

Intelligent system development for detecting suicidal intentions based on Ukrainian literary works analysis

Victoria Vysotska^{1,†}, Lyubomyr Chyrun^{2,†}, Serhii Vladov^{3,†}, Sofia Chyrun^{1,†}, Oksana Markiv^{1,†}, Liubov Kolyasa^{1,†}, Valerii Sokurenko^{4,†}, Oleksandr Muzychuk^{4,†}, Arsen Blyzniuk^{1,†} and Rostyslav Fedchuk^{1,†}

¹ Lviv Polytechnic National University, Stepan Bandera 12, 79013 Lviv, Ukraine

² Ivan Franko National University of Lviv, University 1, 79000 Lviv, Ukraine

³ Kremenchuk Flight College of Kharkiv National University of Internal Affairs, Peremohy Street 17/6 39605 Kremenchuk, Ukraine

⁴ Kharkiv National University of Internal Affairs, L. Landau Avenue 27 61080 Kharkiv, Ukraine

Abstract

The project aims to create a comprehensive methodology for the early detection of suicidal intentions, using the literary analysis of Ukrainian author's works, especially Mykola Khvylovy's. The object of the research is the linguistic and psychological aspects of literary works, which may indicate the suicidal intentions of the author. The subject of the study comprises language patterns and emotional indicators in Mykola Khvylovy's texts, which may be associated with suicidal thoughts. The scientific novelty consists in the development of the methodology that allows to analyze the literary texts to identify suicidal tendencies. This approach has not been used in this area before, but it can significantly contribute to psychological science and literary studies. The technique has great practical value, as it can be used to prevent suicide, providing the tool for early detection of suicidal intentions based on the analysis of written works.

Keywords

NLP, text analysis, author style Ukrainian literary works analysis, big data analysis

1. Introduction

The relevance of the given project is based on the need to develop innovative methods for detecting and preventing suicidal behaviour, which is becoming an increasingly urgent problem in the modern world, especially in Ukraine during and after the war. In recent years, there has been an alarming increase in cases of depression and suicidal thoughts, especially among young people. In conditions of this social context, the development of practical tools for the early detection of suicidal intentions is becoming critical. The introduction of this initiative can help to identify persons who are at high risk in time and provide them with the necessary support, which can contribute to the prevention of the tragic consequences of suicidal behaviour. Many algorithms and techniques for suicide detection and prevention have already been created. However, most of them are based on the analysis of the patient who has already expressed suicidal intentions or, unfortunately, has already had negative cases. These are mainly surveys, the questions for which are developed and the results of which are

CIAW-2024: Computational Intelligence Application Workshop, October 10-12, 2024, Lviv, Ukraine

* Corresponding author.

† These authors contributed equally.

✉ Victoria.A.Vysotska@lpnu.ua (V. Vysotska); Lyubomyr.Chyrun@lnu.edu.ua (L. Chyrun); serhii.vladov@univd.edu.ua (S. Vladov); sofia.chyrun.sa.2022@lpnu.ua (S. Chyrun); oksana.o.markiv@lpnu.ua (O. Markiv); kolyasa.liubov@gmail.com (L. Kolyasa); (V. Sokurenko); o.muzychuk23@gmail.com (O. Muzychuk); arsen.blyzniuk.sa.2020@lpnu.ua (A. Blyzniuk); rostyslav.b.fedchuk@lpnu.ua (R. Fedchuk)

ORCID 0000-0001-6417-3689 (V. Vysotska); 0000-0002-9448-1751 (L. Chyrun); 0000-0001-8009-5254 (S. Vladov); 0002-2829-0164 (S. Chyrun); 0000-0002-1691-1357 (O. Markiv); 0000-0002-9690-8042 (L. Kolyasa); 0000-0001-8923-5639 (V. Sokurenko); 0000-0001-8367-2504 (O. Muzychuk); 0009-0002-9294-174X (A. Blyzniuk); 0009-0002-6669-0369 (R. Fedchuk)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

processed by many psychologists, and analyses of social networks where people can reveal their intentions or links to the possibility of such intentions.

The project is aimed at the creation of a comprehensive methodology for the early detection of suicidal intentions, using the literary analysis of Mykola Khvylovy's works. The following tasks have been set to achieve this goal:

- Development of the algorithm for quantitative text analysis, which includes measuring the length of sentences and words;
- Use of statistical methods to identify abnormalities in language use and what may indicate suicidal tendencies;
- Analysis of the emotional colouring of language, in particular verb endings, to identify psychological states;
- Formation of the rules set that allows determining potential suicidal intentions based on the linguistic features of the texts.

2. Related works

In the beginning, it is necessary to consider the source [1]. It describes algorithms for detecting suicidal intentions. They are based on the analysis of the already existing methods expressing signs of suicidal intentions. When signs are detected, the person is asked questions and the suicide risk screening is carried out. If the screening result is negative (that is, the person has no intention to commit suicide), the algorithm ends its work with a recommendation to "continue normal self-care". If the screening result is positive (the person has intentions to commit suicide), the algorithm is divided into two more branches. The root of the ramification is the result of the question, "Is there a threat to safety that requires urgent treatment?". If so, the algorithm goes to algorithm B. If not, to C. Algorithm B is designed to determine the level of severity of the risk of suicide:

- High (suicidal ideation with suicidal intent and inability to remain safe regardless of support/assistance);
- Medium (suicidal ideation with suicidal intent and failure to maintain safety regardless of support/assistance);
- Low: (no suicidal intentions, no specific suicide plan and preparatory actions and high confidence, for example, of a family member, in the person's ability to independently maintain safety).

Further, the document presents additional information in panels (tables). For example, in Panel 2 (Main characteristics of risk stratification), the main characteristics and actions for acute and chronic risks are described. Moreover, Algorithm C is divided into three branches for three types of severity of risk from the B algorithm, and actions to reduce risk through treatment are prescribed. The idea is to reduce the risk to a low acuity level and move to a management step. Next, the document presents several panels and the table of recommendations: Panel 3 (Modified risk factors); Panel 4 (Treatment to reduce suicides); Panel 5 (Action plan in crisis responses); Panel 6 (Intervention to improve relationships). That is, the research provides interesting algorithms for identifying and reducing the level of risk understanding, as well as an excellent theoretical basis for the issues of the subject of the work.

Moreover, the following research is presented [2]. It is the voluminous and detailed work related to the psychological prevention of suicidal tendencies. The essence, types and means of suicidal behaviour are described in the study. Factors of occurrence protective anti-suicidal factors are also presented. More attention is paid to identifying the intentions of teenagers and young adults. The work describes how to form an adequate attitude of surrounding people to suicidal manifestations and the possibilities of their detection and overcoming. However, the most essential part of this work

[2] is page number 13, with the so-called "suicide risk determination map (V.M. Priymenko)". In the result, the purpose is to determine the risk of committing suicide.

Moreover, a form of conduct is individual. The equipment comprises a suicide risk card form, and the duration is 30 minutes, taking into consideration age from 18 years. The suicide risk map is used to identify the risk of committing suicide and the degree of such risk for persons who find themselves in a difficult life situation. The card has 31 suicide risk factors, the presence of which must be detected in the subject. It is filled out by a psychologist who is sufficiently familiar with the client's personality based on a free conversation with him. When filling out the card, there is no need to rely on the subjective assessments of the client but only on the impressions that the psychologist has received during the study of the anamnesis. With the help of this map, it is possible to determine the presence of suicidal intentions in people aged 18 years.

The following work is "Guide to Forensic Psychiatry" by A.A. Tkachenko, namely, the section concerning the diagnosis of suicidal intentions. In the work, patients are divided into two groups: those who have already attempted suicide and pre-suicide patients with specific manifestations of suicidal intentions. The primary attention of psychologists is given to the first group and the detection of the second group. The study proposes the use of an integrated analysis of two factors developed at the All-Union Scientific and Methodological Centre for Suicidology: suicidal and anti-suicidal [3]. Moreover, suicidal factors are divided into:

1. Group as socio-demographic (gender, age, professional and family status, history of illegal acts); medical (presence of one or another form of mental pathology);
2. Personal and situational as conflicts (localization, content, orientation, dynamics); degree of suicidal manifestations in the past and present;
3. Individual personnel as predisposing suicidal personality complexes, maladjustment forms and levels; the nearest (suicide-dangerous positions and conditions); immediate (suicidal tendencies depth and activity).

Anti-suicidal factors include intense emotional attachments to significant ones, parental duties, expressed sense of duty, preoccupation with one's health, dependence on public opinion and the desire to avoid condemnation from others, and having life plans. At the same time, it must be considered that most of the listed factors are variable. The level of suicidal risk diagnosed in a specific person cannot be automatically extrapolated to his future but requires careful and systematic re-examination. It means there is a need to study patients and investigate their lives, characters, families, work, conflicts, etc.

The next source of analysis is the address of practical psychologist O.M. Gurova, "Methodological materials for curators of academic groups regarding the recognition of suicidal thoughts and their effective actions at the stages of identifying and preventing destructive forms of behaviour among student youth." [4]. The work is also aimed at identifying suicidal intentions, which has the "prejudice - fact" structure. For example, the first prejudice is that most suicides are carried out with little or no warning. The fact is that most people give warning signals about possible suicide in the form of direct statements, physical, body signs, emotional reactions or behavioural manifestations. They report the possibility of choosing suicide as a means of relieving pain and tension, maintaining control, or compensating for loss. These signals often can be considered as "cries for help". However, the most important for the research are the indicators that present the growth of suicidal tendencies among student youth. The author singles out the following indicators: situational indicators (any life situation subjectively perceived by a person as a crisis can be considered a situational indicator of suicidal risk), behavioural indicators of suicidal risk, communicative indicators, cognitive indicators, and emotional indicators. Suicide prevention in the student environment is also described (advice for curator): establishing a connection; identification of risks; notification of senior management; referral to vocational assistance; interaction with family; support for spasticity in programs; systematic control and consideration of the dynamics of changes in the student's personality and

behaviour. That is work on important "prejudices-facts", indicators of the growth of suicidal intentions and an algorithm of actions for the prevention of suicide.

The following work is [5] "Social and psychological factors and risk factors of suicide among young people" by A.R. Ivats, O.P. Romaniv, and B.Ya. Nagy. The research aims to analyze the leading causes of suicide among young people, identify risk factors that lead to suicide attempts, and highlight the characteristics of the behaviour of persons with a tendency to suicide. The work mentions four types of suicides: Selfish, Altruistic, Anomic, and Fatalistic. It also determined that the socio-psychological risk factors for the development of suicidal behaviour of young people include the following: family history of suicide; family history of violence; family history of psychoactive substance abuse; family history of mental health problems; feeling of hopelessness; feelings of isolation or loneliness; issues with the law; the influence of alcohol or drugs; the teenager disciplinary, social problems or difficulties at school; the problem with the use of psychoactive substances; mental disorder or mental illness; attempted suicide in the past; tendency to reckless or impulsive behaviour; weapon ownership; sleep deprivation; identification of oneself as being related to a person who committed suicide; psychomotor agitation, anxious or tense behaviour; changes in habits, sleep patterns, appetite; talking about one's own worthlessness, guilt, shame; consuming more alcohol than usual or starting to drink alcohol by people who have previously avoided him; careless or risky behaviour (reckless and dangerous acts); buying means for committing suicide (pills, weapons, poisonous substances); tendency to solitude, avoidance of close people; psychomotor excitement; statement about the desire "not to burden" loved ones; talks about own death and willingness to die; repentance and self-criticism; behavioural changes characteristic of people with suicidal thoughts and tendencies; frequent mood swings. As a result, the work demonstrates that the assessment of suicidal risk in a specific case is carried out by a specialist who, based on the data obtained during the interview, concludes the degree of formation of intentions, the available resources for solving the problem and the ability to use these resources for the benefit of the patient effectively. From the above, we can determine that the main idea of most research is questionnaires, communication and monitoring of a person [6-9]. The works demonstrate their exciting views on solving problematic situations, provide a robust basis for delving into the topic of research, and provide concrete solutions for the prevention, prevention and avoidance of committing suicidal acts [10-12]. A large number of works on the topic of the detection of suicidal intentions demonstrate the problem's importance since it is about a person's life [12-21]. It reveals the positive effect of solving the problem and is another unique way of determining suicidal intentions [21-27].

3. Materials and methods

The research's primary goal is to develop an innovative system for early detection of suicidal intentions using linguistic analysis. The primary function of the system is to assess the tendency for suicidal intentions based on the study of creativity. This function includes the following tasks: data collection/receiving data; data preparation (tokenization, lemmatization, removal of stop words, removal of punctuation marks); conducting analyses (content analysis; tonality analysis; lexical; psychological) [28-36]; visualization of results; output of the result. The program code consists of 4 components (blocks of code), each of which performs its direct functions and outputs changed data presented in Fig. 1. During the research, the state diagram is also created. The state diagram shows the states of the execution of certain parts of the code, and when an error is received, there is a transition to the beginning. The result of creating this diagram is shown below in Fig. 2. The system accepts the input data as docx files. This data is then prepared and combined into one object of type string. Then, the data is pre-processed. The analysis is conducted based on the prepared data. As a result, visualization and the result in percentages are displayed. Also, two files are created to remove stop words and punctuation, where the corresponding values are written. That is, files with works and two files with stop words and punctuation marks are submitted to the system. For work files, there is a requirement for file extension. It must be docx. At the output, the system provides a visualization of the analysis and the percentage of the author's propensity to suicidal intentions.

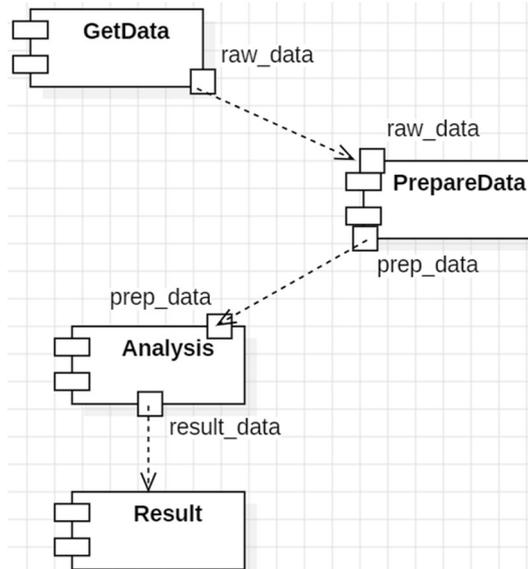


Figure 1: Component diagram.

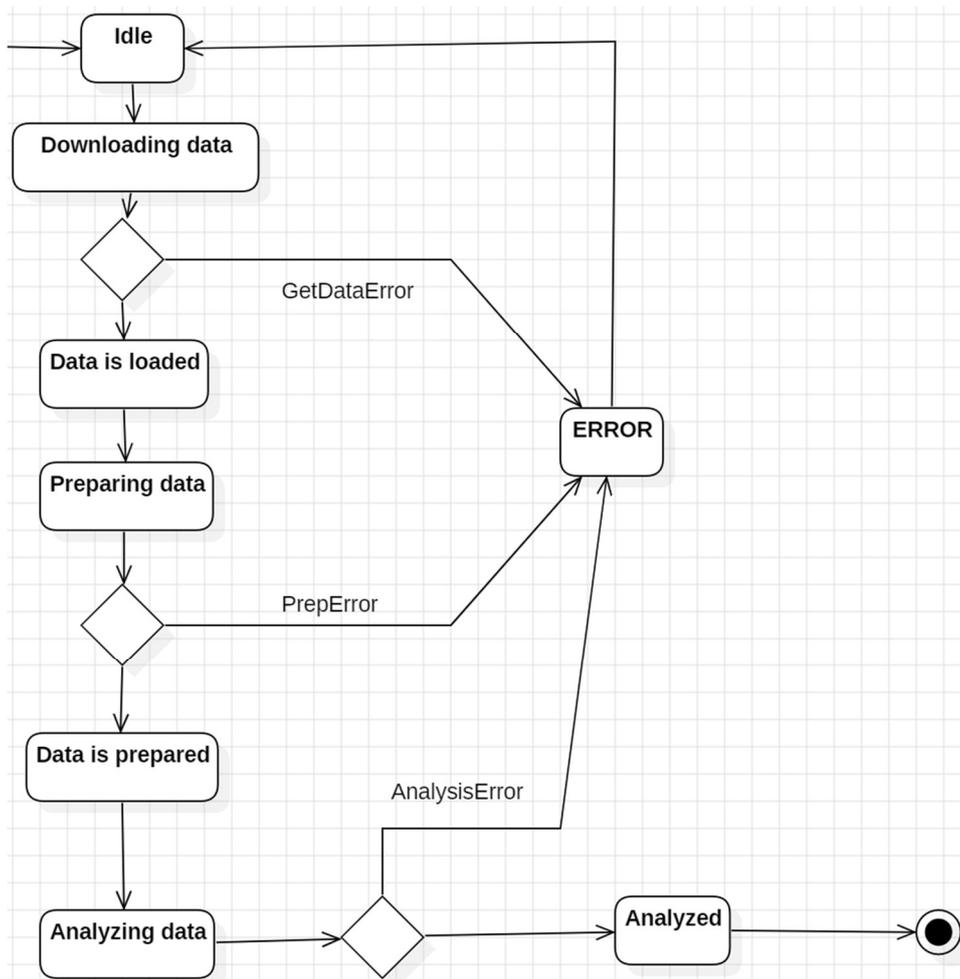


Figure 2: State diagram

1. Methods for text analysis are the following:

- The content analysis method allows the analysis of the content of the text, revealing the frequency of certain words or phrases that may indicate suicidal thoughts;

- Tonality analysis uses software to determine the emotional colour of the text, which may indicate depressive states or suicidal moods;
- The lexical analysis uses specific lexical units that may be associated with suicidal tendencies;
- Psychological analysis studies the psychological motives described in the works and their possible connection with the author's suicidal intentions.

2. The following methods are used to prepare the text for specific analyses:

- Tokenization is the process of dividing text into separate elements called tokens. Tokens can be words, numbers, symbols, or even sentences. This process is the first step in many natural language processing (NLP) tasks because it allows programs to parse and interpret text more easily.
- Lemmatization is reducing the word to its basic form, or lemma. It allows combining different word forms, such as verb tenses or noun cases, under one basic form, which facilitates further text analysis.
- Remove stop words, such as usually common words in the language that do not carry significant information for text analysis, such as "and", "but", "or", etc. Removing stop words can help focus on more meaningful words in the text.
- Removal of punctuation marks such as periods, commas, and parentheses are often removed from the text before parsing, as they may not be needed for specific NLP tasks and may complicate text processing.

The programming language used for text analysis is Python, and the following libraries for NLP:

- NLTK (Natural Language Toolkit) is one of the oldest and most used libraries for NLP. It contains packages for various NLP tasks, including tokenization, stemming, part-of-speech tagging, and stop-word removal.
- spaCy is a modern library that focuses on speed and efficiency. spaCy is used for complex NLP tasks such as named entity recognition and automatic tagging of parts of speech and dependencies.
- TextBlob is a simple library for processing text data. It makes performing NLP tasks such as tokenization, stemming, tonality analysis, and translation easy.
- The Gensim library is focused on topic modelling and semantic analysis. It is excellent for working with large text corpora and discovering structure in text data.
- Hugging Face Transformers is a library that provides access to pre-trained models such as BERT and GPT that can be used for various tasks, including text classification, text generation, and natural language understanding.
- Pymorphy3 is the morphological analyzer (POS-tag + inflectional engine) for Ukrainian languages. It enables a variety of NLP tasks, such as part-of-speech identification, lemmatization, and word form generation. Pymorphy3 is a fork of the pymorphy2 library by Mykhailo Korobov and follows all the rules of the MIT license.
- The os library provides numerous functions that interact with the operating system. It allows one to perform operations related to the file system, such as changing the current working directory, extracting information about the execution environment, starting own processes, and more.
- The Python-docx library allows the creation, reading, and modification of Microsoft Word (.docx) documents using Python. It can help automate the tasks of processing text documents.
- The matplotlib.pyplot is a module of the Matplotlib library that provides a MATLAB-like interface for creating plots. It is widely used to visualize data through graphs, histograms, scatter plots, etc.

- The re library provides tools for working with regular expressions. It allows the performing of complex searches and manipulations of text, using patterns to define sequences of characters.

Therefore, the features of the method are the following:

- Stop words and punctuation marks are custom, meaning a manually created list;
- The data is divided into three parts and has the docx type.

When choosing programming tools such as IDEs or Python environments, it is essential to compare them to other available options to determine why a choice might be best for a given project. Here are some benchmarks for comparison:

- Speed and performance. VS Code is known for its high performance and speed, which makes it ideal for Python development, especially when it is needed to test and execute code quickly.
- Flexibility and customization. VS Code allows the customization of the environment to special needs with extensions, which can be helpful when working with NLP and data analysis.
- Community Support. VS Code has a large community of users and developers who create and support extensions, which can be helpful for troubleshooting and learning new features.
- Integration with other tools. VS Code easily integrates with tools and services, such as Git and Docker, and supports Jupyter Notebooks through extensions.
- Jupyter Notebooks allows you to execute code in parts and visualize data in a single document, which is ideal for experiments and research results.

Unlike other IDEs like PyCharm or Eclipse, VS Code and Jupyter Notebooks are more accessible and flexible, especially for scientific research and data analysis.

4. Experiments, results and discussion

As a result, the following files are included in the directory of the created software tool: files of works, files with a list of stop words, files with a list of punctuation marks, and files with the extension code .ipynb. Moreover, working files comprise collected works of Mykola Khvylovy.

These files are taken from the website <https://www.ukrlib.com.ua/books/author.php?id=20>. Moreover, the extension for files is .doc, and all the works provide text exclusively in Ukrainian. In the course of research, the dataset is created. Three subdirectories have been created in the leading directory, which correspond to the third period. The files are sorted into folders with names according to the periods of their writing:

1. Early works (1920-1925) - the beginning of Khvylovy's work. Early experiments with form and style and the search for one's voice in literature. The works of this period often reflect the optimism and idealism characteristic of a young artist who has not yet faced the harsh reality.
2. Mature works (1926-1930) - Khvylovy is already recognized in literary circles, and his works reflect greater self-confidence and a deeper understanding of social and political realities. It is also when he actively participates in public discussions and expresses his views on the development of Ukrainian culture.
3. Depressive works (1931-1933) - the last years of Khvylovy's life were marked by increasing repression and censorship. His works from this period often have a darker tone and reflect internal conflicts, disillusionment with ideals, and struggles with depression. It may be related to his suicide in 1933.

The stop word list file contains stop words collected by the developer to filter out meaningless words and combinations. The punctuation list file contains punctuation marks collected by the developer for filtering in works. The last file is the main one.ipynb. The file has the following structure: get data from files (part of the code responsible for importing the files mentioned above into the environment); prepare data (block for processing and preparing imported data for analysis); analysis (part of the code for data analysis); data visualization (visualization of analysis results); conclusion (a conclusion about the presence of suicidal intentions). before using the code, the following libraries are imported (Fig. 3):

```
import os
from docx import Document
import nltk
import pymorphy3
morph = pymorphy3.MorphAnalyzer(lang='uk')
from collections import Counter
import matplotlib.pyplot as plt
import re
```

Figure 3: Importing libraries for work

Next, there is the data block (Fig. 4-6):

```
def extract_words_by_row(file_path: str) -> str:
    stopwords = []
    if os.path.exists(file_path):
        with open(file_path, 'r', encoding='utf-8') as file:
            stopwords.extend([word.strip() for word in file.readlines()])
    return ' '.join(stopwords)
```

Figure 4: Use of function for getting words from files

```
# extract text from 1 file .docx
def extract_text_from_docx(file_path : str) -> str:
    doc = Document(file_path)
    full_text = []
    for para in doc.paragraphs:
        full_text.append(para.text)
    return ' '.join(full_text)

# extract all files from 1 folder (with using extract_text_from_docx)
def extract_text_from_directory(directory_path : str) -> str:
    full_text = []
    for filename in os.listdir(directory_path):
        if filename.endswith(".docx"):
            file_path = os.path.join(directory_path, filename)
            doc_text = extract_text_from_docx(file_path)
            full_text.append(doc_text)
    return ' '.join(full_text)
```

Figure 5: Functions to get a merged string of text from all files in a directory

```

# get text from folder as 1 str-object
raw_text_early = extract_text_from_directory('C:/Users/arsen/OneDr
raw_text_mature = extract_text_from_directory('C:/Users/arsen/OneDr
raw_text_depressive = extract_text_from_directory('C:/Users/arsen/

raw_text = [raw_text_early, raw_text_mature, raw_text_depressive]

# get stop-words and punctuation from custom files
stop_words = extract_words_by_row('ukrainian_stopwords')
punctuation = extract_words_by_row('punctuation')

```

Figure 6: Functions to import data

After importing data, analysis is needed (Fig. 7-9).

```

# def for tokenizing raw text | input - raw text; output - tokenized text
def tokenize(raw_text):
    tokens = nltk.word_tokenize(raw_text) # nltk method for tokenize
    return ' '.join(tokens)

# def for lemmatizing tokenized text | input - tokenized text; output - tokenized & lemmatized
def tokenize_and_lemmatize(token_text):
    lemmatized_tokens = [morph.parse(token)[0].normal_form for token in token_text] # pymorphy
    return ' '.join(lemmatized_tokens)

```

Figure 7: Declaration of functions for tokenization and lemmatization

```

# tokenize & lemmatize text
token_text_early = tokenize(raw_text_early).split()
token_text_mature = tokenize(raw_text_mature).split()
token_text_depressive = tokenize(raw_text_depressive).split()

# tokenize & lemmatize text
token_lemm_text_early = tokenize_and_lemmatize(token_text_early).split()
token_lemm_text_mature = tokenize_and_lemmatize(token_text_mature).split()
token_lemm_text_depressive = tokenize_and_lemmatize(token_text_depressive).split()

token_text = [token_text_depressive, token_text_mature, token_text_depressive]
token_lemm_text = [token_lemm_text_depressive, token_lemm_text_mature, token_lemm_text_depressive]
period_names = ['early', 'mature', 'depressive']

```

Figure 8: Using tokenization and lemmatization functions, as well as defining names for periods within the code

```

def delete_stop_words(stopwords : str, splitted_text : list) -> str:
    filtered_text = [word for word in splitted_text if word.lower() not in stopwords]
    return ' '.join(filtered_text)

def delete_punctuation(punctuation: str, text: str) -> str:
    return ''.join(char if char not in punctuation else ' ' for char in text).split()

def clean_text_from_stopwords_and_punctuation(stopwords : str, punctuation: str, splitted_text : list) -> list:
    no_stopwords = delete_stop_words(stopwords, splitted_text)
    return delete_punctuation(punctuation, no_stopwords)

```

Python

```

cleaned_early = clean_text_from_stopwords_and_punctuation(stop_words, punctuation, token_lemm_text_early)
cleaned_mature = clean_text_from_stopwords_and_punctuation(stop_words, punctuation, token_lemm_text_mature)
cleaned_depressive = clean_text_from_stopwords_and_punctuation(stop_words, punctuation, token_lemm_text_depressive)
cleaned_text = [cleaned_early, cleaned_mature, cleaned_depressive]

```

Figure 9: Declaration and application of functions for cleaning objects from stop words and punctuation marks

After the Prepare data block, there is an Analysis block (Fig.10-15):

```

len_early = len(cleaned_early)
len_mature = len(cleaned_mature)
len_depressive = len(cleaned_depressive)
len_words = [len_early, len_mature, len_depressive]

```

Figure 10: Calculation of the number of words in each period

```

word_freq_early = Counter(cleaned_early) # calc
word_freq_early = sorted(word_freq_early.items(), key=lambda x: x[1], reverse=True) # sort

word_freq_mature = Counter(cleaned_mature)
word_freq_mature = sorted(word_freq_mature.items(), key=lambda x: x[1], reverse=True)

word_freq_depressive = Counter(cleaned_depressive)
word_freq_depressive = sorted(word_freq_depressive.items(), key=lambda x: x[1], reverse=True)

word_freq = [word_freq_early, word_freq_mature, word_freq_depressive]

```

Figure 11: Calculation of the occurrence number of each word

```

# input - non-tokenized text; output - average sentence length
def calc_average_sentence_length(raw_text:str, period_name:str, print_out:bool):
    # text to s
    s = nltk.sent_tokenize(raw_text)
    # len of each s
    s_len = [len(sentence.split()) for sentence in s]
    # ave len
    ave_s_len = sum(s_len) / len(s_len)
    if print_out: print(f"In the {period_name} period, the average sentence length was \t {ave_s_len:.2f} words")
    return ave_s_len

```

Figure 12: Declaration of function to calculate the average sentence length

```

# def | input - tokenized text; output - average word length
def calculate_average_word_length(raw_text:str, period_name:str, print_out:bool):
    # text to s
    s = nltk.word_tokenize(raw_text)
    w_len = [len(w) for w in s]
    ave_w_len = sum(w_len) / len(w_len)
    if print_out: print(f"In the {period_name} period, the average word length was \t {ave_w_len:.2f} letters")
    return ave_w_len

```

Figure 13: Declaration of function to calculate the average word length

```
# http://www.senyk.poltava.ua/projects/ukr_stemming/ukr_endings.html
endings = ['я', 'ся', 'ня', 'ося', 'бся', 'ися', 'еся', 'шся', 'ася', 'вся', 'юся', 'ння',

# get a number of each endings
def count_endings(text, endings):
    endings_count = {ending: 0 for ending in endings}

    for ending in endings:
        pattern = re.escape(ending) + r'\b'
        endings_count[ending] = len(re.findall(pattern, text))

    return endings_count
```

Figure 14: Declaration of verb endings list and the function to count those endings into periods (regular expressions)

```
def count_verbs_by_endings(text, endings):
    verbs_count = 0
    max_ending_length = max(len(ending) for ending in endings)
    words = text.split()

    for word in words:
        if len(word) >= max_ending_length:
            ending_to_check = word[-max_ending_length:]
            if ending_to_check in endings:
                verbs_count += 1

    return verbs_count
```

Figure 15: Declaration of the function to calculate the number of verbs

Interim results and the following two blocks, along with visualization and conclusions, should be shown in the subsequent investigations. To view the main research document, the user needs to download and open the main.ipynb file. To run the code in the environment, it is necessary to open the project directory containing all the work files (listed at the beginning of the report). During the work, the average length of sentences of each period has been determined (Fig. 16). Next, the average word length of each period has been determined (Fig. 17). Based on the determined data, a graph of the ratio of the average length of sentences and periods has been constructed (Fig. 18a).

```
ave_sentence_len = []
for period, per_name in zip(raw_text, period_names):
    ave_sentence_len.append(calc_average_sentence_length(period, per_name, print_out=True))
```

In the early period, the average sentence length was	9.43 words
In the mature period, the average sentence length was	10.79 words
In the depressive period, the average sentence length was	12.22 words

Figure 16: Determination of the average length of sentences of different periods

```
ave_word_len = []
for period, per_name in zip(raw_text, period_names):
    ave_word_len.append(average_word_length(period, per_name, print_out=True))
```

In the early period, the average word length was	4.05 letters
In the mature period, the average word length was	4.11 letters
In the depressive period, the average word length was	4.30 letters

Figure 17: Determination of the average length of words in different periods

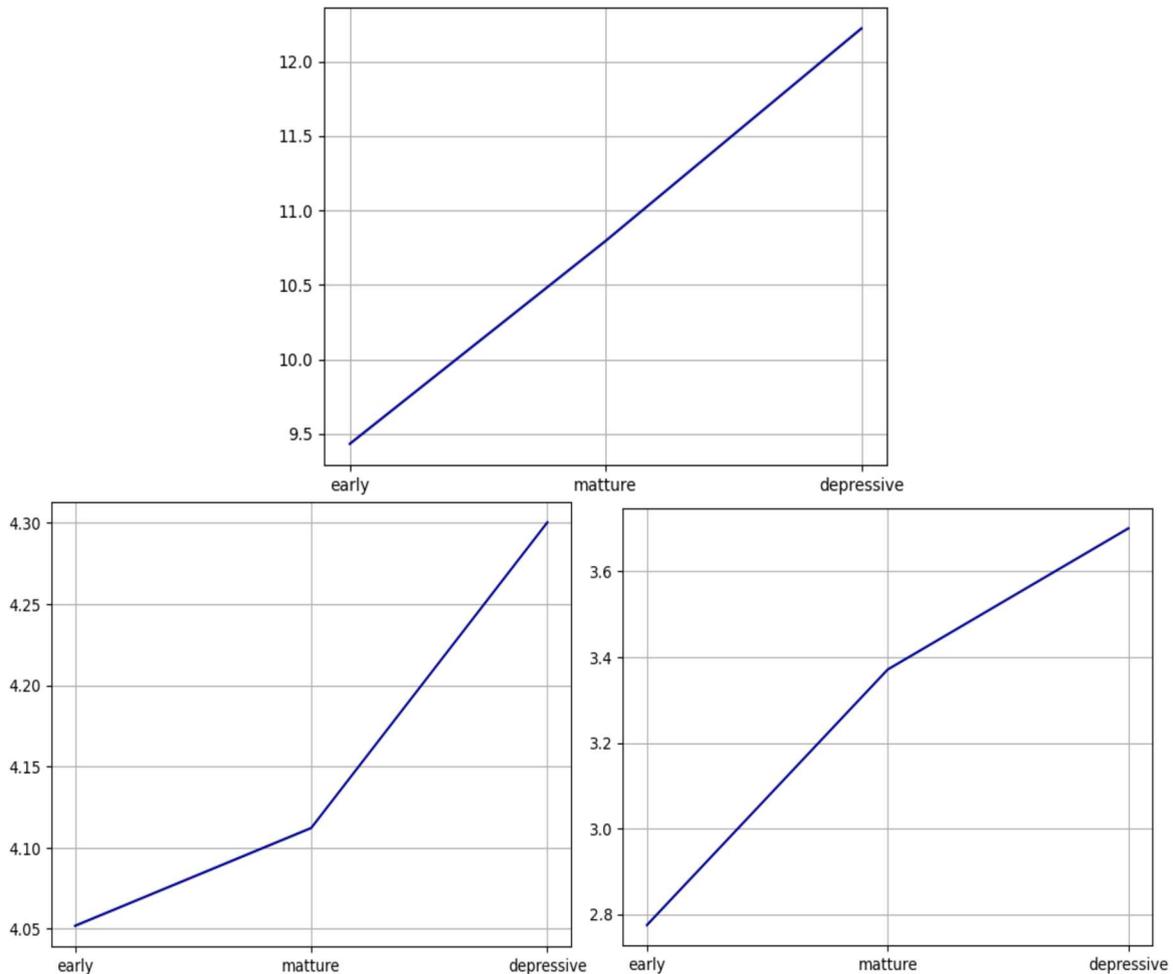


Figure 18: Graph of the ratio of the average length of sentences, average length of words, passivity percentage and periods

The graph of the ratio of the average length of words and periods is also developed and shown in Fig. 18b. Next, the number of words in each period has been calculated, and the number of verbs has been found and counted. The percentage of passivity/activity has been entered as the frequency of verb use according to the number of words. The result of determining the rate of passivity for each period is shown in Fig. 18c. The reshaped dataset needs to be run through the system. The results can be seen below in Fig. 19-21. Also, due to the imbalance of the number of literary works in 3 periods, it is decided to reform the dataset into 2 periods of the same number of literary works. There is a need to run the reshaped dataset through the system.

```

ave_sentence_len = []
for period, per_name in zip(raw_text, period_names):
    ave_sentence_len.append(calc_average_sentence_length(period, per_name, print_out=True))
✓ 1.4s
In the period 1 period, the average sentence length was      10.04 words
In the period 2 period, the average sentence length was      11.11 words

```

Figure 19: Determination of the average length of sentences of different periods

```

ave_word_len = []
for period, per_name in zip(raw_text, period_names):
    ave_word_len.append(calculate_average_word_length(period, per_name, print_out=True))
✓ 4.4s
In the period 1 period, the average word length was      4.09 letters
In the period 2 period, the average word length was      4.15 letters

