Information Sciences 332 (2016) 72–83

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Account classification in online social networks with LBCA and wavelets

Rodrigo Augusto Igawa [a,*], Sylvio Barbon Jr [a], Kátia Cristina Silva Paulo [c], Guilherme Sakaji Kido [a], Rodrigo Capobianco Guido [b], Mario Lemes Proença Júnior [a], Ivan Nunes da Silva [c]

[a] Londrina State University, Rod. Celso Garcia Cid km 380, Londrina-PR, Brazil
[b] Instituto de Biociências, Letras e Ciências Exatas, Unesp - Univ Estadual Paulista (São Paulo State University), Rua Cristóvão Colombo 2265, Jd Nazareth, 15054-000, São José do Rio Preto - SP, Brazil
[c] Departament of Electrical Engineering, School of Engineering at São Carlos,University of São Paulo, 13566-590, São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 3 December 2014
Revised 18 August 2015
Accepted 18 October 2015
Available online 10 November 2015

Keywords:
Account classification
Wavelets
Online social networks
Multilayer perceptrons
Random forests

ABSTRACT

We developed a wavelet-based approach for account classification that detects textual dissemination by bots on an Online Social Network (OSN). Its main objective is to match account patterns with humans, cyborgs or robots, improving the existing algorithms that automatically detect frauds. With a computational cost suitable for OSNs, the proposed approach analyses the distribution of key terms. The descriptors, a wavelet-based feature vector for each user's account, work in conjunction with a new weighting scheme, called Lexicon Based Coefficient Attenuation (LBCA) and serve as inputs to one of the classifiers tested: Random Forests and Multilayer Perceptrons. Experiments were performed using a set of posts crawled during the 2014 FIFA World Cup, obtaining accuracies within the range from 94 to 100%.

1. Introduction

Online Social Networks (OSNs) are considered suitable environments in which to discuss and express thoughts on any subject [61]. Currently, OSNs are a relevant resource for exploring a diversity of subjects, such as customer relations management and public opinion evaluation. The knowledge obtained from OSNs, such as Twitter and Facebook, has been shown to be extremely valuable for marketing research companies, public opinion organizations, and other text mining purposes [3,47,60,63]. Since millions of opinions on a given topic are expressed with simplicity, posting methods provide rich, easy and unbiased content comprehension [19]. Thus, the contents of OSNs form a valuable dataset for decision making on marketing research, business intelligence, stock market prediction, and image monitoring [20,32].

The wide popularity of OSNs and their ease of access have resulted in the misuse of their services. In addition to privacy issues, OSNs face the challenge of dealing with undesirable users and their malicious activities, with spamming for product promotion being one of the most common [5]. An example that shows the presence of these undesirable users is Facebook's noticing an alarming 8.7% of fake users, which represents more than 83 million accounts worldwide [14]. Twitter users also suffer from the same issue. As a result of its popularity and the culture of following individuals with interesting content, Twitter has become a major target for marketing and social manipulation due to its use for business promotion, customer service, political campaigning, and emergency communication [11]. Spam accounts are also presently posting links to paid content, and users

* Corresponding author. Tel.: +5599059844.
E-mail address: igawa.rodrigo@gmail.com, igawa@uel.br (R.A. Igawa).

http://dx.doi.org/10.1016/j.ins.2015.10.039
0020-0255/

shill for companies while pretending to be their independent fans. On Twitter, social robots, called "bots", pretend to be human beings in order to gain followers and replies from target users and promote a product or agenda [54]. Frauds in OSNs, such as these, could lead to uncontrolled dissemination of fake information, inaccurate content, promotional ads, and phishing. Thus, OSN users could become victims of tricky scams or harassment, causing, due to their dissatisfaction, a decline of service [5,30].

In view of these discussions, the main goal in this paper is to describe an algorithm that performs account classifications in order to avoid the uncontrolled information stream in OSNs. Experiments were performed with a Twitter data set in order to determine the influence and performance of binning, weighting schemes, wavelet components, and different kinds of classifiers. The first experiment matches the accounts with humans, cyborgs, and bots. On the other hand, the second experiment classifies the accounts as belonging to a human or not. Due to the low computational costs required by OSN environments, Discrete Wavelets Transforms with (O(N)) are used instead of the classic Fast Fourier Transform (FFT) algorithms with (O(N log(N))) [33] to analyse Twitter posts, i.e., tweets. Additionally, the classifications are based on text mining approaches that require, in fact, only text as input. These approaches do not need extra features from the OSNs, such as age, city, or user frequency. A Multilayer Perceptron Artificial Neural Network and Random Forest learning are used as classifiers.

This paper is organized as follows. Section 1.1 exposits related work on frauds in OSNs and Section 1.2, the relations between text mining and wavelets. In Section 2 we present the proposed approach complemented with related concepts. Section 3 explains, in detail, the tests and experiments. In Section 4, the corresponding results are discussed. Lastly, in Section 5, we present the conclusions and limitations of this work.

1.1. Online social networks and frauds

Currently, OSNs are the subjects of much research. Treated as an environment, they have been shown to be relevant to a variety of research in the literature. For example, concerning the success of specific products, [32] developed method to evaluate customer satisfaction. Still with a similar goal, [52] showed that using OSNs can be very helpful even to small companies attempting to ascend. Focusing on educational ends, there is [6,7], which evaluated OSNs initial usages and explored ways to better apply OSNs to teaching proposals. Concerning network research, OSNs have also encouraged the development of a literature oriented towards DTNs. [24] studied a more efficient way, where mobile nodes would be delivering messages between each other. Still in terms of Delay Tolerant Networks (DTNs), mobile nodes have been investigated in papers like [21,25], where a selfish node might not cooperate with information distribution among the others.

Considering another different goal to research on OSN environments, this paper is limited to classifying accounts in OSNs. The fact that encouraged this research on automatic account classification, mainly fake user detection, is the known policy for removing OSN accounts [9]. OSN policies depend excessively on a real user to report fake ones. By the time the qualified company takes action about a suspicious account, it might be too late, even more when thousands of bots are involved [14]. However, bots are not always a problem. The research described in [11] has already analysed bots as being a double edged sword. Malicious bots spread malicious content and spamming, while legitimate bots generate a lot of benign tweets by delivering news. This is the reason why we want to detect both kinds of bots. The contribution from this work relies on the fact that non-automated tweets form a data set cleaned enough to perform any information retrieval, opinion mining, clustering, and classification task based solely on humans.

Some research on users' identity on OSNs has been carried out with the aid of classification tasks using classic data mining schemes. In these studies, bots, or fake accounts, were identified by performing a combination of selected features in order to compare the data patterns of real users with those of bots [11,14,53]. As to account classification on Twitter, some papers have discussed the type of users that interact with bots, who could be those responsible for organizing fake accounts or who should be warned about [54,55]. On Facebook, a similar experiment was performed in order to detect suspicious accounts from a pool of spamming activities [5,14].

Another approach in managing fake user identities is to perform a pure text mining method. This is formally called author identification, and is used to analyse types of unstructured text like emails, text messages, and online products [23,62]. On Twitter, this kind of procedure, called writeprint, was carried out in order to identify malicious users. In [23], it was shown that if the correct set of features was extracted, a very promising result could be achieved just by using a crawler to extract tweets and part-of-speech taggers.

