Instant Messaging Platforms News Multi-task Classification For Attitude, Sentiment, And Discrimination Detection

Denilson Barbosa[1], **Taras Ustyianovych**[2]
[1]Department of Computing Science, University of Alberta, Canada,

[2]Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Ukraine

[1]denilson@ualberta.ca, [2]taras.o.ustyianovych@lpnu.ua

Introduction

Background: the Russian aggression against Ukraine has notably intensified online discourse on this subject. This surge in online activity inevitably results in the spread of content, some of which may be unreliable or manipulative. The identification of such content with information distortion is crucial to mitigate bias and promote trustworthiness.

Objectives: improve the efficiency of distinguishing between stance/attitude, and determining the sentiment of messages related to such events as the Russian war in Ukraine.

Importance: from a practical perspective, our piece of research can significantly contribute to cognitive security and cyber hygiene purposes. In terms of Ukrainian NLP, it would be an advancement to accurately classify geopolitical attitudes and sentiment in text data with puzzling language usage and linked to complex topics.

Related Works

- Investigation of the dynamics of online activity associated with significant geopolitical events, highlighting the capacity of strategic communication efforts to engage with and influence digital communities (Courchesne et al., 2022).
- A recent study by Park et al. (2022) describes the VoynaSlov dataset that was collected from two social networks, Twitter and VKontakte, to analyze and detect media opinion manipulations related to the Russian war in Ukraine. It consists of more than 38 million posts based on Russian media statements and expressions.
- Fedushko et al. (2023) proposed innovative methods to support real-time decision-making about antagonistic user behavior on social networks. The proposed method helps to moderate and reduce destructive content in an online community.
- The HQP dataset (Maarouf et al., 2023) specifically collected to facilitate the identification of misinformation by incorporating 30 thousand tweets related to the war between Russia and Ukraine. The dataset consists of high-quality labels obtained through human review, ensuring the trustworthiness and accuracy of the data. The trained models have a Area Under the Curve (AUC) score of 92.25.
- Exploration of various computational approaches to address challenges with the Ukrainian language, including classification (Solopova et al., 2023), text summarization (Galeshchuk, 2023), and topic modeling (Ustyianovych et al., 2023), particularly when dealing with low-resource languages (Gomez et al., 2023).

Related Work

Gaps we address:

- Telegram as a platform has a distinct user base and communication style, which significantly differ from other platforms. Its channels offer a rich amount of data in varied tones - ranging from news and factual reports to blog posts and opinion pieces.
- The multi-task model based on NLP transformers would provide cost-efficient and required insights on a given input message without the need to apply multiple single-task models.
- Scarcity of publicly available labeled data related to the Russian war in Ukraine.

Telegram War News Dataset

Dataset collection

- The dataset has been collected from thoroughly selected pro-Russian and pro-Ukrainian Telegram channels.
- The selection of channels is based on the lists of sources provided by the Ukrainian Center for Countering Disinformation and the Institute of Mass Information.
- The total number of messages the research is based on equals to 252,677. Data was collected from January 2022 until December 2023.
- At this point, we continue the data collection, enhancing our processes and expanding the number of channels and features.

Channel	Stance	Count	Fraction	Mean token count	Mode sentiment
rian_ru	Pro-Russian	79,663	28.83%	28.26	neutral
ROSSIYA_SEGODNIA	Pro-Russian	69,238	25.05%	55.16	negative
uniannet	Pro-Ukrainian	67,727	24.51%	48.58	negative
radiosvoboda	Pro-Ukrainian	33,225	12.02%	108.63	negative
UkrPravdaMainNews	Pro-Ukrainian	22,860	8.27%	46.34	negative
ZE_kartel	Pro-Russian	3,601	1.30%	74.91	negative

