UNIVERSITY OF TARTU
Institute of Computer Science
Cyber Security Curriculum

Ghada Zakaria

# Detecting Social Spamming on Facebook Platform

Master's Thesis (30 ECTS)

Supervisor: Innar Liiv
Supervisor: Raimundas Matulevičius

Tartu 2018

## Detecting Social Spamming on Facebook Platform

**Abstract:**
OSNs (Online Social Networks) are dominating the human interaction nowadays, easing the communication and spreading of news on one hand and providing a global fertile soil to grow all different kinds of social spamming, on the other. Facebook platform, with its 2 billions current active users, is currently on the top of the spammers' targets. Its users are facing different kind of social threats everyday, including malicious links, profanity, hate speech, revenge porn and others. Although many researchers have presented their different techniques to defeat spam on social media, specially on Twitter platform, very few have targeted Facebook's. To fight the continuously evolving spam techniques, we have to constantly develop and enhance the spam detection methods. This research digs deeper in the Facebook platform, through 10 implemented honeypots, to state the challenges that slow the spam detection process, and ways to overcome it. Using all the given inputs, including the previous techniques tested on other social medias along with observations driven from the honeypots, the final product is a classifier that distinguish the spammer profiles from legitimate ones through data mining and machine learning techniques. To achieve this, the research first overviews the main challenges and limitations that obstruct the spam detection process, and presents the related researches with their results. It then, outlines the implementation steps, from the honeypot construction step, passing through the data collection and preparation and ending by building the classifier itself. Finally, it presents the observations driven from the honeypot and the results from the classifier and validates it against the results from previous researches on different social platforms. The main contribution of this thesis is the end classifier which will be able to distinguish between the legitimate Facebook profiles and the spammer ones. The originality of the research lies in its aim to detect all kind of social spammers, not only the spreading-malware spammers, but also spamming in its general context, e.g. the ones spreading profanity, bulk messages and unapproved contents.

**Keywords:**
social, honeypot, Facebook, spam , detection, machine learning

**CERCS:** P170, Computer science, numerical analysis, systems, control

## Sotsiaalse rämpspostituse avastamine Facebooki platvormil

**Lühikokkuvõte:**
Tänapäeval toimub väga suur osa kommunikatsioonist elektroonilistes suhtlusvõrgustikes. Ühest küljest lihtsustab see omavahelist suhtlemist ja uudiste levimist, teisest küljest loob see ideaalse pinnase sotsiaalse rämpsposti levikuks. Rohkem kui kahe miljardi kasutajaga Facebooki platvorm on hetkel rämpsposti levitajate üks põhilisi sihtmärke. Platvormi kasutajad puutuvad igapäevaselt kokku ohtude ja ebameeldivustega nagu pahavara levitavad

lingid, vulgaarsused, vihakõned, kättemaksuks levitatav porno ja muu. Kuigi uurijad on esitanud erinevaid tehnikaid sotsiaalmeedias rämpspostituste vähendamiseks, on neid rakendatud eelkõige Twitteri platvormil ja vaid vähesed on seda teinud Facebookis. Pidevalt arenevate rämpspostitusmeetoditega võitlemiseks tuleb välja töötada järjest uusi rämpsposti avastamise viise. Käesolev magistritöö keskendub Facebook platvormile, kuhu on lõputöö raames paigutatud kümme „meepurki" (ingl honeypot), mille abil määratakse kindlaks väljakutsed rämpsposti tuvastamisel, et pakkuda tõhusamaid lahendusi. Kasutades kõiki sisendeid, kaasa arvatud varem mujal sotsiaalmeedias testitud meetodid ja informatsioon „meepurkidest", luuakse andmekaeve ja masinõppe meetoditele tuginedes klassifikaator, mis suudab eristada rämpspostitaja profiili tavakasutaja profiilist. Nimetatu saavutamiseks vaadeldakse esmalt peamisi väljakutseid ja piiranguid rämpsposti tuvastamisel ning esitletakse varasemalt tehtud uuringuid koos tulemustega. Seejärel kirjeldatakse rakenduslikku protsessi, alustades „meepurgi" ehitusest, andmete kogumisest ja ettevalmistamisest kuni klassifikaatori ehitamiseni. Lõpuks esitatakse „meepurkidelt" saadud vaatlusandmed koos klassifikaatori tulemustega ning võrreldakse neid uurimistöödega teiste sotsiaalmeedia platvormide kohta. Selle lõputöö peamine panus on klassifikaator, mis suudab eristada Facebooki kasutaja profiilid spämmerite omast. Selle lõputöö originaalsus seisneb eesmärgis avastada erinevat sotsiaalset spämmi, mitte ainult pahavara levitajaid vaid ka neid, kes levitavad roppust, massiliselt sõnumeid, heakskiitmata sisu jne.

**Võtmesõnad:**
sotsiaalmeedium, meepurk, Facebook, spämm, rämpspost, tuvastus , masinõpe

**CERCS:** P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

# Contents

# List of Figures

# List of Tables

# 1 Introduction

Since the beginning of the Social Media in the 90's, concerns were raised about its security, ethical aspects, potential for misuse, psychological effects and more. Since then, continuously, there were always updated statistics showing its dangerous side and discussing solutions to minimize the threats. In this thesis, the author summarizes some of the efforts done before to detect and decrease the threat of social spamming using data mining and machine learning algorithms.
Although the word *spam* usually refers to the malicious contents e.g. malware, in this research, spam will be considered in its general context, e.g. any unwanted text, requests, posts, and photos, including for example, unwanted advertisement, requests targeting wrong persons, nude photos targeting under-aged children, blackmailing ex-partners with intimate photos, profanity and so on ..
The research starts with defining the problems needed to be solved and the scope of the implementation. It then covers the evolution of the spam detection approaches done through the years 2008 till 2017- and the accuracy of their results, as well as the development of the Facebook platform itself. Implementation of the honeypots and the observation driven from its findings are presented in sections 3 and 4. Finally, the results from the implemented classifiers are presented and validated against results from the previous researches.

## 1.1 Problem Statement

As will be highlighted in the background section, most of the researches dealing with social spamming problems were performed on the Twitter platform for different reasons including, the ease of implementing and gathering information through the honeypot and its already available databases to use. The aim of this research is to integrate different spam detection attributes and feed the results to a classifier that can detect spam profiles on Facebook Platform.

Facebook platform was chosen as it currently has more than 2 billion users [1,2] verses 330 millions on Twitter[3]! The statistics' results crown Facebook platform on the top of the most used application worldwide[4]. Also Facebook integrates with and access the data of a lot of other applications, including Instagram, Goodreads, Snapchat and more.

For set of users $U = \{u_1, u_2, u_3, ..., u_i\}$ who each has a profile $p$ with attributes $A = \{a_1, a_2, a_3, ....a_j\}$ which maps to different profile's attributes, like ID, name, gender,

[1]Mark Zuckerberg - Post #10103831654565331
[2]Statista - Active Facebook users worldwide as of 3rd quarter 2017
[3]Statista - Active Twitter users worldwide from 1st quarter 2010 to 3rd quarter 2017
[4]Techcrunch - Facebook now has 2 billion monthly users

friends' number and so on, the target classifier $C$ should be able to distinguish if user $u_k$ is a spammer, or legitimate user given the information from profile $p_k$ .

$$C|_{p_k} \rightarrow \{Spammer,\ Legitimate\ User\}$$

## 1.2 Research Questions

The research targets answering specifically the below questions :

- What is the applicability of implementing a stable honeypot profile on the Facebook platform?
- What is the reliability of a social honeypot to attract only social spammers on Facebook platform?
- What is the effectiveness of the targeted classifier to identify spam profiles on Facebook?
- What contribution does the classifier add to better spam detection more over than the currently implemented spam detection techniques on Facebook?
- Finally, what is the correctness of the proposed classifier with respect to results from Twitter platform, the most tested social platform in the previous researches, from the accuracy prospective.

## 1.3 Main Tools

### 1.3.1 R-Language

The language and environment of the classifier will be *R*[5]; which is most suitable for statistical computing and graphics and provides different ways to integrate between Facebook and some Linux commands in the back-end. Also *R* will be used in the text-mining steps needed to compare the contents of a profile information or a shared post against some keywords commonly used by spammers.

### 1.3.2 Graph API Explorer

The API[6] is the primary way that data is retrieved from or posted to Facebook. The current version of the API 2.11 and was released in November'2017. The API provides the developer with a token, for 2 hours every time, through which the developer can connect to the API through various environment including *R*, *PHP SDK*, and *curl* command.

[5]R-project
[6]Facebook Graph API

Through the API, information of the honeypot can be extracted, its friend-list, posts available to it, and much more.

### 1.3.3 CURL command

For extracting the information from the Graph API, sometimes the *curl* command is more convenient than *R*. The command has all the Graph API's action available to be done through it, while *R* is limited to fewer permissions. While *R* is the main environment, *curl* command can be used to pull the data, and then easily integrated with more commands to prepare the data and save it to be used later by R in the analysis stage. Also, the command will be used to communicate with the various virus-scanners APIs and the results will be cleaned, organized and prepared to be used by the R-based classifier.

## 1.4 Scope and Limitations

### 1.4.1 Honeypot

According to the Facebook community standards, people should connect through their real personal identity, and Facebook team has the right to disable a profile if the owner could not prove their identity[7]. Furthermore, connecting with several suspicious profiles and accepting many friend requests in a short time period questions the authenticity of the honeypot profile and Facebook team will directly disable the honeypot till I am able to prove that the profile belongs to a real person with this exact name [8]. As a result, I am planning to implement several parallel honeypot profiles and continuously create more as each is taken down by Facebook team.

### 1.4.2 Facebook Graph API

Until version 2.3[9], the Graph API supported a lot of features to extract information. Starting from V2.4 the security concerns and restriction disabled many permissions, e.g. manually extracting friends' list and their basic information, extracting the home feed, and extracting the private messages. The target from using the Graph API will be to utilize as much function as needed to automate the extraction of maximum amount of data, while the rest of information will be extracted manually or by the aid of other codes and extensions, e.g. DataMiner extension on Chrome[10].

[7]Facebook: Authentic-Identity

[8]Facebook Help Center - What types of ID does Facebook accept?

[9]Facebook for Developers - Changelog

[10]Data Miner

### 1.4.3 Data Sets' Sizes

Since Graph API will not help extracting enough information for the honeypot's friends, messages, or home feed, I will have to manually extract the information into a database. Information include friend's age, occupation, education, number of friends, and activity rate. This will consume much time, and will force me to be bounded by the amount I can gather in a given period of time. Since the training and testing database are expected to be small, in the classification stage, I will use the K-fold cross validation to test the classifier against several random data sets.

### 1.4.4 Legal and Ethical Concerns

There are different legal and ethical considerations regarding communicating with the added profiles, both spammers and legitimate, with a false identity, and of course, the usage of their information in the research without their explicit approval. Most of the researchers do not precisely present how they overcame this issue. Nevertheless, several authors have dedicated complete researchers to address the legal and ethical concerns related to scientific researches on the OSNs and data mining technologies [1, 2, 3].

Researchers around the world seem to have reached consensus about the importance of independent research on social networks and how they contribute to general social good and act as a safeguard against the rise of information monopoly and abuses by for-profit platforms [4, 5, 6]. Also, many entities and conferences are calling for papers to discuss the OSNs and their impacts in different fields, e.g. education, business, marketing, and social.[11,12,13,14,15,16]

Even the Facebook team itself is continuously announcing how far researching on the platform and using several AI tools to analyze the data is increasing the social quality that the platform is targeting. Very recently, in November-2017, the team has announced upgrading their AI tools to identify people who have suicidal thoughts, and be ready to help them in early stage[17].

In March 2017, MIT announced an award for rule-breakers!. "You don't change the world by doing what you're told, you don't get a Nobel Prize for doing what you're told,

[11] IEEE Statistical Signal Processing Workshop
[12] ASONAM 2018
[13] ELSEVIER
[14] Journal of Theoretical and Applied Electronic Commerce Research
[15] WikiCFP
[16] Emerald Journals
[17] Mark Zuckerberg - post #10104242660091961

you get it for questioning authority." said Joi Ito, the director of MIT's Media Lab[18]. Ito and many other academics and researchers have been fighting a long battle against rigid laws and rules that are not continuously evolving to meet the exponential advance in technology. Through the *EU Data Protection Directive*, the EU has been trying to shape a frame to regulate the collection and processing of electronic data since 1995. Still, progress of changing the law to cope with the everyday growth in the digital world is struggling way behind.

Under EU law, 'personal data' are defined as "information relating to an identified or identifiable natural person"[19]. Following the principle of limited retention of data, contained in the Data Protection Directive as well as in Convention 108, data gathered in this thesis was completely anonymized before even the processing stage starts, since the personal identification of the profiles are not directly serving the research purpose.

European data protection law specially treats the cases where data collection and processing are done for research purposes, as stated in Fig. 1. It sets the ground rules to follow during data collection and processing, then leaves exact outlines to be defined in the national laws for each EU country. Estonian law permits the processing of personal data for scientific research without the approval of its owner in a strict frame that aims at preventing the identification of the person and avoids further storage of data after the objective research or statistics have been achieved (Fig. 2).

> In relation to science, European data protection law is aware of the special value of science to society. Therefore, the general restrictions for the use of personal data are diminished. The Data Protection Directive and Convention 108 both permit the retention of data for scientific research once they are no longer needed for the initial purpose of their collection. Furthermore, the subsequent use of personal data for scientific research shall not be considered an incompatible purpose. National law is charged with the task of developing more detailed provisions, including the necessary safeguards, to reconcile the interest in scientific research with the right to data protection.

Figure 1. Segment from section 'Freedom of the arts and sciences' - Handbook on European data protection law [7]

Personal data protection is cautiously considered in this work; data involved consists mainly of bots and spammers. Yet, all data has already been anonymized and will be completely deleted as soon as this work is published.

