO' wins Mexican presidency.. https://t.co/BVX2zZn36m"**
> **URLS**

4.5 Tweets with URLs

One of the points that characterizes the tweets generated by a bot is the presence of URLs. Therefore, we decided to separate the tweets that have URLs from those that do not, and we found that the total of AMLO tweets containing URLs was 210,029, from Anaya was 101,488, from Meade was 54,810 and from El Bronco was 14,209 (fig. 5).

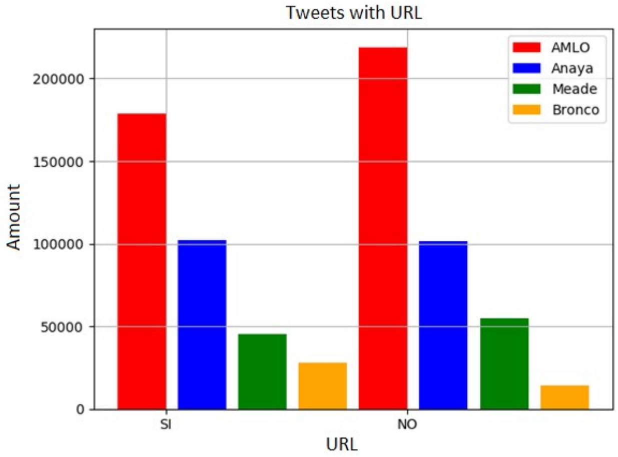

Fig. 5. Number of tweets found with URLs per candidate.

4.6 Tweets with Hashtags

Another variable that increases the probability that a tweet is produced by a bot is the presence of hashtags. Since they can search to generate a trending topic using a certain hashtag to influence the opinion of the users. From the tweets obtained, we found that 212,067 AMLO's tweets contains a hashtag, from Anaya 133,876, from Meade 43,908 and from El Bronco 31,731 tweets include hashtags (fig. 6).

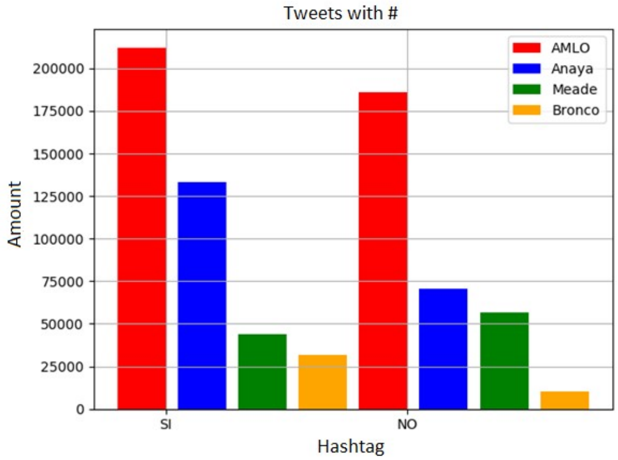

Fig. 6. Number of tweets found with hashtags per candidate.

4.7 Platform

The origin of the tweet is an important variable to take into account when detecting a bot. Since many times in this field we find that the tweet was published from an API or from an application destined to the publication of tweets in an automated way. It allows identifying with more precision the tweets that were published by bots.

Within the Source field of Twitter, we found that most of the tweets came from Android or iPhone mobile applications. However, we could find some that come from other sources, such as the following case (fig. 7).

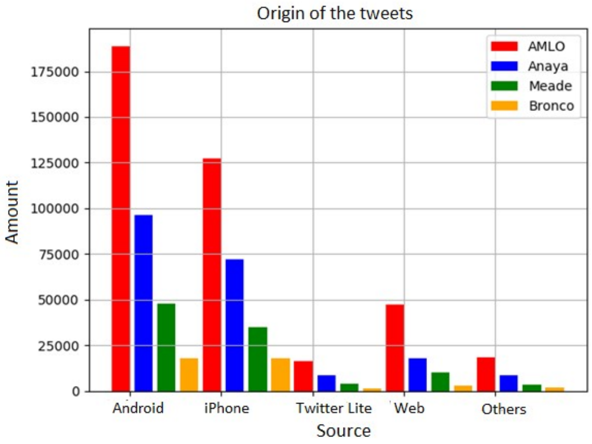

Fig. 7. Number of tweets found with URLs per candidate.

The @AMLONEWS account, in the source section has "AMLO bot political" and has an average of 315 tweets per day. This account is controlled by a bot and is dedicated to retweet and tweet about López Obrador.

4.8 Followers ratio

Another variable that can be applied using the followers and followings is the followers ratio. When it is greater than 1, it means that there is a greater probability that it is a bot. Since generally the bots follow a large number of users. However, they do not have a large number of followers (fig. 8).

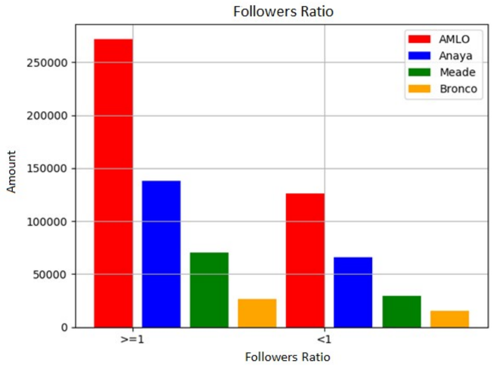

Fig. 8. Followers ratio of the candidates.

