Cardiff University School of Computer Science and Informatics

Evaluating Machine Learning Methods for Malware Detection and Classification

CM3203 – Initial Plan– 40 Credits

Author:

Reema Abaoud

Student Number:

1660906

Supervisor:

Philipp Reinecke

Moderator:

Matthew J W Morgon

Table of Contents

1.Project Description:

1.1 Introduction:

Malware can be called on any software that is designed to perform malicious actions that is intended to cause damage to the computer systems (Aycock, 2011). Malwares can spread to computers in different ways such as links, files, or emails that contain viruses that are waiting for users to click and download them to automatically download malicious programs to the system (Rouse 2017). Moreover, due to the exponential growth of internet technology and computer networking the threats caused by malwares are arising day by day to both individuals and organizations. The number of new malware on the internet keep on increasing at an alarming rate according to AVTest which is an IT security institute more than 350,000 new malware programs are registered everyday (2018). This highlight the great demand of the cybersecurity task to guarantee malware protection of computer systems for users and business, since a single attack can cause in compromised system and sufficient data losses. An effective way of detecting and classifying malware files is using machine learning algorithms that is a promising technique in terms of accuracy and high efficiency during execution. Thus, there is a crucial demand to question and evaluate the performance of the well known Machine Learning algorithms used for malware detection and compare

the different methods in terms of accuracy, and the time required for an algorithm to give a result.

1.2 Brief Description:

The aim of this project is to create a program using the platform Python to implement machine learning algorithms that takes a dataset as an input and detect, identify, files that contain malware. Also, be able to classify the types of malware families. Each algorithm will be examined in terms of accuracy and speed of performance to produce a result. At the final stage of this project the results of accuracy and time required to produce a result of all the methods will be compared in order to determine the most efficient method to use in detecting malware.

2. Project Aims and Objectives:

This section will show in detail the main aims and objectives of this project in the given time frame.

1- Gain an understanding in several aspects that are relevant to the project
   - Different types of malware classes
   - The machine learning techniques used to detect malware
   - Identify a software framework, tools, and techniques required to develop the engine for the project

2- Implement machine learning algorithms to detect malware
   - Design a reader for the data intake
   - Design a data transformation that normalize the data to be suitable for the algorithm, in this stage the data will be separated into a "Training set" and "Test Data".
   - Training set well be used to build the model and later evaluated using test model, the produced result will be used to build a new

model

3- Result evaluation and Discussion
    - acquire the detection accuracy of each method implemented
    - calculate the time needed for each method to detect malware
    - compare and evaluate the results and determine which method gave the highest accuracy, and higher performance in terms of time
4- Documentation of every aspect of the project and submit the final report
    - Provide the final report that include background research, implementation process, evaluation results, future work

3. Project Considerations:

One important thing to consider through out the semester is the time boundary, and in order to manage time effectively and efficiently a work plan and Gantt chart are created. Another consideration is to keep the supervisor updated throughout the semester and inform him if any changes or updates are made to the project and plan. At the end of this project assure that the project aims and objectives are achieved in terms of evaluating the machine learning algorithms in detecting malwares.

4.Supervisor Role - Dr.Philipp Reinecke:

A short meeting with the supervisor of approximately 15-20 minutes will be taken place every week except the Easter break. In order to evaluate the progress of each stage weekly, discuss any concerns, and receive constant feedback and advice.

5. Ethics:

After discussing the ethical aspect with my supervisor and since the data that I'm using is publicly published, the data that will be used will not require an ethical approval.

6.Work Plan:

In order to manage the final project efficiently and to finally produce the final report a work plan has been created to shape the stages of accomplishing the project. This work plan will divide the project timeframe in a weekly basis, which will include the weekly task and progress. Moreover, the work plan will include regular meetings with the supervisor to discuss the project progress and any further changes that need to be made.

Weeks and Tasks	Milestone
Week 1: 27 -31 January Task: 1- Background research on used machine learning methods to detect Malware, best platform used. 2- Start writing the initial report. 3- Arrange a meeting with supervisor to discuss project objective, aims, ethics.	-Initial plan draft

