Received: March 1, 2020. Revised: May 13, 2020.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Framework for Spam Detection in Twitter Based on Recommendation System

Fatna Elmendili[1]* **Younès El Bouzekri El Idrissi[1]**

[1]*Systems Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco*
* Corresponding author's Email: f.elmendili@gmail.com

Abstract: The rapidly growing online social networking sites have been infiltrated by a large amount of spam. Spammers are a particular kind of ill-intentioned users who degrade the quality of OSNs information through misusing all possible services provided by OSNs. Social spammers spread many intensive posts/tweets to lure legitimate users to malicious or commercial sites containing malware downloads, phishing, and drug sales. Given the fact that Twitter is not immune towards the social spam problem, different researchers have designed various detection methods, which inspect individual tweets or accounts for the existence of spam contents. Today, social networks are exposed to various threats that exploit their vulnerability. However, although of the high detection rates of the account-based spam detection methods, these methods are not suitable for filtering tweets in the real-time detection because of the need for information from Twitter's servers. At tweet spam detection level, many light features have been proposed for real-time filtering; however, the existing classification models separately classify a tweet without considering the state of previous handled tweets associated with a topic. First, they propose the identification of spam tweet by the security approach based on social honeypots and then they propose a method based on an algorithm "content filtering" in order to detect those that are similar to spam tweet detected by the approach of honeypots. Our approach has greatly improved the quality of abstraction in terms of performance and design. The algorithm is also fast and simple to implement. Experimental results show the stability and accuracy (over 99%), F-measure 98% of our approach.

Keywords: Tweets spam, Content filtering, Social honeypot, Recommender system.

1. Introduction

In the recent years, online social networks have become increasingly prevalent platforms for users to post their messages and share ideas around the world [1].

Twitter is a great platform of communication and sharing, it attracts the profiles while providing services to disseminate messages of 140 characters. [1].

Every month, over 42 million new accounts are created in Twitter; the openness of Twitter also leads to the popularity of spamming activities on Twitter [1].

Unfortunately, the attackers have made their attention on the OSN and exploit them in carrying out various type of attacks like the phishing [2-4], the injection of malicious codes, the dissemination of malicious software [2, 3, 5]. These malicious behaviors can cause serious privacy and economic problems. User's private data are popular on the black market and access to them may lead to economic crimes.

Detecting reconnaissance activities is very difficult since usually it is performed outside of the organization's premises and without direct interaction with the organizational resources. At some point, the reconnaissance phase enables the attacker to find an entry point into the organization leading to the next phases [6].

Based on the statistics of APT 2013 92% of the researchers believe that the use of social networks increases the likelihood of a successful attack [7]. Social media is already ripe with threats: between

8%–10% of all social media, profiles are malicious in nature [8].

The popularity of social networks allows them to come out of places for the execution of malicious activities. Due to the enormous popularity of social networks, it is easier for the cybercriminal of abusing them. These can be in the form of media, thread or spam tweet that do not belong to a user. These tweets after the click will lead the user to other pages created by a malicious user [9].

Social networking sites provide limited mechanisms to stop the exposure of data profiles to the applications. For example, in the case of Facebook, when users visit an application for the first time, they must allow this application to access all the data in the profile required. This unique choice is not to use or visit the request. However, even this does not guarantee any genuine security [9].

Main contributions:

(1) we show that the honeypot detects the malicious activities of this profile, (e.g., by crawling the profile of the user sending the unsolicited friend request plus hyperlinks from the profile to pages on the Web-at-large).
(2) Social honeypots display certain sensitive keywords in their tweets, which are attractive to attackers.
(3) We apply different machine learning techniques
(4) and compare their performance on large datasets. Also, we measure and compare the performance as well as the number of extracted features (top features).
(5) we Propose a layer based on the content filtering to calculate the similarity between the spam tweet detected by the Layer based on honeypot's and the layer of content filtering in the aim to present the spam tweets that are similar to spam tweets detected by social honeypots.

This paper is organized as follows: In section 2 related work in setection spam tweets is presented. The proposed framework is presented in section 3. Next, in section 4, we describe how the model is evaluated and present the results of our experiments. Section 5, we evaluate the proposed model in comparison with two other models. Finally, section 6 includes the conclusions and future work.

2. Related work

Many efforts have been made to develop spam detection techniques on Twitter in the last decade. In this section, the authors explain the state of the art for detection malicious users in social network [1]. Jasek et al. [12] suggested the general concept of using honeypots (not social net. In our research, they suggest a solution that speciworks honeypots) to detect activities associated with APTsfically targets social networks and takes into consideration their logistical concerns. Several previous studies [13,14] have focused on the identification of spammers that use social honeypots and the creation of classifiers in order to distinguish social malicious users from legitimate profiles. Spammers generally write tweets that contain a hashtag and URL according to the following research studies that analyzed commonly used hashtags and URL: [15-18]. COMPA [19] detected compromised accounts that wrote spam tweets based on the tweeting language of the user's account, the tweeting time window, the URL, and the mention" receiver. This approach for late profiles learns the previous behavioral pattern of each user. Benevenuto et al. [16] and Martinez-Romo et al. [17] the proposal of a classification of models which Apis the number of hashtags and URL [16] or well of the URL of spam that are used in a field of spam tweets.

Yardi and Al. [18] studied spammers' strategic behavioral patterns and concluded that the use of hashtags related to trending topics is a very effective spamming strategy.

Gao et al. [20] built a template based on the sentence structure of spam ground truth tweets and used template matching to filter out spam tweets. Existing techniques in spammer detection typically use a pre-classified data set and a combination of behavioral (content, user information, network and topic) to create a classifier that can accurately differentiate spammers from legitimate users with accuracies obtained of around 90%. The main difference in the majority of these approaches is in the features used for classification [16, 21-23]. Chakraborty et al. [24] proposed a slightly different systems to detect users posting abusive content such as harmful URLs, porn URLs, and phishing links as part of the friend request process. The solution was tested on 5 000 accounts with the SVM classifier performing the best, achieving an accuracy of 89%. Miller et. al. [25] attempt to treat the identification of spammers as an anomaly detection and not classification problem where outliers are flagged as spammers. They use a combination of characteristic of a user and a text feature. They then test two algorithms: DBSCAN, which uses a density, based on similarity metric and K-Means, which uses a Euclidean distance based on metric. These approaches achieved an 82% and 71% F1 score respectively with high accuracy but low precision [11]. Considering user profiles, Lee et al. [26] a

proposed study which distinguishes the profiles spammer and the legitimate profiles, such an approach presents the number of relations, the age etc. as account features. Another example is Martinez et al. who proposed a new and comprehensive system, this proposal focuses essentially on the tweets containing malicious links, then a step of automatic learning (ML) has been used to extract the links in each message posted by the users of the system. Still considering the detection of spammers, Benevenuto et al. [28] have proposed a method based on a ML, which is based on many characteristics textual like: The textual and non-textual features who represent hashtags, the number of words and the number of Links in each post have been used.

The classification algorithm used was the support vector machines, Santos et al [29] have proposed an approach of exploration of text that is based on a method of level of the character, the authors have used sequences of characters that have been called documents that represent a content produced by each profile. Therefore, these approaches can deal with the problem of the detection of malicious profiles, these latter cannot distinguish the fake account and accounts compromise, hang and Wang [30] have proposed a graphical model to assess the trust between users. The false accounts exist just to manipulate the statistics in the online social networks. Companies, politicians and celebrities use the NSOS to disseminate news and promote products or services. In this situation the profiles can interact with false Commercial Accounts [31]. Cal et al. [32] have a method to detect false accounts based on the graphics. By analyzing the graph, and their system, which should be able to classify accounts, which were more likely to be false. In their fifty thousand First accounts, almost all accounts have been in fact false [33].

Other approaches focus on the analysis of profile for detecting false accounts have been applied in the Twitter and Facebook. In particular, On Twitter, Cresci et al. [34] have proposed an analysis taking into account the features such as the presence of name, to the timing of other NSOS, the actual address and the number of references of other accounts. Later, different algorithms for ML as forest random (RF), Naive Bayes (N.-B.), and the decision trees (J48 of Weka) treated these features. in Facebook also , Fong et al. [35] have taken into consideration the analysis of the avatar on a profile, sex, age, and the name. Noha et al. [42], introduces a classification model based on supervised machine learning techniques and word-based N-gram analysis to classify Twitter messages automatically into credible and not credible. The best performance is achieved using a combination of both unigrams and bigrams, LSVM as a classifier and TF-IDF as a feature extraction technique. The proposed model achieves 84.9% Accuracy, Jiang et al. [36] have proposed a method to check the relationship between false accounts. Once they should promote the same account, the author refers to this kind of accounts as "zombies". The proposed hybrid system seeks to contribute to the advance of state of art for detecting malicious profiles. Xiao Sun et al. [43], proposed a hybrid neural network model called Convolutional Neural Network-Long-Short Term Memory(CNN-LSTM), the model to sentiment analysis on a microblog Big-data platform and obtains significant improvements that enhance the generalization ability. Based on the sentiment of a single post in Weibo, this study also adopted the multivariate Gaussian model and the power law distribution to analyze the users' emotion and detect abnormal emotion on microblog, anomaly detection accuracy of an individual user is 83.49%.Vanyashree Mardi et al.[44],proposed a framework to detect the text-based spam tweets using Naive Bayes Classification algorithm and Artificial Neural Network. Performance study of these two algorithms shows that Artificial Neural Network performs better than Naive Bayes Classification algorithm. The proposed model achieves 92% Accuracy.

Basically, our approach considers issues related to the deployment of the social honeypot, collected data, and detected the similar data by collaborative filtering Thus, they could identify it if a user was legitimate or malicious.

3. Proposed hybrid system

This Section presents details about the proposed hybrid system architecture. As it is shown on Fig. 1, the model architecture can be divided into three parts:

- Security layer based on the social honeypots.
- Security layer based on content filtering.
- Classification layer.

Security layer based on the social honeypots is the first step of the proposed model. The aim is to detect spam tweets on twitter.

The second step in our hybrid system is the Security layer based on content filtering; the content, which are similar to spam tweets, can recommend this step.