Analyzing the accounts, we find the following. @Roberto73202922 that has a followers ratio of 2.76, which indicates that it is more likely to be a bot. This counts 51 followers, and it follows 141 people. In addition, analyzing other parameters of the account, we can see that it has an average of tweets per day of 229, of which most are tweets about the López Obrador campaign.

4.9 Relationship of the account's age with the average number of tweets

In the following graphs we can visualize the accounts according to their number of days from their creation and the average of tweets. An account that has few days and a high average of tweets is more likely to be a bot. Since a significant percentage of bots are new accounts and they have a greater flow of tweets than a human user (fig. 9).

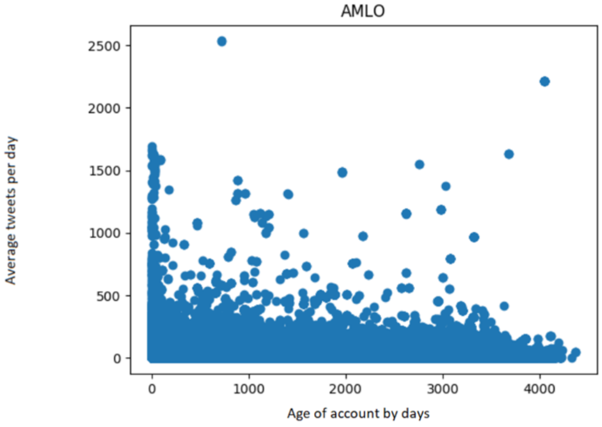

Fig. 9(a). Days old of the AMLO's accounts and their average in tweets.

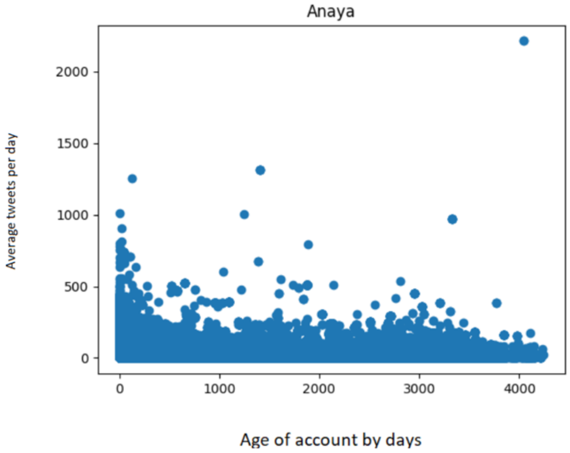

Fig. 9(b). Days old of the Anaya's accounts and their average in tweets.

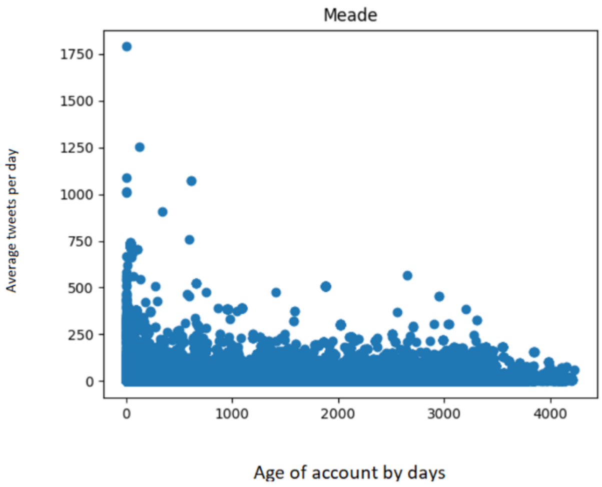

Fig. 9(c). Days old of the Meade's accounts and their average in tweets.

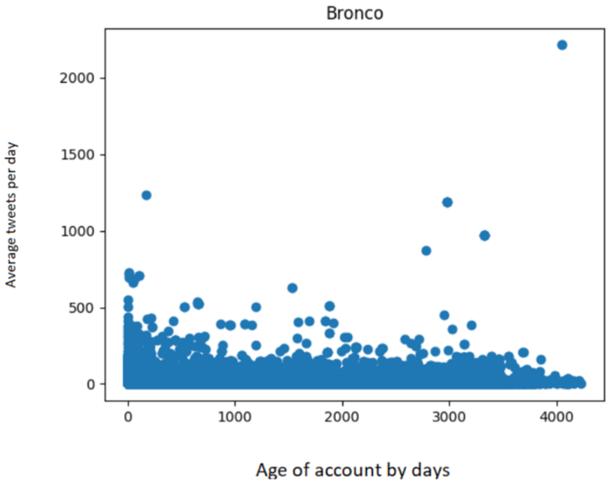

Fig. 9(d). Days old of the El Bronco's accounts and their average in tweets.

An example we found was an account with the name @DeMenK, which had 54,454 tweets in 35 days. It gives an average of 1,555 tweets per day, a very high amount that indicates a high probability in question of a bot. The text of the tweet is as follows:

> **“#MEXICO ABRE LOS OJOS\n\nNo te dejes engañar por\n#AMLO es el mismo #Socialismo que ha quebrado naciones ricas como #VENEZUELA Y #ARGENTINA\n\nSolo mira quienes lo apoyan, como dicen; mira con quien andas y te diré quien eres.\n\n#EleccionesMexico #amLOVEnezuela https://t.co/iXugaHgdUk”**

4.10 Human capabilities versus bot activity

In the study of the bots two types of goal activities were identified. One of them aims to make retweets and another to generate tweets of a specific topic in order to express an opinion, idea, etc. Given this, the following is analyzed.

Figure 10 represents the number of tweets that had 100 different accounts during the collection period. Given that the following is proposed to help detect a bot. Only in the sample that was taken were accounts that had more than 500 tweets and retweets. For example, there is an account named @Daniel that generated 418 retweets, that translated it represents 7667 words read, and posted 128 tweets that translated into words written were 1699. All this activity was generated in less than 6 days.

While the number of words tells us about a person's activity on an account, in addition with the average tweet per day we can see if the activity of an account can be humanly possible. An example we found was an account with the name @DeMenK, which had 54,454 tweets in 35 days, which gives an average of 1,555 tweets per day. If we convert it to words would be a total of 24,880 words in a single day. A person could not achieve such a large number of words written on a social network in a single day, considering it is not his only activity.

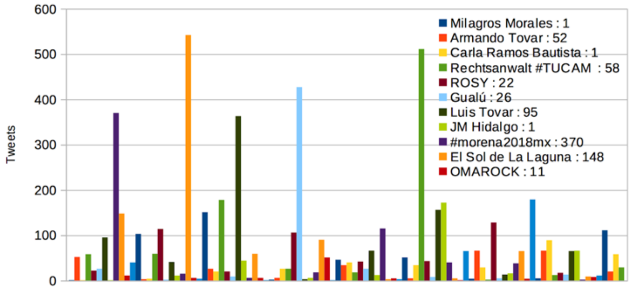

Figure 10. Number of tweets generated by one hundred accounts in the collection period.

5. Conclusion

In this paper we present a strategy aimed at the detection of bots in social networks. Our approach established a series of measures on data related to twitter accounts participating in the topics of interest. Based on the classification of the tweets and graphing the variables previously defined, we were able to realize for the tendency to be bot of the different tweets sent towards the candidates. There are variables that are more useful to identify a bot, for example, the average of daily tweets or the source of the tweet, since this allows us to quickly know if it is a bot or a person. This analysis can be applied when you want to know the opinions of people on a certain topic, allowing us to exclude tweets produced by bots to have a more reliable and realistic source of data. This work can be extended by applying sentiment analysis in the contents of the tweets, so that it is more accurate to detect bots. In addition, to know if the opinion in the tweet is positive, negative or neutral towards any of the candidates. This extension would define more precisely, which are the tweets, within the database, that were published by bots.

Acknowledgments

We thank the Universidad Autónoma de Yucatán, as well as the program "Delfin - Verano de la Investigación Científica y Tecnológica del Pacífico", Mexico, for all the support received to carry out this work.

References

[1] Instituto Nacional de Estadística y Geografía. *“Promedio diario de recolección de tuits en México”.* Instituto Nacional de Estadística y Geografía, INEGI, Mexico, 2018.

[2] Kaur, P., Singhal, A., & Kaur, J. (2016). “*Spam detection on twitter: a survey”.* 3rd International Conference on Computing for Sustainable Global Development (2016).

[3] Dickerson, J. P., Kagan, V., &; Subrahmanian, V. S. (2014). *“Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?”.* IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (2014).

[4] Velayutham, T., & Tiwari, P. K. “*Bot identification: Helping analysts for right data in twitter”.* 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA 2017).

[5] Mateen, M., Iqbal, M. A., Aleem, M., & Islam, M. A. “*A Hybrid Approach for Spam Detection for Twitter”.* 14th International Bhurban Conference on Applied Sciences and Technology (2017).

Francisco Moo-Mena is a Professor in Computer Sciences at Universidad Autónoma de Yucatán, in Mérida, Mexico. From the Institute National Polytéchnique de Toulouse, in France, he received a Master’s degree in computer science and a PhD, in 2003 and 2007, respectively. He also received another Master’s degree in

Distributed Systems from the Instituto Tecnológico y de Estudios Superiores de Monterrey, México, in 1997. His research interests include Parallel and Distributed Computing, Self-healing systems, Web services Architectures, Data Science applications**.**

Sofía Robles-Sandoval is currently studying Computer Engineering in Universidad de Guadalajara, Mexico. She has research interests on Big Data and Data Science applications.

Karina González-Magaña is currently studying Computer Systems Engineering in Instituto Tecnológico de Ciudad Guzmán, Mexico. She has research interests on Big Data and Data Science applications.

Oliver Rodríguez-Adame is currently studying Computer Systems Engineering in Instituto Tecnológico de Tepic, Mexico. He has research interests on Big Data and Data Science applications.