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
Oswaldo Cruz Foundation
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 is a reality nowadays. Such an 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).

CCS CONCEPTS

• **Information systems** → **Social networking sites**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

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

ACM Reference Format:
Eric Ferreira dos Santos, Danilo Silva de Carvalho, and Jonice Oliveira. 2021. Pattern Identification of Bot Messages for Media Literacy. In *Brazilian Symposium on Multimedia and the Web (WebMedia '21), November 5–12, 2021, Belo Horizonte / Minas Gerais, Brazil.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3470482.3479452

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) and Fiocruz - Finance Code 001.

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 [22].

The massive use of OSN by more than 40% of the world's population [21] 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 [30].

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; (6) influencing search engine results in a way that particular links are shown first; (7) generating news feeds; and (8) creating an underground marketplace for purchasing social media followers [6].

Not all automated users (bots) are maleficent, some bots spread messages from newspapers and weather applications, as an automatic service to publish the same news in various channels. Unfortunately, OSNs often face the challenge of dealing with undesirable users and their malicious activities. The most common form of this kind of activity over OSNs is spamming, wherein a bot disseminates contents or malware/viruses to legitimate users of the social networks [10].

Rumors and untrue pieces of information are currently called fake news and their propagation is also related to automated users, which brought a new problem to the OSN environment [4], [2], [31]. Such 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 [3], [29], [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. Automatic identification in order to exclude bots has been an arduous task. While the automatic methods are improved, the bots are becoming more sophisticated, creating a conflict between both the creators of the methods and bots. Furthermore, the automatic deletion 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

WebMedia '21, November 5–12, 2021, Belo Horizonte / Minas Gerais, Brazil

ACM ISBN 978-1-4503-8609-8/21/11…$15.00
https://doi.org/10.1145/3470482.3479452

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 [24]. This area covers how the media will be accessed (e.g., by television or internet), how the message will be analyzed according to the reader's previous knowledge, how it will be evaluated and how a new message will be sent forward. 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 AND OBJECTIVES

Taking into account the motivation and previous research (that will be presented in Section 2) 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 the 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.

With the motivation defined, the remainder of the paper is organized as follows. In section 2, we present an overview of related works that present approaches to deal with the problem cited in this section and can be compared with this research. In section 3, we present the proposed method, based on the objectives of this work to answer the research questions. In section 4 the experiment based on the proposal is demonstrated and its results are compared with the related works. Finally, section 5 concludes this paper, summarizing the research, its limitations and future works.

2 RELATED WORK

In this section, related works which aim to bring an approach for bot detection are briefly discussed.

[10] presents a community-based framework to identify bots in an online social network. The study aims to improve bot detection using the density-based overlapping community, where the community evolution is tracked and the user's relationship within the community is analyzed, including topological features. Decision trees, naive Bayes, ADTree and k-NN algorithms were chosen as classifiers. The authors concluded that a user can be considered a bot if sends a message to a user in another hierarchy level within the network. This investigation established that decision-tree and ADTree had a better performance in both datasets. Future works proposed to use this framework in real-world applications, where creating nodes would be unnecessary.

[26] proposed an approach to identify not bots, but bot messages. This follows the authors' premise that text characteristics are more difficult to be manipulated by bots than account features, such as the number of followers and friends, account creation dates and others. Statistical analysis of language was applied to detect spam messages in Twitter trend topics. Their study introduced an architecture that collects trend topics from Twitter API, labelling the messages, extracting textual features, training a classifier and detecting a spam message. Statistical analysis of language plays a vital role in this work. There are three text units in use: i)a set of messages related to a trending topic, ii)a suspicious message, and iii)a page linked to the suspicious message. The method is based on term distribution Kullback-Leibler Divergence. Decision trees, naive Bayes, logistic regression, support vector machines (SVM), decorate and random forest algorithms were applied for classification. Analyzing messages that have a divergent language is another contribution of this work, which does not treat a message as spam just because it has a link.

In [6, 7] studies, the authors created a Twitter dataset with bots, humans and hybrid accounts. This dataset contained nearly 1.8 million accounts, manually labeled by three volunteers. These volunteers could categorize the accounts as English and Arabic content, so the other users were ignored. The analysis of the bot accounts to discover an appropriate set of features to be used in the classification step was another contribution. They included: i)number of hashtags per tweet, ii)number of times a hashtag has been used, iii)number of links, iv)whether the profile picture contains a face, or it is the basic Twitter profile picture, v)mentions of different users within the same text, vi)number of lists in which the user is listed on are examples of selected features. The authors used four machine learning algorithms for the classification step: Decision Tree, Random Forest, Support Vector Machines (SVM), and multilayer neural network. For evaluation, this study considered two scenarios: two classes and three classes. In the first scenario, there were only two categories: human or bot. In the other, there were three: human, bot and hybrid, which is a bot account that is controlled by a human to mislead bot seek algorithms. The multilayer neural network was more precise in both cases.

There are more recent works that use other machine learning techniques, such as deep learning, to classify bot users [23, 25, 34]. These new approaches improve the model classification, but do not present an explanation about how the model learns.

The variety of approaches used in these studies and others are important to create a consistent classifier model to be run by an automated system, but such classifications are not simple for a typical user to understand. A media literacy approach together with a classification method could bring a simpler way of explaining the human/bot choice, although it may not surpass an automatic classification system performance in identifying a bot user or message in the OSNs.

We published a study that combines bot message classification and an approach to generate an user-guide for lay people to understand how the classification was done. In [17], we developed an analysis tool, based on media literacy considerations, that helps the typical user to recognize a bot message using only textual features. Instead of simply classifying a user as a bot or human, this tool presents an interpretable reasoning path that helps to educate the user into recognizing suspicious activity. For the best of our knowledge, this is the first work that combined the bot message classification and a media literacy approach to explain the process.

3 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 [33], 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) to 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 next subsections will explain each step of the process proposed in this work.

3.1 THE DATASET

The first step of the proposed process is to collect data from a source that will be used as the first input. In this research the source chosen was the Twitter OSN in two different events.

The datasets used are: Arab Spring in Libya (2011 - 2013) [28] and Brazil Presidential Election (2018). The first dataset has the messages in English labeled as bot or human. The other was collected using the Twitter API during the election event, and the messages labeled by the API provided by [33]. This API provides a percentage value for the chance that a user is a bot. Similar to the work of Gilani et al. [18], 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.

From the Morstatter et.al. [28] dataset, we decided to select for each bot a group of human users for which messages most closely resemble the bots textual subjects. In this way, the comparison would be ideally done under textual cues that are less related to the topic, since the topic would be the same for both classes. The idea behind such alignment of topics is that by isolating the "topic feature" – a meta feature for the distribution of words in a message – the remaining textual features would be more easily captured by a Machine Learning classifier, improving classification performance. On the other hand, such restriction of the word distribution also limits the ability to capture other possible textual features that are also unrelated to the topic. A semantic method was selected to determine word similarity, since it presented relevant results in previous works [16], [8] and [32].

To proceed with the 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 use of only textual features is a limiting factor regarding model accuracy, but provides way of classification that is reproducible by a human user with no further tools, given enough explanation about the path taken in the tree to reach a decision. With the decision tree simplification, some features showed to be more important than others.

3.2 ANALYSIS & CLASSIFICATION OF MESSAGES

The next two steps of the process is presented in this subsection.

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 the behaviors 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 [5–7, 14, 18–20, 26–28, 33]. 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.

To combine the above-mentioned features, some works proposed sentiment features to better identify the classes [5–7, 33]. In this work, we utilized the concept of positive, negative or neutral message to enrich the classification and report bot and human message pattern. Varol et al. [33] and Al-Qurishi et al. [5] introduced part-of-speech tags in the classification model, which can explain the message composition, allowing the syntactic and semantic analysis. The message construction pattern is a feature to be analyzed in this research as it can collaborate in the identification phase. Table 1 presents the features selected to conduct this research.

Topics on Twitter OSN are marked by a hashtag (#) prefixed word in the message, but not all messages have this mark. The approach from Carvalho et al. [12], where an n-gram model is utilized to improve the legal text classification, is used in this research. An n-gram model is built for each topic/hashtag set, identified by a clusterization of semantically similar hashtags.

The tokenization and the word model creation have the goal to construct clusters that split the words into similar topics. With these clusters, it could be possible to separate the messages by similarity, even if the message has no word in the topics set.

Considering that the datasets are labeled, a supervised learning algorithm is applied to construct a classification model. The algorithm used is a decision tree classifier, similar to our previous work [17] and due to the higher interpretability of the resulting model. Since our main contribution is a system to guide the user on how to detect a bot message, this approach was selected to facilitate the translation of the resulting tree into a guiding sentence for the user. The decision tree algorithm implementation is CART3. Using only textual features is a limiting factor regarding model

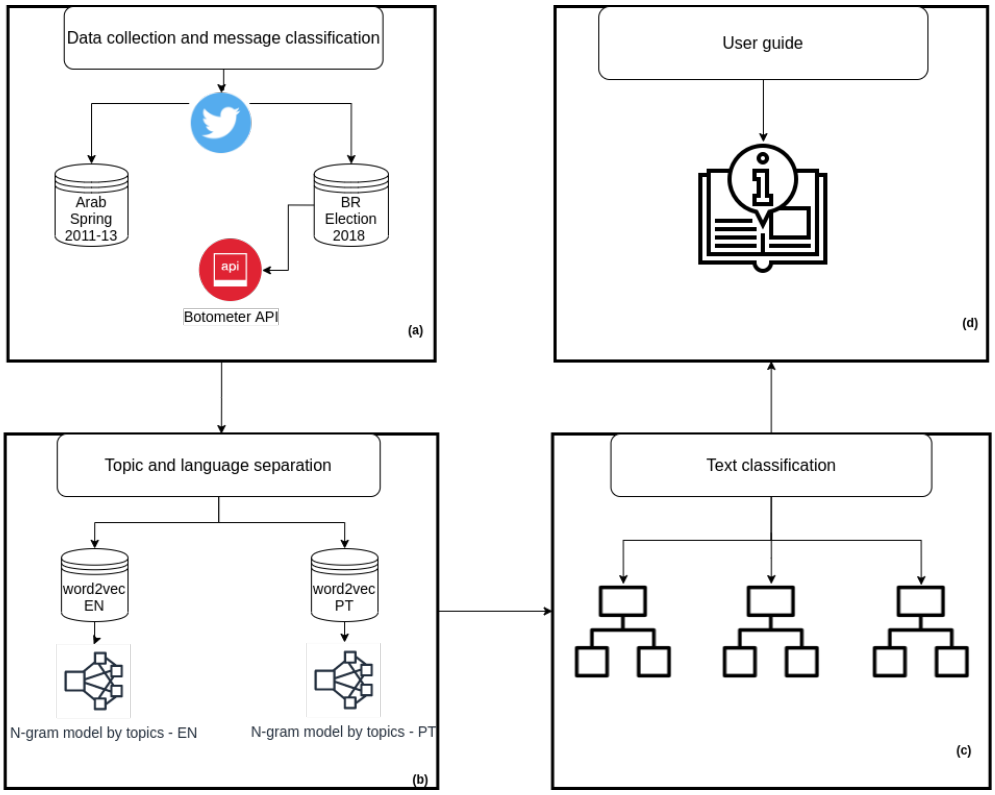

Figure 1: Proposal Process

Features that are used in this work	Description
Hashtags(#)	The number of hashtags present in the message
Link	The number of links present in the message
Retweet (RT)	The number of re-published messages
Mentions(@)	The number of mentions in the message
Emoji	The number of emojis in the message
Text length	Message length (# chars)
Punctuation	The number of punctuation in the message
Upper Case	The number of uppercase letters in the message
Sentiment	Type of sentiment (Positive, Negative, Neutral)
Spelling mistakes	The number of spelling mistakes in the message
POS-tags types	Types of POS-tags present in the text
Number of stop-words	The number of stop-words present in the text
Number of words	The number of words present in the text

Table 1: Textual features analyzed

accuracy but provides a way of classification that is reproducible by a human user with no further tools and even outside the original OSN (Twitter), given enough explanation about the path taken in the tree to reach a decision.

3.3 THE LITERACY GUIDE

As the main contribution of this work, an automatically generated guide is proposed to assist the typical user in manually identifying whether a message is from a bot or human user. This guide is the result of post-processing the model obtained by a machine learning algorithm, that presents the textual features that were determinant in the classification of a given message, in each topic and language, for user appreciation. This is the last step of the proposed process.

The user guide is generated following the path that the features go through in the decision tree (Figure 2). 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."

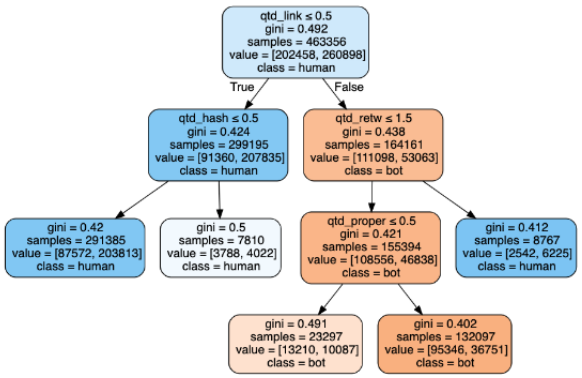

Figure 2: Decision tree generated

The explanatory message was identified as present to the first cluster, created on the Portuguese corpus and the explanation was created upon the decision tree path referent to this cluster. Taking the decision tree path, the message that the proposed literacy guide shows is translated into Figure 3.

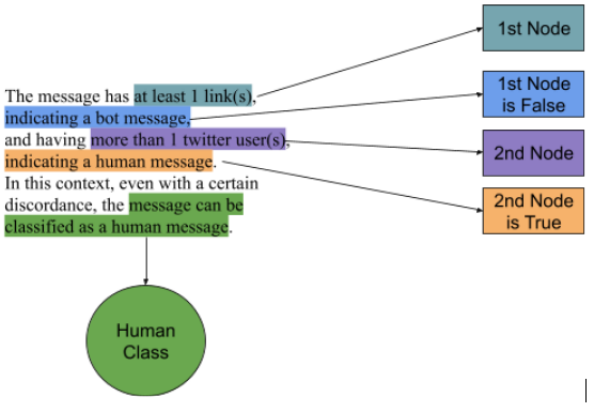

Figure 3: Explanation for the guide message

We traversed the decision path "translating" the attribute names to more understandable descriptions, and used a simple node separation syntax, as the Figure 3 shows.

Following the path, it can be noticed that the message has a link which sends to the right side of the tree. After that, the number of users cited on the message is more than one and the classification path goes to the right side of the tree, finishing the process. After that, the program exhibits that in this specific context the message can be classified as a human message.

It is important to highlight how the program explained the classification. The example has presented a discordance and the program signals this. If the message only had a link, the classification would be different. But it is interesting that the program shows the classification for each feature used in the decision tree.

3.4 DEVELOPMENT

This subsection will describe the tools used to develop the process presented above. The first step was the dataset collection from the Twitter OSN and the data was accessed by the Twitter API through a python script developed for this research. For both datasets, a python script uses the filters given from the API and returns the information in JSON format. The JSON was stored in a non-relational database (MongoDB). From these data collected from Twitter, it was selected the unique user's names to get the Botometer classification.

