Detection of Human, Legitimate Bot, and Malicious Bot in Online Social Networks Based on Wavelets

SYLVIO BARBON JR, GABRIEL F. C. CAMPOS, GABRIEL M. TAVARES, RODRIGO A. IGAWA, and MARIO L. PROENÇA JR, Londrina State University (UEL)
RODRIGO CAPOBIANCO GUIDO, São Paulo State University (UNESP)

Social interactions take place in environments that influence people's behaviours and perceptions. Nowadays, the users of Online Social Network (OSN) generate a massive amount of content based on social interactions. However, OSNs wide popularity and ease of access created a perfect scenario to practice malicious activities, compromising their reliability. To detect automatic information broadcast in OSN, we developed a wavelet-based model that classifies users as being human, legitimate robot, or malicious robot, as a result of spectral patterns obtained from users' textual content. We create the feature vector from the Discrete Wavelet Transform along with a weighting scheme called Lexicon-based Coefficient Attenuation. In particular, we induce a classification model using the Random Forest algorithm over two real Twitter datasets. The corresponding results show the developed model achieved an average accuracy of 94.47% considering two different scenarios: single theme and miscellaneous one.

CCS Concepts: • **Security and privacy → Social network security and privacy**; • **Computing methodologies → Natural language processing**;

Additional Key Words and Pharses: OSN frauds, text mining, writing style, wavelets, bots

ACM Reference format:
Sylvio Barbon Jr, Gabriel F. C. Campos, Gabriel M. Tavares, Rodrigo A. Igawa, Mario L. Proença Jr, and Rodrigo Capobianco Guido. 2018. Detection of Human, Legitimate Bot, and Malicious Bot in Online Social Networks Based on Wavelets. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 1s, Article 26 (March 2018), 17 pages.
https://doi.org/10.1145/3183506

1 INTRODUCTION

Online Social Networks (OSNs) are suitable environments to discuss, express, and argue thoughts on any subject (Zappavigna 2011). Currently, OSNs represent a relevant resource for exploring a diversity of fields, such as customer relationship management and opinion mining. The advent of social networks makes social multimedia sharing very easy and an integral part of society (Ye et al. 2016). Knowledge obtained from thoughts expressed in OSNs such as Twitter and Facebook

Authors' addresses: S. Barbon Jr, G. F. C. Campos, G. M. Tavares, R. A. Igawa, and M. L. Proença Jr, Department of Computing, Londrina State University (UEL), Rod. Celso Garcia Cid km 380, 86057-970, Londrina-PR, Brazil; emails: {barbon, camposg, gtavares, igawa, proenca}@uel.br; R. C. Guido, Instituto de Biociências, Letras e Ciências Exatas São Paulo State University (UNESP), Rua Cristóvão Colombo 2265, Jd Nazareth, 15054-000, São José do Rio Preto-SP, Brazil; email: guido@ieee.org.

https://doi.org/10.1145/3183506

has shown to be valuable for marketing research companies or public opinion analysis by providing information about what people are discussing (Bahrainian and Dengel 2013; Smailović et al. 2014). Since millions of views on a certain topic are expressed, posted messages provide rich and easy comprehension of content (Hassan et al. 2013). Thus, the OSNs content is a valuable data set for decision-making in marketing research, business intelligence, and stock market prediction (Mostafa 2013; Hsieh et al. 2012). Moreover, because of mobile devices growth, people can perform various social activities anytime and anywhere (Ma and Yan 2015).

However, OSNs' wide popularity and ease of access created a perfect scenario to practice malicious activities, compromising their reliability. Just to exemplify some security issues, spamming campaigns for advertisement of products represent a common problem to be dealt on OSNs (Bhat and Abulaish 2013). The existence of such undesirable users is present on both Twitter and Facebook, which are quite popular OSNs. As stated by Chu et al. (2012) on Facebook and by Fong et al. (2012) on Twitter, the amount of fake and malicious accounts in both OSNs is more than 8% worldwide. Thus, the growth of social multimedia underscores potential risks for malicious use (Ye et al. 2016).

Users of OSNs tend to connect with people of their interests, creating a proper environment for marketing and social manipulation, business promotion, customer service, political campaigning, and emergency communication. In the OSN, users create and maintain different social connections, for example, connecting with their real friends, following celebrities, or even liking virtual social entities (Wang et al. 2013). Thus, spam accounts are usually posting a malicious link to obtain users' credentials. On Twitter, the social robots pretend to be human beings to gain followers and replies from target users and promote a product (Wald et al. 2013a). Awareness of third-party applications is also a worry while dealing with user credentials. Once private information is given to such, they are considered a way of exposing private information as well (Pang and Zhang 2015; Shehab et al. 2012; Jin et al. 2013b; Li et al. 2015).

Chen et al. (2015) discussed the fake user problem and the difficulty to judge whether a view is real or fake. Their study exposed the ecosystem for generation of fake views, the kind of tools used and the business motivation for different parties. In conclusion, they highlight the necessity for developing more efficient online algorithms to deal with the issue.

Frauds in OSNs, such as those mentioned above, could lead to uncontrolled dissemination of fake information, inaccurate content, promotional ads, and phishing. This way, OSN users might: (a) access false content, (b) have their credential stolen, or (c) receive unsolicited content. Considering so, users could become victims of tricky scams or harassment, causing a decline of service (Bhat and Abulaish 2013; Miller et al. 2014).

This is where Text Mining (TM) comes into play. Once OSNs provide a tremendous amount of their content in text format (Aggarwal 2011), text mining tasks were performed to analyse the stream of information available with different goals. Some examples were included in Mostafa (2013), which presents a work of mining opinions toward some companies, and Bahrainian and Dengel (2013), which focused on analysing opinions on certain products. Addressing frauds and crimes, Keretna et al. (2013) proposed a text mining approach to recognise false identities, protecting celebrities and politicians from fake announcements. Also concerning the identification of cyber-criminals, Almishari et al. (2014) developed a work to aid the recognition of authors of Twitter messages. In another work in this scenario, Barbon et al. (2016a) addressed the detection of compromised accounts on OSNs based on posts content. They achieved an accuracy of over 93% and highlighted the limitation related to users' post frequency.

Addressing a different ground for text-mining solutions, there are some works that combined concepts from Digital Signal Processing to take advantage of Discrete Wavelet Transform (DWT)

suitable computational cost. In Arru et al. (2013), a solution was presented for user recommendation on Twitter based on signal similarity and Weng and Lee (2011) and Cordeiro (2012) devised DTWs as the ground for the event detection based on selected signalised terms.

Although many text-mining works were developed toward OSNs, none has been solely focused on the distinction of posts' authors between robots (bots) and humans. More specifically, this research aims to classify accounts in human, malicious bot, or legitimate bot. Thus, we aim to contribute to OSN reliability by detecting malicious bots. Our work is a text-mining approach using fast calculated features grounded on DWTs. It does not depend on account features and is the first one to classify between humans and two distinct classes of bots.

The remainder of this work is organised as follows. Section 2 provides a brief review on researches concerning bots on OSNs. Section 3 introduces works toward text mining based on DWTs. Section 4 describes the proposed approach. Section 5 focuses on our experimental settings along with our results, and Section 6 presents our conclusions.

