Social Bots Detection: A Method based on a Sentiment Lexicon Learned from Messages

Samir de O. Ramos, Ronaldo R. Goldschmidt and Alex de V. Garcia
Section of Computer Engineering (SE/9), Military Institute of Engineering (IME), Rio de Janeiro, Brazil

Keywords: Social Bot Detection, Machine Learning, Natural Language Processing, Sentiment Analysis.

Abstract: The use of bots on social networks for malicious purposes has grown significantly in recent years. Among the last generation techniques used in the automatic detection of social bots, are those that take into account the sentiment existing in the messages propagated on the network. This information is calculated based on sentiment lexicons with content manually annotated and, hence, susceptible to subjectivity. In addition, words are analyzed in isolation, without taking into account the context in which they are inserted, which may not be sufficient to express the sentiment existing in the sentence. With these limitations, this work raises the hypothesis that the automatic detection of social bots that considers the sentiment characteristics of the words of the messages can be improved if these characteristics were previously learned by machines from the data, instead of using manually annotated lexicons. To verify such hypothesis, this work proposes a method that detects bots based on Sentiment-Specific Word Embedding (SSWE), a lexicon of sentiment learned by a homonymous recurrent neural network, trained in a large volume of messages. Preliminary experiments carried out with data from Twitter have generated evidence that suggests the adequacy of the proposed methodand confirms the raised hypothesis.

1 INTRODUCTION

The use of social bots - accounts controlled by software that algorithmically generate content and establish connections - in order to deceive and influence other users of a social network (Messias et al., 2013) is a fact known and studied by the scientific community (Lee et al., 2011), (Tang et al., 2014a) (Ferrara et al., 2016). Such accounts are able to behave simi- larly to human users, and can be used to disseminate misleading advertisements and information, leading to error and false perception of reality (Freitas et al., 2014).

Evidences suggest that an increasing amount of social media content is generated by social bots (Statista and partners, 2018), confirming the urgent demand for solutions that automatically detect bots on social networks.

According to (Adewole et al., 2017), one of the main approaches to automatic detection of social bots is based on the analysis of behavioral characteristics or attributes of accounts, such as, for example, number of followers and followers, number of messages posted, sentiment expressed in these messages, among others. Many researches in the area follow this approach, using machine learning algorithms to build classifiers capable of distinguishing legitimate accounts from social bots (Ferrara et al., 2016), (Wagner et al., 2012)(Velázquez et al., 2017) (Kudugunta and Ferrara, 2018) (Davis et al., 2016) (Alarifi et al., 2016) (Zhang et al., 2016) (Grimme et al., 2017) (Clark et al., 2016) (Ramalingam and Chinnaiah, 2018) (Subrahmanian et al., 2016) (Gilani et al., 2017) (Adewole et al., 2017).

As far as it was possible to observe, the works that use sentiment present in messages posted by the accounts among the behavioral attributes for detecting social bots are based on sentiment lexicons whose content was manually noted by humans (Ferrara et al., 2016), (Varol et al., 2017b)(Davis et al., 2016).

Like any process that involves human interpretation, such labeling is susceptible to subjectivity. In addition, in this process, words are analyzed in isolation, without taking into account the context in which they are inserted, which may not be enough to express the sentiment in the message.

In view of these limitations, this article raises the hypothesis that the automatic detection of social bots that considers sentiment characteristics of the words of the messages can be improved if these characteristics were previously learned by machines from the

Ramos, S., Goldschmidt, R. and Garcia, A.
Social Bots Detection: A Method based on a Sentiment Lexicon Learned from Messages.
DOI: 10.5220/0011115000003179
In *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022) - Volume 1*, pages 273-280
ISBN: 978-989-758-569-2; ISSN: 2184-4992

data, instead of use lexicons annotated manually by humans. Such a hypothesis is justified, since, according to (Tang et al., 2014b), there are sentiment classification models with a history of success in several applications and whose characteristics for data representation were learned from large volumes of data.