[18]CNN - MIT offers $250,000 award for breaking the rules

[19]Data Protection Directive, Art. 2 (a); Convention 108, Art. 2 (a).

**§ 16. Processing of personal data for scientific research or official statistics needs**

**(1)** Data concerning a data subject may be processed without the consent of the data subject for the needs of scientific research or official statistics only in coded form. Before handing over data for processing it for the needs of scientific research or official statistics, the data allowing a person to be identified shall be substituted by a code. Decoding and the possibility to decode is permitted only for the needs of additional scientific research or official statistics. The processor of the personal data shall appoint a specific person who has access to the information allowing decoding.

**(2)** Processing of data concerning a data subject without the person's consent for scientific research or official statistics purposes in a format which enables identification of the data subject is permitted only if, after removal of the data enabling identification, the goals of data processing would not be achievable or achievement thereof would be unreasonably difficult. In such case, the personal data of a data subject may be processed without the person's consent only if the person carrying out the scientific research finds that there is a predominant public interest for such processing and the volume of the obligations of the data subject is not changed on the basis of the processed personal data and the rights of the data subject are not excessively damaged in any other manner.

**(3)** Processing of personal data for scientific research or official statistics purposes without the consent of the data subject is permitted if the processor of the personal data has taken sufficient organizational, physical and information technology security measures for the protection of the personal data, has registered the processing of sensitive personal data and the Data Protection Inspectorate has verified, before the commencement of the processing of the personal data, compliance with the requirements set out in this section and, if an ethics committee has been founded based on law in the corresponding area, has also heard the opinion of such committee.

**(4)** Collected personal data may be processed for the purposes of scientific research or official statistics regardless of the purpose for which the personal data were initially collected. Personal data collected for scientific research or official statistics may be stored in coded form for the purposes of using it later for scientific research or official statistics.

Figure 2. Estonian law - Personal Data Protection Act - Section 16[20]

[20]Riigi Teataja - Personal Data Protection Act

# 2 Background

This section aims at over viewing the past, most-related researches that have been done in the social spam detection field. It firsts reviews the papers and articles published by different researchers and summarizes the used techniques and ideas they used to increase the detection precision on social networks, using photos and charts to best visualize the different methods and results.
Later, it outlines the main spam detection techniques used by Facebook team and describes in particular the efforts and results achieved by the Facebook research team in the area of photo segmentation and object detection, which will be later referred to in this research.

## 2.1 Social Honeypots and Classifiers' Different Techniques

In 2008, S. Webb, J. Caverlee and C. Pu proposed the first social honeypot and technique for harvesting deceptive spam profiles from social network community [8]. 51 honeypot profiles were constructed in different geographic locations on MySpace and traffic through four months (from October 1, 2007 to February 1, 2008) was collected and analyzed.
The researchers have created their honeypots using algorithms that balances between staying maximum online time, for MySpace to mark them as active and prioritize them in the search, and employing a sleep timers to avoid being marked as a spam bot by MySpace. Once the honeypot profile receives a new friend request from a profile (total of 1,570 friend requests were received), it downloads the profile's information along with the time-stamp of friend request and rejects the request. The bots also examine all the URLs and pages that are being advertised on these profiles.

The spam profiles collected were categorized as [8] :

- Click Traps : Images are links to other web pages.
- Friend Infiltrators : Their main aim is to collect as many friends as possible.
- Pornographic Storytellers : The "About me" section has random pornographic stories, and pornographic web pages.
- Japanese Pill Pushers : Advertise for male enhancement pills in their "About me" sections.
- Winnies : same headline "Hey its winnie", and links to female's pornographic pictures.

The top interesting conclusions from analyzing all these profiles were (as quoted from their research [8]) :

- The spamming behaviors of spam profiles follow distinct temporal patterns.
- The geographic locations of spam profiles almost never overlap with the locations of their targets.
- 57.2% of the spam profiles obtain their “About me” content from another profile.
- Many of the spam profiles exhibit distinct demographic characteristics (e.g., age, relationship status).
- Spam profiles use thousands of URLs and various redirection techniques to funnel users to a hand full of destination web pages.

Later, in their paper in 2010, Caverlee et al. have discussed the possibility to depend on f implemented social honeypots to attract the spammers and detect them [9]. They have proposed the framework shown in Fig. 3 and tested it on 2 social networks, Myspace and Twitter. They have created 51 generic honeypot profiles on MySpace [10] and one more on Twitter, and for every profile that sent a friend request to their honepots, they stored a local copy of that profile, extracted all the hyperlinks in the "About Me" sections and crawled the pages pointed to by these hyperlinks. From the analysis of these profiles and the malicious URLs included, the researchers confirmed their first hypothesis, that using honeypots can successfully attract spammers across fundamentally different communities.

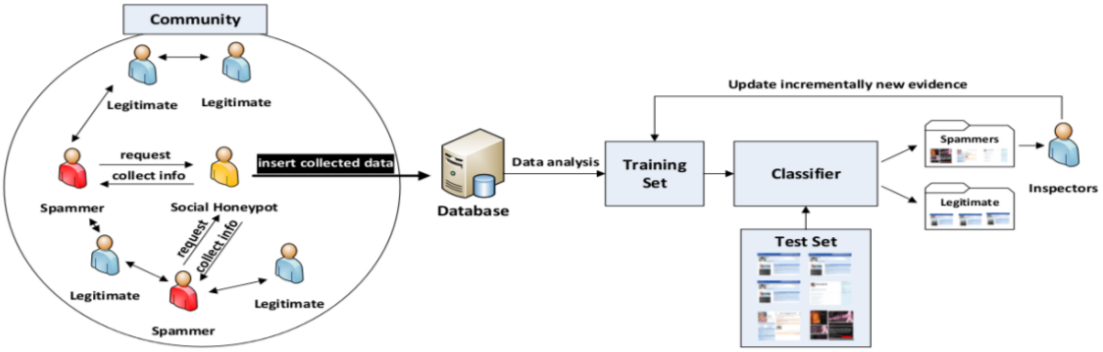



Figure 3. Overall Framework of Social Honeypot-based Approach [8]

The information gathered was categorized to four broad classes of user attributes for each to be analyzed for finding the correlation between this feature’s presence and the possibility of an account to be a spammer. The four main Feature Classes where :

- User demographics: age, gender, location, ...
- User-contributed content: "About Me" text, blog posts, comments on others' profiles, tweets, ...
- User activity features: posting rate, tweet frequency, ...
- User connections: number of friends, followers, following, ...

Around 1300 different profiles were examined for both networks. The classification process was performed using 10-fold cross-validation, and the font analysis was done by Porter Stemmer and Bigrams techniques. The results were very promising, for the most effective features to detect a spammer were independent from the 'personality' a spammer tried to draw in the "About Me" section on the profile (Fig. 4). Instead, the features with the highest True Positive Rate were the Contents/tweet's text, the account age and the shared URLs. Best results showed 99% of accuracy in the MySpace data sample and 88% in Twitter's.

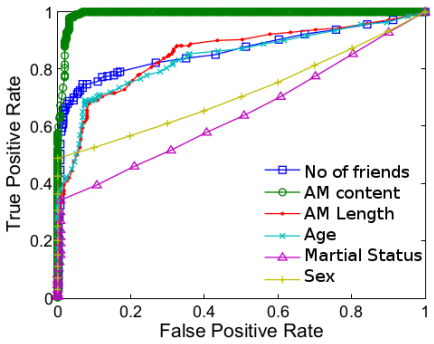



Figure 4. MySpace - Feature Comparison [8](AM : About Me)

The authors tried to deploy their classifier implementation 'in-the-wild' to test if it will be as accurate if implemented any place in the world. The problem in this research challenge was the absence of ground truth data, and so, for evaluating the classifiers, they adopted the spam precision metric

$$\text{Spam Precision} = \frac{\text{true positives}}{\text{true positives}+\text{false positives}}$$

i.e. evaluate only the profiles that the classifier labels as spam. There were 44,000 profiles to test and a human inspector verified whether the newly found predicted spam was actually spam, and accordingly added instances to training set. Fig. 5 shows the

Spam Precision when depending on Sexual contents or Advertisement Content. They have also added a post-filter for the model incorrectly predicted spam labels for profiles containing special profile layout links, e.g."click here to get a theme for your myspace". Lee et al. said that these initial results provide positive evidence of the robustness of the proposed approach.

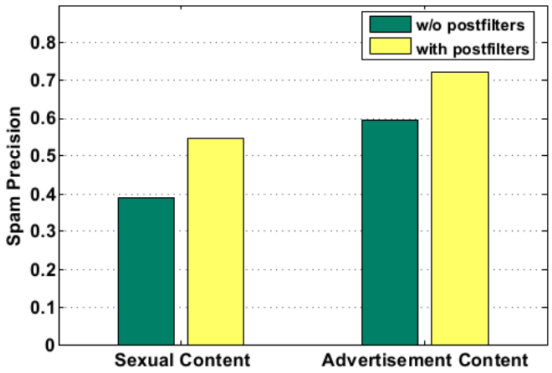
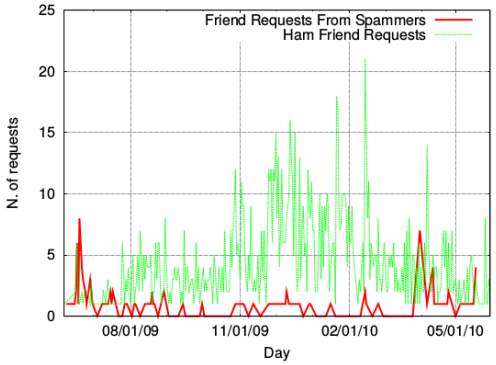


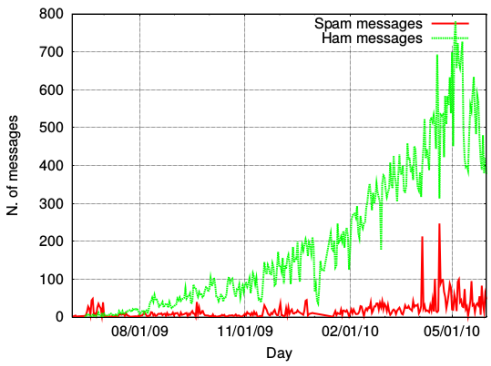

Figure 5. Spam Precision when analyzing Sexual contents or Advertisement Content [8].

On the same year, 2010, Stringhini et al. have created 900 profiles on Facebook, MySpace, and Twitter (300 on each platform) to collect data over 11-12 months, investigate spammers' behavior and characteristics and build a spammer detector based on these information. The pots were decided to be passive, i.e. they will not send any friend request. Once it receives a request, the profile pot logged all the information of the requested profile and all the notifications and private messages it receives. Figures 31 and 7 summarizes Facebook's and Twitter's notification rate through the 11-to-12 months period [11].
From the analysis of the spam profiles' data, the researchers could divide the spammers into four main category:

- Displayer : Spam contents are put on his/her profile, and a friend intentionally visit the profile to check it, and so, it is considered the least effective.
- Bragger : Spammer posts the malicious contents on his own feed, and so, his friends can view it on their feeds.
- Poster : Spammers who send a direct message to each victim, usually by posting on their walls, or sharing the malicious contents in a group or so.
- Whisperer : Those who send private messages to their victims requesting that they, personally and in specific, check a URL to download some files.

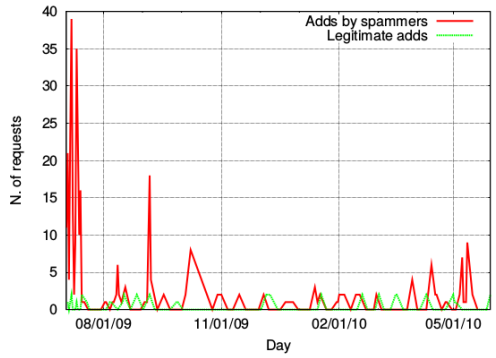



(a) Friend requests received

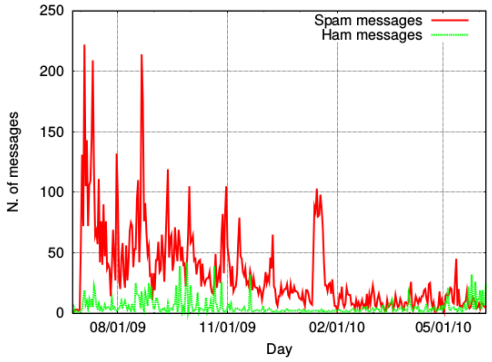



(b) Messages received

Figure 6. Activity observed on Facebook [11]




(a) Users starting following honey-profiles




(b) Messages received

Figure 7. Activity observed on Twitter [11]

These 4 categories of spammers can be further divided according to their behavior :

- Greedy bots : When every message contains a malicious content. They are easy to find and flag as spammers.
- Stealthy bot : Usually sends legitimate messages, and include a spam every now and then.

Out of the 534 spam bots detected, 416 were greedy and 98 were stealthy. Fig. 8 shows a summary of the spammers ratio on the three social platforms.
Using machine learning techniques, the authors have decided to use several indexes to classify spammers and legitimate users (some of them are listed below). Using these

Figure 8. Spammers ratio on the three social platforms [11].

| Network | Overall | Spammers |
|---|---|---|
| Facebook | 3,831 | 173 |
| MySpace | 22 | 8 |
| Twitter | 397 | 361 |

**Table 1: Friend requests received on the various social networks.**

| Network | Overall | Spammers |
|---|---|---|
| Facebook | 72,431 | 3,882 |
| MySpace | 25 | 0 |
| Twitter | 13,113 | 11,338 |

**Table 2: Messages received on the various social networks.**

indexes, a classifier, that uses *Random Forest algorithm* , was created for both, Facebook and Twitter platforms and tested.

- FF ratio : Compares the number of friend requests that a user sends to the number of friends he has. This is based on the idea that this unknown spammers will have lots of the requests rejected because they do not actually know him.
- URL ratio : The high likelihood of a spammer to send out lots of malicious URLs.
- Message Similarity : The similarity of messages sent by the spammer to different friends.
- Number of messages sent : Based on the observation that spammers tend to send more number of messages compared to legitimate users.

The classifier and indexes sensitivity were tunned for each platform for the best results. Finally, the 10-fold cross validation on this training data set estimated 2% false positive ratio and 1% false negative ratio on Facebook platform, Vs. 2.5% and 3%, false-positive ration and false negative ratio respectively on Twitter's platform.

Meanwhile, an algorithm was proposed to estimate the location of a user based on the profile's contents and most used 'local' word [12]. The researchers then calculated the accuracy of the algorithm and error in estimated distance. An example is shown in Fig. 9 for the word "rockets" and its occurrence probability according to the location and Fig. 10 summarizes the most commonly used word in each city in the United States.

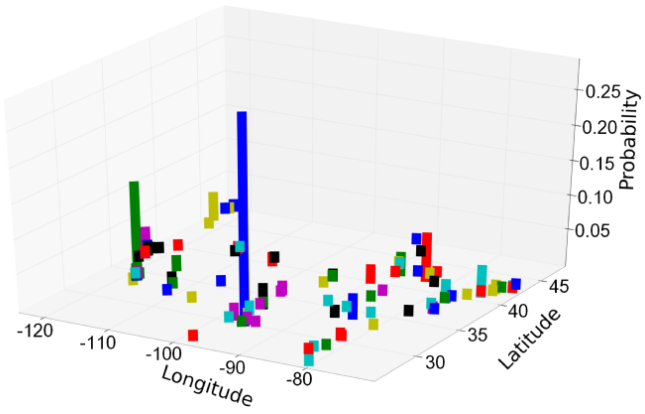



Figure 9. City estimates for the term "rockets" [12]

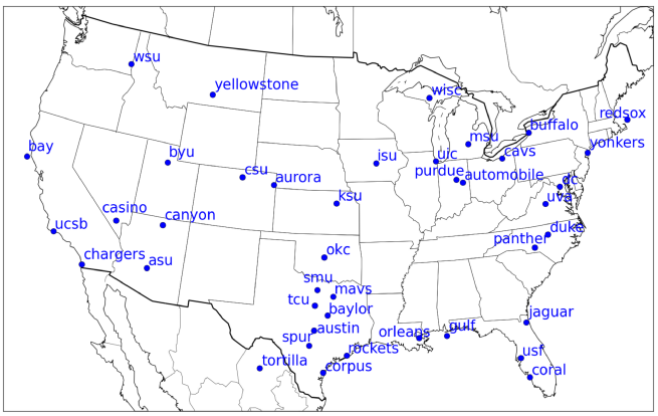



Figure 10. Geographical Centers of Local Words Discovered in Sampled Twitter Dataset [12]

Song et al. suggested a spam filtering approach based on the relationship between the sender and receiver [13]. It depends on the distance, or number of hobs between the message receiver and the message sender, and the correlation between them. The experiment was conducted on 10,000 legitimate messages and 10,000 spam messages. While only 0.9% of the messages are spam at distance=1, 89% of the messages are spam at distance=4. Moreover, the connectivity, i.e. how many common paths/friends, further confirms the spammer detection. Fig.12 shows the ROC curve comparing the results using distance alone, and distance + each of the connectivity algorithm.
Fig. 13 shows the construction of the classification tree and the false positive percentage using different connectivity measurements.

In 2011, Wang et al. suggested a framework that works independent of the social network

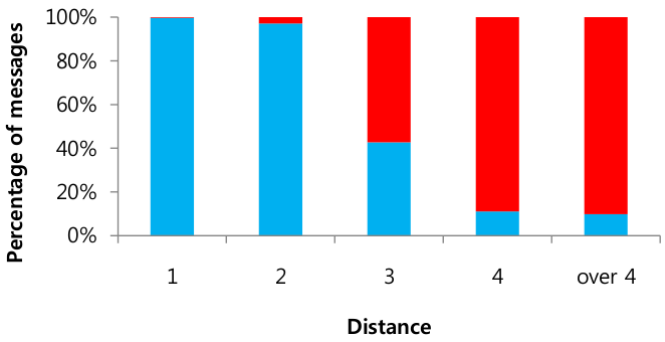



Figure 11. Legitimate(blue) and spam messages(red) Ratio [13]

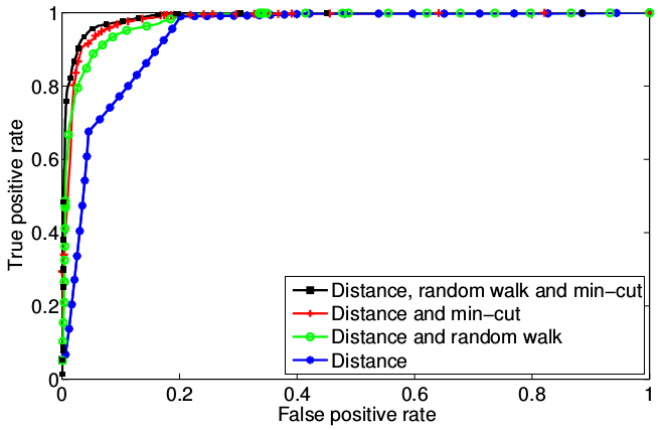



Figure 12. ROC curves for each of the relation features [13]

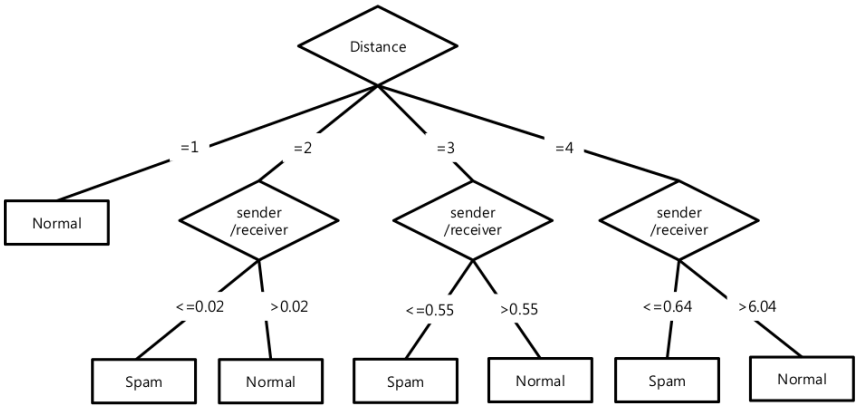
platform, i.e. suitable for all social networks [14]. They claimed that once a new type of spam is detected on one network, it can automatically be identified on the other networks as well. Fig. 14 overviews the suggested framework.
The main idea is based on creating a 3-stages model:

1. Mapping and Assembly Stage, converts the objects of the social network into standard objects for the generalized model ((e.g., profile, message, or webpage)

2. Pre-filtering Stage, searches for already-known hashes, URLs, ... that are known to be spams.

3. Classification Stage, uses supervised machine learning techniques to classify the objects.

The classifier returns a decision per each model (profile, message, and webpage). The decisions then are passed to the combined classifier. The four different combination strategies used were :




A decision tree created by the J48 classifier

| Classifiers | True Positive (%) | False Positive (%) |
|---|---|---|
| Bagging | 94.6 | 6.5 |
| LibSVM | 94.0 | 5.8 |
| J48 | 93.9 | 5.3 |
| BayesNet | 93.5 | 5.5 |
| FT | 93.5 | 5.5 |

The results of the classification using the distance and min-cut

| Classifiers | True Positive (%) | False Positive (%) |
|---|---|---|
| Bagging | 95.1 | 4.7 |
| LibSVM | 94.3 | 4.3 |
| J48 | 94.2 | 4.6 |
| FT | 93.8 | 4.4 |
| BayesNet | 93.4 | 5.9 |

The results of the classification using the distance, random walk and min-cut

Figure 13. Decision tree, structure and results [13]

- AND strategy: classifies an object as spam if all classifier, for all models, classify it as spam.
- OR strategy: classifies an object as spam if any of the classifier, classifies it as spam.
- Majority voting strategy: classifies the object as spam only when majority of classifier, classify it as spam.
- Bayesian strategy, based on the marginal probability of the joint probabilities of the classifiers' output.

Results showed that *Naïve Bayes classifier* showed the best results (Fig. 15).
By manual inspection, some URLs were found to be misclassified as spam while they are legitimate. The classifier was further enhanced by white-listing them and tested again.

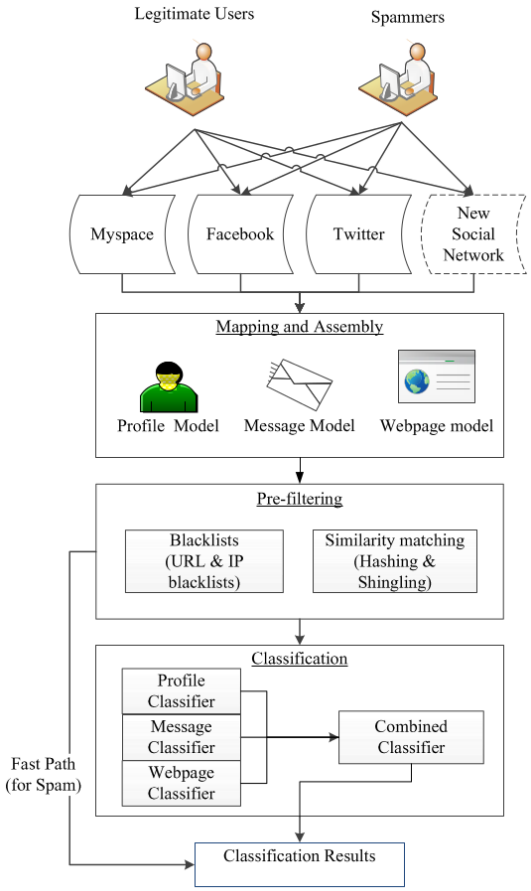



Figure 14. Overview of the spam detection framework [14]

| | Predicted Legitimate | Predicted Spam |
|---|---|---|
| True Legitimate | 3286 | 1714 |
| True Spam | 830 | 4170 |

Figure 15. The results of Naïve Bayes classifier [14]

Fig. 16 shows the enhanced results.

In their paper in 2013, Hu et al. discussed the implementation and accuracy of their proposed algorithm, SSDM, to label social spammers in microblogging [15]. Compared to Least Squares and Elastic Net algorithms, the proposed algorithm showed higher precision (0.865). For further evaluation of the algorithm, it was tested against support vector machine (SVM) and elastic net (EN) using 2 methods, Content-based method and

| | Predicted Legitimate | Predicted Spam |
|---|---|---|
| True Legitimate | 2890 | 121 |
| True Spam | 561 | 4242 |

Figure 16. The results of Naïve Bayes Classifier after whitelisting legitimate sites [14]

Network-based method. Fig.17 compares the SSDM algorithm against the other four algorithms.

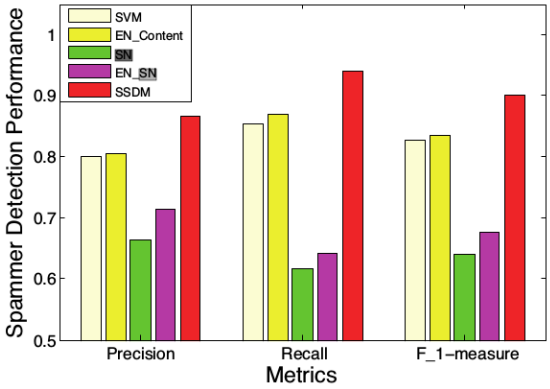



Figure 17. SSDM Performance [15]

In the following year, 2014, the same authors have proposed an online-learning algorithm to detect social spammer and have compared it to other commonly used techniques and to offline learning algorithms [16]. The two main algorithms to compare against is the Least Square, which works on minimizing the error, and BSSD which is the batch version of the proposed online model. The proposed online algorithm, OSSD, showed the second best results when compared to four other algorithms. The best result was obtained using the batch version of the OSSD, BSSD, but the difference was quite close as shown in Fig. 18.

Using statistical analysis of language, Martinez-Romo and Lourdes Araujo detected the malicious tweets in the most trending topics on Twitter [17]. A language model is based on a probability distribution over pieces of text, indicating the likelihood of observing these pieces in a language. As a simplified example, if the analyzed tweet is about "Justin Bieber" and the URL included is dedicated to the sale of pharmaceutical products, the tweet is more likely to be suspicious/malicious.

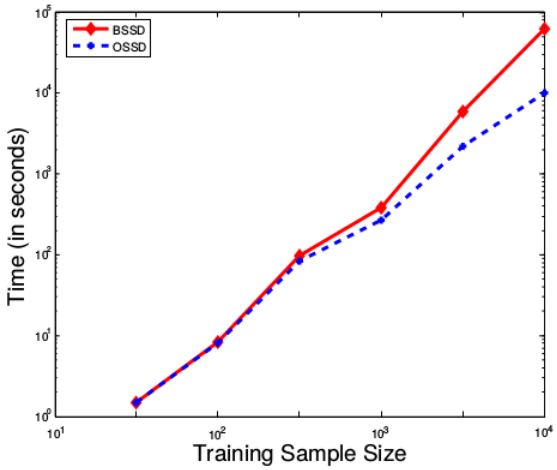



Figure 18. Efficiency Performance on Twitter Sample [16]

The model also includes a feedback mechanism by which the user can confirm, for example, that the spam-labeled tweet is not a spam, and then the model corrects its label for this specific tweet along with all the related tweets. The architecture of the proposed model is summarized in Fig.19. Fig. 20 displays the accuracy of the model using different machine learning techniques and Fig. 21 summarizes the confusion matrix, which shows that the model is able to detect spam content with accuracy of 89.3% and non-spam content with accuracy reaching 93.7%.

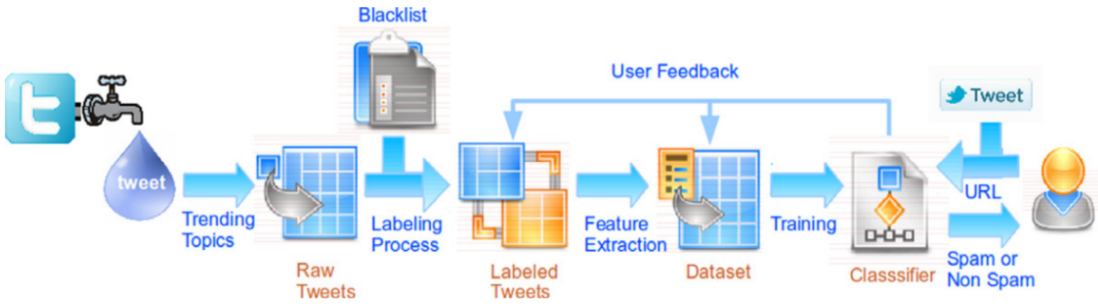



Figure 19. Feedback mechanism [17]

Recently, in January 2016, Ruan et al. approached the spammer-detection problem from the other way around[18]. Instead of analyzing the spammers' behaviors and profiles' contents, they have analyzed and created a model for the legitimate users and have used it to compare against for spammer detection. The authors have first categorized the users to Extroverts, and Introverts, and analyzed each group's behavior independently. The

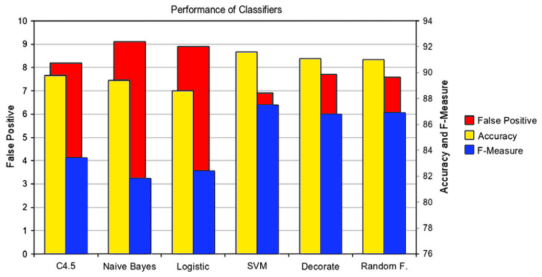



Figure 20. Results for Different classification algorithms [17]

| Twitter spam dataset – confusion matrix | | |
|---|---|---|
| True | Predicted | |
| | *Spam* | *Non-Spam* |
| *Spam* | 89.3% | 10.7% |
| *Non-Spam* | 6.3% | 93.7% |

Figure 21. Confusion Matrix [17]

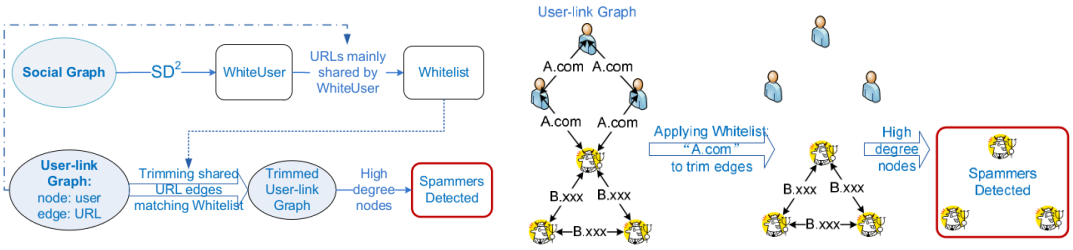
analysis included activity characteristics, browsing preference, sites' visit duration, and others.

Beside the various different supervised techniques and algorithms proposed and studied by different researchers, fewer have tended to use the unsupervised techniques in spammer-detection problem. Unsupervised technique does not require labeling the profiles/contents as legitimate/malicious for the learning phase, and so saves a lot of time and man power. On the other hand, the main disadvantage of the methods based on unsupervised learning is that they usually output less accurate results than those based on supervised ones.

In their paper in 2013 [19], the researchers have discussed the need to find an unsupervised approach with accuracy similar to the supervised ones. The model is constructed based on the graph data approach. First the contents shared by legitimate users creates a Whitelist database and then, upon testing, the nodes with these contents are trimmed from the graph data, so that only the malicious nodes are left behind to be tested. This is can be clearly illustrated in Fig. 22.

In late 2016 and early 2017, scholars continued searching for different methods and algorithms to detect the continuously improving spamming techniques, including:

- Sparse Group Modeling to detect outliers that are not following a specific group structure [20].
- Multi-view learning. The researchers have first worked on proving that singly relaying on Text Features, URL Features or Hashtag Features, will result in the




(a) UNIK workflow

(b) An example of edge trimming in UNIK

Figure 22. UNIK Framework [19]

same distribution of the spammers/legitimate users, i.e. any of the three features can indicate almost-equally confirm the spamming behavior [21]. While multi-viewing increases the detection by around 30%. Also, they have illustrated the need to complete the missing values in the data sets to assure a more precise spamming detection.

- Topic-Based model, to detect real-time spammers that are using the tending topics and hashtags to promote for their malicious links [22]. The research group has used the Hidden Markov model and could prove its more precises results compared to other supervised methods (Random Forest and J48 decision tree).
- Comparing the results for online learning and batch learning with considering different combinations feature sets [23]. The researchers concluded that the 'user network feature' and 'user activity features' were the most robust features against the different spamming patterns.
- Detecting spammers on Facebook platform based on contents, especially the comments [24].The researchers have constructed the classifier using Maximum Entropy method.
- Analyzing public features on Twitter platform to detect the spam tweet's most commonly used words, pattern, text-to-link ratio, and related text attributes. [25].
- Detecting spam messages by analyzing the relation between the users involved and the messages instead of depending on the contents itself [26].
- Detecting spam in closed Facebook groups using social features [27].

## 2.2 Facebook Publications and Researches

Facebook Research team is continuously contributing to the research field by finding solutions to the problems derived from real world [21]. The research team also shares the softwares, platforms, and codes. to be downloaded [22].
Their top research fields are :

- Applied Machine Learning.
- Computer Vision.
- Connectivity.
- Data Science.
- Economics & Computation.
- Facebook AI Research (FAIR).
- Human Computer Interaction & UX.
- Natural Language Processing & Speech.
- Security & Privacy.
- Systems & Networking.
- Virtual Reality.

In this section, the author aims at highlighting the main researches that increases the platform's security and helps detect the anomalies .

### 2.2.1 Object segmentation and Refining

Through several publications [28, 29, 30], Facebook research team have worked on the image segmentation on both, the object level and the pixel level. And on 25 of August, 2016, Facebook officially published its artificial intelligence (A.I.) software for segmenting objects within images on GitHub under a BSD license [23]. The goal is to give the visually impaired persons a different experience when browsing their Facebook's account content. The 3-staged detection technique is summarized by Piotr Dollar in the simple, yet informative, article 'Learning to Segment' [24] as follows :

[21] Facebook: Research Areas
[22] Facebook Immune System
[23] GitHub: Deepmask
[24] Facebook Research - Learning to Segment

1. DeepMask generates initial object masks.

2. SharpMask refines these masks.

3. MultiPathNet identifies the objects delineated by each mask.

Objects' relationship was addressed in the 2017's publication 'Relationship proposal networks' [31], which aim at finding the relationship between objects within the same image. The researchers try to identify 3 types relationships: Interactive : which has at least one object identified as a living being, e.g. a boy flying a kite. Positional : which tried to identify a position of an object in the space, e.g. a kite in the sky. Attributive : which aim at finding the relationship between one small-sized object, and another large one, e.g. a brick in the building.
Fig. 23 visualizes the results targeted by the research team [31].

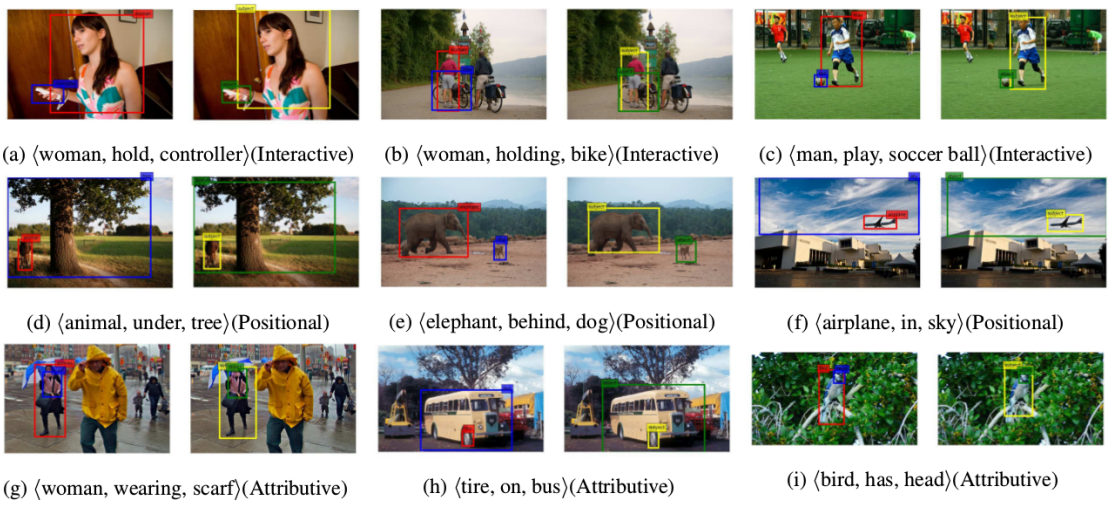



Figure 23. Object segmentation and Refining: Red and blue boxes are ground-truth subject and object, yellow and green boxes are outputs from our model [31]

### 2.2.2 Facebook Spam Detection

In April 2011, the Facebook team has published a paper describing Facebook's immune system at that time. Their main target was to protect the Facebook social graphs from attackers, by first, categorizing the suspicious accounts into three main categories, then taking the corrective actions accordingly.

- Compromised accounts : that have been stolen from real user, should return to their legitimate main user.

- Fake accounts : with no real user / user information, should be blocked.
- Creepers : who are causing inconvenience to other users, e.g. by spreading too much messages that are considered spam by others, or sending friend requests to people they dont know. They need to be educated and reminded of the Facebook's rules.

It is interesting to note that, although the paper is more than 6 years ago, with Facebook expanding to more than double number of users (from 750 millions to 2 billions) and many more spam techniques have been introduced since then, the categorization is still valid to target. The embedded classifier also considers both, scanning the URLs posted and their re-occurrence as well as the users' marking as a spam. Exact details for the technology and flowchart for the detection system is represented in their paper "Facebook Immune System" [32]. Later, Facebook strengthened its spam fight using pure functions[25], and Haskell[26] to cope with the increase in scalability and complexity of the platform.

### 2.2.3 FastText

In August 2016, Facebook team has announced open-sourcing the fastText code. FastText is used for text classification and word representation using a hierarchical classifier that categorizes the words in a binary tree structure instead of a flat list using Huffman code. It participates in the spam detection process by categorizing the bag of words into spam and non-spam. Word representation also detects the rare words in the context and accordingly, understand the aim of the text in general. Facebook also uses this information to promote for related Ads [33, 34]. [27,28]

### 2.2.4 Non-Consensual Intimate Image

The most recent Facebook's attempt to detect and prevent spam was publicly announced early in November 2017. Facebook, in cooperation with the Australian eSafety Commissioner's Office are publishing a pilot that aim at preventing 'revenge porn' on Facebook platform. "We don't want Facebook to be a place where people fear their intimate images will be shared without their consent.", said Antigone Davis, Global Head of Safety at Facebook, starting her explanatory article on the 9th of November, 2017[29]. The article aimed at illustrating the targeted technique and privacy and justifying the need to proceed with this step, after several writers have outraged their privacy concerns of the process in

[25]Facebook: Fighting Spam with Pure functions
[26]Facebook: Fighting Spam with Haskel
[27]Facebook: FastText
[28]GitHub: FastText
[29]News Room - The Facts: Non-Consensual Intimate Image Pilot

different articles and described the project as 'far-fetched'.

It is still too early to evaluate the project and predict the public's feedback towards it. Yet, it is very interesting to note that depending on the Hash value of the picture will definitely not be enough to match against, since it changes with on single pixel change/crop. Also, it is not clear yet how far will there be interaction from Facebook personnel to evaluate the results without having access to the main picture. Finally, how and for how long will Facebook store these images before deleting them from the database !

## 2.3 Summary

Since 2008, many researchers have examined different methods to detect spamming behavior and/or contents. Most of the techniques showed an accuracy above 75%. Yet, none of them have explained in details how could they manage to not violate the right-to-privacy of the spammers while using their profiles and contents!

The development of technology and technique every day will require a continuous research and enhancement in the methods used to detect the spammers, and there will always be no 100% one reliable algorithm to follow when dealing with human behavior.

Although many have examined well the twitter platform because of its nature that by default shows all the contents in public, fewer have tested the Facebook platform that is extensively taking over all the other social network platforms and even less have tried unsupervised techniques with it.
On the other hand, Facebook research team has achieved a lot of progress in the spam detection area over the years, that definitely helped the Facebook to be on the top of all the social applications and platforms, with most number of users nowadays.

# 3 Honeypot Implementation and Inspection

In this section, the author views the practical implementation of the social honeypots and the peripheral tools used to gather the data and extract meta-data from the contents derived from these honeypots to be used later as an input to the constructed classifier.

## 3.1 The Honeypot

### 3.1.1 Basic Implementation

For this research, 10 honeypot profiles were implemented on the Facebook platform, out of which six were marked female profiles, and 4 males'. The profiles were created using 10-minutes mail addresses [30]. The honeypots were claimed located in USA, Russian, Estonia, Latvia and Poland. Profile pictures, 'about me' sections, 'intro about me' section, information about education and more information were varied, some times similar to the spammers that had already been found, sometimes to attract particular spammer group the author had previously detected. Eight honeypots were deactivated by Facebook in less than 4-weeks, including one that was deactivated the second step after registration!. One honeypot, the very first, lasted 6.5 weeks and the other, the very last, is still active in its $7^{th}$ week as per the $1^{st}$ of December, 2017.

