Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Daryna Dementieva

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Methods for fighting with harmful multilingual textual content

Supervisor: Associate Professor Alexander Panchenko

Name of the Reviewer:

I confirm the absence of any conflict of interest Ivan Oseledets	**Date: 15-09-2022**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense. *If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.*
Reviewer's Report

The thesis of Daryna Dementieva has 228 pages, 2 big parts (methods for fake news detection and methods for text detoxification). The first part has 3 chapters, the second has 5 parts. The thesis also includes introduction, background and two supplementaries. The results are published in 8 papers as a main authors and 3 papers as a coauthor.

This is an impressive and very productive work in terms not only the quantity (i.e., the number of publications is amazing) but also quality: the papers were published in top conferences in the field.

The overall topic of AI4SG Is challenging - it is clear, that the same technology can be used for all of the malicious purposes, but it is nice to see that the author indeed tries to use AI for good purposes, which is a dangerous ground in some sense. The risks are nicely highlighted in the introduction part.

Comments.

- I would like to see technical challenges in the introduction discussed more (besides the grand challenges). Overall, the goal of a dissertation is to do research which can be put in form of algorithms, methods and even theory.
- The result for the fake news detection is explicitly stated in the beginning, which is nice
- The picture 3.1 describes Ukrainian and Russian media, whereas it would be good to add, i.e., EU, US and China sources of different type, because in this example the difference is in one word.
- Chapter 3 ends without a conclusion.
- In social media, one of the channels of transmitting fake are «bot farms». I think, that the information about the connection between users could be useful for fake news detection, not only the texts in different languages.
- In multilingual similarity, how recent multilingual models will help? What the role of better pretrained models in the similarity?
- Can we fine-tune the backbone models as it is done In metric learning for example?
- Chapter 7 really describes the dataset collection in details. It would be very useful for NLP practioners: how to use Yandex.Toloka, how to setup data collection, etc. Very nice.
- Chapter 8 presents a conditional Bert for detoxification. The idea is to mask words according to the toxicity of the word. Does the size of the model play a role? Is it better to have a pretrained model at hand?
- I have a question regarding (8.2), which seems to be important. What are the meaning of individual terms in the product? Are those probabilities, or log-probabilities? I.e., product typically corresponds to probabilities of independent events, whereas the sum - to the logarithm. So, explanation why this combination of individual scores is given would be useful
- Table 8.1: Why BART zero-shot gets so low value of J? But being good at SIM_a
- Table 8.3 with cherry-picked examples is very nice.
- Chapter 8 partially answers the question about the model size, but not clear why in English different model sizes are not considered (in Chapter 7), but the discussion is moved to Chapter 8.

Overall, this is a very solid modern-NLP work. Its main contributions are efficient data engineering techniques, combined with new loss functions and models, as well as extensive comparison. It would be great if that later some algorithmic/theoretical insights will grow from the results and datasets collected in this work, but we will see it. The comments above do not influence the overall very high evaluation of the thesis work and the amount of results and efforts put into it. This is one of the best theses in Skoltech I have seen so far, and the author could be proud of her work.

Provisional Recommendation

I recommend that the candidate should defend the thesis by means of a formal thesis defense

I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report

The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense