УКРАЇНСЬКИЙ КАТОЛИЦЬКИЙ УНІВЕРСИТЕТ

upf.
Universitat
Pompeu Fabra
Barcelona

Hidden Persuasion: Detecting Manipulative Narratives on Social Media During the 2022 Russian Invasion of Ukraine

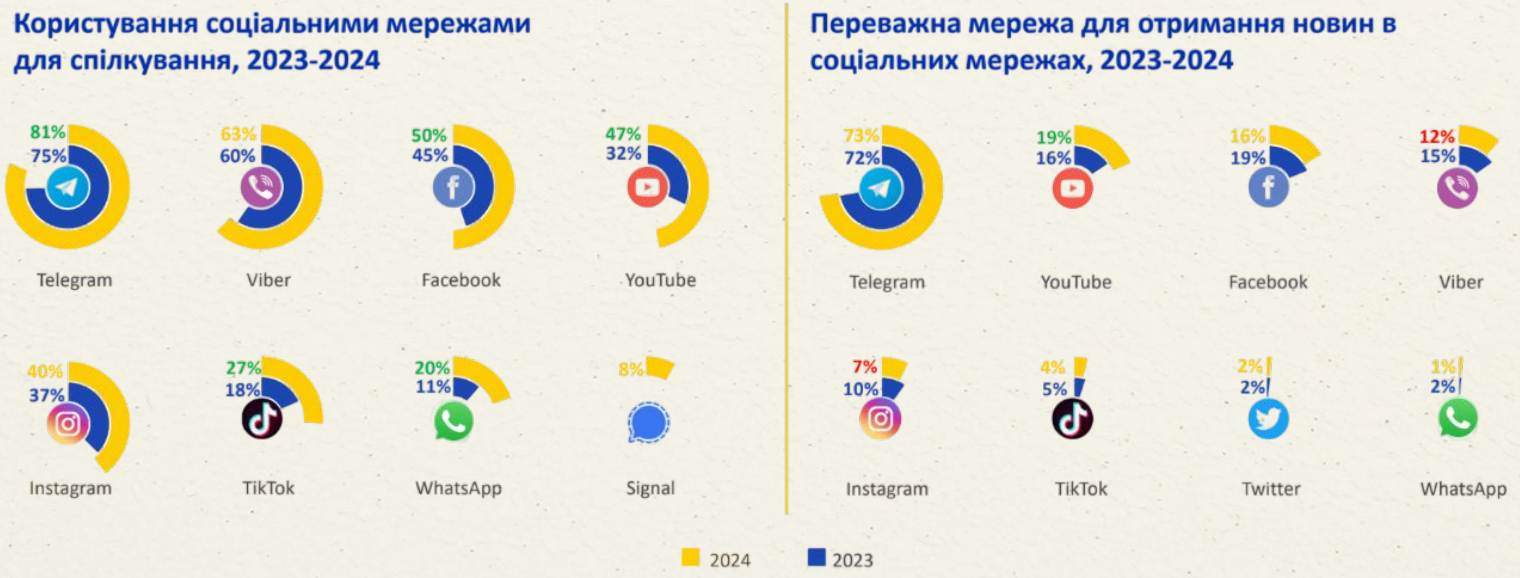
Kateryna Akhynko
Ukrainian Catholic University

Oleksandr Kosovan
Ukrainian Catholic University

Mykola Trokhymovych
Pompeu Fabra University

Background

73% of Ukrainians use Telegram as a primary source of information *(USAID-Internews, summer, 2024)*:

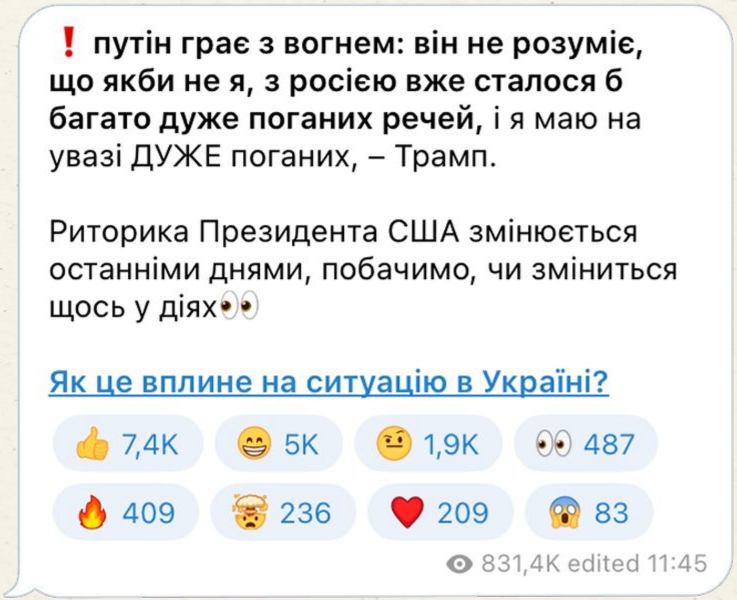


Користування соціальними мережами для спілкування, 2023-2024

Переважна мережа для отримання новин в соціальних мережах, 2023-2024

Background

Are we reading facts
or interpretations disguised
as facts?