That said, the objective of this article is to obtain experimental evidence that the detection of social bots that considers sentiment characteristics of messages can be improved if these characteristics are learned from large text bases. For this purpose, a method of detecting social bots is proposed, which allows the calculation of sentiment characteristics present in messages from the sentiment-specific word embedding (SSWE), a lexicon of sentiment learned bya homonymous recurrent neural network trained ina large historical volume of messages (Tang et al., 2014a). The proposed method combines the sentiment information extracted from the messages with other behavioral attributes associated with the accounts and applies machine learning algorithms to classify accounts into bot and not bot. Experiments carried out in a popular database in social bots detection studies generated preliminary evidences that confirm the hypothesis raised, among them the improvement of the accuracy of the detection models bymore than 5 p.p.

The article has five other sections. Section 2 presents the theoretical foundation necessary to understand the concepts used in this article; Section
3 describes related work; Section 4 the proposed method; Section 5 reports the experiments and the results obtained; Section 6 highlights the contributionsof the article and points out some future works.

2 BACKGROUND

The VAD Lexicon - VAD is the lexicon resulting from the work described in (Warriner et al., 2013) and consists of a list of 13,915 words of the English language and their respective values of valence (pleasantness of a stimulus, ranging from pleasure to dis- pleasure), arousal (intensity of emotion caused by a stimulus, ranging from calm to excited) and dominance (degree of control exercised by a stimulus, ranging from weak/submissive to strong/dominant), characteristics that form a three-dimensional hyper- space to represent sentiments. It was built from questionnaires filled out by humans (Bradley and Lang, 1999).

The calculation of the VAD average (or sentiment) of a message consists of the weighted average of the values of valence-arousal-dominance for each word of the message found in the lexicon. The weighting factor is the frequency of the word in the message.

The SSWE Lexicon - SSWE, in turn, is the lexicon obtained by the works reported in (Tang et al., 2014a) (Tang et al., 2015) and consists of the vocabulary described by 50 sentiment dimensions (sentimental embeddings) learned by a homonymous recurrent neural network trained from 10 million Twitter texts (ie, tweets) with sentiment (positive or negative) labeled using emoticons.

The SSWE average calculation follows the same formula used for the VAD average calculation. The only difference is that it was applied to the fifty existing dimensions while, in the case of VAD, to only three.

In resume, as reported by (Tang et al., 2014a), SSWE is a method that learns word embedding to classify message sentiment. Most existing algorithms for learning continuous word representations usually only model the syntactic context of the words, but ignore the sentiment of the text. This is problematic for sentiment analysis, as they usually map words with similar syntactic context, but opposite in sentiment, such as ”good” and ”bad”, to neighboring word vec- tors. This problem is solved by learning specific- sentiment word embedding (SSWE), which encodes the sentiment information in the continuous representation of words, so that it is able to separate the ”good” and ”bad” at opposite ends of the spectrum. To this end, the existing word embedding learning algorithm developed by (Collobert et al., 2011) and three neu- ral networks were developed from the text’s sentiment value (for example, phrases or tweets) in its lossfunctions. The specific-sentiment word embedding is learned from tweets, taking advantage of massive tweets with emoticons (see Box 3 below) as *corpora* remotely supervised, without any notes manuals(Tang et al., 2014a).

Positive Emoticons	Negative Emoticons
:) :) :-) :D =)	:(: (:-(

The first advantage of using sentiment characteristics from the data is the low cost of the process, since there is no need to hire note labelers. In addition, it does not involve subjectivity; it is also not limited to a predefined set of words. Finally, it takes into account the semantics of the relationship between words, since SSWE originates from a dynamic process of word embedding that captures, through neural networks, the sentiment information.