### 3.1.2 The Star Honeypot

This was the first honeypot to be implemented. By far, it had the most traffic, received most messages, highest number of friends and the most interaction with others. The honeypot 'Avy' was designed to be for an attractive girl who shares several of her photos on her profile. The profile also included, available to the friend-list, her Gmail address and a created blog that is listed as her website. Following a common pattern for such created profiles, Avy had "Searching for Love" as an 'Intro' in the About Me section and marital status as Single.

The honeypot 'Avy' was decided to be a passive one, i.e. it will not approach and add friends, instead, will wait for other profiles to add her. Meanwhile, the profile joined several groups, political, singers, and adult ones. Also Avy was actively liking and commenting others' posts in the groups and pages it joined. It also marked several events as 'Going' and registered for events through Eventbrite application[31].

For a month, no one has added Avy as friend or approached it anyhow. Then, through the profile, I have clicked on a malicious link on a suspicious profile shown in Fig. 24.

[30] 10 Minute Mail
[31] Eventbrite

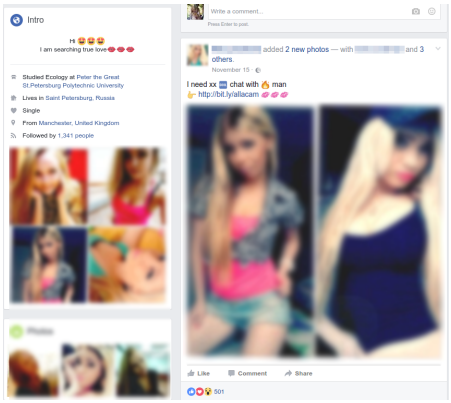



Figure 24. Avy's first malicious link

For not infecting my machine, this was done on a virtual machine to avoid the download of any malware. The suspicious profile had 4115 friends and 1346 followers at that time. Eight hours later, the profile had received 205 friend requests and more than 50 followers (Fig. 25). All the profiles that sent the friend request had the owner of the malicious post as a friend. In one week, Avy had more than 2000 friends, more than 1000 followers and more than 900 friend requests pending approval.

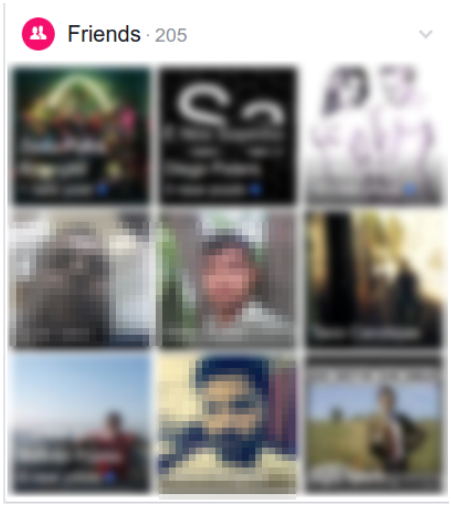



Figure 25. Early stage - Avy's friends

## 3.2 The Blog

In order to provide more channels for the spammers to target the honeypot, the author has created a blog on Blogger [32] and inserted its URL on the About Me section of the honeypot Avy. The blog is monitored by Google Analytics [33] to analyze all the traffic that passes through it. Google analytics also provides the feature of showing the referrals, which are the times when someone visits the blog not through typing the direct link, but instead by clicking on the blog's URL from another page. For example, if someone clicks on the blog's address from the About Me section on the honeypot, the analysis will show that someone was directed from a Facebook page to the blog.

To further ensure that the data is accurate, I have excluded my IP so that the analysis will not consider my own hits. In 12 days there were very interesting results. Total number of sessions on the blog were 28, out of which 7 only were through direct link, and the blog was attacked twice!. The overview report of the blog's traffic showed 2 suspicious languages, *Vitaly rules google* and *Congratulations to Trump and all americans*, as shown in Fig. 26. Interestingly, in the referrals report, same number of sessions were showing suspicious entries because I did not place the blog address on these sources, *twitter.com* and *washingtonpost.com*, as shown in Fig. 27. Both of the spams seem to be spreading and many pages are being infected. Analytics-Toolkit explains the spams and how to get rid of it in 2 main articles [34,35].

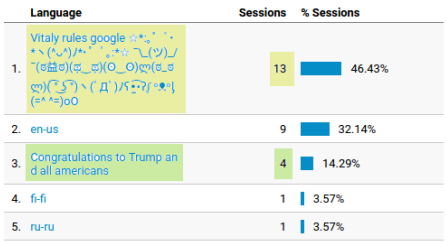



Figure 26. Languages of different sessions

Only 4 real users have clicked on the blog's link placed on the Facebook honeypot. The visitor were from Russia, Finland and USA.

[32] Blogger
[33] Google Analytics
[34] Analytics Toolkit - Vitaly rules google
[35] Analytics Toolkit - Language Spam

| Source | Acquisition<br>Sessions |
|---|---|
| | 21<br>% of Total: 75.00% (28) |
| 1. twitter.com | 13 (61.90%) |
| 2. washingtonpost.com | 4 (19.05%) |
| 3. facebook.com | 3 (14.29%) |
| 4. m.facebook.com | 1 (4.76%) |

Figure 27. Referral sources of different sessions

## 3.3 Data Collection

As mentioned earlier in section 1.4.2, Facebook graph API has limited the information that can be extracted by the user, including the information they can graphically see on their friend's profiles. Therefore, most of the spammer and legitimate profiles' information were extracted manually by the researcher.
From the honeypot's friends, 100 profiles were selected from all the honeypot, all of which had their attributes manually extracted to the database, then marked according to the categories mentioned later in section 4.1, i.e. the social spammers' different types. Legitimate users were picked from the researcher's circles of friends and colleagues to ensure their legitimacy and assures their profiles' variety, including home towns, different posting patterns, and presence of advertisers. Luckily, this latter group had 30 profiles' information extracted by the graph API as their owners had already installed the graph API application before.

## 3.4 Analyzing the Contents

Using Facebook API, VirusTotal API and the curl command, the author have built a code that extracts all (according to the set limit) the contents for the posts the honeypots shared or were tagged in, including the comments. The code then organize the output in JSON format using jq command and later extracts the URLs and/or the targeted words, along with the name and ID for the user who posted it. If the extracted content is URL, the code scan it over VirusTotal and detect its malicious ratio. The chosen words include, 'Click Here', 'Vote' , 'chat', ... which were observed the most on the spammer profiles harvested as well as the words detected earlier by Hailu Xu, Weiqing Sun, and Ahmad Javaid [35] summarized in Fig. 28. The main limitation of the code is that the Facebook API needs to be requested and updated in the code every 2-hours from the Graph API Explorer[36]. The general flow of the code is summarized in Fig. 29.

[36]Facebook for Developers - Tools and Support

| Top 20 Word Features of Spam Posts in Facebook | | | |
|---|---|---|---|
| ***Freq.*** | ***Word*** | ***Freq.*** | ***Word*** |
| 374 | http(s) | 24 | fuck |
| 250 | money | 24 | Internet |
| 80 | sex | 24 | marketing |
| 72 | mobi | 24 | fast |
| 50 | sexy | 22 | photo |
| 40 | justinbieber | 20 | cutshare |
| 36 | online | 20 | video |
| 30 | business | 20 | pregnant |
| 30 | new | 18 | blog |
| 30 | free | 18 | model |

Figure 28. Top 20 words features in the spam of Facebook [35]

```
| Facebook API token = "xxx"
| curl get last 500 content blocks posted related to the user
| jq
| grep  url & saved words
   | curl post url to VirusTotal
| If comment was marked malicious, code outputs the owner ID & name.
```

Figure 29. General flow of the contents code.

## 3.5 Image Inspection

As mentioned earlier in section 2.2.1, Facebook developers have already reached an advance level in detecting the objects in a posted photo. The algorithms' results can be checked using the 'inspect' option while browsing Facebook using Google-Chrome. Fig. 23 was captured from one of the honeypots friends' wall.
Similarly, the objects were identified in the photos shown in Fig. 31 as per each's caption. Further examples extracted from photos with humans from both legitimate and spammer profiles are shown in Fig. 32.
Although the algorithm is showing a high success rate in detecting objects (exceeds 95% true-positive), it had not yet output the exact objects' relationships, and orientation, e.g. the algorithm does not detect if the photo contains full nudity or pornography.
As mentioned later in details, all the detection steps until the moment, does not detect the photos with full nudity and pornography materials which violates the Facebook community standards, and are widely used to attract users to click on malicious links.


Figure 30. Image segmentation example: 'alt="Image may contain: tree, sky, plant, outdoor and nature"'

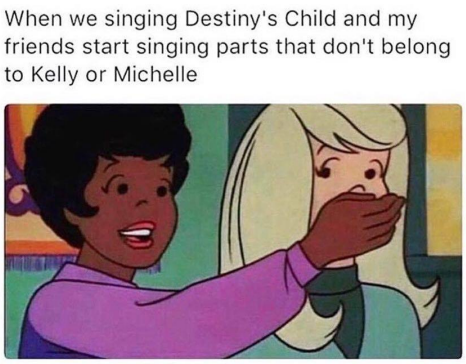
## 3.6 Pornography detection

The Facebook has community standard when it comes to nudity [37]. To by pass being detected when sharing an explicit sexual content, the spammers sometimes show extremely descriptive pictures as gray scale or as a cartoon. The nudity detection scripts available, so far, detects nudity by checking the ratio of 'skin color' in the photo. Having the photo as grey scale or greenish cartoon easily pass the detection, unless it was personally reported.

In many researches, nudity and pornography materials were proven the most attractive bait to attract the spamming victims. There are several online codes that detect nudity, unfortunately, all of them fail to differentiate between the pornographic materials or full nudity and the swimming wear legitimate photos.
Still, the output of these codes along with the rest of indicators, may increase the precision of the spam detection. For example, if the URL posted on an image is malicious and

[37] Facebook Community Standards: Nudity


(a) Image may contain: text and food




(b) Image may contain: meme and text

(c) Image may contain: plant, sky, grass, tree, mountain, cloud, outdoor and nature

(d) Image may contain: dog

Figure 31. Examples for image segmentation. Photos captured from honeypot's Home-Feed

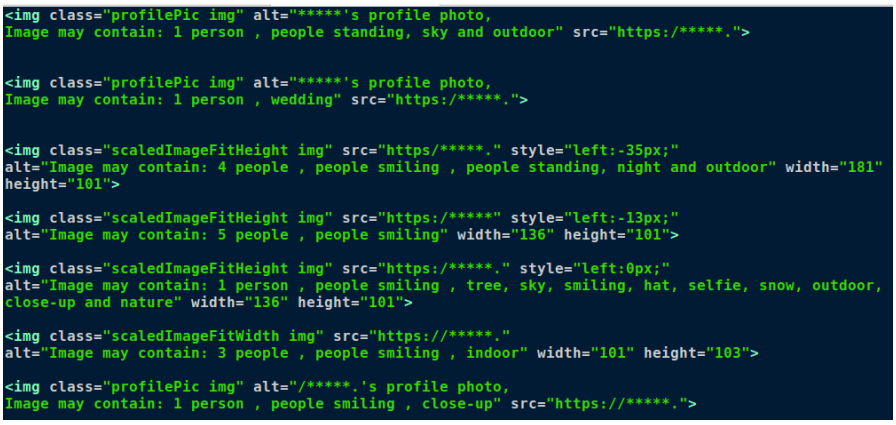



Figure 32. Image May Contain: Examples of photos containing humans

the caption have the word 'click' from one side and contains nudity from another side, it is more likely to be a spam.

Different solutions include :

- Sightengine[38] API can be used to detect nudity. More interestingly, the analyzer can also detect if the photo contains a minor person (babies, children and teenagers under 18). By testing it on the photos extracted from under-aged profiles attracted by the honeypot, it successfully detected with at least 70% confidence that the photos contained nudity and a minor person. Sightengine' service can be integrated using curl command, PHP, python, or Node.js.
- PicPurify[39] could detect the difference between pornographic colored photos and photos with nudity perfectly. Unfortunately, it severely fails to detect nudity or pornography if the photo is in grey scale, or black-and white drawing, or in cartoon-like display. PicPurify's service can be integrated using curl command, python, or Node.js.

There are many other online tools that provide their APIs for ease integration of their service with your code, to either scan your local photos or the online ones but there is always limitation of the maximum free calls you can make per time period. Generally, the code (for curl) looks as follows :

```
curl -X GET -G 'https://api.analyzerserviceprovider.com/' \
     -d 'models=nudity,porn_detection' \
     -d 'UserAPI=abc123' \
     -d 'url=https://xyz.jpg'
```

[38] Sightengine
[39] PicPurify

# 4 Observations

In this section, the author presents the main observations made based on the manual analysis of the approximately 4000 profiles that have added the different honeypots. The author also summarizes the spam changing patterns that took place through the research time. These observations were built on the author's personal monitoring of the profiles and the related traffic through the period December'2016 until October'2017. Luckily, scanning the profiles and documenting their major insights proceed with faster rate than manually extracting every single attribute into the database to further process the information. After building an image of the observations outlined in this section, it was easier to pick the 100 spammer profiles that participated in the first phase of the classifier implementation for better distinguish results as mentioned in section 3.3.

## 4.1 Profile Authenticity

Below is a classification of the profiles based on the ownership of the profile by the person claiming to be.