❗ **путін грає з вогнем: він не розуміє, що якби не я, з росією вже сталося б багато дуже поганих речей,** і я маю на увазі ДУЖЕ поганих, – Трамп.

Риторика Президента США змінюється останніми днями, побачимо, чи зміниться щось у діях👀

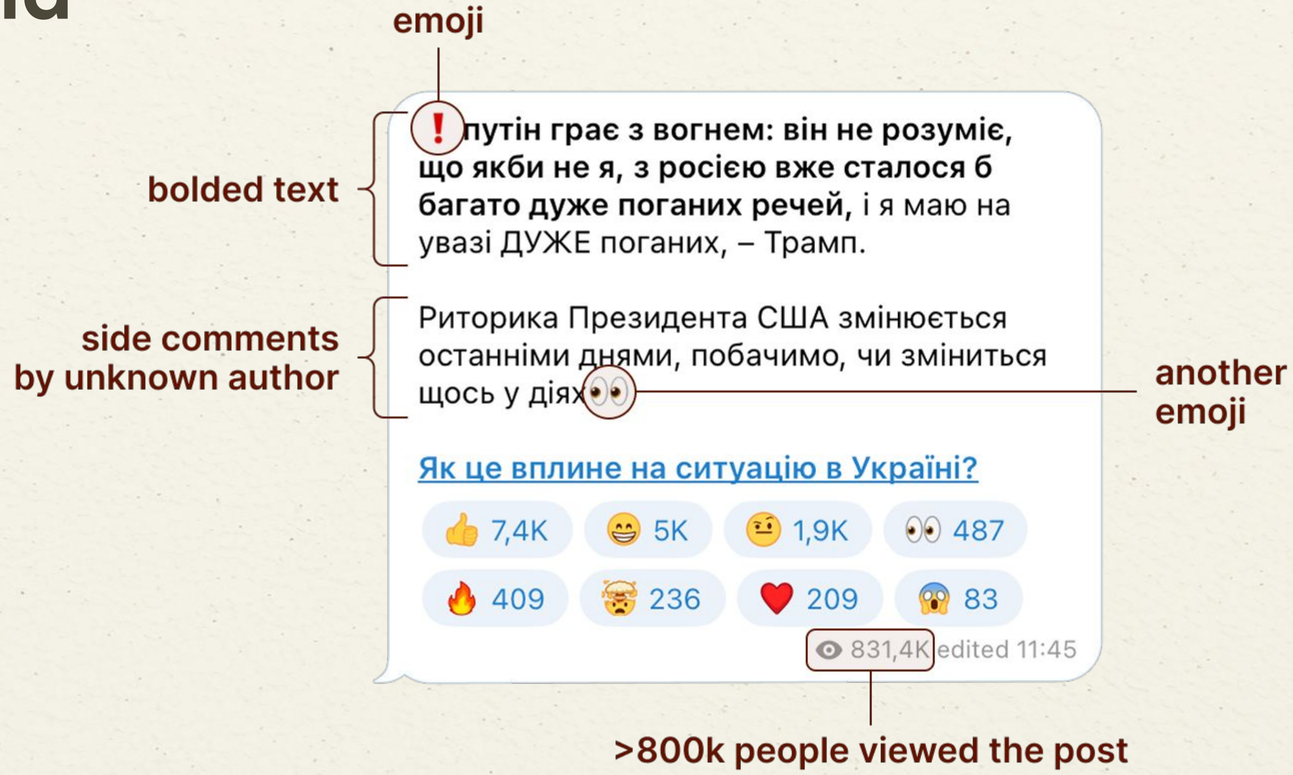
Як це вплине на ситуацію в Україні?

👍 7,4K 😁 5K 🤨 1,9K 👀 487
🔥 409 🤯 236 ❤️ 209 😱 83

831,4K edited 11:45

Background

*Are we reading facts
or interpretations disguised
as facts?*



❗ **путін грає з вогнем: він не розуміє, що якби не я, з росією вже сталося б багато дуже поганих речей,** і я маю на увазі ДУЖЕ поганих, – Трамп.

Риторика Президента США змінюється останніми днями, побачимо, чи зміниться щось у діях👀

Як це вплине на ситуацію в Україні?

👍 7,4K 😁 5K 🤨 1,9K 👀 487

🔥 409 🤯 236 ❤️ 209 😱 83

831,4K edited 11:45

emoji

bolded text

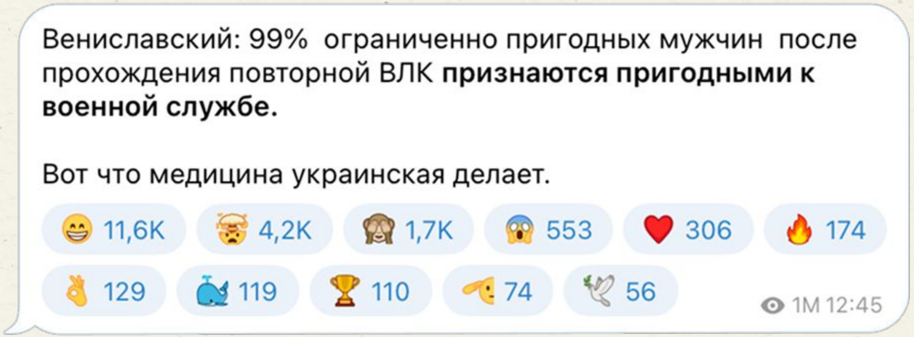
side comments
by unknown author

another
emoji

>800k people viewed the post

Background

Are we reading facts
or interpretations disguised
as facts?



Вениславский: 99% ограниченно пригодных мужчин после прохождения повторной ВЛК **признаются пригодными к военной службе.**

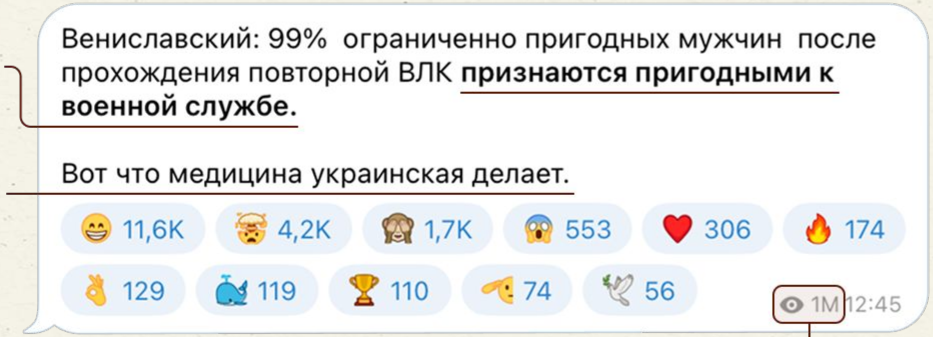
Вот что медицина украинская делает.

😁 11,6K 🤯 4,2K 🙈 1,7K 😱 553 ❤️ 306 🔥 174
👌 129 🐳 119 🏆 110 🫵 74 🕊 56

👁 1M 12:45

Background

Are we reading facts
or interpretations disguised
as facts?



Венислaвский: 99% ограниченно пригодных мужчин после прохождения повторной ВЛК **признаются пригодными к военной службе.**

Вот что медицина украинская делает.

bolded text

side comments
by unknown author

1 million people
viewed the post

Problem formulation

Goal:

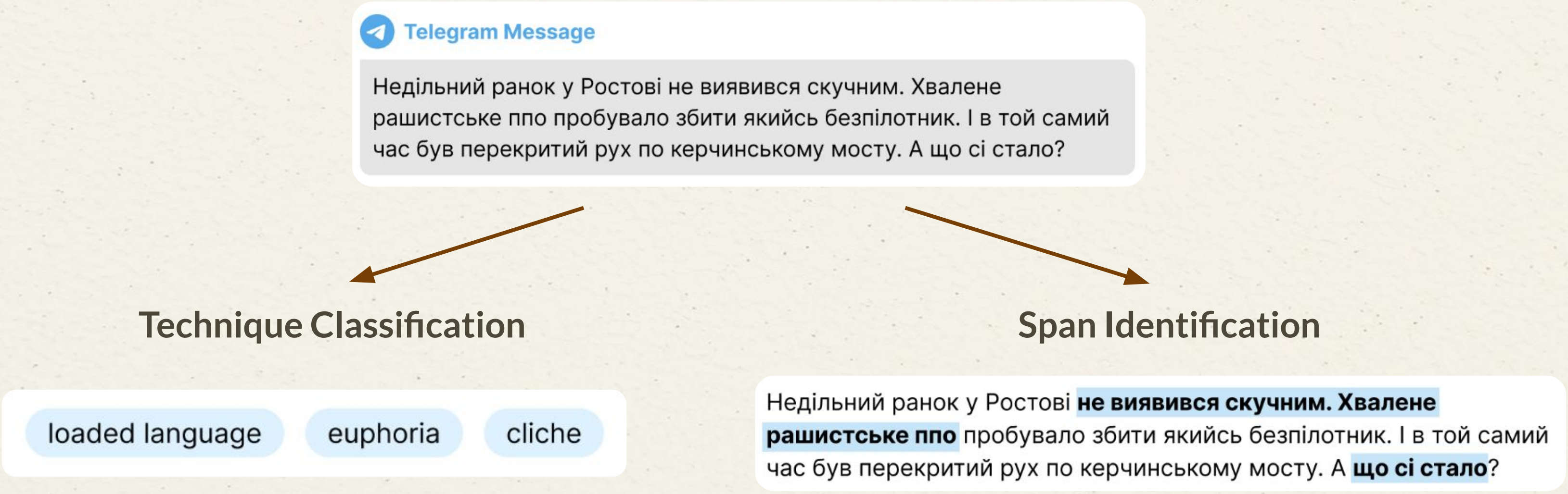
Detect **manipulative narratives** in Ukrainian Telegram posts.

Manipulation defined (UNLP 2025):

The **use of rhetorical or stylistic techniques** to influence readers’ opinions or behavior — **without relying on factual evidence**.

Problem formulation

Tasks



Telegram Message

Недільний ранок у Ростові не виявився скучним. Хвалене рашистське ппо пробувало збити якийсь безпілотник. І в той самий час був перекритий рух по керчинському мосту. А що сі стало?

Technique Classification

loaded language

euphoria

cliche

Span Identification

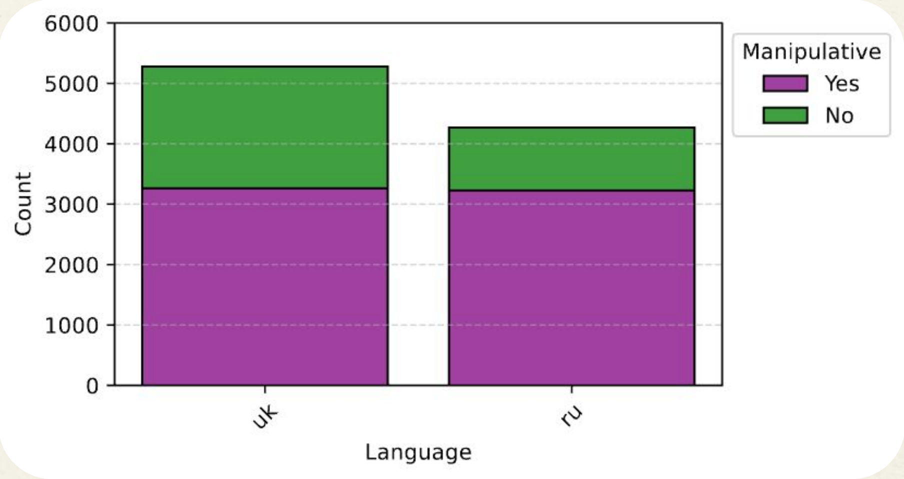
Недільний ранок у Ростові **не виявився скучним. Хвалене рашистське ппо** пробувало збити якийсь безпілотник. І в той самий час був перекритий рух по керчинському мосту. А **що сі стало**?

Data Overview

- **Source:** Telegram posts from Ukrainian channels *(provided by UNLP 2025 Shared Task)*
- **Each post** is annotated with **manipulation techniques** and **manipulative spans**
- 2 **languages** present:

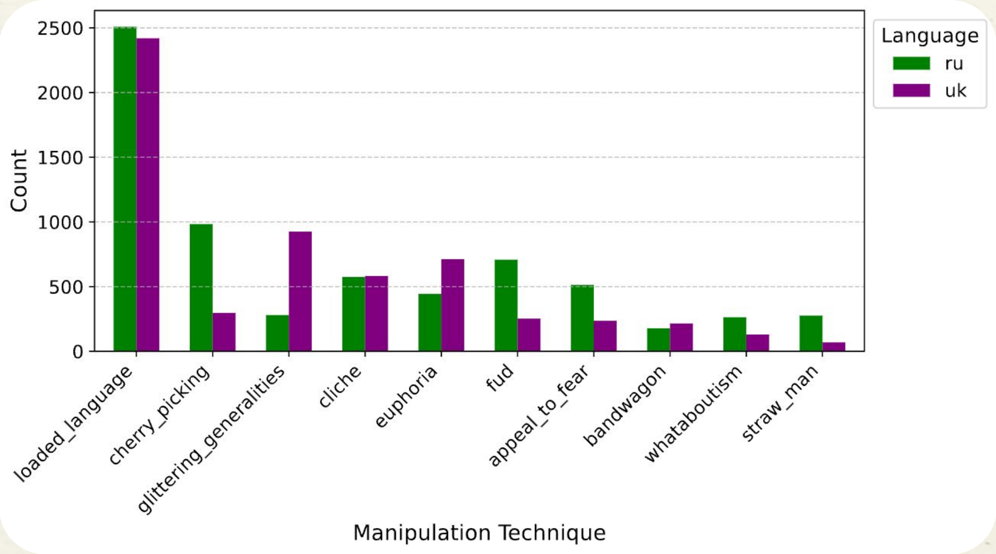
  Ukrainian: 5,278 posts

  Russian: 4,269 posts

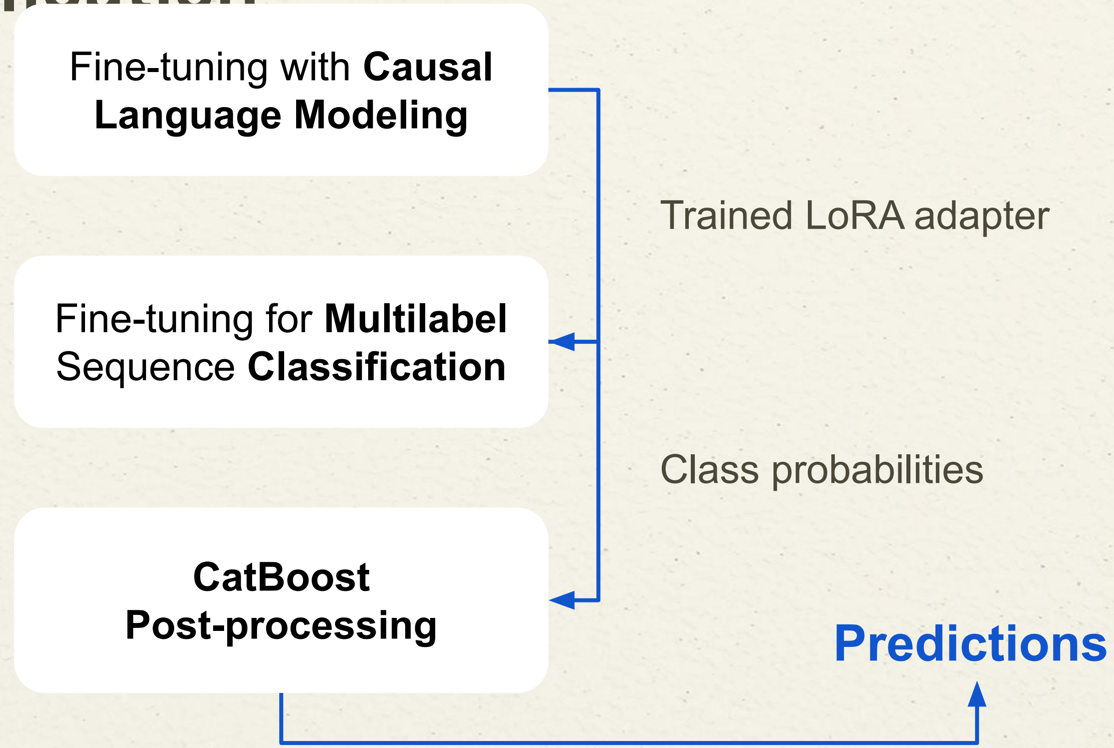


Data Overview

- **Source**: Telegram posts from Ukrainian channels *(provided by UNLP 2025 Shared Task)*
- **Each post** is annotated with **manipulation techniques** and **manipulative spans**
- 10 predefined **techniques**:



Proposed Solution – Technique Classification



Fine-tuning with **Causal Language Modeling**

Trained LoRA adapter

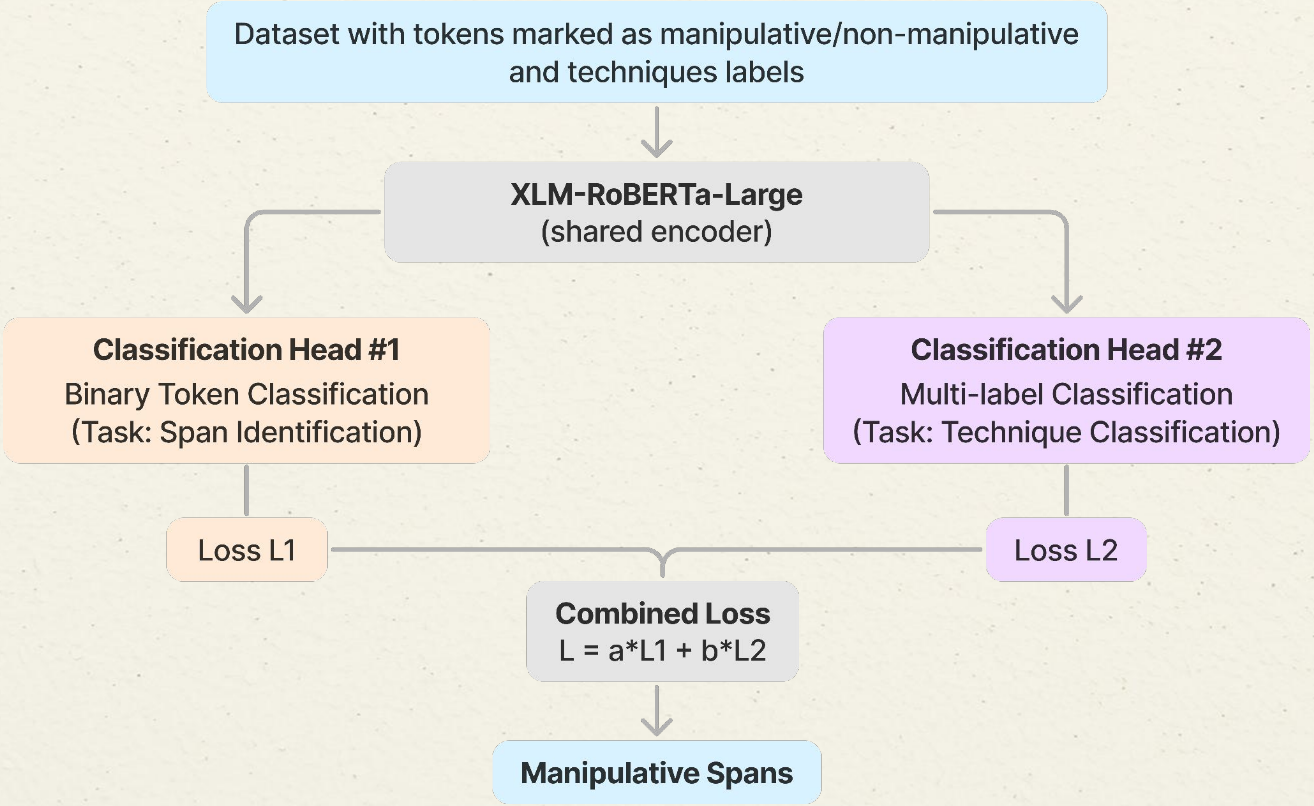
Fine-tuning for **Multilabel** Sequence **Classification**

Class probabilities

CatBoost Post-processing

Predictions

Proposed Solution for Span Identification



Dataset with tokens marked as manipulative/non-manipulative and techniques labels

XLM-RoBERTa-Large
(shared encoder)

Classification Head #1
Binary Token Classification
(Task: Span Identification)

Classification Head #2
Multi-label Classification
(Task: Technique Classification)