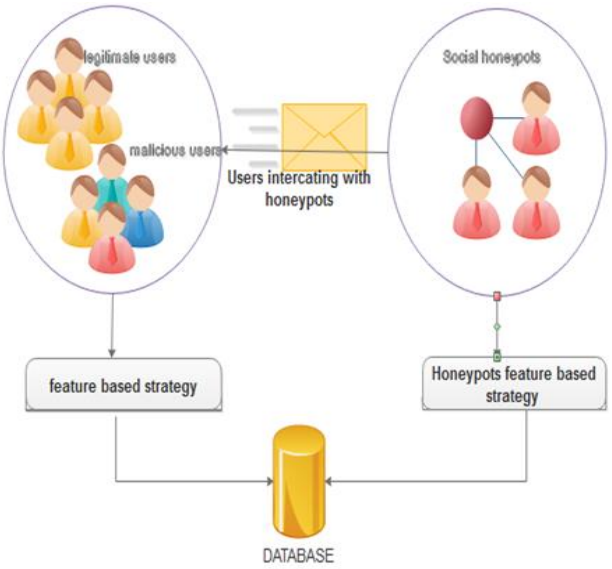

Figure. 1 The deployment of social honeypots for harvesting

The third step called classification layer is its name indicates, it allows you to classify the results found by the two layers.

Social honeypots and recommendation system algorithms are a new concept used for: on the one hand the detection of spam tweets shared between 2 malicious profiles and on the other hand the detection of spam tweets that are similar. This method accelerates the detection of spam tweets. Details concerning each one of these steps are described in the next sections. Section 3.1 presents details about the first layer; section 3.2 presents details about the second.

3.1 Security layer based on the social honeypots

This layer of security based on the social honeypots allows you to detect the spam tweets it is based on the characteristics of the honeypots and characteristics of profiles. The process of this layer is the following:

- The deployment of social honeypots for harvesting information of malicious profiles.
- Analysis of the characteristics of these malicious profiles and those of deployed honeypots for creating classifiers that allow to filter the existing profiles and monitor the new profiles [37].

A spam is malicious content, the problem is to predict if a message m posted on a trend topic (i.e., By including the hashtag or associated keyword) is a spam message via a classifier c:

$$C: m\_i \rightarrow \{spam, non\text{-}spam\} \quad (1)$$

Spammers are characterized by the speed of posting messages over time; we can notice that spammers have quickly posted unwanted messages with a popular topic when it becomes popular. The intention of spammers is to expose spam messages to a larger number of users interested in trends.

This layer for detecting malicious profiles based Social Honeypot deploys and maintains social honeypots for trapping evidence of malicious profile behavior. In our system, it inserts a honeypot which plays a role of a malicious profile and legitimate, if the honeypot attracts a profile, this last is malicious, then the honeypot detects the malicious activities of this profile, (e.g., by crawling the profile of the user sending the unsolicited friend request plus hyperlinks from the profile to pages on the Web-at-large). Social honeypots display certain sensitive keywords in their tweets, which are attractive to attackers. According to Sridharan's survey, attackers will choose their targets based on the content of users' tweets. For example, a spam campaign that wants to promote a kind of diet pill can target users who have the word "weight lose", "slim" or "fat" in their tweets. However, our honeypots must also display sensitive keywords in their tweets.

To prove this idea, a study was conducted using "Term Frequency-Inverse Document Frequency" (TF-IDF) which reflects the importance of a word in a document to extract the most commonly used keywords from honeypots tweets and other influential accounts tweets that are randomly selected. The results of this study show that "follow" and "retweet" are the most commonly used keywords in tweets for honeypots and other influential accounts. Indeed, the acquisition of more followers and more retweet is the common need between most Twitter accounts.

What entails suspicious user behavior can be optimized for the particular community and updated based on new observations of spammer activity [37]. As honeypots collects the characteristics spam content, (for example spam Number of friends, the text on the profile, age, etc.), it is easy to detect at hers to a community of the characteristics of legitimate profiles in the aim of classifying the malicious profile with spam that propagate in the social networks [37]. This is called type of strategy by "Feature based strategy."[38]. A new method used in our approach to improve our classification and increase the ability to detect an attacker on the social networks that is "honeypot feature based strategy", this strategy uses the whole of characteristics of honeypots that interact with users to refine our ranking [38]. The whole data collected

Table 1. Characteristic of honeypots profiles

The number of accounts which malicious interact with a honeypot in the last period of time	5000 accounts
The number of honeypots with which an account interacts	200 accounts
Check if a honeypot posted tweets in return an account or not	yes
Number of honeypot profile	200 accounts
Number of profiles	100000 accounts

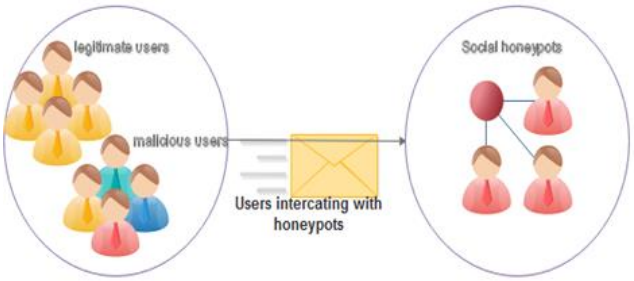

Figure. 2 Collection data

is becoming an integral part of the training of a classifier of malicious profiles. By an iterative refinement of selected characteristics using a set of algorithms for automatic classification, which are implemented on "Weka Machine Learning. Toolkit" the authors can explore the wider space of spam tweets.

First, the researchers have to create 200 profiles Twitter for using as social honeypots and 100000 Twitter profiles, which contains the profiles malicious and legitimate.

Some social honeypots have personal information, such as the biography, the location etc. While others do not have this personal information. To gather information more relevant and increase the probability of being targeted by these malicious profiles, you will create the bots custom Twitter to ensure that all our profiles are connected to Twitter 24 hours per day and 7 days per week. After implementing our honeypots and interact with different types of users, they selected a set of profiles, and for each profile, they extracted the traditional characteristics Feature Based Strategy and the features based on honeypots honeypot features.

The honeypots post tweets that contain links and sensitive keywords to attract malicious profiles to

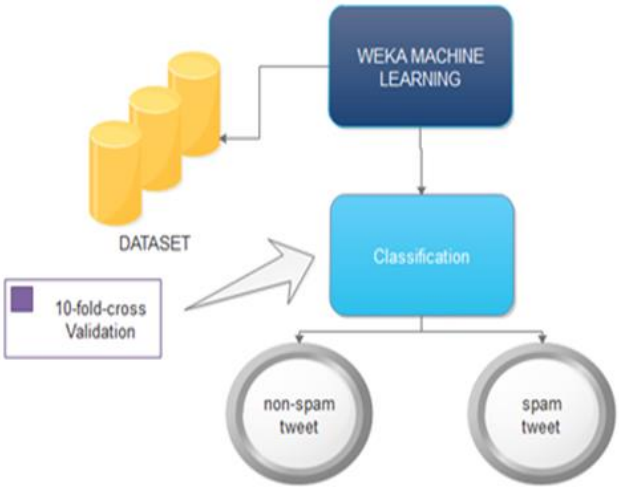

Figure. 3 Classification

react to these tweets, malicious profiles post spam tweets in return.

Through the collection of malicious profile characteristics, we collected the spam tweets posted by all users in order to then apply content filtering to propose a cluster of spam tweets.

The Table 1 represent the characteristic of honeypots profiles. The Fig. 2 present the second step of first security layer based in social honeypots: After collecting the characteristics of profiles, the classification is a necessary step to classify the legitimate tweets and those who are malicious. The researchers chose four types of classification algorithm to make a comparison between them and use the one that gives good results [39].

This step allows you to evaluate the performance of our classification and compare our proposal to the other current approaches. They use the recall, precision-measure, TP rate, FP rate confusion matrix. Recall (sensitivity) is defined as the ratio of correctly classified spam in total real spam. Fig. 3 presents the third step of the first layer of the system proposed.

3.2 Security layer based on content filtering.

Content based systems work with user profiles that are created at the beginning. A profile contains information about a user and his or her taste. Taste is based on how the user has chosen the elements. In general, when creating a profile, referral systems conduct a survey, in order to obtain the first information about a user to avoid the end user's new problem.

In the recommendation process, the engine compares the elements that were already well ranked by the user with the elements that it did not raise and looks for similarities. These items, which are similar, will be recommended to the user.

Content based recommendation systems work by analyzing the characteristics of the objects to be recommended (products, etc.) and then grouping them together. Subsequently, the system will suggest to users who have purchased/consumed any product in the past, the objects/products considered similar, we have used the systems of recommendations in our system for the detection of spam tweets in social networks by proposing spam tweets that are similar to the result of the layer based on the honeypots.

This layer is operated as follows:

The development of the systems layer of recommendation is based on the Algorithm: content filtering, to detect similar tweets that this is legitimate or malicious well.

The idea is to find identical tweets to the layers based on the social honeypot's and collect the maximum of tweets who share the same behaviors, tastes, and operations on the platforms of communication including Twitter.

The functioning of the content filtering algorithm is as follows:

Step 1: Content Analyzer - Depending on the nature of the data to be recommended (text, multimedia, web pages, commercial products, etc.) a pre-processing step is required to describe the objects to be recommended and extract their characteristics. The content analysis module is responsible for producing a structured description of these objects. This description will be used as an input element for the other modules. In our system, we will process the content of 140-character spam tweet.

Step 2: The Profile Learning Module - This module is responsible for analyzing the user's past interactions with system objects. Using methods from the world of learning, this module builds a description of user preferences.

Step 3: The filtering module: From the user profiles and descriptions of the objects to recommend, this module builds lists of suggestions to present to users. The layer of security based on the content filtering gives the tweets that are similar to the spam tweets; Fig. 4 shows the operation of the second layer.

The layer of security based on the content filtering gives the tweets that are similar to the spam tweets; Fig. 4 shows the operation of the second layer.

Fig. 5 presents our hybrid system architecture for detecting spam tweets in the social network.

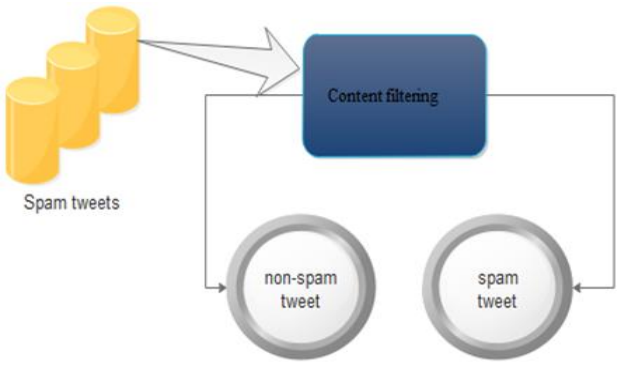

Figure. 4 Security layer-based recommender system

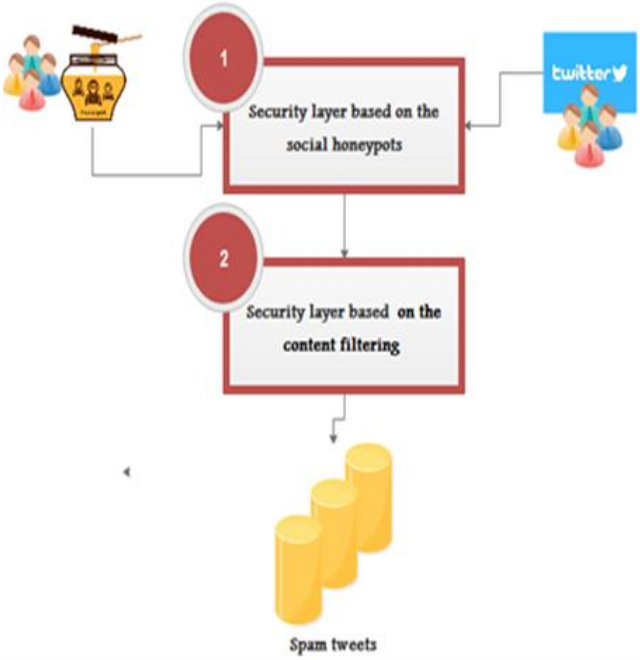

Figure. 5 Hybrid system architecture

4. Results and discussion

This section focuses on the preparation of data that the authors will use in our test. The social network Twitter is known by the publication of thoughts of users in short texts, that is to say the messages published does not exceed 140 characters, which is calls the Tweets that are available to a mannered public. [33].

The data received from Twitter API contains many attributes, for example, the message identification number, ID number of tweets. Our study is interested in different classifiers to make a comparison between the different algorithms for classification and find the one that gives a better result. Using the API streaming.

Twitter from the 1 June 2017 to 20 June 2017 to collect the characteristics of profiles of the system, and the authors collect 100000 Twitter profiles and integrate 200 profiles social honeypots in the system.

Table 2. Analysis of Performance Measures 1

	Random Forest	**Classification via regression**	**Naive bays**	**Trees J48**
Correctly instance	95900	100000	98900	99900
Incorrectly instance	3	0	29	4
Kappa statistic	0.928	1	0.3345	0.9044
Relative absolute erro	14.7544 %	1.7965 %	87.3622 %	12.4222 %
Root relative squared erro	28.9246 %	1.7965 %	105.2429 %	42.5039 %
Total Numbe of Instances	100000	100000	100000	100000

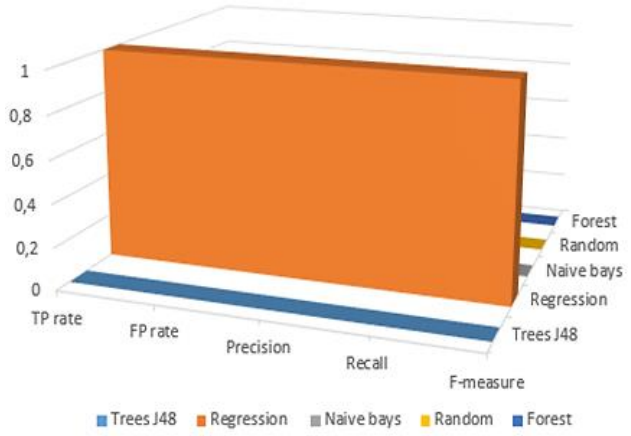

Figure. 6 Analysis of performance measures

For testing our framework, the authors have used a machine learning WEKA to make the classification. The overall results of our hybrid system evaluation are presented in Table 2 and Fig. 6.

We applied 10-fold cross validation on the entire dataset and use different performance measurements to evaluate the results. Accuracy, Precision, Recall, and F-measure as follows:

$$\textbf{Accuracy} = (TP+TN)/(TP+FP+TN+FN) \quad (1)$$

$$\textbf{Precision} = TP/(TP+FP) \quad (2)$$

$$\textbf{Recall} = TP/(TP+FN) \quad (3)$$

$$\textbf{Fmeasure} = 2(Precision \times Recall)/(Precision+Recall) \quad (4)$$

Where TP is the number of tweets correctly identified as Spam, FP is the number of tweets incorrectly identified as spam, TN is the number of tweets correctly identified as no-spam and FN is the number of tweets incorrectly identified as non- spam. The following sections present the results of our experiments.

Our study is interested to apply different classifiers to find the one that is effective and efficient in term of efficiency measures they have applied our proposal on a set of 100000 Twitter profiles.

First, they have tested our new system by four algorithms type including Random forest, Bayes Naive, TreesJ48, a classification via the regression, where the latter is been considered to be the most accurate classifier. Secondly, our comparison of the four classifications reveals on the one hand, the characteristics of a malicious profile and on the second, the efficiency measures, being readily available for each algorithm applied. Although the different measures of effectiveness are applied for each algorithm to choose the more efficient in order to detect the malicious users in Tweet. Our comparison shows that the classification via the regression gives a Precision 99% to identify malicious profiles and (a positive rate, negative rate (100%), recall and F-Measure 99%) compared to the other algorithms of Classification this comparison leads to an improvement of the performance of result. Our choice of algorithm the regression revolves around the measures, which provide effective results and especially this classification tests the entire data with an error rate of 1.7965%.

Table 3. Analysis of performance measures 2

	Trees J48	**Classifi-cation via regression**	**Naive bays**	**Random Forest**	**Category**
TP rate	0.94	1	0.70	O.96	Spam tweets
	0.97	1	0.64	0.97	No-spam tweets
FP rate	0.02	1	0.36	0.02	Spam tweets
	0.05	1	0.29	0.03	No-spam tweets
Precision	0.98	0.99	0.59	0.98	Spam tweets
	0.91	0.99	0.74	0 .94	No-spam tweets
Recall	0.94	0.99	0.70	0.96	Spam tweets
	0.97	0.99	0.64	0.97	No-spam tweets
F-measure	0.96	0.98	0.64	0.97	Spam tweets
	0.94	0.98	0.68	0.95	No-spam tweets

Table 4. Analysis of performance measures

Recommendation system		
TP rate	O.96	Similar
	0.97	Not similar
FP rate	0.03	Similar
	0.04	Not similar
Precision	0.99	Similar
	0 .95	Not similar
Recall	0.97	Similar
	0.98	Not similar
F-measure	0.98	Similar
	0.95	Not similar

The application of content filtering in our contribution increases and improves the detection of spam tweets and especially accelerates the detection time and operation. However, it is enough to detect a spam tweets to find those who are similar by the content filtering. Our proposal provides a model of spam tweets that is characterized on the one hand by the characteristics of account and on the other hand the characteristics of recommendation. This hybrid contribution facilitates the operation of the approach and gives other characteristics of recommendation for the malicious profiles. For the future work, they plan to conduct a more thorough evaluation on the way in which our characteristics would work for the profile and content spam shared by legitimate users, in order to fully understand the effects of bias to continue our approach. Our proposal can offer up to 5000 spam tweets that are similar to the spam tweets detected by honeypots.

Table 4 presents the performance measures for the recommendation of spam tweets that are similar to the parameter tweet. The usefulness of content filtering in our proposal or systems of recommendation is a new trend in social networks; generally, the systems of recommendation are used to make the recommendation of items to users who are similar.

4.1 Comparison (vs.syntax-based methods)

In this section, we compare our method to three existing techniques for detecting spammers and malicious account in social network. The Table 5 present the Accuracy for three others method. In this section, we compare the performance of the proposed model with three models existing in the literature. The first model was introduced by Noha Hassan et al [42] who introduces a classification model based on supervised machine learning techniques and word-based N-gram analysis to automatically classify Twitter messages into credible and non-credible. The results in Table 5 show that this model has an accuracy of 84.9%. Our intuition is that the proposed model performs better than this one because the inclusion of social honeypots and recommendation systems and

Table 5. Comparison with related work

Title	Methodology	Accuracy
[42] Noha Y. Hassan et al.	Detection in Twitter Using Word N-gram Analysis and Supervised MachineLearning Techniques	84.9%
[43] Sun, X., et al.	Detecting anomalous emotion through big data from social networks based on a deep learning method	83,49%
[44] Mardi V et al.	Text-Based Spam Tweets Detection Using Neural Networks	92%
The Proposed model	Socialhoneypot+recommender system+ Classification via regression	99%

classification by Machine Learning algorithms gives a better result.

In addition, we wish to compare our work with another work proposed by Sun et al [43]. who proposed a hybrid neural network model called "Convolutional Neural Network-Long-Short Term Memory" (CNN-LSTM), the sentiment analysis model on a large data microblogging platform and obtained significant improvements that enhance generalizability? Table 5 illustrates the accuracy of the algorithm proposed by [43] and the model based on honeypots and recommendation systems presented in this paper. As shown in Table 5, the model based on honeypots and recommendation systems performs better than the LSTM in terms of accuracy.

However, we have compared our model with other methods to give a good clarification of the performance of our proposal.Vanyashree et al [44] proposed a framework to detect text spam using Bayes' naive classification algorithm and the artificial neural network. The study of the performance of these two algorithms shows that the Artificial Neural Network algorithm performs better than the Bayes naive classification algorithm. The proposed model achieves an accuracy of 92%.

The performance of our model is justified by, the use of social honeypots that attract the attention of malicious profiles and attackers, to retrieve spam tweets that are shared between these profiles;

The role of the recommendation system's content-based filtering algorithm adds a strong point in detecting spam tweets that allows to detect a companion of similar spam tweets in terms of content.

Instead of detecting tweet spam, this method will detect a large amount in a minimized amount of time. Our proposed method using honeypot and recommender system performed better than all current work in terms of accuracy.

For accuracy of our method was about 99%, the first method and the second were 99% and the third was 93%. The usefulness is proposed a system of detection based on the social honeypots and systems of recommendations for: Detecting a spam tweets by the social honeypots and use the content filtering for detecting the tweets that share the appraisals and the tastes between them. The result of the application of new the system, which is composed by two layers, provides a model of a spam tweets in the social networks that is characterized by the characters of account and the characters of recommendation. The recommendation facilitates detection of spam tweets in social networks and gives a comprehensive analysis on the spam tweets and those who are similar.

5. Conclusion

In this paper, we are presented a new system to detect spam tweets in social network Twitter; this system based a social honeypots and content filtering. The hybrid system proposed allows detecting the spam tweets by the characteristics of the social honeypots and malicious accounts inserted in the system, also detecting spam tweets who are similar by the content filtering. They used four types of algorithms to test the proposed system in a machine learning Weka. The classification algorithm via the regression, which gives a better result (precision, a positive rate, negative rate, recall, F-Measure equal 99%). The recommendation by the content filtering allows you to give an idea on the characteristics of the recommendation of a spam content in a social network "Twitter", the layer of content filtering shows that a Spam tweets is a content has abnormal characters in the system. The

results show that our proposed hybrid system can reliably detect spam tweets, and can detect similar spam tweets by content filtering. Future work may include several aspects: (1) they will conduct more theoretical studies on the out performance of our methods to better understand the social honeypots based on malicious user's detection framework. This will in addition, help us improve the performance.

Conflicts of Interest

The authors confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Author Contributions

This paragraph specifying our individual contributions: "conceptualization, Fatna El Mendili and Younès El Bouzekri El Idrissi; Methodology, Fatna El Mendili and Younès El Bouzekri El Idrissi; Software, Fatna El Mendili; Validation, Fatna El Mendili and Younès El Bouzekri El Idrissi; Formal analysis, Fatna El Mendili and Younès El Bouzekri El Idrissi; Investigation, Fatna El Memdili and Younès El Bouzekri El Idrissi; Resources, Fatna El Mendili; Data curation, Fatna El Mendili; Writing—Original draft preparation, Fatna El Mendili; Writing—review and editing, Fatna El Mendili and Younès El Bouzekri El Idrissi; Visualization, Fatna El Mendili and Younès El Bouzekri El Idrissi; Supervision, Younès El Bouzekri El Idrissi".

References

[1] T. Wu, S. Wen, S. Liu, J. Zhang, Y. Xiang, M. Alrubaian, and M. M. Hassan, "Detecting spamming activities in twitter based on deep-learning technique", *Concurr. Comput. Pract. Exp*., Vol. 29, No. 19, p. e4209, 2017.

[2] B. Eshete, A. Villafiorita, and K. Weldemariam, "BINSPECT: Holistic analysis and detection of malicious web pages", In: *Proc. of International Conference on Security and Privacy in Communication Systems*, Berlin, Heidelberg, pp. 149–166, 2013.

[3] B. Eshete, A. Villafiorita, K. Weldemariam and M. Zulkernine, "EINSPECT: Evolution-Guided Analysis and Detection of Malicious Web Pages", In: *Proc. of 2013 IEEE 37th Annual Computer Software and Applications Conference*, Kyoto, pp. 375-380,2013.

[4] A. Aggarwal, A. Rajadesingan and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on twitter", In: *Proc. of 2012 eCrime Researchers Summit*, Las Croabas, pp.1-12, 2012.

[5] M. S.rahman, T. k.huang, H. V.madhyastha, and M.faloutsos, "Efficient and scalable socware detection in online social networks", *In: Proc. of the 21st USENIX Conference on Security Symposium*, Bellevue, WA, pp. 32-32, 2012.

[6] S. Kemp, "DigitalGlobal Overview", accessed on Jan. 24, 2017. [Online]. Available: https://wearesocial.com/special-reports/digital-in- 2017 global-overview.

[7] ISACA. Advanced Persistent Threat Awareness, accessed on 2013. [Online]. Available: http://www.trendmicro.com/cloudcontent/us/pdfs/ business/datasheets/wp_apt-survey-report.pdf

[8] A. Irfan, "How Many Internet and #SocialMedia Users are Fake?", Accessed on Apr. 2, 2015. [Online].Available:http://www.digitalinformationworld.com/2015/04/infographic-how-many-internetsusers-are fake.html.

[9] N. M and J. Prakash, "Detecting Malicious Posts in Social Networks Using Text Analysis", *International Journal of Science and Research*, Vol. 5, No. 6, pp. 345-347, 2016.

[10] M. Verma, Divya and S. Sofat, "Techniques to detect spammers in twitter-a survey", *International Journal of Computer Applications*, Vol. 85, No. 10, pp. 27–32, 2014.

[11] M. A. Fernandes, P. Patel, and T. Marwala, "Automated detection of human users in Twitter", In: *Proc. of 2015 INNS Conference on Big Data*, San Francisco, CA, USA, pp. 224–231,2015.

[12] R. Jasek, M. Kolarik, and T. Vymola, "APT detection system using honeypots", In *Proc. of 13th Int. Conf. Appl. Informat. Commun. (AIC)*, Czech Republic, pp. 25–29, 2013.

[13] S. Webb, J. Caverlee, and C. Pu, "Social honeypots: Making friends with a spammer near you", In: *Proc. of the Fifth Conference on Email and Anti-Spam*, Mountain View, California, USA, pp.1-10, 2008.

[14] K. Lee, B. D. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on twitter", In: *Proc. of the ifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, pp. 1–8, Jul. 2011.

[15] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Compa: Detecting compromised accounts on social networks", *In: Proc. of Network and Distributed System Security*

Symposium, San Diego, CA, United States, pp. 1-18, 2013.

[16] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter, in: Collaboration, electronic messaging, anti-abuse and spam", In: *Proc. of CEAS 2010 - Seventh annual Collaboration, Electronic messaging, AntiAbuse and Spam Conference*, Redmond, Washington, US, p. 12, 2010.

[17] J. Martinez-Romo and L. Araujo, "Detecting malicious tweets in trending topics using a statistical analysis of language", *Expert Systems with Applications*, Vol. 40, No. 8, pp. 2992-3000, 2013.

[18] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, "Detecting spam in a twitter network", First Monday, Vol. 15 No. 1, 2010.

[19] M. Egele, G. Stringhini, C. Kruegel and G. Vigna, "Compa: Detecting compromised accounts on social networks", In: *Proc. of Network and Distributed System Security Symposium*, San Diego, CA, United States, pp. 1-18, 2013.

[20] H. Gao, Y. Yang, K. Bu, Y. Chen, D. Downey K. Lee, and A. Choudhary, "Spam ain't as diverse as it seems: throttling osn spam with templates underneath", In: *Proc. of the 30th Annual Computer Security Applications Conference*, New Orleans Louisiana ,USA, pp. 76-85, 2014.

