ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

REVIEWER'S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Should I click on a link? Machine Learning to Protect from Cyber Attacks on the Web**
Author's name:	**Střasák František**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Computer Science
Thesis reviewer:	Carlos A. Catania
Reviewer's department:	Department of Computer Science, National University of Cuyo, Mendoza, Argentina

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment — challenging

Evaluation of thesis difficulty of assignment.

This work attempts to provide a solution to a complex and significant problem in network security. The process for achieving the solution required the learning of several algorithms and techniques included in the state of the art of Artificial intelligence.

Satisfaction of assignment — fulfilled

Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.

The work in the thesis met the assignment. The central aspects of a web site classification method have been developed. From the several data models and algorithms discussed in the thesis, a more in-depth analysis could have been conducted for the CEDT-DOM-2 data model. However, it can be analyzed in the future, as mentioned in the conclusions section of the thesis.

Method of conception — correct

Assess that student has chosen correct approach or solution methods.

The student has followed the correct methodology used in Machine Learning. He has analyzed state of the art, and proposed several sets of features that could deal with the problem. Then, a set data models were carefully designed for evaluating its hypothesis following machine learning standard procedure

Technical level — A - excellent.

Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.

The student has proven himself capable of dealing with a new problem and provided a valid solution using a different set of
tools. He has showed expertise in several areas such as software development, machine learning and network security

Formal and language level, scope of thesis — A - excellent.

Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.

In general, thesis was well written. The student expressed in a clear language the different aspects involved in the process of building a website classification method using formal notation when required.

Selection of sources, citation correctness	A - excellent.
Present your opinion to student's activity when obtaining and using study materials for thesis creation.Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.	
The student has always made reference to third party articles and software applications used for meeting the thesis assignment. All references used in the work followed the proper quality standards.	

Additional commentary and evaluation
Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.
Please insert your commentary(voluntaryevaluation).

III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

Summarize thesis aspects that swayed your final evaluation. Please present apt questions which student should answer during defense.

In this thesis, the student has proposed a new method for dealing with the web site classification problem based on the application of machine-learning techniques to the website content. The student was involved in all the different stages of building a machine learning solution. The student has placed particular emphasis on the construction of two different datasets made available to the community. Four different sets of features and three different algorithms were analyzed and carefully evaluated following the standard machine learning methodology. Finally, and perhaps the most valuable contribution of the present thesis is the resulting method was implemented in the service http://shouldiclick.org and provided for free to our society

Apt questions:

1.) It is known that the default Variable Importance measure used in the original Breiman's paper is biased towards continuous or high-cardinality categorical variables. The algorithm used for ranking the features Importance on Random Forest considered this problem?

2.) What does the student think could be the explanation behind the poor performance of feature-set-2 despite the good UMAP representation? The techniques used in UMAP representation and transformation are perhaps not valid for tree-based algorithms? Have the student tried a simpler PCA to see the discrimination power of feature-set-2 assuming linear relations?

3.) What is the rationale behind the selection of 55 bins in the graph representation approach for DOM-2 data model? Has the student tried a different number of bins?

4.) The example in Figure 5.9 shows only a small portion of the figure containing actual information. It seems that most of the cells in the matrix have low or zero information. Does the student have an idea of the distribution of the channels information along with the CEDT-DOM Dataset?

5.) According to the results provided in the thesis, the XGBoost algorithm using feature-set-2 seems to outperform Random Forest (for UWD and CEDT data models). However, from the tables, it not possible to assure such a difference is statistically significant. Has the student considered the application of some statistical tests to confirm XGBoost better performance?

6.) During the analysis of the related works, several approaches used Machine Learning for website classification (based on URL and content). Has the student considered implementing some of the approaches and testing them against data models?

I evaluate handed thesis with classification grade A - excellent.

Date:

Signature: