Pattern Identification of Bot Messages for Media Literacy

Eric Ferreira dos Santos
Federal University of Rio de Janeiro
Rio de Janeiro, Rio de Janeiro, Brazil
eric.ferreira@ppgi.ufrj.br

Danilo Silva de Carvalho
Federal University of Rio de Janeiro
Rio de Janeiro, Rio de Janeiro, Brazil
danilo.carvalho@ppgi.ufrj.br

Jonice Oliveira
Federal University of Rio de Janeiro
Rio de Janeiro, Rio de Janeiro, Brazil
jonice@dcc.ufrj.br

ABSTRACT

The massive use of online social media networks is a reality nowadays. Their increasing usage also raises growth in malicious activities in social media, one of which is the use of automated users (bots) that disseminate false information and can insert bias in analyses done on gathered social media data. Based on the concept of media literacy, this research presents a method to teach the human user to identify a pattern of a text produced by a bot, providing a tool (guide) to analyze social media text. Users who learned to identify a bot user with the guide had an average of 90% accuracy in the classification of new messages, against 57% of the participants who had no contact with the guide. The produced guide received a usefulness rating between 4 and 5 by the participants (scale from 1 to 5, with 5 being the highest value).

KEYWORDS

bot detection, message pattern, online social media, media literacy

1 INTRODUCTION

Online social networks (OSNs) became a relevant communication media and are used in many ways. They provide the ability to share videos, images, and news, which can reach many users around the world quickly and easily. The number of people who expose their lives on OSN sharing photos, visited places, political opinions, among others, has vastly increased over the last decade [13].

The massive use of OSN by more than 40% of the world's population [12] indicates the importance and the reach of this environment. Companies can advertise on OSNs, reaching different people on a large scale and evaluating their popularity.

Increasing people's usage also leads to the problem of rising malicious activity on social media. An example is the use of automated users to disseminate false or misleading information, affecting other network users because they are exposed to information that may interfere in their democratic, civil and behavioral processes [20].

These automated users can be used with various objectives, including (1) providing advertisements; (2) promoting politically oriented views and opinions; (3) promoting financial trends; (4) generating product reviews; (5) spreading malware (malicious software abbreviation), SPAM (unsolicited mass messaging), and harmful links; among other harmful and harmless activities, such as publishing news and weather reports in various channels. Such harmful activities can interfere with OSN user decisions and compromise any meaningful analysis of OSN data, since the data does not represent personal opinions or factual truth, and can bring panic or promote violence among the population [2], [19], [1].

The source identification and users' reliability are important resources to provide better information consumption and analysis of public opinion. In this direction, some efforts for automatic bot detection have shown up, using different syntactical, semantic and behavioral characteristics, such as textual style, network topology and users' interaction patterns. While improvements on bot detection and counter-detection have evolved like a cat-and-mouse game, automatic detection does not allow the ordinary user to learn how to identify a bot on their own. Consequently, their ability to form an opinion may be impaired.

For the regular user, identifying this kind of message is crucial in forming an opinion about a topic. With all those facts, the facilities in creating users in OSNs and the range a message can reach on the internet, a need arises for not only automatic identifying bots, but also teaching and raising awareness of the common user to identify them and not share this type of message in the OSNs environment.

Media literacy is an area that discusses the ability to access, analyze, evaluate and create messages in various contexts [14]. This area has studied how a media message needs to be sent and how the receptor will understand it. The user should be able to analyze and evaluate the media content, pondering the message's relevance and confidence.

Seeking to clarify bot's pattern of textual features, this work aims to present an easy way for human users to identify a bot message through the textual characteristics on OSN, and use them as input features in bot detection models already in existence.

1.1 RESEARCH QUESTIONS & OBJECTIVES

Taking into account the motivation and previous research the following research questions were developed:

- Research Question 0 (RQ0): What are the textual patterns of a bot message? Do the messages follow some construction format for a specific topic?
- Research Question 1 (RQ1): How much does it differ from human message patterns?
- Research Question 2 (RQ2): How can the user be educated to spot suspicious messages? Are there differences between an educated and uneducated user?

The main objective of this research is to find textual patterns in bot messages, for selected topics in OSN, and generate a user guide for the common user to identify them. To achieve this objective data was collected from OSN in chosen topics, and textual patterns from bot messages were compared with human messages in English and Portuguese languages. Finally, a guide is generated to present the patterns learned to the ordinary user.

In: III Concurso de Teses e Dissertações (CTD 2021), Minas Gerais, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2021.

ISSN 2596-1683

1.2 CONTRIBUTIONS

We argue that this work is aligned with the directives given from the Ministry of Science, Technology, Innovations and Communications (Brazil) [21], such as Technologies "II – Habilitadoras" and "IV - para Desenvolvimento Sustentável". Regarding the Grand Challenges in Computer Science Research in Brazil [16], defined by the Brazilian Computer Society, we can highlight "2. Computational modeling of complex systems: artificial, natural, socio-cultural, and human-nature interactions".

We can highlight as scientific contributions the use of pattern recognition and model explainability to aid media literacy to combat misinformation, one of the main challenges of modern society with impacts on democracy, health and public safety.

The master's thesis generated two publications, a preliminary paper ([8]) was published on the FATES'19 workshop, and a final one, containing the entire approach, was accepted for publication on the main track of this event [7]. The research involved two interns, students from the Federal Center for Technological Education (CEFET-RJ) from Friburgo, contributing to the relationship between school and university and supporting the scientific development of the countryside regions from State of Rio de Janeiro. In addition, this dissertation helped in international collaboration, considering that research groups from Dublin City University (Ireland) and the Knowledge Media Institute/Open University (United Kingdom) are interested in getting involved in the continuation of this work.

This thesis has some technical contributions. This research produced some technology artifacts, such as a dataset with more than 100,000,000 messages about Brazilian presidential elections (2018), a classifier model over this dataset on bot or human messages, and the process code that others can use for further research. Furthermore, the technical knowledge acquired during the research was partially transferred to the market, considering that the student was a part-time student and worked in the industry throughout his master's degree.

For social contributions, this research points to the importance of being aware of social media information, which became a source of fake news and misinformation, and tries to present to the final user how to deal with it. This proposal is currently being applied for the COVID-19 scenario, a problem that the world is facing now. Due to the good preliminary results, this thesis was incorporated and adapted to the extension project "Fighting misinformation about COVID-19 through media literacy", a project selected and approved by the Dean of Extension of UFRJ.

With the motivation defined, the remainder of the paper is organized as follows. In section 2, we present the proposed method, based on the objectives of this work to answer the research questions. In section 3 the experiment based on the proposal is demonstrated and its results are compared with the related works. Finally, section 4 concludes this paper, summarizing the research, its limitations and future works.

2 PROPOSED METHOD

This section presents the proposed process. In the first step of the process, the data is collected from the OSN and messages are classified using the Botometer API [22], generating the output to the next step (Figure 1a). The next step receives the data and splits into different languages, due to particularities, and generates an n-gram model to select a topic for messages that do not have one explicitly (hashtags). Each topic is composed of a cluster of similar hashtags and generates the input (textual features) for the next step (Figure 1b). After that, a decision tree is trained for each cluster using the messages' textual features (Figure 1c). With the trained classifiers, a user guide can be generated for any given message, to elucidate how the user can verify if the message is from a bot or not (Figure 1d).

The datasets used are: Arab Spring in Libya (2011 - 2013) [18] and Brazil Presidential Election (2018). The first dataset has messages in English labeled as bot or human. The other was collected using the Twitter API during the election event, using keywords related to the candidate names, and the messages labeled by the API provided by [22]. This API provides a percentage value for the chance that a user is a bot. Similar to the work of Gilani et al. [9], where the authors selected that a percentage greater than 50% would mean that the user is a bot, and otherwise a human, this research uses the same metric in a balanced dataset.

To proceed with the class analysis, we selected textual features to train a decision tree model classifier. Since one of our contributions is a user guide on how to detect a bot message, this approach was selected due to the ease of interpreting the resulting model. We developed a way of translating the learned model criteria into human interpretable sentences to generate the user guide.

The proposed research aims to study bot and human message patterns in the Twitter OSN, understanding how the messages are created in a specific topic. English and Portuguese messages are analyzed and their characteristics are compared.

A set of relevant textual features indicated by the systematic mapping is initially used. The most used are the number of words, number of characters, number of URLs, number of hashtags, number of mentions and number of special characters [3–6, 9–11, 15, 17, 18, 22]. In those works, at least one of those features was used in some way and is used in this work. However, only these features are not enough to distinguish between bot and human.

Considering that the datasets are labeled, a supervised learning algorithm is applied on the selected features to construct a classification model. The algorithm used is a decision tree classifier, similar to our previous work [8] and due to the ease of interpreting the model. Since our main contribution is a system to guide the user on how to detect a bot message, this approach was selected due to the ease of interpreting the resulting model.

The user guide is generated following the path that the features go through in the decision tree. A simple explanation is created to teach the user how it can be possible to analyze a message. An example can be seen in the message: "@_luaazevedo Você tem muito talento e merece. Acho até que @jairbolsonaro poderia usar essa versão do hino na campanha. https://t.co/BjtM0qolJF " (text in Portuguese), which based on this message generates the following explanation: "The message has at least 1 link(s), indicating a bot message, and having more than 1 twitter user(s), indicating a human message. In this context, even with a certain discordance, the message can be classified as a human message."

Taking the decision tree path, the message that the proposed literacy guide shows is translated into Figure 2.

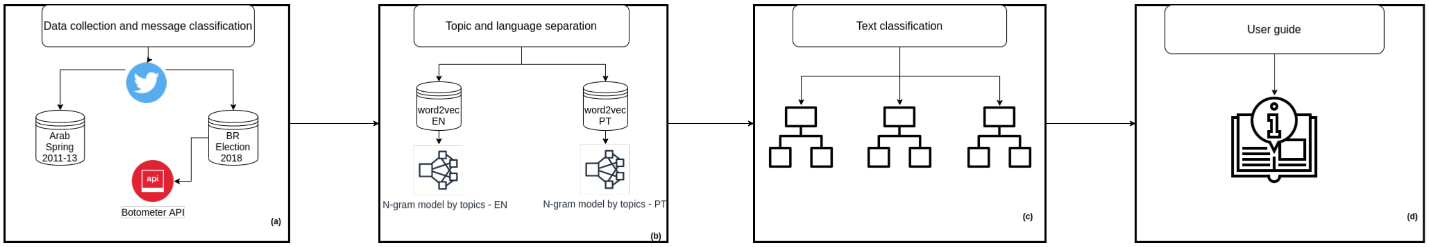
Data collection and message classification
Arab Spring 2011-13
BR Election 2018
api
Botometer API
(a)
Topic and language separation
word2vec EN
word2vec PT
N-gram model by topics - EN
N-gram model by topics - PT
(b)
Text classification
(c)
User guide
(d)

Figure 1: Proposal Process

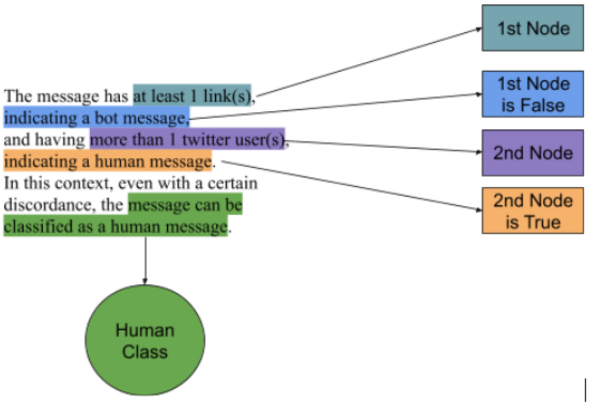
The message has at least 1 link(s),
indicating a bot message,
and having more than 1 twitter user(s),
indicating a human message.
In this context, even with a certain
discordance, the message can be
classified as a human message.
1st Node
1st Node is False
2nd Node
2nd Node is True
Human Class

Figure 2: Explanation for the guide message

We traversed the decision path "translating" the attribute names to more understandable descriptions, and used a simple node separation syntax, as the Figure 2 shows.

3 EXPERIMENTATION

A small review of the experimentation is presented in this section, but the focus is in the literacy guide evaluation and the research question answers.

The messages were separated in clusters, following the topic relation. After that, each cluster messages were used as input in the classification model, generating different classifiers for each cluster.

To answer the RQ3 - "How can the user be educated to identify suspicious messages? Are there differences between an educated and uneducated user?", we created a quasi-experiment, where two groups participated. The first group received the media literacy approach generated in this research and the second group (control group) did not. The evaluation result is present in the next section.

3.1 LITERACY GUIDE: EVALUATION EXECUTION

The experimentation evaluation was conducted between Dec 11, 2019 and Jan 03, 2020, according to the volunteers availability. Each participant was randomly set to a different group and received the form according to the group. All the participants had up to 30 minutes to fill up the form.

