Detection of Bot Accounts in a Twitter Corpus: Author Profiling of Social Media Users as Human vs. Nonhuman

Detección de cuentas de bot en un corpus de Twitter: Elaboración de perfiles de autor de usuarios de redes sociales como humanos vs. no humanos

María José Díaz Torres
Universidad de las Américas Puebla
maria.diazto@udlap.mx

Antonio Rico-Sulayes
Universidad de las Américas Puebla
antonio.rico@udlap.mx

Abstract

This paper presents a successful series of experiments on the detection of SPAMBOTS in Twitter, based on the use of linguistic features. For these experiments, we built a small corpus and classified its contents with the help of human annotators, who achieved a high rate of agreement. We identified linguistic features previously tested in the literature and adapted them to the language and contents of our database. High accuracy, (90%), was achieved in the spambot detection task. Our best results were obtained with a very small feature set produced with automatic reduction techniques. This outcome supports our contention that feature reduction is crucial in text classification tasks. All experiments were conducted by means of software packages with GUIs that do not require programming skills. Our results highlight the fact that language experts can, with a little training, utilize their knowledge and expertise in the very important fight against malicious technologies.

Keywords: *author profiling, bot detection, machine learning, spambots, Twitter.*

Resumen

Este artículo presenta una exitosa serie de experimentos sobre la detección de BOTS DE SPAM en Twitter, que parten del uso de características lingüísticas. Para estos experimentos, construimos un corpus de corta extensión y clasificamos su contenido con la ayuda de anotadores humanos, quienes alcanzaron un alto nivel de acuerdo. Identificamos características lingüísticas probadas previamente en la literatura y las adaptamos al idioma y al contenido de nuestra base de datos. De esta manera, se obtuvieron resultados de exactitud prometedores, (90%), en la tarea de detección de bots. Nuestros mejores resultados se lograron con un conjunto de rasgos muy limitado, producido con técnicas de reducción automática. Estos resultados apoyan la idea de que la reducción de rasgos es crucial en las tareas de clasificación de texto. Más aún, todos los experimentos se llevaron a cabo usando paquetes de software con interfaces gráficas que no requieren de conocimientos de programación. Esto muestra que los expertos en el lenguaje tienen conocimientos y experiencia que, con un poco de entrenamiento, pueden aplicar en la importante lucha contra las tecnologías maliciosas.

ARTÍCULO RECIBIDO EL 06/10/2020, ACEPTADO EL 20/09/2021 *LENGUA Y HABLA* 25. ENERO-DICIEMBRE, 2021

Palabras clave: *perfilado del autor, detección de bots, aprendizaje automático, spambots, Twitter.*

1. INTRODUCTION

This study approaches the task of author profiling with the aim of detecting Twitter BOTS or SPAMBOTS through the identification and analysis of a number of linguistic features that are commonly used by tweeters in Spanish. Defined as the task of identifying characteristics typical of a particular anonymous author across a number of social categories, such as gender or age (Rico-Sulayes, 2018), author profiling attempts to recognize general patterns that characterize the text production of a specific group of authors. We use this approach to identify text produced by bot accounts as well, by assembling a corpus of tweets written in Spanish and manually annotated with the binary categorization *human* vs. *non-human*, or *bot*.

Twitter bots can be defined as semi-automated or fully automated accounts that tweet pre-written or algorithmically generated content without direct human intervention (Mowbray, 2014). While many bots provide useful services or entertainment for Twitter users, the most common type of automated tweets is probably spam (Mowbray, 2014). Therefore, it is important to detect these malicious accounts because not only do they violate Twitter's terms of service, but most importantly, they can be and have been used to create nuisance and even security concerns by "spreading malware, sending spam, and advertising activities of doubtful legality" (Cresci et al., 2015:2). Moreover, these bots "abuse online social networks" in a number of ways including artificially boosting the popularity of individuals, usually that of people in power, or influencing public opinion to further personal agendas (Varol et al., 2017:280). Some examples include the use of bots for corrupt political purposes, like the attempts by allegedly repressive governments to drown out Twitter political conversations by protesters in Russia (Thomas et al., 2011) or Mexico (Santiesteban, 2021), in 2011 and 2012, respectively.

