Cardiff University

School of Computer Science and Informatics

CM3203 - One Semester Individual Project - 40 Credits

Initial Plan

Adversarial Reasoning in Machine Learning for Natural Language Processing: the case of spam emails

Author - Hasna Ahmed Al Jufaili

Supervisor - Federico Cerutti

Moderator - Philipp Reinecke

Table of Contents

Project Description

Background information

Nowadays, emails became a major part of individuals daily life as a vital communication tool. Every second, billions of emails are sent around including spam emails, where spam emails are simply commercial emails that contain links that might look familiar to the user but leads to phishing web sites or malware hosted sites and contain malwares as scripts or executable file attachments that may harm the user. According to IBM Security, spam emails are considered as a primary tool in each attacker's toolkit. In addition, research done by IBM security showed that the number of spam emails is increasing rapidly, and the volume increased by 4x (Fourfold) in 2016 (IBM Security, 2017). Moreover, the spam email rate raised significantly during 2016, from 1 in 220 emails in 2015, to 1 in 131 emails in 2016. The increase was driven by botnets such as, Locky, TeslaCrypt and Dridex which are used to distribute and send spam emails (Symantic, 2017).

Therefore, to protect the users from the threat of spam emails, email spam filters and detectors are used and provided by email services and several software's to keep the spam emails out of the user's inbox. These filters, use statistical machine learning classification techniques and algorithms such as Decision trees (Bhowmick & Hazarika, 2016) and Naïve Bayes classifier (Bhowmick & Hazarika, 2016) to produce intelligent decisions on the classification of the data in the emails to decide whether an email is spam or not. These statistical algorithms are implemented into a machine learning model that adapt the application to the changes in data by learning and training constantly on given datasets, which then produces a set of patterns and rules that will be followed to classify future data (Saini, 2008). However, security issues arise from the machine learning model's adaptability in a presence on an adversary that can subvert the machine learning model and manipulate the attack samples during the model test time (Saini, 2008). Possible attacks could be, causative attacks, in which the adversary alters the learning model by influencing the training datasets and exploratory attacks, where the adversary exploits errors in the classification model without altering the training datasets (Huang et al, 2011).

Brief Description

This project will investigate the security issues that can arise from a spam filter that depends on machine learning algorithms and techniques in the presence of an adversary. To investigate the security issues, Adversarial Machine Learning (AML) library [1] will be reviewed and python Scikit-learn library [2] will be understood to aid in the implementation phase. Moreover, the project will identify and analyse data related to spam checking and will explore a range of possible attacks. In addition, the project will produce multiple software, each representing a machine learning model that will work as the spam filter and will be implemented on a specific statistical classification machine learning algorithm. Each model will be trained by providing email datasets. Furthermore, several attacks will be implemented and tested against the implemented machine learning models. All the implementation of the project will use python as programming language and will be implemented and tested on a windows device and a documentation and reflective exercises on the achievements will be included. The overall aim of this project is to prove that spam filters depending on machine learning techniques are not fully secure in a presence of an adversary.

Project Aims and Objectives

This section will present a list of more detailed aims and objectives of the proposed project.

- Aim: Implement two machine learning models that will work as a spam filter

- Objectives: Use python Scikit-learn library [2] (Scikit-learn developers, 2017) since it provides classes for machine learning algorithms to implement Naïve Bayes classifier and Decision tree classifier (Aski & Sourati) for the machine learning models as they have been used in practice in multiple spam filters and then train the models on the emails dataset.

- Aim: Implement a range of possible adversary attacks and use them to attack the implemented machine learning models

- Objectives: Research about adversary attacks and implement them, such as Good word attack (Lowd & Meek, 2005), which has been used in practice in AML python library [1] against statistical machine learning spam filters and then attack the models.

- **Aim:** Documentation and reflective exercises on the achievements

- Objectives: by critical analysing the critical path of the project and justification and evaluation of the results from the testing of the implementations.