Experiences in the application of SSWE for a set

of sentiment classification data from Twitter reference in the task competition called SemEval 2013 it shows that: (1) the SSWE resource has a performance comparable to the resources made by hand in the best performance system; (2) the enhancement is further enhanced by concatenating SSWE with the existing feature set (Tang et al., 2014a).

3 RELATED WORKS

This article considers papers that, in some way, build (or cite) automatic bot detection models that include sentiment analysis in their scope, even if this construction is not the main objective of these works.

One of the most relevant and comprehensive work in this area and that meets the requirements presented is that of Ferrara et al., 2017, which warns of the massive presence of bots in social networks, especially those created to harm, adulterate, manipulate and deceive social media users. These malicious bots have been used to infiltrate political discourse, manipulate the stock market, steal personal information and spread misinformation. It also presents a taxonomy of the different social bots detection systems proposed in the literature, dividing them into network-based techniques, crowdsourcing strategies, supervised learning based on resources and hybrid systems (a combination of the previous ones).

Although there are several studies in the area of automatic detection of social bots, those that have shown the most promising performance are those that follow the analysis approach based on binary classification models learned from behavioral attributes ac- counts (Varol et al., 2017b). Among them are those that use sentiment lexicons to measure the sentiment of messages posted by the accounts. The advantage of the behavioral patterns approach is that they can be easily encoded in characteristics and adopted us- ing machine learning techniques to learn the patterns of human and bot behavior.

This allows later to classify accounts according to this behavior. To capture orthogonal dimensions of behavior, different classes are commonly employed, including: network, user, friendships, timing, content and sentiment. The latter is the focus of this work and includes happiness, arousal-dominance-valence and emoticon scores.

An example of a feature-based system is ”Bot or Not?” (Davis et al., 2016). Developed in 2014 for Twitter, this system implements a detection algorithm based on highly predictive features that capture a variety of suspicious behaviors and separate social robots aspects of humans. It uses ready-to-use supervised learning algorithms, trained with examples of human behavior and bots, based on the Texas A&M 24 dataset, which contains 15,000 examples from each class and millions of tweets. Experiments with the “Bot or not?” system report detection above 95%, measured by AU-ROC via cross-validation.

Table 1 presents a comparison of these works considering the following characteristics:

- (P1) Behavioral Attributes;
- (P2) Sentiment analysis;
- (P3) Sentiment from annotated lexicon;
- (P4) Sentiment learned by machines from data.

As far as it was possible to observe, none of the papers (among those that use sentiment analysis) currently existing in the literature explicitly considers sentiment characteristics of the words of the messages learned from the data to detect social bots.

In view of the above, it was identified that, among the related papers, none explicitly considers the sen- timent characteristics of the words of the messages learned from the data to detect social bots.

Table 1: Resume of Related Works.

Reference	P1	P2	P3	P4
(Wagner et al.,)	-	X	-	-
(Velázquez et al.,)	X	X	-	-
(Kudugunta and Ferrara, 2018)	-	X	-	-
(Davis et al., 2016)	X	-	X	-
(Varol et al., 2017b)	-	X	X	-
(Ferrara et al., 2016)	-	X	-	-
(Alarifi et al., 2016)	-	-	X	-
(Zhang et al., 2016)	-	X	-	-
(Grimme et al., 2017)	-	X	-	-
(et al.,)	-	X	X	-
(Ramalingam and Chinnaiah, 2018)	-	-	X	-
(Subrahmanian and et al.,)	X	X	-	-
(Gilani et al., 2017)	X	X	-	-
(Davis et al., 2016)	X	X	X	-
(Adewole et al., 2017)	X	-	X	-

4 PROPOSAL

Called RGG-BD, the method proposed in this work builds machine learning models that detect social bots using sentiment analysis. For this purpose, the RGG-BD allows the use of two sentiment lexicons, one being VAD (annotated manually) and the other being SSWE (learned from large masses of data). In addition to sentiment information, the proposed

Table 2: Set of Behavioral Attributes.