```

Figure 20: Determination of the average length of words in different periods

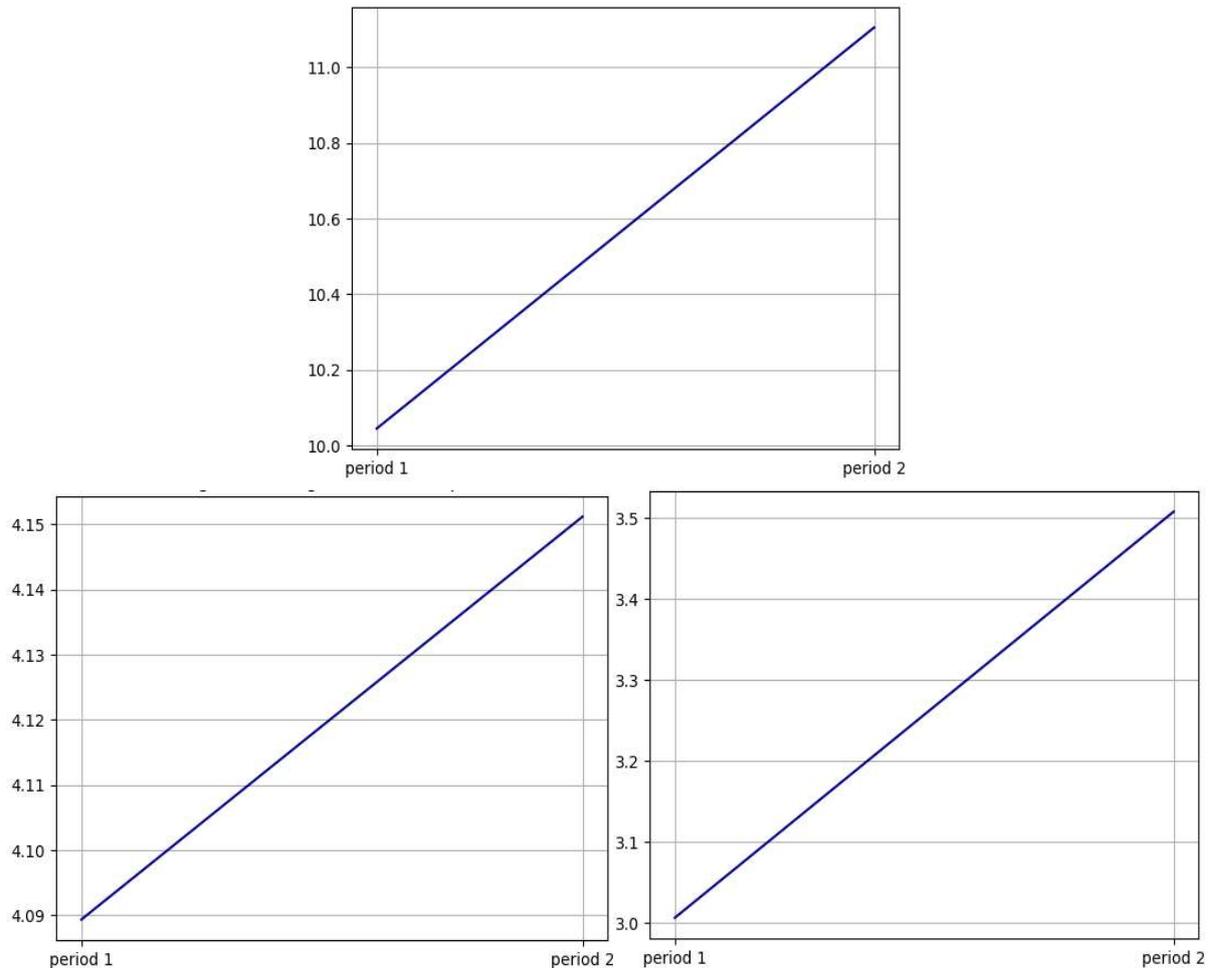


Figure 21: Graph of the ratio of the average length of sentences, average length of words, passivity percentage and periods

The interim conclusion on the balanced dataset of Khvylovy's works shows that all indicators have been increased the same way as on the unbalanced dataset. The software has been developed to analyze specific writers such as Mykola Khvylovy. However, testing has been performed on the works of Valerian Pidmogylny and Ivan Bahryany. So, periods similar to those of Khvylovy have been taken. But there is also a nuance here: for Khvylovy, these were periods divided explicitly according to the theme of his literary work; for Pidmohylny, it was only a selection of works from these periods; and for Bahryany, these were conditionally divided periods of creativity of his years. During the testing, two datasets were created, each containing 3 periods. The program execution result is shown below (Fig. 22-24):

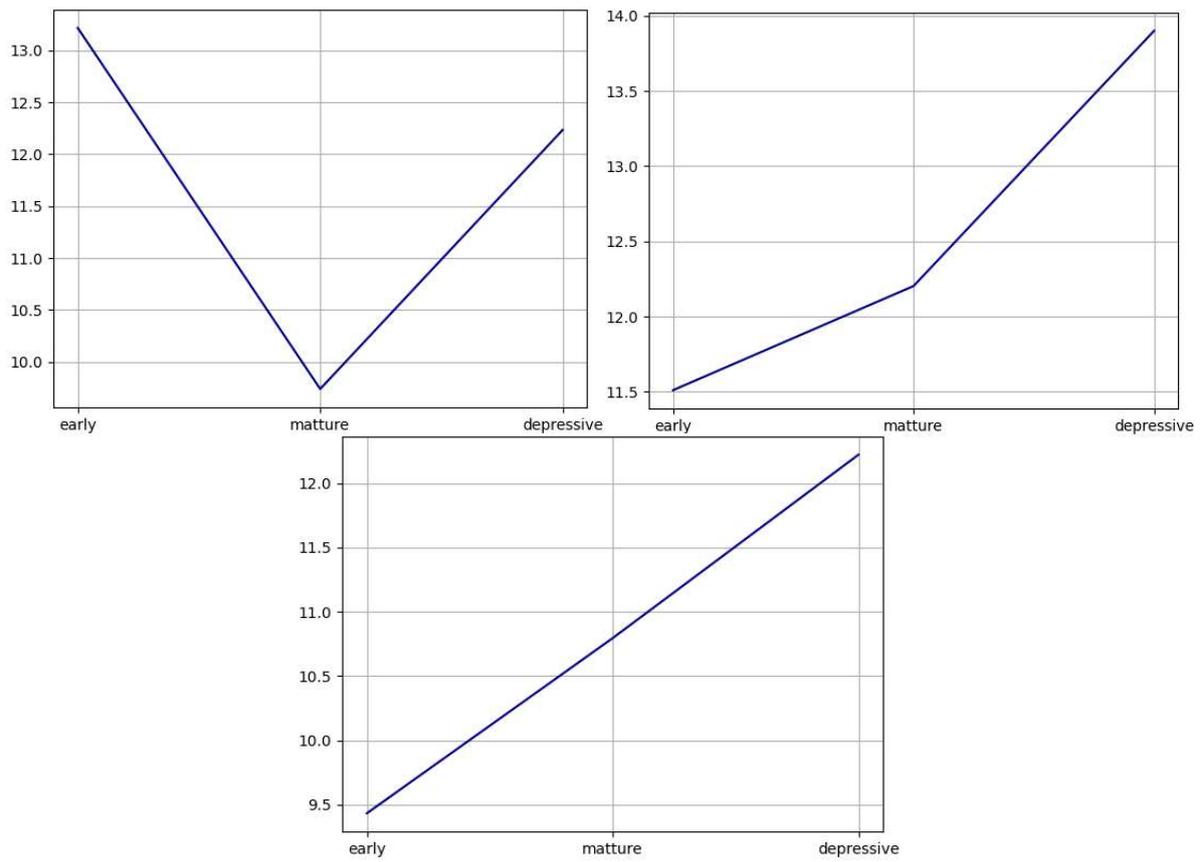


Figure 22: The sentences and periods average length ratio (Pidmogylny, Bahryany and Khvylovy)