In each cluster (Figure 1c), there is a set of more frequent n-grams (in this work are used 3, 4 and 5 grams), and these frequencies are used to classify a text into a certain topic. The term frequency - inverse document frequency (TF-IDF).

To generate the user guide based on the decision trees trained, a python script was developed to go through the trees and create phrases according to each level of the classifier as described in the subsection above. The idea is to provide the user guide result in an API service, serving a variety of end-user interfaces, as presented in the previous subsection. To achieve this result, host services need to be evaluated as well as the front-end technologies.

4 EXPERIMENTATION

The experimentation conducted in this research is described here. It covers the steps presented in Figure 1. In this section, the proposal is properly used with real data to evaluate the approach, given results that help to understand what works and what can be improved. Each subsection explains the sequence presented in the 1.

First, we present the data collection and how the messages were classified1(a). Next comes the process to separate the messages by languages and topics, before training a classifier for each topic 1(b).

Then, we present the messages classification step 1(c) and the guide construction 1(d). This last step is broken into two parts: how it is evaluated and its results.

4.1 DATA COLLECTION AND MESSAGE CLASSIFICATION

As cited before, in this work was utilized two datasets: Arab Spring [28] and Brazil Presidential Election of 2018. The first dataset is already classified, and [28] provided a list which contains the user identification from Twitter OSN and the label, whether bot or human. This dataset has 230.145 messages and their classifications.

The second dataset was collected in 2018, between June and October, during the presidential election in Brazil on real time and the keywords used were related to the candidate names. 106.532.062 messages were collected in this period and were selected as the unique users in these messages to get the classification through the Botometer API[1].

From a total of 3.407.154 unique users found in the collected messages, 842.474 were classified as bot or human according to the result returned from Botometer. With the users classified, 24.953.837

[1]https://botometer.osome.iu.edu/

messages were selected, which belong to these users. For dataset balancing, an undersample technique was used.

4.2 TOPIC AND LANGUAGE SEPARATION

In this step, every task is the same for each language and it is explained once. With the data selected in the previous step, it was necessary to create a model to represent the messages. The choice was to transform words into vectors, to find a similarity between the words in the vector space.

By collecting all hashtags present in the entire corpus and the word-to-vector model, an unsupervised learning algorithm (k-means) was executed to cluster similar ones. Each cluster contains a set of hashtags, and these clusters represent a topic. The hashtags were distributed by clusters, creating topic clusters. The next task was to collect the messages that contain the same hashtag for each cluster and create an n-gram model. 3, 4 and 5-grams were used for the model construction.

After that, all the messages that do not contain hashtags were classified in each cluster according to its n-grams. This classification task used the TF-IDF approach, which selected a cluster for which the n-gram was most related.

At the end of this step, all the messages are distributed among the clusters, which would become inputs to the decision tree algorithm. Each cluster produces a separated classifier.

4.3 MESSAGE CLASSIFICATION

Receiving the messages from each cluster, the next step consists in transforming the text into a set of features that are used in the decision tree classifier. Table 1 presents the textual features that were used. All the features were converted to discrete values, which improved the classification rate.

After a model is created, the important features from that model (characteristics that were used in the model) were gathered and the model is trained again using only those features. This process was done to generate a compact model, aiming an easily-readable user guide program to be developed. The models were evaluated by mean of the metrics generated in each cross-validation iteration. The metrics selected were accuracy, precision, recall and F1 score.

Nevertheless, the main contribution of this research is the explanation for the classification path, contributing to the common user guidelines, which other works did not address. It is important to remember that in this work only textual features were utilized, which is different from the other related research.

The metrics achieved in this work are closer to the ones in Martinez-romo, J. and Araujo, L. [26]. As in this research, the authors there used only content features to the classifiers, but in a more controlled scenario and the F1 measure is lower than the obtained in this research.

From the [26] work, it is seen that the authors used feature combinations and powerful classifiers to conduct this task, with the aim of creating better models. These works have no intention to explain how the classification process behaves. Starting from [26] as a baseline, this research intended to find what could be done to improve the classifier not only in terms of performance, but mainly in terms of interpretability, renouncing in this process most classification performance gains obtained in the last 7 years.

This research, as explained earlier, has no intention to overcome these results but to create an approach to explain how the classification task was conducted to typical users. To achieve this objective, we opted on using an algorithm that does not bring the best result but offers a way to understand the classification task and present this to the people.

The result generated by this research is relevant because it is one of the only works (so far found) that takes into account the media literacy approach in the features selection to the classification process.

4.4 LITERACY GUIDE: EVALUATION DESIGN

The main contribution of this work is an easy explanation of how a typical user can identify a possible bot in a political topic on OSN, improving literacy in this media type. As the principal focus of media literacy is the people, who consume the information, the approach developed in this work must be evaluated by humans. This section is to answer the RQ3 - How can the user be educated to identify suspicious messages? Are there differences between an educated and uneducated user?

Aiming to answer this research question, a quasi-experiment was created, 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 groups had no contact with each other and the participants did not know which group they belonged to, since each person was associated with a group randomly. All the participants did not know what method that would be evaluated, nor the content that would be presented.

The first group received the literacy guide, which explained all textual patterns of bots and how to recognize them. Then, the users classified new messages following the directions described on the guide. The second group did not receive the guide. They had classified a set of messages using only their own knowledge and previous experience. Both groups received the same set of messages, which means that all the messages were evaluated by both groups.

The goal of this evaluation was to verify how the methodology adopted here impacted the participants. For this purpose, we developed two quantitative measurements to evaluate the proposed method: How the explanations generated by the model are evaluated by the users? (E1) This measurement was designed to show how the users interacted with the guide, whether it was easy to understand and useful and; Did the person learn how to classify a message by himself/herself? (E2) This is a metric that rates the success in manual classification of the messages. In this case, both groups are evaluated and the main objective is to compare the assertiveness between them.

Aiming to collect data, an online form was created to evaluate the proposed model. The form is separated into 3 sections: General questions, Media Literacy, and Message classification. The first section was created to know the subjects participating in the evaluation, having general questions such as genre, study field, knowledge on the OSNs and other questions. The second section had the message examples taken from Twitter OSN and how the proposed model classified them. With this, the person could be educated in how he/she can identify the possible message source. The person who

receives the form with this section, would have a set of 20 messages from Twitter, and their explanations from the guide. The last section aims to evaluate the degree of learning on the subjects through the model proposed, where each person classifies 10 new messages.

The messages presented in the media literacy section and the messages selected to be classified in the form were chosen randomly from the datasets used in this research.

4.5 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 degrees 4 and 5 of utility (scale from 1 to 5, where 1 is the lowest value).

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 by themselves a new message correctly 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.

There are some limitations in this proposal. Starting with the chosen scope, which was the political topic in a determined period. It is a limitation because it does not ensure that the result will be the same or close in another scope or another period. A set of different scopes needs to be used in the approach developed in this research to generate insights about how stable the process is.

A limitation in the model evaluation by humans was the number of participants, their study fields and schooling degrees. Despite having a relevant number of participants, statistically, it is not conclusive. To overcome this problem, more people, from different know-how backgrounds, must complete the forms to evaluate the proposal and bring new insights into further improvements.

4.6 Answering the Research Questions

In this subsection we will explain how the research questions were answered.

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 were answered in the 1(c) step, 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. With the classification model built, some patterns have emerged and could be interpreted.

RQ2 is related to how the human user can interpret the patterns found. This RQ is answered in the 1(d) step, where a "guide" to help the typical user to distinguish a bot message in the OSN, was created. A survey was conducted (see Section 4.5) 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.

