Method for neural network detecting propaganda techniques by markers with visual analytic

Iurii Krak[1-2], Olha Zalutska[3], Maryna Molchanova[3], Olexander Mazurets[3], Eduard Manziuk[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

The paper is devoted to the creation and approbation of the method for neural network detecting propaganda techniques by markers with visual analytic, which allows converting input data in the form of text for analysis and supervised machine learning models into output data containing numerical estimates of the presence of each propaganda technique and marked-up text with visual analytical presence of detected propaganda markers. Research was conducted that allows us to detect 17 main propaganda techniques. The study compared the 3 most commonly used approaches: A traditional machine learning approach, an approach based on recurrent neural networks, and an approach based on transformer models. The highest results were achieved by the transformer model approach, which uses self-attention mechanisms that allow each element of the sequence to interact directly with all other elements. This ensures efficient capture of long-term dependencies, which is typical for propaganda techniques. This approach allowed us to detect propaganda techniques with an accuracy of 0.96.

Keywords

BERT, RNN, propaganda techniques, detecting propaganda, propaganda markers, visual analytics

1. Introduction

Propaganda disguised as regular news has been spreading for many decades, but the modern digital age additionally creates the conditions for its faster, more massive and effective dissemination [1]. New methods are being developed to generate texts that are increasingly not much different from those created by humans [2], which leads to a rapid increase in the amount of content. Therefore, all of this emphasizes the importance of creating automated methods for detecting propaganda manipulations that will help users receive information more consciously.

The aim of the research is to improve detecting propaganda techniques accuracy by developing the method for detecting propaganda techniques by markers based on the set of machine learning models, separate for each propaganda technique, trained on modified marked data.

The main contributions of the paper can be summarized as follows:

- An approach to training data preparation has been developed that allows training machine learning models for individual propaganda techniques;
- A method for detecting propaganda techniques is proposed, which allows to find the strength of each of the 17 propaganda techniques, as well as to visually interpret the result using the LIME model.
- The effectiveness of using neural network transformer models in comparison with recurrent models and traditional machine learning approach is experimentally demonstrated.

[1] *ICST-2024: Information Control Systems & Technologies, September , 23 – 25, 2024, Odesa, Ukraine*
yuri.krak@gmail.com (I. Krak); zalutska.olha@gmail.com (O. Zalutska); m.o.molchanova@gmail.com (M. Molchanova); exe.chong@gmail.com (O. Mazurets); eduard.em.km@gmail.com (E. Manziuk); alexander.barmak@gmail.com (O. Barmak)
0000-0002-8043-0785 (I. Krak); 0000-0003-1242-3548 (O. Zalutska); 0000-0001-9810-936X (M. Molchanova); 0000-0002-8900-0650 (O. Mazurets); 0000-0002-7310-2126 (E. Manziuk); 0000-0003-0739-9678 (O. Barmak)

In this paper, the second section provides an overview of related work in the field of propaganda detection according to the two components of the study, including an analysis of existing approaches to the problem of propaganda detection and an analysis of machine learning models for propaganda detection. The third section of the paper contains a scheme and steps of the method for neural network detecting propaganda techniques by markers. The fourth section is devoted to the description of the experiment plan of detecting propaganda techniques by markers and preparation of the dataset. The fifth section contains the results of the experiment, their analysis and discussion.

2. Related Works

2.1. Existing solutions of the detecting propaganda problem

The problem of detecting propaganda remains relevant, as new ways of influencing users to spread propaganda messages are still emerging. In the circumstances, there is a need to continuously monitor new ways of creating propaganda content and improve methods of identifying them, which is an important task of ensuring information security and countering disinformation. Therefore, researchers are working to identify new markers and new methods of propaganda, as well as to improve existing approaches to identify it.

The main methods of analyzing newspaper texts to identify manipulative technologies are investigated, which helps to warn against disinformation and propaganda [1]. A new set of reference data in the Czech language is presented for training and evaluating current and future methods for recognizing 18 manipulative techniques, such as fear-mongering, relativization, and labeling. It is shown that the combination of content analysis with the proposed style analysis increases the accuracy of detecting 15 out of 17 evaluated manipulative techniques from 0.05% to 1.46%. The method was tested on the QCRI propaganda database. Further research will focus on adding new stylometric characteristics, improving existing methods, and using data augmentation techniques to deal with label imbalances. It is also planned to move to fine-grained classification at the level of time intervals rather than at the level of the document as a whole.

Another study presents a multilingual propaganda dataset and conducts an experiment to study the markers that human annotators and classification algorithms use to distinguish propaganda articles from non-propaganda articles on a particular topic [3]. It has been shown that exaggeration, reduced descriptiveness, and lack of adequate sources are common in the propaganda press. The VAGO analyzer confirmed that the use of vague markers significantly correlates with these features. Machine learning models were found to be effective in propaganda detection on a particular topic, but need to be improved in terms of explainability and generalization to other topics. Further work will focus on improving the analysis, developing multilingual models, and improving explainability tools. It is also planned to introduce new labels to refine annotations and identify more stylistic features.

The application of the MVPROP model, which uses multidimensional contextual embeddings, improves the accuracy of propaganda detection. Experiments have shown that the model can be transferred to news articles [4]. For testing purposes, was presented TWEETSPIN, a dataset of tweets containing weak annotations of subtle propaganda techniques, and the MVPROP model for their detection. TWEETSPIN includes only tweet IDs, which is in line with Twitter's terms of use, and contains potentially offensive and hostile statements. The main limitation is weak annotations due to the large scale of the data. In the future, it is planned to study the detection of propaganda at the level of individual fragments.

The researchers applied the RoBERTa language model to detect propaganda techniques in news articles [5]. The model was evaluated using the SemEval-2020 Task 11 reference dataset, demonstrating the ability to detect complex propaganda techniques and outperforming the baseline model with an F1-score of 60.2%. While [6] analyzed the possibilities of using large language models (LLMs), in particular OpenAI's GPT-3.5-Turbo model, to detect signs of propaganda in news articles. Using the technology behind ChatGPT, the researchers analyzed texts to determine the presence of various propaganda techniques identified in previous work [7]. A thoroughly fine-tuned query was developed, which was combined with articles from the Russia Today (RT) network and the SemEval-2020 Task 11 dataset to determine the presence of propaganda

techniques. The study showed that LLM technology can provide reasonable inferences about propaganda, although the detection accuracy is only 25.12% on the SemEval-2022 dataset. However, it shows potential as a propaganda detection tool for end users such as media consumers and journalists.

As confirmed in the above works, propaganda is characterized by techniques that are responsible for certain markers that are inherent in the techniques used. This paper will focus on detecting 17 known propaganda techniques described in detail in [8]. These are: «Appeal to fear-prejudice», «Causal Oversimplification», «Doubt», «Exaggeration», «Flag-Waving», «Labeling», «Loaded Language», «Minimisation», «Name Calling», «Repetition», «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches», «Whataboutism».

2.2. Machine learning models for detecting propaganda techniques

Study will use 3 approaches to machine learning models: traditional machine learning approach, recurrent neural network approach, approach based on transformer models.

Traditional machine learning approach encompasses several methods and algorithms designed to solve various tasks of data prediction, classification and clustering, including the detection of propaganda techniques. Linear regression is used to model linear dependencies between input functions (features) and target values, and is one of the simplest methods of regression analysis and is often used to predict numerical values. Support Vector Machine (SVM) searches for the optimal hyperplane that best separates two classes of data points in the feature space. It is often used for classification tasks, especially when the data has a complex structure [9]. Bayesian network learning is based on a Bayesian probabilistic model where each variable is treated as random and Bayesian inference rules are used to build the model. Traditional machine learning approach to detect propaganda techniques also includes logistic regression, k-NN, and clustering algorithms, such as k-means. In [10], the authors present the results of a study of several classification models, including the multinomial naive Bayesian method, SVM, logistic regression, and K-nearest neighbors.

A recurrent neural network approach to propaganda detection is used to analyze sequential data, including texts that are frequently shared on social media. Recurrent Neural Networks, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are different types of recurrent neural network architectures, each of which has its own characteristics and applications in solving various machine learning tasks, including propaganda detection on social media. RNN is a basic architecture that is capable of processing sequential data, storing information in the form of an internal state (memory) that is updated with each new input [11]. LSTM is an extended version of RNN that includes additional mechanisms such as forgetting gates, update gates, and exit gates. GRU is a simplified version of LSTM that has fewer internal components. GRU is considered to be a less computationally expensive architecture compared to LSTM [12], the results of a study of propaganda identification on the Twitter platform during the COVID-19 pandemic by the authors of [13] showed that the proposed LSTM-based propaganda identification performed better than other machine learning methods considered in the paper. The proposed LSTM-based approach achieves an accuracy of 77.15%. And in [14], the authors use the Bi-LSTM and Bi-GRU deep learning techniques with weakly supervised SVM methods, this approach provided 90% accuracy in identifying propaganda news. The authors argue that this approach is a highly useful and effective one for unlabeled data.

The approach based on transformer models involves the use of such neural network architectures as BERT, RoBERTa, DistilBERT, GPT, etc. [15]. BERT is one of the most famous transformer architectures developed by Google. BERT is capable of achieving good results in natural language processing (NLP) tasks due to its ability to contextualize words and its ability to train regular models for many NLP tasks [16]. RoBERTa is an optimized approach to BERT that improves model training and performance on various NLP tasks by applying different optimization strategies [17]. DistilBERT is considered a lightweight version of BERT that preserves the essence of the original model by reducing the number of parameters and maintaining high performance on various NLP tasks. GPT is a family of transformer models developed by OpenAI. This approach is used in a study to classify propaganda [18].

The authors use three deep learning models, CNN, LSTM, Bi-LSTM, and four transformer-based models, namely multilingual BERT, Distil-BERT, Hindi-BERT, and Hindi-TPU-Electra. The experimental results indicate that the multilingual BERT and Hindi-BERT models provide the best performance with the highest F1 score of 84% according to the experimental study. Also, in [19], the authors study the performance of BERT and RoBERTa, DeBERTa with a combination of different data augmentation methods for detecting propaganda texts. The authors were able to achieve F1 micro score of 60% on the test set using an ensemble of BERT, RoBERTa, and DeBERTa models.

Thus, these approaches find their application in the task of identifying propaganda techniques.

3. Projecting of Method for Detecting Propaganda Techniques

To implement the method for detecting propaganda techniques by markers, it is proposed to create 17 machine learning models, each of which will be responsible for a specific propaganda technique. This approach will allow to train machine learning models in such a way that they can build dependencies inherent in specific types of propaganda.

In general, a scheme of the method for detecting propaganda techniques is shown in Figure 1. The method allows converting input data in the form of text for analysis and trained machine learning models into output data containing numerical estimates of the presence of each propaganda technique and marked-up text with visual analysis of the presence of detected propaganda markers.

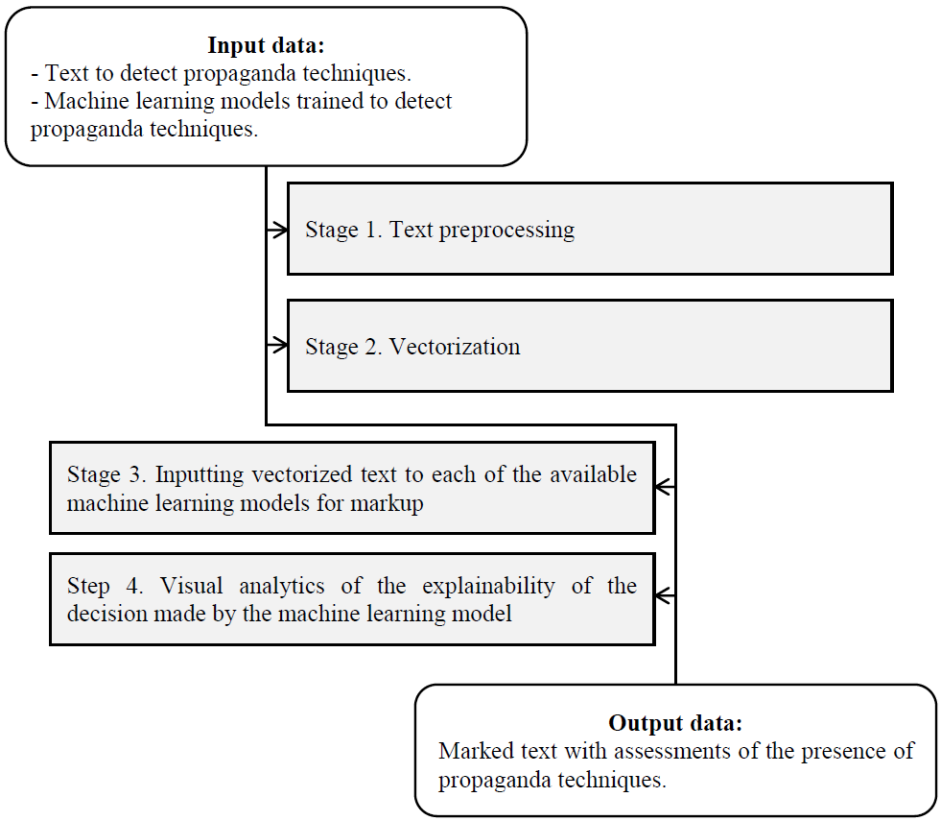

Figure 1: Steps of Method for Neural Network Detecting Propaganda Techniques

The inputs to the propaganda detection method are a text for propaganda detecting and trained machine learning models to detect propaganda.

Text preprocessing includes the removal of punctuation and stop words, although punctuation placed in a certain way can also affect the presence of propaganda [20]. The association of related words was performed by lemmatization, which shows better results than stemming. For lemmatization was used appropriate standard Python library. However, this study will not analyze this effect.

The next step is vectorizing the text after pre-processing. The vectorized representation is given as an input to each trained machine learning model, which predicts the presence of each propaganda technique and its strength (Figure 2).

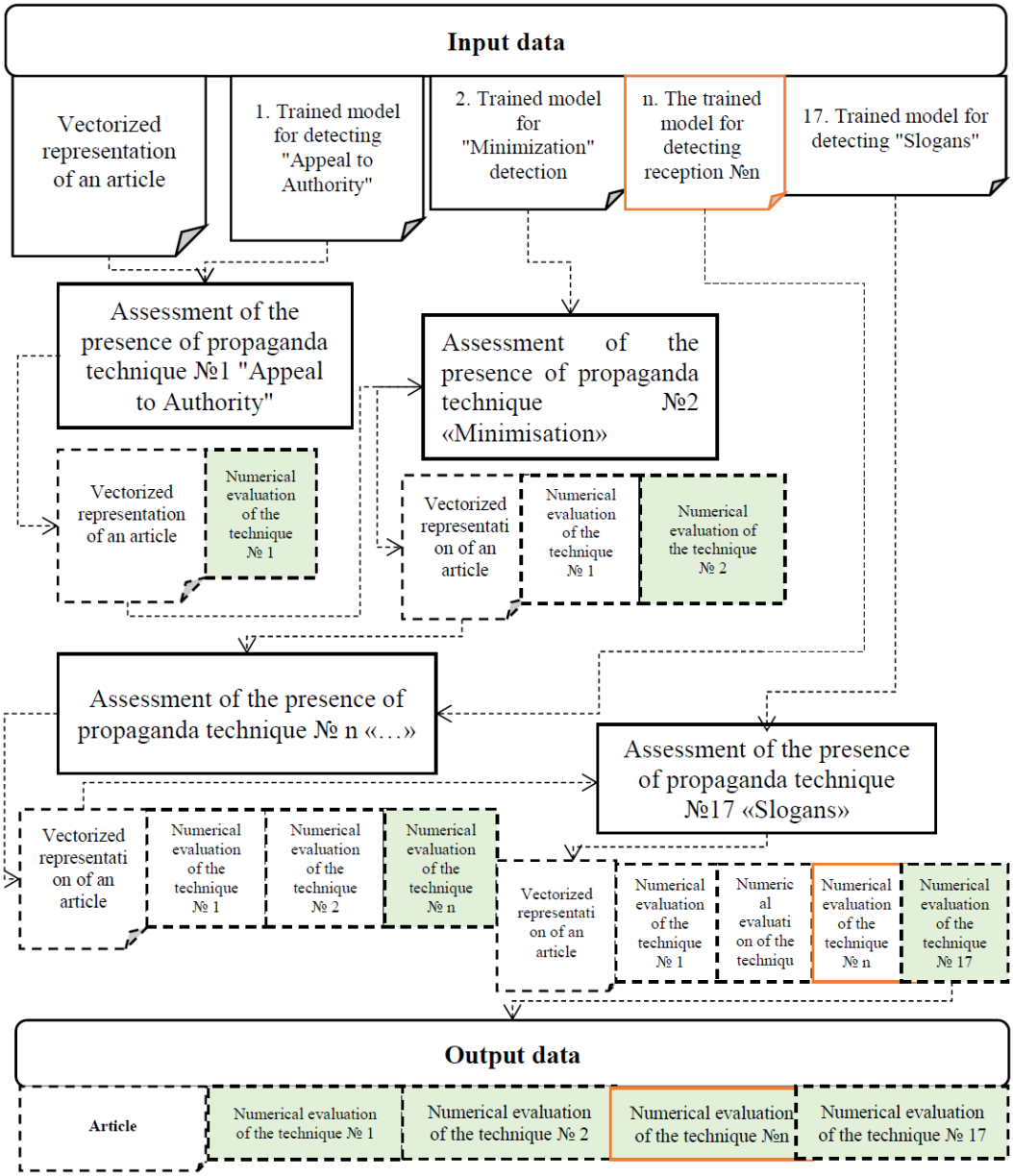

Figure 2: Detailing the step of inputting vectorized text to each of the machine learning models for markup

The last step is to perform visual analytics to explain the decision made by each machine learning model. Visual analytics is used using the LIME method, which is a method for interpreting predictions of machine learning models that is designed to explain individual predictions of complex models [21].

LIME is approximates any black box machine learning model with a local, interpretable model to explain each individual prediction. LIME allows to understand which parts of the input data influenced the model's decision.

The input of the vectorized text representation step to each of the available machine learning models for markup is the vectorized representation of the article and the trained 17 machine learning models.

The models take turns evaluating the vectorized representation of the textual content to analyze for the presence of each of the 17 propaganda techniques. The output data are numerical estimates of the strength of the propaganda techniques inherent in the given vector representation of the text.

Thus, a method for neural network detecting propaganda techniques by markers was created that allows converting input data in the form of text for analysis and trained machine learning models into output data containing numerical estimates of the presence of each propaganda technique and marked-up text with visual analytical presence of detected propaganda markers.

4. Experiment

4.1. Description of the experiment

The experiment studied the use of 3 approaches to training models to identify propaganda techniques: traditional machine learning approach, recurrent neural network approach, approach based on transformer models.

Traditional machine learning approach was used to detect propaganda techniques using regression models, SVM, Random Forest, and Naive Bayes. For the techniques «Appeal to Authority», «Black and White Fallacy», «Reductio ad Hitlerum», «Red Herring», «Slogans», «Thought Terminating Cliches» and «Whataboutism», a study will also be conducted with and without SMOTE balancing.

An approach based on recurrent neural networks includes a comparison of 3 types of architectures: RNN, LSTM and GRU.

An approach based on transformer models includes a comparison of BERT-like models: RoBERTa, BERT, ELECTRA.

To conduct the experiment, the software was created in the Python programming language, using the machine learning libraries Sklearn [22], Tensorflow [23], by LimeTextExplainer [24], Numpy [25], and Pandas [26].

The software consists of a console application for training machine learning models, a console application for detecting propaganda techniques by markers, and a web module for visual analytics of evaluation of accepted results by selected machine learning model with its scores.

4.2. Data set for the experiment

To train machine learning models that will perform the functions of detecting propaganda techniques, the dataset «emnlp_trans_uk_dataset» will be used, which is a translated dataset «emnlp_en_dataset» with the markup in Ukrainian, taken from the Kaggle competition «Disinformation Detection Challenge» [27] with reference to the «Analysis Project».

The Analysis Project team [28] analyzed the texts, detecting all the fragments containing propaganda techniques and their type. In particular, they created a corpus of news articles manually annotated at the level of fragments using eighteen propaganda techniques. The dataset includes 788 articles. For most propaganda techniques, the length of the texts where they are presented does not play a special role. However, «Flag Waving», «Red Herring», «Reductio ad hitlerum» and «Whataboutism» still have a smaller maximum length in the texts where they are presented.

To train machine learning models, this dataset was modified so that the text containing each propaganda technique was placed in a separate catalog. After such a redistribution, the statistics of the available texts representing propaganda techniques were derived. The statistics are shown in Figure 3. Some propaganda techniques such as «Bandwagon», «Confusion», «Intentional Vagueness», «Obfuscation» and «Straw Men» are presented in a critically low number (less than 20 tests), so no separate classifiers will be created for them, this data will be combined into the category «Other propaganda techniques», but in such a way that the existing set does not contain other techniques other than the five listed. For propaganda techniques that are presented in less than 100 documents, but more than 20, SMOTE-balancing will be applied during classifier training [29].

These categories include: «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches» and «Whataboutism».

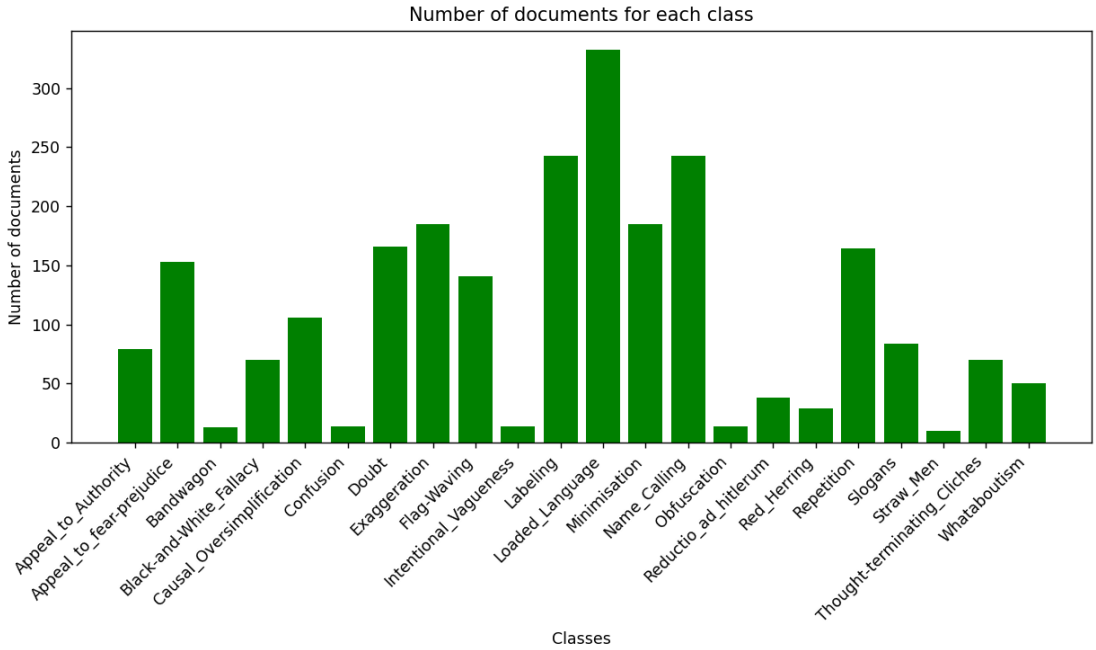

Figure 3: Statistics on the number of texts representing propaganda techniques, pcs

An example of the formation of a data set for detecting the «Appeal to fear-prejudice» technique is shown in Figure 4.

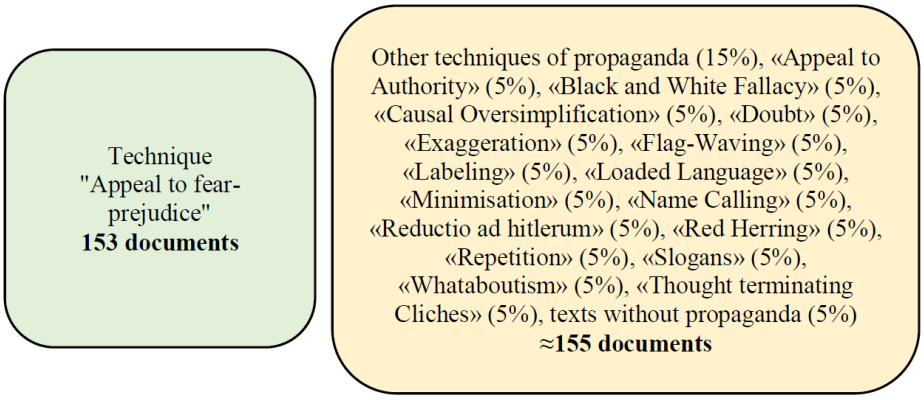

Figure 4: An example of balancing while forming a dataset for training and testing a model for detecting the «Appeal to fear-prejudice» technique

From the above dataset, each of the 17 typical machine learning models will generate its own dependent text set that will meet the following requirements: have texts with a specific propaganda technique; as opposed to using the «Other propaganda techniques» set, supplemented with texts without propaganda and texts representing other propaganda techniques other than the target type.

Thus, the study will use 18 classes: 17 target classes, which are representative in number and correspond to the 17 detected propaganda techniques, and 5 combined into the category «Other propaganda techniques».

5. Results and discussion

The results of the study for traditional machine learning approach for «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches» and «Whataboutism» without using SMOTE balancing by the accuracy metric are shown in Table 1.

Table 1
Traditional machine learning approach for detecting propaganda techniques before SMOTE balancing by the accuracy metric

Techniques of propaganda	Regression	SVM	Random Forest	Naive Bayes
Appeal to Authority	0.57	0.63	0.55	0.64
Black and White Fallacy	0.55	0.64	0.51	0.58
Reductio ad hitlerum	0.68	0.56	0.61	0.59
Red Herring	0.61	0.61	0.59	0.61
Slogans	0.62	0.63	0.56	0.62
Thought terminating Cliches	0.59	0.58	0.63	0.58
Whataboutism	0.62	0.65	0.59	0.57

As can be seen from Table 1, the accuracy of detecting propaganda techniques ranges from 0.51 to 0.68, which is quite low. The next step was to apply SMOTE balancing to these propaganda techniques, thus increasing the number of training samples to at least 100. The result of the experiment is shown in Table 2.

Table 2
Traditional machine learning approach for detecting propaganda techniques after SMOTE balancing by the accuracy metric

Techniques of propaganda	Regression	SVM	Random Forest	Naive Bayes
Appeal to Authority	0.59	0.67	0.54	0.60
Black and White Fallacy	0.61	0.62	0.55	0.64
Reductio ad hitlerum	0.69	0.63	0.62	0.58
Red Herring	0.69	0.64	0.58	0.61
Slogans	0.62	0.63	0.56	0.62
Thought terminating Cliches	0.62	0.68	0.62	0.61
Whataboutism	0.64	0.66	0.58	0.56

As can be seen in Table 2, the application of SMOTE balancing provided positive results for most propaganda techniques, but for «Slogans» there was no improvement. This is due to the fact that the number of training samples is close to the limit and is sufficient to train the proposed machine learning versions.

The next experiment was a study of the use of an approach based on recurrent neural networks, which included a comparison of the use of 3 types of architectures: RNN, LSTM, and GRU. The data of the experiment without using SMOTE balancing are shown in Table 3. As can be seen from Table 3, the results for all propaganda techniques except «Thought terminating Cliches» are higher and range from 0.66 to 0.8.

However, SMOTE balancing will be applied to this propaganda technique in the future, which may improve the score. The next experiment will be to apply SMOTE balancing to the training of neural network models for «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches» and «Whataboutism».

Table 4 shows that SMOTE balancing has a positive effect on the accuracy of detecting propaganda techniques. The detection of «Reductio ad hitlerum» was not improved, where the results before SMOTE balancing were 0.01 higher, and «Appeal to Authority» and «Black and White Fallacy» remained at the same level as before balancing.

The last stage of the study is the use of the approach based on transformer models, which includes a comparison of BERT-like models: RoBERTa, BERT, ELECTRA. Pre-trained models from the Hugging Face resource [30] were used, which were retrained in the above way during 3 epochs of training. The results obtained without using SMOTE balancing are shown in Table 5. So, BERT-like neural network architectures are significantly better at detecting propaganda techniques compared to recurrent and traditional machine learning approach. This is due to the fact that such architectures are context-aware, which is an important aspect for detecting propaganda techniques.

Table 3
An approach based on recurrent neural networks for the detection of propaganda techniques by the accuracy metric

Techniques of propaganda	RNN	LSTM	GRU
Appeal to fear-prejudice	0.69	0.69	0.71
Causal Oversimplification	0.73	0.71	0.76
Doubt	0.75	0.7	0.74
Exaggeration	0.64	0.72	0.75
Flag-Waving	0.69	0.7	0.79
Labeling	0.69	0.73	0.8
Loaded Language	0.71	0.7	0.68
Minimisation	0.78	0.78	0.74
Name Calling	0.76	0.74	0.76
Repetition	0.74	0.75	0.76
Appeal to Authority	0.71	0.72	0.73
Black and White Fallacy	0.7	0.68	0.72
Reductio ad hitlerum	0.75	0.68	0.71
Red Herring	0.65	0.72	0.70
Slogans	0.74	0.68	0.75
Thought terminating Cliches	0.63	0.66	0.65
Whataboutism	0.67	0.69	0.69

Table 4
An approach based on recurrent neural networks for detection propaganda techniques with SMOTE balancing by the accuracy metric

Techniques of propaganda	RNN	LSTM	GRU
Appeal to Authority	0.7	0.72	0.73
Black and White Fallacy	0.72	0.7	0.72
Reductio ad hitlerum	0.73	0.74	0.74
Red Herring	0.69	0.73	0.75
Slogans	0.72	0.76	0.72
Thought terminating Cliches	0.69	0.76	0.78
Whataboutism	0.68	0.7	0.78

Table 5
An approach based on transformer models for the detection of propaganda techniques by the accuracy metric

Techniques of propaganda	bert-base-multilingual-cased	roberta-base	ukr-electra-base
Appeal to fear-prejudice	0.81	0.8	0.87
Causal Oversimplification	0.78	0.79	0.82
Doubt	0.93	0.9	0.87
Exaggeration	0.8	0.8	0.8
Flag-Waving	0.92	0.9	0.89
Labeling	0.96	0.94	0.96
Loaded Language	0.93	0.97	0.94
Minimisation	0.89	0.86	0.9
Name Calling	0.92	0.92	0.91
Repetition	0.93	0.94	0.94
Appeal to Authority	0.87	0.89	0.88
Black and White Fallacy	0.89	0.91	0.88
Reductio ad hitlerum	0.85	0.87	0.86
Red Herring	0.67	0.8	0.78
Slogans	0.84	0.86	0.83

Thought terminating Cliches	0.8	0.73	0.79
Whataboutism	0.79	0.78	0.78

The application of SMOTE balancing allowed to increase the accuracy of detecting «Red Herring» by the ukr-electra-base neural network model to 0.89, and «Whataboutism» to 0.83 using bert-base-multilingual-cased. A comparison of the highest scores on the accuracy metric for the 3 approaches under consideration is shown in Figure 5.

As can be seen in Figure 5, traditional machine learning approach expectedly performed worse, as it is not able to see the context, which is important for detecting propaganda techniques. Recurrent neural network models, although they performed better than traditional machine learning approach, still have problems with processing long dependencies. The highest results from the experiment were found in the approach based on transformer models, which is explained by the self-attention mechanisms used, which allow each element of the sequence to directly interact with all other elements. This allows for the effective capture of long-term dependencies, which is typical of propaganda techniques.

The obtained results ensured the detection of various propaganda techniques with a minimum accuracy of 79.03% (the minimum accuracy values were obtained for the "Whataboutism" technique), which is better than known analogues [8] for detecting propaganda regardless of techniques used.

