Department of Computer Science & Engineering

Detecting Manipulation in Ukrainian Telegram: A Transformer-Based Approach to Technique Classification and Span Identification

Authors: Md. Abdur Rahman and Md Ashiqur Rahman

Paper ID: 44

Department of Computer Science and Engineering,
Southeast University, Dhaka, Bangladesh

Content

Introduction

- The Russia-Ukraine war has intensified information warfare, turning social media platforms like Telegram into critical battlegrounds.
- Telegram is a breeding ground for channels spreading misleading information, Russian-favorable narratives, and falsehoods against Ukrainian interests.
- Detecting these subtle manipulation techniques is an urgent security concern to combat disinformation, protect public consensus, and ensure information integrity.

Challenges

- **Nuance of Manipulation:** Techniques are not just "fake news" but include subtle tactics like loaded language, whataboutism, and emotional appeals, which are hard for models to distinguish.
- **Dual-Task Complexity:** Our work addresses two distinct but related tasks:
  1. Technique Classification: What manipulation is being used?
  2. Span Identification: Exactly where in the text is it?
- **Linguistic Richness:** The dataset contains Ukrainian and Russian, morphologically complex Slavic languages, which poses challenges for tokenization and contextual understanding.
- **Data Imbalance:** Some manipulation techniques are far more common than others, making it difficult to train a model that performs well on rare classes.

Contributions

- Investigation of ML, DL, and transformer-based models. [1]
- Our fine-tuned Transformer-based system like XLM-RoBERTa-Lrge [3] and mDeBERTa [4] achieved competitive results in the UNLP 2025 Shared Task: 3rd Place in Technique Classification and 2nd Place in Span Identification
- We provide a detailed error analysis that offers crucial insights into model performance on Slavic languages and the specific challenges of manipulation detection.

Task & Dataset Description

Task 1: Technique Classification
Objective: Assign one or more of 10 pre-defined manipulation labels to a text.
Metric: Macro F1-Score

Task 2: Span Identification
Objective: Pinpoint the exact start and end character indices of manipulative text.
Metric: Span F1-Score

- A corpus of Ukrainian and Russian Telegram posts provided by Texty.org.ua. [2]

Split	Instances
Train	3,248
Validation	574
Test	5,735
Total Words	805,730
Unique Words	146,410

Table 1: Instance distribution across data splits and dataset word counts.

Proposed Methodology

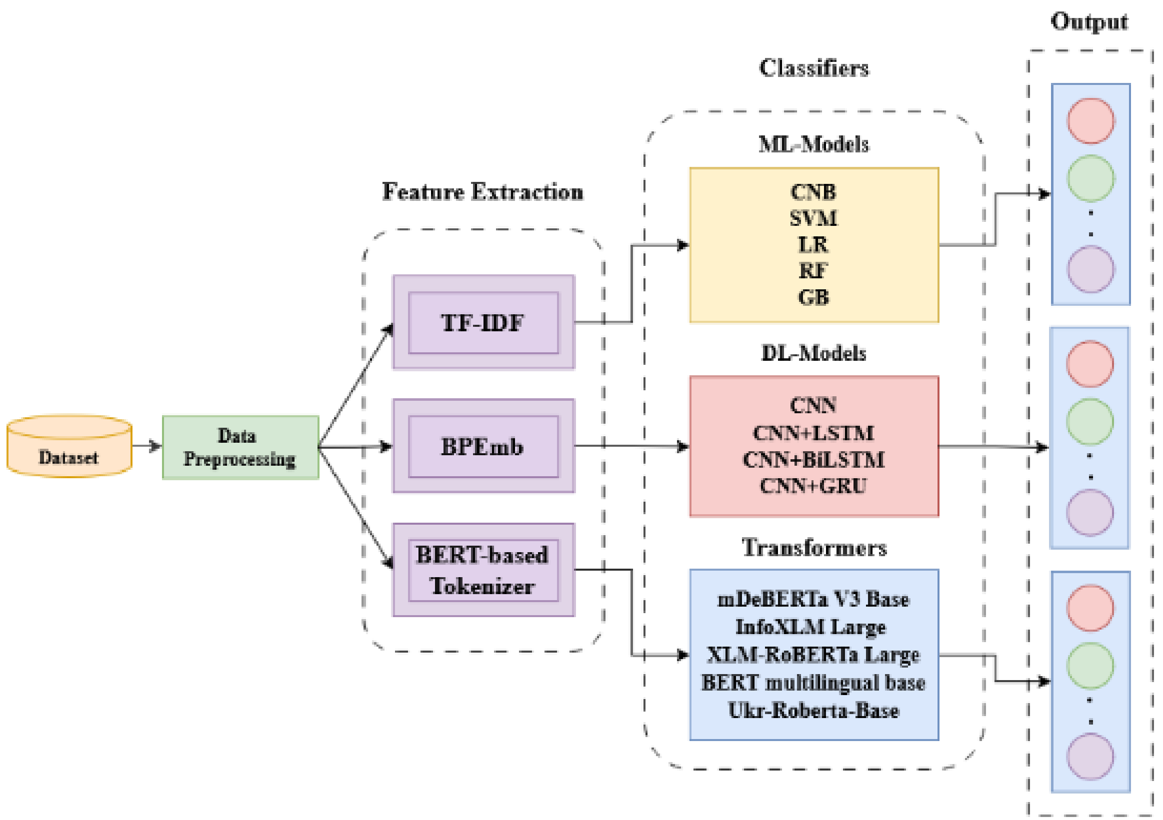

Figure 1: Schematic process for Manipulation Technique Classification

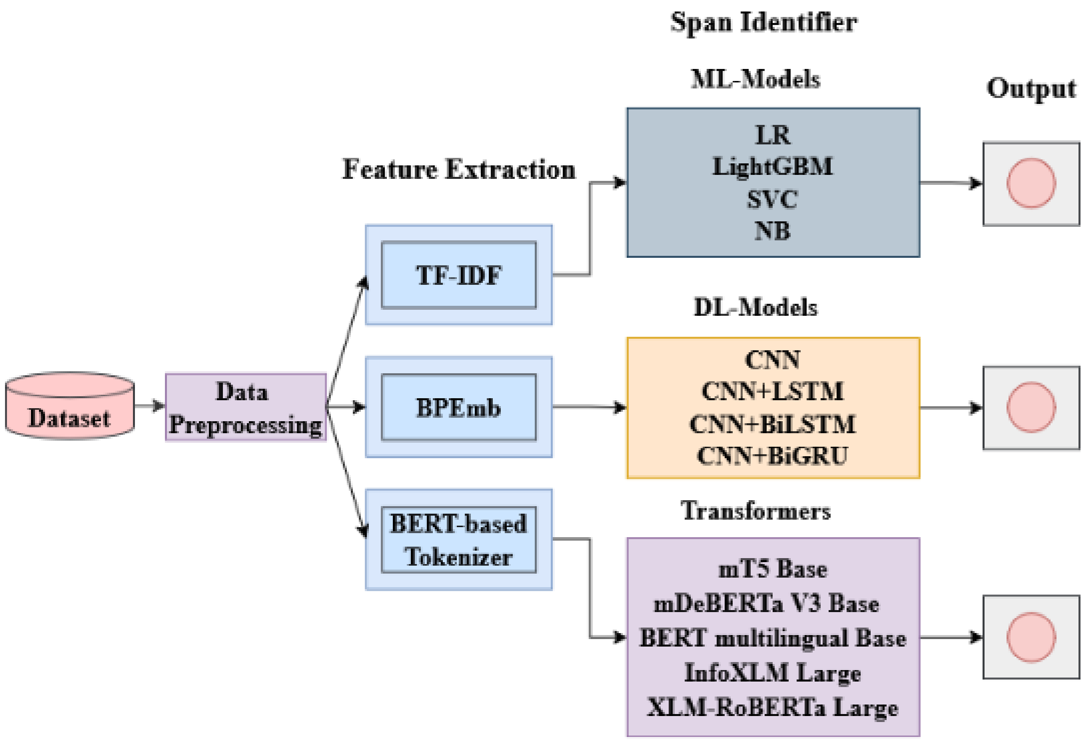

Figure 2: Schematic process for Manipulative Span Identification

Results and Analysis

Classifier	Precision	Recall	F1 Score
Technique Classification			
ML Models			
LinearSVC	0.3543	0.2878	0.3102
CNB	0.2680	0.2818	0.2553
LR	0.2807	0.5433	0.3291
RF	**0.5688**	0.1060	0.1309
GB	0.3926	0.1423	0.1846
DL Models			
CNN	0.2991	0.3287	0.2816
CNN+LSTM	0.3125	0.3388	0.3077
CNN+BiLSTM	0.3403	0.3443	0.3252
CNN+GRU	0.3649	0.3087	0.3179
Transformers			
mDeBERTa V3 Base	0.3453	0.5055	0.3901
InfoXLM Large	0.3855	0.5477	0.4451
XLM-RoBERTa-large	0.3917	**0.5667**	**0.4498**
BERT multilingual base	0.3710	0.3930	0.3772
Ukr-Roberta-Base	0.3687	0.4366	0.3660

Classifier	Precision	Recall	F1 Score
Span Identification			
ML Models			
LinearSVC	0.4020	0.3921	0.3970
LR	0.4169	0.3578	0.3851
MNB	0.4169	0.3578	0.3851
lightGBM	0.3599	0.4794	0.4112
DL Models			
CNN	0.2596	0.8715	0.4001
CNN+LSTM	0.2566	**0.9187**	0.4012
CNN+BiLSTM	0.2878	0.8126	0.4251
CNN+BiGRU	0.2949	0.8023	0.4313
Transformers			
infoXLM-large	0.5646	0.5510	0.5577
mDeBERTa-v3-base	**0.6367**	0.4644	0.5371
XLM-RoBERTa-large	0.5616	0.6500	**0.6026**
BERT-base-multilingual	0.5188	0.5697	0.5431
mt5-base	0.3930	0.6645	0.4939

Table 5: Performance Comparison of ML, DL, and Transformer Models for both tasks

Error Analysis (Quantitative)

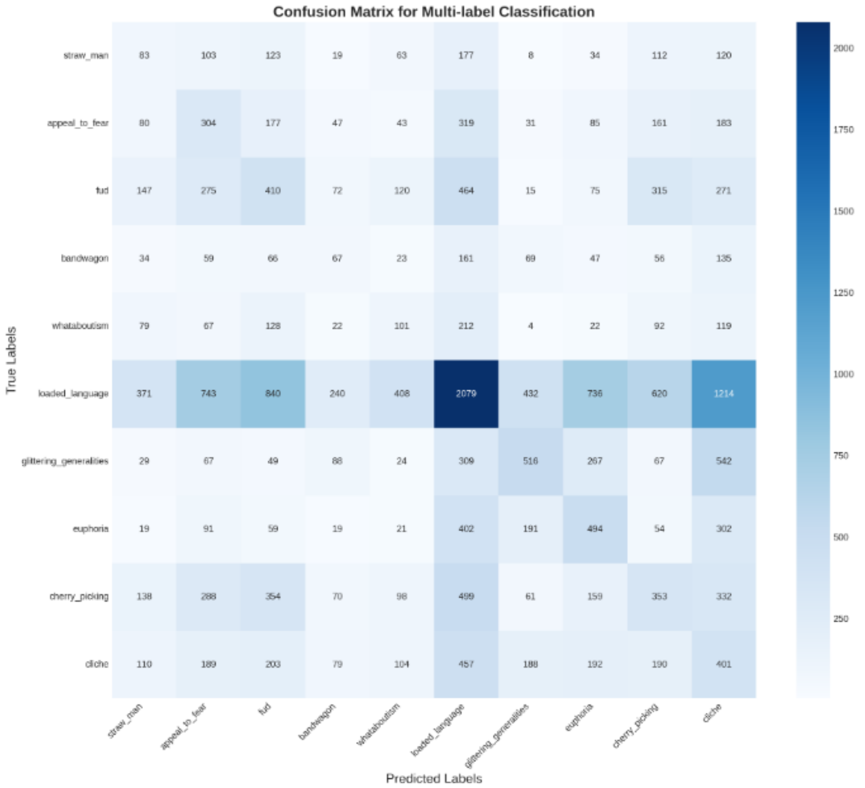

Figure 3: Confusion matrix of XLM-RoBERTa large

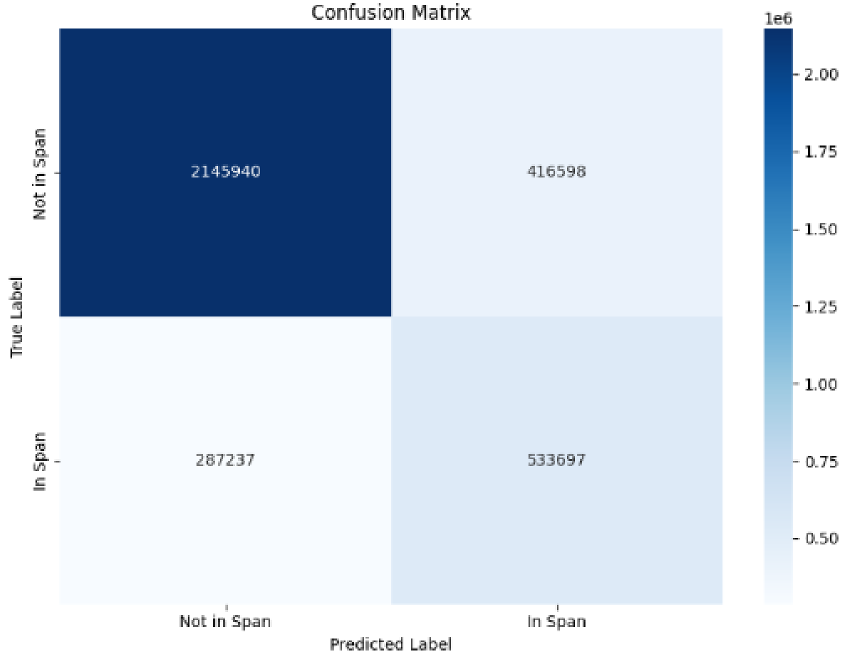

Figure 4: Confusion matrix of the proposed model (fine-tuned XLM-RoBERTa large) for span identification

- ❑ The model excels on common tactics (Loaded_Language) but struggles with rare ones (Straw_Man, Bandwagon). Significant off-diagonal errors show confusion between related techniques (e.g., FUD and Appeal_to_Fear)
- ❑ High False Positives show model tends to over-predict span boundaries, tagging neutral words near manipulative text.

Error Analysis (Qualitative)

Content	Actual Label	Predicted Label
Соловйов, стервятник пропаганди | Реконструкція правди | Віталій Портников https://youtu.be/kB4Kq3yqiXY	Loaded Language	Loaded Language
В Черновцах укроживотные -могилизаторы похитили велосипедиста ... очередной доброволец уехал на фронт...	Appeal_to_fear, loaded_language	Appeal_to_fear, fud, loaded_language
Депутаты Рады, кажется, саму малость без интереса слушают первое выступление нового министра обороны 😁	Loaded_language, cherry_picking	Fud, Whataboutism, Loaded_language, cherry_picking

Figure 5: Few examples of predictions produced by the proposed XLM-R Large model on the technique classification task

Content	Actual Span	Predicted Span
Юзернейм. Если ты радуешься пожару на Новочеокасской ГРЭС - ты расчеловечиваешь электричество. Помни!	[(0, 101)]	[(1, 4), (10, 101)]
Русская весна плавно перейдёт в русское лето и весь Донбасс вернётся домой. Этого мы ждём всей душой.	[(0, 74), (76, 100)]	[(0, 101)]
Сподіваюсь усі зрозуміли хто така русня, а то до цього часу Ізраїль намагався на двох стільцях всидіти.	[(0, 103)]	[(0, 103)]
Соловйов, стервятник пропаганди | Реконструкція правди | Віталій Портников	[(0, 31)]	[(0, 31)]