- A Spammer / Advertiser : Their behavior is quite similar except for the fact that the spammer share malicious contents. They tend to 'market' for their URLs through posts, comments and private messages. More of their behavior is discussed later in the paper.
- A Hisser: As Facebook is continuously improving its spam detection techniques, more suspicious profiles are directed towards inviting 'friends' and their network to chat privately. The higher percentage of this kind of profiles are sharing their Whatsapp numbers on their own profiles and on others' comments that have been posted in public groups or on a friend-of-a-friend profile rather than inviting them to chat over the Facebook messenger. Spam detection over private messaging according to message contents and sender-receiver relation has been covered earlier in several researches [11, 13, 26].
- A Second Profile : From watching this kind of profiles, you can easily notice that this is not the main profile for the person; instead, this profile is made for free navigation through the Facebook pages and adding people away from the notice of the people who already know the person. These people never have their photos on the profiles, most of them do not have a real name set, for example 'lovely bird', and most of their friends are those on the same adult groups and profiles looking like the honeypot, i.e. attractive girls, single, searching for love,....
- A Naive Profile : These are real profiles, adding the honeypot because they are really looking for and believe deeply that they can 'connect' with the girl and

'get married'! They have their real friends, all information and continuously are updating their profile with pictures of them with their families and co-worker at various places. These profiles usually approach the honeypot persistently through posts and private messages even if there is no answer.

Each of the previous types also have distinct messaging and direct-posting behaviors:

- Spammer/Advertiser:
  - Advertise their pages/applications by sending URLs.
  - Sends the malicious URLs.
- Hisser
  - Share their private numbers, mostly Whatsapp numbers.
  - Asks for friends' Whatsapp numbers.
- Second Profile:
  - Asks to send pictures.
  - Sends nude pictures of themselves.
- Naive Profile:
  - Communicate in their native language regardless the fact that the honeypot's language is EN-US.
  - Thanks the honeypot profile for accepting their invitaion then invite her to chat with them and provide their numbers.
  - Gather several of their 'female friends', including the honeypot, in a single thread and message them all 'you are beautiful', or 'I love you', ...

## 4.2 Spamming Behavior

Due to the strong spamming filtration of Facebook, spammers cannot share directly a malicious URL or code on their Facebook profiles. Instead, the URL, always shortened before sharing, placed on Facebook directs to another address, most commonly a blog on Blogger, and from their you are asked to click again on link to 'meet the girl' and the second URL often has a 5 minutes count-down timer, usually accompanied by dummy survey, that is used to skip the scanning of the online anti-viruses, before you finally reach the target URL. Still, the malicious URL can be detected using sandbox. All of the second URLs seen by the honeypot referred to 2 main porn websites through different

profiles and 'girls'.

Two types of malicious posts were noticed:

1. Direct-Content : URLs are shared in the caption of a photo/news/article that is a porn material or contain nudity. And they can be further divided to:
   (a) Those who share these materials on their own profiles, or directly message the selected victims.
   (b) Others whose profiles are very clean and they only share these contents as comments on other pages/groups.
2. Indirect-Content : URLs are shared as caption of religious material! e.g. shared news about the pope from Washington post with an added caption of 'check more details at xxx.com'. The honeypot have also received several posts and private messages addressing her religiously to re-consider her life-style and then at the end of the message a URL of porn website is added.

More points to be noted:

- Once the post/URL is shared, instant virus/malware scanning may detect malicious contents. If the URL is tested couple of days later, usually the scanner does not detect any malicious behavior anymore. This most probably indicates that the owners of the URL, the spammers, monitor the scanners and try to avoid detection continuously.
- Facebook rules mention that "It's against the Facebook Terms to use your personal account to represent something other than yourself" [40]. If the profile is targeted to market a business, it should be converted into a Facebook page instead. In other words, using profiles/fake-profiles to only spread and market a porn website/business should be done through page and not through those profiles seen by the honeypot.

## 4.3 Geographical Distribution

Although the honeypots created were 'living in' Estonia, Latvia, USA, Russia and Poland, as per the information filled in their profiles, it was interesting to note that the thousands who have added the honeypots were located mainly (more than 90% of the profiles) in Russia, India and central and south Africa! These including the spammers and legitimate users who have added the honeypots. Generally speaking, the mentioned country of the profiles attracted to the honeypots almost never overlapped with the honeypot itself!

[40] Facebook Help Center - Converting Your Profile Into a Facebook Page

## 4.4 Sexual Preference

- While female honeypots attracted both female and male profiles to add them, male honeypots received friend requests from male profiles only. All of these male profile explicitly listed their gender as 'male' and their interested-in as 'male'.
- Both straight female and male profiles mostly uploaded female-only pictures, with very few uploading couples-picture as well, but none of them had male-only pictures, even in their profile pictures.
- On the average, suspicious homosexual male profiles have the longest profile ownership duration compared to suspicious straight male profiles, and all female profiles. For example, while most of the harvested non-homosexual male profiles joined only in mid-2017 and later, almost half the homosexual male ones dated to beginning of 2015 and even before, with extensive number of fully-nude and extremely pornographic photos on these profiles. It is to be noted here that one major reason may be the number of 'report profile' each profile receives.
- It is very common to find several homosexual male profiles with different names, sharing same profile picture, most probably a grey-scale pornographic photo.

## 4.5 Child Abuse

According to United Nations Convention on the Rights of the Child, and with 192 countries' ratification, the child is a human being below the age of 18 years. Facebook does not allow children less than 13 years old to create a profile on Facebook. Nevertheless, without reporting the presence of a 'minor' profile, there is no valid method to confirm that the entered age during profile registration is the right one. This being said, it was frustrating to detect clear signs of child abuse:

- Although primary students are maximum 11 or 12 years old, there are a lot of primary students on the friend-list of the profiles marked as 'social spammers'.
- Children are sharing their Whatsapp numbers with strangers upon request! Most of these requests are explicitly sexual and are inviting to talk privately on Whatsapp.
- On their profiles, these children are sharing nude photos for themselves and extremely pornographic materials in black and white photos, that are detected, after further investigation, on other adult friends on their friend-list.
- Three of the extremely most sexual profiles had more than 95% of their friends underaged girls.

## 4.6 Profile's Personal Information

While on the other OSNs other researchers have documented the effort made in the personal information and 'About Me' sections which include location, age, education, interests, quotes, information, and more about the profile's owner, the author noticed that no real interest of spammers to fill in these information on Facebook, instead it is often left 80% empty. Nevertheless, all the spammers (except one profile) have listed their 'martial status' as single.
Few more observations are :

- If occupation is either 'at Facebook', 'Self', or 'at home'. Interesting to question why they did not leave it blank as the rest of their information.
- If education is mentioned, it is usually similar to 'Sex university'.
- If there are few words in the 'Intro to me' section, it definitely contains one of the words {love, sex, lust}.

## 4.7 Spam Pattern Change

During the the research and implementation time period of this research, over the years 2016 and 2017, it was clearly noticed the change in the spread of malicious URLs. Early in 2016, it was easy to detect URLs that instantly direct to malicious sites upon clicking. Later, most of the URLs were shortened URLs that directs to free-porn websites or blogs, and from there you can find photo traps that invite you to click on them to 'meet the girl' and they direct you to malicious site or directly download an exe file. In the second half of 2017, it was rare to find any of these two previous kinds of links. Number of malicious (or pointing to malicious) links has dropped, and if ever found, it directs you to what looks like a legitimate online store for clothes, which also have photo traps.

## 4.8 Photos' Variety

As presented in section 3.5 there is a huge variety of non-nude pictures' categories extracted from the spammers' profiles, including persons, animals, memes, food and more. Other than the pornographic photos, it is hard to define another photo structure or object linked to the spamming behavior.

## 4.9 Summary

In this section, the author has described the observations obtained from an 11-months period of watching the traffic on Facebook platform through the eyes of the implemented social honeypots, and has highlighted the evolution of spamming techniques through the same time period.

# 5 The Classifier

This section views the overall framework of the suggested classifier, describes details of each node and the tunning target, and summarizes the final implemented classifier.

## 5.1 The Proposed Classifier

The overall framework for the classifier is summarized in Fig. 33 and each phase is discussed below. Data collection was discussed earlier in section 3.3.

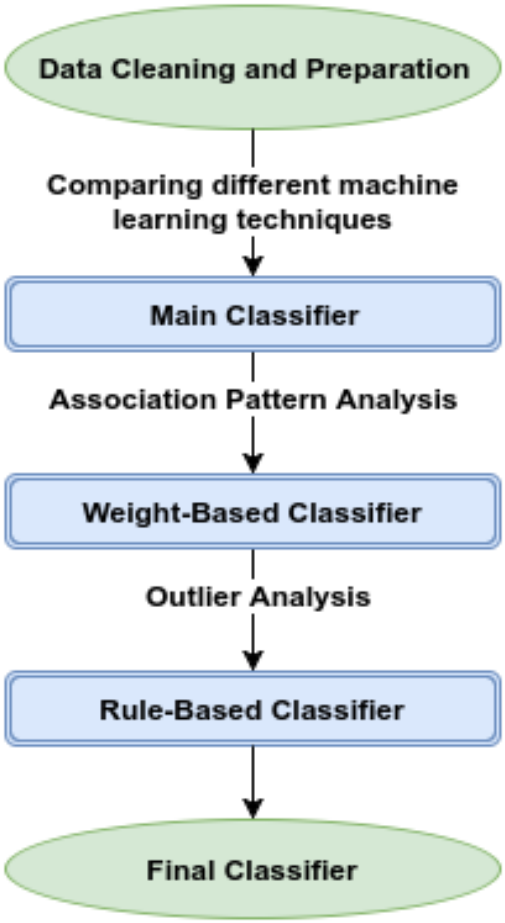



Figure 33. Overall Framework of the Classifier

## 5.2 Data Cleaning and Preparation

In this phase, the data should be prepared to be tested, this will include:

- handling missing entries, e.g. profiles who do not have marital status. As inspired by the paper [21], the missing values were completed using homophily theory [41]. "The similarity-attraction hypothesis [36] predicts that people are more likely to interact with those with whom they share similar traits. The theory of self-categorization proposes that people tend to self-categorize themselves and others in terms of race, gender, age, education, etc., and that they use these categories to further differentiate between similar and dissimilar others[37]. In addition, because interpersonal similarity increases predictability of behavior and reduces communication apprehension [38], so communications among similar others are more likely to occur." [39].

  Following the hypothesis, missing data was completed. For example

  - If the sexual preference is not clearly mentioned, and friend list is mostly gays + photos are similar to the ones shared by other gays profiles –> The profile is marked as Homosexual male.
  - If the profile does not clearly state the age, but most friends are less than 18, + similar shares + similar study conversation + targeted by same spammer –> The profile is marked under 18-years old.

- describing photos in words, i.e. expressing the photos' contents, e.g. natural scene, nude, grey-scale,....

- changing discrete-valued attributes into numbers to decrease the overhead of implementing text-mining steps for easier classification, e.g. 1 for single, 2 for married, ...

## 5.3 Main Classifier

Different supervised machine learning algorithms were tested and results were documented to compare the ones with the highest accurate results. Section 6.2 presents the top successful algorithms examined and their relevant output. The chosen algorithms are listed below with the necessary R-package for each and a pseudo-code :

- KNN

```
for(i in 1:k_valid) {

  N <- trainSize_Now

  for(idx in 1:N) {
    euc_dist[idx]< sqrt(sum((x_train[idx,1:8]    x_valid[i
        ,1:8])^2) )
```

[41] Wikipedia - Homophily

```
    }

    sorted_dist_struct < sort(euc_dist,  index.return = TRUE)
    sorted_dist < sorted_dist_struct[[1]]
    sorted_pos < sorted_dist_struct[[2]]

    knearestneighbors=sorted_pos[1:k]
    knearestdistances=sorted_dist[1:k];

    for (j in 1:k) {
      A[j,] = x_train[knearestneighbors[j],];
      M[i,j] = knearestneighbors[j];
    }

    domColor = mode(A[,9])
    check[i] <  domColor



    if (x_valid[i,9] != domColor)
    { error_calc=error_calc+1 }

  }
```

- Classification Tree

```
| library(rpart)
| fit <  rpart(x_train[,12] ~ ., data = x,method="class")
```

- Naïve Bayes

```
| library(e1071)
| fit < naiveBayes(as.factor(x_train[,9]) ~ ., data = x )
```

- Random Forests

```
| library(randomForest)
| fit < randomForest(as.factor(x_train[,9]) ~ ., x, ntree=50)
```

- SVM

```
| library(e1071)
| fit < svm(x_train[,9] ~ ., data = x,type='C classification')
```

- J48

```
| library(RWeka)
| fit  < J48(as.factor(x_train[,9]) ~ ., data = x )
```

## 5.4 Weight-Based Classifier

Some noticeable association of attributes are used to increase the weight of an entry as a spammer. Below is a list of several attributes' values when associated together the profile is most probably a spammer or a fake profile.

- Works as : self-employed, or, works-at-Facebook.
- Marital status : Single.
- Very short profile life-time, i.e. profile was created less than 5 months ago.
- Intro/Bio : searching for love.
- profile has some keywords, e.g. sins, sensation, hot, ...
- Number of mutual friends with the honeypot is 0 or 1. Each of these cases will be given different weight, e.g. if 0 mutual friend, added weight is 4, and if 1 mutual friend, added weight is 2. Else, no weight will be added.

The result from the previous node is checked, and if the the profile was marked 'legitimate', the classier double checks the profile's attributes, if they matched in 4 out of the 6 aboves values, the profile is marked as a 'spammer'!

## 5.5 Rule-Based Classifier

If outliers were detected, the tunning of the classifier would have been condition-based, i.e. if xxx then 'mark as spammer', or if yyy then 'mark as legitimate' and the condition will over-write the previous result for this specific entry. Unfortunately, there were no exact conditions to add that further decreases the wrongly classified data-points without mis-classifying other points.

## 5.6 Final Classifier

Six different supervised machine learning algorithms were tested and tuned as displayed in the earlier subsections and results were documented to compare the ones with the highest accurate results. The classifier was able to distinguish between the spammers and legitimate profiles but it could not precisely distinguish between different spammers' kinds listed in section 4.1. Section 6.2 illustrates the classifiers' results.

# 6 Results

The final classifier was able to distinguish between the spammers and legitimate profiles but it could not precisely distinguish between different spammers' kinds listed in section 4.1. As mentioned earlier in section 1.4.3, 10-fold cross-validation is used to minimize the inaccuracy resulted from the small data-size. The performance will be presented in terms of four derivations of the confusion matrix (also known as error matrix), **Recall**, **Precision**, **F1-Score** and **Accuracy**. Each terminology is explained below before highlighting the average performance of each classification algorithm. The result of each algorithm will be calculated as the average of results from 'searching for legitimate profiles' and 'searching for 'spammer profile'. More details will be displayed in below, in section 6.1.

## 6.1 Performance Measurements

### 6.1.1 Confusion Matrix

Confusion matrix visualizes the results beyond general accuracy measurement, by summarizing the output in 4 terms, 'True Positive', 'True Negative', 'False Positive' and 'False Negative'. While a general accuracy of 'greater than 90%' may sound good enough, it is important to define the error more. For example, if the data set consists of 100 data points, out of which 90 belongs to class '1' and 10 to class '0', accuracy of 90% looks extremely bad if the 10 errors are in classifying the '0' class, i.e. the model outputs 100% of the '0' class wrong, and does not distinguish 2 different classes. Fig. 34 shows the confusion matrix used to drive the performance values in this section.

| | | True Class | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted | Positive | True Positive (T.P.) | False Positive (F.P.) |
| | Negative | False Negative (F.N) | True Negative (T.N) |

Figure 34. Confusion Matrix Construction

The matrix is mapped below in Fig. 35 to reflect the relevant classes for the problem in hand.

| | | Observed Class | |
|---|---|---|---|
| | | Legitimate | Spammer |
| Predicted | Legitimate | True Legitimate (T.P.) | False Legitimate (F.P.) |
| | Spammer | False Spammer (F.N) | True Spammer (T.N) |

Figure 35. Confusion Matrix - Classes

### 6.1.2 Recall / Sensitivity

The probability that a randomly selected relevant document is retrieved in a search.

$$\text{Recall} = \frac{\sum T.P.}{\sum (T.P. + F.N)}$$

### 6.1.3 Precision

The probability that a randomly selected retrieved document is relevant.

$$\text{Precision} = \frac{\sum T.P.}{\sum (T.P. + F.P)}$$

### 6.1.4 F1-Score

The harmonic mean [42] of precision and recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precision x Recall}}{\text{Precision + Recall}}$$

### 6.1.5 Accuracy

$$\text{Accuracy} = \frac{\text{T.P. + T.N.}}{\text{T.P. + F.P. + T.N + F.N.}}$$

[42] Wikipedia - Harmonic Mean
[43] Wikipedia - Precision and Recall

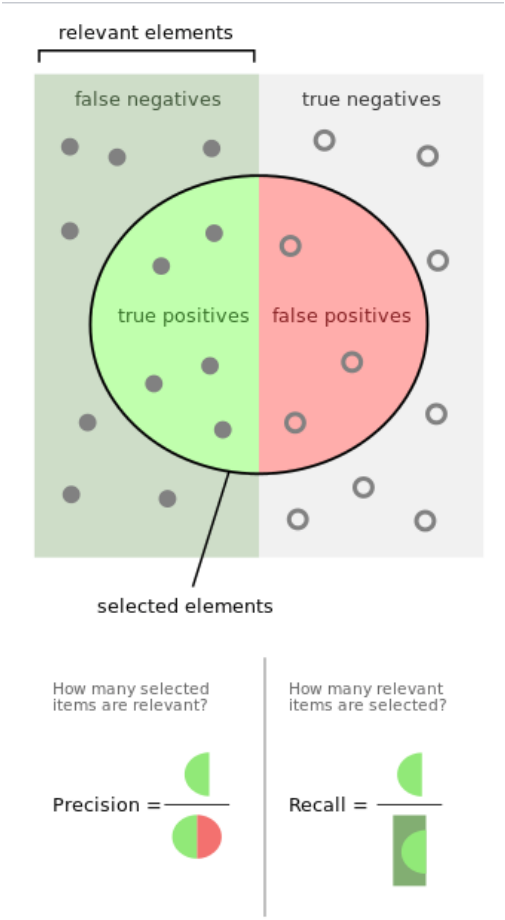



Figure 36. Precision and Recall [43]

## 6.2 Classifiers Evaluation

### 6.2.1 KNN-Classifier

K-nearest neighbor, KNN, classification is considered one of the simplest machine learning techniques, because it depends on Lazy Learning method which evaluates the classification locally for each point in the data set, and so, is usually slower compared to other techniques. KNN function uses distance between K-points in the training set (where K is an integer larger than zero) and the point-to-evaluate to classify the validation/testing sets. There are several metric functions to use in KNN algorithm, and the best function is usually selected depending on the application, e.g. cosine distance is used when comparing documents or vectors, or the best result during the training period, i.e. function outputting least error. The tunning and choosing of the metric function and the value of K should be wisely decided to avoid over-fitting of the model. For the

current classification problem, euclidean distance with k=2 reflected the best average results. Table. 1 shows the average performance of the KNN model at K=2,3,4. As K-value increases further, the performance drops further.

Table 1. Performance of the KNN model

| K-value / Class | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| K=2 (Legitimate) | 0.9278 | 1 | 0.8421 | 0.9143 |
| K=2 (Spammer) | 0.9278 | 0.88 | 1 | 0.9936 |
| K=2 (Average) | 0.9278 | 0.94 | 0.9211 | 0.9540 |
| K=3,4 | 0.9024 | 0.9412 | 0.8421 | 0.8889 |

### 6.2.2 Classification Tree

Decision Trees depend on recursively partitioning the data set to predict the dependent variable. For discrete, finite dependent variables, *Classification trees* are used, while in case of continuous dependent variables, *Regression trees* are used [40].

Table 2. Performance of the Classification Tree model

| Class | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Legitimate | 0.9630 | 1 | 0.9091 | 0.9524 |
| Spammer | 0.9630 | 0.9412 | 1 | 0.9670 |
| Average | 0.9630 | 0.9706 | 0.9546 | 0.9597 |

### 6.2.3 Naïve Bayes

Naïve Bayes algorithm is based on the Bayes' theorem that assumes strong independence of features from each other:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Table 3. Performance of the Naïve Bayes model

| Class | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Legitimate | 0.9630 | 0.9167 | 1 | 0.9565 |
| Spammer | 0.9630 | 1 | 0.9375 | 0.9677 |
| Average | 0.9630 | 0.9584 | 0.9688 | 0.9621 |

### 6.2.4 J48

Decision tree J48 is on of the algorithms developed by the WEKA[44] project team.

Table 4. Performance of the J48 model

| Class | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Legitimate | 0.9630 | 1 | 0.9090 | 0.9524 |
| Spammer | 0.9630 | 0.9412 | 1 | 0.9670 |
| Average | 0.9630 | 0.9706 | 0.9545 | 0.9597 |

### 6.2.5 Support Vector Machine (SVM)

A supervised machine learning technique that solves the classification problem by escalating the input to higher dimensionality space (hyper plane) where separation of data points and classification will be easier.

Table 5. Performance of the SVM model

| Class | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Legitimate | 0.9630 | 1 | 0.9091 | 0.9524 |
| Spammer | 0.9630 | 0.9412 | 1 | 0.9670 |
| Average | 0.9630 | 0.9706 | 0.9546 | 0.9597 |

### 6.2.6 Random Forests

The technique depends on constructing multiple, de-correlated, decision trees (classification trees in this case) and outputting the average [41]. One of the most important advantages of random forests technique is that increasing the number of trees, which further increases the accuracy of the result, does not cause an over-fitting for the model that cause the model to memorize the data rather than learn.

Fig. 37 shows the resultant error percentage as the number of trees differs. At 16 and 25 trees the error drops to 0.00%, while at 50 trees, the error is 3.7%.
Another, less optimistic, way to measure the error in this model will be to consider the confusion matrix that is based on the OOB data. Out-of-bag (OOB) error use the bootstrap aggregating, also known as bagging, to improve the accuracy and stability of the model. Although it usually gives a higher error percentage than when actually tested against the validation data, it outputs a more realistic results on which the model can be better planned .

[44]WEKA

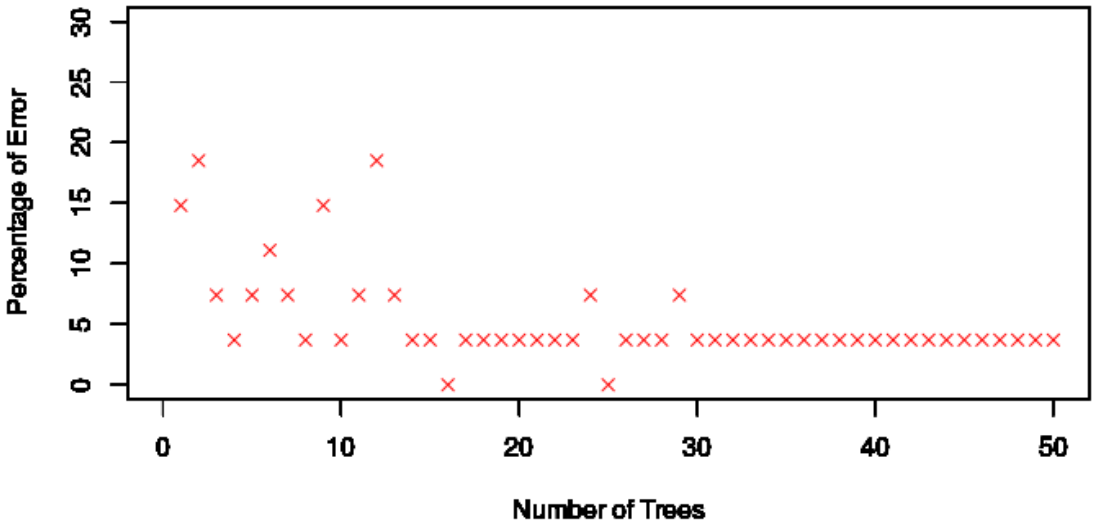



Figure 37. Random Forests: Number of trees vs. error

Considering the OOB error, 18 trees shows an error of 19.2% in 70% of the cases and up until 23.1% in the rest. Less number of trees shows a huge variation in the resultant error, that exceeds 40% in some testing iterations. Nevertheless, As number of trees increases over 18, there is no observed massive improvement in the error percentage. Results driven using 50 to 500 trees are the same.
Table. 6 illustrates the average computations of the model with 18-trees v.s. 50-trees driven from 500 iteration. When implementing the classifier on large-scale, it will be a challenge to decide the social and statistical significances of the results' differences and whether they worth increasing the complexity of the model or not.

Table 6. Performance of the Random Forests model

| # of Trees / Class | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| 18-trees (Legitimate) | 0.8450 | 0.7794 | 0.8766 | 0.8241 |
| 18-trees (Spammer) | 0.8450 | 0.9412 | 1 | 0.9670 |
| 18-trees (Average) | 0.8450 | 0.8603 | 0.9383 | 0.8956 |
| 50-trees (Legitimate) | 0.8573 | 0.7754 | 0.9057 | 0.8349 |
| 50-trees (Spammer) | 0.8573 | 0.9412 | 1 | 0.9670 |
| 50-trees (Average) | 0.8573 | 0.8585 | 0.9529 | 0.9010 |

### 6.2.7 Performance Summary

Table 7. Performance summary of the different classification models

| Classifier | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| **KNN** | 0.9278 | 0.94 | 0.9211 | 0.9540 |
| **Classification Tree** | 0.9630 | 0.9706 | 0.9546 | 0.9597 |
| **SVM** | 0.9630 | 0.9706 | 0.9546 | 0.9597 |
| **Naïve Bayes** | 0.9630 | 0.9584 | 0.9688 | 0.9621 |
| **Random Forests** | 0.8450 | 0.8585 | 0.9529 | 0.9010 |
| **J48** | 0.9630 | 0.9706 | 0.9545 | 0.9597 |

## 6.3 Validation

In order to validate the results, the author compared the classifier's performance against available results from past researches to visualize how far the results lie within an accepted performance range among the previous published researches. The previous results are chosen from some of the most recent research papers:

- Fig. 38a summarizes the results of spam detection in the Facebook posts available in the open public group named as "World of Taylor Swift" on Facebook in the period: July 2015 to August 2015 [35].
- Fig. 38b shows the results driven from analyzing posts on closed Facebook groups using Random Forests algorithm [27].
- Fig. 38c summarizes the results driven from analyzing the profile's public features [25].

| Classifier | Accuracy | Precision | Recall | FM |
|---|---|---|---|---|
| Random Forest | 0.977 | 0.928 | 0.844 | 0.884 |
| Bagging | 0.967 | 0.822 | 0.875 | 0.848 |
| J48 | 0.96 | 0.793 | 0.828 | 0.810 |
| Random Tree | 0.949 | 0.760 | 0.745 | 0.753 |
| Logistic | 0.924 | 0.592 | 0.870 | 0.705 |

(a) Facebook public groups posts [35]

| | Precision | Recall | F1-score |
|---|---|---|---|
| **Spam** | 99.26% | 97.06% | 98.15% |
| **Non-spam** | 97.12% | 99.18% | 98.14% |
| **Avg/Total** | 98.19% | 98.12% | 98.15% |

(b) Facebook closed groups posts [27]

| Classifier | Accuracy + stdv | Precision + stdv | Recall + stdv | F_measure + stdv | AUC + stdv |
|---|---|---|---|---|---|
| J48 | 94.9 ±7.91 | 0.95 ±0.11 | 0.94 ±0.13 | 0.93 ±0.10 | 0.95 ±0.08 |
| K-NN=1 | 94.1 ±8.83 | 0.95 ±0.12 | 0.91 ±0.17 | 0.92 ±0.13 | 0.98 ±0.06 |
| K-NN=3 | 94.0 ±8.58 | 0.95 ±0.12 | 0.91 ±0.15 | 0.92 ±0.12 | 0.97 ±0.07 |
| K-NN=5 | 93.0 ±8.83 | 0.94 ±0.12 | 0.88 ±0.17 | 0.90 ±0.12 | 0.96 ±0.07 |
| MLP | 93.6 ±8.29 | 0.95 ±0.11 | 0.90 ±0.17 | 0.91 ±0.12 | 0.97 ±0.07 |
| NB | 95.7 ±7.18 | 0.94 ±0.11 | 0.96 ±0.11 | 0.95 ±0.09 | 0.98 ±0.05 |

(c) Based on profile public features [25]

Figure 38. Performance results from recent researches.