Some other concerns about bots, spamming, and fraud have been developed. For instance, [51] presented a case in which Twitter bots were a way to attack real Twitter users and transform them into botnets. Worry over real friends is an important point, so, the [46] proposed the use of a tool that would be able to analyse whether or not a friend is an automated user. Other research has focused on what kind of target spammers usually look for [48], bot identification by user name [15], and the influence of a bot with many friends [22]. All of these papers analyse the concern about bots/fake users and the frauds created by these. Both threats, as described in [51], either spamming or even fake news, can lead to full fledge fraud. The present paper aims to propose a solution based completely on textual information. By using only text, we do not require any additional information, such as a complete profile and so on. Therefore, this model would be applicable to any OSN.

1.2. Text mining and the Discrete Wavelet Transform (DWT)

DWTs have been widely used in many different fields of science due to their extensive range of applications [28]. In image processing, wavelets were successfully applied to perform copyright protection schemes [26] and to frequency domain

encryption, ensuring secure and unbreakable forms [49]. De-noising, which is a classical DWT application [56] [37], has been a subject of research, as can be seen in [18,27]. DWTs have also been used in decision making in cancer diagnosis, as described in [42].

For text mining purposes, DTWs can be applied to different tasks, such as information retrieval, document classification, or text visualization. Particularly for information retrieval, DWTs are able to analyse term patterns, i.e., existing text representations that depict a term as a vector of frequencies of occurrences in a number of defined partitions of a document, at different levels of resolution [34]. Inspired by the same principles, the Discrete Cosine Transform (DCT) [35] and the Discrete Fourier Transform (DFT) [16,36,41] have also been used. At the same time, signal based approaches have been employed to retrieve Web-page contents. Link analysis, Page Rank Scoring, and content analysis based on the Fourier Domain Scoring as proposed in [34], have been combined to improve the results in search engines [39,40]. When word signals are used to represent the same documents, instead of the classic Vector Space Model, better results are obtained, as reported in [50]. The use of different signal models, but still based on wavelets, makes another possibility, as explored in [59].

Instead of signal ranking and signal classification, which are fully described in the literature, text visualization techniques intend to enable users to rapidly assess the relevance of documents to their specific interests by providing a tool to connect the text narrative to embedded themes. Although they have also been combined with other tools, visualization techniques have achieved good results in helping the reader's comprehension by rating topics and subject corpora [29,58].

2. Proposed approach

Since this is a paper involving classification, we need to use features that provide a good representation of the original problem. Considering that OSNs are dynamic and heterogeneous, we need an adequate computational cost for both the extracting and classifying steps, a weighting scheme appropriate for classification, and well-known classifiers in this large data scale scenario.

In this section, for which complementary information can be found in Fig. 1, the selected features and details of the algorithms are thoroughly described.

2.1. Profiling

The main idea behind the proposed algorithm is to bring the problem of user representation to a document representation model. In order to perform a wavelet based text mining strategy for account classification, textual content produced by a user in a set, from now on referred to as a *document*, requires a representation, similarly to the research reported in [38]. In an OSN context, definitions are required to explain how to represent each user's textual content as a document. Let $u \in U$ be a user in a set of users. Let $p \in P$ be a published text in a set of texts. A document d from user u is a set of posts p where $p.user == u$, as shown in Eq. 1. Basically, the document d is a cumulative concatenation of all samples from a single user [38].

$$Document(d, u) = \{p | p.user = u\} \tag{1}$$

2.2. Binning

We used the representation described in [34] to transform the document d into a discrete vector. This process, called binning, transforms d by mapping it into a set of term signals. A term signal is a sequence of values that represents a term's occurrence in a particular section of a document. It is computationally represented by a numeric vector. Each vector element, called a *bin*, represents a portion of d. These bins identify the number of occurrences of the term t in each portion, as shown in Eq. 2.

$$\widetilde{f}_{d,t} = [f_{d,t,0}, f_{d,t,1}, \ldots, f_{d,t,b}, \ldots, f_{d,t,n-1}] \tag{2}$$

Eq. 3 illustrates an example where d is represented by the occurrences of a term t in it. Thus, t has appeared twice in the first portion, i.e., *bin*, has not appeared in the second portion, and so on. Another issue addressed in this paper is the ideal number of frames to be used. In structured documents, eight frames have been used [33,34,37], but in web pages corpora, sixteen frames have been preferred [39].

$$\widetilde{f}_{d,t} = [2, 0, 3, 0, 2, 1, 1, 3] \tag{3}$$

2.3. Weighting

Weighting is an approach that complements term signal representation. Many studies have been dedicated to adjusting a term signal representation with weights, as can be seen in [2,33,39]. In the present paper, Table 1 describes all the weights used in experiments, with $s = 0.7$, W_d the document size, and $\bar{W}_d$ the average size of the documents, in numbers of words. Additionally, τ_d is the size of the document and $\bar{\tau}_d$ is the average size of the documents, in numbers of distinct words. $IPF_{t,b}$ and $CF_{d,b,t}$ are described in detail in [2].

All the weighting schemes presented in this paper, with the exception of LBCA, have been used before [2,35,37]. However, these weighting schemes were created with different goals. For example, the weighting scheme used in [2] was created to aid the precision in text mining tasks with recommendation as a goal, while the schemes used in [37] were all created with information retrieval as the main goal. None of the schemes used in the literature were created with a focus on numerically identifying the difference between texts produced by humans and bots.

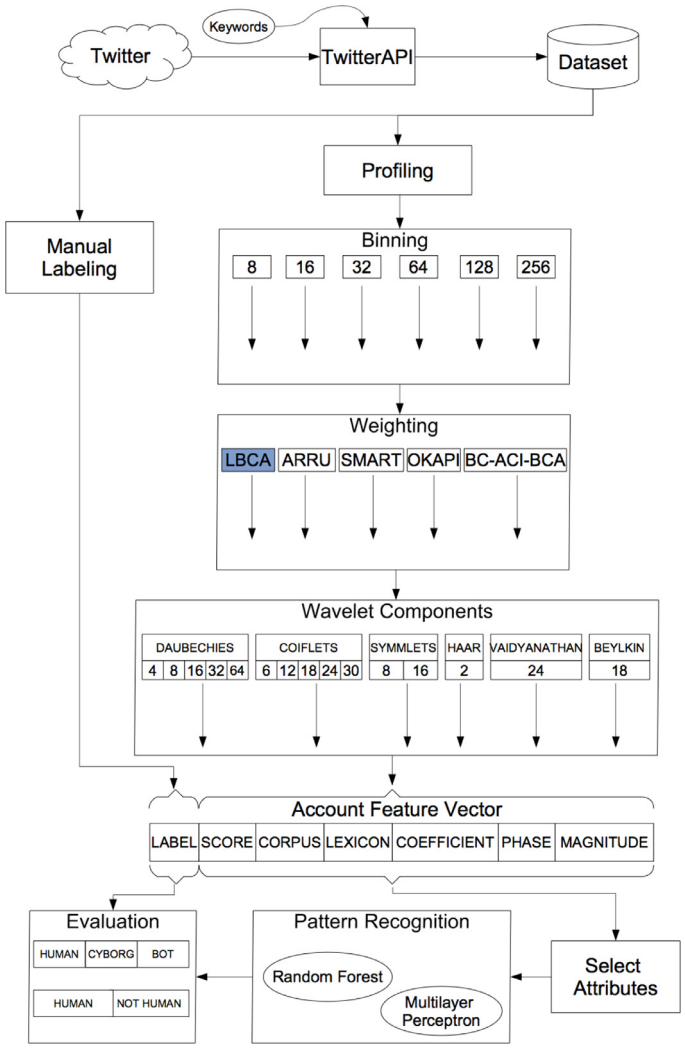


Fig. 1. The proposed approach.