2 LEGITIMATE BOTS AND MALICIOUS BOTS

Automation is considered a double-edged sword on Twitter. Legitimate bots generate a large volume of benign tweets as news and blog updates, which complies with Twitter's original goal of becoming an information network. On the other side, malicious bots have been heavily used to spread spam and unsolicited commercial content or even to phish user's credentials (Chu et al. 2012). OSNs malicious bots detection was the main motivation for our work.

Nowadays, fake or malicious accounts are excessively dependent on real users' reports (Boshmaf et al. 2011) to be removed. This is not a satisfactory solution, since most users are not willing to spend their time reporting other accounts, moreover, this tool can be misused in occasions where a group of users reports accounts based on different opinions, mostly in political matters. Also, when the qualified company reacts over a suspicious account, serious damages had probably already occurred to real and legitimate users, mainly if thousands of malicious bots had been involved (Fong et al. 2012).

Most OSN do not adopt automated ban of accounts, that is, there is no algorithm that either report or ban accounts. This approach is not applied, because it may ban some legitimate accounts and this will probably make the OSN popularity decay in public opinion. Usually, companies hire employees to search the OSN and ban the accounts they judge as malicious, making the process a manual task (Cao et al. 2012). However, this is considered a high monetary investment. This shows the need for an approach that can deal with the high demand of data from OSN toward a small number of false positives to prevent legitimate accounts banning.

2.1 Malicious Bots

Any bot created to cause harm can be considered malicious. In this work, a malicious bot is considered any bot that spreads malicious, redundant, or not desired commercial content in its tweets, characterized mostly by the lack of originality. Authors of Egele et al. (2013) stated that bots were used to post dangerous URL links that could harvest users credentials. Another example, authored by Jiang et al. (2014), stated that bots are "zombies," because thousands of accounts were automatically created to promote specific ones. Usually, such promoted accounts belong to celebrities and companies trying to achieve more popularity.

In both cases, malicious bots cause harm. Regardless their intention, i.e., promoting a specific account or posting malicious content, they could lead to dissemination of fake information, harmful content, and phishing. Such examples of frauds could make users victims of tricky harassment, causing a dissatisfaction addressing OSN security and, therefore, the decline of the service.

Part of the malicious content in OSN rises up from the fact that a user is prone to interact and respond to messages from another OSN user (supposing to be a friend) (Grier et al. 2010); consequently, the social spam is more effective than traditional spam, i.e., e-mail spam (Gao et al. 2010). Furthermore, the majority of spam messages comes from compromised accounts (Gao et al. 2012), that is, considering just the users' profile information can lead to inaccurate conclusions. In this way, the text content analysis could cope with this issue. A suggestion proposed by Jin et al. (2013a) is to incentivize users to migrate to decentralized OSN, stating they could provide a structure capable of protecting sensitive information from leaking out. However, this approach does not attack the malicious accounts themselves. It deviates from the problem by offering a change of behaviour from the user side. Nonetheless, to encourage users to switch to a decentralized OSN is not a trivial task.

2.2 Legitimate Bots

Opposite to malicious bots, legitimate ones help to spread information addressing transit traffic or weather news in cases of urgency. As stated by Chu et al. (2012), this kind of activity matches the original goal of OSNs, such as Twitter.

These kinds of bots, differently from malicious ones, should not be targeted as sources of OSN security problems. So far, there are very few solutions (Fong et al. 2012; Chu et al. 2012) addressing the identification of bots in OSNs, and they depend on a set of different information, limited to a particular scenario. All of them need, along with the textual content produced by the user account, non-textual features such as the quantity of relationships, frequency pattern of use (e.g. days of the week in which the accounts is used), or even analysis of user's avatar image.

Despite having similar frequency characteristics to malicious bots, legitimate ones differ in the content provided. This way, legitimate bots produce content that is substantially close to human content. Chu et al. (2010) states that humans are identified by their originality, intelligence and specificness, usually talking about news and matters of personal interest. Even though legitimate bots have automatic behaviour, the content comes from intelligent sources, like news outlets, which makes this kind of account akin to human accounts.

Textual patterns and characteristics that set apart mischievous and authentic accounts are explored more deeply with DWT, since this technique is able to extract the distribution of the content exposed. Therefore, reinforcing the difference between intelligent and spam-like content that comes from legitimate and malicious accounts, respectively.

3 WAVELETS ON TEXT DESCRIPTION

The combination of DWTs with text mining achieves results in different scenarios such as Information Retrieval (IR), document classification, text visualization, and user content recommendation. Especially for IR, DWTs are capable of analyzing key term patterns as being a vector of frequencies at different levels of resolution (Park et al. 2002). Also, DWT-based approaches were applied to retrieve Web page contents and to improve the results in search engines (Purwitasari et al. 2007; Purwitasari 2008).

For classification, particularly representing a document as a set of key terms, DWTs achieved better results than the classic Space Vector Model as seen in Thaicharoen et al. (2008). Using the same principle, but a different model, Xexéo et al. (2008) also performed document classification successfully, this approach took advance of the dimensionality reduction by DWT and represented the document based on the reorganization of its terms by the analysis of the correlation between them.

Different from classifying and retrieving—fully described in the literature—text visualization techniques were grounded on DWTs intending to give users quick access to the relevance of certain

documents according to their respective interests. Although DWTs were not the only technique adopted, Wong et al. (2003) and Miller et al. (1998) achieved good results for helping the reader's comprehension by rating topics.

One of the most recent examples of Wavelets on text description is toward user recommendation. In Arru et al. (2013), it is shown that by signalizing key terms, mainly hashtags on Twitter, it is possible to discover people who write about the same issue on OSN and, therefore, recommend them to each other.

A wavelet transform can be considered a multi-resolution analysis, where a signal is decomposed into two sub-signals with different scales (Thaicharoen et al. 2008). These sub-signals are approximation and detail expressed by a linear combination of scaling functions a linear combination of wavelet functions, respectively.

This work uses the wavelet transform mostly because, among its advantages, it is (i) independent of the data, i.e., with wavelets there is no need of a syntax analysis, (ii) is able to remove the correlation from a generic large dataset, and (iii) is a fast algorithm (Xexéo et al. 2008). The discrete wavelet transformation extracts the most significant information from the data under analysis, avoiding unnecessary processing and providing an uncorrelated set of values that are more adequate for the proposed classification task. Along these lines, besides feasible computational cost for stream scenario, the DWT emphasize features capable of distinguishing human, legitimate robot or malicious robot toward frequency of terms and tweeting rate based on a small number of tweets. The frequency, e.g., terms and posting, is one of the most important characteristics to understand the bot behaviour (Chu et al. 2012; Arru et al. 2013; Tavares et al. 2017). Chu et al. (2012) applied the time provided by the social network and text analysis, and Tavares et al. (2017) focused only on the frequency by computing day of the week, hour and some statistical inferences of periodicity to bot discovering. All of them worked with frequency information provided by OSN. By applying DWT, we can obtain the frequency information only from the text through a feature vector designed with prevalence, regularity, magnitude, and scale of user behaviour, without any additional information from OSN.

4 PROPOSED APPROACH

The proposed approach was modeled in five steps: Acquisition, Profiling Setup, Features Extraction, Feature Selection, and Classification, as we can see in Figure 1. Our model is suitable to any OSN, i.e., the Acquisition step can be adapted according to each OSN API. Classification step is also flexible. In this work, we adopted Random Forests (RFs) as classifiers. This choice was based on works such as Singh et al. (2014) and del Río et al. (2014), which reported good results by applying RFs in big data environments and Igawa et al. (2016b) that treats specifically bots on OSNs.

4.1 Acquisition and Profiling Setup

The acquisition step consists of getting data from an OSN to create a textual data set. Our model requires the collected textual set to be grounded on one main subject that people are writing about, such as NBA playoffs or election campaigns. This requirement is necessary, because wavelet-based text mining takes analysis of signalized key-terms to obtain any further knowledge from it Igawa et al. (2016a).

In case the dataset is not grounded on one main term, we suggest the extraction of relevant terms. Barbon et al. (2016b) suggest the use of Louvain method for topic modeling, which is the main research area focused on find keywords in textual content.

Considering the use of a supervised machine-learning approach, we need a labelled dataset to induce a method. By the use of a dataset labelled as Human, Legitimate Bot, or Malicious Bot for

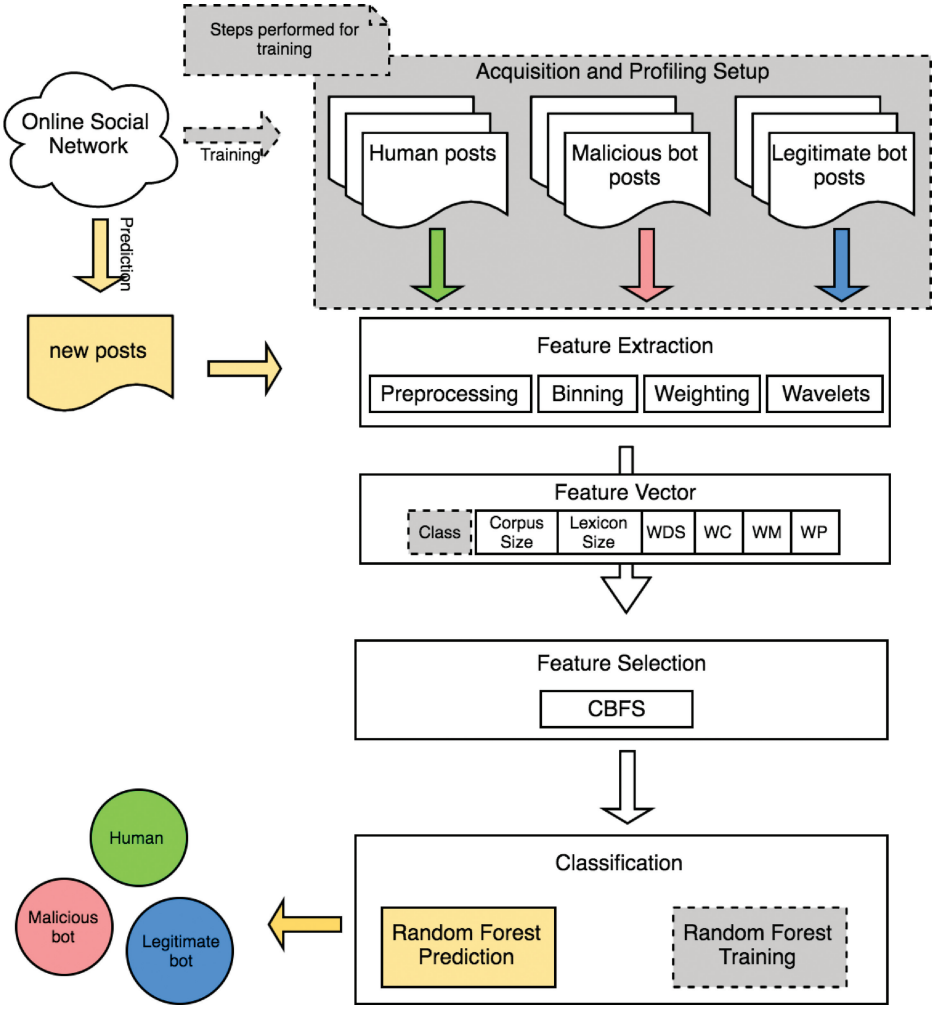

Fig. 1. The proposed approach pipeline: Acquistion and Profiling Setup, Feature Extraction, Feature Vector, Feature Selection, and Classification.

each user, we need to create a user profile. The labelling followed the rules consolidated in related works:

- —Malicious bots: automated accounts with spam-like content, usually redundant and repetitive;
- —Legitimate bots: automated accounts with harmless content, usually posted with the use of auxiliary software;
- —Human: non-automated accounts with intelligent and original content.

The text of each user is concatenated cumulatively following the profile-based paradigm and a textual repository with several users posts. This step is called Profiling Setup and provides the data that will be processed to obtain the feature vector that describes each class (Potha and Stamatatos 2014).

4.2 Binning

We used the representation described in Park et al. (2002) to transform the document d in a discrete vector. Documents consist of a slice of text posted by a single user with limited length. This process, called binning, transforms d by mapping it into a set of term signals. A term signal consists of a sequence of values that represents a term occurrence in a particular section of a document. Each vector's element, called *bin*, represents a portion of d. These bins identify the number of occurrences of the tth term in each portion, as shown in Equation (1):

$$\widetilde{f}_{d,t} = [f_{d,t,0}, f_{d,t,1}, \ldots, f_{d,t,b}, \ldots, f_{d,t,n-1}]. \tag{1}$$

4.3 Weighting

Weighting is an approach that complements term-signal representation. Many studies were dedicated to adjust term signal representation with weights, as can be seen in Purwitasari (2008), Park (2003), and Arru et al. (2013). As stated in Thaicharoen et al. (2008), term signal with mathematical transforms, such as DWT, is better than traditional vector space representation for information retrieval and document classification (Pryczek and Szczepaniak 2006). In this work, the Lexicon-based Coefficient Attenuation (LBCA) weighting scheme was used, being τ_d and $\bar{\tau}_d$, respectively, the *lexicon* size and the average *lexicon* size of d; i.e.,

$$w_{d,t,b} = \widetilde{f}_{d,t,b} \cdot \tau_d / \bar{\tau}_d. \tag{2}$$

LBCA achieves a better description than attenuating the inputs based on *lexicon size* (Igawa et al. 2016b). After this step, the numeric vector containing frequencies should become a vector containing weighted frequencies, as shown in Equation (3):

$$\widetilde{w}_{d,t} = [w_{d,t,0}, w_{d,t,1}, \ldots, w_{d,t,b}, \ldots, w_{d,t,n-1}]. \tag{3}$$

4.4 Wavelets

In this work, the signals provided by the weighting process, ($\widetilde{w}_{d,t}$), are decomposed using DWTs. Among different wavelet families available, Daubechies with support-size 4 was selected. We have selected this family due to its simplicity in terms of comprehension and implementation as stated by Barbon et al. (2007).

