Detecting political manipulations techniques in Internet posts: thresholds auto-selection in multiclass decision space

Iurii Krak[1,2], Maryna Molchanova[3,*], Olexander Mazurets[3,*], Volodymyr Didur[3], Valeriia Klimenko[3], Olena Sobko[3] and Olexander Barmak[3]

[1] *Taras Shevchenko National University of Kyiv, Ukraine*
[2] *Glushkov Institute of Cybernetics of NAS of Ukraine, Kyiv, Ukraine*
[3] *Khmelnytskyi National University, Khmelnytskyi, Ukraine*

Abstract

A solution of the problem of political manipulations techniques detection in Internet posts that uses a thresholds auto-selection optimization was proposed in the paper. Approach consists of using two tracks: the track of neural network models fine-tuning by One-vs-Rest strategy with thresholds auto-selection in multiclass decision space, and the track of detecting political manipulations techniques in Internet posts. The main contribution of paper is development method for neural network models fine-tuning by One-vs-Rest strategy with thresholds auto-selection optimization. The method differs from existing ones by using individual auto-selection thresholds optimization for detecting techniques and using the One-vs-Rest strategy for fine-tuning neural network models. This allows more accurately take into account semantic markers characteristic of each technique and increase the detecting manipulations accuracy. Conducted researches have established that use of One-vs-Rest strategy for fine-tuning the RoBERTa neural network model provided increase of detection accuracy by F_1 macro-metric compared to existing analogues from 0.625 to 0.73; use of One-vs-Rest strategy in combination with thresholds auto-selection optimization provided additional increase in detection accuracy by F_1 macro-metric to 0.76. In general, the proposed approach provides an increase in detection accuracy by macro-metric F_1 by 0.135.

Keywords

political manipulations techniques, One-vs-Rest, NLP, BERT, RoBERTa, thresholds auto-selection

1. Introduction

In modern society, information manipulation has become a widespread practice, covering various areas – from politics to advertising, media and social networks. Due to this, in conditions of information overload, manipulation is increasingly becoming a tool for influencing public opinion and people's behavior [1, 2]. At the same time, the development of artificial intelligence technologies has significantly accelerated the process of spreading manipulation, since now content can be generated not only by people, but also by automated systems that are able to create texts that are almost indistinguishable from materials written by people [3, 4]. This creates new challenges, since manipulative content is becoming more difficult to detect and analyze. Therefore,

CLW-2025: Computational Linguistics Workshop at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), May 15–16, 2025, Kharkiv, Ukraine
* Corresponding author.
yuri.krak@gmail.com (I. Krak); momolchanova@gmail.com (M. Molchanova); exechong@gmail.com (O. Mazurets); pravetz@ukr.net (V. Didur); ler.klimenko.8@gmail.com (V. Klimenko); olenasobko.ua@gmail.com (O. Sobko); alexander.barmak@gmail.com (O. Barmak)
000-0002-8043-0785 (I. Krak); 0000-0001-9810-936X (M. Molchanova); 0000-0002-8900-0650 (O. Mazurets); 0009-0008-2279-1487 (V. Didur); 0000-0001-5869-4269 (V. Klimenko); 0000-0001-5371-5788 (O. Sobko); 0000-0003-0739-9678 (O. Barmak)

along with the development of these technologies, there is a need to develop new artificial intelligence models capable of detecting such manipulative techniques in texts [5, 6].

Currently, there are methods and approaches that allow automatic detection of manipulative techniques, but this issue has not yet been fully explored. One of the main aspects is the definition of a threshold for each manipulative technique at which it can be stated that this technique is present in the text [7]. This threshold is important for the correct classification of manipulations, since different techniques have different semantic markers [8]. For example, manipulation through emotional load will have different signs than a technique aimed at inducing feelings of guilt or fear. The definition of these thresholds is key to the accurate detection and classification of manipulations, since for each technique their presence requires different assessment criteria and different levels of confidence in specific markers in the context of the text [9, 10].

The paper aim is to increase the accuracy of political manipulative techniques classification of in Internet posts by optimizing threshold solutions.

The main paper contributions is the development of method for neural network models fine-tuning by One-vs-Rest strategy, which includes the thresholds auto-selection optimization in multiclass decision space. Method differs from existing ones by the use of individual threshold values for each type of manipulation technique, which allows for more accurate account of various semantic markers characteristic of each of the techniques. In addition, the use of the One-vs-Rest training strategy is proposed, which allows for the classification of manipulative techniques set into separate classes, taking into account the specific characteristics of each of them. This allows for higher accuracy of detecting manipulations in Internet posts, in particular in the conditions of multi-class tasks, where each class may have a different level of manifestation depending on context and specifics of manipulation technique.

2. Related works

The problem of detecting and classifying political manipulative techniques in text messages has been widely studied by scientists in recent years.

So, in [11] noted that previous studies have mainly focused on linguistic features for detecting propaganda in texts, but the role of semantic features in the spread of propaganda remains understudied. In this direction, the authors propose a meta-learning-based method for automatically detecting semantic propaganda at the sentence level in news, using multi-task learning to detect semantic contradictions. The method combines conditional random fields (CRF), bidirectional LSTM networks (BiLSTM) and pre-trained language models, which allows achieving an F_1 score of 0.61 on multilingual data and 0.688 on monolingual data. The authors note that the proposed model outperforms existing approaches, confirming the effectiveness of multi-task learning for detecting disinformation tactics in news.

The study [12] proposes a multilingual system for detecting propaganda that uses ensembles of models with different architectures and prediction aggregation methods (Use-FFN and Skip-FFN). Results in seven languages (English, Arabic, German, Italian, French, Polish, Russian) showed that the MultiProp-Chunk Hybrid model outperformed the others in Arabic and Russian, with F_1-micro results of 0.598 and 0.595, respectively. The MultiProp-Baseline En-B model demonstrated stable results in Polish and Italian (F_1-macro up to 0.625 for Polish), and the MultiProp-ML Hybrid achieved strong results in cross-lingual adaptation, with results for French and German of 0.587 and 0.583, respectively. These results highlight the effectiveness of the system in multilingual propaganda analysis, where the use of meta-learning and specialized models for each language allows for significant improvements compared to traditional methods.

Political manipulations in news often aim to manipulate public opinion through psychological and rhetorical strategies.

In [13] an advanced pre-trained language model RoBERTa is used to detect propaganda manipulations in news articles. The model is evaluated using the SemEval-2020 Task 11 dataset,

which was used for this task. The results show that the RoBERTa model, thanks to use of word vectors, detects complex propaganda techniques and achieves an F_1-score of 60.2%.

The study [14] presents an ensemble model that solves the problem of detecting propaganda techniques in texts extracted from memes. The paper also considers modern pre-trained language models and optimization techniques, such as data augmentation and model ensemble. The model was evaluated using the SemEval-2021 Task 6 dataset, and the results showed that the proposed system achieved an F_1-micro score of 0.604 on the test set.

In the study [15] two architectures for classifying propaganda techniques were considered: one with and without data augmentation (EDA). The models with EDA showed a 3% improvement in F_1-score and achieved 57.57% on the test set. Most propaganda techniques, such as "Appeal_to_fear-prejudice", "Exaggeration, Minimisation" and "Repetition", showed an increase in performance, although some techniques, such as "Doubt" and "Flag-Waving", showed a slight decrease. The largest improvement was observed for the techniques "Causal_Oversimplification" and "Thought-terminating_Cliches". The optimal parameters for the classification tasks were established based on the analysis of epochs, sentence length and learning rate, which allowed achieving an F_1-score of 0.44 for the sentiment detection task and 0.57 for the propaganda technique classification task.

