Comparing Methods for Multi-Label Classification of Manipulation Techniques in Ukrainian Telegram Content

Oleh Melnychuk
Kyiv Aviation Institute / Kyiv, Ukraine
olehmell@pm.me

Abstract

Detecting manipulation techniques in online text is vital for combating misinformation, a task complicated by generative AI. This paper compares machine learning approaches for multi-label classification of 10 techniques in Ukrainian Telegram content (UNLP 2025 Shared Task 1). Our evaluation included TF-IDF, fine-tuned XLM-RoBERTa-Large, PEFT-LLM (Gemma, Mistral) and a RAG approach (E5 + Mistral Nemo). The fine-tuned XLM-RoBERTa-Large model, which incorporates weighted loss to address class imbalance, yielded the highest Macro F1 score (0.4346). This result surpassed the performance of TF-IDF (Macro F1 0.32-0.36), the PEFT-LLM (0.28-0.33) and RAG (0.309). Synthetic data slightly helped TF-IDF but reduced transformer model performance. The results demonstrate the strong performance of standard transformers like XLM-R when appropriately configured for this classification task.

1 Introduction

The volume of online content requires effective methods to identify manipulative language. This work focuses on detecting specific manipulation techniques – defined here as rhetorical or stylistic methods aimed at influencing audiences without clear factual support – within Ukrainian social media content, specifically from Telegram. This investigation is part of our more extensive research on the challenges posed by generative AI in the defense of Sybil's attacks on social media. (Ferrara, 2023; Feng et al., 2024). Understanding these manipulation techniques is therefore crucial for countering coordinated information operations, especially in contexts such as the ongoing hybrid warfare against Ukraine, where social networks are actively used for disinformation campaigns. (Makhortykh et al., 2024).

This paper investigates the effectiveness of different modeling approaches for the specific task of identifying manipulation techniques, using data from the UNLP 2025 Shared Task (Subtask 1). This shared task aims to assess the AI capabilities in detecting the manipulation of social media within the Ukrainian context. We compare:

1. Traditional bag-of-words approaches using TF-IDF features with linear classifiers (Logistic Regression, SVM).
2. A standard fine-tuned transformer model (XLM-RoBERTa-Large).
3. Recent LLMs fine-tuned using Parameter-Efficient Fine-Tuning (PEFT) techniques (LoRA).
4. Retrieval-Augmented Generation (RAG) approach.
5. The effect of augmenting training data with synthetically generated examples.

Our findings indicate that fine-tuning standard transformer models like RoBERTa yields strong performance on this multi-label classification task, especially with limited data. Concurrently, we explored the potential of advanced methods such as Retrieval-Augmented Generation (RAG) and Parameter-Efficient Fine-Tuning (PEFT) using LoRA for smaller LLMs, providing insights into their applicability compared to the established fine-tuning paradigm under data constraints.

2 Methodology

2.1 Dataset

We use the dataset provided for the UNLP 2025 Shared Task on Classification Techniques (Subtask 1) hosted on Kaggle[1]. The dataset was

[1] https://www.kaggle.com/competitions/unlp-2025-shared-task-classification-techniques/overview

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 45–48

provided by the Texty.org.ua team and consists of Ukrainian text snippets from Telegram posts, labeled with one or more of ten manipulation techniques: straw_man, appeal_to_fear, fud, bandwagon, whataboutism, loaded_language, glittering_generalities, cherry_picking, euphoria and cliche. Annotation was performed by experienced journalists, analysts, and media professionals. The training data has a significant class imbalance. We used the provided training data (train.csv), splitting it 90% for training and 10% for validation. Performance is reported on the official competition test set (test.csv) based on the Macro F1 score achieved on the Kaggle leaderboard.

2.2 Preprocessing

For all models, text content was preprocessed by: converting to lowercase, removing URLs, user mentions (@), hashtags (#), and emojis. For the TF-IDF-based models, lemmatization using Mystem was additionally applied.

2.3 Synthetic Data Generation

To address potential data scarcity, we attempted to augment the training set with synthetic examples generated using the mistral-large-latest model via its API (proprietary models were allowed only for data generation per task rules). The prompts were designed to generate diverse and realistic text snippets (approximately 200 words) demonstrating specific manipulation techniques within the Ukrainian Telegram context. The impact of these data varied, as discussed in Section 3.3.

2.4 Models Explored

- **TF-IDF + Classifiers:** We vectorized the cleaned (and lemmatized) text using TF-IDF (char n-grams 3-5, maximum 10k features). We trained separate binary classifiers (Logistic Regression, SVM) for each technique. SMOTE was applied to the training data (original or augmented) for each binary classifier, and threshold adjustment was performed on the validation set.

- **XLM-RoBERTa-Large:** We used xlm-roberta-large (Conneau et al., 2020), a transformer model known for strong performance on various NLP tasks via Hugging Face transformers (Wolf et al., 2020). We used AutoModelForSequenceClassification configured for multi_label_classification. The model was fine-tuned end-to-end.

- **LLMs with LoRA:** We experimented with Gemma-3-1B[2] (Gemma Team et al., 2025) and Mistral-Small/Nemo models[3] (based on architectures like Mistral 7B (Jiang et al., 2023)) (4 bits quantized via unsloth[4]), adhering to the open source model requirement for solutions. LoRA (Hu et al., 2021) was applied (r=8, lora_alpha=8). The models were configured for sequence classification.

- **Retrieval-Augmented Generation (RAG):** We tested a RAG approach using open source components[5]. A vector database (MongoDB + FAISS) was created that contains embeddings of the training data generated using intfloat/multilingual-e5-large was created. Embeddings could be enriched by weighting trigger word positions. For a test input, we retrieved the k (k=5) most similar examples k (k = 5) based on embedding similarity. These retrieved examples (text, techniques, manipulative flag) and the original query were used to construct a prompt for a generator LLM (mistral-nemo, potentially related to (Jiang et al., 2023)) accessed via a local API) to predict the applicable manipulation techniques in JSON format.

2.5 Handling Class Imbalance: Weighted Loss

For direct transformer/LLM fine-tuning, we used BCEWithLogitsLoss with a pos_weight calculated for each class i based on the inverse frequency of positive samples in the training dataset:

$$\text{pos\_weight}[i] = \frac{\text{count}(\text{negative\_samples}_i)}{\text{count}(\text{positive\_samples}_i) + \epsilon} \quad (1)$$

This tensor of weights was passed to the loss function.

[2] https://colab.research.google.com/github/unslothai/notebooks/blob/main/nb/Gemma3_(1B)-GRPO.ipynb

[3] https://docs.mistral.ai/capabilities/finetuning/

[4] https://docs.unsloth.ai/get-started/fine-tuning-guide

[5] https://www.kaggle.com/code/woters/building-rag-using-mistral-faiss-v2

2.6 Evaluation Metric

The primary evaluation metric is the **Macro F1-score**. We also monitor Micro F1 and Hamming loss.

3 Experiments and Results

3.1 Experimental Setup

The models were trained on NVIDIA GPUs available via free-tier Google Colab and Kaggle notebooks. Hyperparameters for XLM-RoBERTa included: LR = 2e-5, batch size = 16, epochs = 5-15, weight loss = 0.01, AdamW. LLM used LR = 1e-4, gradient accumulation (effective batch $\tilde{8}$). The RAG approach used E5-large for embeddings and Mistral Nemo for generation. The best checkpoint for fine-tuned models was selected based on Macro F1 score.

3.2 Results

Table 1 shows the performance on the official Kaggle test set for Subtask 1 (Technique Classification). Note that the TF-IDF score reflects augmentation; others use the original data.

Model and Configuration	Macro F1
TF-IDF (LogReg, SMOTE, Tuned Thr.)	0.36
TF-IDF (SVM, SMOTE, Tuned Thr.)	0.32
RAG (E5 + Mistral Nemo, Retr.+Gen.)	0.309
Gemma-3-1B (LoRA r=8, 4b, W. Loss)	0.28
Mistral Small/Nemo (LoRA r=8, 4b, W. Loss)	0.33
XLM-RoBERTa-Large (Std. FT, W. Loss)	**0.4346**

Table 1: Comparison of Macro F1 scores on the Kaggle test set (Subtask 1). TF-IDF+LogReg score reflects augmentation; others use original data. Abbreviations: Thr. (Thresholds), Retr.+Gen. (Retrieval + Generation), 4b (4-bit), W. Loss (Weighted Loss), Std. FT (Standard Fine-tuning).

3.3 Analysis

XLM-RoBERTa-Large fine-tuned with weighted loss outperforms other methods for classifying manipulation techniques. Addressing class imbalance with weighted loss was essential for performance.

Traditional TF-IDF methods serve as baselines. Their limitations arise because methods based on simple textual patterns struggle against content that avoids repetition and mimics human writing, a known challenge with LLM-generated text (Feng et al., 2024). Increasing the training data with synthetic examples from Mistral Large slightly improved TF-IDF + Logistic Regression (Macro F1 increasing from $\tilde{0}$.30 to 0.36). However, these same synthetic data reduced performance (10-20% F1 drop) when used to train the XLM-R and LoRA LLM models, suggesting issues with the quality or distribution of the generated examples or perhaps greater model sensitivity. Consequently, synthetic data were omitted for the final transformer/LLM runs.

The RAG approach, which combined E5-large embeddings for retrieval and mistral Nemo for generation, yielded a Macro F1 score of 0.309. Although showing the feasibility of RAG, this performance was lower than the TF-IDF baselines and the fine-tuned XLM-R, suggesting difficulties in using retrieved examples effectively for this multi-label classification task within our setup, perhaps requiring different prompting or retrieval strategies (e.g., (Zhang et al., 2024)).

4 Conclusion

This paper compared several methods for multi-label classification of manipulation techniques in Ukrainian Telegram content. A standard fine-tuned XLM-RoBERTa-Large model with weighted loss achieved the highest performance (0.4346 Macro F1), outperforming the TF-IDF baselines, PEFT-tuned LLMs (Gemma, Mistral) and an RAG approach. The attempted augmentation of synthetic data using Mistral Large slightly benefited TF-IDF but harmed transformer/LLM performance, which shows challenges in generating effective synthetic data for complex models. Our results show the continued effectiveness of appropriately tuned standard transformer architectures for specific classification tasks, especially when addressing dataset properties like class imbalance.

Although our RAG implementation performed poorly here, the strategy shows potential, particularly for its ability to incorporate up-to-date information, which is important for dynamic analysis tasks. We suggest that RAG could be useful in a production pipeline, perhaps using LLMs fine-tuned with a dedicated classification head. Some competition participants reportedly achieved results that exceeded our XLM-R score, possibly employing such custom LLM classifiers, indicating room for improvement over standard transformers.

Furthermore, the increasing sophistication of AI-generated content used for targeted manipulation (Goldstein et al., 2023; Yang and Menczer, 2023), requires the development of adaptive, potentially

hybrid detection systems. A key focus for future work will be improving the adaptability to new manipulation campaigns and evolving language, favoring further investigation of RAG and reasoning models.

Limitations

Our study had several limitations, mainly dictated by shared task rules and available resources. First, the prohibition on using external Telegram data restricted our training set to the provided corpus. Although external data was allowed, procuring high-quality, relevant, and appropriately licensed data for Ukrainian Telegram content proved challenging. Second, the requirement to use only open-source models for the final submitted solutions constrained our model choices, although proprietary models like Mistral Large were permitted and used for experimental data generation.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *Preprint*, arXiv:1911.02116.

Shangbin Feng, Herun Wan, Ningnan Wang, Zhaoxuan Tan, Minnan Luo, and Yulia Tsvetkov. 2024. What Does the Bot Say? Opportunities and Risks of Large Language Models in Social Media Bot Detection. *Preprint*, arXiv:2402.00371.

Emilio Ferrara. 2023. Social Bot Detection in the Age of ChatGPT: Challenges and Opportunities. *First Monday*, 28(11).

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, and 1 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *Preprint*, arXiv:2301.04246.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mykola Makhortykh, Maryna Sydorova, Ani Baghumyan, Victoria Vziatysheva, and Elizaveta Kuznetsova. 2024. Stochastic Lies: How LLM-Powered Chatbots Deal with Russian Disinformation about the War in Ukraine. *Harvard Kennedy School Misinformation Review*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. HuggingFace's Transformers: State-of-the-Art Natural Language Processing. *Preprint*, arXiv:1910.03771.

Kai-Cheng Yang and Filippo Menczer. 2023. Anatomy of an AI-powered malicious social botnet. *Preprint*, arXiv:2307.16336.

Lechen Zhang, Tolga Ergen, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. SPRIG: Improving Large Language Model Performance by System Prompt Optimization. *Preprint*, arXiv:2410.14826.