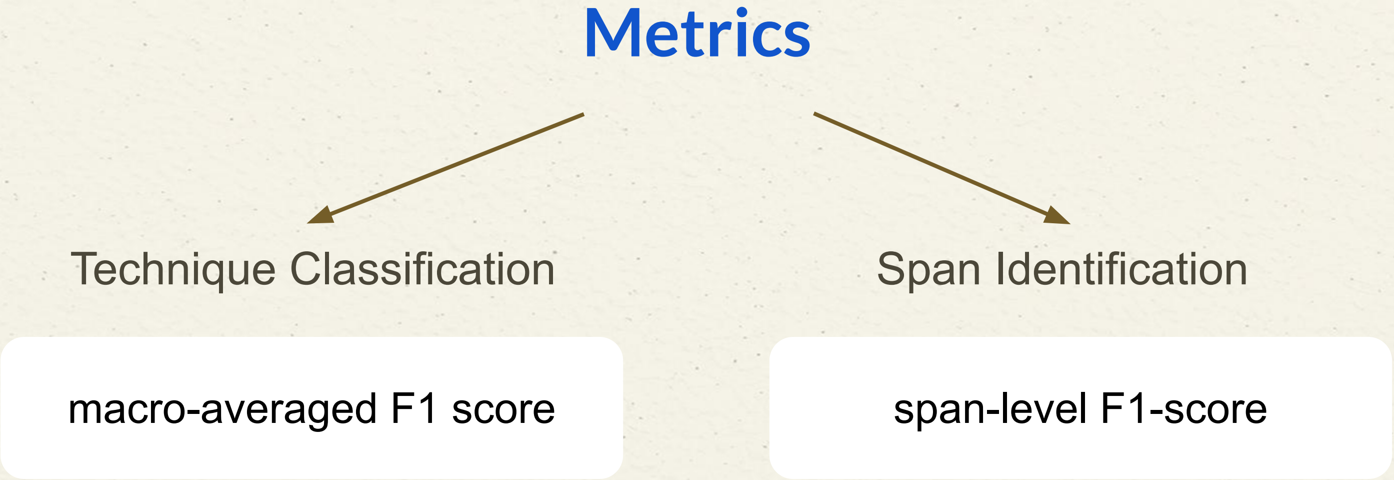
Loss L1

Loss L2

Combined Loss
L = a*L1 + b*L2

Manipulative Spans

Results



Metrics

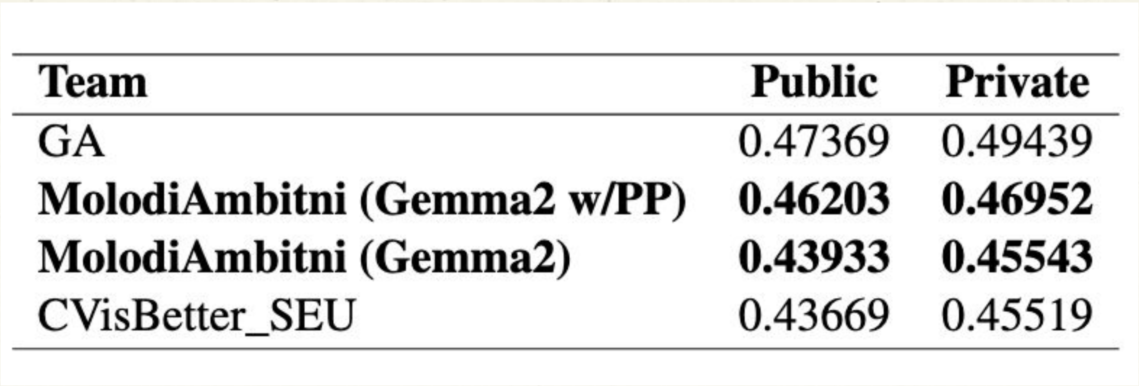
Technique Classification

macro-averaged F1 score

Span Identification

span-level F1-score

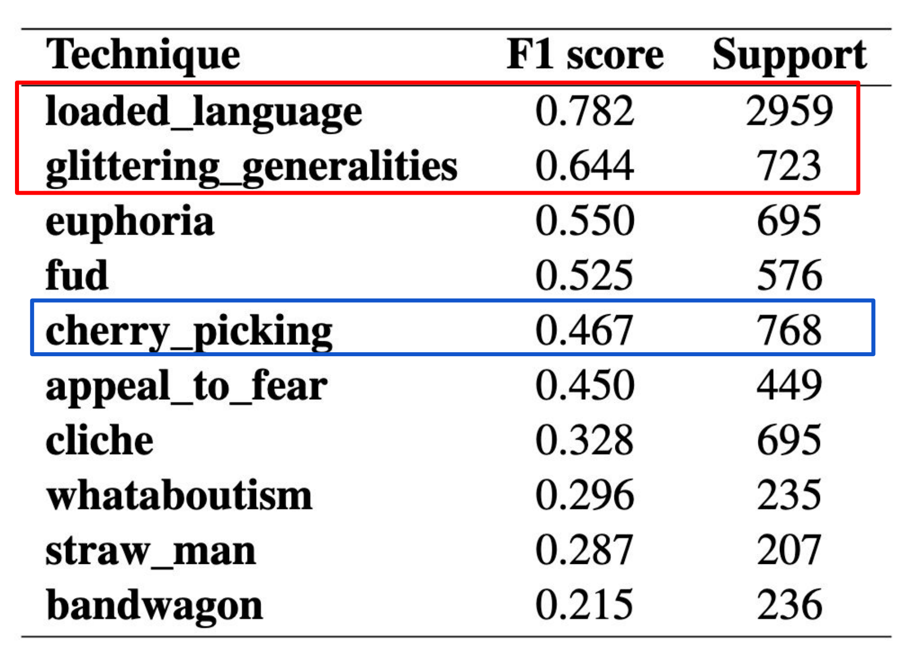
Results – Technique Classification



Team	Public	Private
GA	0.47369	0.49439
MolodiAmbitni (Gemma2 w/PP)	**0.46203**	**0.46952**
MolodiAmbitni (Gemma2)	**0.43933**	**0.45543**
CVisBetter_SEU	0.43669	0.45519