In the context of machine learning models studied in [16] for detecting propaganda content, the Stacking Classifier using feature processing methods such as Word2Vec and TF-IDF demonstrates high versatility and flexibility. This approach integrates multiple representations of complex features and predictive models, making it effective for solving complex text classification tasks. The performance analysis of different models shows that the Stacking Classifier outperforms other models, including Naive Bayes, SVM, KNN, Logistic Regression, and Random Forest. The inclusion of feature engineering significantly improves the performance of the model, as evidenced by the increase in Accuracy, Precision, and F_1 scores compared to other methods. The Stacking Classifier with TF-IDF and Word2Vec achieved Accuracy of 87%, Precision of 81%, and F_1 score of 84%. These indicators exceed the results of other tested models; however, the testing was carried out on a sample with a large class imbalance. The model demonstrates the ability to adapt to various text categorization tasks, making it an effective tool for detecting propaganda content, particularly in poster headlines.

In the study [17], a two-step process was used to evaluate the model's performance in determining the optimal threshold for classifying propaganda techniques. First, experiments were conducted with macrothresholds ranging from 0.1 to 0.9, the threshold with the highest F_1 score was selected, and then microthresholds were added for further optimization. The XLM-RoBERTa models were trained using the Adam optimizer, and early termination was used to prevent overtraining. The performance metrics accuracy, precision, recall, and F_1 score were used at each stage. The results of the study showed that the model effectively classified propaganda content, but the accuracy for the "propaganda" and "non-propaganda" classes had a significant difference, indicating an imbalance in the data. The following results were obtained by evaluating the model at different thresholds for classification. The standard threshold of 0.5 allowed to achieve an accuracy of 0.85 and an F_1 measure of 0.83 for the main techniques. For some categories, such as "Enemy Creation," increasing the threshold to 0.7 increased the accuracy to 0.92, although it decreased the sensitivity on less represented classes. This confirms the importance of tuning the threshold to optimize the results depending on the specific classification goals, in order to achieve a balance between accuracy and sensitivity.

The related works review found that existing methods for classifying political manipulative techniques in Internet posts have several significant limitations. One of the key shortcomings is the use of universal threshold values for all classes, which does not take into account the specifics of each manipulative technique. This leads to decrease in accuracy, since different techniques have different levels of semantic expressiveness and frequency in text content.

In addition, most approaches use traditional multi-class classification strategies that do not provide adequate separation between classes with high level of feature overlap. This makes it

difficult to correctly identify combined or weakly expressed manipulative techniques. Insufficient attention to the individual characteristics of manipulations techniques complicates the results interpretation and reduces the models practical effectiveness. Study hypothesizes that One-vs-Rest training strategy application in combination with thresholds auto-selection optimization for each manipulative technique will improve classification accuracy.

3. Methods and materials

3.1. Approach to detecting manipulations techniques using thresholds auto-selection

Approach to detecting political manipulations techniques in internet posts with thresholds optimization consists of using two tracks: the track of neural network models fine-tuning by One-vs-Rest strategy with thresholds auto-selection in multiclass decision space [18] and the track of detecting political manipulations techniques [19]. Generalized diagram of man interaction with tracks of political manipulations techniques detection is shown in Figure 1.

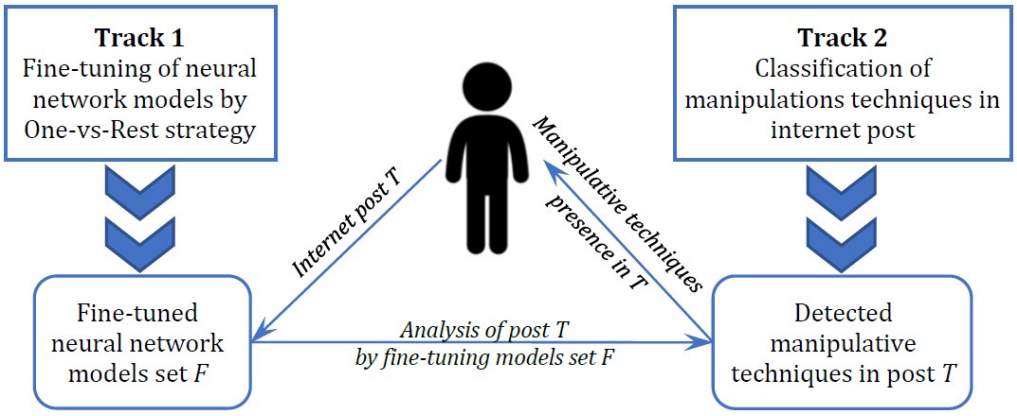

Figure 1: Man in the loop of tracks of detecting manipulation influences

The result of running Track 1 is set of finely tuned models F, each of which is trained using the One-vs-Rest strategy for corresponding manipulative technique. Set of models is defined as:

$$F = \{f_{t_1}, f_{t_2}, \dots, f_{t_{|T|}}\} \quad (1)$$

where each model f_{t_i} corresponds to separate manipulative technique t_i.

The manipulative techniques set T has a dimension equal to the number of political manipulations techniques identified in research:

$$|T| = 10 \quad (2)$$

In turn, political manipulation techniques set considered in research:

$$T=\{\textit{Loaded Language, Glittering Generalities, Euphoria, Appeal to Fear, FUD (Fear, Uncertainty, Doubt), Bandwagon/Appeal to People, Thought-Terminating Cliche, Whataboutism, Cherry Picking, Straw Man}\} \quad (3)$$

The limitation of set T is explained by the composition of the available dataset [20], which contains annotated examples only for the specified manipulative techniques. Thus, each technique t_i corresponds to a separate model f_t, which allows for independent training and classification using the One-vs-Rest strategy.

The result of Track 2 execution are the identified manipulative techniques in Internet posts. The political manipulations technique is considered used if the neural network value is above the

threshold. The threshold is determined adaptively for each manipulative technique on Track 1 using the Youden criterion [19].

The implementation of Track 1 as a method for neural network models fine-tuning will be considered in detail in Section 3.2. At the same time, the implementation of Track 2 consists in using the results of Track 1 using the previously developed method for political propaganda detection [20].

3.2. Method for neural network models fine-tuning by One-vs-Rest strategy

Method for neural network models fine-tuning by One-vs-Rest strategy with thresholds auto-selection optimization in multiclass decision space is intended for further use for political manipulations techniques detection. Scheme of method for neural network models fine-tuning by One-vs-Rest strategy is shown in Figure 2. Step 1 of the method for neural network models fine-tuning is devoted to datasets preprocessing for training by One-vs-Rest strategy, which extracts fragments from the input dataset that are annotated by the authors as expressing manipulative techniques, processes them accordingly, and places them in the appropriate catalogs. Step 2 consists of fine-tuning neural network binary classification models, which involves training separate binary classification by transformer neural network models for each manipulation technique. Step 3 consists in optimization of classification threshold for each manipulative technique, which allows improving the performance of the corresponding model. Step 4 evaluating of models efficiency by metrics and the trained neural network models will be evaluated using the following set of metrics: Accuracy, Precision, Recall, F_1 measure.

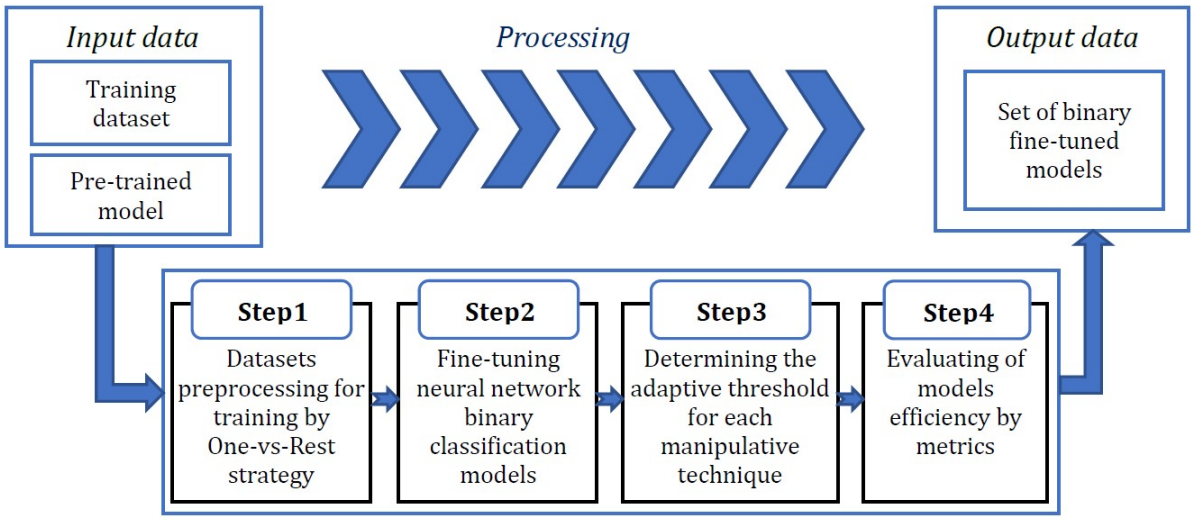

Figure 2: Scheme of method for neural network models fine-tuning by One-vs-Rest strategy

The input data is the dataset [21], created as part of the Fourth Ukrainian Workshop on NLP (UNLP 2025), in which the authors participated. The workshop was dedicated to solving the problem of detecting political manipulation techniques in social networks. The dataset contains marked data at the fragment level. The distribution of documents in the dataset without pre-processing is shown in Figure 3.

The total number of unique documents in the dataset is 3822, of which 2589 are marked as having manifestations of the use of manipulative techniques, and 1233 as not having manifestations of manipulative influences. Each document that has manifestations of the use of manipulative techniques can have more than 1 label. The records in the dataset are presented as Ukrainian, however, after the analysis it was found that records are also found in other languages. Before training the neural networks, typical text preprocessing operations [22] are performed on the elements of the dataset.

Also, the input data is a pre-trained model of the transformer architecture. The study compared the BERT [23] and RoBERTa [24] models, the choice of which is due to the support of the Ukrainian language. The choice of BERT and RoBERTa allows comparing a universal and specialized transformer model for the analysis of the Ukrainian language, which makes the study more representative.

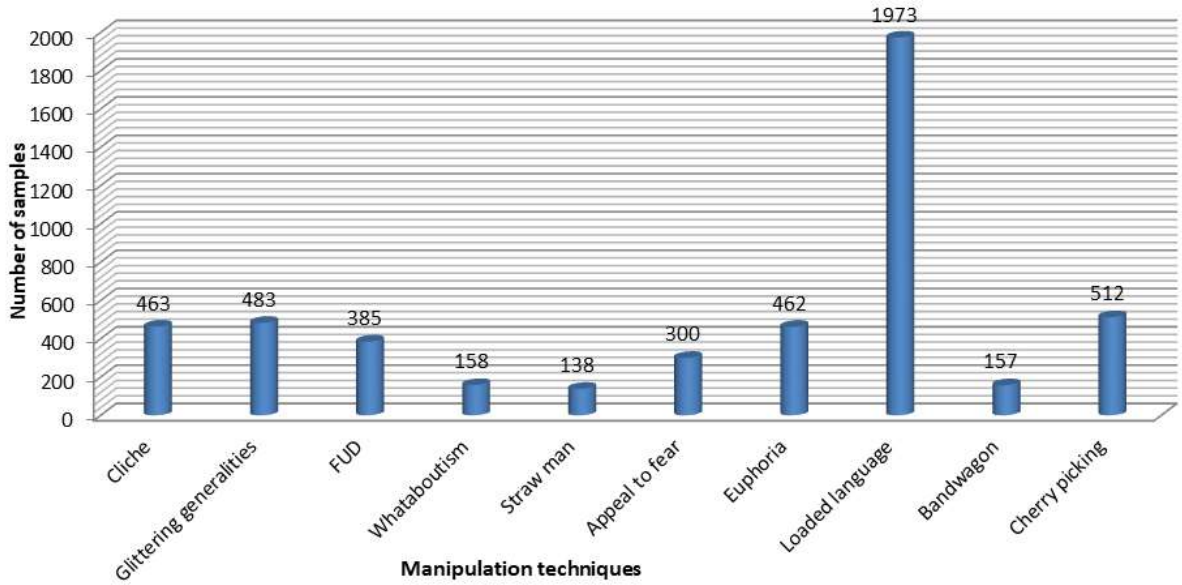

Figure 3: Distribution of text documents by manipulative techniques in dataset

Method consists of steps: datasets preprocessing for training by One-vs-Rest strategy (described in section 3.2.1), fine-tuning neural network binary classification models (described in section 3.2.2), determining the adaptive threshold for each manipulative technique (described in section 3.2.3) and evaluating of models efficiency by metrics (described in section 3.2.4).

3.2.1. Datasets preprocessing for training by One-vs-Rest strategy

Step 1 of method for neural network models fine-tuning by One-vs-Rest strategy is datasets preprocessing for training. From the input dataset D containing text fragments x_i and their labels $Li \subseteq T$, fragments annotated by authors as expressing manipulative techniques are extracted. Each fragment x_i that has label t is added to the corresponding catalog C_t:

$$C_t = \{x_i | t \in L_i\} \quad (4)$$

Accordingly, each fragment is placed in a separate file, which is placed in a directory with the same name as the name of the manipulative technique. If there are several techniques within the document, fragments are duplicated in directories with the same name for the techniques expressed. Figure 4 shows the process of distributing fragments into directories. After the specified distribution, the following distribution was obtained, shown in Figure 5.

