A SILENT HACK DETECTION BASED ON DEEP-LEARNING TECHNIQUE

Nuha Almozaini, Yasmin Alateeq, Noura Alrajeh and Saleh Albahli

IT Department, College of Computers, Qassim University, KSA

ABSTRACT

Sharing information has been democratized with the rise of social networks. Consequently, increasing the usage of Social Network, especially Twitter platform, leads to growing malicious activities. With a silent hack, a hacker can continuously dig around to control over victim's account. In this paper, an observed direct impact to users' security and privacy has been identified. Therefore, we address hidden tactics in the problem specific feature engineering with detailed results to show how deep leaning classifiers are promising direction to understand sentiment than classical machine learning. Thus, we focus on the state of the art Deep learning techniques by constructing a model to detect behavioral changes of users. Our evaluation shows that working with just classical machine algorithms to analyse social data do not achieve higher performance than deep learning algorithms. This will open directions for using deep learning for similar problems. Moreover, our results demonstrate the shortages of classical Machine Learning classifiers compared to Deep learning and how they can be mitigated.

KEYWORDS

Deep learning, Machine learning, silent hacking, social data, behaviors, analysis, Twitter.

1. INTRODUCTION

With the rise of social networks, some people exploit them for bad behaviors. Therefore, it becomes a huge risk to young people's ethics because of their limited capacity for self regulation and susceptibility to peer pressure. With millions of tweets every day, these lead to growing malicious activities.

Technology now is more than collecting and preparing information to users and organizations. Technology seeks insight and knowledge, and that is what we look for in this paper. With Machine Learning behind the scenes, the field that concerns with how machine can learn from experiences, handling the intelligence part to go deeply in data and predict what's coming is our way of the paper.

Activities of users by flooding the Internet with data and sharing content in social networks specifically as a lifestyle must have implicit things in their activities. These are interesting, significant and most importantly abnormal which is a goal to either governments or organizations that target users. These activities that users act online are commonly called behaviors.

David C. Wyld et al. (Eds) : IPPR, SOENG, DaMi, CSIT, AIS, CSE, CSIP, CCNET - 2018
pp. 169–176, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80513

Therefore, this paper aims to find such behaviors that threaten privacy and ethics in communities, countries and people by collecting enough data, setting it up and then finally subjecting it to analysis stages involving machine learning algorithms and predictions. Then, it will end up with a data product that captures what would risk people all around the world.

Thus, we hypothesize that, on one side, working with just classical machine algorithms to analyse social data do not achieve higher performance than deep learning algorithms. As such, this paper attempts to show how deep leaning classifiers are promising direction to understand sentiment than classical machine learning. Besides, we study the effects of a behavioral change of users by using Python with trained different models to expose behaviors and get subjectivity of micro-blogging content. Accordingly, we attempt in this paper to find out more about attitudes, relationships and connectivity between users of Twitter by taking the powerful of deep learning techniques.

Problem Statement: Social network penetration worldwide is ever-increasing. The increased worldwide usage of Online Social Network (OSN) that leads to growing malicious activities. Twitter is one of the most social networks have infected accounts. Therefore, silent hacking on Twitter try to reach active users resulting in negatively impacting other genuine Twitter users. Thus, we aim to focus on the nature of social networking and how users among social media react and behave. We also aim to detect whoever acts abnormal and penetrates user's privacy putting both acts in deeper analysis stages to bring insights from these behaviors. Ultimately, we attempt to protect normal users to avoid damaging their image.

Research Questions: Our research questions focus on two main theme:

- What are the main challenges when analyzing behavior in Social Networks and Twitter interactions?

- How can these challenges be addressed by using state-of-the-art machine learning technologies and algorithms?

Proposed Approach: People are concerned with the security of their accounts and their private information in the accounts and they might not know if they were hacked because some attackers don't leave a trace, so we are going to seek this trace.

Our solution is to capture behaviors that meet our expectations of social media behaviors and lead to reasoned view of how people interact over the internet. Practically, our solution focus on the state of the art deep learning classifiers by constructing a model to detect behavioral changes of users. We label collected tweets as spam or non-spam and adverts or not. In addition, identifies four features (URL, Hashtag, Media, sensitive information) that lead us inferencing them as normal or abnormal. Then, used mechanism to analyze each collected tweet manually and independently to show how Deep Learning classifiers achieve better performance than classical Machine learning classifiers.

Contributions: The primary contributions of our paper may be summarized as follows:

- To construct a model to detect behavioral changes of users in online social media.

- To show how Deep Learning classifiers has an advantage over classical Machine learning classifiers.

Organization: Section 2 provide the related work, followed by the discussion of the benefit of using Semantic Analysis over Sentiment Analysis. Section 3 gives an overview of data collection. The candidate features to identify hacked tweets are presented in section 4. The experimental design, evaluation and results are shown in section 5. Finally, we conclude the paper in section 6 with an outlook on future work.

2. RELATED WORK

Since millions of users on Twitter tweet constantly in their informal languages, the issue with this type of analysis is analyzing words and how to deal with informal phrases. It may contain acronyms, abbreviations, slang words, misspelled words, non-opinion words, sarcasm, etc. Hence, the challenge is to detect neutral words that will help removing non-opinion words [1]. In addition, Twitter often keeps sensitive information about users like their locations secret. Therefore, privacy issue plays a major role of collecting dataset for this type of information [2].

Sentiment Analysis has been used for long time to classify words based on linguistic artifacts that show sentiment and syntax patterns to link subject with the sentiment classification. On the other hand, a more reliable approach is called semantic analysis which help to cluster different data rely on similarity instead of current classification such as positive/negative/neutral [3][4][5].

Savage et al.[6] use graph based analysis to observe suspicious behaviors in Online Social Networks (OSN) and categorized them into three different types: inappropriate content share, silent hacking, and fake promotional accounts. For the first type, they use it to catch accounts/nodes that have very high in-degree as well as very less number of friends; users with such behavior are usually hiding themselves using fake identities. For the second type, since hackers will not be visible and obvious, an active node gets a pattern different from usual which it increases visibly but the out-degree remains the same and therefore it is called “silent”. The third and final type is handled by seeking out-degrees for a node. Hence, if most of the out-degrees targeting one node, this act will be categorized to the fake promotional accounts since this type of accounts intentionally promote specific node/user.

Bravo-Marquez et al. [7] propose word-level classification to show how to generate opinion lexicons from unlabelled tweets. They use Sentiment Analysis to classify words either positive, negative or neutral. Tweets are represented by using two vector: bag of words and a semantic vector based on word-clusters. Finally, they show that the clusterbased vectors are better than the bag of words vectors.

El Kassiri et al. [8] [9] show that semantic similarity measure and RDF graphs achieve high performance for link prediction. They propose an Ontology called ActOnto to share common activities made by communities in social networks. However, as mentioned before, there are other frameworks used to analyses opinions of users in social networks by using the sentiment analysis.

3. DATA COLLECTION

Dataset is always an asset and an indicative part of data science experiments. In the proposed work, dataset prepared continuously and carefully throughout the experiment. Furthermore, a survey was made for 380 twitter users to clarify the kind of actions to be observed from raw data. The main impact of the dataset is sampling (Extracting subset of a dataset), and sampling has been done regularly. When a sample is acquired, labelling manually is applied to the sample, then preparation and preprocessing steps are implemented to make the sample conforms to the experiment requirements. Last step is training the classifiers then evaluate them for prediction and if the prediction accuracy is low, alternatives are taken i.e. obtain more data, redo the experiment and so on.

The experiment is implemented with 3576 tweets, 1694 are labelled abnormal and 1882 are normal.

Data Analysis: In our experiment, we analyze data based on tweet contents. The data were trained using different Machine Learning classifiers (Table 5.1) to be able to compare accuracy of the classifiers regarding the problem domain. 10% of the data was excluded from the training phase to be used to test the models and 90% was in the training phase.

4. CANDIDATE FEATURES

Based on the domain knowledge, we identified hacked tweets based on six candidate features: URL, Spam, Ads, Media, Hashtag and Sensitive information [10][11][12][13].

URL: hackers target users by posting links of malicious websites in their tweets. Since links lead victims to the sites, they often use this method in addition to URL shortened services which can hide the targeted URL. So, hackers tend to use such services to post links to their websites. This feature identified by true/false whether a URL in a tweet or not.

