Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Daryna Dementieva

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Methods for Fighting Harmful Multilingual Textual Content

Supervisor: Assistant Prof. Alexander Panchenko

The thesis document includes the following changes in answer to the external review process.

The title of the work was slightly fixed: from METHODS FOR FIGHTING WITH HARMFUL MULTILINGUAL TEXTUAL CONTENT to METHODS FOR FIGHTING HARMFUL MULTILINGUAL TEXTUAL CONTENT.

In the Section "4.3.3 Automatic Fake News Detection" on pages p. 65-66 there is no description on figures 4-4, 4-5, 4-6 how a score for "All ling." method was calculated
The description of All ling. features are added to the baselines description section (Section 4.3.3, Baselines).

5.7, p.80-86 outdated methods are reported
The explanation of the usage of classical ML regression models is added into the text in Section . The main reasons are the model's lightweight and possibility to gain explanation of models' decisions.

"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) unclear
This Section 5.8 was reconsidered and the results were reformulated. We did the more fair comparison between fake news detection systems and the new proposed metric.

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.
The discussion of the limitations of the proposed method in terms of time delay is added into the Summary (Section 4.4).

Table 5.9 Cap)on: 0.95% confidence intervals or 95% confidence intervals?
The typo was fixed.

How do you think your method based on cross-lingual relevance be extended to the social media domain? A discussion on this would be useful.
The discussion about this further research step is added into the Summary (Section 4.4).

Rewrite into: "Warning: this part contains texts with rude, obscene texts only for example illustra)on. We have no intent to offend the reader."
The typo was fixed.

For the style transfer task, do you think the recent prompt engineering based approaches could be useful? A discussion on this would be useful.
Indeed, the prompt engineering methods can be useful for cross-lingual style transfer. The discussion about this further step in cross-lingual style transfer research is added into the Summary (Section 8.7)

Include technical challenges in the introduction (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.
While the Introduction section was not changed, the more detailed explanation of the methods used in the work was added into Background Chapter under Section 2.2. We hope that this will add more formalization to the work. Besides, for each part – fake news and detoxification – the introduction chapters (Chapter 3 and Chapter 6) formally define the corresponding tasks.

Figure 3-1 – unclear example.
The example of news presented in the picture was changed.

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
In Section 8.2.1, there was added a detailed explanation of the formula and the meaning of each term in it. The meaning of the whole metric is also explicitly described.

Table 8.1: Why BART zero-shot gets so low value of J? But being good at SIM_a
The explanation of such behavior of the BART-zero-shot model was added to Section 8.3.4.

Publications – incorrect ranking
The rankings of publications were cross-checked.

The new cases of toxicity such as racism and sexism can be considered as future directions.
In Section 10.2.2, these types of toxicity were added with corresponding publications.

Additionally, the whole text was cross-checked in terms of typos, misspellings, and visibility of inappropriate language.

Addressing the comments about the manuscript structure:
- The Glossary was moved to the beginning of the text after the table of the content;
- The Glossary, List of Figures, and List of Tables were added to the Table of Contents;
- The warning about obscene words in the examples was moved to the second page of the manuscript;
- The long titles of sections were placed so they do not overlap with Chapter names.