Table 1
Weighting schemes for term signals.

Name	Formulation	Work described in
BD-ACI-BCA	$w_{d,t,b} = \frac{1+log(f_{d,t,b})}{(1-s)+s\cdot W_d/\bar{W}_d}$	[33]
OKAPI	$w_{d,t,b} = \frac{f_{d,t,b}}{f_{d,t,b}+\tau_d/\bar{\tau}_d}$	[33]
SMART	$w_{d,t,b} = \frac{(1+log(f_{d,t,b}))/(1+log(\bar{f}_{d,t}))}{(1-s)+s\cdot\tau_d/\bar{\tau}_d}$	[33]
ARRU	$= w_{d,t,b} = IPF_{t,b}\cdot CF_{d,b,t}$	[2]
LBCA	$w_{d,t,b} = f_{d,t,b}\cdot\tau_d/\bar{\tau}_d$	Proposed

For this reason, concerning a better description, a new weighting scheme named LBCA is proposed. It is formulated with a particular goal: to distinguish automated produced content from human produced content. Our scheme achieves a better description than attenuating the signal produced based on the size of the lexicon. We believe that human texts differ from bot texts in one special characteristic: the number of different words caused by the informal use of language, so that emphasizing the number of distinct words used in a content is an important complement to detecting bot texts. After this step, the numeric vector containing the frequencies should become a vector containing the weighted frequencies, as shown in Eq. 4.

$$\widetilde{w}_{d,t} = [w_{d,t,0}, w_{d,t,1}, \ldots, w_{d,t,b}, \ldots, w_{d,t,n-1}] \tag{4}$$

Table 2
Wavelet families and their supports, i.e., low-pass and high-pass filter lengths, used herein.

Family	Support
Haar	2
Daubechies	4, 8, 16, 32 *and* 64
Coyflets	6, 12, 18, 24 *and* 30
Symmlets	8 *and* 16
Beylkin	24
Vaidyanathan	18

2.4. Wavelets

Wavelets form an orthonormal basis to write a signal, decomposing it into different frequency components. Compared to the traditional Fourier analysis, wavelets are the preferred option to analyse non-stationary signals containing discontinuities and sharp spikes [33]. In this paper, the signals provided by the weighting process, ($\widetilde{\zeta}_{d,t}$), are decomposed using Discrete Wavelet Transforms. The complete list of Wavelet Families used in experiments, including Haar, Daubechies, Coyflets, Symmlets, Beylkin, and Vaidyanathan, is shown in Table 2. We have selected these families due to their well-known reputation, as described in [4]. 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:

$$\widetilde{\zeta}_{d,t} = \mathit{Discrete\ Wavelet\ Transform}(\widetilde{w}_{d,t}) = [\zeta_{d,t,0}, \zeta_{d,t,1}, \ldots, \zeta_{d,t,b}, \ldots, \zeta_{d,t,n-1}] \tag{5}$$

2.4.1. Magnitude components

As proposed in [33], the magnitude components are obtained by applying the absolute value function on each wavelet component: $H_{d,t,b} = |\zeta_{d,t,b}|$.

2.4.2. Phase components

Another important attribute that represents the posts are the phase components, as proposed in [33]. They are obtained by checking the signal of each wavelet component: $\phi_{d,t,b} = \frac{\zeta_{d,t,b}}{H_{d,t,b}}$.

2.4.3. Document score

The Wavelet Domain Score presented in [33,36,37] is also used in this paper due to its capacity to describe text relevance in relation to a given key terms search, T. By calculating the score from a user's tweets, we are able to know how relevant the text produced is in relation to the given words T. To calculate this, we use the Phase and the Magnitudes described in the preceding paragraphs. The detailed calculations are shown in Eqs. 6–8, where $\#(T)$ is the number of search key terms used.

$$\bar{\Phi}_{d,b} = \left| \frac{\sum_{t \in T} \phi_{d,t,b}}{\#(T)} \right| \tag{6}$$

$$s_{d,b} = \bar{\Phi}_{d,b} \cdot \sum_{t \in T} H_{d,t,b} \tag{7}$$

$$s_d = \sum_{i \in b} s^2_{d,b_i} \tag{8}$$

In the present proposed approach, the score is calculated by using the same terms that represent the document.

2.5. Pattern recognition

The first step of an efficient pattern recognition algorithm is the correlation-based feature selection, as proposed by [17], to select a subset of features. This is achieved through the analysis of a value for assessment called *merit*, that depends on correlation measures, such as MDL or *reliefs*.

In the proposed approach, Best First is performed to search for features. Dimensionality reduction is an important aspect considering the amount of information in OSNs. The features adopted in the proposed approach are the phase components, magnitudes components, wavelet components, document score, count amount of the signalized terms, corpus size, and lexicon size, as shown in Fig. 1. Both the classifiers Random Forest and Multilayer Perceptron were used in order to achieve the system output.

In particular, Random Forest was adopted for this approach due to its run-time efficiency and accuracy [13,45]. Both characteristics are essential in OSN scenarios. Also, OSNs present a dynamic environment for big data, as stated in [44]. MLPs are used to solve the problem in situations for which the classifier has to adapt itself to new instances from experience [10,31].

Table 3
Example of collected tweets.

#	Content	Class
1	Copa completa 13 dias e 7 seleções da América já estão garantidas na próxima fase http://glo.bo/1jfPlDf#G1naCopapic.twitter.com/9DdXWs05IH	Bot
2	Trânsito no Distrito Federal é alterado para o jogo decisivo do Brasil. http://bit.ly/V5rXIW#copa2014	Bot
3	Quando eu falei, a Alemanha vai ganhar a copa, os cara deram risada kkkkkkkkkkkk	Human
4	Pensei que o brasil ia fazer mais gols mais td bem kkk	Human
5	I'm at Boulevard Londrina Shopping (Londrina, PR) w/ 3 others http://4sq.com/1qXI6dw Brasil, Londrina	Cyborg
6	tive a impressão de que estava impedido... impressão ASIODHAOSIDHIOASH #Brasil	Cyborg

3. Methodology

Twitter, the social media network studied in this paper, is considered a micro blogging service. Unlike other social media networks, Twitter is known for the fact that its users can publish their thoughts in short texts, i.e., no more than 140 characters, using the Web or mobile devices [61]. These short texts, called tweets, are, by default, made available publicly, and are immediately broadcast to the user's followers [8].

The Twitter developer team offers a streaming set that gives other developers low latency access to Twitter's global stream of Tweet data. The tweet sets used as samples in this experiment were collected using such a service. The data received from Twitter API contains many fields, for example, the message identification number, the tweet's author's ID number, a short text field, i.e., the tweet, and meta data fields [8]. In this paper, the most important information consists of the 140 characters of each tweet that are used to inspect the frequency distribution of the terms. All the data received from Twitter API was stored in a MySQL database. Samples already profiled are available online[1].

The data set was collected during the FIFA World Cup 2014 to retrieve only tweets about the World Cup. We used three words as a query sequence: "BRASIL" (Brazil, in portuguese), "COPA", "COPA2014". The data set is written entirely in Brazilian Portuguese and, considering that the event is known all over the world, it was possible to find all three occurrences of classes, i.e., Human, Bot and Cyborg. It is possible to notice the difference between the posts from bots and humans and also the mixed behaviour of cyborgs as shown in Table 3.

Humans, cyborgs, and bots were possible labels for the data set we obtained. Considering the very few number available that address cyborgs and bots, there is no other option than to manually create a data set. Therefore, manual labeling was carried out according to [11,22,53,55], which also treated such automatic behaviour on OSNs.

Another issue considering the collected data set is the amount of text collected per user. Once every user writes different amounts of words per post/tweet, it was not possible to normalize the user content by the number of posts. Instead, we performed the experiment with different corpus sizes per user. In these experiments, the smallest corpus size was around 500 words and the biggest ones around around 2500 words.

Then, a data set composed of 100 users, of which 36 were humans, 36 cyborgs, and 28 bots, was mounted. Although the amount of users might look small in comparison to the milions of active Twitter accounts, some of the literature has used about the same number of users for classification goals, as seen in [12]. In this paper, this was also possible due to the fact that the range of variation of written texts about a single subject is not that big.

Next, each user in the dataset is represented by a document containing all of their samples cumulatively concatenated as described in Section 2.1 and shown in Table 4. Then, the data transformation proceeded from binning up to the feature vector as previously described.

To perform the classification, we explored many different experimental settings. In particular, we performed two experiments using the same data set. The main purpose of the first experiment was the classification into three groups of accounts: humans, cyborgs and bots. The second experiment involved the classification into only two groups of accounts: humans and non-humans. Thus, in the second experiment, cyborgs and bots were considered non-humans.

The training data set was formed by many combinations. For instance, from binning, we combined the following possibilities: 8, 16, 32, 64, 128, 256. The intention was to discover the best size for binning in order to solve the problem since tweets are different from Web Pages (16 bins) or Structured Documents (8 bins). Then, regarding weightings, 5 schemes found in the literature were presented in Section 2.3 and used in this research. Each of them has been introduced in the previous literature on text mining, except for one, which is now proposed. We also explored many wavelet families, such as Haar, Daubechies, Coyflets, Symlets, Beylkin, and Vaidyanathan, in order to discover an optimal descriptor of account writting patterns. Along with the wavelet coefficients, the magnitudes, and the phases, we also adopted the corpus size, the lexicon size, and a document score from [33], sending as a query the same words passed to the Twitter API.

Before classifying, a Correlation-based Feature Subset Selection proposed by [17] was applied in order to reduce the dimensionality. MLP architectures and Random Forest were used to check the best class identification. The training step was carried out for each one of these possible settings.

[1] http://www.barbon.com.br/wp-content/uploads/2015/07/SourceFilesIS.zip

Table 4
Profiling example.

Author - Post	Content
#1 - #1	Parabéns pelo jogo, pelo gol, por representar o Brasil, por jogar com raça e amor á camisa, por orgulhar o povo brasileiro @neymarjr #bom_Diiiia #DeusnocomandoSempre #viaçãoUtil #copa2014 #job #utilnaCopa #Util #Brasil
#1 - #2	Só por que a copa começa no dia 12, o dia dos namorados tem que ser outro dia, wtf?
#1 - #3	@neymarjr Todo mundo tá confiando em ti e na seleção inteira! O Brasil tá com vocês!
Document	Content
#1	Parabéns pelo jogo, pelo gol, por representar o Brasil, por jogar com raça e amor á camisa, por orgulhar o povo brasileiro @neymarjr #bom_Diiiia #DeusnocomandoSempre #viaçãoUtil #copa2014 #job #utilnaCopa #Util #Brasil Só por que a copa começa no dia 12, o dia dos namorados tem que ser outro dia, wtf? @neymarjr Todo mundo tá confiando em ti e na seleção inteira! O Brasil tá com vocês!

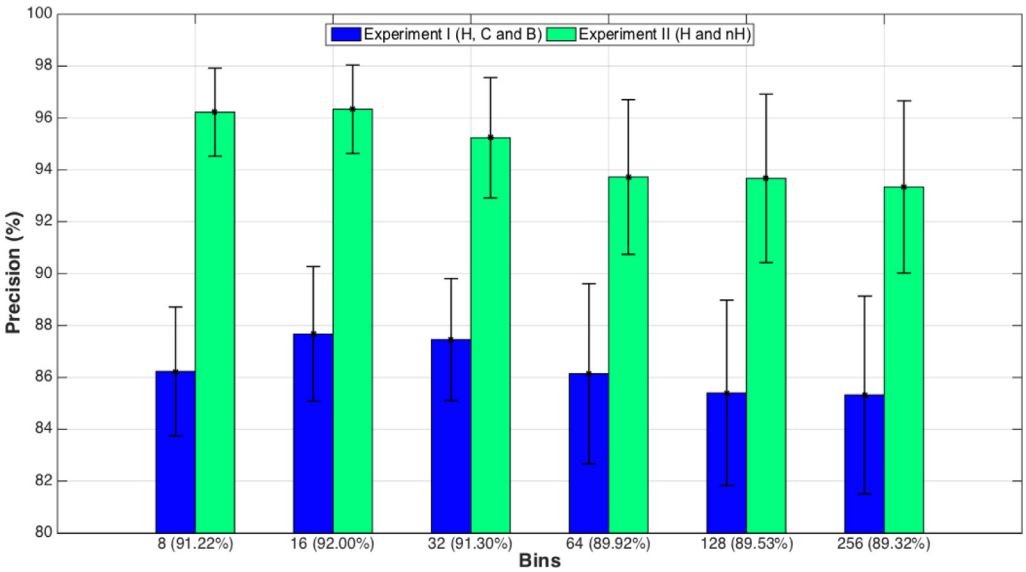


Fig. 2. Bins and mean accuracy of experiments 1 and 2.

Concerning the evaluation, we adopted 10-fold cross-validation as described in [57]. By doing so, we ensure that each instance in the data set is used at least once in either training or testing and also each setting is trained and tested 10 times using different parts of the data set every time. The setting accuracy is computed by averaging the 10 rounds of training and testing.

The final results were computed from a confusion matrix to evaluate the classifiers. Therefore, an example of an experimental setting would be a classification instance using: binning in 8 parts, LBCA weighting scheme, Haar Wavelet Transform, and a Random Forest as a classifier. In this research, we tested many possible combinations of binning, weighting schemes, wavelet families, and classifiers. The results are discussed in Section 4.

4. Results and discussion

As to binning, some paper, such as [33,36], have used 8 bins. In contrast, [40] proposed different numbers of frames to find the best quantity to represent a document based on web pages. The results of experiments on varying the binning led us to realize that 16 and 32 partitions have a noticeable advantage over 8. Fig. 2 shows the mean accuracy of each partition number along with a mean for both experiments, independent of the wavelet family, weighting scheme, and classification approach. Binning based on 8, 16 and 32 partitions has a better mean accuracy, with 91.22, 92.00 and 91.30%, respectively. Also these three partitions has small standard deviations, which is another significant result addressing stability, independent of the other parameters.

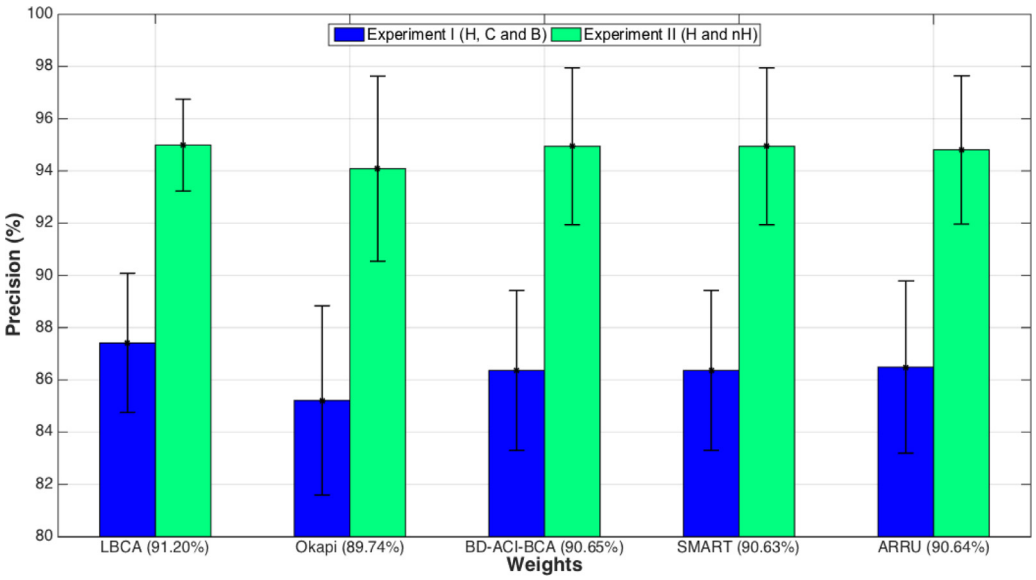


Fig. 3. Weighting schemes and mean accuracy for experiments 1 and 2.

Table 5
Classificator results for Experiment 1.

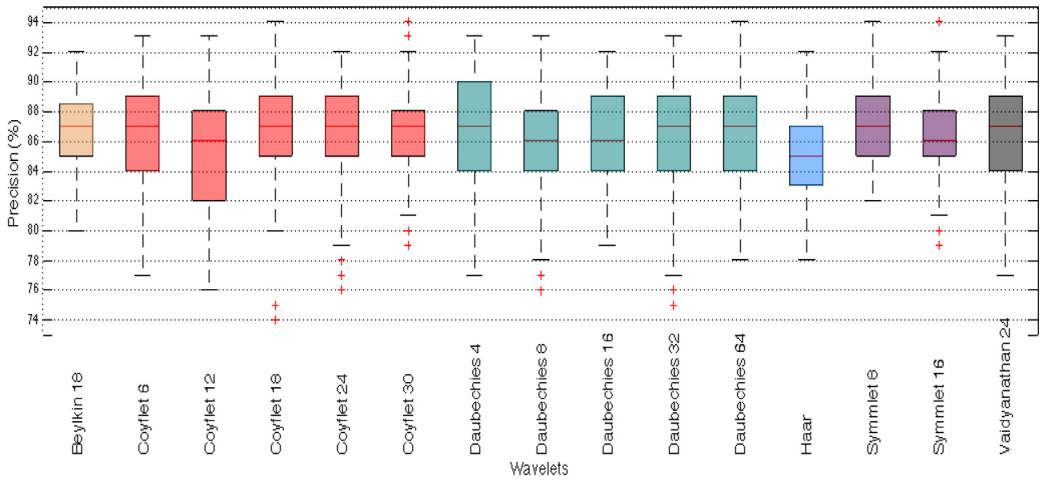
	Approach	Mean (%)	Std. Dev. (%)	Median (%)	Max. (%)	Min. (%)
Top	Random Forests	88.7	2.13	89.0	94.0	82.0
	MLP[9-5-1]	88.1	2.34	88.0	92.0	83.0
	MLP[13-20-1]	87.4	2.40	87.0	91.0	83.0
	MLP[9-9-1]	87.1	2.06	87.0	92.0	83.0
	MLP[13-7-1]	87.0	2.01	87.0	90.0	83.0
Bottom	MLP[46-46-1]	82.1	2.63	83.0	87.0	74.0
	MLP[57-57-1]	82.1	2.43	83.0	87.0	74.0
	MLP[57-81-1]	82.1	2.43	83.0	87.0	74.0
	MLP[34-34-1]	82.1	1.94	82.0	85.0	78.0
	MLP[41-62-1]	82.1	2.35	83.0	85.0	74.0

For 64, 128, and 256 partitions, the mean accuracy presented a descending gradient. This is justifiable considering that a great number of partitions implies a large number of descriptors, which could disrupt the description of the contents.

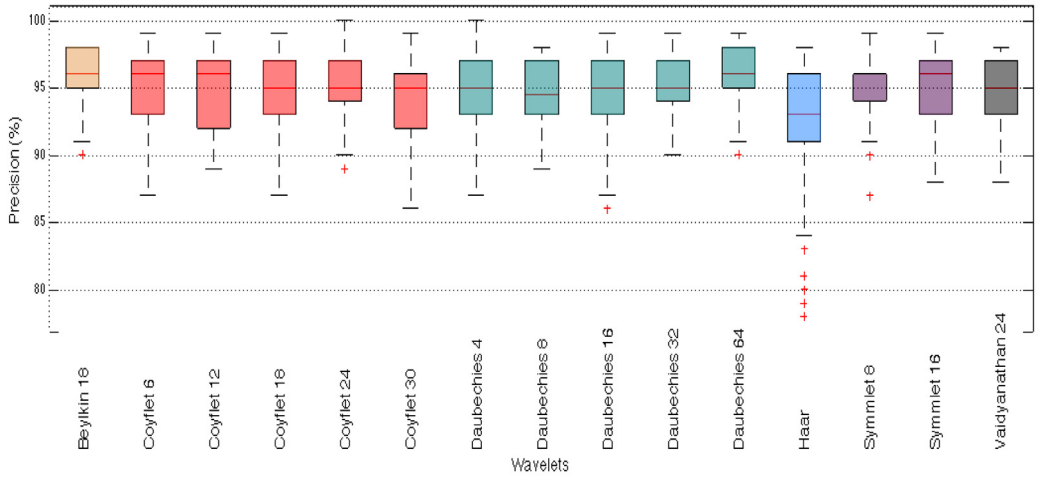
The results as to the weighting schemes are presented in Fig. 3, where it is possible to observe that the proposed LBCA achieved a higher accuracy, with a mean accuracy of 91.20%, and among the smallest standard deviations out of all the weighting schemes for both experiments I and II. The BD-ACI-BCA, Smart, and ARRU schemes also achieved good results, with 90.65, 90.63 and 90.64%, respectively. However, for experiment 1, LBCA surpasses all other schemes in mean accuracy.