Attribute	Description
NumberOfFollowings	Number of accounts *c* follows
NumberOfFollowers	Number of followers account *c*
NumberOfTweets	Number of messages (tweets) posted by *c*
LengthOfScreenName	*c* name size on screen
LenDescrInUseProf	*c* profile description size

method also considers behavioral attributes of accounts when sending messages. Figure 1 presents a macro-functional view containing the steps of the RGG-BD.

The RGG-BD receives as input a set of accounts C and a set of messages M posted by elements of C. Each account $c \in C$ has a unique ID stored in $c.id$ and $c.a_1, c.a_2, \ldots, c.a_n$, behavioral attributes that depict the profile of c, such as exemplified in Table 2. Each message $m \in M$, in turn has 4 attributes: $m.i$ which contains the index of the message, $m.c$ which contains the identification (id) of the account $c \in C$ that posted m, $m.t$ which contains the text posted in m and $m.d$ which contains the date the message was posted.

4.1 Pre-processing

Regarding the existing noise in messages posted on social networks caused by informal language, it can be observed that the posts present the standards mentioned, demanding, for the analysis of the messages, the following pre-processing or cleaning steps, in thisorder:

- URLs Removal
- Usernames Removal
- White spaces Removal
- Hashtags Removal
- Character Repetitions Removal
- Removal of contractions and stop words

For each $m \in M$, the step called Pre-processing removes from $m.t$ urls, usernames, hashtags, repetitions (of letters and words), contractions, spaces, special characters and stop words using Natural Language Processing techniques described in (Sarkar, 2016).

4.2 Vector Calculation based on VAD Lexicon

Then, for each $m \in M$, this step aims to identify all existing words in $m.t$, generating an ordered set $Tk_{m.t} = \{p_1, p_2, \ldots, p_{|Tk m.t|}\}$, where each p_i is a *token* (word). Then, for each *token* $p_i \in Tk_{m.t}$, this step applies three functions $\beta_V(p_i)$, $\beta_A(p_i)$ e $\beta_D(p_i)$ that theyretrieve from the VAD lexicon the values of the valence (v_i), arousal (a_i) and dominance (d_i) dimensions associated with p_i, respectively. After processing all *tokens* in $Tk_{m.t}$, the weighted average of the values for each dimension is calculated. The Equation 1 illustrates the calculation made for the dimension corresponding to the valence V of $m.t$. The weighting factor f_i is the frequency with which p_i appears in$Tk_{m.t}$.

$$V(m.t) = \frac{\sum_{i=1}^{|Tkm.t|} v_i * f_i}{\sum_{i=1}^{|Tkm.t|} f_i} \quad (1)$$

4.3 Vector Calculation based on SSWE Lexicon

Similar to the previous step, for each $m \in M$, this step calculates, based on the SSWE Lexicon, the value of each of the dimensions $S_1, S_2, \ldots, S_{50}$ associated with $m.t$, as illustrated by Equation 2 for an arbitrary dimension S_j. In this equation, $s_{j,i}$ corresponds to the value of the dimension s_j retrieved from the SSWE lexicon for the *token* $p_i \in Tk_{m.t}$.

$$S_j(m.t) = \frac{\sum_{i=1}^{|Tkm.t|} s_{j,i} * f_i}{\sum_{i=1}^{|Tkm.t|} f_i} \quad (2)$$

4.4 Formation of the Structured Dataset

Then, for each message $m \in M$, this step generates a tuple gathering the values of the sentiment dimensions calculated in the previous two steps, as indicated in Equation 3.

$$ME_m = (m.c, V(m.t), A(m.t), D(m.t), S_1(m.t), \ldots, S_{50}(m.t)) \quad (3)$$

Once the tuples containing the sentiment data extracted from all M messages have been generated, the set of messages formed by structured sentiment data ME is formed, as indicated by Equation 4.

$$ME = \{ME_m / m \in M\} \quad (4)$$

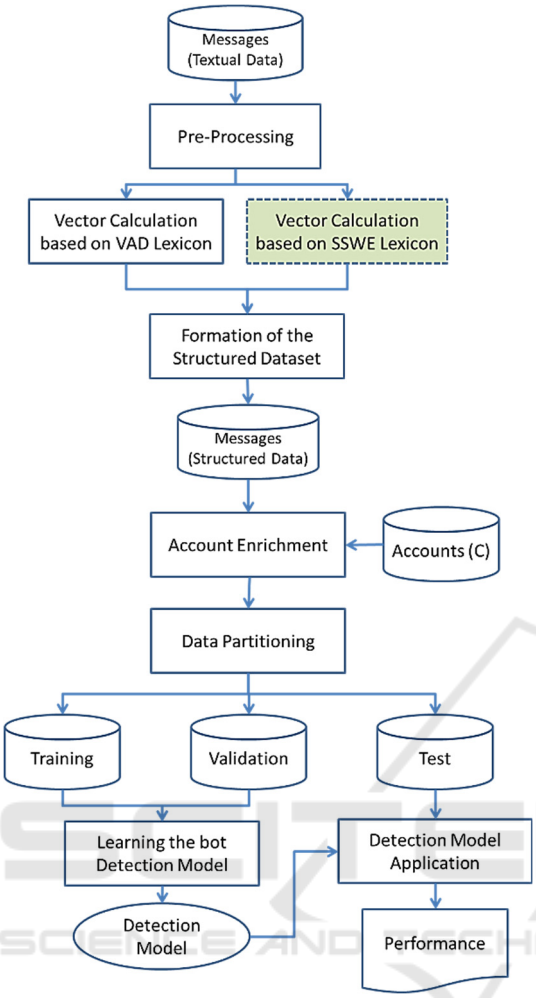

Figure 1: Macro-Functional View of Proposed Method.

4.5 Account Enrichment

In its first step, this enrichment step is responsible for separating, for each $c \in C$ account, the set $ME_c = \{ME_m \in ME / ME_m.c = c.id\}$ which contains all tuples with sentiment dimensions, structured and associated with the c account. Then, it consolidates the values of each dimension by means of statistical measures such as average, maximum and minimum as- sociated to each set. Equations 5 e 6 illustrate the calculation of the average and the maximum for the dimensions Valencia and S_{50}, respectively.

$$AVG(V(c)) = \frac{\sum ME_m \in ME_c \; ME_m.V(m.t)}{|ME_c|} \quad (5)$$

$$\text{MAX}(S_{50}(c)) = Max\{ME_m.S_{50} / ME_m \in ME_c\} \quad (6)$$

Then, this step is responsible for integrating the behavioral data of each $c \in C$ account with the statistical measures calculated from the sentiment of messages posted by c. To do so, it generates, for each ac- count c a tuple concatenating the values of the behav ioral attributes associated with c with the calculated statistical measures, as indicated in Equation 7.

$$T_c = (c.a_1, ..., c.a_n, AVG(V(c)),.., MAX(S_1(c)),...,MIN(S_{50}(c))) \quad (7)$$

Finally, this stage builds the set of enriched accounts CE, as indicated in Equation 8.

$$CE = \{T_c / c \in C\} \quad (8)$$

4.6 Data Partitioning

This step is responsible for separating CE into the training, validation and test sets. Such separation occurs at random, in a proportion that can be parameterized by the user.

4.7 Learning the Bot Detection Model

Then, in this step the RGG-BD trains the classification algorithm indicated by the data analyst with the messages from the training set. The validation set is used in order to select the best model generated by each algorithm, from different parameter settings specified by the analyst.

4.8 Detection Model Application

Finally, this step is responsible for applying and evaluating the performance of the classification model generated by the algorithm in the previous step in the test set. The same metric used in training the models should be used in this step.

