The UNLP 2025 Shared Task on Detecting Social Media Manipulation

Roman Kyslyi[1], Nataliia Romanyshyn[2], Volodymyr Sydorskyi[1]
[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
[2]Texty.org.ua
kyslyi.roman@lll.kpi.ua, nataliia.romanyshyn@texty.org.ua, v.sydorskyi@kpi.ua

Abstract

This paper presents the results of the UNLP 2025 Shared Task on Detecting Social Media Manipulation. The task included two tracks: Technique Classification and Span Identification. The benchmark dataset contains 9,557 posts from Ukrainian Telegram channels manually annotated by media experts. A total of 51 teams registered, 22 teams submitted systems, and 595 runs were evaluated on a hidden test set via Kaggle. Performance was measured with macro F1 for classification and token-level F1 for identification. The shared task provides the first publicly available benchmark for manipulation detection in Ukrainian social media and highlights promising directions for low-resource propaganda research. The Kaggle leaderboard is left open for further submissions.

1 Introduction

The disinformation and manipulative content on social media platforms poses significant challenges to information integrity. In Ukraine, the spread of propaganda through channels like Telegram has underscored the need for advanced NLP techniques to detect and mitigate such content. Recent studies have emphasized the importance of automatic approaches for identifying disinformation, including work focused on russian- and Ukrainian-language content (Taras et al., 2024; Grabar and Hamon, 2024; Zeng et al., 2024; Golovchenko et al., 2023).

To address these challenges, the Fourth Workshop on Ukrainian Natural Language Processing (UNLP) 2025, together with Texty.org.ua[1], organized a Shared Task focused on the detection of social media manipulation in Ukrainian information space. The task comprised two subtasks:

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

The dataset for this shared task was created by Texty.org.ua and consists of 9,557 Ukrainian Telegram posts annotated by media experts for manipulation techniques. This initiative aims to encourage the development of NLP models capable of understanding and detecting nuanced manipulative strategies in Ukraine.

Participants received the datasets, task descriptions, and evaluation metrics via the official GitHub repository[2]. Both subtasks were hosted as Kaggle competitions: Technique Classification[3] and Span Identification[4].

This paper presents an overview of the shared task, including the dataset, evaluation methodology, and a synthesis of participants' approaches and results. By analyzing the outcomes, we aim to highlight the progress in Ukrainian NLP and identify areas for future research and development.

The remainder of this paper is organized as follows. Section 2 reviews previous work on propaganda detection and span-level manipulation identification. Section 3 outlines the UNLP 2025 shared-task setup. Section 4 presents the dataset and manipulation-technique taxonomy. Section 5 describes the evaluation metrics and ranking procedure. Section 6 reports the leaderboard results and summarises the submitted systems. Section 7 concludes the paper, while Section 8 provides an ethics statement and Section 9 discusses current limitations and future work.

[1] https://texty.org.ua/p/about-en/

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

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

[4] https://www.kaggle.com/competitions/unlp-2025-shared-task-span-identification

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 105–111

2 Related Work

Early work in domain of disinformation detection focused on identifying biased or manipulative rhetoric in English-language news sources (Barrón-Cedeño et al., 2019). Subsequent shared tasks such as SemEval 2020 Task 11 (Da San Martino et al., 2020) and the NLP4IF workshop (Alam et al., 2021) further advanced the field by providing benchmark datasets and introducing more fine-grained classification of propaganda techniques.

Span-based propaganda detection, introduced in Da San Martino et al. (2020), treats the problem as a sequence labeling or span extraction task and remains a challenging low-resource setting. In multilingual contexts, limited annotated data has led to the adoption of transfer learning approaches using multilingual transformers like XLM-R (Conneau and Lample, 2019) and fine-tuned mBERT (Devlin et al., 2019) for classification and span identification.

3 Task Description

3.1 Technique Classification

In this shared task, the goal was to build a model capable of identifying manipulation techniques in Ukrainian social media content (specifically, Telegram). In this context, "manipulation" refers to the presence of specific rhetorical or stylistic techniques aimed to influence the audience without providing clear factual support (Da San Martino et al., 2019b).

Given the text of a post, participants had to identify which manipulation techniques were used, if any. This is a multilabel classification problem; a single post could contain multiple techniques (Table 2).

3.2 Span Identification

In the second track, the goal was to identify the specific spans of manipulative text, regardless of the manipulation technique. This is a binary named entity classification task, focusing on pinpointing exactly where the manipulative content occurs. This required systems to accurately detect and localize phrases that exhibit rhetorical or deceptive strategies within the broader context of the post.

4 Data

The dataset consists of 9,557 Telegram posts annotated for the presence of manipulation techniques. The content was collected from Ukrainian news and political blog channels on Telegram, comprising texts in Ukrainian and russian languages. This bilingual composition provides diverse examples of manipulative language used across different segments of the Ukrainian information space.

The dataset includes both manipulative and non-manipulative posts, with the distribution by language shown in Table 1.

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.

The dataset is available through the official repository of the shared task[5] and is licensed under the CC BY-NC-SA 4.0 License.

4.1 Manipulation Techniques

The list of manipulation techniques was compiled by Texty.org.ua. First, the team relied on existing Ukrainian expertise of russian propaganda, especially on the prior work of Detector Media[6] — to ensure that the labels were valid and relevant for the Ukrainian information space. Second, Texty conducted a focus group discussion with Ukrainian journalists, editors, and media analysts to resolve contentious cases:

1. decide which rhetorical patterns should be considered manipulation
2. distinguish manipulations that may be acceptable during the active phase of the war
3. identify the techniques viewed as most destructive on Ukrainian Telegram

The resulting corpus, therefore, combines prior expert research with the practical insights of local media professionals.

Table 2 lists the distribution of each technique in the dataset. Manipulative posts may contain any number of manipulation techniques, so the overall frequency of the techniques exceeds the total number of posts.

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

[6] https://disinfo.detector.media/en/theme/tactics-and-tools

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).

4.2 Dataset Split

Given the highly imbalanced distribution of manipulation techniques (Table 2), we employed the Multilabel Stratification algorithm (Sechidis et al., 2011). The entire dataset was initially split into five approximately equal folds, each containing 20% of the data (1911–1912 samples per fold), with the distribution of techniques preserved across all folds.

Subsequently, the first and second folds were combined to form the training set, the third and fourth folds constituted the private test set, and the fifth fold served as the public test set. As a result, the dataset was split as follows:

- **Training set:** 3822 samples
- **Private test set:** 3824 samples
- **Public test set:** 1911 samples

Importantly, the train/public/private splits remained identical for both competition tracks to prevent any potential data leakage between them.

Thanks to this split strategy, the correlation between public and private leaderboard scores was high (Table 3, Figure 1).

5 Evaluation

5.1 Evaluation Methodology

The evaluation methodology follows the standard Kaggle evaluation protocol, which utilizes both public and private test sets[7]. The public test set is available to participants throughout the competition and serves as an additional evaluation set for real-time feedback. In contrast, the private test set remains hidden until the competition ends and is used to determine the final leaderboard rankings. The main motivation behind using two separate test sets is to prevent overfitting to the public test data and to ensure that participants develop robust validation strategies and build models that generalize well.

5.2 Metrics

For the Technique Classification track, the standard F1 score with macro averaging[8] was used. For the Span Identification track, the F1 score was also used, but computed at the token level[9].

First, tokens are extracted from both the ground truth and predicted spans, where a token is defined as a full text chunk corresponding to a single span. Then, true positives (TP), false positives (FP), and false negatives (FN) are calculated based on the total number of predicted and ground truth tokens and their overlaps. Finally, precision, recall, and the F1 score are computed.

The motivation for using token-level F1 rather than span-level (with an overlap threshold) is to reduce sensitivity to formatting differences such as whitespace and punctuation, which can disproportionately affect short spans. This evaluation approach is inspired by (Da San Martino et al., 2019a).

6 Results and System Descriptions

The shared task drew broad engagement: **51 teams** registered, and **22** ultimately submitted solutions. Nine of these teams participated in both subtasks, while eleven entered only the Technique Classification track and two focused solely on Span Identification. In total, 595 submissions were evaluated — 386 for Technique Classification and 209 for Span Identification.

6.1 Overall Results Summary

This section provides an overview of the top performing systems submitted to the UNLP 2025 Shared Task.