Supervisor roles – Federico Cerutti

There will be a meeting each week on Tuesday for a short duration of 20-30 minutes with the supervisor where advice, guidance and feedback will be provided.

Ethics

The project as discussed with the supervisor is likely not to require an ethical approval. Data used in the project in the implementation and testing phase is the spam email datasets provided by Unitec [3] that has been used in several researches and does not utilize personal information. Furthermore, there will be ethical consideration during the entire major and in the final report.

Work plan

To manage the report efficiently and produce a successful project, a work plan has been created. The work plan divides the project work duration weekly, which results into fifteen weeks. In each week there will be a regular meeting with the supervisor for 30 minutes each Tuesday excluding the Easter recess, therefore, it is not included in the work plan. In addition, all the tasks needed to complete the project with the deliverables, milestones and progress review meetings has been included. The work plan is presented in the next page.

Weeks	**Tasks**	**Deliverables/ Milestones**
W1/ 29th Jan - 4th Feb	- Meet with supervisor - Prepare Initial plan report draft - Finish Initial plan report after getting feedback from supervisor	- Initial plan draft
W2/ 5th Feb - 11th Feb	- Submit Initial plan report - Start background research on: • Adversarial reasoning • Machine learning process • Machine learning algorithms for spam email filtering • Email checking process • Python machine learning libraries and modules • Possible attacks on machine learning algorithms that are suitable for spam email filtering • Email datasets	- Initial plan report: project description, aims and objectives and work plan - Written background research (Background section) **- Milestones:** • Initial plan report submitted by 5th of February
W3/ 12th Feb - 18th Feb	- Review the adversarial machine learning library - Gather understanding of Scikit-learn library - Start the Approach section (Specification and Design): • Describe current approaches of designs for the chosen machine learning algorithms/attacks • Describe the chosen approaches with justification • State the requirements and specification of the software that will be developed - Decide upon python libraries for implementing the learning model and attacks - Evaluate the design and libraries chosen	- Approach section (Specification and design)

W4/ 19th Feb - 25th Feb	- Design UML diagram/ Class diagram for the machine learning models - Design UML diagram/ Class diagram for the attacks	- Diagrams for the machine learning models and the attacks
W5/ 26th Feb - 4th Mar	**- Progress review (Special Meeting)** - Collect data (Emails including spam emails) - Identify and analyse data related to spam email checking - Implementation: • Set up a Git repository for the project and a workspace environment • Start implementing two different machine learning models (two different algorithms) using python	---------------
W6/ 5th Mar - 11th Mar	- Implementation: • Continue implementing the two different machine learning models using python	- Two machine learning models code in python **<u>- Milestones:</u>** • The machine learning models works and functions correctly as the chosen algorithms
W7/ 12th Mar - 18th Mar	- Implementation: • Train both implemented machine learning models on emails dataset • Start implementing the attacks chosen in python	----------------
W8/ 19th Mar - 25th Mar	- Implementation: • Continue implementing the attacks chosen in python	----------------
Easter Recess **W9/** 26th Mar – 1st Apr **W10/** 2nd – 8th Apr **W11/** 9th Apr - 15th Apr	- Implementation: • Continue implementing the attacks chosen in python - No meetings will be scheduled for these weeks - Create testcases - Testing:	- Results and Evaluation section - Multiple attacks implemented in python **<u>- Milestones:</u>** • Software has been tested and evaluated

	- Test the implemented machine learning models using the email dataset and train it - Attack the machine learning models using implemented attacks - Fix any errors and bugs that arise from testing - Evaluate the software and the results from testing - Implement additional functions / Improve some functions	- Results has been captured and written
W12/ 16th Apr - 22nd Apr	**- Progress review (Special Meeting)** - Write Future work section: - Future improvements - Social engineering method to attack spam filters - How the project could be taken further	- Future work section
W13/ 23rd Apr - 29th Apr	- Write abstract, introduction, acknowledgements and conclusion sections	- Abstract, introduction, acknowledgements and conclusion sections - Final report first draft **<u>- Milestones:</u>** - Final report first draft completed
W14/ 30th Apr - 6th May	**- Progress review (Special Meeting)** - Compile and update table of contents, table of figures, glossary, table of abbreviations, appendices and references after combining all sections - Review the report and fix any errors	- Table of contents, table of figures, glossary, table of abbreviations, appendices and references
W15/ 7th May - 13th May	- Proof read the Final report - Submit the Final report	-Final Report with all sections **<u>- Milestones:</u>** - Final report submitted by 11th May

Milestones and deliverables

As seen in the work plan, there will be several milestones and deliverables during the project duration. The following list illustrates the most significant milestones and deliverables for this project,

- Initial plan report (Complete version) – submitted by 5th February
    - The complete version of the report including project description, project aims and objectives and a complete work plan.
- First draft of the following sections
    - Background research, approach (design and specification), implementation, results and evaluation and future work that will be delivered in different weeks according to the work plan
- Implementation
    - Python code on the implementation of two machine learning models
    - Python code on the implementation of the attacks chosen
- Final report draft (Continuing the previous sections in the first draft)
    - Final report with complete main body sections and supporting sections (Title page, abstract, acknowledgements, table of contents, table of figures, glossary, table of abbreviations, appendices and references)
- Final report (Complete version) – submitted by 11th May
    - The complete version of the final report with all the sections included and proof read.

Conclusion

In conclusion, this report presented the proposed project description, aims and objectives and a clear work plan that will be followed in the next phase of the project with the deliverables and milestones.

References

Aski, A & Sourati, N, 2016. Proposed efficient algorithm to filter spam using machine learning techniques. *Pacific Science Review A: Natural Science and Engineering 18 (2016) 145-149.* [Online]. Available at: https://ac.els-cdn.com/S2405882316300412/1-s2.0-S2405882316300412-main.pdf?_tid=2dd713de-09c5-11e8-b6f1-00000aab0f26&acdnat=1517760492_835e31e479844fd2f5f50f37ff036aec [Accessed: 04/02/2018]

Bhowmick, A & Hazarika, S, 2016. *Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends*. [Online]. Available at: https://arxiv.org/pdf/1606.01042.pdf [Accessed: 04/02/2018]

Huang, L et al, 2011. *Adversarial Machine learning.* [Online]. Available at: https://people.eecs.berkeley.edu/~tygar/papers/SML2/Adversarial_AISEC.pdf [Accessed: 31/01/2018]

IBM Security, 2017. *IBM X-Force Threat Intelligence Index 2017.* [Online]. Available at: https://www.ibm.com/security/data-breach/threat-intelligence [Accessed: 31/01/2018]

Lowd, D & Meek, C, 2005. *Good Word Attacks on Statistical Spam Filters.* [Online]. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.9846&rep=rep1&type=pdf [Accessed: 04/02/2018]

Saini, U, 2011. *Machine Learning in the Presence of an Adversary: Attacking and Defending the SpamBayes Spam Filter.* [Online]. Available at: https://people.eecs.berkeley.edu/~adj/publications/paper-files/EECS-2008-62.pdf [Accessed: 31/01/2018]

Scikit-learn developers, 2017. *Scikit-learn user guide*. [Online]. Available at: http://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf [Accessed: 04/02/2018]

Symantic, 2017. *Internet Security Threat Report*. [Online]. Available at: https://digitalhubshare.symantec.com/content/dam/ent/collat/reports/RPT_ISTR-Main-Report_EN.pdf?aid=elq_&elqTrackId=a4536141fdba4e639c7e80d98f3e522f&elqaid=3783&elqat=2 [Accessed: 31/01/2018]

Libraries and Datasets

[1] Scikit-learn library: http://scikit-learn.org/stable/index.html

[2] AML library: https://github.com/vu-aml/adlib

[3] spam email datasets: http://csmining.org/index.php/spam-email-datasets-.html