Table 8 compares the results from the honeypots against the ones driven from others.

Table 8. Summary of results driven from different attributes: Honeypots (implemented in this research), PGP (Public Group Posts [35]), CGP (Closed Group Posts [27]), and PPF(Profiles Public Feature [25])

| Classifier | Honeypots | PGP | CGP | PPF |
|---|---|---|---|---|
| **Bagging** | | 0.848 | | |
| **Classification Tree** | 0.9597 | | | |
| **J48** | 0.9597 | 0.810 | | 0.93 |
| **KNN** | 0.9540 | | | 0.92 |
| **Logistic** | | 0.705 | | |
| **MLP** | | | | 0.91 |
| **Naïve Bayes** | 0.9621 | | | 0.95 |
| **Random Forests** | 0.9010 | 0.884 | 0.9815 | |
| **Random Tree** | | 0.753 | | |
| **SVM** | 0.9597 | | | |

# 7 Conclusion

Different researches have hunted spammers from different prospectives and using different data and algorithms, but most commonly, the objective spammers were those who use the social network to spread malicious contents like malwares. This research has pursued a wider range of social spammers, including those who use the social platforms to spread unapproved contents, harass other users, blackmail others for revenge, or start a child pornography chain.

The results summarized in section 6.2 and validated in section 6.3 show that different algorithms have resulted in F1-score value over 0.9 (most algorithms have output values exceeding 0.95) which is compared to the other researches, lie on the top best results, and is a promising start to tune the classifier even more for further better results. Still it is needed to highlight that the spam pattern is continuously changing, and there is no one technique that is always best detecting spam without any improvements.

On a different thought, it is complicated to evaluate how far the honeypot attracts only spammers on Facebook platform. Reviewing the observations, it is obvious that the most interested profiles are for people who are looking to start relationships or to browse groups they are uncomfortable browsing publicly in front of their friends through their main profiles. On one hand, this majority is a victim for the real spammers who aim at spreading malicious harmful contents, but on the other hand, they are still considered 'social spammers' to the rest of the Facebook community and are considered 'creepers' by the Facebook team, as mentioned earlier in section 2.2.2.

The contribution of this work can be summarized as below:

- Illustrates the obstacles that inhibit the stability of the social honeypots on the Facebook platform compared to another OSNs.
- Highlights the different spammers kinds trapped by the social honeypot and the hazardousness of each.
- Presents a classifier implementation that distinguishes between legitimate Facebook profiles and different social spamming profiles.
- Demonstrates characteristics and social attitudes of spammers on the Facebook platform.
- Compares between the results from extracted from Facebook platform and those from Twitter, and features the different spamming behavior on the different platforms.

Overall, the research answers in details the research questions mentioned in section 1.2, and below are brief answers to those questions.

- Social honeypot implementation on Facebook platform is doable, with more challenges than for those implemented on the other OSNs, because of the higher security scan on Facebook platform that detects fraud profiles.
- The observations driven from the honeypots indicate their success to attract social spammers only.
- The results illustrated in section. 6 suggests a promising classification of spammers using the built classifier.
- The implemented classifier does not target only the spammers through text mining to detect the ones spreading malicious contents, it also targets all kind of social spammers who cause social inconvenience on the social networks.
- Finally, as shown in details in sections 2.1 and 6.2.7, the results of the implemented classifier lie in the range of the results driven from Twitter platform, and mostly exceed an F1-score of 0.95 .

## Future Work

The work presented in this thesis has revealed different new challenges and more questions to address. This section is based on and extends unpublished work and research done by the author to increase the robustness and applicability of the end classifier.

### Create a PHP Application

In this research, bash scripting was used as a convenient main tool to integrate the preparation steps before the classifiers implementation in R-Language, including APIs calling, image inspection and URLs scan. Nevertheless, PHP will also ease the integration of the manual extraction of data from the Facebook graph API as well as provide an environment to perform automatically the rest of the steps[45]. It will be left later to decide if the classifier as well will be implemented in PHP, or the application will integrate the R code in the back-end. The main importance of the application is that applications are given more freedom to use the Facebook graph API, after the approval of the application's user. Fig. 39 visualizes the ease to extract information through the Graph API. Information provided are the 'about me' section, full name, education, birthday, gender, hometown, last checked-in location and work (including employer, position, projects, colleagues, and start and end dates.

[45]Facebook SDK v5 for PHP


Figure 39. Information extracted for one of the legitimate users using FacebookGraph API

### Define Spammers Closely

The classifier should be improved and tuned to output the exact kind of spammer detected as per the classification observed in section 4.1 because every kind of spammers needs to be dealt with differently. For example, a spammer who spread a malware or threatens a child needs to be stopped immediately and reported, while creeper who adds everyone and send them man messages needs to be educated about how to use the platform to be convenient place for all.

# References


[1] Lita Van Wel and Lambèr Royakkers. Ethical issues in web data mining. *Ethics and Information Technology*, 6(2):129–140, 2004.

[2] Michal Kosinski, Sandra C Matz, Samuel D Gosling, Vesselin Popov, and David Stillwell. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543, 2015.

[3] Yuval Elovici, Michael Fire, Amir Herzberg, and Haya Shulman. Ethical considerations when employing fake identities in online social networks for research. *Science and engineering ethics*, 20(4):1027–1043, 2014.

[4] Robert Edmund Wilson and Sam Gosling. A review of facebook research in the social sciences. 2016.

[5] Selami Aydin. A review of research on facebook as an educational environment. *Educational Technology research and development*, 60(6):1093–1106, 2012.

[6] Stanley Wasserman. *Advances in social network analysis: Research in the social and behavioral sciences*. Sage, 1994.

[7] Philippe Boillat and Morten Kjaerum. Handbook on european data protection law. *European Union Agency for Fundamental Rights.–2014.–199 p*, 2014.

[8] Steve Webb, James Caverlee, and Calton Pu. Social honeypots: Making friends with a spammer near you. In *CEAS*, 2008.

[9] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2010.

[10] James Caverlee and Steve Webb. A large-scale study of myspace: Observations and implications for online social networks. In *ICWSM*, 2008.

[11] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 1–9. ACM, 2010.

[12] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[13] Jonghyuk Song, Sangho Lee, and Jong Kim. Spam filtering in twitter using sender-receiver relationship. In *Recent Advances in Intrusion Detection*, pages 301–317. Springer, 2011.

[14] De Wang, Danesh Irani, and Calton Pu. A social-spam detection framework. In *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, pages 46–54. ACM, 2011.

[15] Xia Hu, Jiliang Tang, Yanchao Zhang, and Huan Liu. Social spammer detection in microblogging. In *IJCAI*, volume 13, pages 2633–2639. Citeseer, 2013.

[16] Xia Hu, Jiliang Tang, and Huan Liu. Online social spammer detection. In *AAAI*, pages 59–65, 2014.

[17] Juan Martinez-Romo and Lourdes Araujo. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 40(8):2992–3000, 2013.

[18] Xin Ruan, Zhenyu Wu, Haining Wang, and Sushil Jajodia. Profiling online social behaviors for compromised account detection. *Information Forensics and Security, IEEE Transactions on*, 11(1):176–187, 2016.

[19] Enhua Tan, Lei Guo, Songqing Chen, Xiaodong Zhang, and Yihong Zhao. Unik: Unsupervised social network spam detection. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 479–488. ACM, 2013.

[20] Liang Wu, Xia Hu, Fred Morstatter, and Huan Liu. Adaptive spammer detection with sparse group modeling. In *ICWSM*, pages 319–326, 2017.

[21] Hua Shen, Fenglong Ma, Xianchao Zhang, Linlin Zong, Xinyue Liu, and Wenxin Liang. Discovering social spammers from multiple views. *Neurocomputing*, 225:49–57, 2017.

[22] Mahdi Washha, Aziz Qaroush, Manel Mezghani, and Florence Sedes. A topic-based hidden markov model for real-time spam tweets filtering. *Procedia Computer Science*, 112:833–843, 2017.

[23] Phuc Tri Nguyen and Hideaki Takeda. Online learning for social spammer detection on twitter. *arXiv preprint arXiv:1605.04374*, 2016.

[24] Thi-Hong Vuong, Van-Hien Tran, Minh-Duc Nguyen, Thanh-Huyen Pham, Mai-Vu Tran, et al. Social-spam profile detection based on content classification and user behavior. In *Knowledge and Systems Engineering (KSE), 2016 Eighth International Conference on*, pages 264–267. IEEE, 2016.

[25] Al-Zoubi Ala'M, Hossam Paris, et al. Spam profile detection in social networks based on public features. In *Information and Communication Systems (ICICS), 2017 8th International Conference on*, pages 130–135. IEEE, 2017.

[26] Fangzhao Wu, Jinyun Shu, Yongfeng Huang, and Zhigang Yuan. Co-detecting social spammers and spam messages in microblogging via exploiting social contexts. *Neurocomputing*, 201:51–65, 2016.

[27] Nattanan Watcharenwong and Kanda Saikaew. Spam detection for closed facebook groups. In *Computer Science and Software Engineering (JCSSE), 2017 14th International Joint Conference on*, pages 1–6. IEEE, 2017.

[28] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015.

[29] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *European Conference on Computer Vision*, pages 75–91. Springer, 2016.

[30] Sergey Zagoruyko, Adam Lerer, Tsung-Yi Lin, Pedro O Pinheiro, Sam Gross, Soumith Chintala, and Piotr Dollár. A multipath network for object detection. *arXiv preprint arXiv:1604.02135*, 2016.

[31] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *CVPR*, volume 1, page 2, 2017.

[32] Tao Stein, Erdong Chen, and Karan Mangla. Facebook immune system. In *Proceedings of the 4th Workshop on Social Network Systems*, page 8. ACM, 2011.

[33] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[34] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[35] Hailu Xu, Weiqing Sun, and Ahmad Javaid. Efficient spam detection across online social networks. In *Big Data Analysis (ICBDA), 2016 IEEE International Conference on*, pages 1–6. IEEE, 2016.

[36] Donn Erwin Byrne. *The attraction paradigm*, volume 11. Academic Pr, 1971.

[37] John C Turner, Michael A Hogg, Penelope J Oakes, Stephen D Reicher, and Margaret S Wetherell. *Rediscovering the social group: A self-categorization theory.* Basil Blackwell, 1987.

[38] Herminia Ibarra. Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative science quarterly*, pages 422–447, 1992.

[39] Y. Connie Yuan and Geri Gay. Homophily of network ties and bonding and bridging social capital in computer-mediated distributed teams. *Journal of Computer-Mediated Communication*, 11(4):1062–1084, 2006.

[40] Leland Wilkinson. Classification and regression trees. *Systat*, 11:35–56, 2004.

[41] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[42] Marzelas Panagiotis. A social media honeypot to detect spearphishing (master dissertation). Tallinn University of Technology, 2016.

# Appendix

# I. Terms and Notations

### K-fold Cross-Validation

The original sample is divided into k equal sized subsamples. The classification techniques then takes place K-times where each sample get to be the "Validation Sample" once and the other K-1 samples are used as Training Set. This gives the advantage to test the robustness and the accuracy of the technique independent of the training/validation sets' structures.

### Porter Stemmer

Stemming is the process of reducing the words to their word stem, base or root form. For example, "fishing", "fished", and "fisher" will be reduced to the root word, "fish". Porter stemmer, named after Dr. Martin Porter, is widely used as the standard algorithm used for English stemming.

### Bigrams Model

Bigrams algorithm depends on calculating the occurrence possibility of a word given the preceding word.

$$P(W_n|W_{n-1}) = \frac{P(W_{n-1},W_n)}{W_{n-1}}$$

### Supervised Learning

The learning algorithms that considers both the inputs and outputs. The model keeps tunning for the best match to the output. On the contrary, unsupervised learning takes in the inputs only and best fit the input according to its algorithms with no ground truth information of the required output.

### Huffman code

Huffman code is used for lossless data compression, because it assigns fewer number of bits to encode the more frequent characters and vice-versa.

# II. Abbreviations

| Abbreviation | Phrase / Meaning |
|---|---|
| OSN | Online Social Networks |
| URL | Uniform Resource Locator |
| exe | file extension for an executable file format |
| API | Application Programming Interface |
| OOB | Out-of-Bag |
| SDK | Software Development Kit |
| SVM | Support Vector Machine |
| BSD | Berkeley Software Distribution |
| A.I. | Artificial Intelligence |

# III. Acknowledgments

The author would like to thank *Heiki Pikker* for his help and tips in ways to attract spammers to the honeypots and usage of sandbox to detect malicious links. Also, want to thank *Panagiotis Marzelas* for his help with implementing the honeypot itself and the guidance provided in his master thesis [42].

# IV. License

## Non-exclusive License to reproduce thesis and make thesis public

I, **Ghada Zakaria**,

1. herewith grant the University of Tartu a free permit (non-exclusive License) to:

   1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

   of my thesis

   **Detecting Social Spamming on Facebook Platform**

   supervised by Innar Liiv and Raimundas Matulevičius

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 10.01.2018