ISSN 2224-087X. Електроніка та інформаційні технології. 2023. Випуск 24. С. 46–56
Electronics and information technologies. 2023. Issue 24. P. 46–56

UDC 004.932 DOI: https://doi.org/10.30970/eli.24.5

USING FEATURE ENGINEERING IN MACHINE LEARNING MODELS FOR FAKE NEWS DETECTION

I. Olenych, M. Prytula, Ya. Boyko, O. Sinkevych, O. Khamar

Ivan Franko National University of Lviv,
50 Dragomanov Street, 79005 Lviv, Ukraine
igor.olenych@lnu.edu.ua

In this study, the analytical system for processing Ukrainian and Russian texts and automatically detecting fake news was developed. The effectiveness of text message classification using the naive Bayes classifier, support vector machine, *k*-nearest neighbors, random forest and logistic regression methods was studied. It has been established that adding to the feature vector the number of positive and negative words, the text tone, and the aggression presence makes it possible to increase the accuracy of detecting fake news for developed machine learning models. The methods of support vector machine and logistic regression demonstrate the highest effectiveness of text message classification.

Keywords: computer text analysis, fake detection, machine learning, feature engineering.

1. Introduction

The large volume of information that has become traditional for the information society creates new challenges for humanity. The usual newspaper, the radio and TV news are beginning to give way to new means of disseminating information on a global scale and in real-time due to the invention of the Internet and the emergence of social networks. Social media not only provide instant transmission of information, photos, and video content but also encourage users to participate in online discussions of news. This activism around Internet news can lead to serious societal consequences. The problem of difficult access to information, which was relevant earlier, is being replaced by new problems: structuring and analyzing information and determining its reliability. Disinformation has become one of the biggest problems in today's digital world, a threat to democracy and freedom of expression [1, 2]. Manipulative distortion of facts and fake news as tools of information warfare can influence the public's views on specific world events, military conflicts, and relations between states. As a result, significant political and economic losses can be caused. In particular, a large flow of disinformation was related to the Brexit referendum and the US presidential election in 2016.

Kremlin propaganda actively produces fake news to carry out cultural, religious, and informational expansion. With the beginning of a full-scale Russian invasion of the territory of Ukraine, an even stronger wave of aggressor disinformation began, and cyberspace became a battleground for the preservation of the country's viability. Therefore, the development of new, adapted to modern conditions, technical means of identifying sources of fake news, aggressive rhetoric, and provocative calls for unauthorized actions are necessary measures to neutralize digital threats.

Social and psychological factors are important for the spread of fakes. Most of the fake news in social media usually contains manipulative content and is aimed at people who take information emotionally without proper analysis and verification. Emotional perception of information, especially of a negative nature, makes critical analysis impossible, so a person's ability to recognize fake news is low [3]. In addition, the credibility of news on social networks cannot be verified manually due to the large volume of information and its rapid spreading. However, false content has certain characteristics, such as informal thinking, use of abbreviations, transfer of less information, negative tone of messages, etc. [4]. The detected features make it possible to develop quite effective methods of computer analysis of news messages and automated detection of fakes. In particular, artificial intelligence technologies are widely used for this purpose [5–7].

To increase the efficiency of the classification of information materials, modern machine learning models take into account contextual and syntactic information, as well as the visual representation that accompanies news [8–10]. However, the accuracy of machine learning models significantly depends on the quantitative and qualitative characteristics of the data sets used to train the models. Most available datasets are only suitable for training English-language NLP models. Therefore, the effectiveness of the application of various machine learning models for the automatic analysis of information flows and the detection of fake messages in the Ukrainian and Russian languages is studied in the work. Special attention is focused on ways to improve the accuracy of the classification of news materials. Taking into account the revealed correlation between fake and aggressive messages [11], we have suggested expanding the feature vector of classification models with information of an emotional nature.

2. Methods and means of implementation

Text data sets with news related to full-scale Russian aggression were used for the construction of NLP models and their implementation. Data sets were formed in cooperation with specialists of the Faculty of Journalism of Ivan Franko Lviv National University in the period from June to December 2022. The text data were divided into two classes: fake and true news. In total, 1615 and 1458 messages in Ukrainian and Russian have been analyzed, respectively. Since the amount of misinformation was slightly higher, the training sampling was balanced by reducing the number of fake news due to their random exclusion from the data sets. Besides, the texts were classified as aggressive or neutral messages, which was used as an additional parameter to improve the effectiveness of detecting fakes.

The proposed models for the classification of text information were implemented in accordance with the algorithm, which is shown in Fig. 1. The developed analytical system makes it possible to download news materials obtained from various sources: publications in electronic media, messages on social networks, etc. At the initial stage, the text is divided into tokens [12]. After the stemming procedure and exclusion of punctuation marks and stop words, the resulting array of words was used for further analysis. Vectorization of the text was carried out using the statistical indicator Term Frequency – Inverse Document Frequency (TF–IDF), which reflects not only the frequency of its appearance but the importance of the word in the text [13]. In addition, a search for each word from the received array was carried out in Ukrainian and Russian tonality dictionaries [14–16] to determine the tone of the news messages. Additional features for binary classification of news were the number of positive and negative words, the aggressiveness and tone of the message, and the total number of words in the text. As a result, we obtained the extended TF–IDF$^+$ vector. The next step of the fake detection algo-

rithm is the implementation of machine learning methods using test data sets. The naive Bayes classifier, support vector machines (SVM), k-nearest neighbor (KNN) algorithm, random forest and logistic regression methods were used in the work. The final stage of the proposed algorithm is the evaluation of the model's effectiveness.

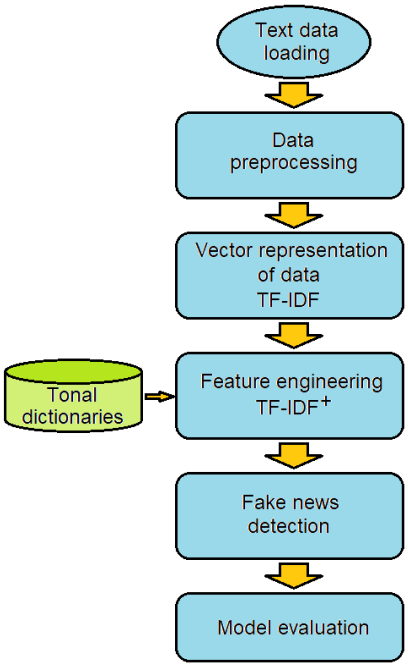
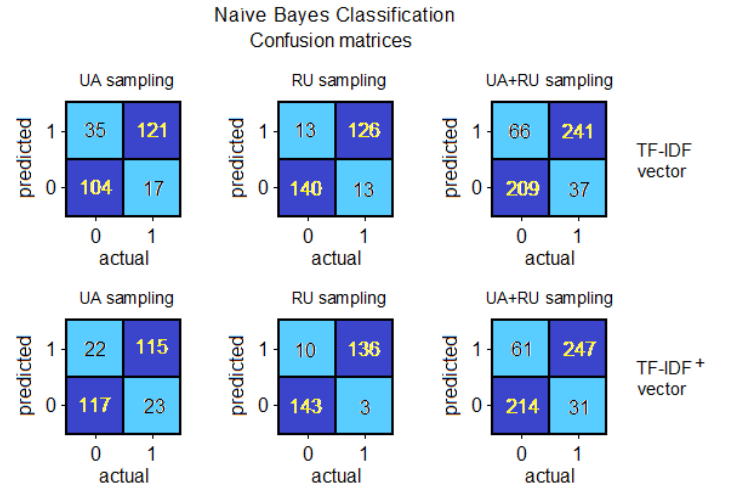


Fig. 1. The algorithm for detecting fakes in text messages.