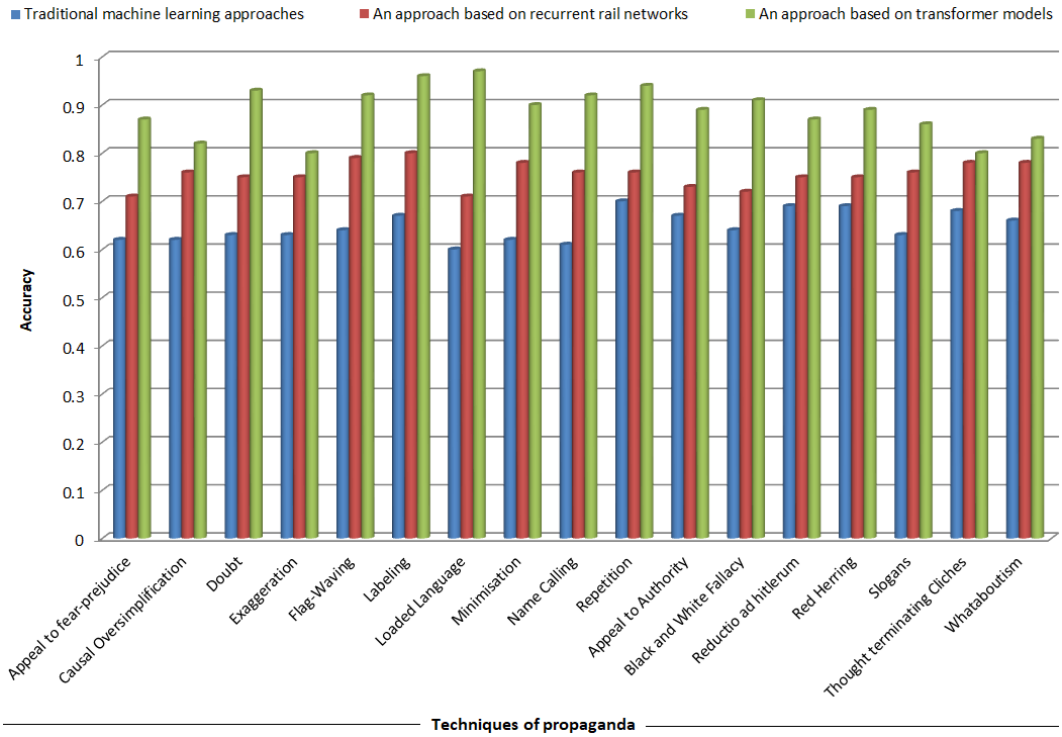

Figure 5: Comparison of the accuracy of models of alternative approaches for detecting propaganda techniques

An example of a visual explanation for identifying the «Repetition» propaganda technique is shown in Figure 6 (the original text in everyday Ukrainian language is used with preservation of spelling and errors).

As can be seen from Figure 6, there are multiple repetitions of phrases such as «economic migrants» (ukrainian: «економічні мігранти»), «Muslim» (ukrainian: «мусульманський»), «Orban» (ukrainian: «Орбан»), etc. According to the definition of the propaganda type «Repetition», it is «repeating the same message over and over again so that the audience eventually accepts it». Thus, the proposed method allows for effective detection of propaganda techniques and has an advantage in accuracy compared to the proposed models that use a multi-class classification approach.

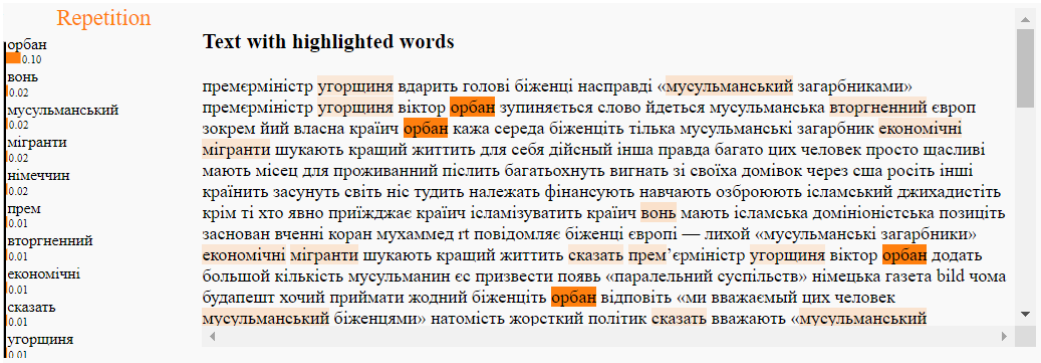

Figure 6: Visual analytics on the detection of the «Repetition» propaganda technique by the developed software

The experiments presented in the paper were carried out using various capabilities of the SKLearn library. This paper presents the maximum results that were achieved by authors empirically. The issue of configuration and selection of hyperparameters is a separate problem that goes beyond the scope of the issues under consideration.

6. Conclusions

Research was conducted that allows us to detect 17 main propaganda techniques, such as: «Appeal to fear-prejudice», «Causal Oversimplification», «Doubt», «Exaggeration», «Flag-Waving», «Labeling», «Loaded Language», «Minimisation», «Name Calling», «Repetition», «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches», «Whataboutism».

The study compared the 3 most commonly used approaches: A traditional machine learning approach, an approach based on recurrent neural networks, and an approach based on transformer models. Traditional machine learning approach expectedly showed worse results, as they are not able to take into account the context, which is important for detecting propaganda techniques. The achieved accuracy for the traditional approach ranged from 0.60 to 0.67. Recurrent neural networks, while outperforming traditional approaches, still have difficulty processing long dependencies. For this approach, the accuracy ranged from 0.66 to 0.80. The highest results were achieved by the transformer model approach, which uses self-attention mechanisms that allow each element of the sequence to interact directly with all other elements. This ensures efficient capture of long-term dependencies, which is typical for propaganda techniques. This approach allowed us to detect propaganda techniques with an accuracy of 0.96.

The obtained results ensured the detection of various propaganda techniques with a minimum accuracy of 79.03% (the minimum accuracy values were obtained for the "Whataboutism" technique), which is better than known analogues [8] for detecting propaganda regardless of techniques used. Compared to known analogues [7], the accuracy of detection of various propaganda techniques has improved: detection accuracy increased minimum by 9.81% (for the "Appeal to Authority" technique), maximum by 62.31% (for the "Reductio ad hitlerum" technique).