An important advantage of applying DWTs is the spectral analysis they provide. Basically, a DWT decomposes a signal into subsignals of different sizes and resolutions. By doing so, it is possible to describe patterns and use these subsignals as features to Machine-learning algorithms. Particularly, the DWT-based feature vectors are more likely to contain information capable of distinguishing between the behaviour of bots and humans, because, while filtering, both the tweeting rate and the frequency at which possible key terms were used are easily captured. This justifies the adoption of such a relevant tool.

From this point onwards, a document, i.e., a set of tweets, d, produced by u, is represented as a set of spectrum signals containing wavelet components:

$$\begin{aligned} \widetilde{\zeta}_{d,t} &= \mathit{Discrete\ Wavelet\ Transform}(\widetilde{w}_{d,t}) \\ &= [\zeta_{d,t,0}, \zeta_{d,t,1}, \ldots, \zeta_{d,t,b}, \ldots, \zeta_{d,t,n-1}]. \end{aligned} \tag{4}$$

4.5 Feature Vector

Our classifications are based on text mining approaches, in fact, only text is needed. But the Feature Vector is composed by text information described as:

—Class, consisting in Human, Legitimate Bot, or Malicious Bot;
—Corpus Size, i.e., total number of words in a document;

—Lexicon Size, i.e., total number of unique words in a document;
—Wavelet Component (WC), i.e., a set of spectrum that represents a document as calculated in Equation (4);
—Wavelet Magnitude (WM), which is obtained by applying an absolute function on each wavelet component: $H_{d,t,b} = |\zeta_{d,t,b}|$, described in Park (2003).
—Wavelets Phase (WP), which is obtained by checking the signal of each wavelet component: $\phi_{d,t,b} = \frac{\zeta_{d,t,b}}{H_{d,t,b}}$, as in Park (2003).
—Wavelet Domain Score (WDS), presented in Park (2003) is also used in this work due to its capacity to describe text relevance in relation to a given key terms search, T. By calculating the score from a certain user tweets, we are able to know how relevant the text produced is in relation to the given words T. To calculate such a value, we use the Phase and the Magnitudes described in preceding paragraphs. The detailed calculations are shown in Equations (5), (6), and (7), where $\#(T)$ is the number of key terms search used:

$$\bar{\Phi}_{d,b} = \left| \frac{\sum_{t \in T} \phi_{d,t,b}}{\#(T)} \right|, \tag{5}$$

$$s_{d,b} = \bar{\Phi}_{d,b} \cdot \sum_{t \in T} H_{d,t,b}, \tag{6}$$

$$s_d = \sum_{i \in b} s^2_{d,b_i}. \tag{7}$$

4.6 Feature Selection

Before classifying the data, a Correlation-based Feature Subset Selection (CBFS) proposed by Hall (1999) was applied to reduce the dimensionality of the Feature Vectors. Addressing Feature Selection, the literature offers a collection of different algorithms. In this work, we choose CBFS considering that the algorithm is a member of Feature Selection Algorithms called Subset Selection. This implies that the output of the algorithm is a subset of features from the original one, without creating any more features. To know the selected features is also important to have a better comprehension of the problem.

4.7 Classification

We propose a general framework used for online classification and offline training. A classification problem consists of taking an input vector with data and deciding which of N classes they belong to. It follows a supervised learning process based on training from instances of each class (Marsland 2009). The most important learning feature is the *generalization*: the algorithm should produce sensible output for inputs that were not encountered during learning.

To evaluate our proposal for user writing style's features, we suggest the Random Forest (RF) method. RF method was proposed by Breiman (2001) combining Decision Tree classifiers in an ensemble. This classifier was selected due to high performance in different multi-class scenarios.

4.8 Numeric Example

This subsection exemplifies several steps described on Section 4. Profiling is illustrated on top of Figure 2. As stated, profiling is basically a concatenation of all messages available from only one user. Thus, one document is generated for each user. The next step, Binning, transforms the profiled document in a numeric vector called term signal. An interesting point of using term signals is that each entry in the numeric vector corresponds to the occurrence of that term in that part of the document. Hypothetically, considering a term t (i.e., *superbowlxlviii*), it occurred 1 time in the

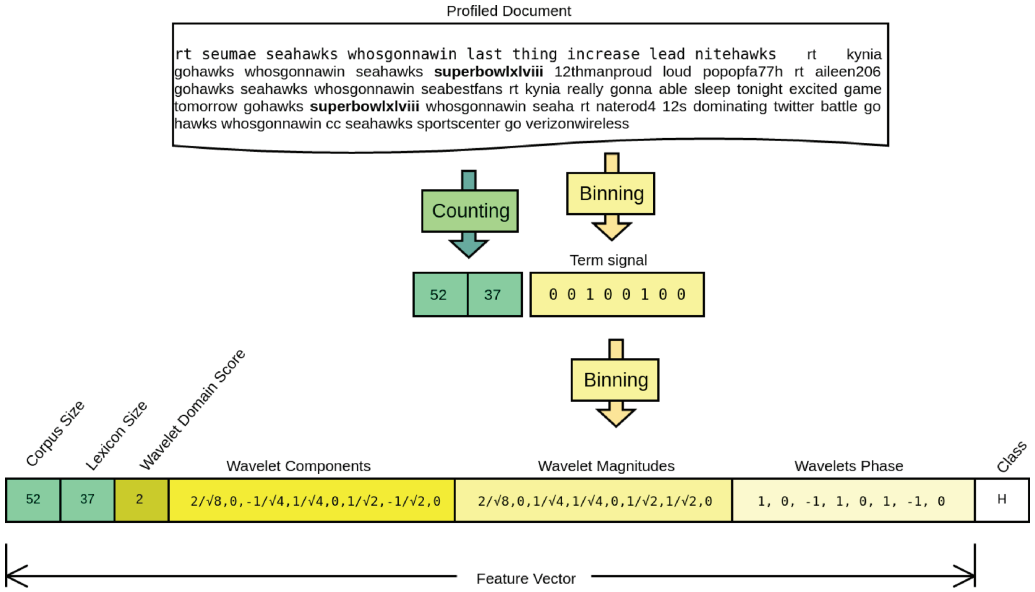

Fig. 2. Example Overview: Profiling, Binning to obtain the term signal, and DWT features extraction to achieve Wavelet Components, Wavelet Magnitude, and Wavelet Phase.

first portion of the document and one time in the second part of the document, and so on. In the example illustrated with Figure 2, we used eight bins to discretize the profiled document. Along the term signal, two additional features are used: the corpus size and lexicon size.

As proposed in Section 4.5, the complete feature vector is formed by the class, corpus size, lexicon size, WDS, WC, WM, and WP. Regarding wavelets features, the wavelets components are basically the values obtained by performing a wavelet transformation of the term signal. Wavelets magnitudes are obtained by performing a absolute function on each entry on the wavelets components. The last wavelets features, phase components, are obtained by taking in consideration the signal of each component.

