Models for Effective Categorization and Classification of Texts into Specific Thematic Groups (Using Gender and Criminal Themes as Examples)

Nina Khairova[1,2,†], Yevhen Kupriianov[1,†], Anzhelika Vorzhevitina[1,*,†], Oleksandr Shanidze[1,†]

[1] National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str. 2, Kharkiv, 61002, Ukraine
[2] Umeå University, Umeå, 901 87, Sweden

Abstract

An analysis of existing automated methods for text classification, used to develop an effective approach for automated text classification by thematic groups in the context of information related to criminal and gender themes, was conducted. Based on the analysis of classification methods, an algorithm for classifying texts by types of crime and gender was developed, information-linguistic and software for the task of distributing texts into thematic groups were developed, and the effectiveness of the developed application was assessed.

Keywords

Classification, categorization, thematic groups, criminal justice theme, gender stereotypes, social practices, gender criminology

1. Introduction

Distributing texts into defined groups is a relevant task as it allows for organizing and structuring information, and also increases the efficiency and accuracy of searching for necessary texts, even based on vague criteria. Text distribution into thematic groups is an actively developing area of research in the field of Natural Language Processing (NLP) and machine learning. This allows for improving information retrieval processes, texts classification, and content analysis. Solutions based on thematic text distribution have many practical applications, including creating indexed databases, automatic categorization of articles, content filtering, or personalized search [13].

The article suggests considering models for effective categorization and classification of texts into specific thematic groups (using the examples of gender and criminal themes). The task of distributing texts into thematic groups, particularly criminal texts, is relevant today and holds great potential in improving the work of law enforcement agencies for several reasons:

1. Due to the expansion of the volume of criminal materials. Due to the expansion of increase in the number of crimes and the volume of materials related to crimes, it is necessary to use more effective methods of analysis and classification of texts.
2. The need for automation of the analysis process arises from the fact that manual processing of materials takes a lot of time and requires significant effort. Classification methods allow for the automation of processing and analysis.
3. Increase in efficiency. By using classification methods, it is possible to categorize criminal materials, enabling the establishment of dependencies and patterns in criminal activities. This ensures an increase in efficiency in crime prevention and investigation.

CLW-2024: Computational Linguistics Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2024), April 12–13, 2024, Lviv, Ukraine
* Corresponding author.
† These authors contributed equally.
nina_khajrova@yahoo.com (N. Khairova); eugeniokuprianov@gmail.com (Y. Kupriianov); anzhelika.vorzhevitina.sgt.khpi@gmail.com (A. Vorzhevitina); s.alexandr21@gmail.com (O. Shanidze)
0000-0002-9826-0286 (N. Khairova); 0000-0002-0801-1789 (Y. Kupriianov); 0009-0001-0562-0191 (A. Vorzhevitina); 0000-0002-0484-6746 (O. Shanidze)

Gender issues have long gone beyond the limits of only sociological science and today can be considered in the aspect of various scientific practices. Science studies the relationship between modern social practices among gender individuals and the historical context that has led to the emergence of certain gender stereotypes, gender discrimination, or gender inclusion. Legal and political sciences explore the practical, legislative aspect of attitudes towards a gendered society. However, one of the promising vectors in the direction of gender issues is research based on linguistic, language, and other scientific [16-19].

From sociology, it is known that gender is the social sex of a person, at least within the framework of modernist thinking of scholars. Modern gender studies and perspectives on this issue suggest that gender is the personal state of a person's identity, which consists not only in the free choice of one's preferences, but also in applying to social practices - subjectivity on the part of each participant in such practices. Almost from birth, gender stereotypes are imposed on us, which can not only affect the quality of a person's life (in a negative way), but also become the cause of criminal offenses. Of course, there is no direct path from ”girls wear pink” to perceiving domestic violence as a norm. However, most often domestic violence is precisely the result of stereotypical endowment of a man with “strength”, “patriarchy”. However, not only “domestic violence” is part of gender discrimination, which can lead to involvement in criminal records.

The aim of the research is to create an effective approach to automated classification of texts into thematic groups in the context of information related to criminal and gender themes. At the same time, urgent tasks arise: researching automated methods of text classification, developing an algorithm for classifying texts by types of crime and gender, creating information-linguistic and software provisions for distributing texts into thematic groups, and evaluating the effectiveness of the effectiveness of the developed application.

The object of the research is the methods of classifying texts related to gender and crime.

The subject of the research is the methods of categorization and classification of texts.

The article provides a detailed overview of the subject area and formulation of the research task. Methods for developing thematic classification of texts are considered. Information-linguistic and software provisions tasks are described. An analysis and evaluation of the obtained results are conducted.

The theoretical significance of this work lies in the search for methods and approaches of the classification of texts based on machine learning methods, which could potentially be used to address a range of tasks in the fields of sociology and law.