Further research will be aimed at expanding the dataset for training and searching for additional labels in texts that characterize propaganda techniques, such as the presence of bullying, emotional tone, etc., which will make the decision of the machine learning model more explanatory and allow for more accurate detection of techniques.

7. References

[1] A. Horak, R. Sabol, O. Herman, V. Baisa, Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis, Expert Systems with Applications 251 (2022). doi:10.1016/j.eswa.2024.124085.

[2] A. Bhattacharjee, H. Liu, Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text? SIGKDD Explor, Newsl 25 (2023) 14–21. doi:10.1145/3655103.3655106.

[3] G. Faye, B. Icard, M. Casanova, J. Chanson, F. Maine, F. Bancilhon, G. Gadek, G. Gravier, P. Egre, Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification, in: Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language, Malta, 2024, pp. 62–72.
[4] P. Vijayaraghavan, S. Vosoughi, TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics, 2022, pp. 3433–3448.
[5] 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), Irbid, Jordan, 2022, pp. 301-308. doi: 10.1109/ICICS55353.2022.9811117.
[6] D.G. Jones, Detecting Propaganda in News Articles Using Large Language Models, Eng OA, 2 1 (2024) 1-12. URL: https://www.opastpublishers.com/peer-review/detecting-propaganda-in-news-articles-using-large-language-models-6952.html.
[7] G. D. S. Martino, A. Barron-Cedeno, H. Wachsmuth, R. Petrov, P.Nakov, SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1377–1414.
[8] G. Martino, S. Yu, A. Barron-Cedeno, R. Petrov, P. Nakov, 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), 2019, pp. 5640-5650. doi:10.18653/v1/D19-1565.
[9] Analytics Vidhya, Guide on Support Vector Machine (SVM) Algorithm, 2024. URL: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/.
[10] S. Mann, D. Yadav, D.Rathee, Identification of Racial Propaganda in Tweets Using Sentimental Analysis Models: A Comparative Study, in: Swaroop, A., Kansal, V., Fortino, G., Hassanien, A.E. (eds) Proceedings of Fourth Doctoral Symposium on Computational Intelligence. DoSCI 2023. Lecture Notes in Networks and Systems, vol 726, Springer, Singapore, 2023. doi: 0.1007/978-981-99-3716-5_28.
[11] Krak I., Zalutska O., Molchanova M., Mazurets O., Bahrii R., Sobko O., Barmak O. Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network. CEUR Workshop Proceedings, vol. 3387, 2023, pp. 16-28.
[12] Geeks for geeks, What is LSTM – Long Short Term Memory, 2024. URL: https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/.
[13] A.M.U.D. Khanday, Q.R. Khan, S.T. Rabani, M.A. Wani, M. ELAffendi, Propaganda Identification on Twitter Platform During COVID-19 Pandemic Using LSTM, in: Abd El-Latif, A.A., Maleh, Y., Mazurczyk, W., ELAffendi, M., I. Alkanhal, M. (eds) Advances in Cybersecurity, Cybercrimes, and Smart Emerging Technologies, CCSET 2022, Engineering Cyber-Physical Systems and Critical Infrastructures, Springer, Cham, vol 4, 2022. doi:10.1007/978-3-031-21101-0_24.
[14] Dev, Large Language Models: Comparing Gen 1 Models (GPT, BERT, T5 and More), 2024. URL: https://dev.to/admantium/large-language-models-comparing-gen-1-models-gpt-bert-t5-and-more-74h.
[15] Hugging Face, BERT, 2024. URL: https://huggingface.co/docs/transformers/model_doc/bert.
[16] Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings, vol. 3387, 2023, pp. 344-356. doi:10.15407/jai2024.02.085.
[17] Hugging Face, DistilBERT, 2024. URL: https://huggingface.co/distilbert/distilbert-base-uncased.

[18] D. Chaudhari, A. V. Pawar, Empowering Propaganda Detection in Resource-Restraint Languages: A Transformer-Based Framework for Classifying Hindi News Articles, Big Data and Cognitive Computing, 7 4 (2023) 175. doi:10.3390/bdcc7040175.
[19] A. Malak, D. Abujaber, A. Al-Qarqaz, R. Abbott, M. Hadzikadic, Combating propaganda texts using transfer learning, IAES International Journal of Artificial Intelligence (IJ-AI) 12 (2023) 956-965. doi: 10.11591/ijai.v12.i2.pp956-965.
[20] I. 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, vol.1631, 2016, pp. 237–245. doi:10.15407/pp2016.02-03.237.
[21] C3.ai, What is Local Interpretable Model-Agnostic Explanations (LIME)?, 2024. URL: https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/.
[22] Scikit-learn, Machine Learning in Python, 2024. URL: https://scikit-learn.org/stable/.
[23] Tensorflow, An end-to-end platform for machine learning. Machine Learning in Python, 2024. URL: https://www.tensorflow.org.
[24] GitHub, Lime, 2024. URL: https://github.com/marcotcr/lime.
[25] Numpy, The fundamental package for scientific computing with Python, 2024. URL: https://numpy.org.
[26] Pandas, Pandas, 2024. URL: https://pandas.pydata.org.
[27] Kaggle, Disinformation Detection Challenge, 2024. URL: https://www.kaggle.com/competitions/disinformation-detection-challenge/data?select=emnlp_trans_uk_dataset.
[28] Propaganda, Propaganda Analysis Project, 2024. URL: https://propaganda.qcri.org/index.html.
[29] Y. Elor, H. Averbuch-Elor, To SMOTE, or not to SMOTE. URL: https://arxiv.org/abs/2201.08528.
[30] Hugging Face, The AI community building the future, 2024. URL: https://huggingface.co/.