The advantage of using wavelets on textual representation is the meaning of each entry in the wavelets components. Each entry carries an interesting feature. For example, in Figure 2, the first value on wavelets components is $\frac{2}{\sqrt{8}}$. Two is the exact value of the sum of all entries in the term signal (0,0,1,0,0,1,0,0). The second wavelet component entry is 0 and 0 is the difference between the sum of first half of term signal (0,0,1,0) and the sum of second half of term signal (0,1,0,0). The third wavelet component entry has the difference between the sum of the first quarter of term signal and the sum of second quarter of the term signal on its numerator. Concretely, each entry of the wavelet components represents the difference between a part of the original term signal. This way, wavelets features aid the feature extraction by giving information about the organization of the frequencies components considering each desired term.

5 EXPERIMENTAL EVALUATION

5.1 Evaluation Methodology

As stated before, our model requires a textual set grounded on the main subject or theme. To ensure that the topic used is discussed enough on an OSN, we used two different datasets. The first one, with a single theme addressing the *Super Bowl XLVIII* event, is a game about *Denver Broncos*

Table 1. An Example of Single Tweet from a Member of Each Class in Experiments

Content	Class
I can't sit still!!! It's mutha fuckin Superbowl time!!! #**WhosGonnaWin** #**Seahawks** #**SB48**"	Human
Fox News to interview Obama before SuperBowl http://t.co/FfNkUnrYCk	Legitimate Bot
@PLDTHome DENVER BRONCOS FOR SURE!!! #FibrSuperBowl #**SB48** 310	Malicious Bot

and *Seatle Seahawks* and was called $Dataset_S$. The second dataset, $Dataset_M$, was composed by miscellaneous themes obtained randomly from Twitter. For this second one, the main subject was extracted based on the simple frequency of terms after preprocessing.

Particularly, $Dataset_S$ is a subsample of a complete base available online.[1] The set also contains usual information retrieved through Twitter API[2] such as latitude, longitude, and if post was re-tweeted. $Dataset_M$ was formed by the use of Twitter API and contains simple information: post time and content. It is important to highlight that the datasets (single and miscellaneous) were used, because they present hundreds of legitimate and malicious bots. The proposed approach is applicable in any dataset presenting both classes of accounts and a main subject.

To perform our experiments, each dataset sample was labelled as human, legitimate or malicious bot. This was done to perform a supervised machine-learning method, i.e., Random Forest. In this work, we perform a manual tagging task as in Chu et al. (2012), Igawa et al. (2016b), and Wald et al. (2013b) to set the post's user class addressing automatic behaviour on OSNs. Chu et al. (2012) describe the importance of spam and non-spam data content to determine the type of user among human, cyborg or bot. Our work does not handle the cyborg type (a hybrid of human and bot), but introduces a division of bots into legitimate and malicious. The bots carry spam tweets and spam external URLs on posts. Another type of advanced spam bots (malicious) intentionally injects non-spam tweets to confuse the recognition. Both types of bot tend to present a much more formal use of language than humans, furthermore, most human tweets do not carry spam. Within bots, a malicious bot tends to present a redundant behaviour. Table 1 shows in the third row a post from a malicious bot. This is one example of spam-like behaviour. The term "310" means that this same content has been posted for 310 times. The writing style, mainly from malicious bots, is redundant, so it is not necessary to have thousands of users. Works as Cingiz et al. (2015) obtained satisfactory results from selected data amount similar to used in our experiments. Thus, we analyze each user's post to identify the type of tweets in both datasets and attach a label. Table 2 shows the datasets specification after tagging.

One important observation about legitimate bot is illustrated at the second row of Table 1. There, it is possible to note that a legitimate bot matches the initial goal of OSN like Twitter, which is to spread news. Such kind of account is usually known by a very formal use of language rules.

In summary, Legitimate bots (e.g., sample 2 in Table 1), corresponding to their original objective, disseminate news and usually well-written texts regarding grammar language and lexical issues. While malicious bots tend to spam the same message, as shown in sample 3, the same tweet was

[1]Data set used in this work is available at: http://www.techtunk.com/index.php?dove=downloads.
[2]Available Twitter APIs: https://dev.twitter.com/overview/api/twitter-libraries.

Table 2. Experimental Dataset Description: Dataset I was Composed by a Single Theme Tweets and Dataset II was Composed by Miscellaneous Theme Tweets Tagged Among Human (H), Fraudulent Bot (FB), and Legitimate Bot (LB)

	$Dataset_S$			$Dataset_M$		
	H	FB	LB	H	FB	LB
Total users	30	40	30	35	90	42
Max tweets/user	27	44	55	28	54	73
Avg tweets/user	13	26	36	15	31	47
Min tweets/user	1	3	7	5	5	1

posted hundreds of times in Table 1. Considering humans, it is possible to realize a third example of writing style differing from both types of bots.

Another issue toward the developed model is the use of key terms to retrieve numerical vectors that will be input to DWTs. In this work, to avoid retrieving non-significant terms, we performed stopwords removal. Our list of terms to be removed included: articles, prepositions, conjunctions and adverbs. This way, removing words that do not represent useful content decreases the size of the arrays and increases the performance of any method performed on them. Such words are called Stopwords[3] (Feldman and Sanger 2007). It is usual to consider pronouns, articles and prepositions as stopwords. In practice, all words considered stopwords are gathered in list and, then, are removed from all arrays of words in the experiment.

Concretely, in this experiment, we perform a classification intending to separate the three types of writing style shown in Table 1, where the terms in bold represent the signalized terms. These signalized terms are processed and several features are extracted obtaining a descriptor vector of a user writing style. The machine-learning classification algorithm used in experiments was RF implemented in the `R` environment[4]. The corresponding package used to implement the algorithm was `randomForest` with standard parameters. The RF was performed 100 times using a tenfold Cross Validation re-sampling strategy on the same data partitions for both datasets. We extracted the performance measures of predictive accuracy balanced according to the number of classes.

5.2 Results and Discussion

Given the perspective of developing a solution adequated to OSN scenario, we explore the influence of the number of users in ML method induction, the number of tweets per users to obtain high accuracy and what are the best features in several scenarios. Finally, the overall accuracy to classify the user account based on its writing style.

5.2.1 Number of Users Importance. We conduct an analysis to evaluate how discriminant was our user writing style's features and the importance of the number of users to induce the Machine-learning method. This comparison was performed with $Dataset_S$ and $Dataset_M$, varying the number of users from 3 to 69. Figure 3 shows the accuracy variation from each number of users. It is possible to detect a point where both datasets obtain a similar accuracy; this point is highlighted by horizontal and vertical green lines in Figure 3. Thus, we can state that, based on 20 users, it was possible to reach an stability point to perform the user classification with 85% of accuracy. By

[3] http://dev.mysql.com/doc/refman/5.7/en/fulltext-stopwords.html.
[4] https://www.r-project.org/.