5 EXPERIMENT AND RESULTS

5.1 Experiment

The RGG-BD prototype used in the experiments was implemented in Python, using the Pandas, Numpy, Scikit-Learn, Matplotlib, Seaborn libraries and the MySQL database.

5.1.1 Public Dataset

The *corpus* Caverlee 2011 (Lee et al., 2011) was chosen as the database, as it comes from *Twitter*, composed of a balanced set of accounts between bots and human, for having the texts complete of the messages in the English language, in addition to having

behavioral data from the accounts (see Table 4 for behavioral attributes used). Table 3 presents a statistical summary of this *corpus*. This table also indicates the total accounts and messages actually used in the initial experiments. Such subsets were obtained from a random sampling process without replacement in or- der to reduce the volume of data to be used.

The public dataset Caverlee 20111 is nothing more than a set of data acquired through social *honey-pots* collected from December 30, 2009 to August 2, 2010 on Twitter, containing 22,223 content polluting accounts, the number of his followers and followed over time, in addition to 2,353,473 *tweets* posted by them; similarly, 19,276 legitimate (human) user accounts, the number of their followers and followers over time and 3,259,693 *tweets* (Varol et al., 2017a). These data are divided into 6 tables.

Table 3: Accounts and messages used (Lee et al., 2011).

	Bot	**Human**
Total accounts	22,223	19,276
Total accounts used	**1,637**	**2,981**
Total tweets	2,353,473	3,259,693
Total tweets used	**238,455**	**436,798**

5.1.2 Lexicons

According to (Warriner et al., 2013), information on the affective meanings of words they are used by researchers working on emotions and moods, word recognition and text-based sentiment analysis.

In this sense, the first lexicon used in experiments of this type was the Affective Norms for Words in English (ANEW), developed to provide a set of emotional assessments for **1304** words in the English language. Subsequently, for more robust and comprehensive studies, affective assessments were collected for **13,915** well-known words in English, forming a solid basis from which the values of the remaining words can be automatically derived (Bestgen Vincze, 2012). This lexicon is known as Warriner, Kuperman Brysbaert (called in this work simply from the VAD lexicon).

In summary, for calculations of valence (V), arousal (A) and dominance (D), a table [1]containing 64 VAD values for 13,915 English words was used. In the case of this experiment, we opted for the average values (for each word) in the columns V-Mean-Sum,A-Mean-Sum and D-Mean-Sum.

The second lexicon used in this work was, in fact, the vocabulary of words represented by 50-dimensional *word embeddings*, resulting from the trained SSWE algorithm, when submitted to 10 million *tweets*. It is a table containing 134,915 *(embeddings)* records, but for reasons of algorithm performance, it was reduced to 11,665 records. This embedding table of sentiment-specific word embeddings was used in a similar way to the VAD lexicon, in or- der to search for sentiment values for the words of the message and, in an approximation, for the message asa whole.

5.1.3 Classification Algorithm

The classification algorithm used in the implementation of the learning steps and application of the bots detection model was random forest (scikit learn.org, 2020) (datcamp.com, 2020). This algorithm was chosen due to its history of good performances presented in other research in the area of social bot detection (Varol et al., 2017b), (de Souza, 2018). For the experiments now reported, the default values of the parameters random forest (datcamp.com, 2020) were used, eliminating the use of a validation set for the purpose of calibrating the detection model. Thus, the partitioning of the data generated only the training and test sets, in the proportion of 80% — 20%. The metric adopted for the evaluation of the classification model was the accuracy(Varol et al., 2017b). Basically it ex presses the number of correct answers produced by the model.

5.2 Results

In order to verify the hypothesis raised in this work, the four scenarios indicated in Table 4 were analyzed. This table summarizes the results of the experiments.

A first point to be highlighted is the improvement in the performance of the model whenever data on the sentiment present in the message texts are considered together with the behavioral data of the accounts (scenarios 2, 3 and 4 exceeded the scenario 1). This con- firms the results reported in (Ferrara et al., 2016).

Table 4: Accuracy obtained in the analyzed scenarios.

Scenario	**Accuracy**
1.Behavioral Data	89.5(%)
2.Behavioral Data + VAD	94.0(%)
3.Behavioral Data + SSWE	99.57(%)
4.Behavioral Data + VAD + SSWE	99.78(%)

Another fundamental aspect pointed out by the results is the confirmation of the hypothesis raised in this work that the detection of social bots that con-siders characteristics of sentiment in the words of

the messages can be improved if these characteristics have been learned from the data. Such an aspect can be observed when comparing the superiority of the performance of the model of the scenario 3 (lexicon of sentiment learned from the data) in relation to the scenario 2 (lexicon of sentiment noted manually by specialists).

Finally, it is worth highlighting an interesting result that the combination of sentiment data obtained based on an annexed lexicon and another learned from the data (4 scenario) was able to surpass the performances of the models generated based on the lexicons separately (2 and 3 scenarios). Such a result seems to suggest that the sentiment information learned from the data and that noted by experts can be complementary in supporting the identification of social bots.

6 FINAL CONSIDERATIONS

The use of social bots for malicious purposes has grown in recent years. Among the information used by the state of the art methods in the automatic detection of social bots are those that represent the sentiment existing in the messages propagated by the network. This information is calculated based on lex- icons of sentiment with content annotated manually by a restricted set of people. Therefore, this article proposed an approach that, based on a lexicon whose content has been learned from the diversity oflarge sets of messages, would be able to better rep- resent the existing sentiment in new messages and, consequently, increase effectiveness in detecting social bots. More specifically, the main objective of this work was to obtain experimental evidence that the detection of social bots that considers sentiment characteristics of messages can be improved if these characteristics are learned from large text bases.

An experiment was carried out using part of the public dataset Caverlee 2011. Like baseline, information on characteristics or behavioral attributes inherent in the accounts that send the messages was used. For the determination of sentiment values, two lexicons of sentiment were used, one being Warriner, Kuperman Brysbaert (also called in this work VAD), annotated manually, and the SSWE (learned by ma- chines from large masses of data).

The results obtained in the experiment showed an increase of 5 percentage points in scenario 3 where the SSWE lexicon was used in conjunction with the behavioral data. This may be due to the fact that the aforementioned algorithm, composed of three neural networks, incorporates sentiment information into thecontinuous vector representation of the lexicon.

Analyzing the results it is also possible to verify that the last scenario (4) showed a significant increase that may point to a combination of the two lexicons (noted and learned by machines with distant supervision) as a solution to improve the detection of social bots that use features of feeling analysis.

Thus, it can be observed that the use of the proposed approach reached its initial objective, verifying increases that show the hypothesis and validity of theproposal and justify further research in this direction. Possibilities for future work include: using all ofthe dataset Caverlee 2011, in addition to experimenting with new databases; optimization of the approach through the suppression of characteristics with little contribution to the prediction process, using, for this purpose, the information of importance of the characteristics; adaptation of the method to consider messages written in Portuguese; and use of other lexicons resulting from other methods.

REFERENCES

Adewole, K. S., Anuar, N. B., Kamsin, A., Varathan, K. D., and Razak, S. A. (2017). Malicious accounts: Dark of the social networks. Journal of Network and Computer Applications, 79(Supplement C):41 – 67.

Alarifi, A., Alsaleh, M., and Al-Salman, A. S. (2016). Twit- ter turing test: Identifying social machines. Inf. Sci., 372:332–346.

Bradley, M. and Lang, P. (1999). Affective norms for english words (anew): Instruction manual and affective ratings.