Fig. 4 , uses boxplots to show the results concerning another choice: the wavelets. In Text Mining approaches, the related literature has used only Haar wavelets and Daubechies wavelets [37,39]. In this paper, we evaluated the application of several wavelet families. First, Coiflets and Daubechies are more accurate. In a specific configuration, they were able to achieve 94% accuracy in experiment 1 (Fig. 4a) and 100% in experiment 2 (Fig. 4b). However, as the boxplot is characterized by an observation of wavelets that achieved symmetric size of quartiles, Daubechies 4 shows a good result in relation to quarter size for experiment 1 and 2, along a high median. Regarding the Coiflet family, the results lead us to conclude that supports close to 6 and 12 achieve good results while bigger supports lead to the occurrence of outliers as seen in Fig. 4a. Symmlets presented good maximum results in both experiments. However, both Symmlets used produced outliers: Symmlet 16 for experiment 1 and Symmlet 8 for experiment 2. Vaidyanathan, Haar, and Beylkin provided symmetric results in experiment 1. The same results were not achieved in the second experiment. In sum, results show that, in a experiment between humans and non-humans, Haar, known for its simplicity [4], might not grant the appropriate description to each problem. Vaidyanathan and Beylkin are specific for audio [4], and so, the results for textual data are justifiable. This is of great interest due to the fact that many previous papers used Daubechies and Haar wavelets. Results show that Coiflets 6, Coiflets 12, and Daubechies 4 achieved a good median and almost had a symmetric result considering that these three did not produce outliers.

In relation to the classifiers, we present results in Table 5 for experiment 1 and in Table 6 for experiment 2. Both tables contain the results of the five top and worst classifications, in terms of mean accuracy.



(a) Experiment Human, Cyborg and Bot



(b) Experiment Human and Non-Human

Fig. 4. Wavelets and accuracy.

Accuracies higher than 87% were achieved when distinguishing between human, cyborg, and bot accounts. Random Forest is one of them, achieving 88.7% accuracy along with median at 89.0% and maximum at 94.0%. Along with 88% mean accuracy, we have MLP, producing good results. We consider it a suitable result for OSN situations because Random Forests need only the raising cost and MLPs, after being trained, can also readjust their weights. In our case, the best MLPs, in terms of accuracy, are simple architectures, which keep the computational cost of the proposed approach suitable for the OSN scenario.

For Table 6, in experiment 2, Random Forest was not present in the top results (mean: 96.6%, standard deviation: 1.19%, median: 97%, maximum: 100%, minimum: 93%) but reached the maximum result with a specific set (binning: 32, weighting scheme: LBCA, wavelet: Daubechies 4). Simple architectures, as in experiment 1, obtained better results and stability, as the standard deviations in Table 6 show. Results are also suitable in experiment 2 for OSN scenarios.

Another important observation is in experiment 2, which separates the instances into humans and non-humans. This would be useful for Sentiment Analysis, which would benefit from retrieving only human accounts. In some particular cases, such as the Random Forest that was not present in the top 7, it had 96% mean accuracy and 100% of maximum accuracy in a classification task of humans and non-humans. It also had the best results from: binning, weighting scheme, wavelet along support, and classifiers. The classification in experiment also obtained a good result in terms of false positive, as can be seen in Table 7. The separability between humans an non-humans in experiment 2 achieved results more accurate than in experiment 1.

Analysing every result at once, we propose some settings composed of the most appropriate combinations in Table 7. These combinations take into consideration not only a single setting that achieved a very high result just once, but also require each part of its setting to have obtained a good result in previous discussions. Thus, in Table 7, the weighting schemes used are only ARRU and LBCA (for which both the standard deviaton and mean accuracy were good), the wavelets used are mostly Daubechies

Table 6
Classificator results for Experiment 2.

	Approach	Mean (%)	Std. Dev. (%)	Median (%)	Max. (%)	Min. (%)
Top	MLP[7-11-1]	97.0	1.31	97.5	98.0	94.0
	MLP[8-4-1]	97.0	1.69	97.0	99.0	93.0
	MLP[7-7-1]	96.9	1.32	97.5	98.0	94.0
	MLP[8-16-1]	96.9	1.81	97.0	99.0	93.0
	MLP[7-4-1]	96.9	1.29	97.0	98.0	94.0
[a]	Random Forest	96.6	1.19	97.0	100.0	93.0
Bottom	MLP[35-35-1]	90.8	2.41	91.0	95.0	87.0
	MLP[35-35-1]	90.8	2.41	91.0	95.0	87.0
	MLP[35-70-1]	90.7	2.31	91.0	95.0	87.0
	MLP[35-53-1]	90.7	2.37	91.0	95.0	87.0
	MLP[38-38-1]	90.7	2.94	90.0	96.0	86.0

[a] Exceptional case for Random Forest.

Table 7
Best settings.

	Bins	Weight	Wavelet	Class.	Prec. (%)	FPR (%)
(Exp. I)	16	LBCA	Coiflet 6	Random F.	94	6.3
	8	LBCA	Daubechies 4	Random F.	93	7.4
	16	ARRU	Daubechies 16	MLP[9-9-1]	92	8.6
	16	ARRU	Daubechies 8	MLP[9-5-1]	92	8.6
(Exp. II)	32	LBCA	Daubechies 2	Random F.	100	0
	32	ARRU	Daubechies 32	MLP[8-8-1]	99	1.2
	8	ARRU	Coiflet 6	MLP[8-16-1]	99	1.2
	16	LBCA	Beylkin 18	MLP[7-7-1]	97	2.5

Fig. 5. Proposed setting for both experiments.

and Coiflets, the classifiers were mostly present in Tables 5 and 6 as top results. Therefore, we ensure that the settings proposed in Table 7 are not only made up of parts that achieved higher results as outliers, but parts that always performed well in the experiments.

In case both experimental scenarios are needed, we recommend: [binning: 32, weighting: LBCA, wavelet: Daubechies 4, classifier: Random Forest]. In experiment 1, this setting achieved 91 and 100% in experiment 2, averaging 95.5% accuracy. We recommend this setting because each of its parts appears in Table 7, on which is based our recommendation. Fig. 5 illustrates our proposal.

The last discussion is about the computational complexity: we are concerned about the cost of each method used in this paper. The binning step has a constant cost using the document as a vector. All the weighting schemes adopted, including LBCA, depend only on counting and, therefore, have a linear cost. The WTs are linear, following [33]. Regarding the classifiers, Random Forest costs $O(T*K*N*log(N))$ where T is the number of trees, K is the number of features, and N is the number of samples [1]. MLP with Back-propagation, as used in this paper, works with a training time of $O(W^3)$, where W is the number of weights, and works with $O(W)$ after trained [43].

5. Conclusion

As presented in Section 1.1, there has been a developing concern about frauds in Online Social Networks (OSNs) in the literature. However, this paper is the first pure text mining approach and would be of interest to the community as it needs only text and, thus, is applicable to any OSN.

As a proposal for a pure text mining solution for finding bot accounts in today's OSNs, we developed a model that was tested with many settings, as shown in Section 3. The features addressed for this research were presented in Section 2. Along with the proposed approach, this paper presented a new weighting scheme—an essential step in many studies—which was fully supported by results. Lastly, this same weighting scheme also contributes to the distinction between humans, cyborgs, and bots.

