Deep learning approaches to detect disinformation across news platforms and social media

NATO OTAN TIDE Hackathon Warsaw, Poland 20 - 24 February 2023

AIM OF THE RESEARCH

The aim of this research is to detect and analyse disinformation in digital media, particularly focusing on critical global issues like the Russo-Ukrainian war. For this, machine learning techniques including LDA, BERT, and RoBERTa were employed, using a dataset of news articles and social media messages.

DATASET

- The disinformation detection dataset encompassed 505 news articles and 624,982 social media messages, curated by POL Cyber Command and further refined at the NATO TIDE 2023 Hackathon.
- It included 141 articles from outlets like *Sputnik*, *TASS*, and *NewsFront*, with 108 flagged as containing false information.
- Text pre-processing included the removal of non-English content, excessively brief messages, and elements like mentions, tags, and emojis.

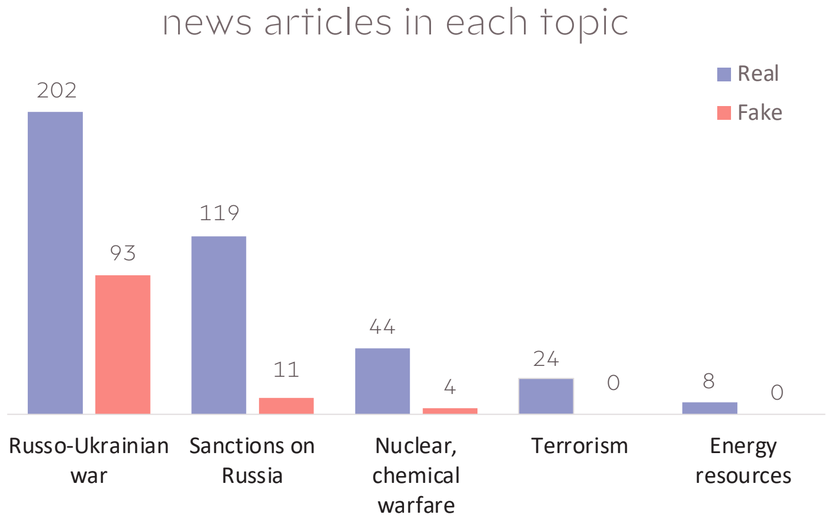
Type of text	Total items
Fake news articles	108
Factual news articles	397
Social messages	624,982

DATA ANALYSIS

Data analysis in our study employed text-based methods to pinpoint disinformation trends in collected news articles and social media messages.

- Employed LDA for Topic Modelling to categorize content in news articles.
- Noted high disinformation within the Russo-Ukrainian war-related articles.
- No disinformation detected in topics concerning NATO-aligned issues like energy and terrorism.
- Social media analysis indicated a wider range of less distinct topics, demanding different disinformation detection methods.

The distributions of fake and real news articles in each topic



DEEP LEARNING FOR DISINFORMATION DETECTION

We used BERT model architecture, specifically DistilBERT and RoBERTa, for detecting disinformation in news articles due to their effectiveness in understanding textual context and pre-training on fake news classification. Two selected specific models were RoBERTa-fake-news and DistilBERT for fake news detection.

We split this collected dataset into 85% for training and 15% for testing. The results of the models without fine-tuning were:

Model	Accuracy	F1-score
DistilBERT	0.79	0,72
RoBERTa	0.74	0,51

After fine-tuning the models with the 85% of the collected dataset we got the results:

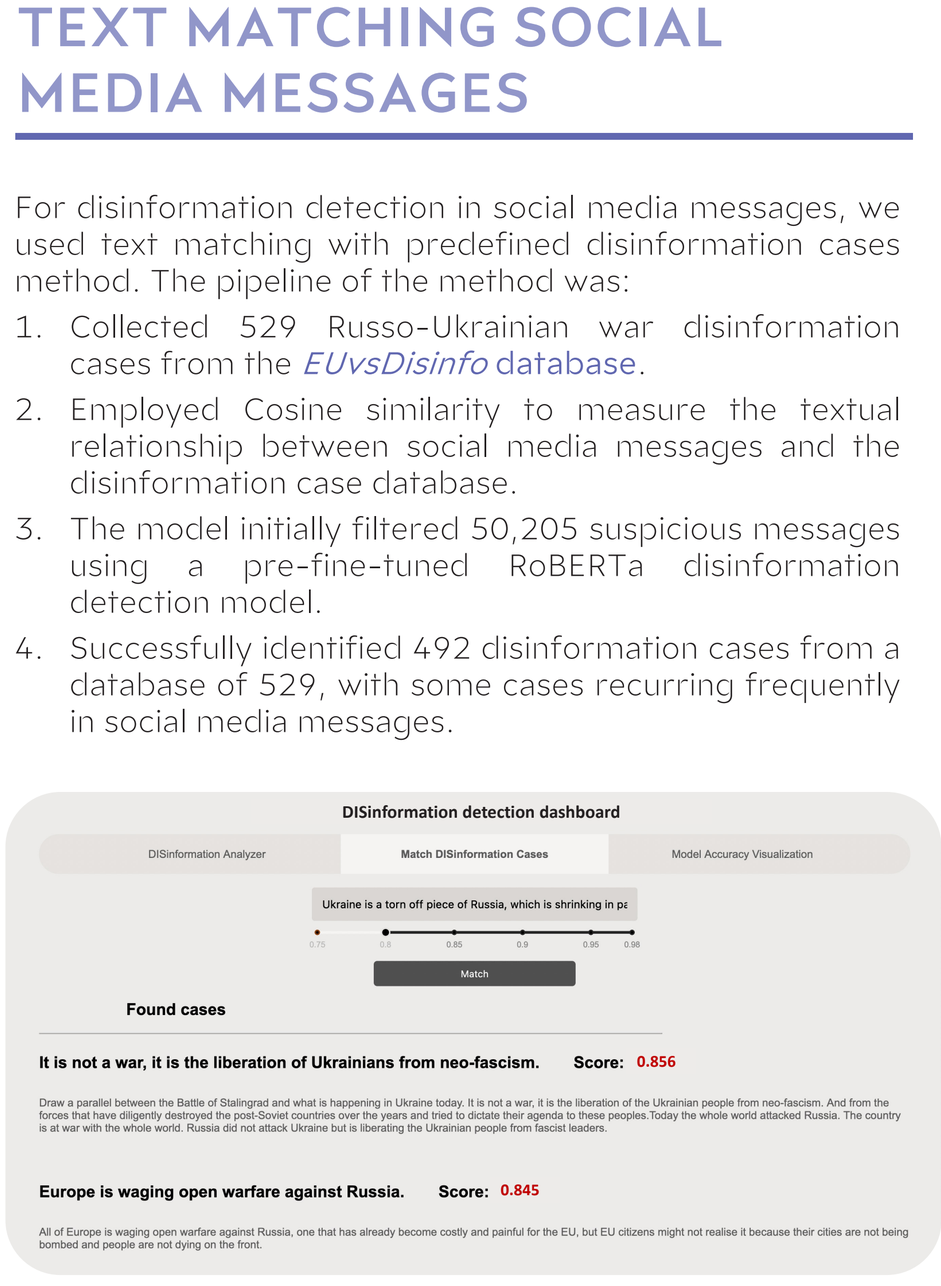
Model	Accuracy	F1-score
DistilBERT	0.96	0,95
RoBERTa	0.99	0,98

Finally, both models demonstrated a marked increase in accuracy and other metrics. However, RoBERTa model showed better overall results.

TEXT MATCHING SOCIAL MEDIA MESSAGES

For disinformation detection in social media messages, we used text matching with predefined disinformation cases method. The pipeline of the method was:

1. Collected 529 Russo-Ukrainian war disinformation cases from the *EUvsDisinfo* database.
2. Employed Cosine similarity to measure the textual relationship between social media messages and the disinformation case database.
3. The model initially filtered 50,205 suspicious messages using a pre-fine-tuned RoBERTa disinformation detection model.
4. Successfully identified 492 disinformation cases from a database of 529, with some cases recurring frequently in social media messages.



CONCLUSIONS

- BERT and RoBERTa models, were identified as the most effective methods for disinformation detection.
- Initially the selected deep learning methods scored less than 75% accuracy in classifying disinformation.
- After additional fine-tuning, both BERT models significantly improved, achieving over 95% accuracy in classifying disinformation.

FUTURE PLANS

- Expand research by training several complex language models on larger and more diverse datasets to enhance disinformation detection across various topics.
- Develop a Lithuanian language model for disinformation detection, addressing the lack of large language models for Lithuanian language and its relevance in local news and social media contexts.

AUTHORS:

Milita Songailaitė
milita.songailaite@vdu.lt

Justina Mandravickaitė
justina.mandravickaite@vdu.lt

Eglė Rimkienė
egle.rimkiene@vdu.lt

Anton Volčok
anton.volcok@vdu.lt

Tomas Krilavičius
tomas.krilavicius@vdu.lt

VYTAUTO DIDŽIOJO UNIVERSITETAS
Informatikos fakultetas

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT