Review of the Ph.D. thesis

Author: Ing. Aliaksandr Barushka

Study programme: Applied Informatics

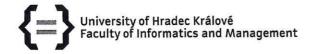**Supervisor:** doc. Ing. Petr Hájek, Ph.D.

Reviewer: doc. Ing. Vladimír Bureš, Ph.D., MBA

Doctoral thesis written by doctoral student Aliaksandr Barushka deals with the current and highly topical issue of spams and their identification and filtering. This issue is tackled with the application of machine learning methods and development of a new approach, which performance is consequently compared to other existing methods.

The thesis is well structured. It is written comprehensively in a readable fashion. Its readability is decreased only by details such as stressing significant results by underlining, which is not suitable for tables. Highlighting by bold text would definitely be a better choice. Although the work focuses on the specific technical issue, it would be understandable to readers not familiar with all details and topics in the field. The reason is that author provides a meaningful flow of his ideas and offers all issues in the order which goes from general to specific topics. Furthermore, different aspects of the thesis are clearly separated. The initial introduction into the subject is concise, however apt. It plainly describes what is going on in the field. Table 1-4 need to be highlighted as they significantly support this fact and their construction was definitely connected with a lot of effort. Applied datasets and the procedure of pre-processing are described, briefly evaluated, and explanation of their main pros and cons is provided. Application of selected methods is not associated with significant drawbacks. However, as this piece of work is a scientific document rather than a life story or a travelogue, it would be more appropriate to use the passive voice instead of the singular.

The dissertation thesis is based on analysis of references which are actual, topical and published in prestigious resources. In-text citations are correct. However, explicit justification of a need for similar research resulting from the literature review is not sufficiently highlighted. Apparently, the development of more efficient tools are methods is required at any time. Nevertheless, the reasons why the current solutions do not represent good enough solutions which are not appropriate, are not stressed convincingly. While the current state of the art is readable and exciting, it does not result in the identification of a research gap that needs to be resolved. Although it can be deduced implicitly, the explicit expression would increase research justification and meaningfulness. The main objective is adequately described. It would be better to define it once in a dedicated section (section 2). Its formulation at the end of the first section on page 14 is a little bit premature and useless. Moreover, it is not in accordance with the context provided in this part of the thesis. Similar can be stated about an associated indication of the applied methodology.

Experimental outcomes of the proposed method are compared with results acquired from the application of other methods. Unfortunately, the detail description of these methods is missing. The author provides merely general characterisation (in the theoretical part and consequently in section 7 with approx. one paragraph devoted to individual method only) without technical and formal details which would enable a real comparison and evaluation of differences among approaches.

Question:

Although Twitter and SMS datasets are associated with a huge number of messages, their average length is 17.7 and 15.6 words, respectively. To which extent are these short messages suitable for analysis from the content perspective? What are the main issues we need to consider once we decide to use them?

Conclusion:

This doctoral dissertation is based on doctoral student's publication activities, which are not numerous, however, at the high level of quality. His supervisors co-authored all published papers. This fact is a little bit contradictory. Although it proves close contact with supervisor who is a recognised expert in the field (which is good from the doctoral study programme perspective), it raises a question how the author has been (dis)abled to become an integral part of the research community and what is the level of his penetration into the existing research network. The thesis is relatively short. The text itself begins on page 12 and ends on page 94. However, a high level of quality and thoroughness of its development compensate this shortcoming. The author demonstrates his ability to work on research topics in the field as well as his competence. One has to agree with the statement of all contributions declared in section 11. Therefore, I RECOMMEND the Ph.D. thesis written by Ing. Aliaksandr Barushka to final defence. I RECOMMEND him to be granted the degree of Ph.D. in Applied Informatics.

In Hradec Králové, May 12th, 2020

Vladimír Bureš

REVIEW OF DISSERTATION THESIS

Title of Thesis: *Machine Learning Techniques in Spam Filtering*
Author: Ing. Aliaksandr B a r u s h k a
                      University of Pardubice, Faculty of Economics and Administration
Field of Dissertation: Applied Informatics
Reviewer: Prof. Dr. Ing. Miroslav Pokorný
                      VŠB-Technical University Ostrava, Faculty of Electrical
                      Engineering and Computer Science

Submitted dissertation thesis contains a total of 116 pages including publication of the author divided into eleven chapters with the introduction and conclusion. Topic of dissertation thesis is focused in the area of identification and filtering unsolicited and unwanted electronically send messages. Work objectives, contained also in the results of the author's research works and appropriate publications presented in the final part of thesis, are explicitly declared.

The sate-of-the-art is elaborated in detail in Chapter 1. The importance of spam eliminating in messages of e-mail services, SMS services, social networks and online shopping is emphasized. A thorough literature search is prepared, supplemented by clear tables. Based on the knowledge from the processing of multi-dimensional datasets, the author proposes his own solution to the problem - uses DFFNNs as base learners in more ensemble learning schemes. Summary involved in the first chapter is also important for pedagogical and next research activities.

The main goal of thesis presented in Chapter 2 is declared: to propose a spam filtering model that integrates a high-dimensional feature selection and a regularized DFFNN model with rectified linear units to capture complex features from the high-dimensional data. Next, the appropriate specific objective are defined, namely: collect and pre-process suitable spam datasets, perform high-dimensional feature selection, propose a regularized DFFNN neural model for spam filtering and finally, benchmark the proposed model against other selected existing models.

All above defied goals fully correspond to the methodology of dissertation theme solving presented in Chapter 3 and depicted in Fig.1.

To evaluate the performance of proposed spam filters in different conditions, several publicly available datasets were used namely as e-mail, SMS, social network and online hotel review domains (Capture 4).

The procedure of pre-processing (Capture 5) is focus on how text pre-processing of messages affects the performance of automated methods for spam detection. To extract both of sender-centric and message-centric text features for detecting spam several text pre-processing procedures were used to improve text mining effectiveness. Research effort of the author is focused on the role of text processing techniques over multiple spam detection domains. To benchmark the pre-processing datasets, several classification methods were applied and the results are presented in tables-form. The author's propose baseline settings appears be effective. Conclusions of pre-processing are presented and summarised. the results demonstrate the central importance of text pre-processing strategies in detecting spam / legitimate messages. The quality of the experiments performed within the pre-processing contributed to the achievement of good results of the next solution.

Deep neural networks (regularized DFFNN) functional analysis and properties is presented and appropriate scheme of DFFNN with ReL units for spam filtering is proposed in the first part of Capture 6. Such solution allows faster and effective training of deep neural architectures on large and complex datasets. It is important that, unlike the traditional machine learning algorithms, the author proposed a new solution, in with both of bag-of-words approach and linguistic word context are taken into consideration in order to enhance the performance of spam detection and to improve a high-dimensional feature representation. The proposed model is represented by a multilayer perceptron NN with one to three hidden layers. In the input layer, two sets of features were extracted from the raw message text, namely the top unigrams, bigrams and trigrams according to their tf.idf weights and average embeddings calculated for each message from the pre-trained embedding weight matrix.

Several comparative approaches developed for spam classification were applied to demonstrate the effectiveness of the proposed spam filtering model in Chapter 7. The methods were assessed in terms of their ability to deal with high-dimensional and sparse datasets. The results are expressed in Table 18 only in two-valued qualitative form (+ positive, - negative).

Important part of Capture 8 is the evaluation measures determination to evaluate the performed testing. Of particular interest is the graphical representation showing the performance of the classification model at all classification thresholds. All numerical measures are suitable for evaluating the effectiveness of the proposed model in next capture.

The final test results (Capture 9) showed that the proposed method shows very good spam detection efficiency. The proposed spam filtering models are more

computationally complex then benchmarked models and that's why more computationally intensive than the other models in terms of training time. I do not consider this one to be a significant shortcomming of solution, because the results indicates satisfactory practical data throughput regardless of data size and control domain. The very good results of the proposed method are also confirmed by the statistical evaluation of its effectiveness applied nonparametric Friedmann test. The effectiveness is confirmed also through the average accuracy obtained was compared with that of previous studies that examined the same datasets.

Regarding a future research (Capture 10) - no doubt that the proposed models have potential also in other text categorization tasks than only in cases the datasets with binary classes were used (legitimate, spam). Therefore, in my opinion, the main perspective is to extend use-case scenarios of messages filtering and to retrain the system on other large corpuses for alternative domain applications.

The evaluation of contributions of dissertation thesis theme solution in both of theoretical and practical fields (Capture 11) is correct. The results show that the proposed approach based on the ensemble methods demonstrate the best performance and outperform the state-of-the-art classification methods. The combination of complex DFFNNs trained on random subsets of pre-processed high-dimensional data seems to be an effective method for spam filtering in different spam filtering domains. It should be emphasized that the proposed and implemented solution procedure is sophisticated and demanding.

Reviewer´s questions:

1. Which quantitative classification parameters were used in Table 18 and what are their thresholds for decision if (+) ore (-)?
2. Could be used for message classification "legitimate / spam" expert knowledge, represented by linguistic fuzzy model (expert system)?

Conclusions

The structure of the work is formally and methodically correct, the scope corresponds to the requirements of the solution. From point of view of the theoretical and application aspects is the dissertation theme very interesting and actual. The content and form of his dissertation thesis demonstrates author´s professional experience, professional level and the ability of scientific work. The list of 9 author´s publications documents his sufficient publication activities. The proposed DFFNN neural model is original contribution of doctoral dissertation thesis. The core of the dissertation was published in a renowned professional scientific journals and conference papers. The text of dissertation thesis is very good structured and it is

very well readable. The presented dissertation thesis is of very good quality also in point of linguistic and pedagogical view. Number of references (188) is high.

I consider presented dissertation work at a very high quality in terms of its content and its form as well. I suppose that submitted dissertation thesis of Mr. Ing. Aliaksandr Barushka demonstrates his good orientation in the field of information theory and applications.

Taking into account I recommend the dissertation thesis of Mr. Ing. Aliaksandr Barushka for defence. I recommend further that - after passing the successful defence of the dissertation thesis in front of the doctoral committee – to award Mr. Ing. Aliaksandr B a r u s h k a - in accordance with Act No. 111/98 Coll. of the Higher Education and the relevant dissertation prescriptions of the University of Pardubice and its Faculty of Economics and Administration – the title of

D o c t o r, PhD.,

in the accredited field of Applied Informatics.

Ostrava, May 20, 2020   Prof. Dr. Ing. Miroslav Pokorný