Tables 4 and 5 present the final private leaderboard scores for both shared task tracks. The top performing teams achieved strong results across both tasks, with Team GA securing first place

[7] https://www.kaggle.com/docs/competitions#making-a-submission

[8] https://www.kaggle.com/code/vladimirsydor/multilabel-f1-macro

[9] https://www.kaggle.com/code/woters/f1-token?scriptVersionId=217767698

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.

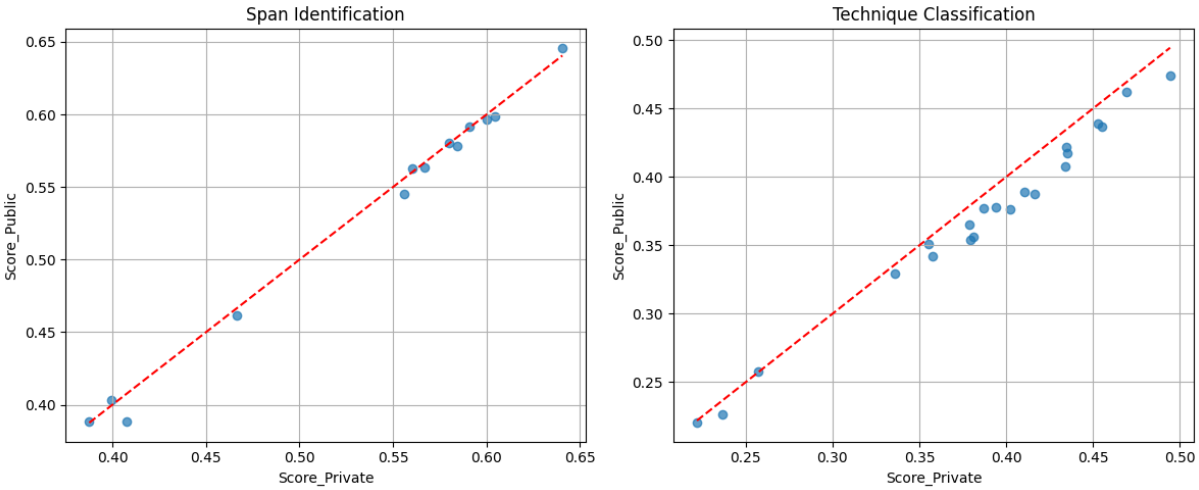

Figure 1: Public and private leaderboard scores for different subtasks.

in each subtask. CVisBetter_SEU and MolodiAmbitni also achieved consistently high rankings, placing within the top three for each task. The competition attracted a diverse set of participants who explored a wide range of modeling approaches, ranging from multilingual transformer baselines to large instruction-tuned language models and custom ensemble pipelines.

6.2 Team GA

Technique Classification Team GA (Bazdyrev et al., 2025) experimented with a range of models, including mDeBERTa[10], Aya101[11], LLaMA3[12], and Mistral Large[13]. Ultimately, they selected Gemma 2-27B (a decoder-only model)[14] due to its superior performance. To address class imbalance, the team optimized classification thresholds using a grid search regularized according to class distribution, replacing the default 0.5 threshold. To improve generalization, the final prediction was obtained by averaging the outputs of models trained on different cross-validation folds (out-of-fold ensemble). This approach led to state-of-the-art results with a significant performance margin.

Span Identification For the span detection task, Team GA explored both encoder-only architectures (mBERT[15], XLM-RoBERTa[16], EuroBERT[17], mDeBERTa) and decoder-only LLMs. Based on their findings, mDeBERTa was the most effective among smaller encoder-based models. However, they hypothesized that large decoder-only models could outperform them due to scale and pretraining advantages. To overcome the uni-directionality limitations of decoder models, the team developed a custom encoder-like architecture for bidirectional attention, using Gemma 2-27B as a base. They pretrained this model on Ukrainian and russian news corpora with a masked language modeling objective, then fine-tuned it on the shared task dataset. The model used a character-level binary labeling approach instead of BIO tagging, and thresholds were again optimized via grid search. The final solution was an ensemble of models from all folds.

6.3 Team MolodiAmbitni

Technique Classification MolodiAmbitni team (Akhynko et al., 2025) used a multistage fine-tuning pipeline based on instruction-tuned Gemma 2-2B using LoRA (Hu et al., 2021). The prompt

[10] https://huggingface.co/microsoft/mdeberta-v3-base
[11] https://huggingface.co/aya-research/aya-101
[12] https://ai.meta.com/llama/
[13] https://mistral.ai/news/mistral-7b/
[14] https://ai.google.dev/gemma
[15] https://huggingface.co/bert-base-multilingual-cased
[16] https://huggingface.co/xlm-roberta-large
[17] https://huggingface.co/ukr-models/eurobert-base

Rank	Team	Score
1	GA	0.49439
2	MolodiAmbitni	0.46952
3	CVisBetter_SEU	0.45519
4	OpenBabylon	0.45265
5	KCRL	0.43518
6	olehmell	0.43460
7	CUET_DuoVation	0.43388
8	Moneypulator	0.41611
9	Affix	0.41065
10	mediguards	0.40224

Table 4: Leaderboard for Subtask 1: Technique Classification. Final rankings are based on private leaderboard scores.

Rank	Team	Score
1	GA	0.64058
2	CVisBetter_SEU	0.60456
3	MolodiAmbitni	0.60001
4	OpenBabylon	0.59096
5	KCRL	0.58434
6	CUET_DuoVation	0.58023
7	LLMInators	0.56686
8	CUET_EagerBeavers	0.56046
9	potato traders v2	0.55578
10	Taleef Tamsal	0.46652

Table 5: Leaderboard for Subtask 2: Span Identification. Final rankings are based on private leaderboard scores.

included class descriptions and similarity-selected examples. Initial training used causal language modeling, followed by sequence classification. The final classifier combined LLM outputs with CatBoost-based metadata features. Class-specific thresholds were optimized via stratified k-fold cross-validation.

Span Identification For span identification, they fine-tuned XLM-RoBERTa-large for binary token classification. The model incorporated a multi-target classification head and used k-fold cross-validation to select optimal thresholds. This hybrid strategy balanced simplicity with effective regularization.

6.4 Team CVisBetter_SEU

Technique Classification CVisBetter_SEU (Rahman and Rahman, 2025) achieved third place in the classification task by fine-tuning XLM-RoBERTa-large[18] in a multilingual setting. To mitigate class imbalance, they applied a weighted binary cross-entropy loss with capped class weights, along with label smoothing (Szegedy et al., 2016) and word-level data augmentation. The architecture was enhanced with a GELU-activated (Hendrycks and Gimpel, 2016) pre-classifier and multi-sample dropout (Inoue, 2019). Training employed AdamW (Loshchilov and Hutter, 2017) optimization with a cosine scheduler, gradient accumulation, and early stopping. Per-class thresholds were dynamically tuned based on F1 score improvements. Additional preprocessing and language heuristics were used to handle Ukrainian and russian text.

[18] https://huggingface.co/xlm-roberta-large

Span Identification For span identification, they used XLM-RoBERTa-large with BIO tagging and formulated the task as token classification. To improve learning across model layers, they employed Layer-wise Learning Rate Decay (Howard and Ruder, 2018). They addressed token-level class imbalance with a weighted focal loss (Lin et al., 2017) and used early stopping to prevent overfitting. Post-processing merged adjacent span predictions with a threshold-based strategy. Training used balanced sampling and a token-level F1 evaluation metric. This system achieved second place in the competition with a private F1 score of 0.60456.

7 Conclusion

We believe that the UNLP 2025 Shared Task is instrumental in facilitating research on propaganda detection and span-level manipulation identification in Ukrainian-language social media content. Teams explored a variety of techniques — from threshold optimization and span post-processing to LoRA fine-tuning and multi-stage inference pipelines — demonstrating the creative potential of the NLP research community when working in low-resource settings.

All datasets used in the shared task are publicly available on GitHub, and all participating teams agreed to open-source their final systems. This ensures the reproducibility of results and contributes to the development of more accessible and transparent models for the Ukrainian language. Top-performing systems employed models such as Gemma 2-27B, XLM-RoBERTa, and mDeBERTa.

We hope this shared task will serve as a foundation for future work in Ukrainian NLP, and that

the tools, data, and approaches developed through this competition will continue to support progress in trustworthy AI systems for media analysis.

8 Ethics Statement

To ensure equal opportunities for all participants and to promote the development of reproducible and accessible solutions for the broader research community, the organizers of the shared task imposed clear restrictions on data and techniques that could be used.

By participating in the shared task, all teams agreed to abide by the following terms and conditions:

- Participants committed to fair and ethical conduct, refraining from the use of any illegal, malicious, or otherwise unethical methods to gain an unfair advantage.
- Participants agreed not to distribute, leak, or share the test data provided during the shared task with any external parties.
- Participants agreed to make their final solutions publicly available after the competition to support open research and contribute to the advancement of Ukrainian NLP.

To the best of our knowledge, all participants complied with these rules throughout the duration of the shared task.

9 Limitations

While the UNLP 2025 Shared Task advances research on propaganda detection in Ukrainian, several limitations must be acknowledged.

Dataset Scope. The dataset used in this shared task is limited to Ukrainian Telegram posts, which may not fully represent the diversity of manipulative content across other platforms (e.g., Facebook, YouTube).

Technique Granularity. Although the task includes ten manipulation techniques, the label set may still be coarse-grained compared to the nuanced range of real-world strategies. Some techniques may overlap semantically or appear jointly in a single sentence, making clear-cut classification difficult.

Dataset Split. Although the dataset split strategy ensured a similar distribution of manipulation techniques across sets and resulted in high score correlations, it does not fully reflect a real-world scenario. Future work should consider incorporating both time and group-based validation strategies. In such settings, there would be no overlap between information sources (e.g., Telegram channels) and no overlap in publication time. Ideally, the private test period should chronologically follow the public one, and the training data should precede both.

Evaluation Metrics. While we used standard metrics, these may not fully capture the interpretability or societal impact of propaganda detection models. Future work could explore human-centered evaluation or robustness under adversarial conditions.

Acknowledgments

We would like to thank the Texty.org.ua team for their crucial contribution to the UNLP 2025 Shared Task. They provided the annotated dataset used in both subtasks, enabling the development and evaluation of systems for manipulation detection in Ukrainian social media.

Parts of this paper were refined with the help of ChatGPT for language clarity and proofreading.

References

Kateryna Akhynko, Oleksandr Kosovan, and Mykola Trokhymovych. 2025. Hidden Persuasion: Detecting Manipulative Narratives on Social Media During the 2022 Russian Invasion of Ukraine. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP) @ ACL 2025*, page to appear. Association for Computational Linguistics.

Firoj Alam, Shaden Shaar, Alex Nikolov, and 1 others. 2021. A survey on nlp for fake news detection. *Computational Linguistics*, 47(4):905–960.

Alberto Barrón-Cedeño, Ibrahim Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 60–64.

Anton Bazdyrev, Ivan Bashtovyi, Ivan Havlytskyi, Oleksandr Kharytonov, and Artur Khodakovskyi. 2025. Transforming Causal LLM into MLM Encoder for Detecting Social Media Manipulation in Telegram. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP) @ ACL 2025*,

page to appear. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Giovanni Da San Martino, Alberto Barrón-Cedeño, James Glass, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414.

Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, and 1 others. 2019a. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

Yevgeniy Golovchenko, Arkaitz Zubiaga, and 1 others. 2023. Detecting propaganda in russian and ukrainian: Challenges and resources. *Computational Propaganda Studies*, 7(2).

Natalia Grabar and Thierry Hamon. 2024. Study of the propaganda techniques occurring in Russian newspaper titles in 2022. In *METAPOL*, Liège, Belgium. université de Liège.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Abdur Rahman and Ashiqur Rahman. 2025. Detecting Manipulation in Ukrainian Telegram: A Transformer-Based Approach to Technique Classification and Span Identification. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP) @ ACL 2025*, page to appear. Association for Computational Linguistics.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 145–158. Springer.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Ivan Taras, Oksana Lytvyn, and Andrii Koval. 2024. Deep learning for disinformation detection in ukrainian telegram channels. *arXiv preprint arXiv:2503.05707*.

Yirong Zeng, Xiao Ding, Yi Zhao, Xiangyu Li, Jie Zhang, Chao Yao, Ting Liu, and Bing Qin. 2024. RU22Fact: Optimizing evidence for multilingual explainable fact-checking on Russia-Ukraine conflict. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14215–14226, Torino, Italia. ELRA and ICCL.