Figure 6: Few examples of predictions produced by the proposed XLM-R Large model on the span identification task

- ❑ The model struggles with technique ambiguity, often predicting extra, related labels.
- ❑ The model frequently makes boundary errors, merging or splitting manipulative spans.

Limitations

- Reliability is low for rare techniques like whataboutism and straw_man due to insufficient training examples.
- The model struggles to precisely identify start/end points in morphologically complex Slavic languages, often resulting in overextended or merged spans.
- Techniques with similar rhetorical purposes (e.g., loaded language, appeal to fear, and FUD) are frequently confused.
- The model was validated only on Telegram data; its performance on other social media platforms or propaganda styles is unknown.

Future Works

- Employ synthetic data augmentation and weighted loss functions to improve performance on rare manipulation classes.
- Implement boundary-aware architectures and targeted post-processing to refine span predictions and reduce boundary errors.
- Use contrastive learning to explicitly train the model to distinguish between semantically similar manipulation tactics.
- Develop custom tokenization and embeddings to better handle code-mixing and dialectical variations present in real-world data.

Conclusion

- We presented a robust system for detecting manipulation in Ukrainian and Russian Telegram posts, achieving top-3 performance in the UNLP 2025 shared task.
- Transformer-based models, especially XLM-ROBERTa-large, proved highly effective, demonstrating the power of large, pre-trained multilingual models for this domain.
- Key challenges remain in distinguishing fine-grained techniques and precisely identifying span boundaries, highlighting areas for future research.
- This work represents a significant step toward developing automated tools to combat information warfare in critical socio-political contexts.

References

[1] https://github.com/borhanitrash/Detecting-Manipulation-in-Ukrainian-Telegram

[2] https://github.com/unlp-workshop/unlp-2025-shared-task/tree/main/data

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettle- moyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. CoRR, abs/1911.02116.

[4] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient disentangled embedding sharing. Preprint, arXiv:2111.09543.

Thank You