Given the malicious use to which bots can be put, methods that make their detection possible, through the identification of the textual features of their messages, are certainly worth pursuing. The work we present in this article is particularly important because, although bot account detection is currently a popular task in the text mining community (e.g., Atodiresei et al., 2018; Inuwa-Dutse et al., 2018; Washha et al., 2017; Zheng et al., 2015), there is still little work specifically tailored to Spanish, and even less focused on the Mexican context as is our study. In addition, our results, which achieved 90% accuracy, obtained through the use of several classifying algorithms, are particularly encouraging compared to the state-of-the-art results that range from 96%-99%, in various related tasks such as spam filtering (Dong et al., 2017) and bot detection (Kudugunta & Ferrara, 2018). Although our corpus is rather small, the fact that we use a much-reduced set of features to identify bot accounts makes our work particularly interesting for the text classification community working with restricted access to relevant text (Rico-Sulayes, 2017). Furthermore, working with small text excerpts and datasets has also been recognized as an important factor to deal with in bot detection research (Kudugunta & Ferrara, 2018).

The article is organized as follows: the second section provides the theoretical framework that includes a review of the literature on the detection and profiling of bots and their typical features. The third section describes the dataset and data collection procedures. In the fourth section, we explain our methodology for detecting spambots through the linguistic features that appear in tweets in Spanish, including feature engineering and feature reduction techniques, and the classifiers we chose, which include several algorithms commonly employed for similar tasks. In section five, we present and discuss the results. Finally, we draw conclusions based on our results and discuss directions for future research.

2. LITERATURE REVIEW

A great deal of research in computer science, linguistics and related fields of study has been published in recent years on the topic of identifying Twitter spambots. This literature reports on the performance of several classifying algorithms based on machine learning techniques through a combination of different measures, such as accuracy, precision, and F-measure. In this literature, accuracy has been reported as high as 93% or more (Mowbray, 2014). Accuracy, or the true positive rate, represents in this context the proportion of times an algorithm correctly classifies a message as produced by a spambot. In this section, we review previous work on the identification and profiling of bots, fake followers, and spam accounts. Drawing from this theoretical framework, we will outline the features found to be characteristic of these automated accounts and their algorithmically generated text.

One of the challenges of automated account detection is Twitter bots to commonly pretend to be humans. Matwyshyn and Mowbray (2012) analyzed the profiles of 727 potential bots that tweeted using unregistered clients. Manual examination confirmed that the accounts were almost certainly bots; however, 37% of them contained human-like indicators and gave no sign that the account was automated (Matwyshyn & Mowbray, 2012). These indicators included recognizably human names and textual content in the Twitter biographies, which suggested that the account was genuine. Furthermore, certain bots' actions may also make them appear human, such as automatically replying to tweets or sending human-like tweets (Matwyshyn & Mowbray, 2012).

Consequently, to tackle the task of bot identification, the approach many authors have adopted consists of analyzing data from the profile of the account. For instance, Cresci et al. (2015) created a baseline dataset of both human and fake follower accounts. The researchers surveyed techniques for spammer and bot detection and identified classification rules, as well as several account profile-related features proposed by both scientific literature and GREY literature, namely, "online documentation that presents a series of intuitive fake follower detection criteria, though not proved to be effective in a scientific way" (2015:4). After the evaluation of their dataset by means of machine learning-based algorithms, the authors found that the most effective profile-related features for the identification of bots were the ratio of friends to followers, the age of the account, the number of tweets, and whether the profile had a name or not.

Similarly, the content of the accounts' posts or tweets has also proved to be useful for the detection of spambots. An example of this approach is in Thomas et al. (2011). In this study,

the authors examined over 1.1 million Twitter accounts suspended for disruptive activities and collected a dataset of 1.8 billion tweets, including 80 million from bots. In the study, they characterized the behavior and lifetime of spam accounts and concluded that indicators of spam include the use of words correlated with spam in tweets or biographical content, the frequent use of mentions, hashtags or trending topics, URLs, and short account age.

There have also been approaches that place an emphasis on identifying linguistic features present in messages from bots. An example of the success of linguistic feature engineering is Laboreiro, Sarmento and Oliveira (2011). They analyzed various elements of the writing style of tweets. They also included features such as the time of posting, the account used to post, the presence of links and user interaction. One of the authors' goals was to achieve characterizations of human, automated accounts (or bots) and semi-automated accounts (denominated CYBORGS). In the analysis of their results, the stylistic features performed best, with a median accuracy of 97%. These features included emotext (such as emoticons), punctuation, length and frequency of tokens, use of capitalization, and beginnings and endings of messages. The authors argued that bots' tweets were overall formal, objective, non-emotional and grammatical. According to the authors, these tweets demonstrate careful use of capitalization and scarce use of punctuation marks. Laboreiro et al. concluded that "some bots will reveal a pattern that is used in all their messages" (2011:11). Our study has been particularly influenced by the results of this study, in which linguistic elements outperform other types of features.