[21] A. H wang, "Don't follow me: Spam detection in Twitter", In: *Proc. of 2010 International Conference on Security and Cryptography (SECRYPT)*, Athens, pp. 1-10,2010.

[22] A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, and C. Yang, "CATS: Characterizing automation of Twitter spammers", In: *Proc. of 2013 Fifth International Conference on Communication Systems and Networks (COMSNETS)*, Bangalore, pp. 1-10,2013

[23] J. Ma, L. T. Yang, F. G. Mármol, L. J. G. Villalba, A. X. Li, and Y. Wang, "Autonomic and Trusted Computing", In: *Proc. of the International Conference on Autonomic and Trusted Computing, Lecture Notes in Computer Science*, Banff, Canada, pp.143-153, 2011.

[24] A. Chakraborty, J. Sundi, and S. Satapathy, "SPAM: A framework for social profile abuse monitoring", *Technical report, Department of Computer Science, Stony Brook University*, Stony Brook, NY 11794-4400, USA, pp. 1-6, 2012.

[25] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H.wang, "Twitter spammer detection using data stream clustering", *Information Sciences*, Vol. 260, No. 1, pp. 64-73, 2014.

[26] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots+ machine learning", In: *Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Geneva Switzerland, pp. 435–442, 2010.

[27] J. Martinez-Romo and L. Araujo, "Detecting malicious tweets in trending topics using a statistical analysis of language", *Expert Syst. Appl.*, Vol. 40, No. 8, pp. 2992–3000, 2013.

[28] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter. In: Collaboration", In: *Proc. of Electronic Messaging, Anti-abuse and Spam Conference (CEAS)*, Redmond, Washington, US, p.12,2013.

[29] I. Santos, I. Minambres-Marcos, C. Laorden, P. Gal , A. Santamarıa-Ibirika, and P. Bringas, "Twitter content-based spam filtering", In: *Proc. of International Joint Conference SOCO13-CISIS13-ICEUTE13*, Salamanca, Spain, pp. 449–458, 2014.

[30] Z. Zhang and K. Wang, "A trust model for multimedia social networks", *Soc. Netw. Anal. Min.*, Vol. 3, No. 4, pp. 969–979, 2013.

[31] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang, "Detecting suspicious following behaviour in multimillion-node social networks", In: *Proc. of the Companion Publication of the 23rd International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee*, Seoul, Korea, pp. 305–306, 2014.

[32] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services", In: *Proc. of the 9th USENIX Conference on Networked Systems Design and Implementation*, San Jose CA, pp. 15–15, 2012.

[33] S. Barbon, R. A. Igawa and B. B. Zarpelao, "Authorship verification applied to detection of compromised accounts on online social networks A continuous approach", *Multimed. Tools Appl.*, Vol. 76, No. 1 pp. 3213–3233, 2017.

[34] S. Cresci, R. D pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "A fake follower story: improving fake accounts detection on twitter", *International Journal of Computer and Information Engineering*, Vol. 10, No. 1, pp. 13-18, 2014.

[35] S. Fong, Y. Zhuang, and J. He, "Not every friend on a social network can be trusted:

classifying imposters using decision trees", In: *2012 International Conference on Future Generation Communication Technology (FGCT)*, London, UK, pp. 58–63, 2012.

[36] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang, "Detecting suspicious following behaviour in multimillion-node social networks", In: *Proc. of the Companion Publication of the 23rd International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee*, Seoul Korea, pp. 305–306, 2014.

[37] K. Lee, J. Caverlee, and S. Webb, "The Social Honeypot Project: Protecting Online Communities from Spammers", In; *Proc. of the 19th international conference on World wide web*, Raleigh North Carolina, USA, pp. 1139–1140, 2010.

[38] Y. El Bouzekri El Idrissi, F. El Mendili, and N. Maqrane, "A security approach for social network sbased on honeypot", In: *Proc. of the 4th IEEE International Colloquium on Information Science and Technology (CIST)*, Tangier, Morroco, pp. 638-643, 2017.

[39] http://www.d.umn.edu/~padhy005/Chapter5.htl

[40] Z. Rafik and N. Khial, "Filtrage Collaboratif Des Objets Pédagogiques", *Université Abou Bakr Belkaid– Tlemcen Faculté Des Sciences*, 2013.

[41] F. Elmendili, Y. E, Elidrissi, and H. Chaoui, "Detecting Malicious Users in Social Network via Collaborative Filtering", In: *Proc. of the 2nd international Conference on Big Data, Cloud and Applications*, Tetouan, Morocco, p. 44, 2017.

[42] N. Y Hassan, W. H Gomaa, G. A Khoriba, and M. H Haggag, "Credibility Detection in Twitter Using Word N-gram Analysis and Supervised Machine Learning Techniques", *International Journal of Intelligent Engineering and Systems*, Vol. 13, No. 1, pp. 291-300, 2020.

[43] X. Sun, C. Zhang, S. Ding, and C. Quan, "RETRACTED ARTICLE: Detecting anomalous emotion through big data from social networks based on a deep learning method", *Multimedia Tools Appl.*, Vol. 79, No. 9687, 2020.

[44] V. Mardi, A. Kini, V. M Sukanya, and S. Rachana, "Text-Based Spam Tweets Detection Using Neural Networks", In: *Proc. of Advances in Computing and Intelligent Systems*, Singapore, pp. 401-408, 2020.