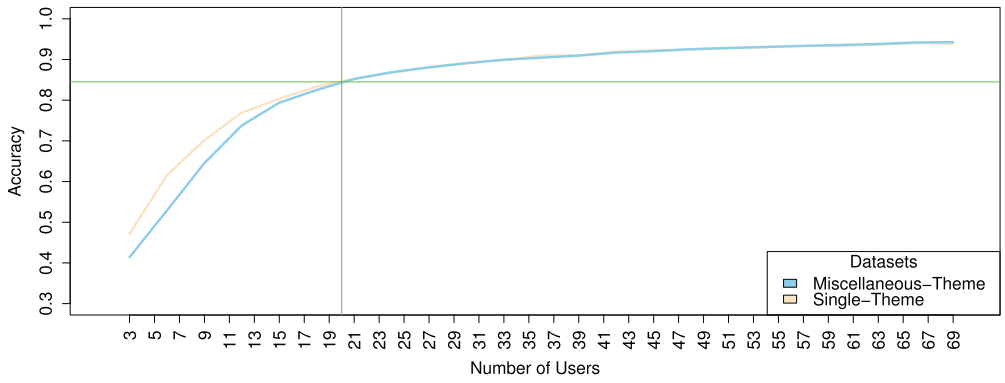

Fig. 3. Comparison of different number of users in training step and the accuracy obtained. The green lines intersection highlight the stability accuracy point to classify users based on a single ($Dataset_S$) or miscellaneous theme ($Dataset_M$).

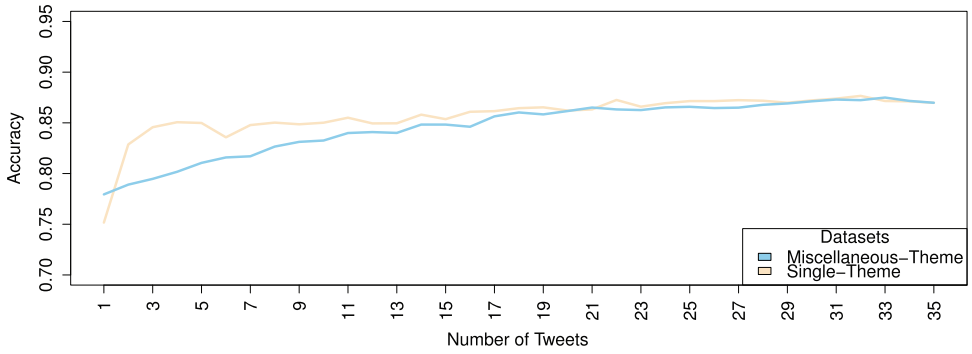

Fig. 4. Comparison of number of tweets influence in $Dataset_S$ (single theme) and $Dataset_M$ (miscellaneous theme).

the use of 69 users, it was possible to achieve 94%; however, from 50 users (93.2%) the accuracy improvement was less than 1% in both datasets.

5.2.2 Number of Tweets Per User. Considering the influence of textual quantity per user to create an effective ML model to classify the writing style, we perform an experiment with a different amount of tweets. Varying from 1 to 35 tweets per user, based on 20 users, the results showed that by the use of 28 tweets the accuracy (86%) had an attenuated growth and equal precision rate for both datasets, as Figure 4 shows.

Regarding a small number of tweets, the $Dataset_S$ shows better accuracy rate (>80%) just with two tweets. $Dataset_M$ required more tweets to achieve an accuracy rate similar to the single theme dataset. This fact is justified due to wavelets features relation to the term distribution and frequency, in other words, $Dataset_M$ needs more terms to express the user writing style.

5.2.3 Features Selected. In regard to evaluating the user's writing style features towards to achieve the relevant ones, we perform the CFS algorithm. It is important to highlight that we assess the merit of a subset considering different number of tweets. This was done to observe the presence of wavelet features along with lexicon and corpus size. Figure 5 shows, for both datasets, the use of wavelet features. In the right side of the triangle, it is possible to observe the best features

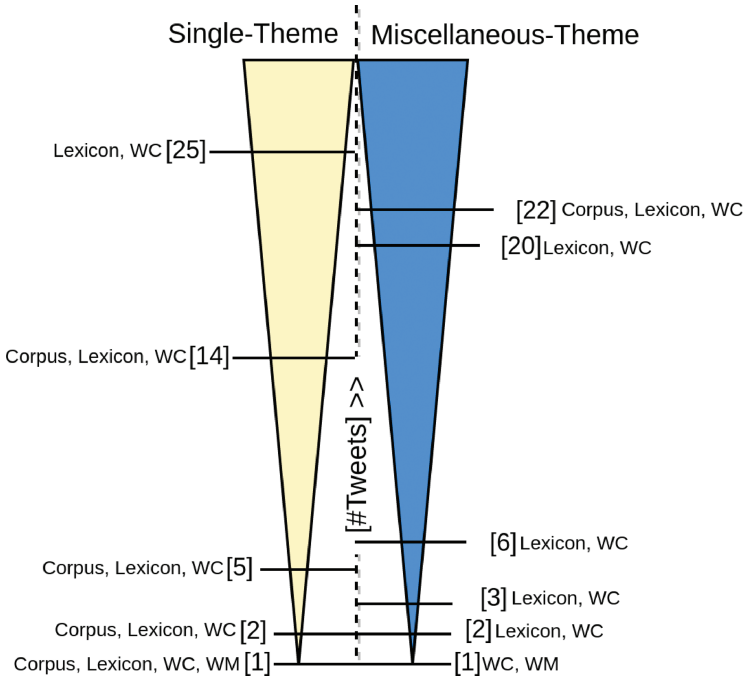

Fig. 5. Best features subset achieved by CFS algorithm varying the number of tweets in single ($Dataset_S$) and miscellaneous theme ($Dataset_M$) datasets.

subset to predict a $Dataset_M$. In opposite side, we have the single theme features, $Dataset_S$. There are several compositions; however, from the bottom of the triangle (Figure 5) it is possible to see the presence of Wavelets Coefficient (WC) and Wavelets Magnitude (WM) for both datasets.

With the increase of tweets, we observe that WM was not in the best subset. For instance, by the use of two tweets, the selected features were WC, corpus, and lexicon size for the single theme; therefore, for the miscellaneous theme, the subset was formed by lexicon size and wavelet components. Another consideration touches the larger number of tweets, which by only the use of lexicon and WC it was possible to achieve high accuracy (>85%). Consequently, we could confirm that WC was relevant to compose the user writing style and WM mainly with a few tweets. Wavelets Phase (WP) was not selected to compose no feature subsets.

5.2.4 Classification Performance. An overview of our results is presented in Figure 6, that shows a high accuracy for both datasets, 95.05% and 93.88%, i.e., $Dataset_S$ and $Dataset_M$, respectively. Based on standard deviation lower than 3.5% for 100 repetitions, we state that we have low variation and our proposed approach is suitable for distinguishing among human, legitimate bot and malicious bots. The premise of a specific term to extract WC, WM, WP, and WDS was highlighted by the higher accuracy of $Dataset_S$. However, the extraction of relevant term in a $Dataset_M$ could be a viable alternative to support the use of the proposed approach.