Spam: spam accounts on Twitter post same content many times and maybe have a small changing of the tweet. Accordingly, they target to send same content to many users.

Ads: it was also noticed that advertisement content usually reflects some malicious content characteristics so, advertisement content has been selected as the third feature. An example of Ads tweet can be shown as follows:

```
Click to #win #Hellraiser: The Scarlet Box on Blu-ray with @HeyUGuys
          https://t.co/afsq2qGB1Q https://t.co/hMq5wZl3G2
```

Media: identified whether media contents are included or not. This feature can be indicated a spam and is like sub-feature of spam feature, therefore media has been selected as the forth feature.

Hashtag: infected accounts try to attract legitimate users to read their tweets by posting multiple unrelated tweets using trending hashtags. These accounts hope to reach more viewers quickly, so they use trending and popular hashtags in their tweets. Therefore, this feature is candidate for detecting spammers too.

Sensitive information: Attackers normally lead users to their malicious contents, so they take advantage of the basis of human behavior. Thus, users are led by their instincts which sometimes overcome morals and ethics. Moreover, the leverage of adult content can be a trigger for human instincts that leads them to explore such content, therefore we take sensitive content as a candidate feature.

After picking appropriate features, tweets have been cleaned feasibly in a way that does not take out the treasure of the problem-related anomaly data [14]. Cleansing process accomplished using Python scripts and luckily the dataset has only English tweets so the cleansing process was for eliminating emojis, symbols as well as irrelevant URLs.

When data was collected directly from Twitter feed, there were URLs in tweets that were not practically activities of users sharing external content. These URLs were "Quote Tweet" activities. A quote Tweet is very normal activity in Twitter platform where users comment on others' tweets but as new tweets published directly to their own feed so that their followers can see what they commented on, But the tweets come with a URL of the tweets being commented-on. Now Twitter replaced these URLs with small boxes containing the commented-on tweets for readability but not at data collection. That is why these URLs were excluded programmatically from the dataset.

5. EXPERIMENTS

This section discusses our experimental evaluation with different popular machine learning models, including Deep Learning, Support Vector Machine (SVM), Random Forest and Naive Bayes as summarized in Table 5.1. The same parameter initialization is utilized when comparing multiple optimization classifiers.

Dataset and Feature Selection:

Acquiring data is an easy task since datasets are publically available online, but acquiring an appropriate and suitable data for specific domain can be very difficult task so data scientists may end up with data generation tools to obtain relevant and appropriate data. Since the proposed work is a classification problem, there was no dataset available online that meets the need of this work and ready for classifying abnormal behaviors. Thus, raw data has been collected from a freely available dataset. The collected data is a subset of the dataset, which is Tweets attribute, also known as feature. Feature selection is a task that requires deeper look into the problem to get the most useful and relevant features to the problem domain. However, for constructing a model to handle text data, the following preprocessing techniques were applied to increase text mining accuracy. First technique is Tokenization which takes lines of text then turns them into individual separated words. The second is Stop Words Elimination, and stop words are the unimportant words that do not add any meaning for text analysis experiments, like the word "an". The third is Stemming, and this technique turns words to their stem like the words "Closed" and "Closing" both will be "Close" after Stemming. The fourth and last technique is Transform Cases which turns uppercase to lowercase so that the "Read" and "read" words will be considered the same after this process.

The dataset has been enhanced several times using features engineering. The models have been improved using text processing techniques such as tokenization to end up with a better accuracy as shown in Table 5.1.

Table 5.1: Overview of experimental results

Deep Learning	**95.57%**
Random Forest	94.90%
Support Vector Machine (SVM)	93.67%
Naïve Bayes	91.72%

Experimental Results and Discussion for classical algorithms: For classical Machine Learning classifiers, Table 5.1. shows that SVM resulted robust classification capabilities that even with major and minor changes of the experiment it stays at high accuracy, near 90% and ended up with 93.67%. **Random Forest**, on the other hand, has shown sensitivity regarding text preprocessing and feature engineering. It shows very low accuracy at the beginning of the experiment when there are four features, tweets, URL, spam and ads and only tokenization for text preprocessing. But it surprisingly increased to 94.90% after optimizing text preprocessing methods and feature engineering. For **Naive Bayes**, it shows good enough result from the beginning and was very flexible at optimizing. It kept increasing while improving the experiment, until it ended up with 91.72% accuracy.

Experimental Results for Deep Learning: In Table 5.1, the Deep learning classifier resulted 95.57% of the tweet correctly, which shows the best performance among other classical ML algorithms. However, Deep Learning did not work well at the beginning, so after enhancing many feature engineering and Neural Network, it is resulted much higher than other compared classifiers. The overall number of layers used is six; two for input/output and four hidden layers. Dropout was applied to all hidden layers with value of 0.2 and dropout is Deep Learning approach to avoid overfitting so that the model can be tested on unseen data. The Adam optimization algorithm [15] used throughout. The ReLU activation [16] used for our experiment for all layers except the output layer which used Sigmoid activation. So, after all these optimizations for layers' architecture, Deep Learning classifier is reached 95.57% accuracy.

6. CONCLUSIONS AND FUTURE WORK

Increasing the usage of Online Social Network (OSN) leads to growing malicious activities. Twitter is one of the most social networks have infected accounts. Our paper aims to show that Deep Learning classifiers has gained advantage over classical Machine learning classifiers. As such, we study the effects of a behavioral change of users by using Python with trained different models to expose behaviors and get subjectivity of microblogging content. Accordingly, we show in this paper how to find out more about attitudes, relationships and connectivity between users of Twitter by taking the powerful of deep learning techniques. In detail, four algorithms were applied to the problem and what was remarkable is that deep learning went beyond expectations for the prepared small portion of data which agree with our hypothesis.

Since we aim at to detect behavioral changes of users in online social network, our further research plans to extend the proposed work to build a semantic-based model using different ontology techniques.

REFERENCES

[1] F. Atefeh and W. Khreich, "A survey of techniques for event detection in Twitter," Computational Intelligence, vol. 31, no. 1, pp. 133–164, 2015.

[2] J. Schmidhuber, "Deep Learning in neural networks: An overview," Neural Networks, vol. 61. pp. 85–117, 2015.

[3] A.R. Guess, "Sentiment Analysis v. Semantic Analysis: A much more statistically reliable approach is semantic analysis." [Online]. Available: http://www.dataversity.net/sentiment-analysis-v-semantic-analysis/. [Accessed:18-Jan-2018].

[4] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012, vol. 7649 LNCS, no.7 PART 1, pp. 508–524.

[5] R. Giovanetti and L. Lancieri, "Model of computer architecture for online social networks flexible data analysis: The case of Twitter data," in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, pp. 677–684.

[6] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Anomaly detection in online social networks," Social Networks, vol. 39, no. 1, pp. 62–70, 2014.

[7] F. Bravo-Marquez, E. Frank, and B. Pfahringer, "From Unlabelled Tweets to Twitter-specific Opinion Words," Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'15, pp. 743–746, 2015.

[8] A. El Kassiri and F. Z. Belouadha, "ActOnto: An extension of the SIOC standard for social media analysis and interoperability," in Colloquium in Information Science and Technology, CIST, 2015, vol. 2015–Janua, no. January, pp. 62–67.

[9] A. El Kassiri, F. B.-I. S. T. and, and undefined 2015, "Towards a unified semantic model for online social networks analysis and interoperability," in 10th International Conference on Intelligent Systems: Theories and Applications (SITA), 2015, pp. 1–6.

[10] X. Ruan, Z. Wu, H. Wang, and S. Jajodia, "Profiling Online Social Behaviors for Compromised Account Detection," IEEE Transactions on Information Forensics and Security, vol. 11, no. 1, pp. 176–187, 2016.

[11] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Towards Detecting Compromised Accounts on Social Networks," IEEE Transactions on Dependable and Secure Computing, vol. 14, no. 4, pp. 447–460, 2017.

[12] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," Neurocomputing, vol. 159, no. 1, pp. 27–34, 2015.

[13] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 502–518, 2017.

[14] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553. pp. 436–444, 2015.

[15] D. P. Kingma and J. L. Ba, “Adam: a Method for Stochastic Optimization,” International Conference on Learning Representations 2015, pp. 1–15, 2015.

[16] V. Nair and G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines,” Proceedings of the 27th International Conference on Machine Learning, no. 3, pp. 807–814, 2010.