The practical value is that the texts classification model in real conditions can allow studying the behavior of the classification algorithm and correct possible errors. Researches results can be a useful tool for criminal police departments, investigative and operational units and human rights organizations, due to a better understanding of the structure and characteristics of criminal activity, and a reduction in the level of criminal cases.

2. Methods of text classification

The modern approaches to automatic text categorization involve the use of various methods and algorithms to automatically classify textual data according to their themes in order to systematize, organize, and facilitate access to large volumes of information to improve the processes of information analysis and retrieval [6]. The distribution of texts by thematic groups allows you to effectively organize texts by various categories, facilitating the process of searching and gathering information, contributes to improving the quality of the results of analytical research, and makes calculations more accurate and meaningful. Methods of text classification based on rules, machine learning, and deep learning are used to group textual data into different categories based on predefined criteria.

In comparison to rule-based classification, machine learning and deep learning methods are more effective for categorizing texts into thematic groups, as they can automatically identify complex dependencies and patterns in data, allowing for more accurate distribution of textual

documents into thematic groups in areas with a large amount of data. Consequently, machine learning and deep learning methods are typically more powerful and efficient for text classification into thematic groups in area with a large amount of data.

Considered modern methods and algorithms of classification based on machine learning for the distribution of texts by thematic groups include the Naive Bayes classifier, Support Vector Machine, neural networks, decision trees, k-Nearest Neighbors Algorithm, Vector Space Model and others. These methods allow for the effective distribution of texts into thematic groups, facilitating the detection and analysis of various aspects within the researched domain.

An analysis of existing classification methods has shown that machine learning based methods are superior for text classification for several reasons. Firstly, these methods have high flexibility and ability to learn on large volumes of data. Secondly, they uncover hidden and complex patterns and relationships between textual data, which is crucial for accurate text classification. Thirdly, they can consider numerous features and relationships among them, such as linguistic characteristics, semantic context, and contextual relationships, which helps improve the accuracy of classification within thematic groups.

For example, the naive Bayes classifier is suitable for classification tasks with large volumes of data, where speed of operation and ease of implementation are of great importance. The classification method using neural networks is a powerful machine learning tool that has a number of advantages and features for working with diverse data and can provide high accuracy and adaptability in classification tasks. The vector space model is a powerful tool for analyzing textual information, especially in cases where accounting for semantic relationships and working with large volumes of text is needed [10]. The choice of classification method depends on the type of data, the volume and diversity of the data, computational resources, and the required interpretation of the results.

The problem of distributing texts into thematic groups is relevant in today's information world. Processing a large volume and variety of topics complicates the task. Multilingualism also complicates the analysis and classification of text due to syntactic, semantic, and cultural differences between languages. The development of new methods and algorithms, including the application of Text Mining methods, can help to address these issues, particularly in the context of criminal themes and gender issues, where automatic classification and analysis of texts can increase the efficiency and speed of making informed decisions based on these data [2, 17].

Methods for determining semantic proximity are important in the modern context of text analysis and categorization into thematic groups. The development of such methods allows for improving the accuracy and efficiency of systems, simplifying the work of sociologists, law enforcement agencies, and other stakeholders. There are several methods for determining semantic proximity, which are based on statistical analysis methods and linguistic methods and rules. The methods of determining semantic proximity, which are based on methods of statistical analysis, include:

1. Vector models of words
2. Using vector representations of words
3. Matrices of relationships between words
4. Machine learning methods for determining the proximity of words

The methods for determining semantic proximity based on linguistic methods, rules, and dependencies include:

1. Use of semantic roles and syntactic structure
2. Comparison of dependencies and relations between words
3. Analysis of semantic components of sentences

Each of these methods has its own characteristics, and the choice of method depends on the task setting and context, available data, and the context of application [9, 14, 18]. In practical applications, statistical methods are often used where large volumes of data are available, such as machine translation, speech recognition, or text analytics. Linguistic methods can be useful in cases where the focus is on accuracy and understanding complex semantics, such as named entity recognition, information extraction, or in the field of deep semantic analysis.

The tasks of text analysis include identifying keywords, extracting information, classifying texts into categories, automatic translation, etc. For the effective performance of these tasks, reliable methods of determining text features are needed to take into account semantic and statistical information. Realization of the relevance of text analysis tasks and the need for accurate and informative methods of extracting features from textual data emphasizes the importance of research and development of models, such as the vector space model, for solving these tasks.

The vector space model is a way of representing text as numerical vectors, where each word is represented as a vector according to its meaning in the context of the text. This approach allows measuring the semantic similarity between different textual documents. Once the text is transformed into vectors, various data analysis methods can be applied, such as clustering, classification, or determining the similarity of textual documents. The vector space model is applied in information retrieval systems to effectively determine similarity between users queries and documents [3, 7, 15].

The cosine similarity measures the similarity between two vectors in a multi-dimensional space and expresses the similarity of their directions by determining the cosine of the angle between them, taking values from -1 to 1, providing a clear understanding of the degree of semantic closeness. The closer the cosine similarity values are to 1, the more semantically similar the textual fragments are. The equation for calculating the cosine similarity between two vectors A and B is defined as the ratio of their dot product to the product of their lengths in the equation (1).

$$similarity = cos(\theta) = \frac{A * B}{\|A\|\|B\|}, \quad (1)$$

where A * B is the dot product of vectors A and B; cos(θ) is the cosine of the angle between vectors A and B; ||A|| is the length of vector A; ||B|| is the length of vector B.

The main advantages of using the cosine similarity metric compared to other semantic proximity metrics is that it is robust to scaling. This means that the determination of angular similarity between vectors is not dependent on their length or scale, making cosine similarity a convenient and effective method for comparing vectors of different lengths or scales. Furthermore, cosine similarity is widely utilized in the field of natural language processing, text classification, and data analysis [5, 12]. Thematic proximity is defined as the degree of semantic similarity between textual documents or text fragments directed towards mainly a specific theme, concept, or field of knowledge.

Stages of determining narrowly thematic closeness by cosine similarity:

1. Representation of texts in vector form
2. Measuring the cosine similarity
3. Comparing cosine similarity results

Representation of texts in vector form. Term frequency (TF) measures how often a term occurs in a document. This is calculated in equation (2).

$$TF_t = \frac{(the\ number\ of\ times\ the\ term\ t\ occurs\ in\ the\ document)}{(the\ total\ number\ of\ terms\ in\ the\ document)}, \quad (2)$$

where TF is term frequency; t is the term occurring in the document.

Inverse Document Frequency (IDF) measures how unique a term is across the entire document corpus. This is calculated in equation (3).

$$IDF_t = \frac{\log(the\ number\ of\ documents\ in\ the\ corpus)}{(the\ number\ of\ documents\ containing\ the\ term\ t)}, \quad (3)$$

where IDF is Inverse Document Frequency; t is the term occurring in the document.

TF-IDF vector. Each word in the document can be represented by a vector, where each component corresponds to the TF-IDF value for the respective term. This allows for comparison of documents based on their content and used for further analytical tasks. Therefore, TF-IDF is used to highlight important words in the text and construct a vector representation of documents for further analysis.

Measuring the cosine similarity between vectors of texts begins with the preparation stage. Two vectors, A and B, representing the texts to be compared, need to be prepared. The calculation

of the dot product in equation (4). For this, each component of vector A is multiplied by the corresponding component of vector B. This gives us the numerator of the equation for cosine similarity as per equation (1).

$$dot\ product = A * B, \quad (4)$$

where A and B are vectors.

Calculation of vector norms. We calculate the Euclidean norm (length) of vector A and vector B. The vector's norm is calculated as the square root of the sum of the squares of its components in equation (5). This will give us the denominator of the equation for cosine similarity as per equation (1).

$$vector\ norms = \|A\|\|B\|, \quad (5)$$

where ||A|| is the length of vector A; ||B|| is the length of vector B.

Calculation of cosine similarity. Divide the dot product of vectors A and B by the product of their norms, which gives us the cosine value of the angle between these vectors, see formula (1). Evaluation of cosine similarity. The result of the calculation will be a value ranging from -1 to 1. A value of 1 indicates similarity, 0 indicates no similarity, and -1 indicates complete dissimilarity between the vectors. This step-by-step calculation provides an assessment of the degree of semantic similarity between texts based on the cosine angle between their vector representations.

Comparing cosine similarity results. After calculating the cosine similarity for each pair of texts, we obtain numerical values reflecting the degree of their semantic similarity.

The TF-IDF vectorization allows for the identification of important terms in a text and the construction of a vector representation of documents, taking into account their significance for analysis. This makes it a valuable tool for detecting and summarizing essential information from textual documents. Overall, the algorithm for representing texts in vector form, followed by the calculation of cosine similarity and comparison of results, allows for the effective assessment of thematic similarity of texts. It is believed that in future research, the prospects of using cosine similarity include refining methods of text vectorization, developing combined models that utilize cosine similarity in combination with other methods to improve the overall accuracy of text data analysis.

3. Information-linguistic and software provisions

In order to develop informational and linguistic support for the task of distributing texts based on thematic directions, specifically in the areas of criminal themes and gender issues, the integrated development environment Sublime Text was chosen, in which the previously created dictionary with vocabulary on criminal topics, which included gender crimes, was improved. The created dictionary contained structured vocabulary, namely: nouns, adjectives, verbs. Vocabulary on criminal topics was provided in Ukrainian, in English languages. Improving the dictionary, phrasal verbs and stable expressions were added to the existing vocabulary. Additionally, another foreign language, French, was incorporated to present the existing vocabulary in the dictionary. Structure of an XML article for a single lexical item with the annotation “phrasal verb” presented in three languages.

The first line is the parent element of a block for all other child elements, with the <term> tag and an “id” attribute containing the ordinal number of the phrasal verb. The second line has a <lemma> tag with the lang=”ua” parameter, which represents the base form of the phrasal verb in the Ukrainian language. The third line contains the <domain> tag, which denotes the narrow thematic group to which the lexical unit of the phrasal verb related to criminal theme is assigned, as shown in Figure 1 [1].

#	Event type	Event subtype
1	TRANSFER	Movement, Traffic Accident
2	CRIME	Injure, Offense
3	POLICE	Arrest, Trial, PD

Figure 1: Narrow thematic groups of criminal themes

The fourth line has an <example> tag with the lang=”ua” parameter - an example sentence using the phrasal verb in the Ukrainian language. The fifth line has a <lemma> tag with the lang=”en” parameter - the base form of the phrasal verb in English. The sixth line has an <example> tag with the lang=”en” parameter - an example sentence using the specified phrasal verb in English. The seventh line has a <lemma> tag with the lang=”fr” parameter - the base form of the phrasal verb in French. The eighth line has an <example> tag with the lang=”fr” parameter - an example sentence using the phrasal verb in French. The ninth line closes the tag of the first line with the tag </term>.

The XML structure of an article for a lexical unit marked as “stable expression”, presented in three languages, bears a significant resemblance to the XML structure for a lexical unit marked as “phrasal verb”, but still has its own specific features.

Type of element <term>. The first line – the <term> tag and the “id” attribute representing the ordinal number of the stable expression. The second line – the <domain> tag, indicating the narrow thematic group to which the lexical unit of the stable expression related to criminal themes belongs. The third line – the <lemma> tag with the lang=”ua” parameter, representing the base form of the stable expression in the Ukrainian language. The fourth line – the <definition> tag with the lang=”ua” parameter, denotes the definition of the stable expression in Ukrainian. The fifth line – the <example> tag with the lang=”ua” parameter, provides a sentence in Ukrainian illustrating the use of the stable expression. The sixth line – the <lemma> tag with the lang=”en” parameter, representing the base form of the stable expression in English. The seventh line – the <definition> tag with the lang=”en” parameter, denotes the definition of the stable expression in English. The eighth line – the <example> tag with the lang=”en” parameter, provides a sentence in English illustrating the use of the stable expression. The ninth line – the <lemma> tag with the lang=”fr” parameter, representing the base form of the stable expression in French. The tenth line – the <definition> tag with the lang=”fr” parameter, denotes the definition of the stable expression in French. The eleventh line – the <example> tag with the lang=”fr” parameter, provides a sentence in French illustrating the use of the stable expression. The twelfth line has a closing tag for the first line with the tag </term>.

In order to develop the software code that performs texts distribution by types of criminality based on user queries where the user searches for texts based on words related to criminal themes, a wide range of libraries, modules, and functions were used, including tkinter, ttk, scrolledtext, messagebox, pandas, pip, BeautifulSoup, TfidfVectorizer, cosine similarity, lxml, openpyxl, and numpy. These libraries are used for various tasks, such as creating GUI, working with structured data, web page parsing, text vectorization, calculating similarity between documents, and much more.

Next, we create the “main_choice” function, which reads data from the Excel file “Criminal Corpus.xlsx”, containing information about criminal texts, and reads data from the XML file “new_voc_tagged_summer_2023.fr.xml”, containing a lexical dictionary. This function performs a search for lexical terms and their corresponding domains for the selected language. After that, the “main_choice” function returns the text corpus “corpus” and the dictionary “word_domain_dict”, where the keys are words and the values are their corresponding domains.

The function “on_word_entry_changed(event)” is an event handler used to respond to changes in the “word_entry” input field. When the text in this field changes, the function is triggered, retrieves the text, and converts it to lowercase. It retrieves the language selected by the user from the “combo_lang” combo box. It then calls the “main_choice(lang)” function, which selects the corresponding corpus of texts and dictionary of domain words. This function is linked to the “find_texts()” function, which is called when the “Find Similar Texts” button is pressed and uses the currents words and language to find similar texts and their domains.

The next function is “find_texts()”, which searches for similar texts based on the words entered by the user. It uses the “main_choice(lang)” function to load the corpus of texts in the selected language - Ukrainian, English, French, and dictionary of domain words for each of these languages. It then processes the entered text, identifies its words and their corresponding

domains, and builds TF-IDF vectors for the corpus of texts. After vectorization, the “find_texts()” function compares the vectors belonging to the user query with the vectors of texts from the corpus using cosine proximity and obtains a ranked list of the most similar texts.

Upon receiving the cosine similarity values, it is important to consider that if words from the dictionary have different “domain”, the cosine similarity is calculated for each word with the text. This allows for more detailed detection of the semantic relevance of each individual word to each text. And as a result, for each text, the “domain” that belongs to the word with the largest coefficient of cosine proximity is displayed.

In order to make the program user-friendly, the program code was created in the PyCharm integrated development environment, which represents a graphical user interface (GUI) that uses the Tkinter library to create the program window and place the interface elements.

In the modern world, trust in criminal data and their analysis is crucial, therefore, the use of Python for classifying criminal texts is one of the tools that can help law enforcement effectively combat crime and ensure public safety. Furthermore, Python offers further possibilities for improving text classification, including the use of images and location data to enhance hypothesis testing and establish connections between different types of crimes. All these capabilities underscore the importance of Python in the development of innovative technologies and their application in law enforcement. With this programming language, experts will be able to identify key words and interpret complex criminal patterns, ultimately contributing to safeguarding public safety.

The invasion of Ukraine by the Russian Federation violates all possible laws, both de jure and de facto. Genocide, ecocide, violation of the laws of warfare, violation of any legal norms. It is clear that gender discrimination is also present in the actions of the aggressor country. In this case, it becomes relevant to create effective methods of categorization and classification of texts according to specific thematic groups, to identify narrowly targeted markers that indicate a criminal offense in the field of gender discrimination, as well as the inclusion of similar crimes or themes in the corpus of relevant texts.

So, for example, in the linguistic aspect, applying the analysis of an array of texts, the following factor will certainly be found: gender issue, used by the aggressor country as a means of propaganda. Scientific facts and research on gender issues are distorted. Additionally, practical aspects of gender sociology suffer at the level of undermining equal opportunities for all citizens of Ukraine, gender-based violence, and discrimination by the aggressor country towards our citizens.

The occupying country utilizes gender stereotypes to discredit Ukrainian statehood, the Ukrainian people, Ukrainian authorities, and defenders of Ukraine. An aggressive (i.e., negative) feminization of Ukrainian politicians and military personnel is occurring. Even the concept of “Ukraine” as a metaphysical entity is portrayed for propaganda purposes as a “weak woman”, when the enemy is depicted through the prism of a “patriarchal man who is always right” [19].

The user guide, in the context of the application's web interface in the PyCharm environment, is essential for providing users with the necessary information and guidance for use. User instructions for using the program for distributing texts by thematic groups on criminal topics:

1. Launch the PyCharm integrated development environment
2. Open the project “criminal_classifier”
3. Run the “main” file of the “criminal_classifier” project
4. The “Text Similarity Analysis” window will open
5. Choose the language for classification. In the “Select language” section, you can select one of the three available languages: “Ukrainian”, “English”, “French”. For example, select the language of the query as “Ukrainian”
6. In the “Enter a word” field, enter any number of words related to criminal topics in the selected language, separated by commas. For example, enter the words “protocol” with the domain “arrest” according to the dictionary. However, if any of the words are not recognized by the dictionary or do not belong to the criminal topic, the web interface will display an error message, as shown in Figure 2

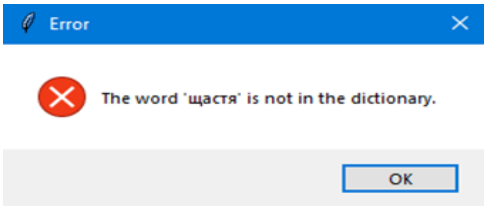

Figure 2: Request error

7. Enter the number of texts that we want to see as a result of the program for distributing texts by thematic groups on criminal and gender topics. This quantity of texts should be entered in the field “Number of documents to be displayed”. For example, 123, as shown in Figure 3

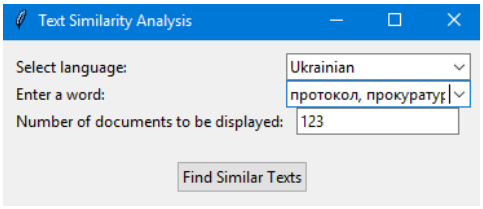

Figure 3: Program interface

8. Click on the “Find Similar Texts” button
9. A new dialog window “Text Similarity Results” will open with the results. As an example, we see the results that correspond to the chosen language, the entered word, and the number of texts. The calculated cosine similarity is displayed under “Similarity”, and the type of crime to which the text is related is displayed under “Domain”, which in this case corresponds to “arrest” for the word “protocol”, as this word is most frequently found in the text, shown in Figure 4

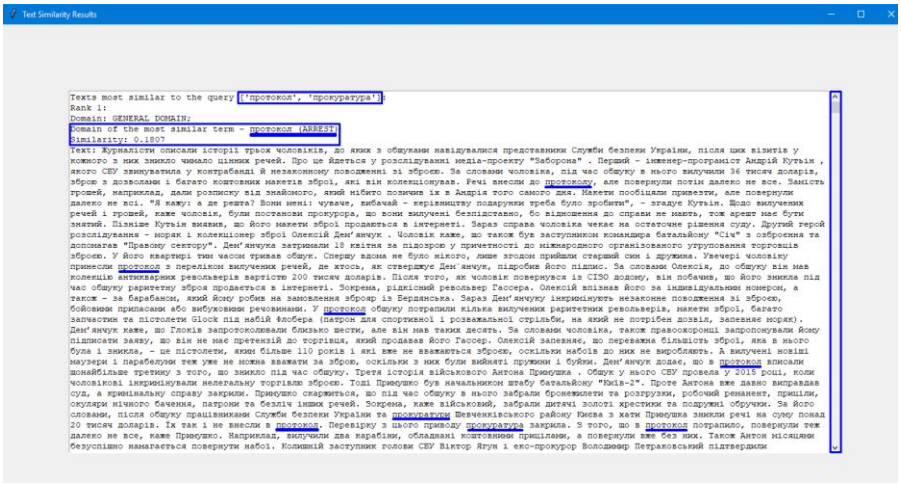

Figure 4: The result of the program

4. Accuracy of the program

Determining the effectiveness of text classification into thematic groups. Cohen's Kappa metric is a statistical measure used to assess the degree of agreement between two or more assessors who classify the same set of objects. It takes into account the randomness of agreement between assessors, making it a more reliable measure than simple agreement. Cohen's Kappa metric is determined using a formula that compares the observed agreement between assessors to the agreement that could be expected by chance. Cohen's Kappa metric is calculated using equation (6).

$$k = \frac{p_0 - p_e}{1 - p_e}, \tag{6}$$

where k is the symbol for Cohen's Kappa metric; p_o is the observed agreement between assessors or the empirical probability of agreement in assigning a class within any sample; p_e is the agreement that could be expected by chance.

The parameter po is measured as the fraction of objects that were classified uniformly by all assessors. The parameter pe represents the agreement that could be expected if the classification was done randomly. The value of Cohen's Kappa can range from -1 to 1, with 1 indicating complete agreement, 0 indicating random agreement, and a negative value indicating less agreement than would be expected by chance.

The Cohen's Kappa metric is widely used in the field of machine learning to measure the level of agreement between classifiers when evaluating classification quality, and it involves several steps:

1. Select a sample of texts
2. Distribute the sample into several categories
3. Select assessment assessors
4. Calculate the Cohen's Kappa metric to determine the degree of agreement between assessors and the effectiveness of classification
5. Analyze the results to determine the effectiveness of classifying texts into thematic groups

So, let's evaluate the program for the distribution of texts by thematic groups according to Cohen's Kappa metric. As experts, students of the National Technical University "Kharkiv Polytechnic Institute" from the Department of Intelligent Computer Systems specializing in Applied and Computer Linguistics were involved. It was decided to determine the expert evaluation of the results of the program for the Ukrainian and English languages, with the aim of greater purity of the experiment. 100 texts were taken for one word, for two and for three words, the total number of texts is 300, the results of consistency and inconsistency are presented in the Table 1.

Table 1
Expert evaluation by Cohen's Kappa metric

	Expert 2				
Expert 1		Ukrainian		English	
		1	0	1	0
	1	257	15	265	6
	0	8	20	8	21

Cohen's Kappa metric is calculated using the formula, see formula (6), according to this formula we get:

- p_o = (Both said "Yes" + Both said "No") / (total number of texts)
- $p_{(matches)}$ = (total number when Expert 1 said "Yes" / total number of texts) * (total number when Expert 2 said "Yes" / total number of texts)
- $p_{(does\ not\ match)}$ = (total number of times Expert 1 said "No" / total number of texts) * (total number of times Expert 2 said "No" / total number of texts)
- $p_e = p_{(matches)} + p_{(does\ not\ match)}$
- $k = (p_o - p_e) / (1 - p_e)$

Let's apply Cohen's Kappa metric for the Ukrainian language:

- p_o = (257 + 20) / 300 = 277 / 300 = 0.923
- $p_{(matches)}$ = ((257 + 15) / 300) * ((257 + 8) / 300) = (272 / 300) * (265 / 300) = 0.906 * 0.883 = 0.799
- $p_{(does\ not\ match)}$ = ((8 + 20) / 300) * ((15 + 20) / 300) = (28 / 300) * (35 / 300) = 0.093 * 0.116 = 0.0107
- p_e = 0.799 + 0.0107 = 0.809
- k = (0.923 – 0.809) / (1 – 0.809) = 0.114 / 0.191 = 0.59

Let's apply Cohen's Kappa metric for the English language:

- p_o = (265 +27) / 300 = 292 / 300 = 0.97
- $p_{(matches)}$ = ((265 + 6) / 300) * ((265 + 8) / 300) = (271 / 300) * (273 / 300) = 0.903 * 0.91 = 0.821
- $p_{(does\ not\ match)}$ = ((8 + 21) / 300) * ((6 + 21) / 300) = (29 / 300) * (27 / 300) = 0.096 * 0.09 = 0.008
- p_e = 0.821 + 0.008 = 0.829
- k = (0.97 – 0.829) / (1 – 0.829) = 0.141 / 0.171 = 0.82

The result of the Cohen's Kappa metric, which is 0.59 for the Ukrainian language, corresponds to “Moderate agreement” according to the interpretation of the results. For the English language, it is 0.82, corresponding to “Near perfect agreement” according to the interpretation of the results of the Cohen's Kappa metric. The interpretation of the results of the Cohen's Kappa metric is shown in Figure 5 [4, 11].

Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

Figure 5: Interpretation of the results of the Cohen's Kappa metric

The test of text classification was carried out on 300 texts, of which the first 100 texts are for text classification, on request by one word. The next 100 texts, for text classification, on request by two words. And for the last 100 texts, for text classification, on request by three words. The classification was carried out so that there were no words in the query that have the same “domain”. The expert assessment of the result is presented in the Table 2.

Table 2
A fragment of the expert evaluation of the results of the program

Text	Ukrainian		English	
	Expert 1	Expert 2	Expert 1	Expert 2
1	1	1	1	1
2	1	1	0	0
3	1	1	1	0
4	0	0	1	1
5	0	0	1	1
6	1	1	1	1
7	1	1	0	0
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1

Let's generalize Table 2 to the general results presented in Table 3.

Table 3
General results of the expert evaluation of the program result

Ukrainian		English	
Matched	Didn`t match	Matched	Didn`t match

100 texts for 1 word	Expert 1	80	20	Expert 1	91	9
	Expert 2	82	18	Expert 2	22	12
100 texts for 2 words	Expert 1	95	5	Expert 1	89	11
	Expert 2	90	10	Expert 2	92	8
100 texts for 3 words	Expert 1	97	3	Expert 1	91	9
	Expert 2	93	7	Expert 2	93	7

Calculation of Accuracy, for this, it is necessary to divide the number that coincides with the classification by the total number of texts, that were investigated for accuracy, as evident from equation (7).

$$accuracy_{1,2} = true \div 300, \quad (7)$$

where $accuracy_{1,2}$ is the accuracy for experts 1 and 2; true is the number of correctly classified texts; 300 is the total number of texts for which expert assessment was conducted.

To determine the average value between the two experts, it is necessary to calculate the $Accuracy_{average}$ as shown in equation (8).

$$accuracy_{average} = accuracy_1 \div accuracy_2, \quad (8)$$

where $accuracy_{average}$ is the average accuracy between the two experts; $accuracy_{1,2}$ is the accuracy for experts 1 and 2.

For the Ukrainian language:

- For expert 1, $accuracy_1$ = (80 + 95 + 97) / 300 = 272 / 300 = 0.906
- For expert 2, $accuracy_2$ = (82 + 90 + 93) / 300 = 265 / 300 = 0.883

For the English language:

- For expert 1, $accuracy_1$ = (91 + 89 + 91) / 300 = 271 / 300 = 0.903
- For expert 2, $accuracy_2$ = (88 + 92 + 93) / 300 = 273 / 300 = 0.91

Now let's calculate the average percentage of expert assessment for the selected languages.

For the Ukrainian language:

- $Accuracy_{average}$ = (0.906 + 0.883) / 2 = 1.809 / 2 = 0.904 = 89.4%

For the English language:

- $Accuracy_{average}$ = (0.903 + 0.91) / 2 = 1.793 / 2 = 0.896 = 90.6%

Thus, based on the expert assessment results of the text classification program, it can be concluded that the accuracy for the Ukrainian language is 89.4%, and for the English language, it is 90.6%. Therefore, we have very good results for the program of classifying texts into thematic groups on criminal gender topics.

5. Conclusions

The article analyzed the subject area of the research, namely, reviewed modern approaches to the automatic distribution of texts according to certain directions, analyzed existing classification methods, reviewed the problems of thematic classification of texts, reviewed the use of Text Mining capabilities in the direction of solving problems related to legal and gender topics. Methods for determining semantic similarity are considered, the use of the Vector Space Model to determine text features, the use of the cosine similarity metric in tasks of determining semantic similarity, and an algorithm for determining thematic similarity of texts is described.

The information and linguistic support for the task of distributing texts by thematic areas has been improved, a multilingual corpus and DataFrame of news article texts has been created using parsing methods, software for distributing texts by types of crime and gender features has been created, using vectorization of texts and user queries and cosine proximity, a web interface has been created user of the task of distributing texts by thematic directions.

The choice of programming language and libraries has been justified, and the user's instructions for using the program for distributing texts by narrowly thematic groups have been drawn up. Cohen's Kappa metric was used to determine the effectiveness of text classification by

certain groups, the results of which are equal to 0.59 for the Ukrainian language, i.e. “Moderate agreement”, for the English language it is equal to 0.82, i.e. “Near perfect agreement”. The expert evaluation of the results of the program was determined, which are equal to 89.4% for Ukrainian and 90.6% for English, which indicates good results of the program for the distribution of texts by certain groups. In the future, it is planned to expand the created system, to add new terms, languages, domains to it. For the web application, it is planned to expand the functionality and transform the web application from local to online.

References

[1] N. Khairova, O. Mamyrbayev, N. Rizun, M. Razno, G. Ybytayeva, A Parallel Corpus-Based Approach to the Crime Event Extraction for Low-Resource Languages, IEEE 11 (2023). doi:10.1109/ACCESS.2023.3281680.

[2] LinkedIn, AI based Data classification modeling techniques using NLP Classification, 2023. URL: https://www.linkedin.com/pulse/ai-based-data-classification-modeling-techniques.

[3] Bilal Abu-Salih, Applying Vector Space Model (VSM) Techniques in Information Retrieval for Arabic Language, 2018. URL: https://arxiv.org/ftp/arxiv/papers/1801/1801.03627.pdf.

[4] B. M. Doddagangavadi, B. N. Murthy, N. Rajpurohit, Statistical Tool for Testing Agreement Level on Continuous Datasets, Current Research in Biostatistics 11 (2021) 1–11. doi:10.3844/amjbsp.2021.1.11.

[5] Medium, Cosine Similarity, 2023. URL: https://medium.com/@TheDataScience-ProF/cosine-similarity-6cbba097b3db.

[6] G. Bryda, A. P. Costa, Qualitative Research in Digital Era: Innovations, Methodologies and Collaborations, Social Sciences 12.10 (2023). doi:10.3390/socsci12100570.

[7] Medium, Text Processing Techniques on Twitter data, 2020. URL: https://towardsdatascience.com/text-processing-techniques-on-twitter-data-69233296c778.

[8] Marketbrew, The Impact of Machine Learning on SEO, 2023. URL: https://marketbrew.ai/the-impact-of-machine-learning-on-seo.

[9] S. Biggins, S. Mohammed, S. Oakley, L. Stringer, M. Stevenson, J. Priess, University_Of_Sheffield: Two Approaches to Semantic Text Similarity, in: Proceedings of the 1th Joint Conference on Lexical and Computational Semantics, Montreal, 2012, pp. 655–661. URL: https://aclanthology.org/S12-1097.pdf.

[10] V. Kant Singh, V. Kumar Singh, Vector space model: an information retrieval system, in: Proceedings of BITCON-2015 Innovations For National Developmentational Conference on Information Technology Empowering Digital India, International Journal of Advanced Engineering Research and Studies, Durg 2022, pp. 141–143. URL: https://www.researchgate.net/publication/362060638_VECTOR_SPACE_MODEL_AN_INFORMATION_RETRIEVAL_SYSTEM.

[11] A. S. Kolesnyk, N. F. Khairova, Justification for the Use of Cohen’s Kappa Statistic in Experimental Studies of NLP and Text Mining, Cybernetics and Systems Analysis 58 (2022) 280–288. doi:10.1007/s10559-022-00460-3.

[12] J. Wang, Y. Dong, Measurement of Text Similarity: A Survey, Information 11.9 (2020) 421. doi: 10.3390/info11090421.

[13] University of San Diego, The Role of Natural Language Processing in AI, 2023. URL: https://onlinedegrees.sandiego.edu/natural-language-processing-overview/.

[14] M. Eminagaoglu, A new similarity measure for vector space models in text classification and information retrieval, 48.4 (2022) 463–476. doi/10.1177/0165551520968055.

[15] M. Biryukova, I. Kyrychenko, N. Shanidze, O. Shanidze, Social Computing of the Social Well-being of Refugees and Internally Displaced Persons in Ukraine Using Data Mining Methods, in: Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems, COLINS-2023, Kharkiv, 2023, pp. 410–422. URL: https://ceur-ws.org/Vol-3403/paper32.pdf.

[16] Y. Kalagin, O. Shanidze, Hendernyi analiz sotsialnykh praktyk ukrainskykh bizhentsiv 2022 roku, Naukovo-teoretychnyi almanakh Hrani 26.3 (2023) 62–68. doi:10.15421/172349.
[17] Y. Kalagin, O. Shanidze, Instytutsializatsiia sotsialnykh praktyk vnutrishno peremishchenykh osib v ukraini u 2014–2022 rokakh, Habitus 53 (2023) 26–32.
[18] I. Gruzdo, I. Kyrychenko, G. Tereshchenko, O. Shanidze, Analysis of Models Usability Methods Used on Design Stage to Increase Site Optimization, in: Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems, COLINS-2023, Kharkiv, 2023, pp. 387–409. URL: https://ceur-ws.org/Vol-3403/paper31.pdf.
[19] DW, Yak propahanda RF atakuie Ukrainu cherez henderni stereotypy, 2023. URL: https://www.dw.com/uk/ak-propaganda-rf-atakue-ukrainu-cerez-genderni-stereotipi/a-64807182.