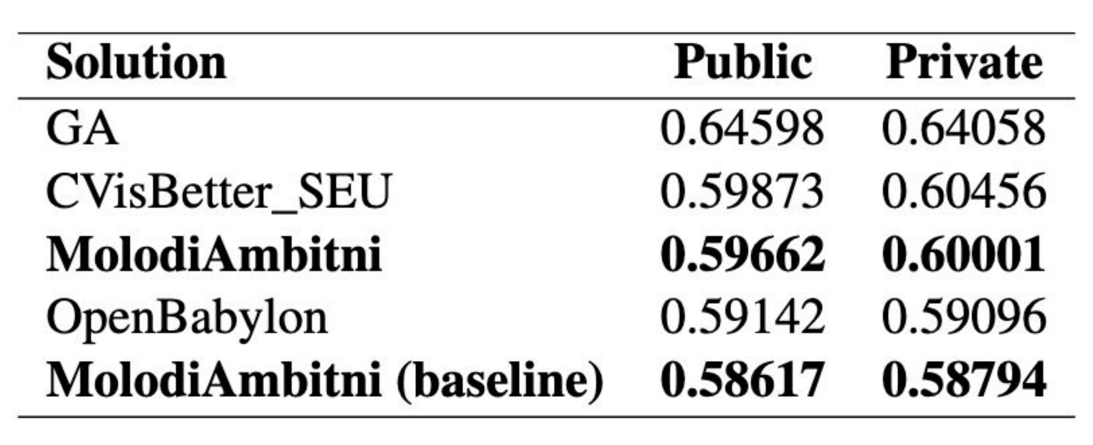
Comparison of metrics for top-3 solutions from competition leaderboard

Results – Technique Classification



Technique	F1 score	Support
loaded_language	0.782	2959
glittering_generalities	0.644	723
euphoria	0.550	695
fud	0.525	576
cherry_picking	0.467	768
appeal_to_fear	0.450	449
cliche	0.328	695
whataboutism	0.296	235
straw_man	0.287	207
bandwagon	0.215	236

Results – Span Identification



Solution	**Public**	**Private**
GA	0.64598	0.64058
CVisBetter_SEU	0.59873	0.60456
MolodiAmbitni	**0.59662**	**0.60001**
OpenBabylon	0.59142	0.59096
MolodiAmbitni (baseline)	**0.58617**	**0.58794**

Comparison of metrics for top-4 solutions from competition leaderboard

Conclusions

- Achieved **2nd place** in **technique classification** and **3rd place** in **span detection** in the UNLP 2025 Shared Task

- Developed fine-tuned Gemma 2 + meta-feature post-processing that **significantly boosted** classification performance

- Showed that a simple **XLM-RoBERTa model**, paired with a **dual-head pipeline**, can achieve **top-tier** span detection **results**

Thank you!

Q&A

Stage 1: Fine-tuning with CLM

Model

Gemma 2B IT

Training Setup

LoRA (Alpha (α): 32, Rank (r): 32)
+ 4-bit quantization

in causal LM setup

Output

LoRA adapter for techniques generation

Stage 1: Fine-tuning with CLM

Prompt Composition

System:
You are an AI trained to detect rhetorical manipulation in social media. Return ONLY the technique names from the list, comma-separated.

User:

Task: *Identify techniques in this post using ONLY the following: <techniques description>*

Examples: *<2 examples of posts + their techniques>*

POST to analyze: *<target post text>*

Assistant:
Predicted output: <technique1, technique2, …>

Stage 2: Supervised Multi-label Classification

Model: Gemma 2B IT + LoRA adapter from Stage 1

Training Setup:

1. Multi-label sequence classification
2. LoRA (Alpha (α): 32, Rank (r): 16) adapter
3. Threshold selection per class

Output: class probabilities for each text

Stage 3: Post–Processing

Model Used: CatBoost

Features Used:

- technique probabilities from Stage 2

- cosine distances **->** current **text** and **centroids** of trigger phrase clusters
- **frequency** of techniques among top-20 **nearest texts** and **trigger phrases**

- meta-features :
  word count, number of question marks, presence of URLs, etc