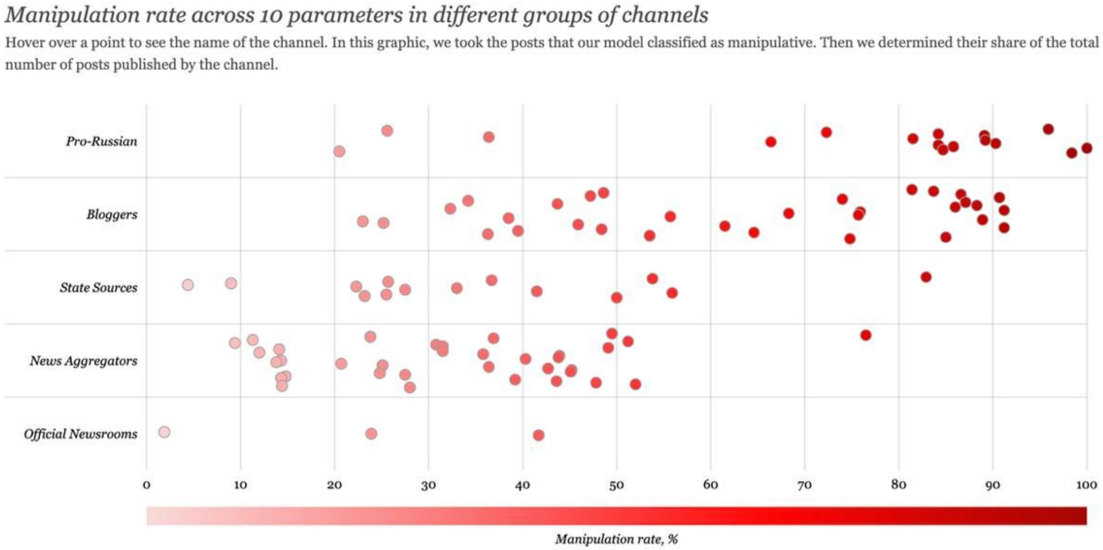
The UNLP 2025 Shared Task on Detecting Social Media Manipulation

Roman Kyslyi, Nataliia Romanyshyn, Volodymyr Sydorskyi

Why Detect Social Media Manipulation?

- Rising influence of disinformation on platforms like Telegram, especially in Ukraine.
- Need for automated tools to support media literacy and counter propaganda.
- Ukrainian is a low resource language, so it lacks dedicated models



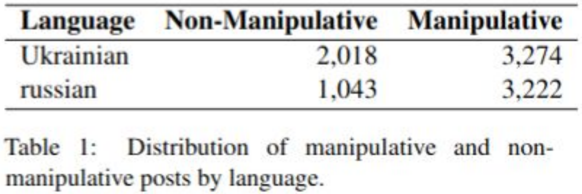
Manipulation rate across 10 parameters in different groups of channels

Hover over a point to see the name of the channel. In this graphic, we took the posts that our model classified as manipulative. Then we determined their share of the total number of posts published by the channel.

Source: https://texty.org.ua/projects/113693/carousel-of-emotions-manipulation-level-of-ukrainian-telegram-channels/

Data

- 9,557 Telegram posts from Ukrainian news and political blog channels on Telegram,
- Included both manipulative and non-manipulative posts
- License: CC BY-NC-SA 4.0 License
- The list of manipulation techniques was compiled by Texty.org.ua:
    - Get initial labels from Detector Media
    - Discussion with focus groups of Ukrainian journalists, editors, and media analysts to resolve hard cases:
        - Decide which rhetorical patterns should be considered manipulation
        - Distinguish manipulations that may be acceptable during the active phase of the war
        - Identify the techniques viewed as most destructive on Ukrainian Telegram



Language	Non-Manipulative	Manipulative
Ukrainian	2,018	3,274
russian	1,043	3,222

Table 1: Distribution of manipulative and non-manipulative posts by language.

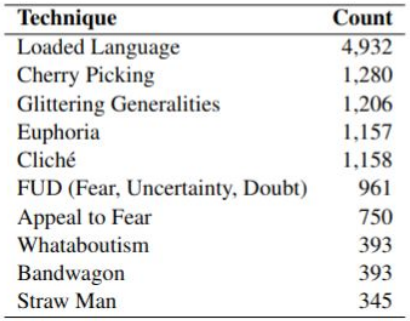
Shared Task Tracks

- **Technique Classification:** identifying the specific manipulation techniques employed within a given text.
- **Span Identification:** locating the exact spans of text that constitute manipulative content, irrespective of the technique used.

