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: Paolo Rosso**

| | |
|---|---|
| I confirm the absence of any conflict of interest<br><br>(Alternatively, Reviewer can formulate a possible conflict) | **Date: 19-09-2022** |

*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 thesis addresses the important problem of harmful textual content from a multilingual perspective, both in English and Russian. The PhD candidate addressed several topics: from the detection of propaganda (a lot during the special operation/invasion/war of Russia in Ukraine), and fake news (what is a fake news? Mentioning words such as invesion or war?), to the detection of toxic language and detoxification (papers on the last topic was published at the top EMNLP-2021 and ACL-2022 conferences). The work done in the framework of this PhD has been already validated being published in several good venues such as the above CORE A/A* conferences (by the way, I'm not sure if also the student workshops at ACL 2021 ans 2022 can be considered as CORE A* publications). Moreover, the PhD candidate co-organised also the RUSSE-2022 shared task on Russian text detoxification based on parallel corpora with Russian toxic messages from Odnoklassniki, Pikabu and Twitter platforms.

In the part on fake news detection, the PhD candidate proposed Multiverse, a new feature that is based on cross-lingual evidence extracted from a multilingual search. It showed to improve the performance of the model in comparison with only monolingual evidence. With respect to toxic language and its detoxification, she introduced Paradetox, a new parallel dataset composed of toxic, non toxic pairs. Moreover, she proposed condBERT in order to address detoxification from an unsupervised perspective, both in English (EN-Detox model) and Russian (RU-Detox model).

The PhD is very interesting although unfortunately the PhD candidate did not have the time to proof-read what she wrote in some parts and the manuscript has plenty of typos (but I understand that in these difficult times it's a matter of priorities and finishing the PhD the fastes as possible in order to start the postdoc was priority number one). The PhD manuscript will be a public document that will be published online and it would be nice to fix at least some of them. Below a list of just some of the typos. A couple of minor comments about the enumeration of pages: often Roman numbers (i, ii, ...) are used for the preliminary parts (acknowledgements, abstract, index, list of figures and tables); every chapter could start in an odd page on the right (a blank page could be left at the end of the previous chapter if necessary).

As future work, at page 176 "more hidden types of toxic language such as sarcasm or passive aggressiveness" (e.g. hate can be conveyed with the use of stereotypes) are mentioned. Maybe the PhD candidate could be interested in having a look at some recent works on these topics (also about the usage of ethnophaulism, e.g. “хохол” for the Ukrainian) and eventually adding them among the references .

Frenda S., Cignarella A., Basile V., Bosco C., Patti V., Rosso P. (2022) The Unbearable Hurtfulness of Sarcasm. In: Expert Systems with Applications (ESWA), vol. 193

Frenda S., Patti V., Rosso P. (2022). Killing me Softly: Creative and Cognitive Aspects of Implicitness in Abusive Language Online. In: Natural Language Engineering (JNLE), pp.1-22 doi:10.1017/S1351324922000316

Sánchez-Junquera J., Chulvi B., Rosso P., Ponzetto S. (2021). How Do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants. In: Applied Science, 11(8)

Pronoza E., Panicheva P., Koltsova O., Rosso P. (2021) Detecting Ethnicity-targeted Hate Speech in Russian

Social Media Texts. In: Information Processing & Management, 58(6)

Prof. Paolo Rosso

Universitat Politècnica de València, Spain

Some of the typos (in upper case the words that should be added)

page 2: texts that has -> have

2: we test THIS new feature

2: the usage... improve+s

2: demostration OF how THE proposed

3: foe -> for

3: After THE parallel dataset creation

3: to be extend+ed

7: but also IN life

8: in my -> my? head

22: used to for?

22: become A quite popular platform

23: a lot OF work is done

23: addresses THE following research questions

24: After new multilingual news similarity system selection -> please rephrase it

25: can be re-used of any other text style transfer task -> possibly to rephrase it

25: two version+s

25: This method address+es

25: a replacement of AN exact toxic part

25: we provide A detailed description of THE evaluation

26: new state-of-the-art model+s

26: its -> THIS kind OF exploration

26: for THE detoxification task

38: development OF THE multilingual model

38: the follows -> the following ?

38: We present A new multilingual feature

40: A lot of system+s

41: that predicts A class

41: In out -> our work

41: to extend THE usual definition

41: the+y rely on

42: several works has -> have

42: combat received by ??

42: explicit emphasize -> emphasis?

42: Also, A different perspective

43: datasets that includes -> include

43: In this dataset THE authors

43: contain only data has been ...or has received -> contain only data that have been...or have received

44: one main limitations -> limitation

44: such datasets can be mentioned -> the following datasets can be mentioned

44: statistics also illustrates -> illustrate

44: difficulties of collection -> collecting

44: and THEmplementation OF AN algorithm

45: gram-ars

45: trained an SVM

46: the set OF emotions that ARE present

46: that A trusted news

46: With THE recent growth

46: in THE NLP field

46: as THE COVID-19 fake news detection task

46: easy in -> to use

46: of THE model performance

47: There was created THE dEFEND system

47: THE Factual...system

47: Talking about THE information verification step in THE fake news detection pipeline

47: One of the source+s

49: are the follows -> are the following ?

49: The code of THE proposed method

90: We have no intense -> intent?

172: domain dataset+s

172: acheiving AN extremely high performance

172: we provided A task formulation of THE fake news classification task

172: on -> an overview

173: in THE TransformerCosSim

173: we provided A motivation of -> FOR THE detoxification task

173: A formal problem statement

173: for THE style transfer task

173: One of the reason+s

173: to solve THE detoxofocation task

173: TO edit text

173: of A parallel dataset

174: We introduced A new

174: outperformed THE baselines

174: For THE English language

174: For THE Russian language

174: of its kind+s OF approaches

174: how THE proposed detoxification

174: in THE game industry

174: of THE detoxification task evaluation

174: That showed, that -> That showed that

175: accroding -> according

175: has A good perspective

175: on THE seq2seq approach

175: but THE cross-lingual model

175: While this dissertation propose+s
175: not all language variety -> varieties is -> are

175: but A very important accomplishment

176: While there is A multilingual version

176: to create A parallel detoxification

176: Also, an additional experiments -> Also, additional experiments

176: As the problem of toxicity dataset available -> As the problem of the availability of a toxicity dataset

176: solution ON/ABOUT how

176: we coverede with obvious toxicity types that is ?

176: done A more accurate description OF

176: can be THE implementation

176: in A more tolerant

176: with THE addition of

176: The -> Some examples and existed datasets

177: in THE detoxification

177: in which THE model is unsure

177: for text generation:

177: For instance: THE model

177: to make step for -> to make a step forward?

**Provisional Recommendation**

X *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*