Our final discussion is toward the available results in the literature. As stated before, there is no related work focused on text mining classification in humans, legitimate bots and malicious bots. However, as considered in Section 2, Chu et al. (2012) and Chu et al. (2010) have classified accounts on Twitter by using at least 100 posts per user and more features than only text aiming at the classification in Humans, Cyborgs, and Bots. Our technique was able to achieve suitable results

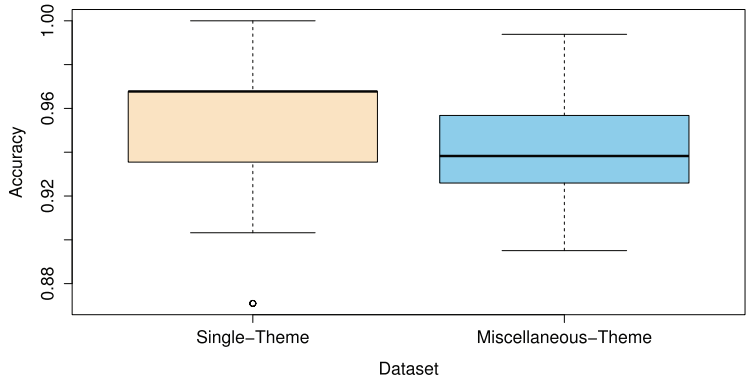

Fig. 6. Box plot of Random Forest method applied to classification of user account in Human, Malicious, or Legitimate Bot by different scenarios: single ($Dataset_S$) and miscellaneous ($Dataset_M$) theme dataset.

with a small number of accounts toward built a classification model with competitive performance. Furthermore, the number of tweets requested in our solution was greatly inferior to the other solutions. Chu et al. (2010) employed 100 tweets per user to cope with classification task.

The first comparison concerns true positive rate to classify humans accounts. On their respective data sets, Chu et al. (2010) reached 94.90% and Chu et al. (2012) obtained 98.6%. Considering a comparison in terms of bots, Chu et al. (2010) achieved 93.70% and Chu et al. (2012) obtained 97.60%. In our proposed approach, we could detect three different types: Human, Legitimate, and Malicious Bot. Our method achieved an average accuracy of 94.47%, considering both datasets. As stated before, our model only needs textual features, thus, it is applicable to any OSN. Despite the fact that we did not obtained rates superior to those similar, we still consider our results satisfactory once the proposed model and related works are supposed to be applied in different situations.

6 CONCLUSION

In this study, we have proposed an algorithm for classifying authors as being a human, a legitimate robot, or a malicious robot, in OSNs. The algorithm was based on Discrete Wavelet Transform to obtain a pattern of writing style embedded in post contents. Experiments have been conducted by classifiers with two different datasets: single and miscellaneous theme. It was observed that the proposed method yields the high average classification accuracies of 94.47% for both datasets. Considering the results, the text-based model we have developed gives promising accuracies in classifying the user type based on its writing style. We believe that the proposed algorithm can be very helpful to combat frauds in OSN. Further exploration of different machine-learning approaches can yield more interesting results.

REFERENCES

Charu C. Aggarwal. 2011. An introduction to social network data analytics. In *Social Network Data Analytics*. Springer, 1–15.

Mishari Almishari, Dali Kaafar, Gene Tsudik, and Ekin Oguz. 2014. Are 140 characters enough? A large-scale linkability study of tweets. *arXiv:1406.2746* (2014).

Giuliano Arru, Davide Feltoni Gurini, Fabio Gasparetti, Alessandro Micarelli, and Giuseppe Sansonetti. 2013. Signal-based user recommendation on twitter. In *Proceedings of the 22nd International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 941–944.

Seyed-Ali Bahrainian and Andreas Dengel. 2013. Sentiment analysis and summarization of twitter data. In *Proceedings of the 2013 IEEE 16th International Conference on Computational Science and Engineering (CSE'13)*. IEEE, 227–234.

Sylvio Barbon, Rodrigo Capobianco Guido, Shi-Huang Chen, Lucimar Sasso Vieira, and Fabricio Lopes Sanchez. 2007. Improved dynamic time warping based on the discrete wavelet transform. In *Proceedings of the 9th IEEE International Symposium on Multimedia Workshops (ISMW'07)*. IEEE, 256–263.

Sylvio Barbon, Rodrigo Augusto Igawa, and Bruno Bogaz Zarpelão. 2016a. Authorship verification applied to detection of compromised accounts on online social networks. *Multi. Tools Appl.* (2016), 1–21. DOI: http://dx.doi.org/10.1007/s11042-016-3899-8

Sylvio Barbon, Guilherme Sakaji Kido, and Rodrigo Augusto Igawa. 2016b. Recognition of compromised accounts on twitter. In *Proceedings of the Annual Conference on Brazilian Symposium on Information Systems*, Vol. 1. SBC, 353–360.

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

Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. The socialbot network: When bots socialize for fame and money. In *Proceedings of the 27th Annual Computer Security Applications Conference*. ACM, 93–102.

Leo Breiman. 2001. Random forests. *Mach. Learn.* 45, 1 (2001), 5–32.

Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. 2012. Aiding the detection of fake accounts in large scale social online services. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (NSDI'12)*. USENIX Association, Berkeley, CA, 15–15.

Liang Chen, Yipeng Zhou, and Dah Ming Chiu. 2015. Analysis and detection of fake views in online video services. *ACM Trans. Multi. Comput. Commun. Appl.* 11, 2s (2015), 44.

Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2010. Who is tweeting on Twitter: Human, bot, or cyborg?. In *Proceedings of the 26th Annual Computer Security Applications Conference*. ACM, 21–30.

Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Trans. Depend. Secure Comput.* 9, 6 (2012), 811–824.

Mustafa Özgür Cingiz, Banu Diri, and Göksel Biricik. 2015. Am I typing fresh tweets: Detecting up-to-dateness and worth of categorical information in microblogs. *Expert Syst. Appl.* 42, 12 (2015), 5256–5263.

Mário Cordeiro. 2012. Twitter event detection: Combining wavelet analysis and topic inference summarization. In *Proceedings of the Doctoral Symposium on Informatics Engineering (DSIE'12)*.

Sara del Río, Victoria López, José Manuel Benítez, and Francisco Herrera. 2014. On the use of mapreduce for imbalanced big data using random forest. *Info. Sci.* 285 (2014), 112–137.

Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2013. COMPA: Detecting compromised accounts on social networks. In *Proceedings of the Network and Distributed System Security Symposium (NDSS'13)*.

Ronen Feldman and James Sanger. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

S. Fong, Yan Zhuang, and Jiaying He. 2012. Not every friend on a social network can be trusted: Classifying imposters using decision trees. In *Proceedings of the 2012 International Conference on Future Generation Communication Technology (FGCT'12)*. 58–63.

Hongyu Gao, Yan Chen, Kathy Lee, Diana Palsetia, and Alok Choudhary. 2012. Towards online spam filtering in social networks. In *Proceedings of the 19th Annual Network and Distributed System Security Symposium (NDSS'12)*.

Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y. Zhao. 2010. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC'10)*. ACM, New York, NY, 35–47. DOI: http://dx.doi.org/10.1145/1879141.1879147

Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. @Spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS'10)*. ACM, New York, NY, 27–37. DOI: http://dx.doi.org/10.1145/1866307.1866311

Mark A. Hall. 1999. *Correlation-based Feature Selection for Machine Learning*. Ph.D. Dissertation. The University of Waikato.

Ammar Hassan, Ahmed Abbasi, and Daniel Zeng. 2013. Twitter sentiment analysis: A bootstrap ensemble framework. In *Proceedings of the 2013 International Conference on Social Computing (SocialCom'13)*. IEEE, 357–364.

