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: Assistant Professor Alexander Panchenko

Name of the Reviewer: Natalia Lukashevich

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 18-09-2022**

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

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

The summary of issues to be addressed before/during the thesis defense

The Ph.D. thesis of Ms. Daryna Dementieva «Methods for Fighting with Harmful Multilingual Textual Content» considers two problems of currently existing harmful text content: automatic detection of fake news and methods for text detoxification. The thesis contains 2 preliminary chapters (Introduction and Background), two task-related parts and conclusion. The first part "Methods for fake news detection" contains three chapters and the second part – Methods for Texts Detoxification includes four chapters.

The candidate has 8 papers as a main author including 2 publications in Core A* conference, other papers are indexed in SCOPUS.

In introduction, tasks of restricting harmful text content are formulated and research questions are stated. In Chapter 2, the author describes a specific subdomain of artificial Intelligence and natural language processing concerning social good goals. Besides, an overview of available models for multilingual natural language processing is given.
Chapter 3 opens discussion about approaches to fake news detection; the chapter contains the description of existing fake news detection datasets and systems. In Chapter 4, the author introduces the proposed method for fake news detection based on cross-lingual search. The main idea of the approach is as follows: a fake news has less support in the international press. The proposed system is tested on fake COVID news dataset. It was shown that the feature based on cross-lingual news similarity yields significant improvements in fake news detection. Chapter 5 is devoted to study of methods detecting cross-lingual news similarity. It was proposed to use named entities as an additional signal to estimate multilingual news text similarity.
The Part II is devoted to text detoxification approaches. It begins from Chapter 6 introducing the task, which goal is to change a toxic message to more neutral, preserving the content if possible. In the chapter the author classifies the detoxification task as a style transfer task and surveys methods of style transfer. This technique can be used as a prompt to a user or for control of messages generated by chat-bots. Chapter 7 describes a new dataset of parallel sentences for training detoxification models - ParaDetox for English and Russian languages. Chapter 8 presents two approaches to detoxification: a new method for unsupervised text style transfer and EN-Detox and RU-Detox – monolingual detoxification models trained on the created parallel detoxification corpora. Also, the task of cross-lingual style transfer is considered. In Chapter 9, the problem of disagreement of human and automatic scores in evaluation of detoxification methods is considered.

I have only one comment.
All fact checking experiments are based on well-known fake stories collected in specialized datasets. But for new fake messages, most evidence including cross-lingual evidence is absent, therefore the proposed models will work much worse than it is presented in current studies.

To conclude, the thesis is well-structured and well-written. Daryna Dementieva proposes new methods for fact checking and automatic detoxification. She created new datasets for training detoxification models.

I recommend to proceed with the public defence. The contribution of Daryna Dementieva deserves awarding the author with the Ph.D. degrees of Skolkovo Institute of Physics and Technology.

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*