UDC 004.942, 519.876.5

Proposing of suggestive influence detection and classification method based on fuzzy logic and feature driven analysis

Yuliia Nakonechna[1]

[1] *National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», 37, Prosp. Beresteyskyi, Kyiv, 03056, Ukraine*

Abstract

This research proposes an approach to the identification and classification of tools used in informational operations aimed at the implementation of suggestive influence, based on existing research on the feature-based informational influence identification. The proposed method combines the theory of fuzzy sets and the methods of fuzzy inference with the approach of analysis based on text features thanks to the author's proposed list of suggestive influence techniques, certain combinations of which are characteristic of various information influence tools. Using this approach, research focuses on identifying and classifying tools such as propaganda, fakes, disinformation, manipulation and artificial narrative. This structure result allows to improve the quality of analysis of similar research cases and to develop optimal countermeasures strategies that will take into account the features of each of the considered information warfare tools in further studies.

Keywords: Suggestive influence, warfare, propaganda, disinformation, manipulation, fuzzy logic, Mamdani.

Introduction

In today's realities, the information domain serves as a significant platform for hybrid warfare, where special information operations employ various tactics, which directly impact the public, security, defense, and public administration by attempting to establish ideological and psychological foundations through the dissemination of propaganda, aiming to create zones of influence. Through the systematic use of information and psychological manipulation, as well as political and economic means, conflicts, social contradictions, and disruptions to social unity can be provoked.

Information warfare uses various influence techniques and technologies including redirecting attention, evoking emotional responses, fabricating problems, controlling information accessibility, presenting one-sided coverage of events, asserting the dominance of an imaginary majority, employing false analogies, engaging in manipulative commenting, and disseminating half-truths.

While the scale of suggestive influence continues to expand, the development and implementation of a coherent state policy to counter the threats of hybrid warfare have only recently begun. Consequently, it is crucial for researchers to focus on developing comprehensive measures to safeguard the information environment, swiftly identify information operations, and deploy effective countermeasures.

Taking into account all of the above, the study aim consists in proposing a method that could take into account the specifics of suggestive influence methods as structural elements of tools of information warfare. The task was defined to combine the feature-based approach with fuzzy logic methods and to consider the possibilities of applying these methods to the process of classification of information influence tools.

1. Problem relevance and research analysis

Detecting instruments of suggestive influence poses a significant challenge due to complex and multi-criteria nature of the task. It is crucial to

develop methods that can distinguish information operations from regular informational activities.

Tools utilized in information operations are constantly evolving and often mimic content of legitimate information sources. This makes detection of suggestive influence a formidable undertaking. Additionally, existing methods often focus on specific types of information operation tools without clearly defining comparative characteristics used for classification. This lack of specificity and justification can hinder transparency of the research process and relevance of the obtained results. It also limits the potential for enhancing the quality of work.

The aim of this paper is to propose a method that takes into account the specifics of the methods of sugestive influence in the tools of information warfare, and to apply the two simplest methods of fuzzy logic to the process of classifying tools of information influence. Such an approach will allow evolve the strategies to counteract these tactics, improving the overall quality of defense against information operations.

2. Existing feature based approaches for suggestive influence instruments detecting

Many techniques have been proposed to identify propaganda, disinformation, fake news and others information distortion techniques, which include data mining and text-mining using ensemble methods [1, 2], linguistic-based detection and social network analysis methods [3], natural language inference approach [4], sentence-level analysis [5] and sentiment analysis techniques [6]. Due to generation of enormous amount of information pieces after full-scale Russian invasion binary classification methods using text-mining and dictionaries also regaining their popularity [7]. Mentioned studies aim to propose methods of fakes, propaganda and disinformation detection and extensively analyze the effectiveness of existing approaches.

Predominantly considered suggestive influence detection techniques utilize a feature-based analysis approach, which consists of addressing the task of suggestive technique detection and classification at defined information features level. Feature sets used in investigations are usually based on existing fake/lying/propaganda detection data sets [2] and include features that were known to be quite effective in lie detection. Features can also be categorized and grouped in accordance with the purpose of conducted analysis, but actual variations in feature essence are minor. Furthermore, most of the reviewed studies did not specify their feature selection method.

E.g. for linguistically inspired propaganda detection process [5] the feature set is following:

- total number of sentences in the article;
- average character-length of article's sentences;
- variance of character length of the article's sentences;
- character-length of current sentence;
- average and variance of character-length of this sentence's words;
- sentence punctuation frequency; sentence capital-case frequency.

Sentiment analysis techniques features involve dictionary-based feature selection in order to assist in classification of test data as positive, negative, and neutral (as context-based customized dictionaries, information polarity detecting features) [6].

The most recent approaches for suggestive influence detection are based on language models and use techniques and markers of specific informational influence instrument, so the analysis output consists of an annotated version of the input text, where the used suggestive influence techniques are detected [8]:

- appeal to authority;
- oversimplification;
- doubt;
- name calling and labeling;
- etc.

So, the result of integrated feature sets approach can be generalized in Table 1 [9]:

Table 1

Extracted Features based on perspectives

Approach	Features	Description
Style-based features	TF-IDF (F1)	Relative frequency of words
	Quantity (F2)	# Characters
		# Words
		# Noun Phrases
		# Sentences
	Complexity (F3)	Average # characters per word
		Average # words per sentence

		Average # punctuations per sentence
	Uncertainty (F4)	# Modal verbs
		# Certainty terms # Generalizing terms # Tentative terms # Numbers and quantifiers # Question marks
	Sentiment (F5)	# Positive words
		# Negative words # Anxiety/angry/sadness words (emotion) # Exclamation marks Content sentiment polarity
	Subjectivity (F6)	# Subjective verbs
	Diversity (F7)	# Unique words
		# Unique nouns, verbs, adjectives, adverbs
	Informality (F8)	# Typos/spellchecks
		# Swear words/ netspeak/assent/fillers
	Additional (F9)	# Hashtags
		# Mentions # Stopwords # URL Mean word length
User engagement features	Popularity (F10)	# Likes
		# Retweets # Replies

Optionally there are also encountered content-based, context-based, stylistic-based approaches etc., but in fact they are represented with similar feature meanings, clustered in other feature groups.

After processing features using selected machine learning mechanisms and using them in substitutions in a suitable model gained result be used to predict the information influence presence or help to classify whether the obtained result is a sample of suggestive influence. But most methods for detecting fake news use post-publication effects on the community to determine whether the news is true or false.

Otherwise speaking, mentioned methods effectiveness can be questionable in the early stages of the information spreading process and can only be used when the data samples which already has spread in the community and potentially left its harmful effects.

3. Proposed feature-based analysis method

Suggestion is a process of influencing human psyche, associated with decrease in consciousness and criticality in perceiving embedded themes, which do not require a detailed personal analysis or motivation evaluation for certain actions, are directed towards individual or public consciousness by informational, psychological or other means, and causes transformation in views, value orientations, stereotypes of the person.

One of the varieties of suggestion are information operations [10]. In hybrid war conditions information operations instruments are the means of suggestive influence. In more detail, suggestive influence and the assessment of strength of informational and suggestive influences are described in [11].

3.1. Suggestive influence as an information manipulation tool

Main suggestive influence instrument categories, which are distinguished among others by their completeness and purpose of use, are propaganda, disinformation, fake, constructed narrative and manipulation of information.

On the other hand, researchers view suggestive influence through the prism of so-called methods of suggestive influence, such as: the method of affirmation, the method of «disinformation», the method of focus on emotions, the method of using stereotypes, the method of repeating information, the method of «myths» and others.

Due to complex nature of instruments of suggestive influence, they use combinations of manipulative methods, which leads to emerging of typical markers by which one or another tool can be recognized [11].

Let's summarize most important markers, relying on investigations [3, 6, 8, 9, 12, 13, 4, 11, 14], and create logically full list of not-intersecting markers:

1. emotionally charged rhetoric;

2. appeal to authority;
3. lack of credible and/or verifiable sources;
4. selective emphasis;
5. unfounded logical leaps;
6. name-calling and other logical fallacies;
7. usage of fear-mongering tactics;
8. repetitive rhetoric and lack of source diversity;
9. presenting inaccurate or misleading information;
10. usage of conspiracy theories as a source;
11. unaddressed internal inconsistency;
12. issue oversimplification;
13. attacking specific social groups;
14. topic polarization;
15. conflation of multiple ideas, terms or concepts;
16. lack of correction even after it has been corrected;
17. bandwagon effect;
18. bias or expert involving to manipulate;
19. informality, poor grammar or spelling;
20. hoaxes or scams.

Listed set of markers takes into account features of each of the above-mentioned tools of information warfare, while some markers are similar to the list suggestion methods given in [11].

3.2. Proposing marker-to-feature characteristics hypothesis

Let us make the following two assumptions to identify the hypothesis that each mentioned marker can be characterized by some set of features:

1. feature-based analysis makes it possible to draw a conclusion regarding the belonging of some piece of information to the tools of information influence;
2. the presence of suggestive influence can be determined by indicating methods described as markers that allow specifying the type of information influence tools.

By establishing meaningful connections between features and markers of suggestive influence, it becomes feasible to differentiate between various types of tools used in suggestive influence. This differentiation can be achieved during the identification or detection process, allowing for a more accurate assessment of the presence of these tools.

Analyzing existing feature-based approaches and relying on the feature selection rationale applicable to specifying information warfare tools, relying on the investigations [3, 6, 8, 9, 12, 13, 4, 11, 14] and stepping on the knowledge base on hybrid warfare tools provided by State Service of Special Communications and Information Protection of Ukraine and Center for Countering Disinformation the author proposed the correspondence of markers of suggestive influence to tools of informational influence.

Thus, the typical suggestive markers for the given tools of information influence will be as following:

1. Propaganda: emotionally charged rhetoric; appeal to authority; lack of credible and/or verifiable sources; selective emphasis; name-calling and other logical fallacies; usage of fear-mongering tactics; repetitive rhetoric and lack of source diversity; presenting inaccurate or misleading information; issue oversimplification; attacking specific social groups; conflation of multiple ideas, terms or concepts; bandwagon effect.
2. Fake: emotionally charged rhetoric; lack of credible and/or verifiable sources; selective emphasis; presenting inaccurate or misleading information; usage of conspiracy theories as a source; issue oversimplification; topic polarization; informality, poor grammar or spelling; hoaxes or scams.
3. Disinformation: emotionally charged rhetoric; lack of credible and/or verifiable sources; presenting inaccurate or misleading information; usage of conspiracy theories as a source; unaddressed internal inconsistency; lack of correction even after it has been corrected; bias or expert involving to manipulate.
4. Manipulation: emotionally charged rhetoric; lack of credible and/or verifiable sources; selective emphasis; presenting inaccurate or misleading information; issue oversimplification; topic polarization; lack of correction even after it has been corrected; bias or expert involving to manipulate.
5. Narrative: emotionally charged rhetoric; appeal to authority; unfounded logical leaps; repetitive rhetoric and lack of source diversity; issue oversimplification; attacking specific social groups; topic polarization.

In accordance with the general specificity of suggestive means, different tools can have both different markers and common ones, while it is not necessary to use all markers specific to the tool at the same time.

3.3. Fuzzy logic driven approaches

Let's consider two of the simplest approaches using fuzzy logic that will allow us to depict and describe the relationships between the sets of features discussed earlier, namely suggestive tools, markers, and textual features.

3.3.1. Cognitive mapping approach

A fuzzy cognitive map is a model of a studied system in the form of a directed graph defined using a set of sets:

$$CM =< C, F, W > \quad (1)$$

where $C = C_i$ is the set of graph vertices i.e., concepts or factors that have the greatest importance in terms of studying the system being considered; $F = F_k$ is the set of directed edges representing the relationships between concepts;; $W = W_{ij}$ is the set of all edge weights (relationships).

It is assumed that the connections between concepts can be positive - «strengthening» the influence of concept C_i on concept $C_j(W_{ij} > 0)$, or negative «weakening» the influence of concept C_i on concept C_j $(W_{ij} < 0)$. In the simplest case, $W_{ij} = +1$ or $W_{ij} = -1$, in which case it is called a signed directed graph.

The values of the weights (strength of connection) W_{ij} can be expressed using a fuzzy linguistic scale, which is an ordered set of linguistic values (terms) representing strength of connection ratings, but in current case we will use only fact of connection between concepts, so defined weights set is {0,1}.

As an signed directed graph cognitive map is fully defined by its adjacency matrix:

$$W = \begin{vmatrix} W_{11} & W_{12} & \dots & W_{1n} \\ W_{21} & W_{22} & \dots & W_{2n} \\ \dots & \dots & \dots & \dots \\ W_{n1} & W_{n2} & \dots & W_{nn} \end{vmatrix} \quad (2)$$

where the elements W_{ij} take values of +1 (positive link), -1 (negative link), or 0 (no link); n - is the number of concepts in the cognitive map [11].

The cognitive mapping methodology has been chosen in this case due to the inability to assess the quantitative influence of manipulation methods on human consciousness. The use of a linguistic scale or even simply defying suggestive influence presence allows us to move from fuzzy information about the state of concepts to the possibility of numerically evaluating the resulting influence of one concept on another.

To construct a weighted cognitive map, we need to first build the adjacency matrix of concepts based on the given graph. Note that in the considered model, there is no influence of a concept on itself, i.e., the weighted graph will not have loops, so we can exclude the part of the adjacency matrix that corresponds to connections of the form $C_i \to C_i$, and set the elements $W_{ij} = 0$ for $i = j$.

Let us consider suggestive influence instruments, markers and features (these terms will be understood within the scope of this study).

Let's define the following sets: suggestive influence instruments set I, markers set M, and features set F.

Suggestive influence instruments set I consists of concepts:

- propaganda I_1
- fake I_2;
- disinformation I_3;
- manipulation I_4
- narrative I_5.

Markers set M contains markers enumerated previously:

- emotionally charged rhetoric M_1;
- ...
- hoaxes or scams M_{20}.

Feature set is built based on Table 1 descriptions and consists of:

- relative frequency of words F_1;
- ...
- replies F_{32}.

Let's construct the adjacency matrix for the sets of tools and markers, as well as for the set of markers and features. The adjacency matrixs are presented in the tables:

Table 2

The adjacency matrix for I and M concept sets

	I_1	I_2	I_3	I_4	I_5
M_1	1	1	1	1	1
M_2	1				1
M_3	1	1	1	1	
M_4	1	1		1	
M_5					1
M_6	1				
M_7	1				
M_8	1				1
M_9	1	1	1	1	

M_{10}		1	1		
M_{11}			1		
M_{12}	1	1		1	1
M_{13}	1				1
M_{14}		1		1	1
M_{15}	1				
M_{16}			1	1	
M_{17}	1				
M_{18}			1	1	
M_{19}		1			
M_{20}		1			

Table 3
The adjacency matrix for M and F concept sets

	M_1	M_2	M_3	M_4	…	M_{20}
F_1	1	1	1	1		
F_2	1					
F_3	1	1	1	1		
F_4	1	1		1		
F_5						
…		…		…		…
F_{30}		1				
F_{31}		1		1		
F_{32}	1	1				

We present a general scheme of the suggestive influence instruments dependencies from test features in the form of a cognitive map, where concepts $I_1 - I_5$ correspond to suggestive influence instruments, concepts M_1 - M_{20} correspond to the markers, and concepts $F_1 - F_{20}$ represents text features.

The obtained fuzzy cognitive map for suggestion methods is shown in Figure 1.

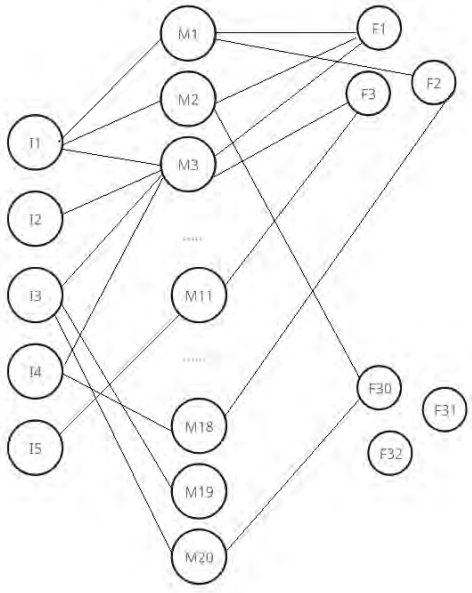

Figure 1: Example figure

The construction of a fuzzy map provides a visual representation of the relationships between text features and the different types of suggestive influence tools, highlighting the patterns in which these features are manifested. By incorporating weights into the generated graph, it becomes possible to classify a specific suggestive influence tool by calculating the weight coefficients of the relevant paths formed by its input connections.

Thus, the final type of suggestive influence can be classified on stage of identifying the presence of suggestion methods in the examined text.

To test the hypothesis, a cognitive mapping approach has been adopted, enabling us to establish a network of connections among features, markers, and types of suggestive influence tools using a cognitive map. After assigning weights to the strength of connections between concepts of a cognitive map using correspondence between linguistic and numeric terms, which brings us to the fuzzy suggestive influence model.

In order to transform the given cognitive map into a fuzzy one, it will be necessary to evaluate the presence of elements of a set of text features in each of the given markers, as well as to develop weight assessments of the identified markers regarding the unambiguity of their belonging to a certain type of information tools (e.g. there are markers like M_5 in connection with I_5 or M_{20} in connection with I_2) which are characteristic only for one type of information tools, so it can help in unambiguously defining one or another tool of informational influence. By the other hand, M_1 or M_9 are typical for most of the listed tools. Same situation we experiencing with connection markers-to-features.

The author suggests using the following approach to assign weights: based on the results of research on the classification of information influence tools and their own, to form a database of correspondence between the presence of text features in certain information influence tools, and then, based on the evaluations assigned according to one of the expert evaluation procedures, to create a correspondence between the presence of a certain text feature and a means of suggestive influence (marker) for which such a feature is characteristic. The described approach will be fully developed by the author and applied in a practical way in future research.

3.3.2. Mamdani model approach

Relying on the set of rules for instrument-marker-feature connection, rule-based fuzzy model can be applied. In these models, the relationships between variables are represented by means of if-then rules with imprecise predicates, like: If the fridge cooling is low then the temperature will lower slow. Qualitive predicate as «high» or «low» is defined by linguistic variable compared to a numerical range. E.g. an usual predicate scale is given in range [0,1] and divided into intervals according to linguistic variables used.

Due to specifics of features analysis use, mentioned before, the result of the analysis presented in the form of numerical coefficients of the intensity of the presentation of one or another characteristic (or features, as we refer to them). This gives us the opportunity to establish an appropriate scale of the intensity of the appearance of this or that feature and to put a linguistic variable in accordance with the intervals. Then, for our instrument-marker-feature case we can create a set of fuzzy rules which in general would be as following [15]:

$$R_i: if\ x\ is\ A_i\ then\ y\ is\ B_i,\ \ i = 1,2,\dots,K, \quad (3)$$

where R is a rule, A and B are linguistic terms (such as «small», «large», etc.), represented by fuzzy sets, and K is the number of rules in the model. E.g. assessing the strength of the connection using linguistic terms is given in the Table 4:

Table 4

Strength of connection values

Linguistic term	Numeric range
does not affect	0
very weak	(0; 0.15]
weak	(0.15; 0.35]
average	(0.35; 0.6]
strong	(0.6; 0.85]
very strong	(0.85; 1]

The linguistic fuzzy model is useful for representing qualitative knowledge such as in the following illustrative example.

Due to adjacency matrix obtained in Ta, let us present concepts I, M, F as variables, then fuzzy rule for suggestive instrument classification via markers and feature would be as follows:

$$R_i: if \left(\begin{array}{c} \left(\begin{array}{c} (F_{i1}\ is\ A_{i1}\ and\ F_{i2}\ is\ A_{i2}\ and\ \dots) \\ then\ M_1\ is\ B_{k1} \end{array} \right) \\ and \left(\begin{array}{c} (F_{i3}\ is\ A_{i3}\ and\ F_{i4}\ is\ A_{i4}\ and\ \dots) \\ then\ M_2\ is\ B_{k2} \end{array} \right) \\ and\ (\dots) \end{array} \right) \\ \dots\ then\ I\ is\ I_i, \quad i = 1,2,\dots,K, \quad (4)$$

where I, M, F are elements of mentioned above instrument, markers and features sets, as well as A, B, I.

In this model, the antecedent (if-part of the rule) and the consequent (then-part of the rule) are fuzzy propositions, so we are getting the Mamdani-similar model. That means we can use fuzzy logic methods to process feature-based analysis results for suggestive influence instruments classification/detection.

Using the fuzzy scores, which were obtained as weighting coefficients for the cognitive map, it is thus possible to create a fuzzy inference system for classifying tools of fuzzy influence in the text processing process and searching for means of suggestive influence in it

Conclusions

The research conducted an in-depth analysis of existing approaches and methods for analyzing and identifying instruments of suggestive influence. It compiled a comprehensive list of typical features commonly utilized in the analysis of such instruments. Furthermore, the study investigated, analyzed, and summarized the characteristic markers associated with major instruments of suggestive influence, including propaganda, fake content, disinformation, manipulation, and artificial narrative.

A hypothesis was proposed, suggesting that a specific set of features capable of detecting suggestive content in information could also characterize the markers inherent in instruments of suggestive influence. To test this hypothesis, a cognitive mapping approach was adopted. This approach facilitated the establishment of connections between features, markers, and types of suggestive influence tools using a cognitive map.

By examining and evaluating the strength of the constructed connections within the map, corresponding to the key concepts explored in this study, the tools of suggestive influence could be classified during the process of identifying suggestive methods within the analyzed text.

The proposed method for identifying and classifying tools of suggestive influence in information operations involves the combination of fuzzy sets theory and fuzzy inference methods with feature-based analysis. This approach enables the processing of results obtained from feature-based text analysis and utilizes fuzzy inference systems to detect and classify instruments of suggestive influence.

We propose a way to identify suggestive influence and classify tools used in information operations by combining fuzzy sets theory and fuzzy inference methods with feature based analysis, allowing to process feature-based text analysis results and use fuzzy inference systems for suggestive influence instrument detection and classification.

References

[1] Pathak A., Rohini K. S., Nihit N. "Disinformation: analysis and identification". Computational and Mathematical Organization Theory 27 (2021): 357–375.

[2] R. Harita, R. Namratha, G. Manali, B. Annappa. "Text-mining-based Fake News Detection Using Ensemble Methods". International Journal of Automation and Computing 17 (2020): 210-221.

[3] Mahyoob M., Algaraady J., Alrahaili M. "Linguistic Based Detection of Fake News in Social Media". International Journal of English Linguistics 11 (2021): 99-109.

[4] Sadeghi F., Jalaly Bidgoly A., Amirkhani H. "Fake News Detection on Social Media using a Natural Language Inference Approach". Multimedia Tools and Applications 81(2022): 33801-33821.

[5] Ferreira-Cruz A., Rocha G., Lopes Cardoso H. "On Sentence Representations for Propaganda Detection: From Handcrafted Features to Word Embeddings". Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda 81(2019): 107-112.

[6] A. Siti-Rohaidah, M. Zakwan, M. Rodzi, N. Syafira Shapiei Nurhafizah Moziyana Mohd Yusop, S. Ismail. "A Review of Feature Selection and Sentiment Analysis Technique in Issues of Propaganda". International Journal of Advanced Computer Science and Applications 11(vol. 10) (2019): 240-245.

[7] V. Solopova, O. Popescu, C. Benzmüller, T. Landgraf. Automated Multilingual Detection of Pro-Kremlin Propaganda in Newspapers and Telegram Posts (2023).

[8] Vorakitphan V., Cabrio E., Villata S. PROTECT: A Pipeline for Propaganda Detection and Classification. Reviews of Modern Physics (2022).

[9] Rastogi S., Bansal D. "Disinformation detection on social media: An integrated approach". Multimedia Tools and Applications 81 (2022): 40675-40707.

[10] L. Kompantseva, E. Skulysh, O. Boyko, V. Ostroukhov. Suggestive technologies of manipulative influence: education help; under the editorship E. D. Skulisha. VIPOL, 2011.

[11] Nakonechna Y., Sviridenko S. "Fuzzy cognitive maps as a means of suggestive risks modeling, analysis and assessment", Materials from all-Ukrainian science conference for students and young scientists "Theoretical and Applied Problems of Math, Physics and Computer Science", 2020.

[12] Garth S. J., O'Donnell V. Propaganda and Persuasion, fifth edition. SAGE, 2012.

[13] Giles K. NATO Defense College Cataloguing in Publication-Data: Handbook of Russian Information Warfare. DeBooks Italia, 2016.

[14] CCD. Center for Countering Disinformation: Glossary (2023)

[15] Robert Babuška. "Identification Using Fuzzy Models". Control systems, robotics, and automation VI (2004).

[16] Nakonechna Y. "Feature based analysis for suggestive influence detection". Materials from all-Ukrainian science conference for students and young scientists "Theoretical and Applied Problems of Math, Physics and Computer Science" (2020).