Text Data Vectorization Model of Ukrainian-Language Internet Communication Content

Vitalii Slobodzian[1], Oleksii Kovalchuk[1], Maryna Molchanova[1], Olena Sobko[1], Olexander Mazurets [1], Olexander Barmak [1], Iurii Krak [2,3].

[1] *Khmelnytskyi National University, Khmelnytskyi, 11, Institutes str., 29016, Ukraine*
[2] *Taras Shevchenko National University of Kyiv, Kyiv, 64/13, Volodymyrska str., 01601, Ukraine*
[3] *Glushkov Cybernetics Institute, Kyiv, 40, Glushkov ave., 03187, Ukraine*

Abstract

This paper proposes a model for analyzing the content of the modern Ukrainian everyday language. The model is built by combining the known and significant, for meaning transfer, frequency dictionaries, which cumulatively cover spheres of activities and types of content. Studies established that the best separability, when classifying texts using the proposed model, is observed at a word vector length of 1500 units; it determines the model's optimal dimension. According to the results of the test classification of more than 400 texts, the Text Rank keyword search method has been established as the most suitable for work with the proposed model of the modern Ukrainian everyday language. To compare the effectiveness of different keyword search methods, it was used the method of visual analytics MDS, which is considered an effective and sufficient tool for visual verification of the results of classification of digital texts into different categories, including both thematic classes categories of emotionally charged texts. The conducted research provides opportunities of using the created model of modern Ukrainian everyday language to solve the text analysis issues and their classification by various grounds. It is essential to prevent suicidal tendencies, detect bullying on social networks, determine the negative emotional charged texts, and warn users about potentially harmful content.

Keywords

text data vectorization, web-content classification, Text Rank, YAKE, TF-IDF, dispersion evaluation, MDS, corpus of the everyday Ukrainian language, Internet communication.

1. Introduction

People have a deep and universal need to interact with others, and the greater their communicative ability, the more satisfying and rewarding will be their lives [1].

Nowadays, Internet communication has a prominent place in the means of nonverbal communication, with more than 4 billion users. Due to the COVID-19 pandemic, social media audiences were significantly increased. On average, Ukrainians spend more than 2 hours a day on social networks, forming a voluminous amount of information available to many users. This is comparable to public address because, on average, one post may be read by a large number of people. Also, recently, alternative communication technologies are being developed, which attract people with disabilities to communicate [2], expanding the audience and volume of communication.

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland.
EMAIL: vitalii.slobodzian@gmail.com (V. Slobodzian); losha.kovalchyk1998@gmail.com (O. Kovalchuk); momolchanova@gmail.com (M. Molchanova); olenasobko.ua@gmail.com (O. Sobko); exe.chong@gmail.com (O. Mazurets); alexander.barmak@gmail.com (O. Barmak); yuri.krak@gmail.com (I. Krak);
ORCID: 0000-0001-8897-0869 (V. Slobodzian); 0000-0001-9828-0941 (O. Kovalchuk); 0000-0001-9810-936X (M. Molchanova); 0000-0001-5371-5788 (O. Sobko); 0000-0002-8900-0650 (O. Mazurets); 0000-0003-0739-9678 (O. Barmak); 0000-0002-8043-0785 (I. Krak);

CEUR Workshop Proceedings (CEUR-WS.org)

Internet content is not only a means of communication but also an critical factor of influence to people because it can cause various collective reactions, create fear, panic, especially in the context of the spread of Covid-19 [3]. Also, internet content has a significant impact on adolescents' risky behaviors. The study [4] shows that adolescents with friends who drink alcohol and promote such behavior on social networks have a higher risk of alcohol consumption than adolescents whose environment has not published photos and posts with similar content. Also, we must not forget the recent regrettable incidence of suicides of teenagers worldwide from the so-called game "Blue Whale Challenge," which is closely associated with depression, bullying, self-mutilation [5]. Therefore, classifying Internet content to prevent socially dangerous manifestations is an essential issue of computational linguistics. Using such a classification will help prevent suicidal behavior, detect bullying on social networks, identify negative emotional charged texts and warn users about possible harmful content.

This research aims to develop a model of the everyday Ukrainian language, within which a hyperplane classification of Internet content will be possible for the issue of preventing socially dangerous manifestations that occur when communicating on the Internet.

2. Related work

We will review recent publications that correlate with the considered problem in one way or another. Reviewed publications present approaches that correlate with the approaches of this study. The authors plan to use a vector model to model the Ukrainian-language segment of Internet communication, which will be based on statistical measures used to assess the importance of the word in the context of the message, which in turn is part of the collection of messages or corpus (TF-IDF, Dispersion evaluation). On the other hand, it is necessary to consider works that offer tasks related to preventing socially dangerous manifestations that occur when communicating on the Internet.

In recent years, the recognition of abuse on social networking platforms has been an active research issue. In non-native English-speaking countries, social media texts are primarily mixed. The study [6] presents experiments using several machine learning models, deep learning, and transfer learning to detect offensive content on Twitter. The experiment results showed that the features of TF-IDF are more suitable for this task. In study [7], it was compared the usage of TF-IDF, n-grams, and pre-trained MuRIL. The issue is to identify the offensive content from YouTube's mixed-comment set. The TF-IDF showed the best results for two languages out of three. In study [8], the research focused on identifying offensive content in Tanglish, Manglish, and Malayalam languages, using four classifiers (SVM, Random Forest, k-nearest neighbors, and Naive Bayes). The proposed model achieved an accuracy of 76.96% when using a linear SVM with the TF-IDF feature presentation technique.

In study [9], it was conducted experiments in different languages, which show that images complement natural language processing models (including BERT), taught without prior external training. Text classification studies were focused on Wikipedia articles, as images usually are complemented with text, and Wikipedia pages can be written in different languages. In study [10], it was researched online conversations: their course, arguments, and how they are resolved. The main emphasis in the work is on the identification of "sarcasm." The authors show that identifying sarcasm in the message helps to understand whether the author of this message agrees or disagrees with the statement under discussion. The experiment was performed based on functions, using the logistic regression model from Scikit-learn. In study [11], was considered the impact and effective measures to counteraction the spread of disinformation in the context of the COVID-19 pandemic. The research was conducted among Twitter users (analysis of posts). The study explains how to use the BERT-based model to match facts to tweets and identify misinformation.

In study [12], was carried out the allocation of emotional moods and classification of their polarity. The publication conducted experiments with 8 data sets in English. The study shows the potential importance of annotating phrases for small data sets about emotional moods. At the same time, the results show that the performance of modern models for the prediction of polar expressions of language is poor, which prevents the use of this information in practice. The authors [13] proposed using emojis to represent abusive words to reveal abusive meaning in social networks because emojis are extralinguistic information. This approach does not depend on manual annotation and does not require expensive resources to download (for example, WordNet). In many experiments, the authors used BERTLARGE as a basis for the most modern text classification to detect offensive posts and the BERT model.

Social networks face a severe issue of forming rumors and fake news. This is due to their internal nature of connecting millions of users to millions of others without proper identification. Therefore, the authors of [14] propose automated detection of rumors by the method of constructing semantic opposites. The Glove was used to embed and initialize word vectors and construct opposites.

Therefore, considering the research, this issue is relevant for the Ukrainian everyday language. It is also prospective to use a vector model to achieve this goal, which will be characterized by statistical measures used to find the importance of the word in the content of the message, which in turn is part of the message collection or corpus (TF-IDF, Dispersion evaluation, etc.).

3. Methods and materials

The Ukrainian language used for communication on the Internet is mixed with other languages. This mixed language is named "surzhyk." There are no statistically significant corpora for this language. This fact leads to constraints in using standard approaches in vectorization of language and searching for features of a text that can be used for further processing by machine learning.

In this study, it is proposed to use the variant of the bag of words (BOW) model [15, 16] to vectorize the text. In this variant, the list of feature words is selected from various frequency dictionaries that are currently available [17-21]. The set of words is generated by discarding stop words and words that are not often used from the frequency dictionaries. The threshold of such discarding is defined experimentally. Different frequency dictionaries are combined to get a more complete list of words.

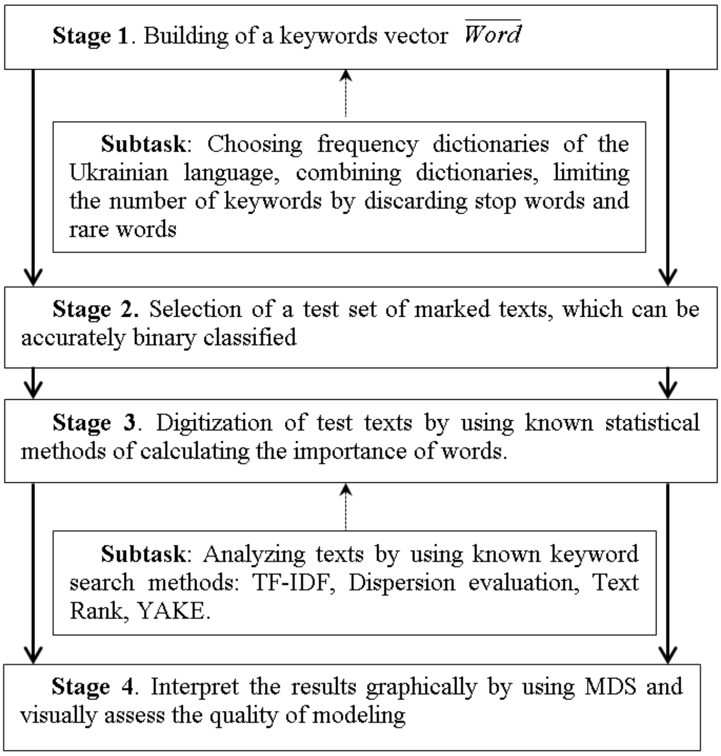

Stage 1. Building of a keywords vector $\overline{Word}$
Subtask: Choosing frequency dictionaries of the Ukrainian language, combining dictionaries, limiting the number of keywords by discarding stop words and rare words
Stage 2. Selection of a test set of marked texts, which can be accurately binary classified
Stage 3. Digitization of test texts by using known statistical methods of calculating the importance of words.
Subtask: Analyzing texts by using known keyword search methods: TF-IDF, Dispersion evaluation, Text Rank, YAKE.
Stage 4. Interpret the results graphically by using MDS and visually assess the quality of modeling

Figure 1: The scheme of the approach for the creation of the model of the Ukrainian language in the tasks for the analysis of communication on the Internet

Statistical measures are considered in the work for digitizing texts (determination of numerical values for words included in BOW). Statistical measures are used for assessing the importance of words in document context, which is a part of the document collection or corpus: TF-IDF [22], dispersion evaluation [23] etc.

Visual analytics methods are used to verify the obtained sets of words and numerical values (measures) of vectors that represent texts in the ability to classify texts. The Multidimensional scaling (MDS) [24, 25] method is used in this work. Visual assessment of the ability of the proposed implementation of the model allows assessing the quality of results.

3.1. The description of the proposed approach

The scheme of the proposed approach for building a model for the content analysis of the Ukrainian-language segment of Internet communication is shown in Figure 1.

Next, consider each of the stages in detail.

3.2. Building of a vector of keywords for the common lexicon of the Ukrainian-language segment of the Internet

The first stage of the approach is to build a vector of keywords for the transmission of meaning in Internet communication. Building a vector of keywords is a separate subtask because of the mentioned feature of language (using "surzhyk," distorted words, and profanity). The vector should not include only keywords from papers. It should be a balanced data set because we need to analyze the inflected Ukrainian language. That is why frequency dictionaries of the Ukrainian language were used in works [17-21] ($\overline{W_j} = \{\overline{w_1}, \overline{w_2}, \ldots, \overline{w_n}\}, j = \overline{1..n}$), where j – serial number of the dictionary, n – number of dictionaries. Every dictionary is a set of words $\overline{w} = \{word1, word2, \ldots, wordn\}, i = \overline{1..n}$, where n – number of words in the dictionary. In the study, words were removed from each frequency dictionary that would not significantly affect the model in the authors' opinion (Words are inherent in a vast number of texts (stop words that have official meaning and are used to connect words in the text), infrequent words, etc.). As a result, $\overline{W_j}$ contains only selected words. The keyword vector $\overline{Word}$ will be a combination of such frequency dictionaries:

$$\overline{Word} = \cup_{i=1}^{n} \overline{W_i}. \qquad (1)$$

The second stage is the selection of texts that are subject to ideal classification when it is unambiguously possible to conclude that the selected text belongs to a specific category. A set of texts *D* is formed where each text $d \in D$ can correspond to a specific category $c \in C$, where *C* is a set of categories. In this case, binary classification will be used, as our goal is to identify texts with negative color content as a category.

In the third stage, each text document is vectorized by known keyword search methods.

At the preprocessing stage, each text document $d_i \in D$, is converted into a word vector, here *i* – is the number of documents in the collection. After that, a corresponding vector of estimates of occurrences of keywords $\overline{Wd_i}$, which are in the $\overline{Word}$ vector, is formed by using each of the keyword search methods.

TF-IDF [22], Dispersion evaluation [23], Text Rank [26, 27, 28], YAKE [29] keyword searching methods are proposed to use for forming estimates.

Classical *TF-IDF* has the following formula:

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,d,D) \qquad (2)$$

where *IDF(t,d,D)* is the weight of the term *t* of document *d* of the corpus *D*, and *TF*(*t*,*d*) is the value of the frequency of the term *t* in document *d*.

Dispersion evaluation is presented as an evaluation of the importance of each word in the analyzed text by using the dispersion evaluation method. This method evaluates the discriminant force of words and allows getting words that are evenly distributed from the general set of commonly used words in the text [23]. According to [23] if a word *A* in a text consisting of *N* words is denoted as A_k^n, where the index k is the number of occurrences of the word in the test, and n is its position in the text then the interval between sequential occurrences of the word is calculated by the following formula:

$$\varDelta A_k^m = A_{k+1}^m - A_k^n = m - n, \qquad (3)$$

where m is iteration and n is a position of A word that met $k+1$ and k time in the text. So, the dispersion evaluation is calculated by the formula:

$$\sigma = \sqrt{(\Delta A2\ \ ) - (\Delta A)^2} \Big/ _{(\Delta A)}, \qquad (4)$$

where (ΔA) is the average value of sequence ΔA_1, ΔA_2,.., ΔA_k . K is the number of occurrences of the word A.

The Text Rank method is intended for modeling text as an undirected weighted graph $G=(V,E,W)$. In this method, the keyword candidates are vertices V of the graph, and the relationship between two words is considered as an edge E. W represents the frequency of occurrence in relation to E [26, 27]. The following formula is used for iterative calculation of weights of vertices:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in ln(V_i)} \frac{w_{i,j}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j), \qquad (5)$$

where d is damping factor (default value is 0.85) [26]. $ln(V_i)$ $ln(V_i)$ is the inbound link of V_i, $Out(V_j)$ is the outgoing link of V_j. Formula (5) shows that the weight of V_i vertex depends on the weight of the edge between V_j to V_i vertices and the sum of the weight of outgoing edges from V_j to others.

The YAKE algorithm consists of the following four steps [29]: preprocessing and term-candidate generation; determining the features of terms; counting points of the term; association of similar terms. At the first stage, the division is performed at the level of sentences, which are further divided into terms. At the stage of determining the features of terms, each term is evaluated by using special functions [29]. At the stage of calculating points for the term, the following formula is used:

$$S(t) = (Trel * Tposition)/Tcase + ((Tnorm/Trel) + (Tsentence/Trel)), \qquad (6)$$

Where *Tcase* – the importance of capitalization and acronyms, *Tposition* – more importance is given to the words that are present at the beginning of the document, *Tnorm* – word frequency, *Trel* – checks for the diversity of context in which this word is used, *Tsentence* – determines how often the candidate word occurs with different sentences. The highest score is given to words that are often found in different sentences. The last step is to combine the evaluation of morphologically similar words. According to this method, the better word has minimal evaluation.

The approach to the interpretation of the obtained results (stage 4) is given below.

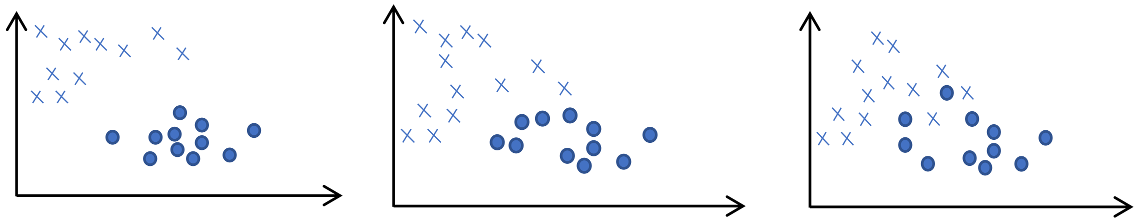
3.3. Criteria for the quality of visual analysis modeling

Multidimensional scaling (MDS) [24, 25] method is proposed to assess the quality of the obtained models for the classification task. MDS is one of the methods of reducing the dimension of vector space. The purpose of the method is to reduce the dimension to make it possible to visualize (3 or 2-dimensional). For example, the criterion for reducing dimensionality is the Euclidean distance between the vectors. Solving the optimization problem, we find the mapping $R^n \rightarrow R^2$, which makes it possible to obtain a two-dimensional graph of the relative position of the points-vectors and visually assesses the quality of the model for the classification task.

Visual criteria for assessing the quality of modeling are proposed (Figure 2).

Criterion 1 – High level model for text classification. The Figure 2 (a) shows, that the two classes are clearly separated. It indicates the correctness of the proposed model.

Criterion 2 – An acceptable level of model for classifying texts. Figure 2 (b) shows that two classes are tangent with each other. This result can be considered as workable, but it will require additional expert review to confirm the classification.



(a) (b) (c)

Figure 2: Model quality levels for classification task: (a) – high, (b) – acceptable, (c) – unsatisfactory

Criterion 3 – Unsatisfactory level of model for text classification. The Figure 2 (c) shows that the two classes are almost inseparable. The distance between classes is insignificant, and there is an intersection in some places. This model can't be considered workable, and it needs refinement.

It is proposed to use the above criteria to check the quality of the proposed model of the everyday Ukrainian language. We will consider the model correct if the results' value is in the range between the first and second criteria.

3.4. Dataset

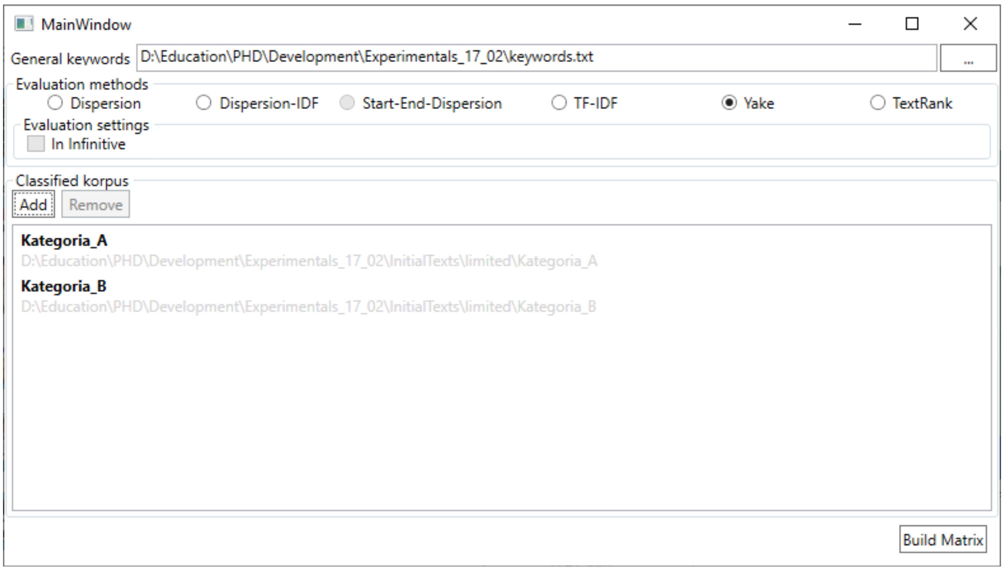
The following corpora of the Ukrainian categorized texts and frequency dictionaries were used to build the model and its validation (ability to classify texts of communication on the Internet):

1. ***БрУК*** – an open genre-balanced corpus of modern Ukrainian language with volume of 1 million word usages. The corpus is built on the foundations that formed the basis of the famous English corpus Brown [17]. This corpus also include the VESUM dictionary (URL: https://r2u.org.ua/vesum/), which includes defective words, profanity, "surzhik", which is an integral part of the everyday Ukrainian language.
2. ***Corpus of Ukrainian language MOVA.info*** – designed to search for tokens and word forms in Ukrainian texts of a specific style (for some part of corpus can be used to search for morphemes and syntactic structures) [18]. In this paper, it was used to construct a keyword vector.
3. ***UA-GEC*** – the first annotated GEC-corpus of the Ukrainian language. This is a collection of texts written by ordinary people. It includes texts like essays, blog posts, social networks, reviews, letters, etc. These texts contain grammatical, stylistic, and spelling mistakes that bring them as close as possible to everyday language [19].
4. ***Ukrainian News collection*** – a collection of over 150,000 news articles collected from over 20 news resources. Data sets are divided into the following 5 categories: politics, sports, news, business, technology. The data set was provided by the non-profit student organization FIdo.ai (FIdo Machine Learning Research Department of the National University of Kyiv-Mohyla Academy) for research purposes of data analysis (classification, clustering, keyword selection, etc.) [20].
5. **Ukrainian web-building of the University of Leipzig** – the corpus of Ukrainian language texts. Contains frequency dictionaries. Dictionaries were created based on Wikipedia, news sites, web documents. Text can be downloaded with different size (words): 10 000, 30 000, 100 000, 300 000, 1000 000 [21].
6. ***Karpaty bud karkas*** – set of articles with information about the progress of construction, news of modern architecture, technology. The corpus contains more than 200 texts containing 500 words in an average (https://karpatybud.com.ua/statti/).
7. **Gardener's blog** – site contains more than 200 texts of about 500 words each, dedicated to the topic of gardening (https://agro-market.net/ua/news/).

This number of sources is needed because the everyday Ukrainian language covers many areas of life. That is why taking only one of the corpora for the study cannot cover the entire vocabulary of everyday language.

4. Results and discussion

The software application in C# was developed to validate the proposed model in its ability to classify texts in the everyday Ukrainian language. This application converts the textual content of files from the training set into a digital representation. The main window of the developed application is shown in Figure 3.



General keywords D:\Education\PHD\Development\Experimentals_17_02\keywords.txt

Evaluation methods

○ Dispersion ○ Dispersion-IDF ○ Start-End-Dispersion ○ TF-IDF ● Yake ○ TextRank

Evaluation settings

☐ In Infinitive

Classified korpus

Kategoria_A
D:\Education\PHD\Development\Experimentals_17_02\InitialTexts\limited\Kategoria_A

Kategoria_B
D:\Education\PHD\Development\Experimentals_17_02\InitialTexts\limited\Kategoria_B

Figure 3: Experimental application for converting textual content into a digital representation

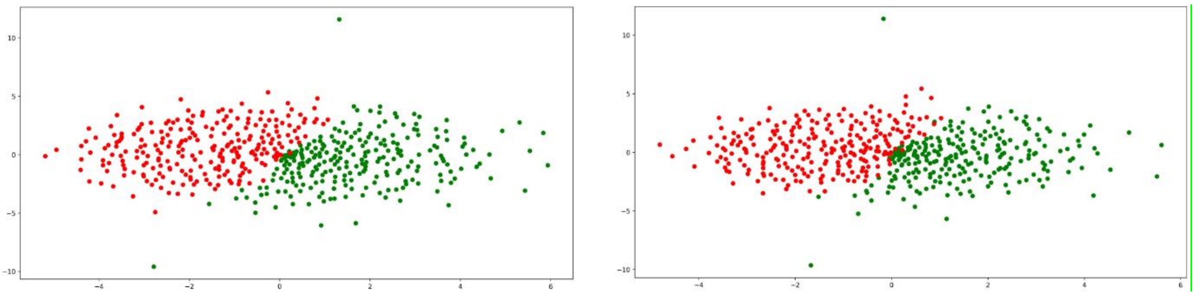
The created program allows setting following parameters:

1. The path to the file containing the key terms.
2. Choose a method of evaluating terms in the text.
3. Choose the corpora of texts to be processed.

The result of the application is a file that contains a digital representation of each text from the selected corpora.

Parameters of the text data analysis environment. The resulting file is sent for processing to a software application developed in the Python programming language using the Manifold library. This application reads data from the input file from the previous step and processes the received data using the MDS method from the Manifold library. The result obtained from the MDS method is visualized on the graphical interface.

The basis of the proposed generalized vector $\overline{Word}$ is a dictionary MOVA.info [18], as it is closest to the tasks of classifying Internet content with the addition of words from other dictionaries. Filtering by noun was also performed to achieve this goal because the dictionary had different parts of speech and was heterogeneous. The resulting vector consists of 1500 words.

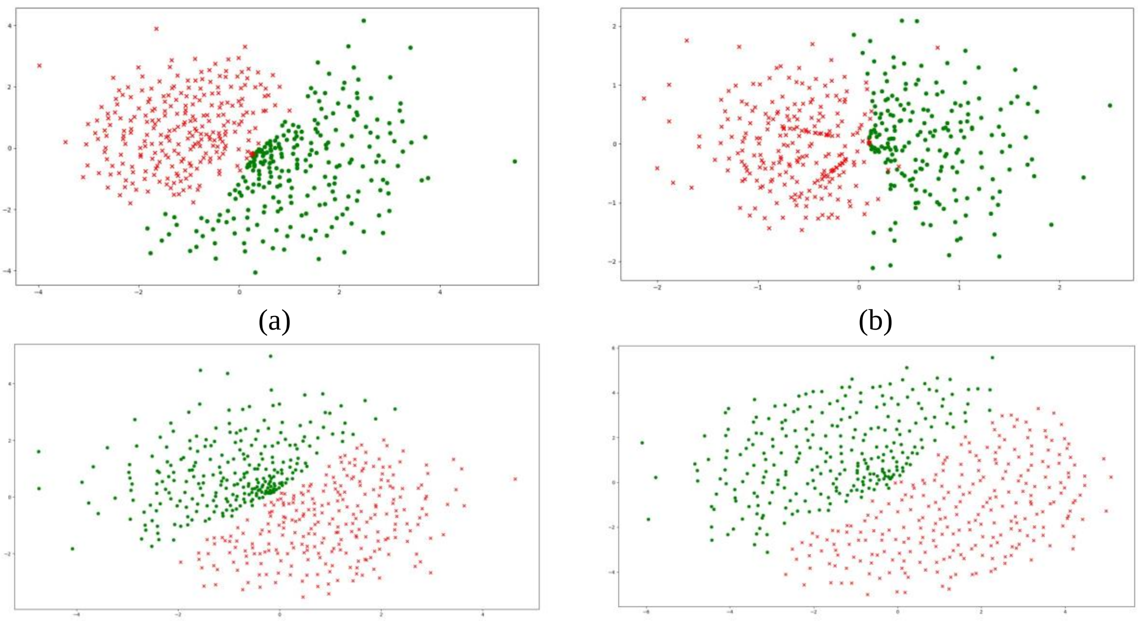


(a) (b)

Figure 4: Visualization of binary classification of texts by TF-IDF and vector with (a) 3000 words and (b) 2000 words length

The words vector length of the model for vectorization of the Ukrainian-language segment of the Internet in 1500 units were established based on research results. For example, Figure 4 (a) shows the binary classification of the described set of texts by TF-IDF by vector length of 3000 words, and Figure 4 (b) shows the visualization of the binary classification of the described set of texts by TF-IDF by vector lengths of 2000 words. Studies have shown that the best ability to separate in the classification of texts is observed at a word vector length of 1500 units.

Texts of two following categories were collected for experimental research: construction (https://karpatybud.com.ua/statti/) and gardening (https://agro-market.net/ua/news/). These collections contain 200 texts each and an average length of 500 words for each text. It is enough to use two categories to validate the proposed model and conduct experimental research because the research goal is binary classification.



(a) (b) (c) (d)

Figure 5: Validation of model quality by methods: (a) TF-IDF, (b) Dispersion Evaluation, (c) Text Rank, (d) YAKE

Result 1. The result of validation of the quality of the model of everyday lexicon of the Ukrainian-language segment of the Internet using the TF-IDF method is shown in Figure 5 (a). The result can be assessed as unsatisfactory because the categories of some texts were defined incorrectly and the area of division/delimitation is blurred. It is due to the characteristics of this method because this method is high sensitivity to the selection of texts of alternative categories. Texts of alternative categories may be outside the classes used for classification. It may lead to cases of incorrect classification of texts.

Result 2. The results of validation of the quality of the model of everyday lexicon obtained by using the method of dispersion evaluation are shown in Figure 5 (b). The result can be assessed as acceptable because there are texts that are contained on the class boundaries. The method of dispersion evaluation requires as many appearances of meaningful words in individual texts for the effective computing of value of the semantic importance of words. Everyday communication is characterized with using the small number of meaningful words, that is why some time unsatisfactory statistical values of unique words of the text do not provide sufficient separation of values.

Result 3. The result of using TextRank method is shown in Figure 5 (c). The quality of the result can be assessed as high because both categories have a clear separation. To calculate the semantic importance of words, the Text Rank method uses not only the actual position of words in the text, but also the relationships between words and the relationship between the frequencies of occurrence of words in relation to the relationships between words. This allows us to take into account more parameters of the text than other methods, but the topics of everyday communication do not allow obtaining many primary indicators for evaluation. That is why this method has shown a fairly high efficiency of division of texts by category.

Result 4. The result of using YAKE method is shown in Figure 5 (d). Such results can be interpreted between high and satisfactory. The categories are divided, but there is no clear boundary. This method takes into account a number of important indicators of the text, such as word frequency, variety of context of word occurrences and frequency of word occurrences in different sentences.

However, the characteristic features of everyday communication do not allow the method to use its advantages. Such advantages include taking into account using of capital letters and abbreviations, the presence of words at the beginning of the text. Also, when the number of texts increases, the capabilities of the YAKE method decrease and its use may lead to fuzzy classification. That is why this method is effective for texts of another type, like scientific articles. The method showed a less satisfactory result during analysis of texts on everyday communication.

Result 5. This research was performed by using the set of words that consists of the intersection of sets of keywords found by different methods. The resulting set of words was about 500 words, which was common for the methods considered. The obtained words can be considered sufficient for modeling the task of classification by the sets of texts. The obtained set can be used as a basis for the formation of a vector model in the future. The formation of the model can be done by supplementing the basic set of words with words that are specific to the specific tasks like detection of suicidal ideation, bullying, negative emotional coloring of texts, negative content and etc.

The results of validation of the model of the everyday lexicon of the Ukrainian-language segment of the Internet show that the classification of everyday texts is the most effective with using the Text Rank method for keyword searching. Further research will aim at improving and modifying the described approaches by testing assumptions, using methods for vectorization of texts, and so on. Further research will also be aimed at improving the general vector of words of the everyday Ukrainian language and solving problems of determining the negatively colored text messages in the segment of Internet communication.

5. Conclusion

The paper proposes a modern model of everyday Ukrainian language built by combining meaningful frequency dictionaries. The model should be used to analyze the content of the Ukrainian-language segment of the Internet, which includes using swear words and grammatically and syntactically incorrect words because existing frequency dictionaries do not fully cover this segment. To create a correct model of the modern everyday Ukrainian language, frequency dictionaries of the existing corpora of the Ukrainian language were used, covering different areas of activity, types of text, different areas of communication (including everyday communication on the Internet). The vector of words is obtained by combining frequency dictionaries of these corpora of the Ukrainian language. The resulting vector was filtered by a noun and limited in number.

Two orthogonal sets of texts, more than 200 in each, were taken to validate the proposed model. Each text was vectorized by one of the four proposed keyword search methods (TF-IDF, variance, text rank, YAKE) and was assigned to a specific category. To assess the quality of the obtained models, the MDS method was used, and the criteria for the interpretation of the obtained results on a three-level scale were proposed. This allowed us to determine the keyword searching method that is best suited for use with the proposed model of modern Ukrainian everyday language.

The studies identified the following results:

1. It is established that the best ability for separation during classifying texts using the proposed model of modern Ukrainian everyday language is observed at a word vector length of 1500 words. It determines the optimal dimension of the model.

2. According to the results of test classification of more than 400 texts and using visual verification of classification results by MDS, it was determined that the Text Rank method for keyword searching is best suited for using the proposed model of modern Ukrainian everyday language.

3. It was confirmed that the MDS method for visual analysis is an effective and sufficient tool for visual verification of the classification results of digital texts into various categories, which include both thematic categories and categories of the emotional coloring of texts.

The points above allow arguing about the possibility of using the created model of the modern everyday Ukrainian language to solve the tasks of analysis of textual content of Internet communication and its classification according to various grounds. This is especially important to prevent suicidal tendencies, detect bullying on social networks, determine the negative emotional color of texts, and warn users about possible harmful content.

A characteristic feature of the approach considered in the paper is high efficiency in working with the Ukrainian language as a characteristic representative of inflectional languages. Also, the features of the approach allow being effective for analytical languages, including English. This increases the value of the approach to working with Ukrainian-language content of everyday communication because it often contains English words and proper names as borrowings. The approach's effectiveness for agglutinative languages such as Hungarian is lower because identifying and working with formats is somewhat different from working with inflections of inflected languages and requires additional solutions. Thus, it creates a separate area of further research on working with everyday communication texts with mixed, multilingual content.

In further research, it is planned to improve the above approaches to vectorization of text with verification and assumption about the use of the composition of methods and so forth. There is also a need to focus research on improving the general vector of words of the everyday Ukrainian language and solving problems of determining the negative color of text messages during online communication.

6. References

[1] O. Hargie, Skilled Interpersonal Communication. Research, Theory and Practice, 7th. ed., London, 2021. doi:10.4324/9781003182269.

[2] I.Kryvonos, I. Krak, O Barmak, R. Bagriy, Predictive text typing system for the Ukrainian language, Cybern. Syst. Anal. 53(4) (2017) pp. 495–502. doi:10.1007/s10559-017-9951-5.

[3] A. Y. Lee, R. Katz, J. Hancock, The Role of Subjective Construals on Reporting and Reasoning about Social Media Use, Social Media + Society (2021). doi:10.1177/20563051211035350.

[4] G. C.Huang, J. B.Unger, D. Soto, K. Fujimoto, M. A. Pentz, M. Jordan-Marsh, T. W. Valente, Offline Friendship Networks on Adolescent Smoking and Alcohol Use, 54(5) (2014) pp. 508-514. doi:10.1016/j.jadohealth.2013.07.001.

[5] R. J. Moreira de Freitas, T. N. Carvalho Oliveira, J. A. Lopes de Melo, J. do V. e Silva, K. C. de Oliveira e Melo, S. Fontes Fernandes, Adolescents' perceptions about the use of social networks and their influence on mental health, Enfermería Global, 20(64) (2021) pp. 324-364. doi:10.6018/eglobal.462631.

[6] S. Saumya, A. Kumar, J. P. Singh, Offensive language identification in Dravidian code mixed social media text, Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, 2021, pp. 36–45.

[7] B. Dave, Sh. Bhat, P. Majumder, IRNLP DAIICT@DravidianLangTech-EACL2021: Offensive Language identification in Dravidian Languages using TF-IDF Char N-grams and MuRIL, Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, 2021, pp. 266-269.

[8] D. Sivalingam, S. Thavareesan, OffTamil@DravideanLangTech-EACL2021: Offensive Language Identification in Tamil Text, Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, 2021, pp. 346-351.

[9] Ch. Ma, A. Shen, H. Yoshikawa,T. Iwakura, D. Beck, T. Baldwin, On the (In)Effectiveness of Images for Text Classification, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2021, pp. 42-48. doi:10.18653/v1/2021.eacl-main.4.

[10] D. Ghosh, R. Shrivastava, S. Muresan, «Laughing at you or with you»: The Role of Sarcasm in Shaping the Disagreement Space, 2021, pp. 1998-2010. doi:10.18653/v1/2021.eacl-main.171.

[11] Zh. Zhu, K. Meng, J. Caraballo, I. Jaradat, X. Shi, Z. Zhang, F. Akrami, H. Liao, F. Arslan, D. Jimenez, M. S. Saeef, P. Pathak, Ch. Li, A Dashboard for Mitigating the COVID-19 Misinfodemic, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2021, pp. 99–105. doi:10.18653/v1/2021.eacl-demos.12.

[12] J. Barnes, L. Øvrelid, E. Velldal, If you've got it, flaunt it: Making the most of fine-grained sentiment annotations, Proceedings of the 16th Conference of the European Chapter of the

Association for Computational Linguistics, Association for Computational Linguistics, 2021, pp. 49–62. doi:10.18653/v1/2021.eacl-main.5.

[13] M. Wiegand, J. Ruppenhofer, Exploiting Emojis for Abusive Language Detection, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2021, pp. 369–380. doi:10.18653/v1/2021.eacl-main.28.

[14] N. de Silva, D. Dou, Semantic Oppositeness Assisted Deep Contextual Modeling for Automatic Rumor Detection in Social Networks, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2021, pp. 405–415. doi:10.18653/v1/2021.eacl-main.31.

[15] D. Yan, K. Li , Sh. Gu, L. Yang, Network-Based Bag-of-Words Model for Text Classification, IEEE Access, Volume 8 (2020). doi:10.1109/ACCESS.2020.2991074.

[16] C. Macdonald, N. Tonellotto, S. MacAvaney, IR From Bag-of-words to BERT and Beyond through Practical Experiments, CIKM '21: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Association for Computing Machinery, New York, United States, 2021. doi:10.1145/3459637.3482028.

[17] Corpus of Modern Ukrainian Language (BRUK). URL: https://r2u.org.ua/corpus.

[18] MOVA.info: about the Ukrainian language, linguistics and more. URL: http://www.mova.info/.

[19] UA-GEC: the first annotated GEC-corpus of the Ukrainian language. URL: https://ua-gec-dataset.grammarly.ai/.

[20] Ukrainian News is a collection. URL: https://github.com/fido-ai/ua-datasets/tree/main/ua_datasets/src/text_classification.

[21] Deutscher Wortschatz. Corpora Ukrainian. URL: https://wortschatz.uni-leipzig.de/en/download/Ukrainian#ukr_mixed_2014.

[22] Zh. Jiang, Bo Gao, Y. He, Y. Han, P. Doyle, Q. Zhu, Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports, Mathematical Problems in Engineering (2021). doi:10.1155/2021/6619088.

[23] I. Krak, O. Barmak, O. Mazurets, The practice investigation of the information technology efficiency for automated definition of terms in the semantic content of educational materials. CEUR Workshop Proceedings, 2016, vol.1631, pp. 237–245. doi:10.15407/pp2016.02-03.237.

[24] I. Krak, O. Barmak, E. Manziuk, Using visual analytics to develop human and machine-centric models: A review of approaches and proposed information technology. Computational Intelligence. 2020; pp. 1–26. doi:10.1111/coin.12289.

[25] E. L. Fink, D. A. Cai, Multidimensional Scaling, The International Encyclopedia of Media Psychology. Hoboken, NJ: Wiley, 2020. doi:10.1002/9781119011071.iemp0282.

[26] A. Kazemi, V. P'erez-Rosas, R. Mihalcea, Biased TextRank: Unsupervised Graph-Based Content Extraction, Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 1642–1652. doi: 10.18653/v1/2020.coling-main.144.

[27] Zh. Huang, Zh. Xie, A patent keywords extraction method using TextRank model with prior public knowledge, Complex Intell. Syst. (2021). doi:10.1007/s40747-021-00343-8.

[28] M. Zhang, X. Li , Sh. Yue, L. Yang, An Empirical Study of TextRank for Keyword Extraction, IEEE Access, Volume 8 (2020). doi: 10.1109/ACCESS.2020.3027567.

[29] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, YAKE! Keyword extraction from single documents using multiple local features, Information Sciences, Volume 509, 2020, pp. 257–289. doi: 10.1016/j.ins.2019.09.013.

[30] R. A. Yunmar, A. Setiawan, H. Tantriawan, The Combination of YAKE and Language Processing for Unsupervised Term Extraction Ontology Learning, International Conference on Science, Infrastructure Technology and Regional Development, 2019. doi:10.1088/1755-1315/537/1/012023.