Liang-Chi Hsieh, Ching-Wei Lee, Tzu-Hsuan Chiu, and Winston Hsu. 2012. Live semantic sport highlight detection based on analyzing tweets of twitter. In *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo (ICME'12)*. IEEE, 949–954.

Rodrigo Augusto Igawa, Alex Almeida, Bruno Zarpelão, and Sylvio Barbon Jr. 2016a. Recognition on online social network by user's writing style. *iSys-Revista Brasileira de Sistemas de Informação* 8, 3 (2016), 64–85.

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. 2016b. Account classification in online social networks with LBCA and wavelets. *Info. Sci.* 332 (2016), 72–83.

Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. 2014. Detecting suspicious following behavior in multimillion-node social networks. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 305–306.

L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos. 2013a. Understanding user behavior in online social networks: A survey. *IEEE Commun. Mag.* 51, 9 (September 2013), 144–150. DOI: http://dx.doi.org/10.1109/MCOM.2013.6588663

Lei Jin, James B. D. Joshi, and Mohd Anwar. 2013b. Mutual-friend-based attacks in social network systems. *Comput. Secur.* 37 (2013), 15–30.

Sara Keretna, Ahmad Hossny, and Doug Creighton. 2013. Recognising user identity in twitter social networks via text mining. In *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC'13)*. IEEE, 3079–3082.

Yan Li, Yingjiu Li, Qiang Yan, and Robert H. Deng. 2015. Privacy leakage analysis in online social networks. *Comput. Secur.* 49 (2015), 239–254.

Sixuan Ma and Zheng Yan. 2015. PSNController: An unwanted content control system in pervasive social networking based on trust management. *ACM Trans. Multi. Comput. Commun. Appl.* 12, 1s (2015), 17.

Stephen Marsland. 2009. *Machine Learning: An Algorithmic Perspective* (1st ed.). Chapman & Hall/CRC.

Nancy E. Miller, Pak Chung Wong, Mary Brewster, and Harlan Foote. 1998. TOPIC ISLANDS TM-a wavelet-based text visualization system. In *Proceedings of Visualization'98*. IEEE, 189–196.

Zachary Miller, Brian Dickinson, William Deitrick, Wei Hu, and Alex Hai Wang. 2014. Twitter spammer detection using data stream clustering. *Info. Sci.* 260 (2014), 64–73.

Mohamed M. Mostafa. 2013. More than words: Social networks text mining for consumer brand sentiments. *Expert Syst. App.* 40, 10 (2013), 4241–4251.

Jun Pang and Yang Zhang. 2015. A new access control scheme for facebook-style social networks. *Comput. Secur.* 54 (2015), 44–59.

Laurence A. F. Park. 2003. *Spectral-based Information Retrieval*. Ph.D. Dissertation. The University of Melbourne.

Laurence A. F. Park, Marimuthu Palaniswami, and Kotagiri Ramamohanarao. 2002. A new implementation technique for fast spectral-based document retrieval systems. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'03)*. IEEE, 346–353.

Nektaria Potha and Efstathios Stamatatos. 2014. A profile-based method for authorship verification. In *Artificial Intelligence: Methods and Applications*. Springer, 313–326.

M. Pryczek and P. S. Szczepaniak. 2006. On textual documents classification using fourier domain scoring. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06)*. 773–777. DOI: http://dx.doi.org/10.1109/WI.2006.125

Diana Purwitasari. 2008. A study on ranking method in retrieving web pages based on content and link analysis: Combination of fourier domain scoring and pagerank scoring. *Jurnal Ilmiah Teknologi Informasi* 7, 1 (2008), 9–18.

Diana Purwitasari, Yasuhisa Okazaki, and Kenzi Watanabe. 2007. A study on web resources_ navigation for e-learning: Usage of fourier domain scoring on web pages ranking method. In *Proceedings of the 2nd International Conference on Innovative Computing, Information and Control (ICICIC'07)*.. IEEE, 458–458.

Mohamed Shehab, Anna Squicciarini, Gail-Joon Ahn, and Irini Kokkinou. 2012. Access control for online social networks third party applications. *Comput. Secur.* 31, 8 (2012), 897–911.

Kamaldeep Singh, Sharath Chandra Guntuku, Abhishek Thakur, and Chittaranjan Hota. 2014. Big data analytics framework for peer-to-peer botnet detection using random forests. *Info. Sci.* 278 (2014), 488–497.

Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. 2014. Stream-based active learning for sentiment analysis in the financial domain. *Info. Sci.* 285 (2014), 181–203.

Gabriel Marques Tavares, Saulo Martielo Mastelini, and Sylvio Barbon Junior. 2017. User classification on online social networks by post frequency. In *Proceedings of the 12th Brazilian Symposium on Information Systems on Brazilian Symposium on Information Systems, Volume 1 (SBSI'17)*. Brazilian Computer Society, 464–471.

Supphachai Thaicharoen, Tom Altman, and Krzysztof J. Cios. 2008. Structure-based document model with discrete wavelet transforms and its application to document classification. In *Proceedings of the 7th Australasian Data Mining Conference, Volume 87*. Australian Computer Society, 209–217.

Randall Wald, Taghi M. Khoshgoftaar, Amri Napolitano, and Chris Sumner. 2013a. Predicting susceptibility to social bots on Twitter. In *Proceedings of the 2013 IEEE 14th International Conference on Information Reuse and Integration (IRI'13)*. IEEE, 6–13.

Randall Wald, Taghi M. Khoshgoftaar, Amri Napolitano, and Chris Sumner. 2013b. Which users reply to and interact with twitter social bots? In *Proceedings of the 2013 IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI'13)*. IEEE, 135–144.

Zhi Wang, Wenwu Zhu, Xiangwen Chen, Lifeng Sun, Jiangchuan Liu, Minghua Chen, Peng Cui, and Shiqiang Yang. 2013. Propagation-based social-aware multimedia content distribution. *ACM Trans. Multi. Comput. Commun. Appl.* 9, 1s (2013), 52.

Jianshu Weng and Bu-Sung Lee. 2011. Event detection in twitter. In*Proceedings of the International Conference on Web and Social Media (ICWSM'11)*. 401–408.

Pak Chung Wong, Harlan Foote, Dan Adams, Wendy Cowley, and Jim Thomas. 2003. Dynamic visualization of transient data streams. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'03)*. IEEE, 97–104.

Geraldo Xexéo, Jano de Souza, Patrícia F. Castro, and Wallace A. Pinheiro. 2008. Using wavelets to classify documents. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'08)*, Vol. 1. IEEE, 272–278.

Conghuan Ye, Hefei Ling, Zenggang Xiong, Fuhao Zou, Cong Liu, and Fang Xu. 2016. Secure social multimedia big data sharing using scalable JFE in the TSHWT domain. *ACM Trans. Multi. Comput. Commun. Appl.* 12, 4s (2016), 61.

Michele Zappavigna. 2011. Ambient affiliation: A linguistic perspective on Twitter. *New Media Soc.* 13, 5 (2011), 788–806.

Received November 2016; revised September 2017; accepted November 2017