Week 2: 3-7 February Task: 1-Supervisor meeting for a feedback of the initial plan. 2- Further research on: Machine learning process, Malware types, Detection methods, Related work, Python machine learning libraries and modules. 3-Malware dataset.	-Initial plan submission by 3rd of February
Week 3: 10-14 February Task: 1-Write the initial draft of introduction part of the final report "Background and research" 2-Start the Methodology section: ▪ Illustrate the chosen machine learning algorithms, Malware types ▪ Indicate the chosen algorithms with the reasons of choosing them ▪ Indicate the requirements and specification of the software that will be developed ▪ Choose the python libraries for implementation ▪ Evaluate the design and libraries chosen 3-Meeting with supervisor	-First draft of Introduction part First draft on the Methodology section
Week 4: 17-21 February Week 5: 24-28 February Task: 1-Desing UML diagrams for general workflow process, machine learning models, malware attack schema 2- Set up a Git repository for the project and a workspace environment	-UML diagrams

3-Start implementation (python): ▪ Import libraries 4-Meeting with supervisor	
Week 6: 2-6 March Week 7: 9-13 March Task: 1-Continue implementation: ▪ load data ▪ Dataset transformation: divide data into training and test data 2-Start writing implementation draft of the final report 3-Meeting with the supervisor	-Implement machine learning models using python
Week 8: 16-20 March Week 9: 23-27 March Task: 1-Continue implementation: training machine learning methods into malware types chosen 2-Meeting with supervisor	-----
Week 10: 30-27 March Week 11: 23-27 March Task: 1-Continue implementation: training machine learning methods into malware types chosen 2- Finalise the implementation part of the final report	- Draft of the implementation part of the final report

Week 12: 30 March- 3 April Week 13: 6-10 April Task: 1-Continue implementation: training machine learning methods into malware types chosen 2-Produce Use cases 3- Produce initial final report draft 4- Special meeting with supervisor to discuss the results and the further evaluation process before Easter	- Use Cases - First Final Report Draft - Results of machine learning detection performance
Easter Break: 13 April - 5 May Task: 1-Create Test cases 2- Testing process: ▪ Test the implemented models using the malware dataset and train it 3-Evaluation process: ▪ Detection Accuracy ▪ Classification of malware type Accuracy ▪ Acquire and record the time of each algorithm to produce a result ▪ Compare the results ▪ Indicate which machine learning algorithm achieved the best performance 4-Combine the report section, compile and update table of contents, figures, glossary, appendices and references 5-Review the report and fix errors	-Result analysis -Final report sections combined
Week 17: 6-8 May Task: 1-Proof read final report 2-Submit Final report	-Submission of Final

	report
Viva Week Task: 1-Viva preparation	-Project completion

7. Milestones and Deliverables:

The work plan specified the project milestones within a determined time frame, the most significant ones will be highlighted in this section.

1. Initial plan report submission – 3/2/2020

2. Initial draft of certain aspects of final report such as:

   Background research, Related work, Approach and implementation, Diagrams.

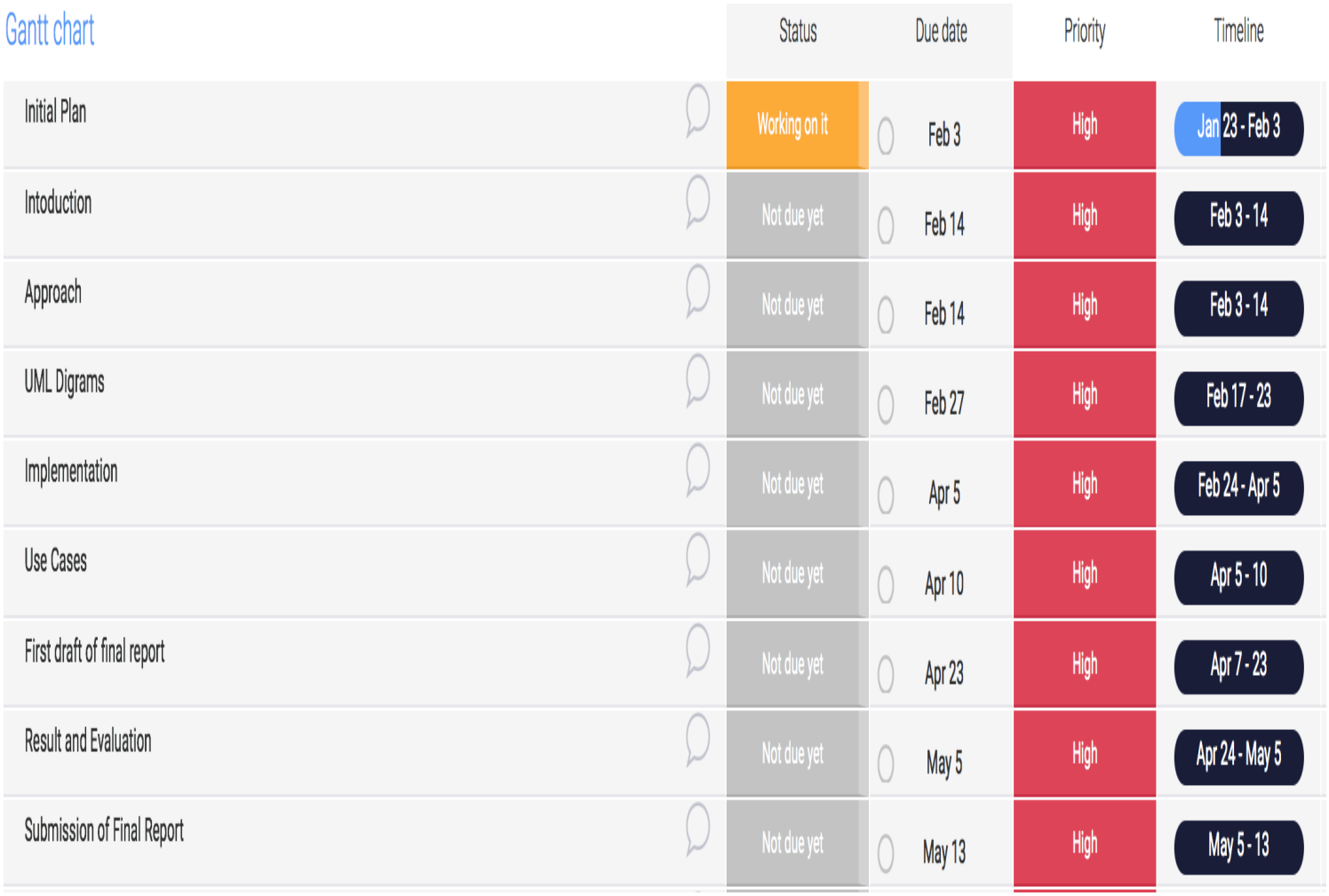
3. The completion of the implemented Software.

4. Second Final Report Draft that contain the full project parts such as: Result and Evaluation, Enhancements and Future Work.

5. Final report submission – 7/5/2020

6. Project completion: Viva presentation

8.Gantt chart:



9.Risk Management:

Risk	Probability of occurrence	Effect of occurrence	Plan to reduce severity
Failure to implement one of the machine learning algorithm	Medium	Low	Before implementing an algorithm study it and examine my abilities to do it. Also, give my self a timeframe to do so, if I'm unable to do it look for an alternative.
Unable to determine the speed of each	Medium	Medium	Discuss with supervisor for an alternative approach to evaluate the performance of machine learning

algorithm			algorithms
Hardware failure or loss of data	Low	High	Back up my work on different locations, such as Google drive, one drive
Failure to achieve milestones on time.	Medium	Medium	Follow the work plan and Gantt chart and assure each milestone is achieved on time.
Sickness	Low	High	Stick to the Gantt chart and work plan to make sure that the workload is distrusted between the weeks to have enough time to recover.

10.Refrences:

1. Av-test.org. (2020). *Malware Statistics & Trends Report | AV-TEST*. [online] Available at: https://www.av-test.org/en/statistics/malware/ [Accessed 29 Jan. 2020].
2. Aycock, J. (2011). *Computer viruses and malware*. New York, NY: Springer.
3. Rouse, M. (2017). *What is Malware? - Definition from WhatIs.com*. [online] SearchSecurity. Available at: https://searchsecurity.techtarget.com/definition/malware [Accessed 29 Jan. 2020].

School of Computer Science and Informatics

Coursework Submission Cover Sheet

Please use Adobe Reader to complete this form. Other applications may cause incompatibility issues.

Student Number

Module Code

Submission Date

Hours spent on this exercise

Special Provision

(Please place an x in the box above if you have provided appropriate evidence of need to the Disability & Dyslexia Service and have requested this adjustment).

Group Submission

For group submissions, *each member of the group must submit a copy of the coversheet.* Please include the student number of the group member tasked with submitting the assignment.

Student number of submitting group member

By submitting this cover sheet you are confirming that the submission has been checked, and that the submitted files are final and complete.

Declaration

By submitting this cover sheet you are accepting the terms of the following declaration.

I hereby declare that the attached submission (or my contribution to it in the case of group submissions) is all my own work, that it has not previously been submitted for assessment and that I have not knowingly allowed it to be copied by another student. I understand that deceiving or attempting to deceive examiners by passing off the work of another writer, as one's own is plagiarism. I also understand that plagiarising another's work or knowingly allowing another student to plagiarise from my work is against the University regulations and that doing so will result in loss of marks and possible disciplinary proceedings.