- Input texts were reused across both competitions but targets were different

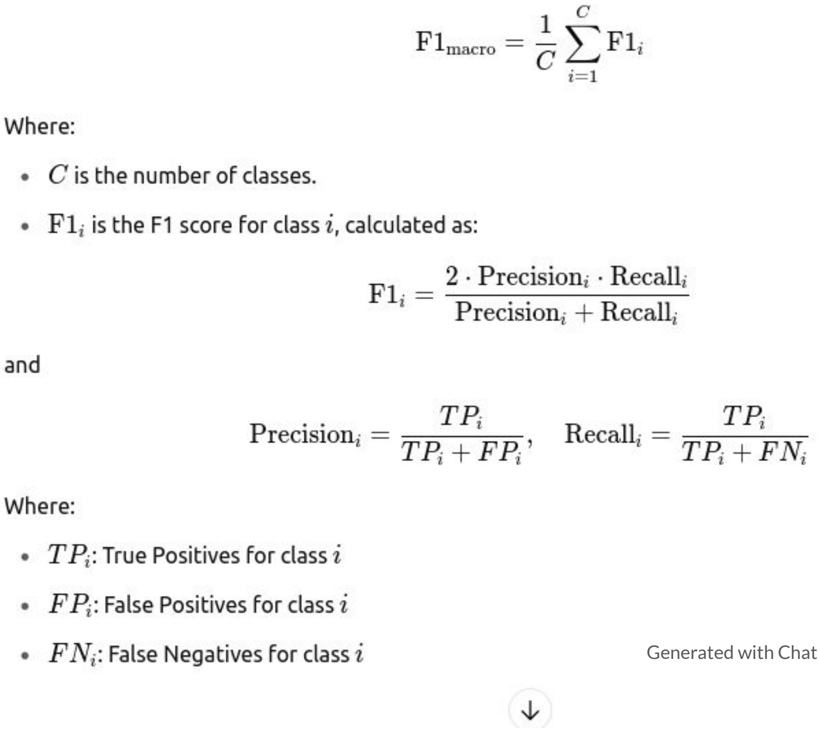
Technique classification

- Multilabel task
- Labeled on the post level
- Evaluated by F1 Macro



Technique	Count
Loaded Language	4,932
Cherry Picking	1,280
Glittering Generalities	1,206
Euphoria	1,157
Cliché	1,158
FUD (Fear, Uncertainty, Doubt)	961
Appeal to Fear	750
Whataboutism	393
Bandwagon	393
Straw Man	345

Table 2: Frequency of manipulation techniques (a post may contain multiple techniques).



$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C}\sum_{i=1}^{C}\mathrm{F1}_i$$

Where:

- C is the number of classes.
- $\mathrm{F1}_i$ is the F1 score for class i, calculated as:

$$\mathrm{F1}_i = \frac{2 \cdot \mathrm{Precision}_i \cdot \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$$

and

$$\mathrm{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \quad \mathrm{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

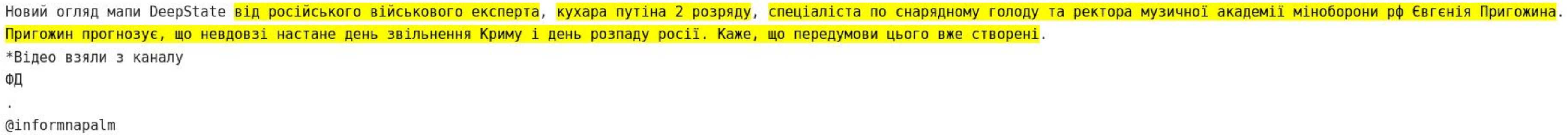
Where:

- TP_i: True Positives for class i
- FP_i: False Positives for class i
- FN_i: False Negatives for class i

Generated with ChatGPT

Span Identification

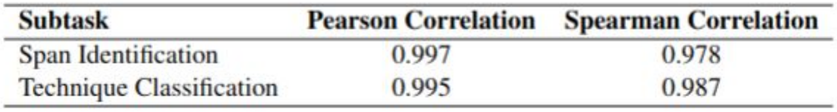
- NER type task but in binary format
- Metric Token Level F1 - in order to reduce sensitivity to missed punctuation, white spaces and errors in several characters



```
Новий огляд мапи DeepState від російського військового експерта, кухара путіна 2 розряду, спеціаліста по снарядному голоду та ректора музичної академії міноборони рф Євгєнія Пригожина.
Пригожин прогнозує, що невдовзі настане день звільнення Криму і день розпаду росії. Каже, що передумови цього вже створені.
*Відео взяли з каналу
ФД
.
@informnapalm
```

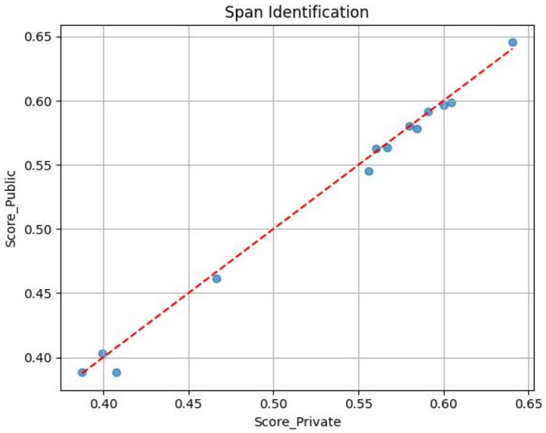
Data Split

- Performed using Multilabel Stratification algorithm based on manipulation techniques
- Split into 5 Folds (each 20% of all data):
    - **Training set:** 3822 samples (1 and 2 Folds)
    - **Private test set:** 3824 samples (3 and 4 Folds)
    - **Public test set:** 1911 samples (Fold 5)

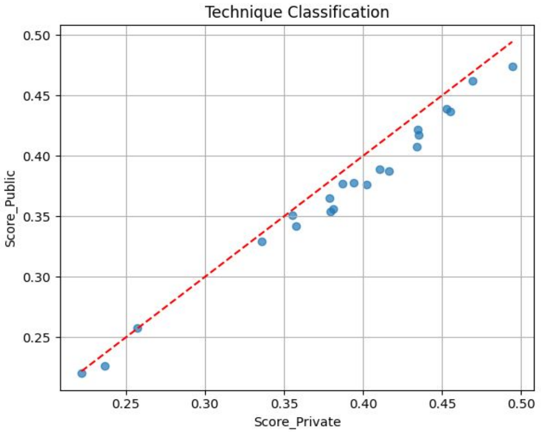


Subtask	Pearson Correlation	Spearman Correlation
Span Identification	0.997	0.978
Technique Classification	0.995	0.987

Table 3: Correlation of public with private leaderboard scores for different subtasks.



Span Identification



Technique Classification

Participation

Span Identification:

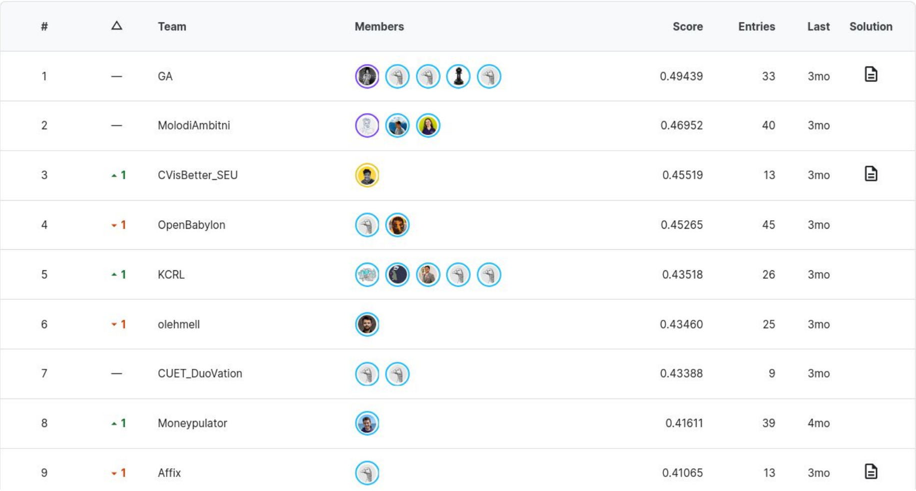
- Entrants: 53
- Participants: 32
- Teams: 12
- Submissions: 216

Technique Classification:

- Entrants: 65
- Participants: 46
- Teams: 20
- Submissions: 386

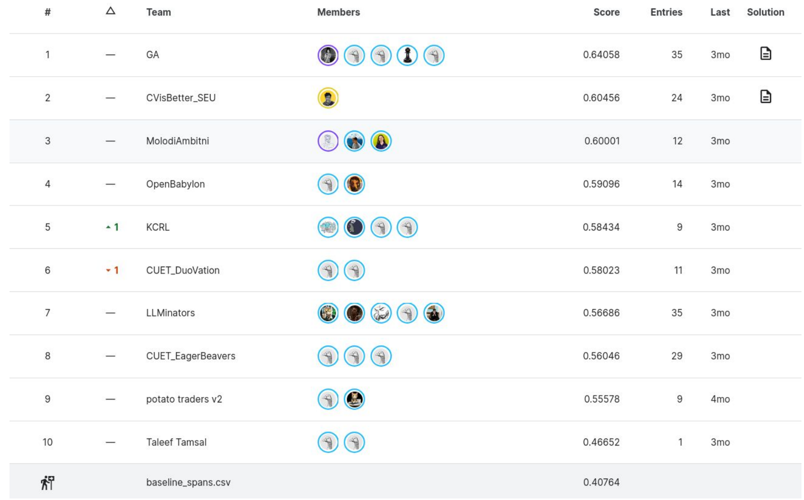
Results

Technique Classification



#	△	Team	Members	Score	Entries	Last	Solution
1	—	GA		0.49439	33	3mo	📄
2	—	MolodiAmbitni		0.46952	40	3mo	
3	▲1	CVisBetter_SEU		0.45519	13	3mo	📄
4	▼1	OpenBabylon		0.45265	45	3mo	
5	▲1	KCRL		0.43518	26	3mo	
6	▼1	olehmell		0.43460	25	3mo	
7	—	CUET_DuoVation		0.43388	9	3mo	
8	▲1	Moneypulator		0.41611	39	4mo	
9	▼1	Affix		0.41065	13	3mo	📄

Span Identification



#	△	Team	Members	Score	Entries	Last	Solution
1	—	GA		0.64058	35	3mo	📄
2	—	CVisBetter_SEU		0.60456	24	3mo	📄
3	—	MolodiAmbitni		0.60001	12	3mo	
4	—	OpenBabylon		0.59096	14	3mo	
5	▲1	KCRL		0.58434	9	3mo	
6	▼1	CUET_DuoVation		0.58023	11	3mo	
7	—	LLMinators		0.56686	35	3mo	
8	—	CUET_EagerBeavers		0.56046	29	3mo	
9	—	potato traders v2		0.55578	9	4mo	
10	—	Taleef Tamsal		0.46652	1	3mo	
		baseline_spans.csv		0.40764			

Team GA

Technique Classification

- Explored multiple models: mDeBERTa, Aya101, LLaMA3, Mistral Large.
- Chose Gemma 2-27B (decoder-only) for best performance.
- Handled class imbalance by optimizing classification thresholds via grid search, regularized by class distribution (instead of default 0.5).
  Improved generalization by using out-of-fold ensemble (averaging predictions across CV folds).

Span Identification

- Evaluated encoder models: mBERT, XLM-RoBERTa, EuroBERT, mDeBERTa.
- Found mDeBERTa most effective among smaller encoders.
- Hypothesized large decoder-only LLMs could outperform due to scale and pretraining.
- Built a custom encoder-like architecture based on Gemma 2-27B to enable bidirectional attention.
- Pretrained on Ukrainian and russian news corpora using masked language modeling.
- Used character-level binary labeling instead of BIO tags.
- Optimized span thresholds via grid search.
- Final model: ensemble across all folds.

Team MolodiAmbitni

Technique Classification

- Used instruction-tuned Gemma 2-2B with LoRA for parameter-efficient fine-tuning.
- Followed a multistage fine-tuning pipeline:
  - Stage 1: Causal language modeling.
  - Stage 2: Sequence classification.
- Prompts included:
  - Class descriptions.
  - Similarity-selected examples.
- Final classifier combined:
  - LLM outputs.
  - CatBoost-based metadata features.
- Class-specific thresholds were optimized using stratified k-fold cross-validation.

Span Identification

- Fine-tuned XLM-RoBERTa-large for binary token classification.
- Added a multi-target classification head.
- Employed k-fold cross-validation to optimize thresholds.

Team CVisBetter_SEU

Technique Classification

- Fine-tuned XLM-RoBERTa-large in a multilingual setting.
- Addressed class imbalance with:
  - Weighted binary cross-entropy (with capped class weights).
  - Label smoothing.
  - Word-level data augmentation.
- Architecture enhancements:
  - GELU-activated pre-classifier.
  - Multi-sample dropout.
- Training strategy:
  - AdamW optimizer with cosine scheduler.
  - Gradient accumulation and early stopping.
- Dynamic threshold tuning based on per-class F1 score.
- Used language-specific preprocessing and heuristics for Ukrainian and russian.

Span Identification

- Used XLM-RoBERTa-large with BIO tagging for token classification.
- Training techniques:
  a. Layer-wise Learning Rate Decay for better layer utilization.
  b. Weighted focal loss to address token-level imbalance.
  c. Early stopping to prevent overfitting.
- Post-processing: Merged adjacent spans using a threshold-based strategy.
- Employed balanced sampling and token-level F1 for evaluation.

Limitations

- **Dataset Scope:** The dataset used in this shared task is limited to Ukrainian Telegram posts
- **Technique Granularity:** Some techniques may overlap semantically or appear jointly in a single sentence, making clear-cut classification difficult.
- **Dataset Split:** Do not consider grouping by channels (data sources) and correct time split

Conclusion

- Established new benchmarks in task of Manipulation Technique detection for Ukrainian media field
- Introduced new Manipulation Technique detection dataset
- Gathered strong baselines in Manipulation Technique classifications and Manipulation Span Detection Challenges

Q&A