Clark, E. M., Williams, J. R., Jones, C. A., Galbraith, R. A., Danforth, C. M., and Dodds, P. S. (2016). Sift- ing robotic from organic text: a natural language approach for detecting automation on twitter. Journal of computational science, 16:1–7.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of machine learning research, 12(ARTICLE):2493–2537.

datcamp.com (2020). Understanding random forests classi- fiers in python.

Davis, C., Varol, O., Ferrara, E., Flammini, A., and Menczer, F. (2016). Botornot: A system to evaluate social bots. CoRR, abs/1602.00975.

de Souza, V. O. (2018). Avaliação de Modelos de Aprendizado de Máquina para Detecção Reativa e Preventiva de Botnets

et al., E. M. C. Sifting robotic from organic text: A natural language approach for detecting automation on twitter.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, (2016). The rise of social bots. *Commun. ACM, 59(7):96–104.*

Freitas, C., Benevenuto, F., and Veloso, A. (2014). Socialbots: Implicações na segurança e na credibilidade de serviços baseados no twitter. SBRC, *Santa Catarina, Brasi*l, pages 603–616.

Gilani, Z., Kochmar, E., and Crowcroft, J. (2017). Classification of twitter accounts into automated agents and human users. In 2017 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM), pages 489–496.

Grimme, C., Preuss, M., Adam, L., and Trautmann, H. (2017). Social bots: Human-like by means of human control? CoRR, abs/1706.07624.

Kudugunta, S. and Ferrara, E. (2018). Deep neural networks for bot detection. CoRR, abs/1802.04289.

Lee, K., Eoff, B. D., and Caverlee, J. (2011). Seven months with the devils: a long-term study of content polluters on twitter. In In AAAI Int'l Conference on Weblogs and Social Media (ICWSM.

Messias, J., Schmidt, L., Oliveira, R. A. R., and Souza, F. D. (2013). You followed my bot! transforming robots into influential users in twitter.

Ramalingam, D. and Chinnaiah, V. (2018). Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Comput. Electr. Eng., 65:165–177.*

Sarkar, D. (2016). Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data. Apress, *Berkely, CA, USA, 1st edition.*

scikit learn.org (2020). A random forest regressor description.

Statista and partners (2018). Bot traffic share. 06 maio de 2018.

Subrahmanian, V., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., and Menczer, F. (2016). The darpa twitter bot chal- lenge. Computer, 49(6):38–46.

Subrahmanian, V. S. and et al. The DARPA twitter bot challenge.

Tang, D., Wei, F., Qin, B., Yang, N., Liu, T., and Zhou, M. (2015). Sentiment embeddings with applications to sentiment analysis. *IEEE transactions on knowledge and data Engineering*, 28(2):496–509.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014a). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the* Association for Computational Linguistics (Volume 1: Long Papers), pages 1555–1565.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014b). Learning sentiment-specific word embedding for twitter sentiment classification. In *Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification,* pages 1555–1565, Baltimore, Maryland. Association for Computational Linguistics.

Varol, O., , and al. (2017a). Datasets.

Varol, O., Ferrara, E., Davis, C., Menczer, F., and Flammini, A. (2017b). Online human-bot interactions: *Detection, estimation, and characterization.*

Velázquez, E., Yazdani, M., and Suárez-Serrato, P. Social- bots supporting human rights.

Velázquez, E., Yazdani, M., and Suárez-Serrato, P. (2017). Socialbots supporting human rights. arXiv preprint arXiv:1710.11346.

Wagner, C., Mitter, S., Ko¨rner, C., and Strohmaier, M. (2012). When social bots attack: Modeling susceptibility of users in online social networks. In # MSM, pages 41–48.

Wagner, C., Mitter, S., Strohmaier, M., and Ko¨rner, C. When social bots attack: Modeling susceptibility of users in online social networks.

Warriner, A., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. Behavior Research Methods.

Zhang, J., Zhang, R., Zhang, Y., and Yan, G. (2016). The rise of social botnets: Attacks and countermeasures. PP.