3. DATA COLLECTION AND DESCRIPTION

In order to develop a classification model for the detection of text produced by bot accounts, we assembled a corpus of tweets written in Spanish. The tweets gathered were manually annotated with the binary categorization *human/non-human*. The data was collected by searching for tweets with the trending topic *#felizlunes* '#happymonday', as this tag represented a popular Twitter discussion with a large number of users taking part when we collected the data. These kinds of trending topics have been found to be attractive to spambots, since they allow them to reach a larger audience (Thomas et al., 2011).

One of our goals was to gather both human and non-human generated data. With this purpose in mind, we selected 25 accounts with 10 tweets from each, for a total of 250 tweets. Three human annotators analyzed and classified the 25 accounts (as *human/nonhuman*) with an overall agreement of 94.7%. Based on the annotators' ratings, we decided to use 10 bot accounts (which were classified as non-human by at least one annotator, and with a general accuracy of 90%) and 10 human accounts (which were judged as human with an accuracy of 100%). Although the overall agreement would have allowed us to get a larger human user set, we only selected 10 human accounts to make the task balanced.

4. LINGUISTIC FEATURE ENGINEERING FOR SPAMBOT DETECTION IN SPANISH

Our Twitter bot detection system is based on a supervised learning model that classifies tweets as human or nonhuman based on the presence of a number of linguistic features and the frequency with which they appear. Based on our review of related research, we identified four feature categories: content words, function words, tokens and a miscellaneous category made up of numerical features. As we will discuss later, the latter two categories contained mostly structural information. Next, we selected a number of the most frequently occurring individual features for each category to be used in the classification. This resulted in a total of 34 features shown in Table 1 and further described below.

Our first category included in Table 1 is content words, or words with a lexical meaning that is relatively independent of the position of the word in the sentence. A prototypical example of content words are entities, or names of people, places and things. We selected the ten most frequent content words in our corpus, which included: *año* 'year', *vida* 'life', *gente* 'people', *méxico*, *amlo* (the acronym of Mexican president Andrés Manuel López Obrador), *amor* 'love', *chávez* (in reference to Hugo Chávez, former president of Venezuela), *ciudad* 'city', *noviembre* 'November', and *semana* 'week'. It should be noted that all words were previously changed to lower case, as is customary for information retrieval text preprocessing (Manning, Raghavan, & Schütze, 2009).

The second category of features shown in Table 1 is function words, which are the counterpart to content words. Namely, they join content words to make up phrases and sentences, so their meaning is derived from their function in the sentence. The ten most frequent function words selected include: prepositions, *de* 'of', *a* 'to', *en* 'in', and *para* 'for'; conjunctions, *que* 'that' and *y* 'and'; articles, *la* 'the (feminine)', *el* 'the (masculine)', and *los* 'the (masc. plural)'; and a contraction *del* 'of the'.

Additionally, the most frequent tokens, which are neither a content nor a function word, make up the third feature category in Table 1. In this particular corpus these tokens are structural features because they are dependent on or characteristic of the communication medium (Rico-Sulayes, 2018). The ten most frequent tokens were the hashtag *felizlunes* 'happymonday', the internet-related sequences *https* and *com* used to form URLs, the hashtag *noticiasmatrix* (a news related topic) and its abbreviation *mtx*, the hashtag *metrobuscdmx* (about Mexico City's rapid transit bus system), the hashtag *diainternacionaldelhombre* 'internationalmensday', the sequences *youtu* and *be* from YouTube's URLs, and the hashtag *felizviernes* 'happyfriday'.

Lastly, the fourth category of features in Table 1 represents a set of miscellaneous numerical features derived from various types of data, such as the frequency of hashtags and the number of users mentioned in the text. These two features were selected because they are a key Twitter attribute likely to be used distinctively by groups of authors, as found in previous work (Thomas et al., 2011). In this category we also noted the presence of uppercase letters and the length of the tweets because we observed important differences between groups during data collection and took into account their reported effectiveness for this task in the literature (Laboreiro et al.,

2011). We believe that the features in this last category are closely related to the structural nature of the conversation in Twitter.

