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: Mikhail Burtsev

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

Reviewer's Report
The thesis "Methods for fighting with harmful multilingual textual content" by Daryna Dementieva is well written and has solid overall structure that fits requirements for this type of work. The title of the thesis correctly reflects the main idea of research presented in the main text which consists of studies for problems of fake news detection and texts detoxification. To address the task of fake news detection the thesis explores a hypothesis that fake news have lower tendency to spread globally due to filtering by journalists and news editors on the international level. This allows to use a multilingual presence of some news as a feature for a fake prediction. To test the hypothesis a methodology and a system to collect the cross-lingual evidence for a specific news were developed. The evidence was collected and manually verified on a small scale. After the pilot confirmation the large-scale automatic evaluation on well known datasets demonstrated that addition of proposed cross-lingual feature significantly improves quality of fake news detection over a range of approaches and results in new state of the art performance. This is an important outcome of the study that lays ground for highly demanded fake news detection applications. The second part of the thesis is focused on the problem of texts detoxification. It is proposed to approach detoxification by rewriting the source text trained to change style from toxic to neutral with either a pipeline based on unsupervised masked language modelling or sequence to sequence transformer. For training and evaluation of implemented systems the author collects parallel dataset of toxic\neutral phrases called ParaDetox in English and Russian. The presented condBERT model for unsupervised selective correction of rough texts demonstrated reasonable quality. The sequence-to-sequence training on ParaDetox resulted in EN-Detox and RU-Detox models which demonstrate state of the art results. This confirms the initial hypothesis of the thesis that training on parallel corpus should improve performance

on detoxification task. EN-Detox and RU-Detox models can be successfully applied to many real-life use cases some of which considered in the thesis in a greater detail.

All research results included in the thesis were presented on leading peer-reviewed NLP conferences or workshops and published in conference proceedings.

There are a few questions I would like to be addressed during the thesis defense.

1. In the Section "4.3.3 Automatic Fake News Detection" on pages p. 65-66 there is no description on figures figures 4-4, 4-5, 4-6 how a score for "All ling." method was calculated. What was a method to calculate this score?
2. Results section (5.7, p.80-86) of the chapter on multilingual text news similarity metrics includes results for a lot of outdated ML methods rarely used in NLP applications today. These methods include linear regression, SVR, decision trees, gradient boosting, etc. Quite expectedly all these methods show poor performance. Why were they selected in the first place?
3. Statement - "Such good performance of the models based on Transformer-based embeddings can be explained with the origin of the data. The datasets that we use for comparison are all originally in English." (p.86) is unclear. Is there something specific in transformer architecture related to English language? To the best of my knowledge transformer architecture is language agnostic.

These questions do not affect the overall high quality of the work. I deeply believe that the thesis is of high quality and Daryna Dementieva clearly demonstrated research skills to be qualified for PhD degree.

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*