UDC 519.688; 004.89; 004.9; 78.071.4: 78.078(477)+316.74
DOI: 10.20535/SRIT.2308-8893.2025.2.05

DETERMINING THE LEVEL OF PROPAGANDA IN OPERA LIBRETTOS USING DATA MINING AND MACHINE LEARNING

I. DATS, O. GAVRILENKO, K. FESHCHENKO

Abstract. The article presents an adapted multifactorial model that can be used to determine the level of propaganda in librettos to world operas. This model was created using the linear convolution method, for which eight indicators were selected that are most effective in identifying elements of propaganda in the text, taking into account the subject area's peculiarities. Each of the selected indicators was calculated using statistical analysis, data mining, and machine learning methods. As a result of applying the proposed method, the value function is calculated for each libretto, based on which a conclusion is made as to whether it contains elements of propaganda or not.

Keywords: art, propaganda, opera, libretto, multivariate model, statistical analysis, Data Mining, Machine Learning, information technology.

INTRODUCTION

Propaganda in art is the use of artistic forms to influence public opinion, shape ideas, and spread specific ideologies or political views. It can be both explicit and subtle, serving as an instrument of the state, religion, or social movements.

When studying the factors that influence human opinions in various areas of activity, it is worth paying attention to the vast and diverse realm of "agitation" in art. Since classical times, this has included visual and monumental art; during the Renaissance, masterpieces carried propaganda of a new era for humanity. Later, theatrical art acquired a dual meaning, while musical compositions and cinema, with their strong emotional impact, took on a special role in global propaganda.

The development of propaganda in art is based on:

- The promotion of an individual or a collective's creative activity (promotional advertising), which helped advance the careers of "useful" figures in the creative field.
- The involvement of specialists in the propaganda of artistic products, where musical content and literary foundations contributed to patriotic songwriting (particularly from the perspective of socialist state leaders).

Quite often, musical works used for propaganda incorporated compositions by other composers or folk songs, embedding entirely new meanings into them. For example:

The anthem of the USSR (at least its musical material) was taken from Mykola Lysenko's "Epic Fragment", whose impact and emotional depth made it highly suitable for Soviet state propaganda.

The agitational song "Far Beyond the River", which fully adopted a Ukrainian insurgent song about a fallen hero, was repurposed by the Red Army to promote the fight against what they considered old and bourgeois elements.

These are just a few examples of musical works that, in addition to raising issues of plagiarism in music, also highlight the problem of identifying propaganda.

Due to the vast diversity of art forms, this article focuses on the propagandistic impact on opera audiences, considering opera as a genre with a long history, an elite form of art, and a significant part of world culture.

Propaganda in opera has been particularly evident in productions staged in China [1], Nazi Germany, and the Soviet Union. For example, the works of Richard Wagner, which glorify ancient Germanic legends, were used to emphasize the superiority of the German nation and the Aryan race, reinforcing the ideology of world domination.

Similarly, the soviet regime implemented propaganda slogans by repurposing older russian operas and creating new, ideologically charged soviet works that praised and glorified the soviet government, its achievements, and way of life. In Ukraine: "The Death of the Squadron" by Yuliy Meitus, "Standard-Bearers" by Oleksandr Bilash. In russia: "In the Storm and Alpine Story" by Tikhon Khrennikov, as well as the film-operetta "Wedding in Malinovka*"* and the film-musical "Three Fat Men", among others.

Another intriguing aspect of musical propaganda is its presence in modern advertising. Commercials often feature simple, easily memorable melodies consisting of just a few notes, making them instantly recognizable and associated with the promoted product or message. In instrumental, vocal, and stage music, propaganda can be embedded in an emphasized form, calling for specific conclusions or even radical actions.

Overall, propaganda in opera has significant historical importance, particularly in societies where culture was used as a tool of ideological influence. As a synthesis of music, drama, and visual art, opera has a strong emotional impact, making it an effective medium for conveying political and ideological messages. Given the large volume of textual data in opera librettos and arias, identifying propagandistic elements requires advanced technologies.

Therefore, addressing this issue necessitates the integration of artistic expertise, including the work of playwrights, directors, actors, composers, and poets, along with information technologies such as mathematical modeling, Data Mining, statistical analysis, and Machine Learning techniques. This combination will enable systematic detection of propaganda in opera librettos, providing new insights into how ideological messages are embedded in classical and modern operatic works.

ANALYSIS OF LITERARY SOURCES AND PROBLEM STATEMENT

Research on propaganda detection demonstrates a variety of approaches and conclusions in this field. Scholars are increasingly leveraging modern techniques, particularly machine learning models such as *BERT* and *GPT*–4, to analyze and detect propaganda in textual data streams. These models can identify and classify different propaganda techniques across various texts.

Study [2] used a pre-trained *BERT* model to improve the detection of propaganda in news articles. The model processed text at the word level and integrated sentence-level features, effectively distinguishing between propagandistic and non-propagandistic content. However, issues such as data imbalance were identified, leading researchers to employ methods like oversampling and data augmentation to address them.

Study [3] focused on annotating and detecting propaganda using *GPT–*4. The research involved a multi-stage annotation process to ensure high-quality data, compiling a dataset of annotated paragraphs from diverse news sources to analyze propaganda techniques across different topics.

Study [4] examined the impact of propaganda on the political landscape in the U.S., revealing that disinformation in mass media significantly influenced social discourse and policymaking. This study proposed further research through ontology construction based on interdisciplinary methods from computer science and social sciences.

Study [5] conducted detailed text analysis, identifying 18 propaganda techniques in manually annotated news articles. The research also introduced a new BERT-based neural network to enhance propaganda detection.

Study [6] presented a credibility assessment methodology for questionable information, using semantic similarity metrics on knowledge graphs to calculate the shortest paths between conceptual nodes.

Study [7] explored the history and evolution of information warfare methodologies, comparing American, British, and Russian models while introducing the concept of "semantic warfare" in the modern world.

A crucial limitation of current machine learning models is their reliance on supervised learning, meaning they require human-labeled training datasets. This introduces an element of subjectivity, as the classification of certain texts as propaganda depends on human judgment.

Additionally, social media plays a significant role in propaganda dissemination today [8]. For example:

Study [9] introduced the *CatRevenge* model, designed to identify active and passive revenge communication in social media, which aligns with propaganda detection. The model used *Slangzy* (an internet slang dictionary) for preprocessing, assigning *TF–IDF*-based weights to words and employing a *CATBoost* classifier to reduce overfitting.

Study [10] investigated influential individuals in knowledge-sharing processes within internal social networks, predicting future knowledge flow patterns and analyzing propaganda's ideological impact through a four-phase methodology combining social network analysis and structural modeling.

Study [11] analyzed how social media posts by influential figures affected cryptocurrency markets, highlighting an example of propaganda in commerce.

Study [12] deals with the problem of detecting propaganda in text files. The authors consider methods for solving the problem of classifying textual information for spam filtering, contextual advertising, news categorization, and creating thematic catalogs.

Study [13] presents a multifactorial model for determining the level of propaganda in a publication. The publications used were text news and social media posts. The model was created based on the linear convolution method. This model considered 10 indicators, a high level of each of which indicates the presence of propaganda in the publication. This model is based only on statistical data and calculations made using Data Mining, statistical analysis and Decision Theory algorithms.

Study [14] provides an overview of multilingual models for working with limited data sets and analyzes their development. The following models are considered: *XLM–RoBERTa*, *mBERT*, *LASER*, *MUSE*.

These studies emphasize the importance of using sophisticated Machine Learning, Statistical Analysis, Data Mining, and careful data annotation processes to detect and analyze propaganda. They provide valuable insights into methodologies that can improve the accuracy and reliability of propaganda detection systems, which is crucial for understanding and mitigating the impact of propaganda.

It should be noted that the process of propaganda detection continues to require the development of various mathematical models to better identify this form of communication.

In addition, it should be emphasized that none of the proposed models has been used to identify propaganda in the musical and theatrical arts in general and opera in particular.

The authors of this study propose a modified version of the Multi-Factor Propaganda Detection Model (MMDP) from Study [13], adapted specifically for evaluating propaganda levels in opera librettos. Additionally, the study examines how propaganda detection results from MMDP correlate with the assessments of opera experts. An information technology was developed to conduct experimental research. By integrating Machine Learning, Statistical Analysis, and Data Mining with artistic expertise, this study aims to fill the existing gap in identifying propaganda in opera as an elite and historically significant art form.

OBJECTIVE AND TASKS OF THE RESEARCH

The objective of this research is to adapt the MMDP [13] for processing and analyzing the libretto of world operas to identify signs of propaganda within them. To achieve this objective, the following tasks have been set:

- Compile a dataset of libretto from well-known world operas that differs from the dataset presented in study [13].
- Select from the 10 propaganda indicators outlined in study [13] those that are most relevant to the chosen artistic domain.
- Improve methods for determining indicators that are characteristic of propaganda detection in publications.
- Utilize the MMDP to calculate the level of propaganda content within the compiled dataset.
- Draw conclusions regarding the presence of propaganda indicators in the libretto texts.

MATERIALS AND METHODS OF RESEARCH

The object of the study is the process of identifying propaganda in opera libretto (hereafter referred to as publications) based on an analysis of information about them. Specifically, the study considers the following factors:

- Primary source of the publication (in this context, the literary work that served as the basis for the opera libretto).
- Brief description of the primary source.
- Word count in the publication.
- Sentence count in the publication.
- Syllable count in the publication.

- Total number of opera productions currently available on streaming platforms.
- Number of productions of operas based on libretto contained in the dataset.
- Number of reviews of opera performances based on the libretto in the dataset.
- Number of re-posts of the publication (in this context, the number of video recordings of the opera based on the given libretto on a streaming platform).
- Number of likes under the video recording of the opera based on the given libretto on a streaming platform.
- Number of comments under the video recording of the opera based on the given libretto on a streaming platform.
- Sources of re-posts (in this context, channels that share opera video recordings based on the given libretto).

The set of publications and all necessary information for this research was obtained from [15].

The successful completion of the study requires both basic statistical data and data obtained using Data Mining and Machine Learning techniques

Fig. 1 illustrates the main steps involved in determining the level of propaganda in publications.

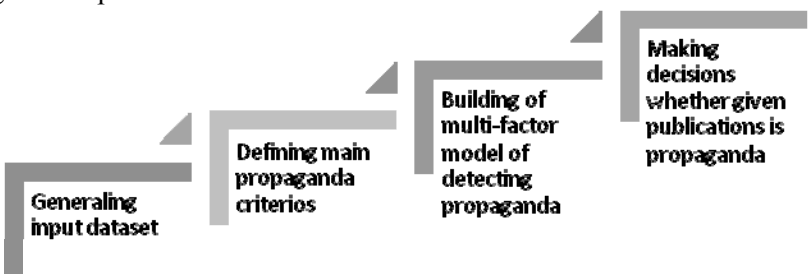

Fig. 1. Main Steps for Determining the Level of Propaganda in Publications

The proposed propaganda detection principle is based on calculating a metric that reflects the degree of correspondence between a given publication and pre-selected propaganda indicators. This is achieved using the convolution method.

To compute the values of the indicators, the study employs statistical analysis methods, as well as Data Mining and Machine Learning techniques. Additionally, specialized software was developed for conducting intelligent analysis and obtaining results based on these methods.

ADAPTATION OF A MULTIFACTOR MODEL FOR CALCULATING THE LEVEL OF PROPAGANDA IN OPERA LIBRETTO

The process of constructing a multifactor model for calculating the level of propaganda in publications, based on the convolution method, can be outlined in the following stages [16; 17].

Stage 0: Preprocessing of the publication text.

Stage 1: Calculation of numerical indicators for the model.

Stage 2: Calculation of importance coefficients for each indicator.

Stage 3: Calculation of the value function.

Stage 4: Formulation of conclusions regarding whether the given publication is propagandistic.

These stages are illustrated in Fig. 2.

Fig. 2. The process of building a multifactor model

Step 0

Input: a set of publications. $P=(P_1,\ldots,P_l)$.

Output: sets of words $A_1,A_2,\ldots,A_l$ used in the publications $P_1,\ldots,P_l$ respectively.

To form each set, it is recommended to preprocess the text of the publications using lemmatization and stemming processes. This helps reduce the size of the set by eliminating root-related words and auxiliary parts of speech.

PHASE 1

Input: a set of publications $P=(P_1,\ldots,P_l)$ a set of propaganda features $X=(x_1,\ldots,x_8)$ [13]:

x_1 = {attempts to manipulate the audience};

x_2 = {the publication is aimed at evoking emotions};

x_3 = {frequent repetition of a specific idea in the publication};

x_4 = {frequent reposting of the publication};

x_5 = {simplicity of the publication's text};

x_6 = {a high level of propaganda in the original source};

x_7 = {belonging to a specific topic that is particularly susceptible to propaganda};

x_8 = {the publication has an impact on the viewer}.

It is necessary to calculate the levels of propaganda for each of the given features.

At the output, a set is formed $K^j=(K_1^j,\ldots,K_8^j)$, where K_i^j, $i=1,\ldots,8$; $j=1,\ldots,l$ — the values of the metrics that indicate the level of propaganda in publication according to feature x_i.

1. **Calculation of the metric** K_1^j. A numerical assessment of manipulative attempts in texts can be based on methods of computational linguistics, sentiment analysis, lexical analysis, and machine learning.

Emotional tone analysis. Manipulative texts often contain emotionally charged words (e.g., fear, threats, exaltation). Emotional dictionaries $(SentiWordNet, LIWC, NRC, VADER)$ are widely used to determine the emotional tone of a text.

For example, if a publication contains a negative tone (fear, anger), manipulation is possible. If a publication contains excessive positivity, propaganda is possible. If the negative emotion index is higher, and the aggregated sentiment score is too low, deliberate escalation is possible.

Detection of logical fallacies and manipulative techniques. Manipulators use certain rhetorical techniques:

- Appeal to Fear (e.g., the phrase "Either you are mine, or death!" from G. Puccini's opera Tosca).
- False Dilemma (e.g., the phrase "Who does not fall at my feet will perish!" from G. Verdi's opera Nabucco).
- Ad Hominem (e.g., the phrase "God, who has placed a ray of His divinity within us, created man to rule!" from G. Verdi's opera Don Carlos).

Lexical patterns and Machine Learning methods are used to detect emotional fallacies. The model is trained on datasets containing labeled manipulative phrases. If a text contains excessively negative predictions, it may be an attempt at manipulation.

Lexical Analysis: Frequency of Manipulative Constructions. Manipulative texts often contain:

- Generalizations ("Everyone knows this!" from G. Verdi's opera Rigoletto; "No sinner will escape God's judgment!" from G. Verdi's opera Don Carlos).
- Evaluative Judgments ("There has never been a more ruthless tyrant!" from G. Verdi's opera The Sicilian Vespers).
- Appeals to Authority ("The law is the law!" from G. Puccini's opera Tosca).

If a text contains many generalizations and emotionally charged evaluative judgments, it may be manipulative.

Text Style Analysis (Stylometry). Manipulative texts may contain a high number of exclamations, many interrogative sentences (rhetorical questions), as well as excessively long or very short sentences.

Thus, score K_1^j for a publication P_j, $j = 1,\ldots,l$ is calculated as follows:

$$K_1^j = \alpha_1 S^j + \beta_1 L^j + \gamma_1 F^j + \delta_1 C^j, \qquad (1)$$

where S^j — sentiment of the text, determined using a word dictionary with specific polarity (positive, negative, neutral) ($S^j = 1$ for a positive or negative tone, $S^j = 0$ for neutral text); L^j — relative frequency of manipulative clichés (lexical features) compared to their total variety; F^j — relative frequency of logical fallacies (fallacies detection) compared to their total variety; C^j – relative frequency of identified stylistic characteristics (stylometry) compared to their total variety; $\alpha_1, \beta_1, \gamma_1, \delta_1$ — weight coefficients. In this study $\alpha_1 = 0{,}4$; $\beta_1 = 0{,}3$; $\gamma_1 = 0{,}2$; $\delta_1 = 0{,}1$. The values of the weight coefficients, as well as those in the subsequent models, were chosen according to the specifics of the subject area and agreed upon with an expert — M.I. Hamkalo, director of a musical-dramatic theater and associate professor at the Tchaikovsky National Music Academy of Ukraine.

It is evident that $0 \le K_1^j \le 1$, and the closer its value is to one, the more manipulative features the given publication contains. Thus, based on x_1 criterion, it can be considered propagandistic.

2. **Calculation of the Metric** K_2^j. The emotional orientation of a text indicates the extent to which it evokes specific emotions (fear, joy, anger, etc.). It can be assessed using the following approaches:

1. Sentiment Analysis.

2. Emotion Detection.

3. Lexical Analysis of Emotional Intensity.

4. Deep Learning (*NLP* -models) (models: $BERT, GPT, LSTM$).

In this study, sentiment analysis was used to evaluate the emotional orientation of the text.

Thus, the metric K_2^j for a publication P_j, $j = 1,\ldots,l$ is calculated as the overall emotional score:

$$K_2^j = \alpha_2 S^j + \beta_2 E^j + \gamma_2 C^j, \quad (2)$$

where S^j — sentiment of the text (this parameter was described earlier); E^j — proportion of emotional words in the text; C^j — atext style analysis (this parameter was also described earlier); $\alpha_2, \beta_2, \gamma_2$ — weight coefficients. In this study $\alpha_2 = 0{,}5;\ \beta_2 = 0{,}3;\ \gamma_2 = 0{,}2$.

It is evident that $0 \le K_2^j \le 1$ and the closer its value is to one, the higher the level of emotional intensity in the given publication. Thus, based on this criterion, it can be considered propagandistic.

3. **Calculation of the Metric** K_3^j. If an idea is expressed using different words, vector models $(Word2Vec, BERT)$ can be used to find similar expressions.

In this study, a vector model $Word2Vec$ was used, with cosine similarity as the similarity measure. Thus, the metric K_3^j for a publication P_j, $j = 1,\ldots,l$ is calculated as follows:

$$K_3^j = \cos(\theta) = \frac{(B^j D^j)}{\| B^j \| \| D^j \|}, \quad (3)$$

where B^j and D^j — are vectors representing objects (word vectors extracted from the publication j); $(B^j D^j)$ — is the dot product of the vectors; $\| B^j \|$, $\| D^j \|$ — are the magnitudes (norms) of the vectors; $\cos(\theta)$ — represents the cosine of the angle between the vectors.

It is evident that $0 \le K_3^j \le 1$ and the closer its value is to one, the more frequently a particular idea is repeated in the given publication. Thus, based on this criterion, it can be considered propagandistic.

4. **Calculation of the Metric** K_4^j**.** The frequency of reposting a publication refers to the number of video recordings of opera performances based on the analyzed libretto found on streaming platforms (Netflix, YouTube, etc.).

Thus, the metric K_4^j for a publication P_j, $j = 1,\ldots,l$ is calculated as the relative frequency of the opera performance's P_j on a streaming platform using the following formula [18; 19]:

$$K_4^j = \frac{n_j}{n}, \quad (4)$$

where n_j — the number of video recordings of the opera based on the given libretto j; n — the total number of operas found on the platform.

It is evident that $0 \le K_4^j \le 1$ and the closer its value is to **one**, the more frequently the given publication is reposted. Thus, based on this criterion, it can be considered propagandistic.

It should be noted that the accuracy of this metric K_4^j depends on the choice of the streaming platform. The more popular the platform, the larger audience it covers within the study. On the other hand, major platforms require processing a large volume of statistical data, which may introduce additional complexities in calculating this metric.

For example, on the OperaVision website [20], 264 video recordings of opera performances were found. G. Verdi's opera Aida was represented in 8 videos. Thus, for the libretto of this opera, $K_4^j \approx 0{,}03$. On other platforms, this metric may have a different value due to variations in statistical data.

5. **Calculation of the Metric** K_5^j. This metric indicates the readability of the given publication's text.

The metric K_5^j for a publication P_j, $j = 1,\ldots,l$ is calculated as follows:

$$K_5^j = \left(206{,}835 - 1{,}015\frac{a_j}{b_j} - 84{,}6\frac{c_j}{a_j}\right)0{,}01, \qquad (5)$$

де a_j — total number of words; b_j — total number of sentences; c_j — total number of syllables.

This metric K_5^j is known as the Flesch Reading Ease Index [21].

The interpretation of this metric's values is shown in Table 1.

T a b l e 1. Interpretation of Flesch Reading Ease Index Values

Score	School level	Notes
0,9 – 1,0	Grade 5	Very easy to read. Easily understood by an average 11-year-old student
0,9 – 0,8	Grade 6	Easy to read. Conversational language for consumers
0,8 – 0,7	Grade 7	Fairly easy to read
0,7 – 0,6	Grades 8-9	Standard language. Easily understood by 13–15-year-old students
0,6 – 0,5	Grades 10-12	Fairly difficult to read
0,5 – 0,3	College	Difficult to read
0,3 – 0,1	Technical Graduate	Very difficult to read. Best understood by university graduates
0,1 – 0,0	Professional	Extremely difficult to read. Best understood by university graduates

It is evident that $0 \le K_5^j \le 1$ and the closer its value is to one, the easier the given publication is to read. Thus, based on this criterion, it can be recommended as propagandistic.

6. **Calculation of the Metric** K_6^j. The primary source refers to the literary work that served as the basis for the libretto (publication).

The metric K_6^j for a publication P_j, $j = 1,\ldots,l$ is calculated as follows:

Step 1. Identify the primary source.

Step 2. Find a brief description of this work.

Step 3. Use a model *Word2Vec* to determine the key words from the text description.

Step 4. Calculate the cosine similarity (equation 3) between the key word vector and a predefined reference vector.

$$Q = (fight; danger; trait; enemy; alliance; tragedy;$$
$$destruction; patriot; glory; unite; fame;$$
$$recretion; hero; unbeatable). \quad (6)$$

In equation (6), the vector Q used in this study was constructed based on a set of words characteristic of propaganda detection. It was reviewed and approved by an expert — M.I. Hamkalo, director of a musical-dramatic theater and associate professor at the Tchaikovsky National Music Academy of Ukraine. This vector can be adjusted or modified depending on the specific subject area of analysis.

It is evident that $0 \le K_6^j \le 1$, and the closer its value is to one, the higher the likelihood that the given publication has a propagandistic nature. Thus, based on criterion x_6, it can be considered propagandistic.

As an example, we can consider the opera "The Golden Ring" by Ukrainian composer Borys Lyatoshynsky, based on the libretto by Yakiv Mamontiv, which was inspired by Ivan Franko's novel "Zakhar Berkut". It is well known that the novel contains a call to struggle against external and internal enemies. This leitmotif was transferred into the libretto and, consequently, into the opera.

Thus, according to criterion x_6, the opera "The Golden Ring" exhibits propaganda elements.

7. **Calculation of the Metric** K_7^j. Consider a publication P_j, $j = 1,\ldots,l$; the set of words used in the publication A_j; the set of topics $S = (s_1; s_2; \ldots; s_r)$, in which propagandistic publications are most frequently found, and the dictionaries of characteristic words for these topics $T_1; T_2; \ldots; T_r$. The topics and their corresponding dictionaries should be predefined. Some of these topics include:

- Politics: "power", "tyranny", "monarchy", "autocracy", "rebellion", "discord", "revolutionary movement", "coup", "betrayal", "intrigue", "enemies", "opponents",...
- Military Conflicts: "army", "legion", "foreign rule", "tyranny of conquerors", "conquest",...
- Ideology: "people", "nation", "society", "unity", "solidarity", "cohesion", "alliance", "threat", "danger", "monarchy",...
- Conspiracies and Disinformation: "the real truth", "triumphant truth", "secret conspiracy", "treacherous plan", "spies", "accomplices",...

The metric K_7^j indicates whether the publication P_j belongs to one of the topics in the set S. It is calculated as follows:

Step 1. Compute the Jaccard similarity coefficients between the set of words A_j and each topic dictionary T_k, $k = 1,2,\ldots,r$ [22]:

$$J(A_j, T_k) = \frac{\left|A_j \cap T_k\right|}{\left|A_j \cup T_k\right|}; \quad (7)$$

Step 2. Select the maximum Jaccard coefficient:

$$J_{max}(A_j, T_k) = \max_{k=1,2,\ldots,r} J(A_j, T_k), \quad j = 1,\ldots,l$$

and establish which topic T corresponds to this maximum value $s_k \in S$.

Step 3. The metric K_7^j is defined as:

$$K_7^j = J_{\max}(A_j, T_k). \quad (8)$$

It is evident that $0 \le K_7^j \le 1$ and the closer its value is to one, the more closely the publication aligns with topics that are most susceptible to propaganda. Thus, based on this criterion x_7 it can be considered propagandistic.

As an example, we can again consider the opera "The Golden Ring". The libretto of this opera can be categorized under the "Ideology" topic, which is frequently influenced by propaganda. Therefore, for this libretto (publication), the value of the metric K_7^j is quite close to 1.

8. **Calculation of the Metric** K_8^j. To assess the audience reach and its impact, the overall score is calculated as follows:

$$K_8^j = \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3, \quad (9)$$

where X_1 — relative number of likes to the total number of opera views; X_2 — proportion of opera views relative to the most popular opera in the dataset; X_3 — relative number of comments to the total number of opera views; $\lambda_1, \lambda_2, \lambda_3$ — weight coefficients.

It is evident that $0 \le K_8^j \le 1$ and the closer its value is to one, the greater the level of influence the given publication has on the audience. Thus, based on criterion x_8, it can be considered propagandistic.

It should be noted that the accuracy of the metric K_8^j similar to K_4^j depends on the choice of the streaming platform.

PHASE 2

Input: indicators K_i^j, $i = 1,\ldots,8$; $j = 1,\ldots,l$, calculated using formulas (1)–(9).

It is necessary to calculate importance coefficients for each criterion to determine the value function.

Output: coefficients ω_i.

To compute these coefficients ω_i, the following steps must be performed:

Step 1: Form statistical samples from the indicators K_i^j with corresponding names.

Step 2: Select a threshold value, exceeding which a publication can be considered propagandistic.

In this study, by analogy with Chaddock's scale [18; 19], which defines the strength of correlation between two random variables, the following scaling was proposed:

$0{,}0 - 0{,}1$ — no propaganda;

$0{,}1 - 0{,}3$ — low level of propaganda;

$0{,}3 - 0{,}5$ — noticeable level of propaganda;

$0{,}5 - 0{,}7$ — moderate level of propaganda;

$0{,}7 - 0{,}9$ — high level of propaganda;

$0{,}9 - 1{,}0$ — very high level of propaganda.

In this study, all levels of propaganda starting from the noticeable level were considered. Thus, the threshold value was set at $\overline{K_i} = 0{,}3$.

This threshold was introduced to facilitate further statistical calculations and ensure the convenient comparison of results with expert opinions.

It should be noted that no universally defined percentage threshold exists in scientific sources that explicitly determines when a text is considered propagandistic [23]. This study emphasizes the importance of qualitative analysis and the recognition of specific influence techniques rather than establishing a universal quantitative threshold.

In future research, a more personalized approach is planned for each propaganda characteristic.

Step 3: If $K_i^j \geq \overline{K_i}$, the given publication P_j is considered propagandistic based on the feature x_i. Otherwise, it is classified as non-propagandistic. Each publication is assigned the value «1», if it is propaganda based on this feature, and «0» otherwise.

$$P_j \to \widetilde{K}_i^j = \begin{cases} 1, & \text{if } K_i^j \geq \overline{K_i}; \\ 0, & \text{if } K_i^j < \overline{K_i}. \end{cases}$$

The transition from quantitative values K_i^j to boolean functions $\widetilde{K}_i^j$ was made to facilitate the comparison of results with expert opinions.

Step 4: Calculate the Relative Frequency of Propagandistic Publications for Each Feature x_i.

$$w_i = \frac{m_i}{n},$$

where m_i — the number of propagandistic publications based on feature x_i; n — the total number of publications in the dataset.

Step 5: Normalize the Relative Frequencies w_i:

$$\omega_i = \frac{w_i}{w_1 + w_2 + \ldots + w_8}.$$

PHASE 3

Input: a set of publications $P = (P_1, \ldots, P_l)$, indicators K_i^j, $i = 1, \ldots, 8$; $j = 1, \ldots, l$ and coefficients ω_i.

It is necessary to calculate the value function for each publication to determine the presence of propaganda features.

Output: the value function result V_j .

The value function V_j , s computed using the linear aggregation method as follows [16; 17]:

$$V_j = \sum_{i=1}^{8} (\omega_i K_i^j). \quad (10)$$

Based on the values of V_j a **statistical sample** of value function results is formed according to equation (10):

$$V = (V_1, V_{2,}, \ldots, V_l).$$

PHASE 4

Input: a set of publications $P = (P_1, \ldots, P_l)$ and a statistical sample $V = (V_1, V_{2,}, \ldots, V_l)$ (see Step 3).

Output: conclusions regarding which publications $P = (P_1, \ldots, P_l)$ єare propagandistic.

Recommendations are made according to the following rule [24]:

- If $V_j \geq \overline{V}$, ($j = 1, \ldots, l$) , then the publication P_j is recommended as propagandistic.
- If $V_j < \overline{V}$, ($j = 1, \ldots, l$) , then the publication P_j is not recommended as propagandistic.

In this rule $\overline{V} = 0{,}3$ — is the threshold value for the sample V (analogous to Step 2).

$$P_j \to \widetilde{K}_i^j = \begin{cases} 1, & \text{if } K_i^j \geq \overline{K_i}; \\ 0, & \text{if } K_i^j < \overline{K_i}. \end{cases}$$

Thus, a publication is assigned «1» , if it is considered propaganda and «0» otherwise.

The correctness of the provided conclusions is evaluated using the *Recall* та *Precision* metrics:

$$Precision = \frac{tp}{tp + fp}, \quad Recall = \frac{tp}{tp + fn},$$

where tp — the number of correctly identified propagandistic publications (true positives); fp — the number of incorrectly identified propagandistic publications (false positives); fn — the number of incorrectly identified non-propagandistic publications (false negatives).

OBTAINED RESULTS

As part of this study, a dataset was compiled, containing the librettos of 10 operas (Table 2).

For these operas, the value function was calculated, based on which conclusions were drawn regarding the presence of propaganda elements in their librettos.

Table 2. Compiled Dataset

Libretto	Opera Title	Composer
P_1	The Huguenots	Giacomo Meyerbeer
P_2	The Mastersingers of Nuremberg	Richard Wagner
P_3	Fidelio	Ludwig van Beethoven
P_4	The Troubadour	Giuseppe Verdi
P_5	A Life for the Tsarc	Mikhail Glinka
P_6	La Traviata	Giuseppe Verdi
P_7	Carmen	Georges Bizet
P_8	Madame Butterfly	Giacomo Puccini
P_9	Turandot	Giacomo Puccini
P_{10}	The Marriage of Figaro	Wolfgang Amadeus Mozart

The obtained results are presented in Table 3.

Table 3. Obtained Results

Libretto	K_1^j	K_2^j	K_3^j	K_4^j	K_5^j	K_6^j	K_7^j	K_8^j	V_j
P_1	1	1	1	0	1	0	1	1	1
P_2	1	1	1	0	1	0	1	1	1
P_3	1	1	1	0	1	0	1	1	1
P_4	1	1	1	1	1	1	1	1	1
P_5	1	1	1	1	1	1	1	1	1
P_6	0	1	0	1	0	0	0	0	0
P_7	0	1	0	1	0	0	0	0	0
P_8	0	1	0	1	0	0	0	0	0
P_9	0	1	0	1	0	0	0	0	0
P_{10}	0	1	0	1	1	0	0	0	0

In Table 3, each publication P_j, $j=1,\ldots,10$ is assigned a value «1», if it is considered propaganda based on feature x_i, $i=1,\ldots,8$ and value «0» otherwise.

DISCUSSION OF RESEARCH RESULTS

The obtained results were compared with the expert opinion of M.I. Hamkalo, associate professor in the field of musical directing at the Tchaikovsky National Music Academy of Ukraine. The comparison is presented in Table 4.

Thus, from Table 4, it is evident that the proposed MMDP identified the presence of propaganda elements in the same opera librettos as the expert. Accordingly, the values of the $Precision=1$, $Recall=1$ metrics, confirm the high accuracy of the MMDP.

Table 4. Comparison of MMDP Results with Expert Opinion

Libretto	V_j	Expert Opinion	Expert's Argumentation
P_1	1	1	Propaganda: Anti-Catholicism
P_2	1	1	Propaganda: German nationalism
P_3	1	1	Propaganda: Liberalism and the struggle for freedom
P_4	1	1	Propaganda: Revolutionary spirit and fight for independence
P_5	1	1	Propaganda: Russian imperial narrative
P_6	0	0	No propaganda: Pure melodrama about personal emotions, without political or social context
P_7	0	0	No propaganda: The opera has no ideological connotations, only depicting emotions and the fatality of destiny
P_8	0	0	No propaganda: A personal tragedy and cultural misunderstandings, without a political message
P_9	0	0	No propaganda: A mythical story not tied to specific political events
P_{10}	0	0	No propaganda: Despite criticism of the feudal system, it is more about romantic twists than politics

CONCLUSIONS

Propaganda in opera is a powerful tool for influencing society, utilizing the impact of music, librettos, and stage performances to shape specific ideological narratives. Throughout different historical periods, opera has served as an instrument of state propaganda, expressing political, social, and nationalist ideas.

In the XIX century, during the era of Romanticism, opera was often used to elevate national spirit and support struggles for independence (for example, "Nabucco" by Giuseppe Verdi became a symbol of the Italian liberation movement). In the XX century, totalitarian regimes actively employed opera to reinforce state ideology: Soviet socialist realism, Nazi Germany, and Maoist China promoted productions that glorified the party, leaders, or the "ideal citizen".

Despite this, opera also served as a means of protest and counter-propaganda. It became a tool for criticizing authority or social structures, often using allegorical plots or hidden messages.

Thus, opera not only reflects historical context but also actively shapes public consciousness, making it a significant instrument of both official and oppositional propaganda.

This study presents an adapted multifactor model, which allows for the assessment of propaganda levels in the librettos of world opera masterpieces. This model is based on the linear aggregation method, for the implementation of which eight indicators were selected. These indicators are the most effective in detecting propaganda elements in a text, taking into account the specific features of the subject area. Each of the selected indicators was calculated using statistical analysis, Data Mining methods, and Machine Learning techniques. As a result of the proposed method, a value function is computed for each publication, based on which a conclusion is drawn regarding whether it contains propaganda elements or not.

Advantages of the Proposed Model:

1. Elimination of Human (Subjective) Influence — the model's calculations rely solely on statistical data or data obtained through Data Mining and Machine Learning methods, ensuring objectivity in detecting propaganda indicators.

2. Scalability — the model can be easily expanded by adding new indicators or removing outdated ones, making it adaptable to evolving research needs.

3. Result Accuracy — the correctness of the obtained results is guaranteed by the use of classical Data Mining and Machine Learning methods.

Disadvantages of the Proposed Model:

1. Large Data Requirements — the model requires the collection and storage of vast amounts of statistical and textual data, which may pose challenges in data management.

2. Continuous Accuracy Monitoring — the reliability of conclusions must be regularly evaluated. In this study, an expert in the subject area was consulted. In other domains, the accuracy of the MMDP model should be validated using multiple propaganda detection methods.

The obtained results can be used as an effective tool in information warfare, both in Ukraine and globally, serving as a powerful element of intent analysis. Additionally, they can assist directors and actors in musical-dramatic theaters, including opera houses and operetta theaters.

Focusing specifically on the concept of artistic propaganda, the proposed methodology can be applied to all forms of art that are in some way related to textual data, such as songs, films, theater, literature, and poetry. For these domains, the methodology would differ only in terms of input statistical data, such as song lyrics, brief descriptions of literary works, or play scripts. It would also vary in the values of weight coefficients in formulas (1), (2), and (9), as well as in the adaptation of propaganda features presented in [13], where some characteristics may be added or removed depending on the specific artistic field.

REFERENCES

1. Zongrui Zhang, "Model Opera" of the 20th Century in Chinese Musical Culture," *Art History Notes*, no. 43, pp. 206–210, 2023. doi: https://doi.org/10.32461/2226-2180.43.2023.286862
2. 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.*, vol. 8, pp. 3603–3612, 2022. doi: https://doi.org/10.1007/s40747-021-00393-y
3. Maram Hasanain, Fatema Ahmed, Firoj Alam, "Can GPT-4 Identify Propaganda? Annotation and Detection of Propaganda Spans in News Articles," *Computation and Language (cs.CL)*, 2024. doi: https://doi.org/10.48550/arXiv.2402.17478
4. K. Hamilton, "Towards an Ontology for Propaganda Detection in News Articles," in *R. Verborgh et al. The Semantic Web: ESWC 2021 Satellite Events. ESWC 2021. Lecture Notes in Computer Science*, vol. 12739. Springer, Cham, 2021. doi: https://doi.org/10.1007/978-3-030-80418-3_35
5. G. Da San Martino, S. Yu, A. Barrón-Cedeño, 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)*, Association for Computational Linguistics, pp. 5635–5645, 2019. doi: https://doi.org/10.18653/ v1/D19-1565
6. G.L. Ciampaglia, P. Shiralkar, L.M. Rocha, J. Bollen, F. Menczer, A. Flammini, "Computational fact checking from knowledge networks," *PLoS One*, 10(6), 15, 2015. doi: https://doi.org/10.1371/journal.pone.0128193
7. G. Pocheptsov, *Modern information wars*. Kyiv: Kyiv-Mogylianska Academy, 2015, 497 p.
8. S. Ghosal, A. Jain, "CatRevenge: towards effective revenge text detection in online social media with paragraph embedding and CATBoost," *Multimed Tools Appl.*, 83, pp. 89607–89633, 2024. doi: https://doi.org/10.1007/s11042-024-18791-y
9. R. Alhajj, J. Rokne (Eds), *Encyclopedia of Social Network Analysis and Mining*. Springer, New York, NY., 2018, 2200 p. doi: https://doi.org/10.1007/978-1-4614-7163-9
10. Ramona-Diana Leon, Raúl Rodríguez-Rodríguez, Pedro Gómez-Gasquet, Josefa Mula, "Social network analysis: A tool for evaluating and predicting future knowledge flows from an insurance organization," *Technological Forecasting and Social Change*, vol. 114, pp. 103–118, 2017. doi: https://doi.org/10.1016/j.techfore.2016.07.032

11. Sergii Telenyk, Grzegorz Nowakowski, Olena Gavrilenko, Mykhailo Miahkyi, Olena Khalus, "Analysis of the influence of posts of famous people in social networks on the cryptocurrency course," *Bulletin of the Polish Academy of Sciences Technical Sciences*, vol. 72(4), 2024. doi: https://doi.org/10.24425/bpasts.2024.150117
12. O. Gavrilenko, Y. Oliinyk, H. Khanko, "Analysis of Propaganda Elements Detecting Algorithms in Text Data," in *Z. Hu, S. Petoukhov, I. Dychka, M. He, (Eds) Advances in Computer Science for Engineering and Education II. ICCSEEA 2019. Advances in Intelligent Systems and Computing*, vol. 938, Springer, Cham, 2020, pp. 438–447. doi: https://doi.org/10.1007/978-3-030-16621-2_41
13. O. Gavrilenko, K. Feshchenko, "Detecting propaganda in news flows," *Adaptive systems of automatic control*, no. 1 (46), pp. 160–177, 2025. doi: https://doi.org/10.20535/1560-8956.46.2025.323759
14. V. Oliinyk, I. Matviichuk, "Low-resource text classification using cross-lingual models for bullying detection in the Ukrainian language," *Adaptive systems of automatic control*, no. 1 (42), pp. 87–100, 2023. doi: https://doi.org/10.20535/1560-8956.42.2023.279093
15. Opera librettos and arias by foreign authors notes. Accessed on: Feb. 23, 2025. [Online]. Available: https://musicinukrainian.wordpress.com/biblio/import_opera/
16. Jürgen Branke, Kalyanmoy Deb, Kaisa Miettinen, Roman Słowiński, *Multiobjective Optimization*. Springer-Verlag Berlin Heidelberg, 2008, 470 p. doi: https://doi.org/10.1007/978-3-540-88908-3
17. Kalyanmoy Deb, *Multi-Objective Optimization using Evolutionary Algorithms*. Wiley, 2001, 536 p.
18. Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, E.Ye. Keying, *Probability and Statistics for Engineers and Scientists*; 9th ed. Pearson, 2016, 816 p.
19. Sheldon Ross, *A First Course in Probability*; 10th ed. Pearson, 2018, 528 p.
20. *Operavision*. Accessed on: Feb. 23, 2025. [Online]. Available: https://operavision.eu/
21. Rudolf Flesch, *How to Write Plain English: A Book for Lawyers and Consumers*. Harper & Row, 1979, 126 p.
22. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, *Mining of Massive Datasets*. Cambridge University Press, 2014, 326 p.
23. Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, Preslav Nakov, "Fine-Grained Analysis of Propaganda in News Articles," *Computation and Language* (*cs.CL*), 2019. doi: https://doi.org/10.48550/arXiv.1910.02517
24. P.G. Preethi, V. Uma, A. Kumar, "Temporal Sentiment Analysis and Causal Rules Extraction from Tweets for Event Prediction," *Procedia Computer Science*, no. 48, pp. 84–89, 2015. doi: https://doi.org/10.1016/j.procs.2015.04.154

Received 01.03.2025

INFORMATION OF THE ARTICLE

Iryna V. Dats, ORCID: 0000-0003-3851-2047, Tchaikovsky National Music Academy of Ukraine, Ukraine, e-mail: irynadats@gmail.com

Olena V. Gavrilenko, ORCID: 0000-0003-0413-6274, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: gelena1980@gmail.com

Kyrylo Yu. Feshchenko, ORCID: 0009-0002-8142-179X, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: fkirill440@gmail.com

ВИЗНАЧЕННЯ РІВНЯ ПРОПАГАНДИ В ОПЕРНИХ ЛІБРЕТО ЗА ДОПОМОГОЮ ЗАСОБІВ DATA MINING ТА MACHINE LEARNING / І.В. Даць, О.В. Гавриленко, К.Ю. Фещенко

Анотація. Подано адаптовану багатофакторну модель, яку можна використати для визначення рівня пропаганди в лібрето до світових опер. Модель створено на основі методу лінійної згортки, для реалізації якого обрано 8 індикаторів, найбільш ефективних для виявлення елементів пропаганди в тексті з урахуванням особливостей предметної галузі. Кожного з обраних індикаторів розраховано з використанням методів статистичного аналізу, Data Mining та машинного навчання. У результаті застосування запропонованого методу для кожного лібрето розраховується значення функції цінності, на основі якого робиться висновок про те, чи містить вона елементи пропаганди, чи ні.

Ключові слова: мистецтво, пропаганда, опера, лібрето, багатофакторна модель, статистичний аналіз, Data Mining, Machine Learning, інформаційна технологія.