The main concern about the accuracy of the classification task was the computational complexity of each step of the process. Therefore, our study has proven that good options of binning divisions are 8, 16 or 32, corresponding to interesting results in terms of computational costs. Regarding weighting schemes, LBCA surpassed the techniques presented in the literature for

experiments 1 and 2. In relation to the classifiers, the fact that smaller architectures obtained better results also reinforces our expectations of keeping the computational cost of the proposed approach suitable for OSNs.

Therefore, the proposed approach obtained suitable results for OSNs. In terms of accuracy, we achieved a mean accuracy of 94% for classifying accounts, distinguishing between Humans, Cyborgs, and Bots. Another remarkable result was in the second experiment, which classified the accounts into Humans and Non-Humans, the mean accuracy for this task being 100.0%. With these experiments, we expect to contribute to the reduction of errors in OSNs and diminish the number of victims of frauds by detecting whose are humans or not.

6. Future research

The main concern of this paper has been the OSNs and their purpose in society in terms of marketing and security. In this respect, bots that are fraudulent need to be found in order to preserve the OSN. Thus, it would be interesting to analyse the behaviour of the content produced by bots for fraud evaluation, improving the detection of spamming activities in OSNs.

7. Acknowledgements

We are grateful to the CNPq and FAP for making this paper possible by sponsoring our research, specifically CNPQ processes 479821/2013-5 and 306811/2014-6, and FAP process 064/2013 - 33785.

References

[1] S.Z. Alborzi, D. Maduranga, R. Fan, J.C. Rajapakse, J. Zheng, CUDAGRN: Parallel speedup of inferring large gene regulatory networks from expression data using random forest, in: Pattern Recognition in Bioinformatics, Springer, 2014, pp. 85–97.
[2] G. Arru, D. Feltoni Gurini, F. Gasparetti, A. Micarelli, G. Sansonetti, 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, 2013, pp. 941–944.
[3] S.-A. Bahrainian, A. Dengel, Sentiment analysis and summarization of Twitter data, in: Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on, IEEE, 2013, pp. 227–234.
[4] S. Barbon Junior, R.C. Guido, S.-H. Chen, L.S. Vieira, F.L. Sanchez, Improved dynamic time warping based on the discrete wavelet transform, in: Multimedia Workshops, 2007. ISMW'07. Ninth IEEE International Symposium on, IEEE, 2007, pp. 256–263.
[5] S.Y. Bhat, M. Abulaish, 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, 2013, pp. 100–107.
[6] H. Bicen, Student opinions regarding twitter usage with mobile applications for educational purposes, Procedia-Social and Behav. Sci. 136 (2014) 385–390.
[7] H. Bicen, N. Cavus, Twitter usage habits of undergraduate students, Procedia—Social and Behav. Sci. 46 (2012) 335–339.
[8] C.A. Bliss, I.M. Kloumann, K.D. Harris, C.M. Danforth, P.S. Dodds, Twitter reciprocal reply networks exhibit assortativity with respect to happiness, J. Comput. Sci. 3 (5) (2012) 388–397.
[9] Y. Boshmaf, I. Muslukhov, K. Beznosov, M. Ripeanu, The socialbot network: When bots socialize for fame and money, in: Proceedings of the 27th Annual Computer Security Applications Conference, ACM, 2011, pp. 93–102.
[10] G. Calcagno, A. Staiano, G. Fortunato, V. Brescia-Morra, E. Salvatore, R. Liguori, S. Capone, A. Filla, G. Longo, L. Sacchetti, A multilayer perceptron neural network-based approach for the identification of responsiveness to interferon therapy in multiple sclerosis patients, Inf. Sci. 180 (21) (2010) 4153–4163.
[11] Z. Chu, S. Gianvecchio, H. Wang, S. Jajodia, Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Trans. Dependable and Secure Comput. 9 (6) (2012) 811–824.
[12] M.Ö. Cingiz, B. Diri, G. Biricik, Am i typing fresh tweets: detecting up-to-dateness and worth of categorical information in microblogs, Expert Syst. Appl. 42 (12) (2015) 5256–5263.
[13] S. del Río, V. López, J.M. Benítez, F. Herrera, On the use of mapreduce for imbalanced big data using random forest, Inf. Sci. 285 (2014) 112–137. Processing and Mining Complex Data Streams. http://dx.doi.org/10.1016/j.ins.2014.03.043
[14] S. Fong, Y. Zhuang, J. He, Not every friend on a social network can be trusted: Classifying imposters using decision trees, in: Future Generation Communication Technology (FGCT), 2012 International Conference on, 2012, pp. 58–63, doi:10.1109/FGCT.2012.6476584.
[15] D.M. Freeman, Using naive Bayes to detect spammy names in social networks, in: Proceedings of the 2013 ACM workshop on Artificial intelligence and security, ACM, 2013, pp. 3–12.
[16] P. Galeas, R. Kretschmer, B. Freisleben, Document relevance evaluation via term distribution analysis using Fourier series expansion, CoRR (2009).abs/0903.0153
[17] M.A. Hall, Correlation-Based Feature Selection for Machine Learning, The University of Waikato, 1999 Ph.D. thesis.
[18] X. Han, X. Chang, An intelligent noise reduction method for chaotic signals based on genetic algorithms and lifting wavelet transforms, Inf. Sci. 218 (2013) 103–118.
[19] A. Hassan, A. Abbasi, D. Zeng, Twitter sentiment analysis: a bootstrap ensemble framework, in: International Conference on Social Computing (SocialCom), 2013, IEEE, 2013, pp. 357–364.
[20] L.-C. Hsieh, C.-W. Lee, T.-H. Chiu, W. Hsu, Live semantic sport highlight detection based on analyzing tweets of twitter, in: IEEE International Conference on Multimedia and Expo (ICME), 2012, IEEE, 2012, pp. 949–954.
[21] P. Hui, K. Xu, V.O. Li, J. Crowcroft, V. Latora, P. Lio, Selfishness, altruism and message spreading in mobile social networks, in: INFOCOM Workshops 2009, IEEE, 2009, pp. 1–6.
[22] T. Hwang, I. Pearce, M. Nanis, Socialbots: voices from the fronts, Interactions 19 (2) (2012) 38–45.
[23] S. Keretna, A. Hossny, D. Creighton, Recognising user identity in twitter social networks via text mining, in: IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2013, IEEE, 2013, pp. 3079–3082.
[24] Y. Li, Y. Jiang, D. Jin, L. Su, L. Zeng, D.O. Wu, Energy-efficient optimal opportunistic forwarding for delay-tolerant networks, IEEE Trans. Veh. Technol. 59 (9) (2010) 4500–4512.
[25] Y. Li, G. Su, D.O. Wu, D. Jin, L. Su, L. Zeng, The impact of node selfishness on multicasting in delay tolerant networks, IEEE Trans. Veh. Technol. 60 (5) (2011) 2224–2238.
[26] T.-C. Lin, C.-M. Lin, Wavelet-based copyright-protection scheme for digital images based on local features, Inf. Sci. 179 (19) (2009) 3349–3358. http://dx.doi.org/10.1016/j.ins.2009.05.022.
[27] J. Lu, Y. Zou, Z. Ye, Enhanced fractal-wavelet image denoising, in: ISECS International Colloquium on Computing, Communication, Control, and Management, 2008. CCCM'08., vol. 1, IEEE, 2008, pp. 115–119.
[28] T. Meng, A.T. Soliman, M.-L. Shyu, Y. Yang, S.-C. Chen, S. Iyengar, J. Yordy, P. Iyengar, Wavelet analysis in current cancer genome research: aA survey, IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) 10 (6) (2013) 1442–14359.