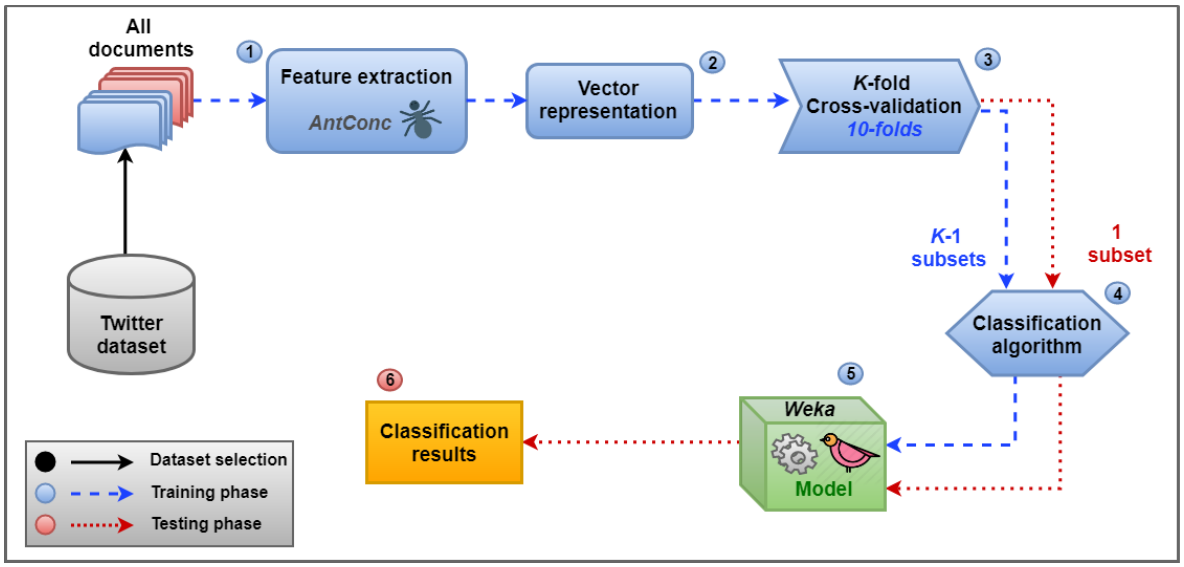
Table 1. The 34 linguistic features selected organized by four categories: content words, function words, tokens, and miscellaneous features.

Category	Feature	Translation	
Content words	*año*	year	
	vida	life	
	gente	people	
	méxico	Mexico	
	amlo	acronym: Andrés Manuel López Obrador	
	amor	love	
	chávez	Hugo Chávez	
	ciudad	city	
	noviembre	November	
	semana	week	
Function words	*de*	preposition	of
	a	preposition	to
	en	preposition	in
	para	preposition	for
	que	conjunction	that
	y	conjunction	and
	la	article	the (fem.)
	el	article	the (masc.)
	los	article	the (masc. pl.)
	del	contraction	of the
Tokens	*felizlunes*	happymonday	
	https	URL	
	com	URL	
	noticiasmatrix	News' related topic	
	mtx	Abbreviation of news topic	
	metrobuscdmx	Mexico City's rapid transit bus	
	diainternacionaldelhombre	internationalmensday	
	youtu	Part of trimmed URL for YouTube	
	be	Part of trimmed URL for YouTube	
	felizviernes	happyfriday	
Miscellaneous features	Hashtags (#), frequency		
	Users mentioned (@)		
	Capitalization		
	Length of tweets		

Another important component of our feature engineering approach is the use of a feature reduction technique. Given the reported success of these kinds of procedures in improving the

accuracy of text classification algorithms when combined with a large set of features (Rico-Sulayes, 2017), we wanted to explore whether such techniques would be beneficial in this context. We were especially interested in exploring whether, in these experiments, that already utilize a small feature set, an even smaller set of particularly discriminatory features could still refine the performance of classification tasks. In this study, this procedure was adopted using the attribute selector tool of Weka[1], a data mining open software (version 3.8.4 for Windows, 2019). The reduction technique used was Correlation-based Feature Subset Selection (CfsSubsetEval), together with the Best First method, because it is considered to be one of the most successful techniques available (Rico-Sulayes, 2017). After the application of this reduction technique, four out of the 34 original features were selected: the hashtag *felizlunes*, the frequency of hashtags, the use of capitalization, and tweet length. The methodology of our classification model is displayed in Figure 1 and explained below.

