Negotiating language attitudes on TikTok - computational challenges

Taras Andrushko, University of Oslo. 2023

Main Idea – using NLP methods, extract attitudes towards the language issue in war conditions and identify the main arguments.

Main Challenges – data extraction, distinguishing UKR from RUS, data preprocessing, dealing with short comments and mixed data.

Workflow:

1. Collecting comments using a crawler written in Python
2. Filtering comments that contain more than 5 words and likes.
3. Separating UKR from RUS using the Langid module
4. Comment preprocessing (NLTK)
5. Topic modeling
6. Argument mining

10.20378/irb-93378

Two datasets

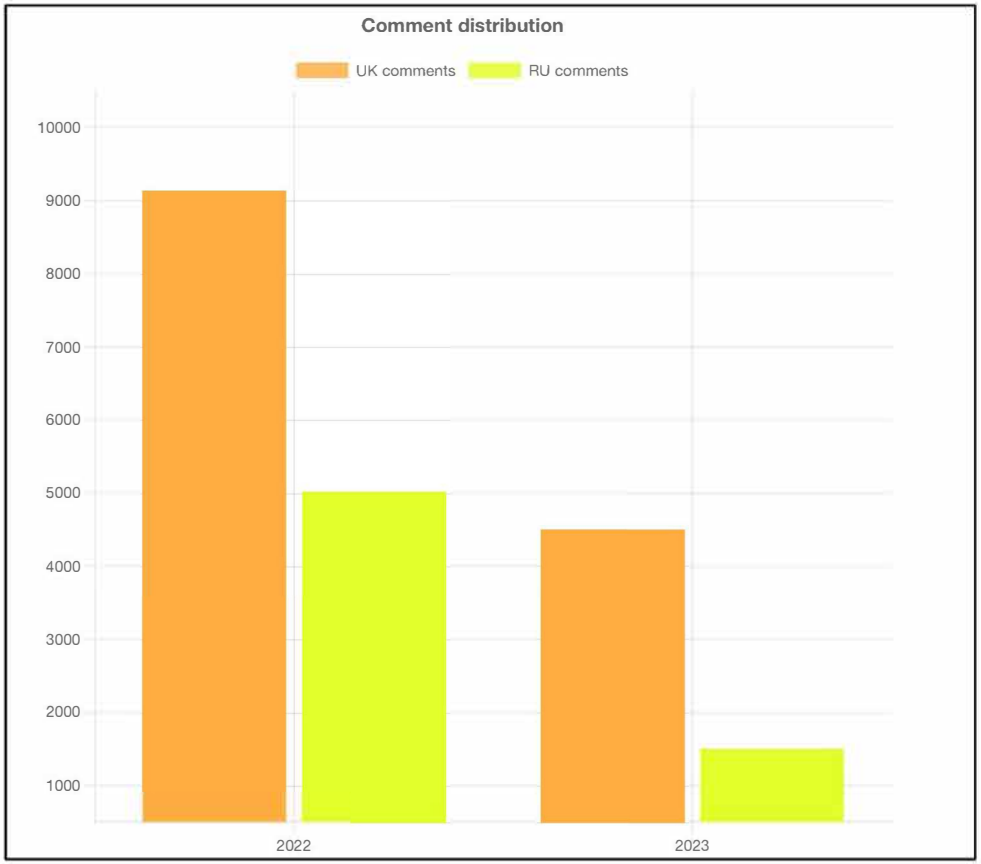

Figure 1. Number and distribution of comments collected in 2023 and 2022

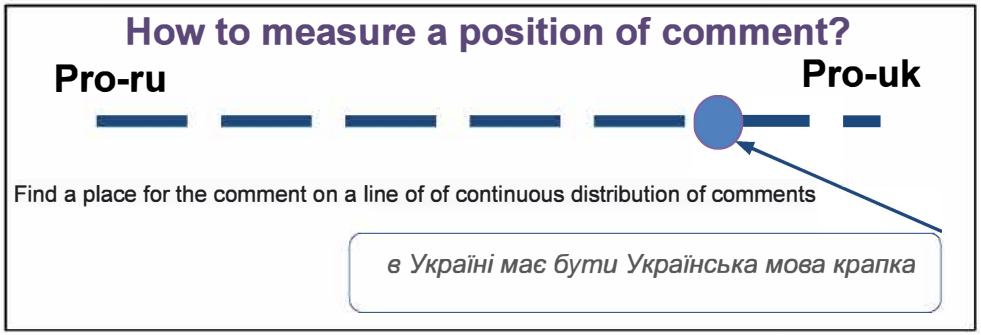

Figure 2. Concept of a line of of continuous distribution of comments

Results:

- Topics that appear most often in comments are extracted (“Language unites”, “language divides”, “UKR only”, “Doesn't matter”, etc.)
- A gold standard corpus has been compiled for further training of the argument mining model

Prospects

- Further training of the machine to distinguish UKR from RUS. This task is complicated by the fact that there is no sensitive division between Ukrainian and Russian, since there is a continuum between them.(Possible approaches: rule based, neural networks.)
- Using the golden corpus, train the model to predict numerical proximity for a comment on the bipolar scale (From **Pro-ru** to **Pro-uk** Figure 2)

References

Racek, Daniel et al. “The Politics of Language Choice: How the Russian-Ukrainian War Influences Ukrainians’ Language Use on Twitter.” Institute of Statistics, Ludwig-Maximilians-University, 2023. https://arxiv.org/pdf/2305.02770v3.pdf.