Table 1: Number and percentage of messages per channel

Telegram War News Dataset

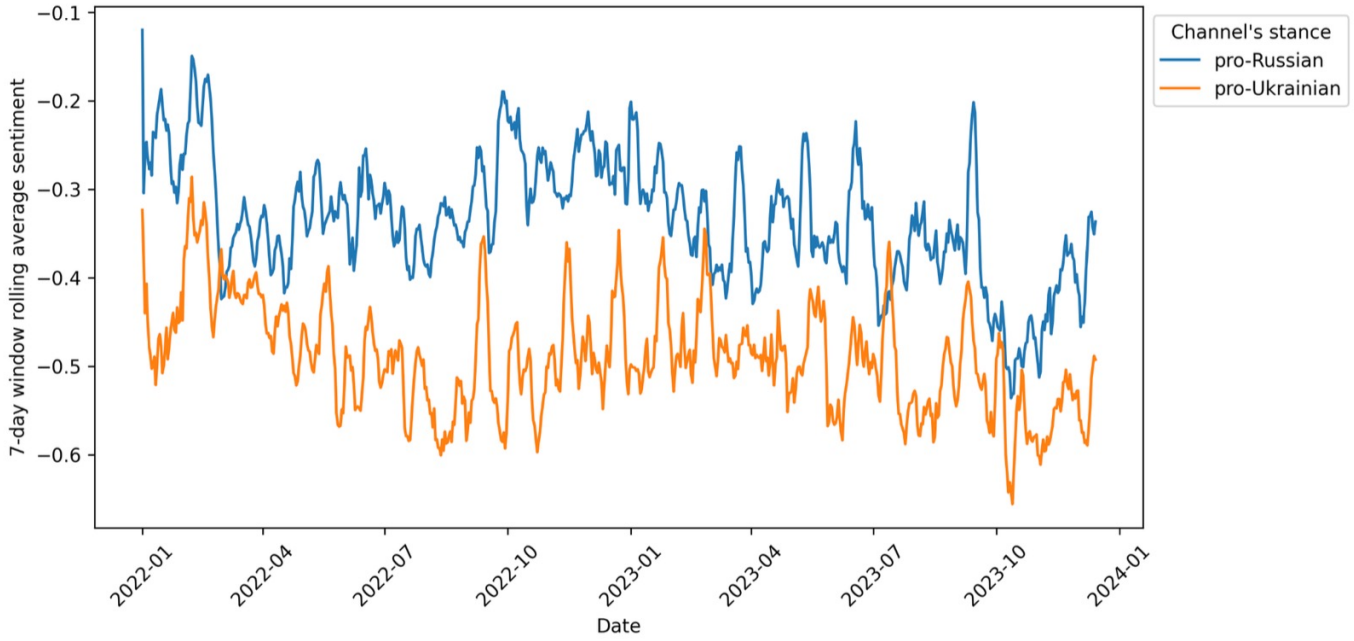
Dataset statistics

- 55.19% of the collected messages are retrieved from pro-Russian sources, whereas 44.80% posts belong to pro-Ukrainian channels.
- Less than 10% of messages were in the Ukrainian language when performing this research. Nevertheless, at this point we managed to achieve a 60-40 percents balance in favor of the Ukrainian language. Messages written in Russian are crucial for this piece of research to make the final solution as versatile as possible.
- The mean and median token count after text preprocessing are 29.52 and 21 respectively, and the standard deviation is 29. This indicates most publications are quite short and highlight mostly the crucial details.
- Majority of messages exhibit a negative or neutral sentiment (more on this in the next slides)

Telegram War News Dataset

Sentiment analysis

- GPT-3.5 AI-agent was utilized for sentiment classification.
- The aggregated sentiment scores represent such patterns:
  - Both pro-Ukrainian and pro-Russian channels inherit a negative sentiment over the analyzed period
  - Sentiment of pro-Ukrainian messages exhibit less variance compared to the pro-Russian counterparts
  - The sentiment values fluctuate over time suggesting external events affecting the web communication.
- GPT-3.5 model tends to interpret themes of war and conflict with a negative sentiment despite the evidence that some messages might be perceived differently by specific users. The used AI agent may not fully capture the context of war-related communications.



Telegram War News Dataset

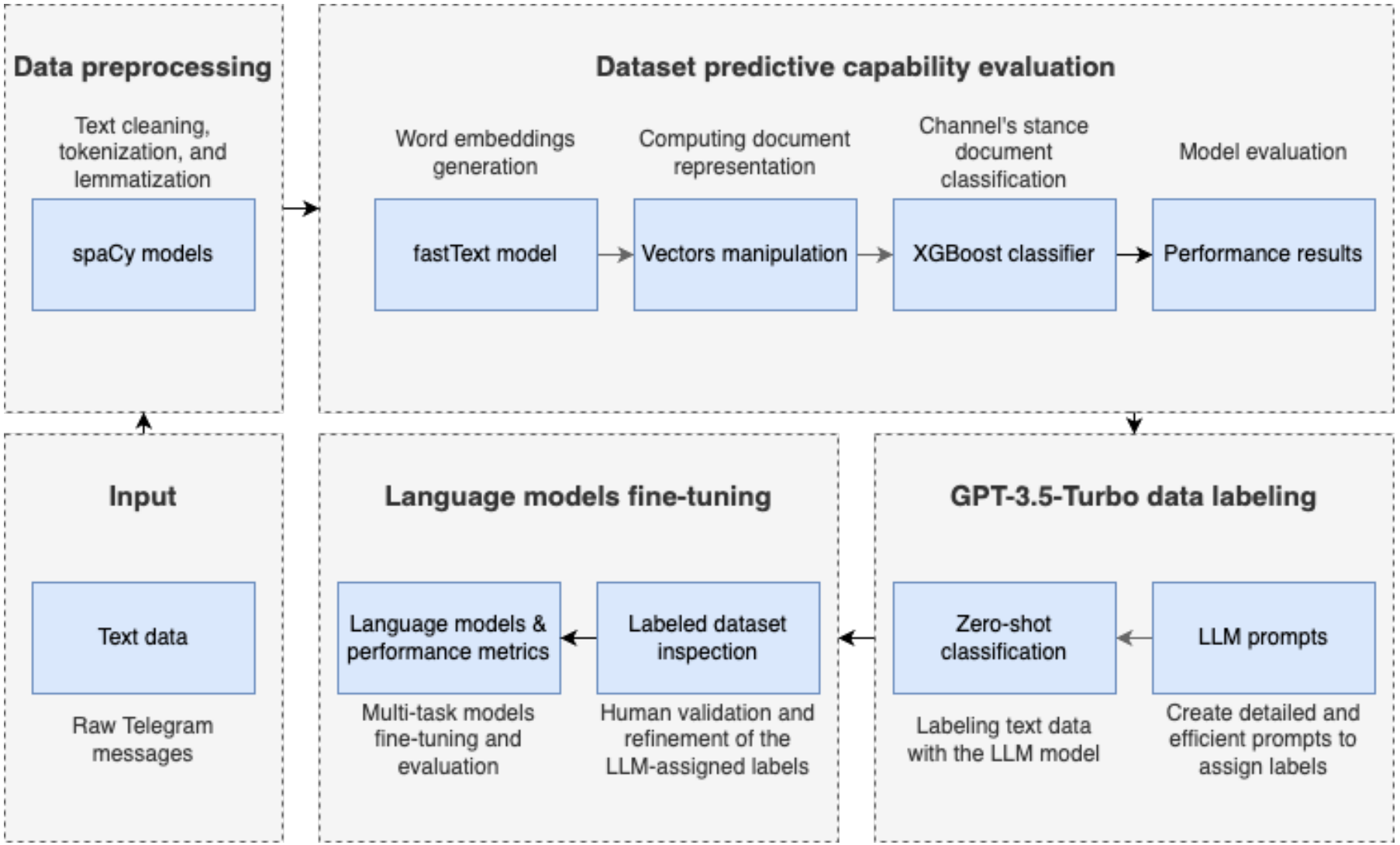
Challenges

- New patterns form constantly. Data/model/concept drifts are likely to occur. MLOps and incremental learning shall be used to assist.
- Scarcity of labeled data.
- Validation of labels assigned by the AI-model.
- Excluding irrelevant messages (not related to the subject area)

Limitations

- Lack of pro-Russian messages written in Ukrainian language.
- Low-resource NLP capabilities of Ukrainian language compared to English.

Data Processing



Data Processing
Instruction to the AI agent

The following prompt instructions were provided to the AI agent:

Analyze the following messages related to the war between Ukraine and Russia. For each message:

1. Determine the sentiment (positive, negative, neutral, etc.) expressed in the message.
2. Identify geopolitical attitude or hate/discrimination and in favor of what side it is expressed: indicate whether it's pro-Ukrainian, pro-Russian, or any other geopolitical stance. Take into account that messages might contain glorification, hate, and discrimination, which should be considered when classifying attitudes.
3. If the message lacks a geopolitical attitude or isn't related to the conflict, mark it as not applicable to geopolitical attitude.

The output should be returned as a Python dictionary array with such keys: message ID, sentiment, detected favorable attitude, and whether a message contains hate or discrimination (yes or no)

Text classification

- The developed multi-task model can distinguish between pro-Ukrainian and pro-Russian messages, determine sentiment, geopolitical stance, presence of discrimination.
- **fastText** word embeddings and **XGBoost classifier** were used to **validate the predictive capabilities** of the dataset. This approach allowed us to achieve 0.92 AUC score when predicting the originating channel's stance.
- The **google/mt5-base** model was used as the final LLM for fine-tuning on the abovementioned task.
- Tokenization restricted the text input to a length of 256 tokens. The training process spanned 10 epochs with batch sizes of 64 for both training and evaluation.
- The multi-task model displayed **above-mediocre performance** with an average accuracy of 0.74. It effectively identified the **originating channel's attitude** with a high accuracy of 0.95 and **detected the presence of discrimination** with an accuracy of 0.94. However, when the model was tasked with simultaneous detection of channel attitude and sentiment, the accuracy slightly reduced to 0.67, and further to 0.51 for combining geopolitical stance and sentiment.

Conclusion and Future Work

- Introduced TRWU dataset, comprising text data from diverse channels on the Russian aggression against Ukraine.
- Built a data pipeline to:
  - process the text
  - assess predictive capabilities
  - perform labeling using zero-shot classification and labels validation
  - fine-tune a multi-task text classification model
- Future directions to work on:
  - advanced stance/sentiment classification with rigorous labeling
  - context-based entity sentiment analysis, especially in conflict-related discourse
  - recognize unknown patterns in online communication
  - pursuing excellence in model performance for both multi-task and single-task objectives

Areas for improvement and practical application

- Data labeling in a single-task manner with iterative & systematic human validation
- Analyzing the user reactions in the form of emojis along with the text data
- Improvement in model performance
- Entity-based sentiment detection

- This study methodology and approaches have been used to develop a solution for text analytics and manipulative content classification that is currently utilized hands-on.

Selected references

1. L. Courchesne, B. Rasikh, B. McQuinn, and C. Buntain. 2022. Powered by twitter? the taliban's takeover of Afghanistan. ESOC Working Paper 30, Emperical Studies of Conflict.
2. C.Y. Park, J. Mendelsohn, A. Field, and Yu. Tsvetkov. 2022. Challenges and opportunities in information manipulation detection: An examination of wartime Russian media. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5209–5235, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
3. S. Fedushko, K. Molodetska, and Yu. Syerov. 2023. Decision-making approaches in the antagonistic digital communication of the online communities users. Social Network Analysis and Mining.
4. A. Maarouf, D. Bär, D. Geissler, and S. Feuerriegel. 2023. HQP: A human-annotated dataset for detecting online propaganda.
5. Veronika Solopova, Christoph Benzmüller, and Tim Landgraf. 2023. The evolution of pro-kremlin propaganda from a machine learning and linguistics perspective. In Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 40–48, Dubrovnik, Croatia. Association for Computational Linguistics.
6. Svitlana Galeshchuk. 2023b. Abstractive summarization for the Ukrainian language: Multi-task learning with hromadske.ua news dataset. In Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 49–53, Dubrovnik, Croatia. Association for Computational Linguistics.
7. T. Ustyianovych, N. Kasianchuk, H. Falfushynska, S. Fedushko, and E. Siemens. 2023. Dynamic topic modelling of online discussions on the Russian war in Ukraine. In Proceedings of International Conference on Applied Innovation in IT, pages 81–89.
8. Frank Gomez, Alla Rozovskaya, and Dan Roth. 2023. A low-resource approach to the grammatical error correction of Ukrainian. In Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 114–120.