ČESKÉ
VYSOKÉ
UČENÍ
TECHNICKÉ
V PRAZE

REVIEWER‘S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Phishing detection using natural language processing**
Author’s name:	**Bc. Radek Starosta**
Type of thesis :	master
Faculty/Institute:	
Department:	
Thesis reviewer:	RNDr. Petr Somol, Ph.D.
Reviewer’s department:	UTIA AV CR, Avast Software

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment — challenging

Evaluation of thesis difficulty of assignment.

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.

Method of conception — outstanding

Assess that student has chosen correct approach or solution methods.

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.

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

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

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.

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.

I read Mr. Starosta‘s thesis with interest, as it touches both key aspects of modern applied AI system challenges: the choice and learning of model, and making the system useful in large-scale high-throughput realistic application scenario.

The thesis is well structured to cover the topic comprehensively. It starts with outlining the problem of Phishing in good breadth. Then it continues with description of various options how to address the problem of phishing detection, from simpler ones, to more sophisticated ones taking use of recent key advances in natural language processing (utilization of pre-trained large scale models of BERT type). All considered approaches have been implemented by author and evaluated against each other to justify the eventual decision about which way to accept as basis for an industrial system. The next chapter covers the design of classification workflow where it is assumed that the best practical results will be achievable through extensible ensembles of detectors; I agree. The workflow is then wrapped in a phishing detection engine API for use in modern cloud environment where the expected throughput is aimed at enabling detections for millions of users. Accordingly, the next chapter addresses remaining issues relevant for maintenance and orchestration of such system for high-throughput and high stability. The final content chapter is quite interesting. It addresses the problem of speed optimization not on the level of the whole system, but on the level of the actual neural inference engine, given its role as building block in the full cloud service pipeline. One particular detector is addressed here, to enable deep analysis and proposal of relatively low-level optimizations. The chapter serves as a very good case study and shows to the reader some of the finer intricacies in the high-end industrial ML systems.

The thesis is written in good English with very few glitches I could recognize. There are only minor points that I would see as worth improvement. In chapter 4 and some other chapters it would be beneficial for the reader to start with outlining the expected design goals more explicitly; as is the text does contain the information but spread throughout the text. The understanding that this is about building a real industrial system for millions of users gets unrecognized for the reader for some time while reading.

Some questions that would be interesting to answer in more detail:

- in section 4.2.3. you concatenate text extracted from elements. Have you analyzed from NLP standpoint whether this may lose some useful dicriminative information? I mean is not there value in keeping the information about concatenation points or about the structure of text? In the same section you discuss speed concerns while from preceding context it was not clear that there should be throughput limitation conditions around this. (cf. My comment above)
- It appears your work contributes to a large system where presumably a team of software engineers participates. Throughout the text you refer to a list of potential or existing detectors. Although it may not be necessary to cover the question within the scope of this thesis, but it would be interesting to understand more clearly the interplay of your detectors with possible other detectors, e.g., if any evaluation/benchmarking of their contribution to the overall system is done. In section 4.4. you mention that ensemble construction has been done semi-automatically or manually, while automated optimization failed to bring improvement. Was this due to principal problems or was not there simply time to investigate this in more depth?
- In 4.2.4 you outline further options of speeding up, one such option is re-implementing part of the system in Rust. Why this would help?
- While reading section 5.5 it appears that you intend to instantiate a large number of instances. It would be interesting for readers to appreciate the scale of this system a bit better. Also, some more details on what Datadog did better for you than open source.
- You put a lot emphasis on the production deployment and production run systems. It would be great to have also the training back-ends covered in a bit more detail.
- The topic of graph optimization in 7.3.2.1 would deserve a bit more introduction, what exactly is represented by graphs (one can guess this is about neural inference steps represented as directed graph but explaining it explicitly would help)
- in 7.3.2.3 you focus entirely on speed gains through optimizations, but it would really help if you addressed in more detail also the other side of the coin, i.e., the resulting lossess of accuracy. It is just mentioned that these are negligible, but quantifying them would give much better picture. This is probably the only point in the thesis where I had a feeling that something important is missing in the text.

Throughout the thesis the author displayed consistent care to cover all that is important with sufficient detail, while successfully striving to keep with the most up-to-date knowledge in all areas, from problem definition over the choice of prediction models and their learning to the architectural decisions and practice of building a scalable, high-throughput and well maintainable system. Bibliography is extensive and very well selected.

To summarize, the command of the problem as well as solutions presented in the thesis are very good, meaningful and stand on good up-to-date grounds. The thesis as a whole is particularly strong in the sections covering system design, benchmarking and system optimization aspects. In this sense the thesis almost reaches the quality of a would-be guidebook for correct industrial quality deployments of high-end high-throughput neural inference systems. I do recommend this thesis to be accepted as Master thesis. I recommend to rate the thesis by mark **A** – excellent.

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.

I evaluate handed thesis with classification grade A - excellent.

Date: 24.1.2021 Signature: Petr Somol