SPAMMER DETECTION AND FAKE USER IDENTIFICATION ON SOCIAL NETWORKS

A.DURGA DEVI, KORADA TEJASWI

1. PG STUDENT, D.N.R. COLLEGE, P.G. COURSES (AUTONOMOUS), BHIMAVARAM-534202.

Email id:- koradatejaswi10@gmail.com

2. Assistant Professor in DEPARTMENT OF MASTER OF COMPUTER SCIENCE, BHIMAVARAM-534202.

Email id:- adurgadevi760@gmail.com

ABSTRACT

Online Social Networks (OSNs) are great environments for sharing ideas, following news, advertising products etc., and they have been widely used by many in the world. Although these are the advantages of social networks, it is difficult to understand whether an account in socialmedia platform such as Instagram, Twitter, Facebook really belongs to a person or organization. Through creating fake and malicious accounts, unwanted content can spread over the social network. Therefore, the prediction of fake accounts is an important problem. In this study, we applied machine learning algorithms to this problem and we evaluated performances of different activation functions. According to the experimental results, use of machine learning algorithmsin detecting fake accounts yielded successful results. The use of various activation functions indifferent layers on the ANN significantly affects the results. In the literature, other classification methods have been widely used for detecting fake accounts and spammers on online social Network. To the best of our knowledge, there is no brief study that classifies fake accounts using ANNs with different activation functions..

1. INTRODUCTION

1.1 PROBLEM DEFINITION

Malicious users produce fake profiles to phish login info from unsuspecting users. A fake profile can send friend requests to several users with public profiles. These counterfeit profiles bait unsuspecting users with photos of individuals and they may misuse them for various use. Once the user accepts the request, the owner of the phony profile can spam friend requests to anyone this user could be a friend.

The fake profile's contents usually have links that result in malicious external web site where there is an attack of virus to the system and when unaware curious user clicks the dangerous link can result in crashing of their systems. The effect of this may be dangerous as putting in a root kit turning the pc into a zombie. Whereas Face book

contains a rigorous screening to stay this fake accounts out, it solely takes one fake profile to wreck the computers of the many. Hence, we came up with a solution by using machine learning algorithms which gave successful results.

1.2 PROJECT PURPOSE

It has become quite unpretentious to obtain any kind of information from any source across the world by using the Internet. Huge volumes of data available on these sites also draw the attention of fake users. Twitter has rapidly become an online source for acquiring real-time information about users, Twitter is an Online Social Network (OSN) where users can share anything and everything, such as news, opinions, and even their moods. Several arguments can be held over different topics, such as politics, current affairs, and important events. When a user tweets something, it is instantly conveyed to his/her followers, allowing them to outspread the received information at a much broader level [2]. With the, evolution of OSNs, the need to study and analyze users' behaviors in online social platforms has intensity. Many people who do not have much information regarding the OSNs can easily be tricked by the fraudsters. There is also a demand to combat and place a control on the people who use OSNs only for advertisements and thus spam other people's accounts. Recently, the detection of spam in social networking sites attracted the attention of researchers. Spam detection is a difficult task in maintaining the security of social networks.

2. LITERATURE SURVEY AND RELATED WORK

2.1 INRODUCTION

Literature survey is the most important step in software development process. Before developing the tool, it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language used for developing the tool. Once the programmers start building the tool, the programmers need lot of external support. This support obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into for developing the proposed system.

2.2 RELATED WORK

Sybil rank was designed in late 2012, to with efficiency establish faux profiles through a ranking graph-based system. The algorithmic rule uses a seed choice technique combined with early terminated random walks to propagate trust. Its machine value is measured in O(nlogn). Profile’s area unit graded consistent with the number of interactions, tags, wall posts, and friends over time. Profiles that have a high rank area unit thought of to be real

with faux profiles having a coffee rank within the system.

Unfortunately, this method was found to be principally unreliable as a result of it did not take under consideration the chance that real profiles may be graded low and faux profiles may be graded high. Sarcode and Mishra projected a special approach that may be a sequence of steps to notice faux profiles. They used the Face book graph API tool to achieve access to varied profiles and wrote a script to extract the viewed data. Later on, this extracted data forms the attributes the classifier can use in their algorithmic rule. First, the information is in JSON format, that is additional parsed to a structured format (CSV) that's easier legible by machine learning techniques. These commas separated values can later build the classifier additional economical. The authors tried unattended and additionally supervised machine learning techniques. They used eightieth of the samples to coach the classifier and therefore the rest to check it. Once the algorithmic rule runs, there's feedback provided to the profile, requiring it to submit identification to prove it's not a faux profile. Profile's area unit processed on mass to extract options. Resilient Back Propagation algorithmic rule in neural networks algorithmic rule combined with support vector machines is employed within the classification of pretend profiles. Sybil Frame uses multi-stage level classification. Approaches embrace content- based mostly} and structure based. Content- based approach explores the dataset and extracts data accustomed calculate previous data regarding nodes and edges. Structure-based approach correlates nodes victimization mathematician randomfield and insane belief propagation that employs previous data.

3. EXISTING SYSTEM

The existing systems use very fewer factors to decide whether an account is fake or not. The factors largely affect the way decision making occurs. When the number of factors is low, the accuracy of the decision making is reduced significantly. There is an exceptional improvement in fake account creation, which is unmatched by the software or application used to detect the fake account. Due to the advancement in creation of fake account, existing methods have turned obsolete. The most common algorithm used by fake account detection Applications is the Naïve bias classifier. The accuracy of existing system is less compared to proposed system..

4. PROPOSED SYSTEM

In the proposed system, the system elaborates a classification of spammer detection techniques. The system shows the proposed taxonomy for identification of spammers on Tw. The proposed taxonomy is categorized into four main classes, namely,

- fake content,
- detecting spam in trending topics, and
- URL based spam detection,
- fake user identification.

Moreover, the analysis also shows that several machine learning-based techniques can be effective for identifying spams on social media. However, the selection of the most feasible techniques and methods id highly dependent on the available data.

- The first category (fake content) includes various techniques, such as regression prediction model, malware alerting system, and Lfun scheme approach.
- In the third category (URL based spam detection), the spammer is identified in URI. through different machine learning algorithms.
- The last category (fake user identification) is based on detecting fake users through hybrid techniques.

ADVANTAGES OF PROPOSED SYSTEM

1.This study includes the comparison of various previous methodologies proposed using different datasets and with different characteristics and accomplishments.

2.Tested with real time data.

3.In proposed system we use Random Forest Algorithm. These algorithms use a smaller number of features, while still being able to correctly classify about 98% of the accounts of our training dataset..

5. METHODOLOGIES

MODULE

The below steps are followed in a Machine Learning process:

1.DEFINE OBJECTIVE:

Define the goal of the Problem Statement At this step, we need to recognize what precisely wishes to be expected. In our case, the goal is to are expecting the opportunity of rainwith the aid of using reading climate conditions. At this degree, it's also crucial to take intellectual notes on what form of records may be used to remedy this trouble or the kind of technique you need to comply with to get to the solution.

2 .DATA GATHERING:

Data Gathering At this degree, you need to be asking questions along with,

• What formof records is wanted to remedy this trouble? Are the records available?

• How can I get the records? Once you recognize the kinds of recordsthis is required, you need to recognize how you could derive these records. Data series may beachieved manually or with the aid of using internet scraping. However, if you’re a novice and also, you’re simply seeking to research Machine Learning you don’t need to fear approximatelygetting the records. There are hundreds of records assets at the internet, you could simply download the records set and get going. Coming returned to the trouble at hand, the records wanted for climate forecasting consists of measures along with humidity level, temperature, pressure, locality, whether or not or now no longer you stay in a hill station, etc. Such records need to beaccrued and saved for analysis.

3. PREPARING DATA:

Data Preparation the records you accrued is nearly by no means within side the properformat. You will stumble upon a number of inconsistencies within side the records set along with lacking values, redundant variables, replica values, etc. Removing such inconsistencies may be very crucial due to the fact they may result in wrongful computations and predictions. Therefore, at this degree, you test the records set for any inconsistencies and also you restore them then and there.

4.DATA EXPLORATION:

Exploratory Data Analysis Grab your detective glasses due to the fact this degree is allapproximately diving deep into records and locating all of the hidden records mysteries. EDA or Exploratory Data Analysis is the brainstorming degree of Machine Learning. Data Exploration includes information the styles and developments within side the records. At this degree, all of the beneficial insights are drawn and correlations among the variables are understood. For example, within side the case of predicting rainfall, we realize that there may be a sturdy opportunity of rain if the temperature has fallen low. Such correlations need to be understood and mapped at this degree.

5.BUILDING A MODEL:

Building a Machine Learning Model All the insights and styles derived all through Data Exploration are used to construct the Machine Learning Model. This degree constantly starts of evolved with the aid of using splitting the records set into parts, schooling records, and checking out records. The schooling records can be used to construct and examine the version. The common sense of the version is primarily based totally at the Machine Learning Algorithm this is being carried out. Choosing the proper set of rules relies upon at the kind of trouble you are seeking to remedy, the records set and the extent of complexity of the trouble. In the imminent sections, we are able to speak the special kinds of issues that may be solved with the aid of using the use of Machine Learning.

6. MODEL EVALUATION:

Model Evaluation & Optimization After constructing a version with the aid of using theuse of the schooling records set; it’s far subsequently time to position the version to a test. Thechecking out records set is used to test the performance of the version and the way appropriatelyit is able to are expecting the outcome. Once the accuracy is calculated, any in addition upgradeswithinside the version may be carried out at this degree. Methods like parameter tuning and cross-validation may be used to enhance the overall performance of the version.

7.PREDICTION:

Predictions Once the version is evaluated and improved, it's far subsequently used to make predictions. The very last output may be a Categorical variable (e.g., True or False) or it is able to be a Continuous Quantity (e.g., the expected cost of a stock).

6. RESULTS AND DISCUSSION SCREEN SHOTS

SCREEN-1

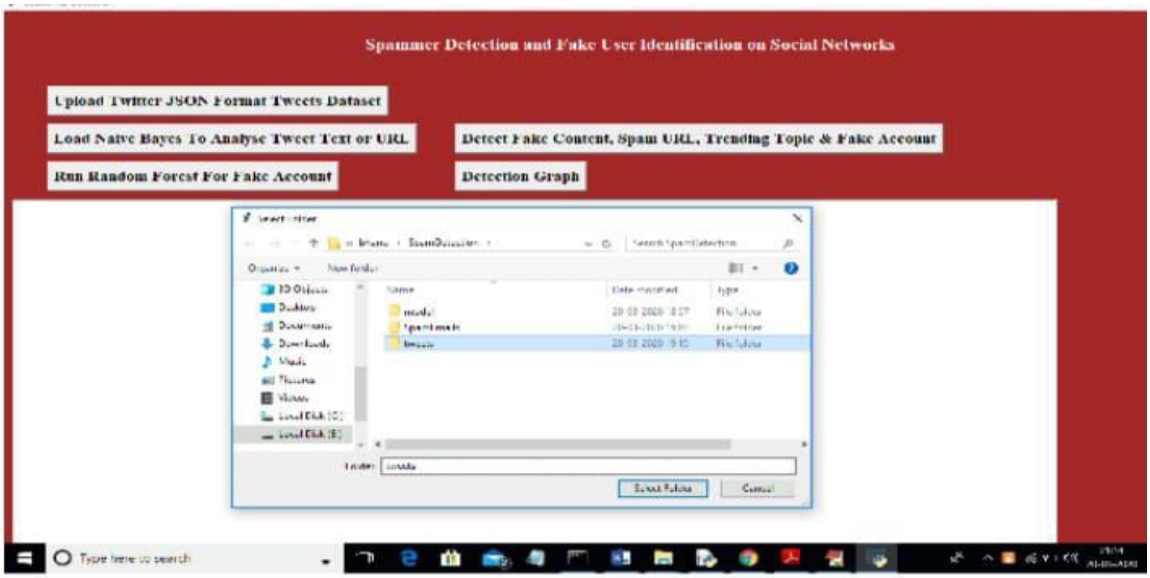

Fig.4. Uploading 'tweets' folder

SCREEN-2

Fig.5. Load Naive Bayes

SCREEN-3

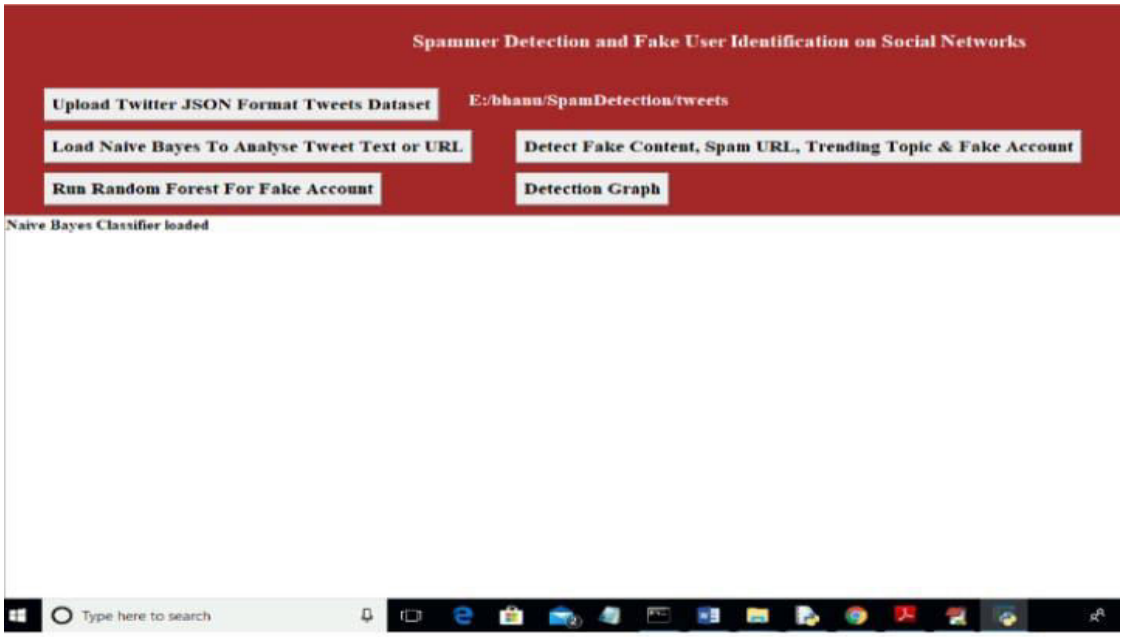

Fig.6. Detect Fake Content

SCREEN-4

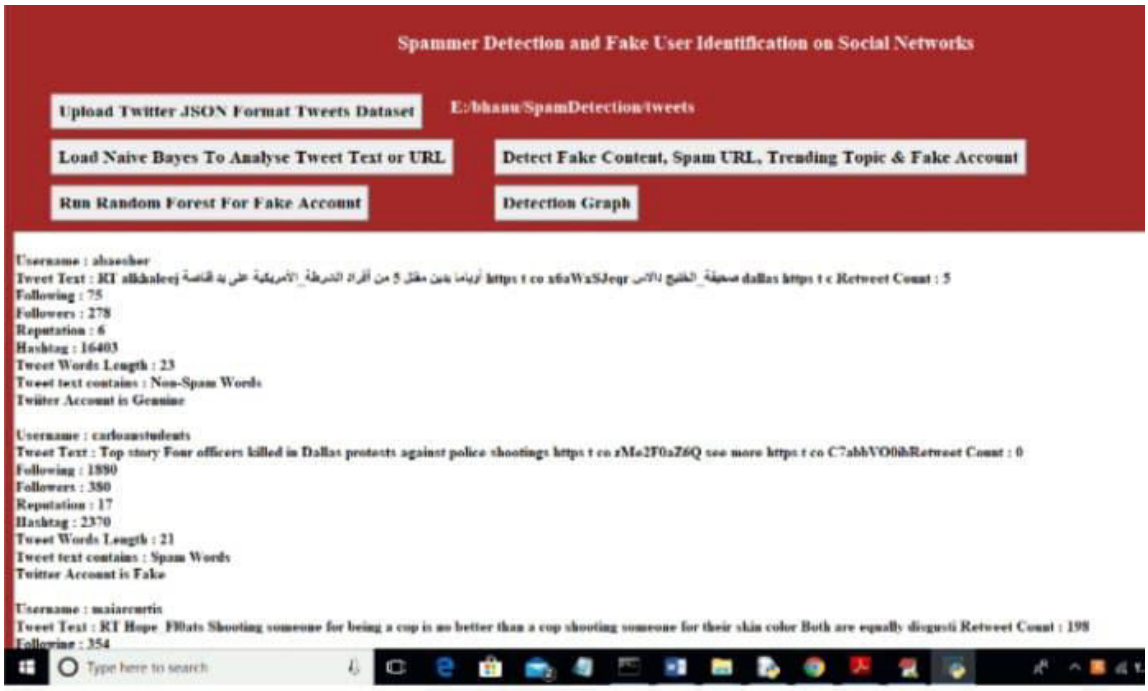

Fig. 7. Features extracted from tweets

SCREEN-5

SCREEN-6

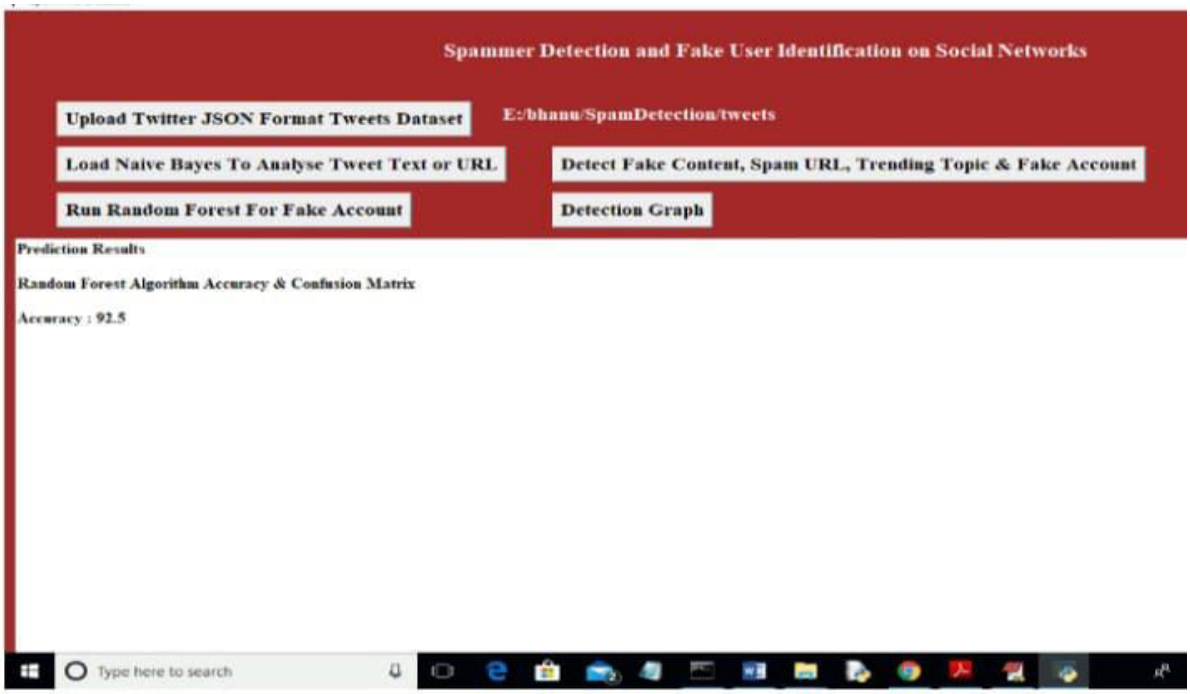

Fig.8. Random Forest Prediction Accuracy as 92%

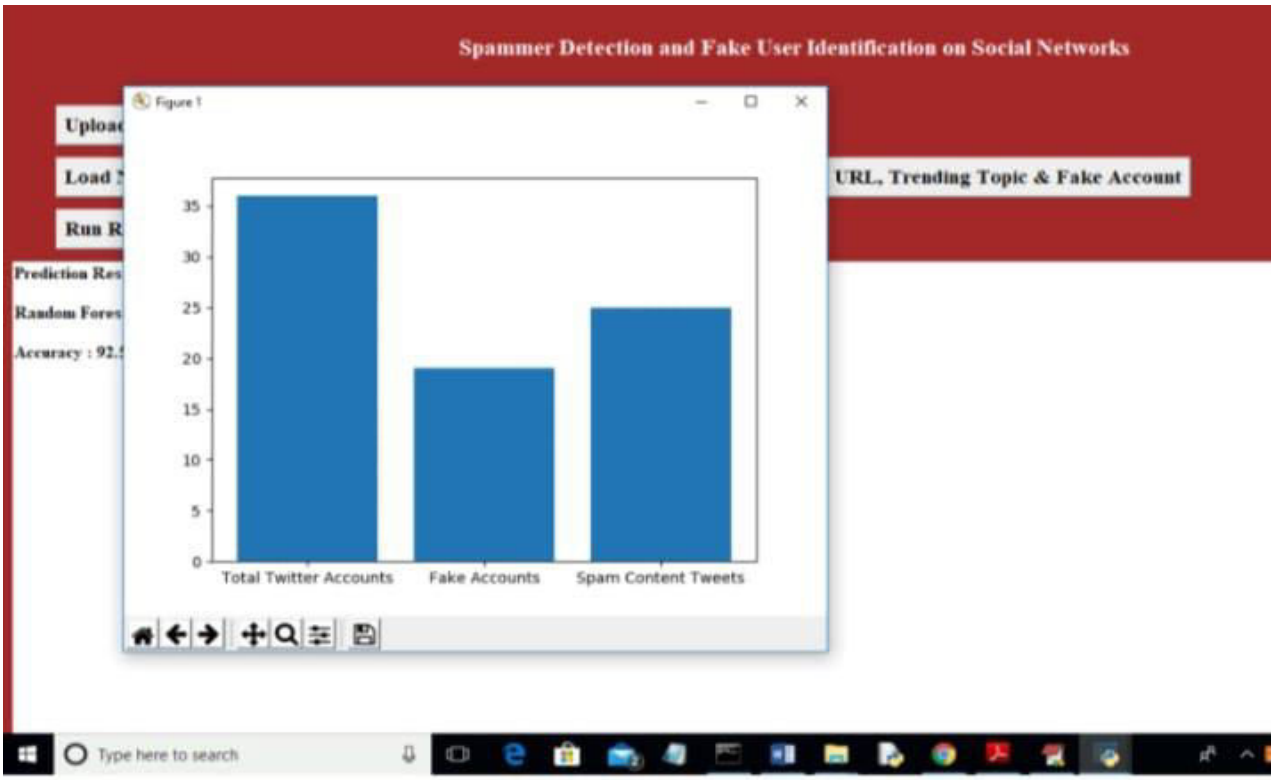

Fig.9. Total Tweets

7. CONCLUSION AND FUTURE SCOPE

We have given a framework which collects data from Twitter using Twitter API and from every tweet, we extract features that we need to feed our classifiers, that binary classification through the Random Forest is more efficient than through any other classifier. Using Decision tree, we have achieved the efficiency of 96%. In the future, we wish to classify profiles by analyzing the behavior of the user by his tweets find out a pattern and classify

Despite the development of efficient and successful ways for spam detection and fake user identification on Twitter , there are still certain gaps in the study that need to be addressed. The following are a few of the issues: Because of the substantial ramifications of false news on an individual and communal level, false news identification on social media networks is a subject that needs to be investigated. The identification of rumour origins on social media is another related topic worth researching.

Although a few studies using statistical methods to discover the origin of rumours have already been undertaken, more complex approaches, such as social network-based approaches, can be used due to their demonstrated efficiency.

8. REFERENCES

1. (2018) Political advertising spending on Facebook between 2014 and 2018. Internet draft. [Online]. Available: https://www.statista.com/statistics/891327/political- advertisingspending-facebook-by-sponsor-category/
2. J. R. Douceur, "The sybil attack," in International workshop on peerto-peer systems. Springer, 2002, pp. 251–260.
3. (2012) Cbc.facebook shares drop on news of fake accounts. Internet draft. [Online]. Available: http://www.cbc.ca/news/technology/facebook-shares-drop-onnews-of-fake-accounts-1.1177067
4. R. Kaur and S. Singh, "A survey of data mining and social network analysis based anomaly detection techniques," Egyptian informatics journal, vol. 17, no. 2, pp. 199–216,2016.
5. L. M. Potgieter and R. Naidoo, "Factors explaining user loyalty in a social media- based brand community," South African Journal of Information Management, vol. 19, no. 1, pp.1–9, 2017.
6. (2018) Quarterly earning reports. Internet draft. [Online]. Available: https://investor.fb.com/home/default.aspx
7. (2018) Statista. Twitter: number of monthly active users 2010-2018. Internet draft. [Online]. Available: https://www.statista.com/statistics/282087/number-of- monthly active-twitter-users/
8. Y. Boshmaf, M. Ripeanu, K. Beznosov, and E. Santos-Neto, "Thwarting fake osn accounts by predicting their victims," in Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security. ACM, 2015, pp. 81–89.
9. (2018) Facebook publishes enforcement numbers for the first time. Internet draft. [Online]. Available: https://newsroom.fb.com/news/2018/05/enforcement-numbers/
10. (2013) Banque populaire dis-moicombiendamistu as sur facebook, je tediraisi tabanquevataccorder un prłt. Internetdraft. [Online]. Available:
11. http://bigbrowser.blog.lemonde.fr/2013/09/19/popularitedis-moi-combien-damis-tu-as- sur-facebook- je-te-dirai-si-ta-banqueva-taccorder-un-pret/
12. S.-T. Sun, Y. Boshmaf, K. Hawkey, and K. Beznosov, "A billion keys, but few locks:the crisis of web single sign-on," in Proceedings of the 2010 New Security Paradigms Workshop. ACM, 2010, pp. 61– 72.