[29] N.E. Miller, P.C. Wong, M. Brewster, H. Foote, TOPIC ISLANDS tm—a wavelet-based text visualization system, in: Proceedings on Visualization'98., IEEE, 1998, pp. 189–196.
[30] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, A.H. Wang, Twitter spammer detection using data stream clustering, Inf. Sci. 260 (2014) 64–73.
[31] S. Mirjalili, S.M. Mirjalili, A. Lewis, Let a biogeography-based optimizer train your multi-layer perceptron, Inf. Sci. 269 (2014) 188–209.
[32] M.M. Mostafa, More than words: social networks' text mining for consumer brand sentiments, Expert Syst. Appl. 40 (10) (2013) 4241–4251.
[33] L.A. Park, Spectral Based Information Retrieval, The University of Melbourne, 2003 Ph.D. thesis.
[34] L.A. Park, M. Palaniswami, K. Ramamohanarao, A new implementation technique for fast spectral based document retrieval systems, in: Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on, IEEE, 2002a, pp. 346–353.
[35] L.A. Park, M. Palaniswami, K. Ramamohanarao, A novel web text mining method using the discrete cosine transform, in: Principles of Data Mining and Knowledge Discovery, Springer, 2002b, pp. 385–397.
[36] L.A. Park, K. Ramamohanarao, M. Palaniswami, Fourier domain scoring: a novel document ranking method, IEEE Trans. Knowl. Data Eng. 16 (5) (2004) 529–539.
[37] L.A. Park, K. Ramamohanarao, M. Palaniswami, A novel document retrieval method using the discrete wavelet transform, ACM Trans. Inf. Syst. (TOIS) 23 (3) (2005) 267–298.
[38] N. Potha, E. Stamatatos, A profile-based method for authorship verification, in: Artificial Intelligence: Methods and Applications, Springer, 2014, pp. 313–326.
[39] D. Purwitasari, 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.
[40] D. Purwitasari, Y. Okazaki, K. Watanabe, A study on web resources navigation for e-learning: usage of Fourier domain scoring on web pages ranking method, in: Second International Conference on Innovative Computing, Information and Control, 2007. ICICIC'07., IEEE, 2007.458–458
[41] D. Purwitasari, N. Suciati, R. Soelaiman, D. Farida, Fourier domain scoring for ranking method in small data set with preprocessing using oracle text, in: In ICTS'07: The 3rd International Conference on Information & Communication Technology and Systems, 2007.
[42] M. Ramaraj, S. Raghavan, A survey of wavelet techniques and multiresolution analysis for cancer diagnosis, in: International Conference on Computer, Communication and Electrical Technology (ICCCET), 2011, IEEE, 2011, pp. 109–114.
[43] R.D. Reed, R.J. Marks, Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, MIT Press, 1998.
[44] S.E. Seker, K. Al-Naami, L. Khan, Author attribution on streaming data, in: 14th International Conference on Information Reuse and Integration (IRI), 2013, IEEE, 2013, pp. 497–503.
[45] K. Singh, S.C. Guntuku, A. Thakur, C. Hota, Big data analytics framework for peer-to-peer botnet detection using random forests, Inf. Sci. 278 (2014) 488–497.
[46] S. Sivanesh, K. Kavin, A. Hassan, Frustrate Twitter from automation: how far a user can be trusted? in: International Conference on Human Computer Interactions (ICHCI), 2013, 2013, pp. 1–5, doi:10.1109/ICHCI-IEEE.2013.6887787.
[47] J. Smailović, M. Grčar, N. Lavrač, M. Žnidaršič, Stream-based active learning for sentiment analysis in the financial domain, Inf. Sci. 285 (2014) 181–203. Processing and Mining Complex Data Streams. http://dx.doi.org/10.1016/j.ins.2014.04.034
[48] V. Sridharan, V. Shankar, M. Gupta, Twitter games: How successful spammers pick targets, in: Proceedings of the 28th Annual Computer Security Applications Conference, ACM, 2012, pp. 389–398.
[49] S. Tedmori, N. Al-Najdawi, Image cryptographic algorithm based on the Haar wavelet transform, Inf. Sci. 269 (2014) 21–34.
[50] S. Thaicharoen, T. Altman, K.J. Cios, 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, Inc., 2008, pp. 209–217.
[51] A.K. Tyagi, G. Aghila, Detection of fast flux network based social bot using analysis based techniques, in: International Conference on Data Science & Engineering (ICDSE), 2012, IEEE, 2012, pp. 23–26.
[52] G.A.N. Vásquez, E.M. Escamilla, Best practice in the use of social networks marketing strategy as in SMEs, Procedia—Social and Behav. Sci. 148 (2014) 533–542.
[53] R. Wald, T. Khoshgoftaar, A. Napolitano, Filter-and wrapper-based feature selection for predicting user interaction with Twitter bots, in: 14th International Conference on Information Reuse and Integration (IRI), 2013, IEEE, 2013a, pp. 416–423.
[54] R. Wald, T.M. Khoshgoftaar, A. Napolitano, C. Sumner, Predicting susceptibility to social bots on Twitter, in: 14th International Conference on Information Reuse and Integration (IRI), 2013, IEEE, 2013, pp. 6–13.
[55] R. Wald, T.M. Khoshgoftaar, A. Napolitano, C. Sumner, Which users reply to and interact with Twitter social bots? in: 25th International Conference on Tools with Artificial Intelligence (ICTAI), 2013, IEEE, 2013b, pp. 135–144.
[56] J.S. Walker, A Primer on Wavelets and Their Scientific Applications, CRC press, 1999.
[57] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005.
[58] P.C. Wong, H. Foote, D. Adams, W. Cowley, J. Thomas, Dynamic visualization of transient data streams, in: IEEE Symposium on Information Visualization, 2003. INFOVIS 2003., IEEE, 2003, pp. 97–104.
[59] G. Xexéo, J. de Souza, P.F. Castro, W.A. Pinheiro, Using wavelets to classify documents, in: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08., vol. 1, IEEE, 2008, pp. 272–278.
[60] S.J. Yu, The dynamic competitive recommendation algorithm in social network services, Inf. Sci. 187 (2012) 1–14.
[61] M. Zappavigna, Ambient affiliation: a linguistic perspective on Twitter, New Media & Society 13 (5) (2011) 788–806.
[62] Y. Zhao, J. Zobel, Searching with style: authorship attribution in classic literature, in: Proceedings of the Thirtieth Australasian Conference on Computer Science-Volume 62, Australian Computer Society, Inc., 2007, pp. 59–68.
[63] X. Zhou, S. Wu, C. Chen, G. Chen, S. Ying, Real-time recommendation for microblogs, Inf. Sci. 279 (2014) 301–325. http://dx.doi.org/10.1016/j.ins.2014.03.121.