
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

(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 invasion 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 fastest 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. “xoxox” for the Ukrainian) and eventually adding them among the references.


Pronoza E., Panicheva P., Koltsova O., Rosso P. (2021) Detecting Ethnicity-targeted Hate Speech in Russian
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
we provide a detailed description of the evaluation of new state-of-the-art models for the detoxification task. The development of the multilingual model follows this kind of exploration. We present a new multilingual feature that predicts a class in our work to extend the usual definition they rely on. Several works have combat received by explicit emphasis? Also, a different perspective. Datasets that include contain only data that have been or have received one main limitation of such datasets can be mentioned the following datasets can be mentioned. Statistics also illustrate difficulties of collection and the implementation of an algorithm. We trained an SVM.
the set of emotions that are present
that a trusted news
With the recent growth
in the NLP field
as the COVID-19 fake news detection task
easy in -> to use
of the model performance
There was created the DEFEND system
THE Factual... system
Talking about the information verification step in the fake news detection pipeline
One of the source+s
are the follows -> are the following?
The code of the proposed method
We have no intense -> intent?
domain dataset+s
achieving an extremely high performance
we provided a task formulation of the fake news classification task
on -> an overview
in the TransformerCosSim
we provided a motivation of -> for the detoxification task
A formal problem statement
for the style transfer task
One of the reasons
to solve the detoxification task
TO edit text
of a parallel dataset
We introduced a new approach that outperformed the baselines for the English and Russian languages. The proposed detoxification in the game industry has a good perspective on the seq2seq approach but the cross-lingual model has a very important accomplishment. While there is a multilingual version, not all language varieties are covered with obvious toxicity types. We created a more tolerant solution on the availability of a toxicity dataset. Also, additional experiments showed that the problem of the availability of a toxicity dataset can be improved with the addition of.
Some examples and existed datasets in the detoxification in which the model is unsure for text generation. For instance: the model to make a 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