Looking at all the participants, the 47 people were distributed randomly in the two groups to participate in the evaluation. Most of the participants were from high school and the second group of subjects was composed from undergraduate students and the majority of the participants are from the computer science area, followed by journalism.

The experiment indicates that the participants gather information daily, which includes the online newspapers, conversation with friends and other sources. The news shared in OSN was the second communication way used by the participants. The last question in the forms general section, about the bot identifications, catches some attention because no one uses any developed tool for this task yet.

The feedback collected from the group that received the form presenting the approach developed in this work was positive. Almost 80% evaluated the explanations between grades 4 and 5 of utility (scale from 1 to 5, where 1 is the lowest value), with the average grade 4.04 and standard deviation of 0.78.

Comparing the results of the classification (last section in the forms) between the two groups, the one who received the explanations got the best performance. The average of correct answers was almost 90%, while the group that did not receive the proposed approach obtained an average of 57% correct answers. This result can point out that the people who received the model path explanation could classify a new message correctly by themselves in almost all cases, which validates the objective of this work.

Another comparison is the difference between the better and the worst performance in each group. In the first group, the best performance was 10/10 (10 correctly answers from 10 questions) and the worst was 7/10. In the second group, the best performance was 10/10 too, but the worst was 3/10. Although the two groups had the maximum hit, the difference between the performance of the lowest hits between the groups illustrates that those who received the media literacy approach obtained more homogeneous results.

3.2 Answering the Research Questions

The RQ0 and RQ1 are related to finding the message patterns in the selected topic for each class presented in the problem. These two RQs could be answered after interpretation of the emerging patterns in the text classification step (Fig. 1(c)), where the bot message features are analyzed for each topic and the differences between bot and human messages are exposed, such as differences between the use of vocabulary or grammar patterns.

RQ2 is related to how the human user can interpret the patterns found. This RQ could be answered after the creation and application of the user guide (Fig. 1(d)) to help the typical user to distinguish a bot message in the OSN. A survey was conducted (see Section 3.1) to verify the guide's impact on users, and this part helps the users to have a healthy interaction with the OSNs, which is proposed in media literacy.

4 CONCLUSION

As mentioned earlier in this work, online social media has become a source of information for many people. Fast and free access to information has allowed it to reach the population with ease. It has also been observed that with the increasing use of OSNs, attacks that users are exposed to have also increased, such as viruses or misinformation that can compromise the decisions users can make.

The major works that deal with bot classification are concerned in creating a better model that identifies the bot users automatically and fast, not taking into account of how the OSN user will understand the classification result. The scientific contribution from this research is the media literacy approach, which differently from the other approaches, tries to teach a regular user how to classify a possible bot message in OSN manually. The results observed from the conducted experiments indicate that the proposed user guide was evaluated satisfactorily and the people who received the literacy guide achieved better results when compared to those who did not receive the guide.

The process from collecting information, separating topics within context, creating classifications and explaining models can be replicated for various scenarios and also improved.

The process developed here is also presented as technical contribution, as each part can be modified by an alternative approach. The data source may differ, as well as its classification, the way the word model is constructed, the way other messages are sorted by topic, and the algorithm used for classification. All those steps can be modified to improve the approach and produce better results.

There are some limitations in this work that point to future improvement. It was based on the premise that the Botometer API is accurate in classifying users, which may present problems, especially in the classification of users who only write in Portuguese. Maybe using another classification source could bring confidence to the message classification. Working with textual features only (syntactic and lexical) is a complex task and impacted in the models' evaluation metrics and the textual characteristics that were selected, being a limitation to this research.

Future work based on this research would not only seek new contexts but propose other techniques to reach a better result. Perhaps most important would be the search for new textual features that best represent the message patterns or other features that can be easily used by the typical user to identify a bot.

Creating solutions that can evolve automatically or semi-automatically is important, so we can keep the users of our guide up-to-date with the new bots' behaviors.

REFERENCES

[1] 2017. "Relembre casos de violência provocados por boatos na rede" in Portuguese. https://blogs.oglobo.globo.com/eissomesmo/post/relembre-casos-de-violencia-provocados-por-boatos-na-rede.html

[2] 2018. Nigerian police say "fake news" on Facebook is killing people. https://www.bbc.co.uk/news/resources/idt-sh/nigeria_fake_news

[3] Muhammad Al-Qurishi, Majed Alrubaian, Sk Md Mizanur Rahman, Atif Alamri, and Mohammad Mehedi Hassan. 2018. A prediction system of Sybil attack in social network using deep-regression model. *Future Generation Computer Systems* 87 (2018), 743–753.

[4] Abdulrahman Alarifi, Mansour Alsaleh, and AbdulMalik Al-Salman. 2016. Twitter turing test: Identifying social machines. *Information Sciences* 372 (2016), 332–346.

[5] Mansour Alsaleh, Abdulrahman Alarifi, Abdul Malik Al-Salman, Mohammed Alfayez, and Abdulmajeed Almuhaysin. 2014. Tsd: Detecting sybil accounts in twitter. In *2014 13th International Conference on Machine Learning and Applications*. IEEE, 463–469.

[6] Isaac David, Oscar S Siordia, and Daniela Moctezuma. 2016. Features combination for the detection of malicious Twitter accounts. In *2016 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*. IEEE, 1–6.

[7] Eric Ferreira Dos Santos, Danilo Carvalho, and Jonice Oliveira. 2021. Pattern Identification of Bot Messages for Media Literacy. In *WebMedia '21: Brazilian Symposium on Multimedia and Web Proceedings*.

[8] Eric Ferreira Dos Santos, Danilo Carvalho, Livia Ruback, and Jonice Oliveira. 2019. Uncovering Social Media Bots: a Transparency-focused Approach. In *Companion Proceedings of The 2019 World Wide Web Conference*. 545–552.

[9] Zafar Gilani, Ekaterina Kochmar, and Jon Crowcroft. 2017. Classification of twitter accounts into automated agents and human users. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. 489–496.

[10] Rodrigo Augusto Igawa, Sylvio Barbon Jr, Kátia Cristina Silva Paulo, Guilherme Sakaji Kido, Rodrigo Capobianco Guido, Mario Lemes Proença Júnior, and Ivan Nunes da Silva. 2016. Account classification in online social networks with LBCA and wavelets. *Information Sciences* 332 (2016), 72–83.

[11] Mücahit Kantepe and Murat Can Ganiz. 2017. Preprocessing framework for Twitter bot detection. In *2017 Int. Conference on Computer Science and Engineering (UBMK)*. IEEE, 630–634.

[12] Simon Kemp. 2019. Report: Social media use is increasing despite privacy fears. https://thenextweb.com/contributors/2018/04/17/report-social-media-use-is-increasing-despite-privacy-fears/

[13] Bernardo Pereira Lauand and Jonice Oliveira. 2014. "Inferindo as Condições de Trânsito através da Análise de Sentimentos no Twitter" in Portuguese. *iSys-Revista Brasileira de Sistemas de Informação* 7, 3 (2014), 56–74.

[14] Sonia Livingstone. 2004. Media literacy and the challenge of new information and communication technologies. *The communication review* 7, 1 (2004), 3–14.

[15] Juan Martinez-Romo and Lourdes Araujo. 2013. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications* 40, 8 (2013), 2992–3000.

[16] Claudia Bauzer Medeiros. 2008. Grand research challenges in computer science in brazil. *Computer* 41, 6 (2008), 59–65.

[17] Amanda Minnich, Nikan Chavoshi, Danai Koutra, and Abdullah Mueen. 2017. BotWalk: Efficient adaptive exploration of Twitter bot networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. 467–474.

[18] Fred Morstatter, Liang Wu, Tahora H. Nazer, Kathleen M. Carley, and Huan Liu. 2016. A new approach to bot detection: Striking the balance between precision and recall. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*. IEEE, 533–540. https://doi.org/10.1109/ASONAM.2016.7752287

[19] The Star Online. 2018. When fake news sparks violence: India grapples with online rumours. https://www.thestar.com.my/tech/tech-news/2018/07/16/when-fake-news-sparks-violence-india-grapples-with-online-rumours/

[20] Mariam Orabi, Djedjiga Mouheb, Zaher Al Aghbari, and Ibrahim Kamel. 2020. Detection of Bots in Social Media: A Systematic Review. *Information Processing & Management* 57, 4 (2020), 102250. https://doi.org/10.1016/j.ipm.2020.102250

[21] Marcos Pontes. 2020. PORTARIA Nº 1.122, DE 19 DE MARÇO DE 2020. https://www.in.gov.br/en/web/dou/-/portaria-n-1.122-de-19-de-marco-de-2020-249437397

[22] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107* (2017).