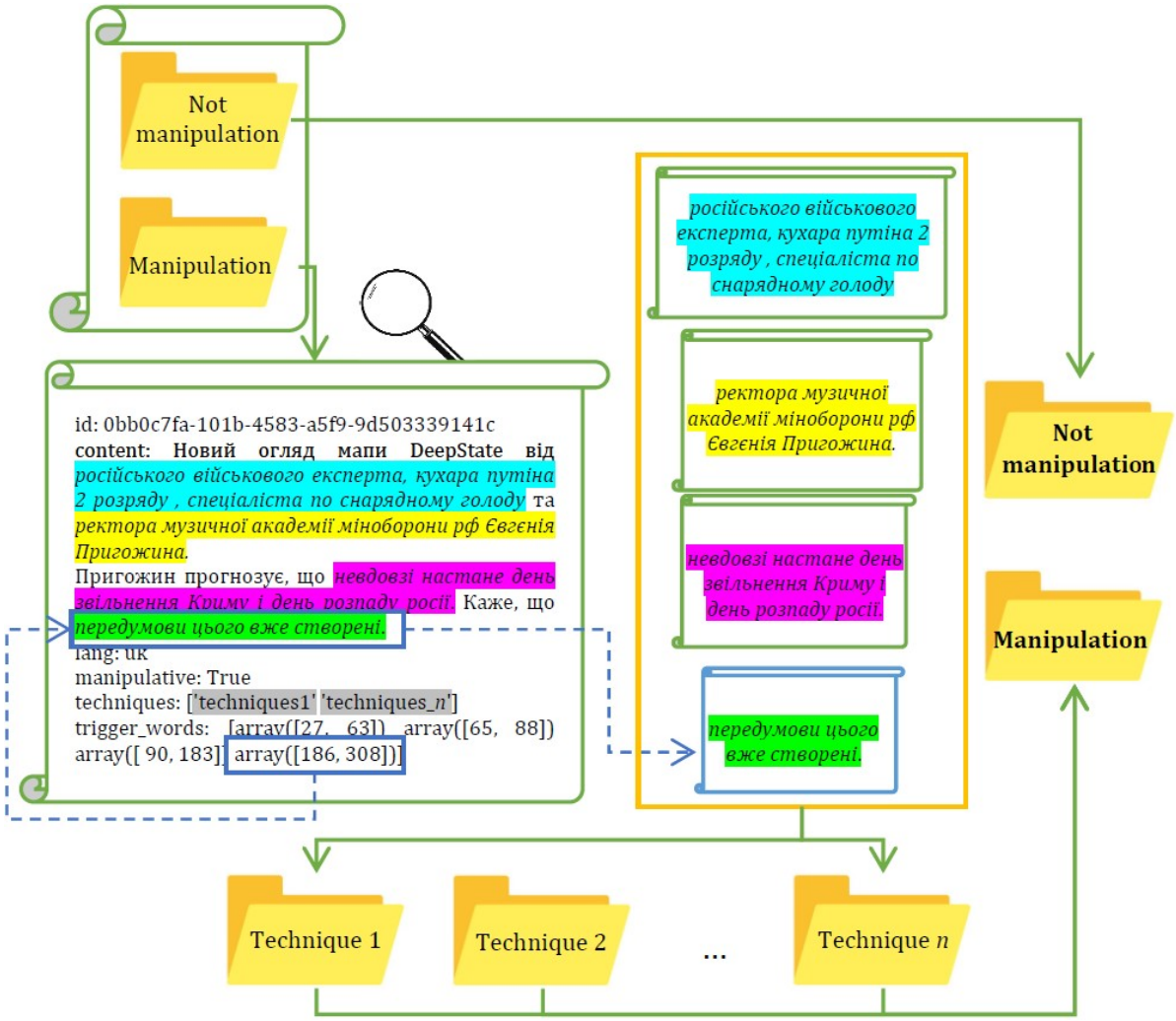

Figure 4: Distribution of text fragments into directories according to manipulation techniques

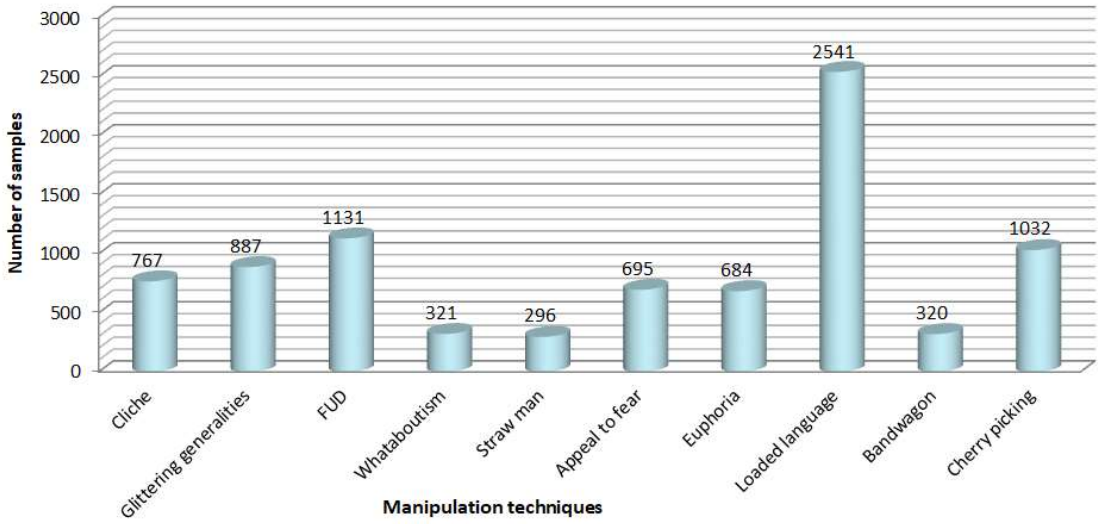

Figure 5: Distribution of texts by manipulative techniques after preprocessing

Files in the resulting directories are filtered by size; if the file is less than 100 bytes, it will not participate in fine-tuning binary BERT models of similar architectures. If the dataset contains texts in languages other than Ukrainian, they are automatically translated.

3.2.2. Fine-tuning neural network binary classification models

Fine-tuning neural network binary classification models is Step 1 of method for neural network models fine-tuning by One-vs-Rest strategy. This step involves training separate binary BERT classification models for each manipulation technique. Since the sample has some imbalance, the child datasets were formed according to certain rules:

- All samples are selected from the target catalog (positive samples) expressing text messages of a certain manipulative influence:

$$D_{t_i}^{pos} = C_{t_i} \tag{5}$$

All samples from the equipment catalog t_i are included in the positive class. Accordingly, the sample will have dimension $|D_{t_i}^{pos}| = N_{t_i}$.

- The non-target catalog is added $\alpha^*100\%$ of texts (from the dimension of the target class) that do not contain manipulative influences, and $\beta^*100\%$ of texts from other catalogs that contain manifestations of other, different from the target, manipulative techniques:

$$D_{t_i}^{neg} = D_{clear}^{sam} \cup D_{\neg t_i}^{sam} \tag{6}$$

where D_{clear}^{sam} are random texts without manipulations (sample size α^*N_t), $D_{\neg t_i}^{sam}$ are random texts with other techniques than the target one. The hyperparameters α and β determine the sample balance and satisfy the equation:

$$\alpha+\beta = 1,\ \alpha,\beta\in[0,1] \tag{7}$$

Thus, the general dataset for fine-tuning the f_{t_i} model will look like this:

$$D_{t_i} = D_{t_i}^{pos} \cup D_{t_i}^{neg} \tag{8}$$

The influence of the values of the parameters α and β for the content of the non-target sample requires a separate study, which will be performed in further work, within the framework of the study the parameters will be: α=0.5 and β=0.5. The use of non-target manipulative techniques in the non-target class will help the model distinguish precisely the target manipulative technique, and not just the presence of manipulation in general [20, 25].

3.2.3. Determining the adaptive threshold for each manipulative technique

Step 3 of method for neural network models fine-tuning by One-vs-Rest strategy is optimization of classification threshold for each manipulative technique t_i, which allows improving the performance of the corresponding model f_t. For each model f_{ti}, the probability p of belonging of text x to class t_i is calculated:

$$p_{t_i}(x) = P(y = t_i|x) \tag{9}$$

Classification is carried out according to the threshold rule:

$$\hat{y}_{t_i} = \begin{cases} 1, & p_{t_i}(x) > \tau_{t_i} \\ 0, & else \end{cases} \tag{10}$$

where τ_{t_i} is the optimal threshold for the manipulative technique t_i. In this study, the Youden criterion [19] will be used, which is used to select a threshold value that provides the optimal balance between «True Positive Rate» and «False Positive Rate»:

$$J(\tau) = TPR(\tau) - FPR(\tau) \tag{11}$$

The Youden criterion [19] is calculated as the difference between the sensitivity and the level of false positives, allowing to find the threshold at which the model demonstrates the maximum

ability to distinguish texts containing a specific manipulative technique from those that do not contain it. Optimization of this indicator helps to avoid an excessive number of false positives and improves the quality of recognition. The adaptive threshold for each technique allows to flexibly adjust the classification according to the specifics of the manipulative effect, since different techniques can have different levels of expressiveness in texts, using a single threshold value for all cases can lead to a decrease in the efficiency of classification. Determining the optimal threshold separately for each class allows to achieve better differentiation and improve the overall performance of the model.

3.2.4. Evaluation of neural network models efficiency by metrics

Step 4 of method for neural network models fine-tuning by One-vs-Rest strategy is evaluation of neural network models efficiency by metrics. In the research, the trained neural network models will be evaluated using the following set of metrics: Accuracy, Precision, Recall, F_1 measure [26]. This set of metrics is sufficient for evaluating neural network models within the framework of the classification of political manipulative techniques, given the specifics of the task, in particular, the multi-class nature of the classification and the imbalance in the distribution of classes.

Since each neural network model is responsible for a specific political manipulative technique, such an assessment will allow us to accurately determine where each model works well and where it needs improvement, allowing us to adapt threshold values and improve the overall results of the classification of political manipulative techniques [20].

4. Experiments

To research the proposed hypothesis about the feasibility of applying the One-vs-Rest learning strategy in combination with an adaptive threshold to each manipulative technique to increase accuracy, a series of experiments will be conducted with BERT-like architectures that support work with the Ukrainian language. The scheme of experiment is shown in Figure 6.