5 CONCLUSION

As presented 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. To achieve this it was necessary to select features that the typical users can analyze without a tool, and can classify the message source by themselves. 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.

This work presents a classification approach based on textual characteristics that can be extended and applied in other contexts. 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. One limitation that the work has is related to the first step of user classification as bot or human. 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 use another classification source, such as "DeBot" [13] or "SaraBotTagger" [9], could bring confidence to the message classification.

Working only with textual features (syntactics 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.

Recurrent Neural Network or another Deep Learning technique could be used for building classification models, improving not only the classification results but removing the need to choose features manually.

In this work we used word2vec as the word model, but for the further works other can be used, such as "fastText" [11] or Bidirectional Encoder Representations from Transformers "BERT" [15]. This change would bring new word representations, and the topic separation and the classification steps would be affected.

Keeping up with the evolution of these patterns, as bots will continue to evolve in an attempt to deceive the average user, is a relevant subject for this research. 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. How is Fake News Spread? Bots, People like You, Trolls, and Microtargeting. http://www.cits.ucsb.edu/fake-news/spread

[3] 2018. Nigerian police say "fake news" on Facebook is killing people. https://www.bbc.co.uk/news/resources/idt-sh/nigeria_fake_news

[4] 2018. Russian hackers targeted Tumblr during the US election. https://www.independent.co.uk/life-style/gadgets-and-tech/news/tumblr-russian-hacking-us-presidential-election-fake-news-internet-research-agency-propaganda-bots-a8274321.html

[5] 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.

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

[7] 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.

[8] Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 809–815.

[9] Carlos Magno Geraldo Barbosa, Lucas Gabriel da Silva Félix, Antônio Pedro Santos Alves, Carolina Ribeiro Xavier, and Vinícius da Fonseca Vieira. 2020. SaraBotTagger-A Light Tool to Identify Bots in Twitter. In *International Conference on Complex Networks and Their Applications*. Springer, 104–116.

[10] Sajid Yousuf Bhat and Muhammad Abulaish. 2013. Community-based features for identifying spammers in online social networks. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. IEEE, 100–107.

[11] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[12] Danilo S Carvalho, Minh-Tien Nguyen, Chien-Xuan Tran, and Minh-Le Nguyen. 2015. Lexical-morphological modeling for legal text analysis. In *JSAI International Symposium on Artificial Intelligence*. Springer, 295–311.

[13] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. 2016. Debot: Twitter bot detection via warped correlation.. In *Icdm*. 817–822.

[14] 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.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Cicero Dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *International Conference on Machine Learning*. PMLR, 1818–1826.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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/

[22] 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.

[23] Greeshma Lingam, Rashmi Ranjan Rout, and Durvasula VLN Somayajulu. 2019. Adaptive deep Q-learning model for detecting social bots and influential users in online social networks. *Applied Intelligence* 49, 11 (2019), 3947–3964.

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

[25] Linhao Luo, Xiaofeng Zhang, Xiaofei Yang, and Weihuang Yang. 2020. Deepbot: a deep neural network based approach for detecting Twitter bots. In *IOP Conference Series: Materials Science and Engineering*, Vol. 719. IOP Publishing, 012063.

[26] 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.

[27] 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.

[28] 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

[29] 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/

[30] 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

[31] Raj Samani and August 31. 2018. The anatomy of fake news: Rise of the bots. https://www.helpnetsecurity.com/2018/08/31/fake-news-bots/

[32] Miran Seok, Hye-Jeong Song, Chan-Young Park, Jong-Dae Kim, and Yu-seop Kim. 2016. Named entity recognition using word embedding as a feature. *International Journal of Software Engineering and Its Applications* 10, 2 (2016), 93–104.

[33] 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).

[34] Feng Wei and Uyen Trang Nguyen. 2019. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. In *2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, 101–109.