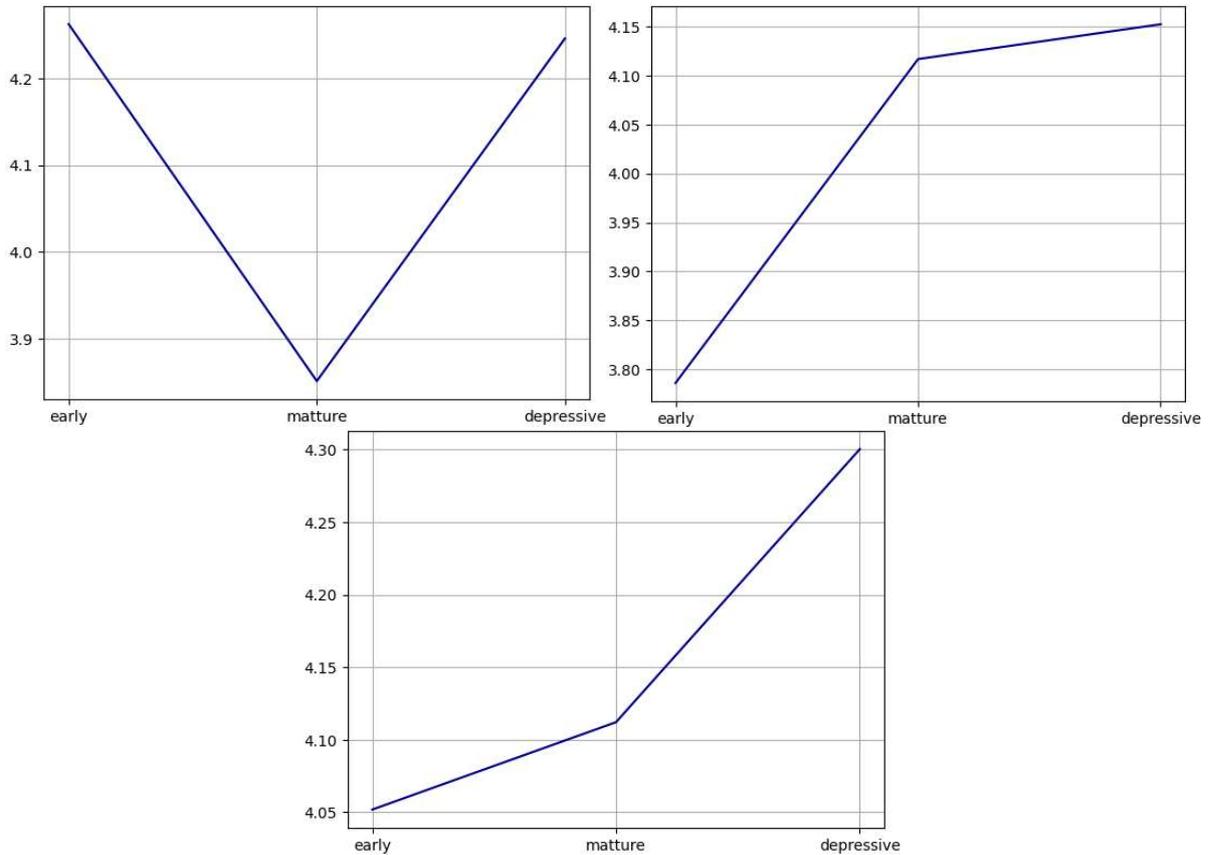


Figure 23: The words and periods average length ratio (Pidmogylny, Bahryany and Khvylovy)

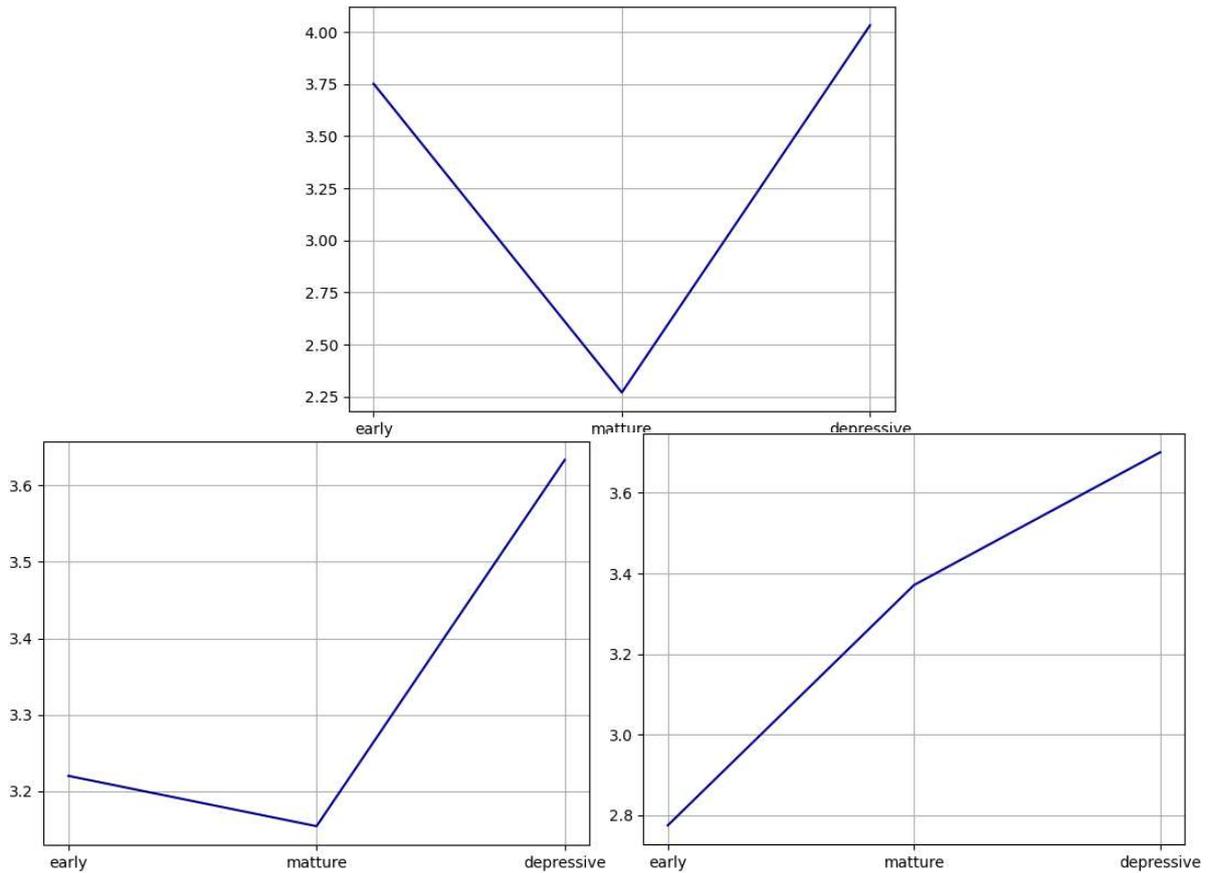


Figure 24: The ratio of passivity percentage and periods (Pidmogylny, Bahryany and Khvylovy)

Table 1

Analysis conclusion

Metrics	Writer	Value	Conclusion	Discussion
Length of sentences	Pidmogylny	13.25→	In mature years, the value has decreased significantly	Khvylovy has the shortest sentences. Instability is revealed only in Podmohylny's works.
		9.25→		
		12.25		
	Bahryany	11.5→	Steady growth in value	
		12.25→		
		13.9		
Khvylovy	9.5→	Stable growth		
	10.75→			
	12.75			
Length of words	Pidmogylny	4.28→	The sharp decrease in the value in the second period	Khvylovy writes the longest words. Stable growth is maintained in Khvylovy's and Bahryany's works; Pidmogylny's work shows a sharp increase in value in the second period.
		3.845→		
		4.25		
	Bahryany	3.78→	Relatively sharp growth until the second period	
		4.12→		
		4.152		
	Khvylovy	4.052→	Almost the same growth, but more between 2-3 periods	
		4.115→		
		4.3		

Passivity rate	Pidmogylny	3.75→ 2.25→ 4.05	Decreased percentage of passivity in the second period by almost 2 times.	Writers who did not commit suicide have a decrease in passivity in the second period. In Khvylovy's works, stable growth of passivity is found throughout his life.
	Bahryany	3.22→ 3.15→ 3.65	Also, a decrease in the percentage of passivity in the second period, and then a sharp increase.	
	Khvylovy	2.77→ 3.38→ 3.7	The growth of passivity throughout the entire literary work activity	

Also, additional research has been conducted on stabilized datasets (performed by the same number of 10 files for 2 periods). The result of determining the average length of sentences for three writers (order: Khvylovy, Bahryany, Pidmogylny; from left to right) is shown in Fig. 25. The result of determining the average length of words is in Fig. 26. The result of determining the percentage of passivity in Fig. 27.

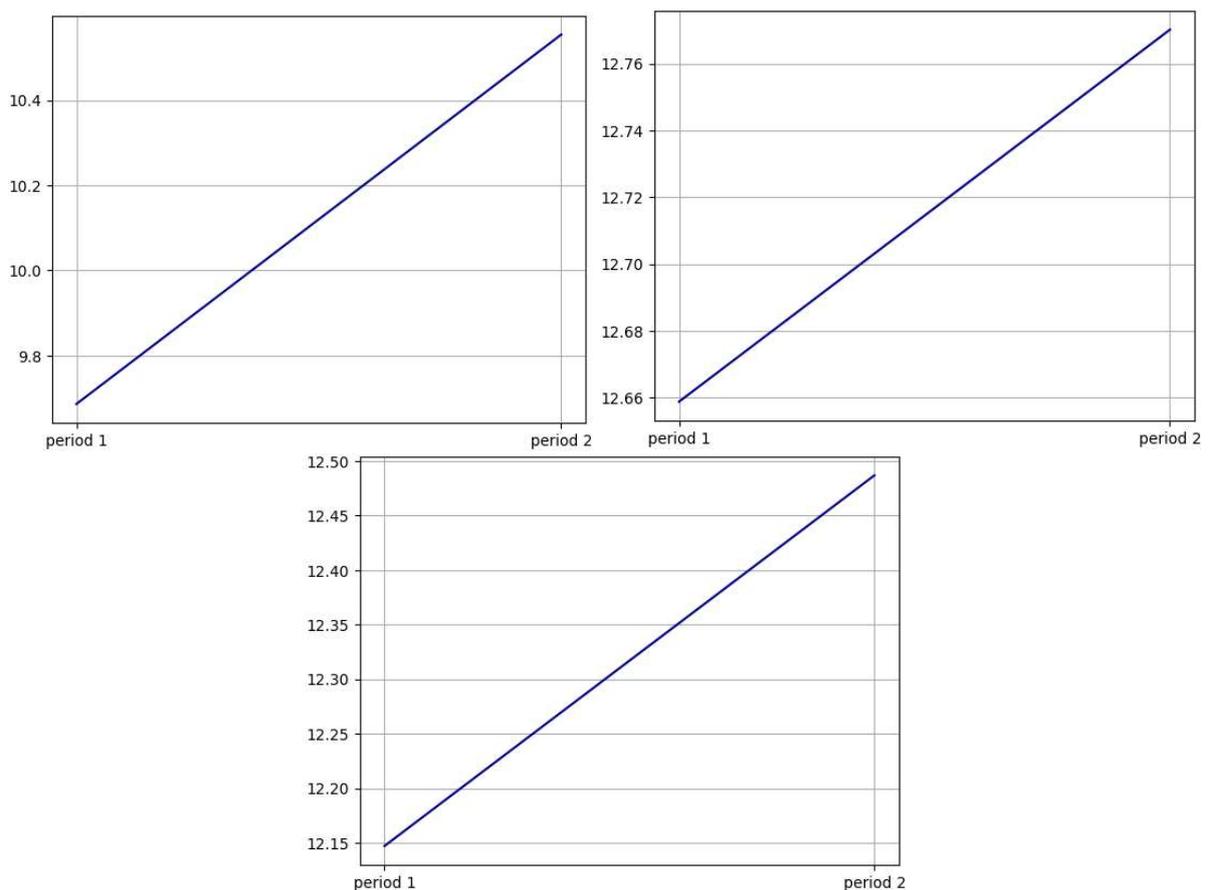


Figure 25: Graph of the ratio of the average length of sentences and periods

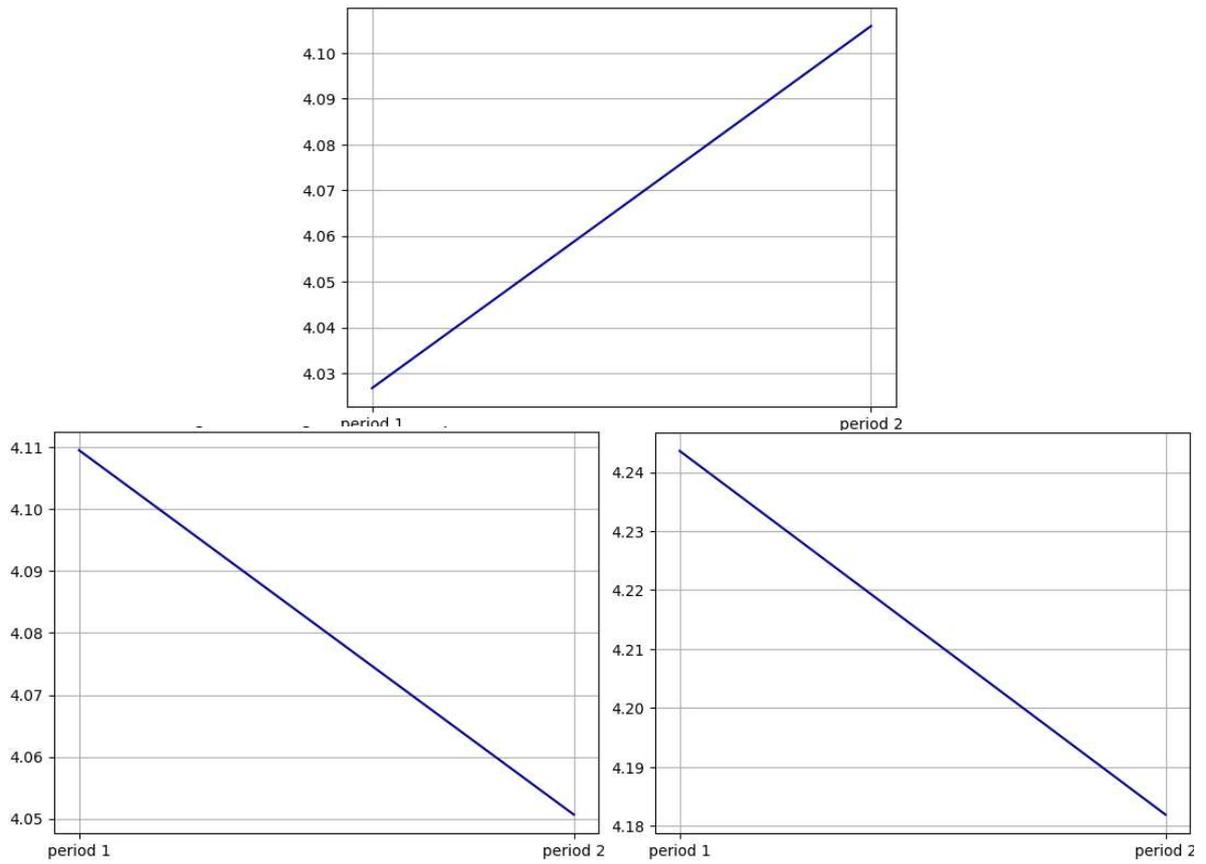


Figure 26: Graph of the ratio of the average length of words and periods

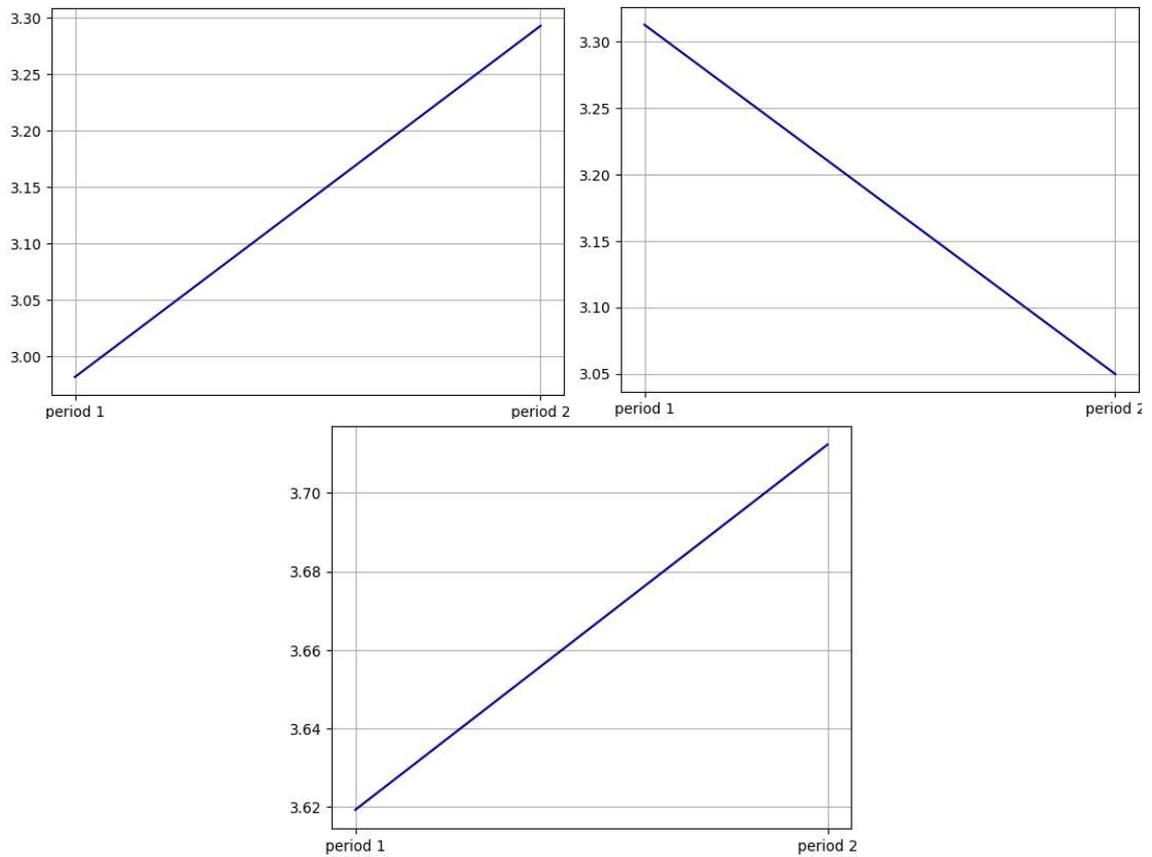


Figure 27: Graph of the ratio of passivity percentage and periods

Interim conclusions based on balanced data comprise the fact that the average sentence length has been increasing steadily in all 3 cases, the average word length has been growing only in Khvylovy's works, and the percentage of passivity has increased the most in Khvylovy's works, next in Pidmogylny. The positive decline in this percentage has been in Bahryany's works.

5. Conclusion

So, the dataset of Mykola Khvylovy's works has been collected during the research. Also, the list of stop words and punctuation marks has been defined. As a result, software has been developed that analyses Khvylovy's works of 3 periods. Based on the implementation of the control example, taking into consideration that the length of words and sentences in Khvylovy's works has been increasing gradually with slight acceleration according to different periods facing decreasing in Khvylovy's activity, the following assumptions have been made:

1. Khvylovy could express his thoughts more voluminously, as expressed by the increased text volume. At the same time, his activity expressed using verbs decreased, which means he had less life energy.
2. Khvylovy could develop himself as a writer, making it possible to compose more complicated texts. The decrease in activity can be due to age or the use of other literature constructions instead of verbs, such as more detailed descriptions of objects, people, and locations.

The software has prospects both for improvement and application. It is possible to improve the software using the following methods: determining the mood of the works, detecting suicidal words, verifying statistical values on different types of works, and using code on different datasets by other writers to find patterns. It is possible to apply the software even on small datasets of modern writers to detect changes in the statistical values of their writing.

Acknowledgement

The research was carried out with the grant support of the National Research Fund of Ukraine "Information system development for automatic detection of misinformation sources and inauthentic behaviour of chat users", project registration number 187/0012 from 1/08/2024 (2023.04/0012).

References

- [1] Recommendations for the evaluation and management of patients at risk of suicide, American College of Physicians, 2019. URL: https://www.dokazovo.in.ua/wp-content/uploads/2021/03/rekomendaciji_z_ocinki_ta_vedennya_pacientiv_z_rizikom_sujicidu.pdf.
- [2] V. V. Rybalka, Psychological prevention of suicidal tendencies in students, 2007. URL: http://poippo.pl.ua/files/pidrozdzily/CPPSR/NP_DOC/Syitcud_2007.pdf.
- [3] A. A. Tkachenko, Manual of Forensic Psychiatry, 2015. URL: https://stud.com.ua/40956/psihologiya/diagnostika_suyitsidalnih_namiriv.
- [4] O. M. Gurova, Methodical materials for curators of academic groups regarding the recognition of suicidal thoughts and their effective actions at the stages of identifying and preventing destructive forms of behavior among student youth, 2024. URL: https://cusu.edu.ua/images/psihologicna-služba/metod_recom_suicid.pdf.
- [5] A. R. Ivats, O.P. Romanov, B.Ya. Nad, Socio-psychological factors and risk factors of suicides among young people, 2024, URL: <https://dspace.uzhnu.edu.ua/jspui/bitstream/lib/21525/1/28-29.pdf>.
- [6] M. Bublyk, V. Lytvyn, V. Vysotska, L. Chyrun, Y. Matseliukh, N. Sokulska, The Decision Tree Usage for the Results Analysis of the Psychophysiological Testing, CEUR workshop proceedings, Vol-2753 (2020) 458-472.

- [7] O. Oborska, V. Andrunyk, L. Chyrun, R. Hasko, A. Vysotskyi, S. Mushasta, O. Petruchenko, I. Shakleina, The Intelligent System Development for Psychological Analysis of the Person's Condition, CEUR Workshop Proceedings, Vol-2870 (2021) 1390-1419.
- [8] O. Veres, O. Oborska, A. Vasyliuk, Y. Brezmen, I. Rishnyak, Problems and peculiarities of the IT project management of ontological engineering for person psychological state diagnosing. In: CEUR Workshop Proceedings, Vol-2565 (2020) 162–177.
- [9] V. Lytvyn, V. Vysotska, A. Rzhеuskyi, Technology for the Psychological Portraits Formation of Social Networks Users for the IT Specialists Recruitment Based on Big Five, NLP and Big Data Analysis, CEUR Workshop Proceedings, Vol-2392 (2019) 147-171.
- [10] L. Chyrun, V. Vysotska, I. Kis, L. Chyrun, Content Analysis Method for Cut Formation of Human Psychological State, in: Proceedings of the International Conference on Data Stream Mining and Processing, DSMP, 2018, pp. 139-144. doi: 10.1109/DSMP.2018.8478619
- [11] A. Dyriv, V. Andrunyk, Y. Burov, I. Karpov, L. Chyrun, The user's psychological state identification based on Big Data analysis for person's electronic diary, in: proceedings of the 16th International conference on computer science and information technologies on Computer science and information technologies, pp. 101–112, 2021.
- [12] L. Chyrun, I. Kis, V. Vysotska, L. Chyrun, Content monitoring method for cut formation of person psychological state in social scoring, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 106-112. DOI: 10.1109/STC-CSIT.2018.8526624
- [13] A. Abdulsalam, A. Alhothali, S. Al-Ghamdi, Detecting suicidality in Arabic Tweets using machine learning and deep learning techniques, Arabian Journal for Science and Engineering, 1-14, 2024.
- [14] E. Pranckeviciene, J. Kasperiuoniene, Global Suicide Mortality Rates (2000–2019): Clustering, Themes, and Causes Analyzed through Machine Learning and Bibliographic Data, International Journal of Environmental Research and Public Health 21(9) (2024) 1202.
- [15] M. A. Allayla, S. Ayyaz, A Big Data Analytics System for Predicting Suicidal Ideation in Real-Time Based on Social Media Streaming Data, arXiv preprint arXiv:2404.12394, 2024.
- [16] M. Ghonge, T. Kachare, S. Kakade, S. Shintre, S. Nigade, Machine Learning and Web Integrated Chatting Forum Which Detected Mental Health of the User, in Proceedings of the International Conference on Smart Technologies in Urban Engineering, pp. 96-106, 2022.
- [17] V. Slobodzian, M. Molchanova, O. Kovalchuk, O. Sobko, O. Mazurets, O. Barmak, I. Krak, An Approach Based on the Visualization Model for the Ukrainian Web Content Classification, in Proceedings of the 12th International Conference on Advanced Computer Information Technologies (ACIT), pp. 400-405, 2022.
- [18] K. Cosic, V. Kopilas, T. Jovanovic, War, emotions, mental health, and artificial intelligence, Frontiers in psychology 15 (2024) 1394045.
- [19] M. Hassib, N. Hossam, J. Sameh, M. Toriki, AraDepSu: Detecting depression and suicidal ideation in Arabic tweets using transformers, in Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP). pp. 302-311, 2022.
- [20] S. Albota, et. al., Linguistic traces of psychological manipulations in discussions of wikipedia talk pages, CEUR Workshop Proceedings, Vol-2386 (2019) 183–193.
- [21] S. Albota, Linguistic and Psychological Features of the Reddit News Post, in Proceedings of the 15th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2020, 1, pp. 295–299.
- [22] S. Albota, et. al., Discussions of Wikipedia Talk Pages: manipulations detected by lingual-psychological analysis, CEUR Workshop Proceedings, Vol-2392 (2019) 309–320.
- [23] A. Morushko, et. al., Determining the psychological portrait of members of web communities through socionic analysis, CEUR Workshop Proceedings 2616 (2020) 112–124.
- [24] V. Pasichnyk, et. al., The model of data analysis of the psychophysiological survey results, Advances in Intelligent Systems and Computing 512 (2017) 271-281.

- [25] V. Vasyliuk, et. al., Information System of Psycholinguistic Text Analysis, CEUR workshop proceedings, Vol-2604 (2020) 178-188.
- [26] S. Fedushko, M. Davidekova, Analytical service for processing behavioral, psychological and communicative features in the online communication, The International Workshop on Digitalization and Servitization within Factory-Free Economy 160 (2019) 509-514.
- [27] S. Fedushko, M. ml. Gregus, T. Ustyianovych, Medical card data imputation and patient psychological and behavioral profile construction, in The 9th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare 160 (2019) 354-361.
- [28] I. Khomytska, V. Teslyuk, N. Kryvinska, I. Bazylevych, Software-based approach towards automated authorship acknowledgement-chi-square test on one consonant group, Electronics (Switzerland) 9(7) (2020) 1–11.
- [29] A. R. Sydor, V. M. Teslyuk, P. Y. Denysyuk, Recurrent expressions for reliability indicators of compound electropower systems, Technical Electrodynamics 4 (2014) 47–49.
- [30] V. Motyka, V. Vysotska, L. Chyrun, O. Markiv, S. Chyrun, L. Kolyasa, System Project for Ukrainian-language Feedback Tonality Analysis in the Health Care Field Based on BERT Model, in: Proceedings of the 18th International Conference on Computer Sciences and Information Technologies, CSIT, Lviv, 19-21 October 2023.
- [31] N. Shakhovska, K. Shakhovska, The Method of Text Tonality Classification, in: International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2020 - Proceedings, 2020, 1, pp. 19–23.
- [32] R. Peleshchak, V. Lytvyn, I. Peleshchak, A. Khudyy, Z. Rybchak, S. Mushasta, Text Tonality Classification Using a Hybrid Convolutional Neural Network with Parallel and Sequential Connections Between Layers, CEUR Workshop Proceedings, Vol-3171 (2022) 904-915.
- [33] S. Voloshyn, V. Vysotska, O. Markiv, I. Dyyak, I. Budz and V. Schuchmann, Sentiment Analysis Technology of English Newspapers Quotes Based on Neural Network as Public Opinion Influences Identification Tool, in: Proceedings of the IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), 2022, pp. 83-88, doi: 10.1109/CSIT56902.2022.10000627.
- [34] V. Vysotska, O. Markiv, S. Tchynetskyi, B. Polishchuk, O. Bratasyuk, V. Panasyuk, Sentiment Analysis of Information Space as Feedback of Target Audience for Regional E-Business Support in Ukraine, CEUR Workshop Proceedings, Vol-3426 (2023) 488-513.
- [35] A. Sachenko, T. Lendiuk, K. Lipianina-Honcharenko, M. Dobrowolski, G. Boguta, L. Bytsyura, Method of Determining the Text Sentiment by Thematic Rubrics, CEUR Workshop Proceedings 3688 (2024) 404-414.
- [36] R. Nazarchuk, S. Albota, Tweets about Ukraine during the russian-Ukrainian War: Quantitative Characteristics and Sentiment Analysis, CEUR Workshop Proceedings, Vol-3426 (2023) 551-560.