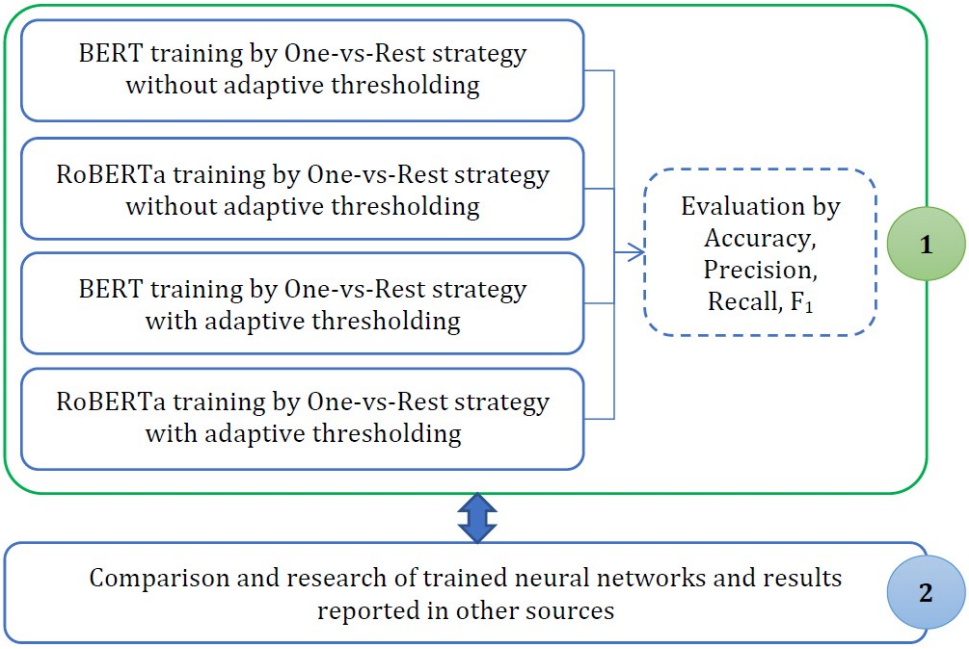

Figure 6: Applied research map of developed method

The first part of the experiment is training neural network models of the BERT and RoBERTa architectures. To train neural network models using the One-vs-Rest strategy with and without adaptive thresholding, a console application was created in Python, which uses the PyTorch [27],

transformers [28], datasets [29] libraries. The result of the application is the saved trained models and their evaluations by metrics and optimal threshold.

Also, as part of the second part of the study, a desktop application was created that uses the PySide [30], PyTorch, transformers libraries and allows you to evaluate the quality of detecting manipulative techniques. The appearance of the developed application is shown in Figure 7.

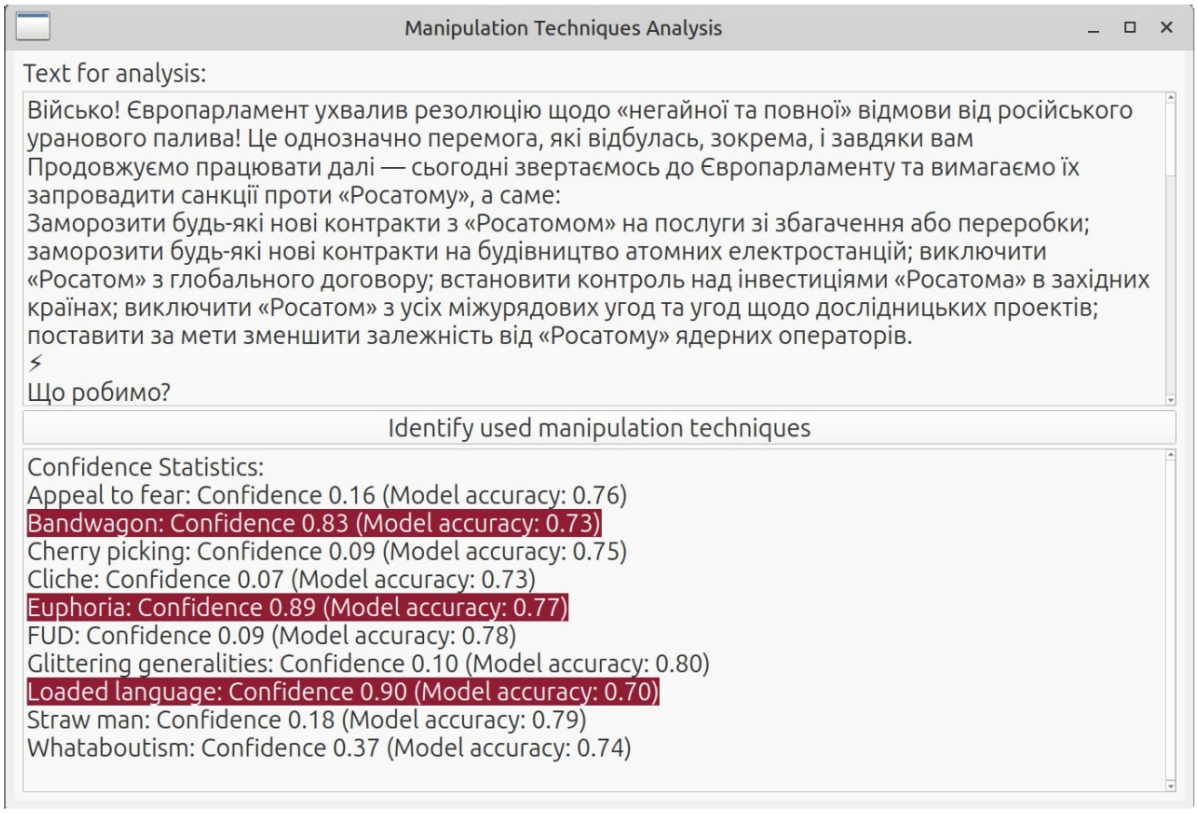

Figure 7: Experimental software for evaluating quality of manipulative techniques detection

The next section will present graphs comparing the performance of the trained versions of neural networks and the main results obtained.

5. Results and discussion

After training the classifiers on the datasets formed according to transformations (4)-(8) for implementing the One-vs-Rest strategy, the results were obtained on the training (80% of the total dataset) and test samples (20% of the total dataset), shown in Table 1.

The analyzed data presented in Table 1 allow to conclude that in general, RoBERTa trained using the One-vs-Rest strategy demonstrates higher performance on the test set for most techniques, especially in the Precision and F1-score metrics. This indicates that this model is more resistant to generalization and better recognizes patterns of manipulative rhetoric.

BERT shows high performance on the training data, but there is a significant decrease in the metrics on the test set for some categories, for example, “Bandwagon/Appeal to People”, which may indicate a certain tendency to overtraining. At the same time, RoBERTa demonstrates more balanced results between Train and Test, especially for techniques such as “Cherry Picking”, “Glittering Generalities” and “Whataboutism”, which confirms its ability to generalize knowledge more effectively.

Table 1
Evaluation of neural networks trained by One-vs-Rest strategy without adaptive thresholding

	Metrics:	Accuracy		Recall		F_1		Precision	
Models	Techniques:	Train	Test	Train	Test	Train	Test	Train	Test
BERT	Appeal to Fear	0.789	0.743	0.789	0.743	0.780	0.732	0.843	0.789
	Bandwagon/ Appeal to People	0.904	0.664	0.904	0.664	0.903	0.658	0.910	0.677
	Cherry Picking	0.747	0.748	0.747	0.748	0.730	0.730	0.825	0.818
	Thought-Terminating Cliche	0.768	0.701	0.768	0.701	0.756	0.685	0.829	0.764
	Euphoria	0.885	0.804	0.885	0.804	0.885	0.804	0.886	0.805
	FUD	0.826	0.765	0.826	0.765	0.821	0.759	0.862	0.808
	Glittering Generalities	0.851	0.764	0.851	0.764	0.849	0.760	0.865	0.777
	Loaded Language	0.685	0.69	0.685	0.690	0.645	0.651	0.789	0.782
	Straw Man	0.763	0.731	0.763	0.731	0.757	0.711	0.796	0.792
	Whataboutism	0.749	0.727	0.749	0.727	0.743	0.717	0.774	0.769
RoBERTa	Appeal to Fear	0.824	0.699	0.824	0.699	0.820	0.688	0.852	0.730
	Bandwagon/ Appeal to People	0.829	0.773	0.829	0.773	0.826	0.768	0.855	0.800
	Cherry Picking	0.872	0.710	0.872	0.710	0.871	0.710	0.874	0.715
	Thought-Terminating Cliche	0.761	0.692	0.761	0.692	0.752	0.684	0.797	0.730
	Euphoria	0.825	0.754	0.825	0.754	0.824	0.751	0.833	0.764
	FUD	0.838	0.779	0.838	0.779	0.838	0.779	0.838	0.779
	Glittering Generalities	0.876	0.783	0.876	0.783	0.876	0.783	0.877	0.788
	Loaded Language	0.717	0.599	0.717	0.599	0.717	0.599	0.717	0.599
	Straw Man	0.835	0.703	0.835	0.703	0.833	0.694	0.855	0.710
	Whataboutism	0.802	0.828	0.802	0.828	0.797	0.822	0.850	0.839

Some manipulative techniques remain difficult for both models, for example, “Loaded Language”, for which RoBERTa has a significant performance decrease on the test data, which may indicate the difficulty of semantic identification of this technique due to its contextual variability.

Thus, RoBERTa is generally a better model for detecting manipulative techniques, demonstrating more stable results between training and test sets.

The evaluation of neural networks using the One-vs-Rest strategy with adaptive thresholding by transformations (9) – (11) is given in Table 2.

Analysis of the results of the classification of manipulative techniques based on the BERT and RoBERTa models demonstrates the effectiveness of the One-vs-Rest strategy with an adaptive threshold. Compared to the baseline indicators (Table 1), the use of this strategy leads to an overall improvement in classification accuracy.

RoBERTa generally demonstrates better results than BERT in most techniques, especially in terms of stability between training and testing metrics. The most noticeable gap in favor of RoBERTa is observed in the manipulative techniques “Whataboutism” and “FUD”, where the F1-score and Precision significantly exceed the similar indicators of BERT. This indicates that the model differentiates these manipulations better and has a lower tendency to overtraining.

Table 2
Evaluation of neural networks trained by One-vs-Rest strategy using adaptive thresholding

	Metrics:	Accuracy		Recall		F_1		Precision	
Models	Techniques:	Train	Test	Train	Test	Train	Test	Train	Test
BERT	Appeal to Fear	0.813	0.764	0.813	0.764	0.813	0.760	0.813	0.787
	Bandwagon/ Appeal to People	0.790	0.727	0.790	0.727	0.782	0.726	0.834	0.727
	Cherry Picking	0.777	0.748	0.777	0.748	0.775	0.734	0.789	0.800
	Thought-Terminating Cliche	0.833	0.730	0.833	0.730	0.832	0.726	0.843	0.749
	Euphoria	0.881	0.775	0.881	0.775	0.881	0.772	0.881	0.799
	FUD	0.824	0.776	0.824	0.776	0.824	0.774	0.826	0.793
	Glittering Generalities	0.881	0.798	0.881	0.798	0.881	0.798	0.882	0.798
	Loaded Language	0.702	0.695	0.702	0.695	0.686	0.666	0.731	0.758
	Straw Man	0.896	0.790	0.896	0.790	0.896	0.781	0.897	0.828
	Whataboutism	0.782	0.742	0.782	0.742	0.782	0.734	0.786	0.779
RoBERTa	Appeal to Fear	0.865	0.743	0.865	0.743	0.865	0.743	0.866	0.743
	Bandwagon/ Appeal to People	0.827	0.797	0.827	0.797	0.825	0.795	0.842	0.808
	Cherry Picking	0.874	0.744	0.874	0.744	0.874	0.742	0.877	0.764
	Thought-Terminating Cliche	0.771	0.705	0.771	0.705	0.770	0.689	0.777	0.783
	Euphoria	0.847	0.768	0.847	0.768	0.847	0.764	0.849	0.790
	FUD	0.856	0.801	0.856	0.801	0.856	0.798	0.856	0.829
	Glittering Generalities	0.879	0.803	0.879	0.803	0.879	0.801	0.880	0.807
	Loaded Language	0.724	0.690	0.724	0.690	0.716	0.650	0.737	0.796
	Straw Man	0.843	0.788	0.843	0.788	0.842	0.777	0.855	0.825
	Whataboutism	0.918	0.820	0.918	0.820	0.918	0.819	0.918	0.820

For BERT, significant improvements are observed in the techniques “Straw Man”, “Glittering Generalities” and “FUD”, where the F_1-score increased compared to the previous approach, indicating the effectiveness of the adaptive threshold in balancing precision and completeness.

Adaptive thresholding had a positive impact on the metrics of the test set. For example, in BERT for the manipulative technique “Bandwagon/Appeal to People” Precision increased from 0.677 to 0.727, indicating a decrease in false positives of the model. Similarly, in RoBERTa Precision for the technique “Straw Man” improved to 0.825, which is an important indicator for recognizing manipulative techniques.

At the same time, adaptive thresholding does not always provide significant improvements for complex techniques such as “Loaded Language”, where in RoBERTa there is still a gap between the training and test metrics (F_1-score dropped from 0.717 to 0.65). This indicates the need for additional optimization of the model for processing context-sensitive expressions.

Comparison of BERT and RoBERTa neural networks on the test set of applying the One-vs-Rest strategy with adaptive thresholding and without the Precision metric is shown in Figure 8.

Comparison of BERT and RoBERTa neural networks on the test set using the One-vs-Rest strategy with and without an adaptive threshold using the Recall metric is shown in Figure 9.

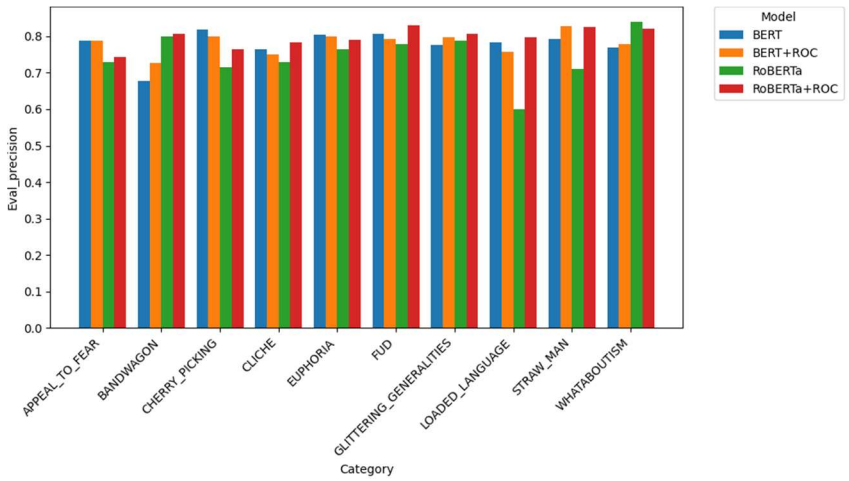

Figure 8: Comparison of detecting political manipulations techniques by BERT and RoBERTa neural networks with and without threshold auto-selection by Precision metric

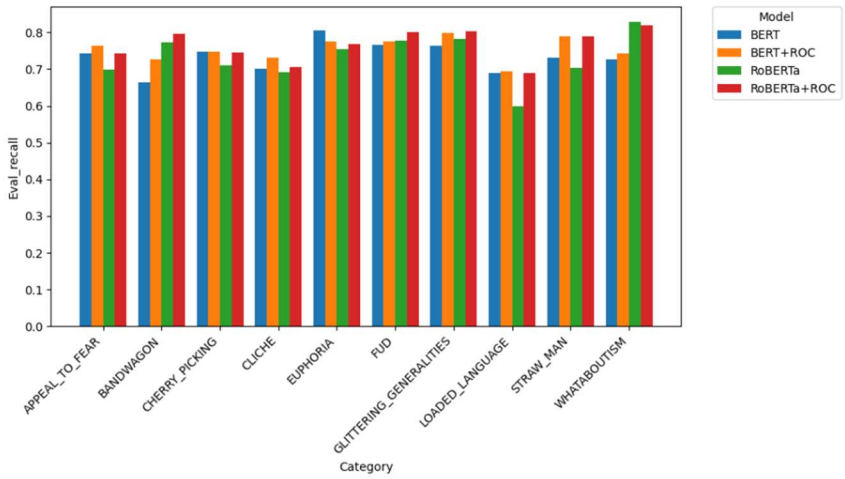

Figure 9: Comparison of detecting political manipulations techniques by BERT and RoBERTa neural networks with and without threshold auto-selection by Recall metric

Comparison of BERT and RoBERTa neural networks on the test set using the One-vs-Rest strategy with and without an adaptive threshold by the F1 metric is shown in Figure 10.

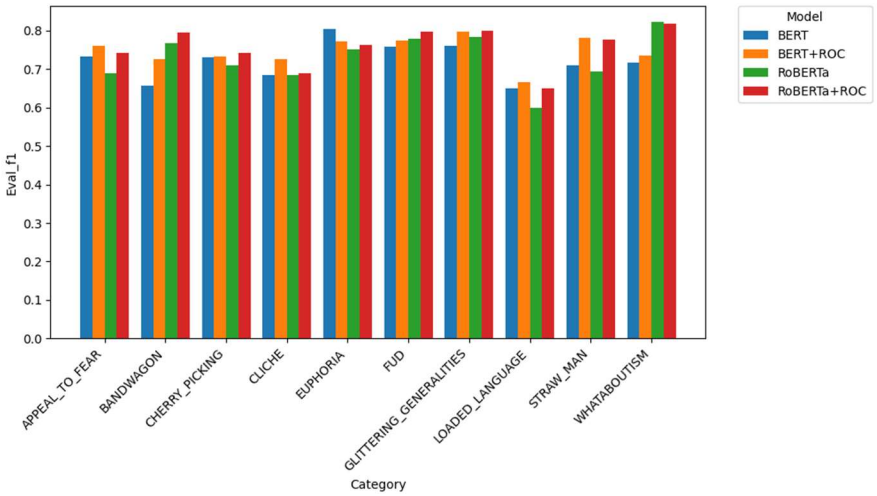

Figure 10: Comparison of detecting political manipulations techniques by BERT and RoBERTa neural networks with and without threshold auto-selection by F_1 metric

Compared to known analogues, this approach allows to obtain higher estimates. Comparison of the approaches with the proposed one is given in Table 3.

Table 3
Comparison of developed methodology and known analogues

Methodology	Language	Macro F_1
BERT One-vs-Rest (developed)	Ukrainian	0.720
BERT One-vs-Rest + τ (developed)	Ukrainian	0.750
RoBERTa One-vs-Rest (developed)	Ukrainian	0.730
RoBERTa One-vs-Rest + τ (developed)	Ukrainian	0.760
CRF+ BiLSTM [11]	multilingual	0.610
MultiProp-Baseline En-B [12]	Polish	0.625
RoBERTa [13]	English	0.602
Ensemble model [14]	English	0.604

According to the data from Table 3, the One-vs-Rest training strategy allows improving the known analogues by at least 0.095 for the implementation of BERT without adaptive threshold. The best increase in the F_1-macro metric is observed for the One-vs-Rest strategy with adaptive threshold, and is 0.135.

The One-vs-Rest strategy with adaptive thresholding improves the generalization ability of the models, which is especially noticeable in the example of RoBERTa, which demonstrates higher resistance to overtraining compared to BERT. The results indicate an improvement in the balance between accuracy and completeness of classification, especially for the techniques "Whataboutism", "Glittering Generalities" and "FUD", where Precision and F_1-score remain consistently high on both training and test samples.

Despite the overall improvement in metrics, for complex categories such as “Loaded Language”, there is a significant gap between training and testing indicators, which may indicate insufficient generalization of the model. This may be a consequence of limited data or high variability of lexical constructions in these cases. In addition, the increase in F_1-score for most techniques indicates that the methodology allows for a more effective balance between detecting positive cases and reducing false positives.

Considering the results obtained, the use of adaptive thresholding in the One-vs-Rest strategy in classification is a promising direction for improving the generalization ability of transformative models in the tasks of identifying propaganda techniques. To further improve the accuracy of classification, it is necessary to consider expanding the training dataset or using additional regularization mechanisms to reduce the gap between training and testing metrics in complex categories.

The proposed method has a number of limitations. Research is based on dataset [16], which contains annotated examples for only 10 manipulative techniques. Accordingly, the methodology does not take into account other possible manipulative strategies. Another limitation is possibility of working only in Ukrainian, other languages were not studied. Method works with text files from 100 to 7514 bytes long with tokenizer size of 512 tokens.

Conclusions

A solution of the problem of political manipulations techniques detection in Internet posts that uses a thresholds auto-selection optimization was proposed in the paper. Approach consists of using two tracks: the track of neural network models fine-tuning by One-vs-Rest strategy with thresholds auto-selection in multiclass decision space, and the track of detecting political manipulations techniques in Internet posts. The One-vs-Rest strategy assumes that each technique is analyzed separately, which allows achieving greater detecting accuracy, avoiding mixing of

various manipulative influences. Pre-trained BERT and RoBERTa transformer neural network models supporting the Ukrainian language were used for fine-tuning. This made it possible to compare universal and specialized architectures for text analysis and determine their effectiveness in detecting manipulations. Optimization of the classification threshold for each model is carried out based on the Youden criterion, which helps to balance between the «True Positive Rate» and «False Positive Rate» indicators. This approach allowed fine-tuning the model to the specifics of each manipulative technique, increasing the overall classification efficiency. Additionally, the use of non-target manipulative techniques in the opposite class to the target one contributes to more accurate distinction between different techniques of political manipulations.

The main contribution of paper is development method for neural network models fine-tuning by One-vs-Rest strategy with thresholds auto-selection optimization. The method differs from existing ones by using individual auto-selection thresholds optimization for detecting techniques and using the One-vs-Rest strategy for fine-tuning neural network models. This allows more accurately take into account semantic markers characteristic of each technique and increase the detecting manipulations accuracy, particularly in multi-class tasks, where each class may have a different level of manifestation depending on the context and the specifics of the political manipulation technique.

To investigate the developed approach, experiments series were conducted on training and evaluating BERT and RoBERTa neural network models using the One-vs-Rest strategy for classifying manipulative techniques in text Internet posts. Software for training models and desktop application were developed to evaluate the models performance. Conducted researches have established that use of One-vs-Rest strategy for fine-tuning the RoBERTa neural network model provided increase of detection accuracy by F1 macro-metric compared to existing analogues from 0.625 to 0.73; use of One-vs-Rest strategy in combination with thresholds auto-selection optimization provided additional increase in detection accuracy by F1 macro-metric to 0.76. In general, the proposed approach provides an increase in detection accuracy by macro-metric F1 by 0.135.

The experiment results showed, that RoBERTa neural network model demonstrates higher overall performance compared to BERT, especially after applying thresholds auto-selection optimization. The largest increase was observed for the manipulative techniques “Whataboutism”, “FUD” and “Glittering Generalities”, where the Precision and Recall metrics remained consistently high on both the training and test samples. The complexity of the classification of some techniques, in particular “Loaded Language”, indicates the need for further research in the direction of expanding the dataset or introducing additional regulation mechanisms.

The methodology has certain limitations. In particular, its practical application depends on the dataset, for example, the dataset used for research contained 10 manipulative techniques. Also, the experiments were conducted exclusively for the Ukrainian language, which does not allow assessing its effectiveness for other languages.

Thus, the results obtained confirm the effectiveness of One-vs-Rest strategy with with thresholds auto-selection optimization in multiclass decision space for improving the accuracy of manipulative techniques classification.

Further research can be aimed at improving the methodology by expanding the dataset, testing on multilingual corpora, and developing more flexible models that can adapt to complex manipulative techniques.

Acknowledgements

This research was made possible thanks to the dataset provided by the UNLP 2025 Shared Task initiative (GitHub repository) [20]. We sincerely appreciate the efforts of the organizers and contributors who curated and shared this valuable resource, enabling further advancements in the study of propaganda detection techniques.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] J. Szwoch, M. Staszkow, R. Rzepka, K. Araki, Limitations of Large Language Models in Propaganda Detection Task, Appl. Sci. 14.10 (2024) 4330. doi:10.3390/app14104330.

[2] M. Hasanain, F. Ahmed, F. Alam, Can GPT-4 identify propaganda? Annotation and detection of propaganda spans in news articles. URL: https://arxiv.org/abs/2402.17478.

[3] S. Vajjala, Generative Artificial Intelligence and Applied Linguistics, JALT J. 46.1 (2024) 55–74. doi:10.37546/jaltjj46.1-3.

[4] A. Sela, O. Neter, V. Lohr, P. Cihelka, F. Wang, M. Zwilling, J. Phillip Sabou, M. Ulman, Signals of propaganda—Detecting and estimating political influences in information spread in social networks, PLOS ONE 20.1 (2025). doi:10.1371/journal.pone.0309688.

[5] M. Hasanain, M. A. Hasan, M. B. Kmainasi, E. Sartori, A. E. Shahroor, G. D. S. Martino, F. Alam, Reasoning about persuasion: Can LLMs enable explainable propaganda detection? (2025). URL: https://doi.org/10.48550/arXiv.2502.16550.

[6] G. Smyth, Nazi 'black' Propaganda to Britain: Secret Radio Stations and British Renegades, Hist. J. Film, Radio Telev. (2023) 1–21. doi:10.1080/01439685.2023.2296215.

[7] A. M. U. D. Khanday, M. A. Wani, S. T. Rabani, Q. R. Khan, A. A. Abd El-Latif, HAPI: An efficient Hybrid Feature Engineering-based Approach for Propaganda Identification in social media, PLOS ONE 19.7 (2024). doi:10.1371/journal.pone.0302583.

[8] A. Horák, R. Sabol, O. Herman, V. Baisa, Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis, Expert Syst. With Appl. (2024) 124085. doi:10.1016/j.eswa.2024.124085.

[9] G. Faye, B. Icard, M. Casanova, J. Chanson, F. Maine, F. Bancilhon, P. Égré, Exposing propaganda: An analysis of stylistic cues comparing human annotations and machine classification. (2024). URL: https://doi.org/10.48550/arXiv.2402.03780.

[10] G. Ramberdiyeva, A. Dildabekova, Z. Abikenova, L. Karabayeva, A. Zhuasbaeva, The Functional and Semantic Category of Appeal as a Linguistic Tool in Political Propaganda Texts (in the Example of the English Language), Int. J. Semiot. Law - Rev. Int. Semiot. Jurid. (2024). doi:10.1007/s11196-024-10115-5.

[11] P. N. Ahmad, L. Yuanchao, K. Aurangzeb, M. S. Anwar, Q. M. u. Haq, Semantic web-based propaganda text detection from social media using meta-learning, Serv. Oriented Comput. Appl. (2024). doi:10.1007/s11761-024-00422-x.

[12] F. Aldabbas, S. Ashraf, R. Sifa, L. Flek, MultiProp Framework: Ensemble models for enhanced cross-lingual propaganda detection in social media and news using data augmentation, text segmentation, and meta-learning, in: Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script, 2025. 7–22. URL: https://aclanthology.org/2025.abjadnlp-1.2.

[13] M. Abdullah, O. Altiti, R. Obiedat, Detecting Propaganda Techniques in English News Articles using Pre-trained Transformers, in: 2022 13th International Conference on Information and Communication Systems (ICICS), IEEE, 2022. doi:10.1109/icics55353.2022.9811117.

[14] M. Abdullah, D. Abujaber, A. Al-Qarqaz, R. Abbott, M. Hadzikadic, Combating propaganda texts using transfer learning, IAES Int. J. Artif. Intell. (IJ-AI) 12.2 (2023) 956. doi:10.11591/ijai.v12.i2.pp956-965.

[15] W. Li, S. Li, C. Liu, L. Lu, Z. Shi, S. Wen, Span identification and technique classification of propaganda in news articles, Complex & Intell. Syst. (2021). doi:10.1007/s40747-021-00393-y.

[16] R. Mahmood, I. A. Shah, T. Hassan, H. Abdullah, T. M. Mubassir, Detecting propagandistic poster title: A machine learning approach, Doctoral dissertation, BRAC University, 2024. URL: https://dspace.bracu.ac.bd/xmlui/handle/10361/24152.

[17] P. N. Ahmad, J. Guo, N. M. AboElenein, Q. M. u. Haq, S. Ahmad, A. D. Algarni, A. A. Ateya, Hierarchical graph-based integration network for propaganda detection in textual news articles on social media, Sci. Rep. 15.1 (2025). doi:10.1038/s41598-024-74126-9.
[18] S. Rose, N. Dethlefs, C. Kambhampati, One-Vs-Rest Neural Network English Grapheme Segmentation: A Linguistic Perspective, in: Proceedings of the 28th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Stroudsburg, PA, USA, 2024, pp. 464–469. doi:10.18653/v1/2024.conll-1.36.
[19] J. Wang, J. Yin, L. Tian, Evaluating joint confidence region of hypervolume under ROC manifold and generalized Youden index, Stat. Med. (2023). doi:10.1002/sim.9998.
[20] I. Krak, V. Didur, M. Molchanova, O. Mazurets, O. Sobko, O. Zalutska and O. Barmak, Method for political propaganda detection in internet content using recurrent neural network models ensemble, in: CEUR Workshop Proceedings, 3806 (2024) 312-324. URL: https://ceur-ws.org/Vol-3806/S_36_Krak.pdf.
[21] UNLP Workshop Data, 2025. URL: https://github.com/unlp-workshop/unlp-2025-shared-task/blob/main/data/techniques-en.md.
[22] Y. Krak, O. Barmak, O. Mazurets, The Practice Investigation of the Information Technology Efficiency for Automated Definition of Terms in the Semantic Content of Educational Materials. CEUR Workshop Proceedings, 1631 (2016) 237-245. doi:10.15407/pp2016.02-03.237.
[23] Hugging Face, BERT Base Multilingual Cased, 2025. URL: https://huggingface.co/google-bert/bert-base-multilingual-cased.
[24] Hugging Face, YouScan Ukr-RoBERTa Base, 2025. URL: https://huggingface.co/youscan/ukr-roberta-base.
[25] I. Krak, M. Molchanova, V. Didur, O. Sobko, O. Mazurets and O. Barmak, Method of semantic features estimation for political propaganda techniques detection using transformer neural networks, in: CEUR Workshop Proceedings, 3917 (2025) 286-297. URL: https://ceur-ws.org/Vol-3917/paper56.pdf
[26] P. Vickers, L. Barrault, E. Monti, N. Aletras, We need to talk about classification evaluation metrics in NLP. (2024). URL: https://doi.org/10.48550/arXiv.2401.03831.
[27] PyTorch, PyTorch Framework, 2025. URL: https://pytorch.org/.
[28] PyPI, Transformers Library, 2025. URL: https://pypi.org/project/transformers/.
[29] PyPI, Datasets Library, 2025. URL: https://pypi.org/project/datasets/.
[30] Python GUI, PySide6 Tutorials, 2025. URL: https://www.pythonguis.com/pyside6/.