Figure 1. Spambot detection through linguistic features in Spanish.



As shown in Figure 1, we first used the AntConc[2] concordance open software package (version 3.5.9 for Windows, 2020) to manipulate our corpus, composed of documents from bots and human tweeters. By means of this tool, we identified and counted the instances of the selected features formerly discussed (1). The great advantage of this kind of tool is that it does not require programming skills, which many social scientists, including many linguists, do not

[1] https://www.cs.waikato.ac.nz/ml/weka/

[2] http://www.laurenceanthony.net/software/antconc/

have. Using a comma separated values spreadsheet (which can be produced in Excel, for example), we represented each document as a vector (or lists of values) based on the frequency of occurrence of each of the features (2). Next, given the fact that our corpus consisted of a single dataset, we selected a *k*-fold cross-validation technique, which divides the data into training and testing sets to explore the performance of the various classifying algorithms. For these experiments, the data was divided into 10 folds (or partitions) by means of Weka (3), in order to train the classification algorithms with *k*-1 subsets (90% of the data), and the remaining subset (10% of the data) was used for testing. This process was repeated 10 times with different testing subsets so that the performance of a classification algorithm was not biased by using data in the model that is later used in the classification (4). All of these components are part of Weka, a software package that has a GUI that also does not require programming skills, at least not with relatively small databases (5). The average of the classification results for the ten folds formerly described is presented by Weka using various metrics, which include the corresponding author categories (*human* or *nonhuman*) associated with each document (6). In order to explore the performance of different machine learning algorithms, we used five classifiers in Weka that are suitable to apply to text mining tasks with binary problems such as ours (*human* vs. *nonhuman*): Naïve Bayes (a Bayesian classifier), Simple Logistic (a function classifier), LMT (a decision tree classifier), MultiClass Classifier Updatable (a meta classifier), and KStar (a lazy family classifier).

5. RESULTS

The main goal of this study is the task of profiling tweets in order to detect bots. We did this by using both all of our features combined, and a reduced subset selected automatically. Our chosen linguistic features were represented in the document vector by the frequency of occurrence in the previously annotated tweets in Spanish. For the experiments that combined the original set of 34 features, the best results were obtained by the Simple Logistic and LMT classifiers that achieved 85% accuracy. Table 2 below shows the performance of all five classifiers tested (Naïve Bayes, Simple Logistic, LMT, MultiClass Classifier Updatable, and KStar) in combination with all of our features.

Table 2. Accuracy and error rates of the classifiers for the original dataset (34 features).

	LMT	**SimpleLogistic**	**NaiveBayes**	**MulticlassClassifier Updatable**	**KStar**
Accuracy	85.00%	85.00%	75.00%	70.00%	70.00%
Error rate	15.00%	15.00%	25.00%	30.00%	30.00%

As for the experiments with feature reduction techniques, the reduced set of four features achieved higher accuracy for all individual classifiers. This is shown in Table 3 below. The first three classifiers, shown from left to right, reached 90% accuracy (increasing the accuracy from between 5 and 15 percentage points) and the last two obtained 80% accuracy (with a 10-percentage-point increase).

Table 3. Accuracy and error rates of the classifiers for the reduced dataset (4 features).

	NaiveBayes	SimpleLogistic	LMT	MulticlassClassifier Updatable	KStar
Accuracy	90.00%	90.00%	90.00%	80.00%	80.00%
Error rate	10.00%	10.00%	10.00%	20.00%	20.00%

6. CONCLUSION

From the results presented in the former section, we can conclude that Twitter bots in Spanish can be identified with high accuracy based on the textual and linguistic features of their messages. This is true even for small datasets with very small, refined sets of features. Based on our experiments in Spanish, we conclude that tweets of automated spam accounts in this language are characterized by heavy use of user mentions and hashtags, and tend to have a fixed format. This kind of automatically generated text in Spanish also tends to be longer and exhibit greater use of uppercase letters.

An important limitation of this study we should acknowledge is that we used a small dataset compared to the large datasets that natural language processing competitions, mainly targeted at computer scientists, tend to employ. However, we have argued that the use of GUIs allows social scientists to model solutions for these kind of problems on small datasets, so we think that demonstrating this is in fact an important contribution of this study, providing an opportunity for engineers and linguists to work together eventually to adapt these solutions to large-scale settings.

In future work, we want to conduct a qualitative syntactic analysis, because we hypothesize that the syntax of bots is simpler than that of human generated text, an observation that arose from an analysis of the most frequent function words present in the tweets. Bots more often used the preposition *de* and the conjunction *y*, characteristics of coordinate structures, whereas humans used the conjunction *que* more, employed in subordinate structures.

As users, we are often able to distinguish the messages that were created by another human being from those that were not, by looking carefully at how they are written. By the same token, the findings of our study suggest that there are certain linguistic traits that characterize spambot texts, and that language experts can contribute to identifying these linguistics features via a statistics-based classifier that can detect nonhuman generated text with a high level of accuracy. We believe that further research on the use of linguistic characteristics for the development of bot automatic detection models is necessary and contend that much of this work can be done by linguists. Finally, it is important to mention that user-friendly experimentation environments such as Weka and AntConc allow language experts to explore machine learning techniques without requiring programming skills. In the present research we made use of these kinds of

software tools and achieved promising results with a database that uses little text with very small sets of features in Spanish.

REFERENCES

Atodiresei, Costel-Sergiu, Alexandru Tănăselea & Adrian Iftene. 2018. Identifying Fake News and Fake Users on Twitter. *Procedia Computer Science* 126. 451-461.

Cresci, Stephano, Di Pietro, Roberto, Petrocchi, Marinella, Angelo Spognardi & Maurizio Tesconi. 2015. *Fame for sale: efficient detection of fake Twitter followers. Decision Support Systems* 80. 56-71.

Dong, Youngsu, Mourad Oussalah & Lauri Lovén. 2017. A on Spam Filtering Classification: A Majority Voting like Approach. *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR 2017)*. 293-301. Madeira, Portugal: Science and Technology Publications.

Kudugunta, Sneha & Emilio Ferrara. 2018. Deep Neural Networks for Bot Detection. *arXiv, 1802.04289v2*. 1-10.

Inuwa-Dutse, Isa, Mark Liptrott & Ioannis Korkontzelos. 2018. Detection of spam-posting accounts on Twitter. *Neurocomputing* 315. 496–511.

Laboreiro, Gustavo, Luís Sarmento & Eugénio Oliveira. 2011. Identifying automatic posting systems in microblogs. *Progress in Artificial Intelligence Conference*. Lisbon, Portugal.

Manning, Christopher D., Raghavan, Prabhakar, & Schütze, Hinrich. (2009). *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University.

Matwyshyn, Andrea & Miranda Mowbray. 2012. Bot or not?: Digital augmentation and personhood. *Internet Law Work-in-Progress*. New York Law School: New York, United States.

Mowbray, Miranda. 2014. Automated Twitter Accounts. In Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (Eds.), *Twitter and Society*. New York et al.: Peter Lang.

Rico-Sulayes, Antonio. 2017. Reducing Vector Space Dimensionality in Automatic Classification for Authorship Attribution. *Revista de Ingeniería Electrónica, Automática y Comunicaciones* 38(3). 26-35.

Rico-Sulayes, Antonio. 2018. *Authorship Attribution on Crime-Related Social Media: Research on the darknet in forensic linguistics*. Aracne Editrice: Rome, Italy.

Santiesteban, Iván. 2021. *Goodbye bots*. Retrieved from http://santiesteban.org/adiosbots/en.html

Thomas, Kurt, Grier, Chris, Dawn Song & Vern Paxson. 2011. Suspended accounts in retrospect: An analysis of Twitter spam. *Internet Measurement Conference*, Berlin, Germany.

Varol, Onur, Ferrara, Emilio, Davis, Clayton A., Filippo Menczer & Alessandro Flammini. 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*.

Washha, Mahdi, Qaroush, Aziz, Manel Mezghani & Florence Sedes. 2017. A Topic-Based Hidden Markov Model for Real-Time Spam Tweets Filtering. *Procedia Computer Science* 112. 833-843.

Zheng, Xianghan, Zeng, Zhipeng, Chen, Zheyi, Yuanlong Yu & Chunming Rong. 2015. Detecting spammers on social networks. *Neurocomputing* 159. 27-34.