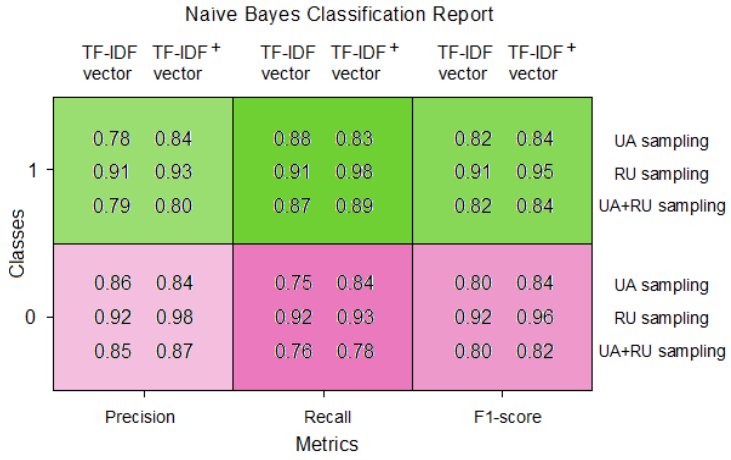
The Python 3.9 language and the Jupyter Notebook environment were used to create the software modules. The machine learning models in the Python language were implemented using the Scikit-learn library. The library provides a wide range of tools for data preparation, dimensionality reduction, classification, regression, clustering, and other machine-learning tasks, including algorithms for model evaluation.

3. Results and discussion

Classified news materials were used to train the machine learning models proposed in the work and to evaluate the effectiveness of the developed analytical system in detecting fakes. We analyzed news in Ukrainian and Russian languages both separately and together. The set of classified text data was divided into training and test samples in the ratio of 80 and 20%, respectively. The trained models were used to classify a test sample of news messages. Usually, the results of solving a binary classification problem are marked as positive or negative, that is, correctly or incorrectly classified instances. These solutions are visually represented in the form of a confusion matrix [17]. Here, class 0 relates to fake and class 1 to neutral messages. The obtained reports on the effectiveness of the models are shown in Fig. 2 – Fig. 6.



a)



b)

Fig. 2. Confusion matrices (*a*) and fake news detection report (*b*) using naive Bayes classification.

Various metrics of classification efficiency are calculated based on the confusion matrix and its values:

• *accuracy*, which measures the ratio of relevant instances to the total number of instances

$$accuracy = (TP + TN) / (TP + TN + FP + FN),$$

where TP and FP are the numbers of correctly and incorrectly classified instances of the positive class, TN and FN are the numbers of correctly and incorrectly classified instances of the negative class, respectively;

• *precision*, which determines the ratio of relevant instances of the positive class to the total instances of the positive class

$$precision = TP / (TP + FP);$$

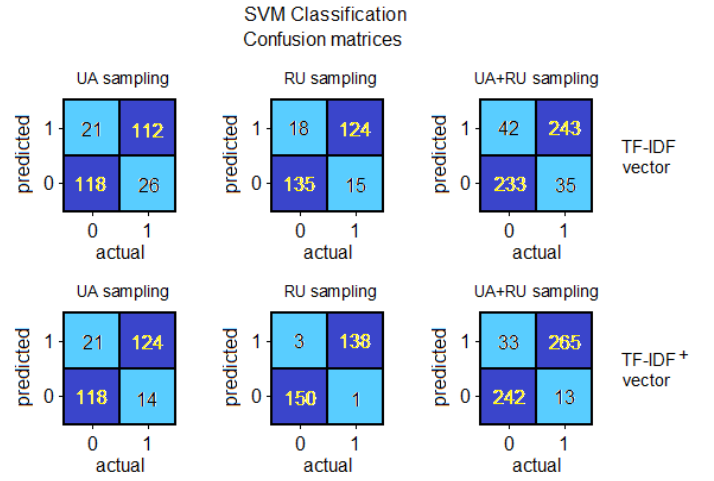
• *recall*, which determines the ratio of relevant instances of the positive class to the total number of truly positive instances

$$recall = TP / (TP + FN);$$

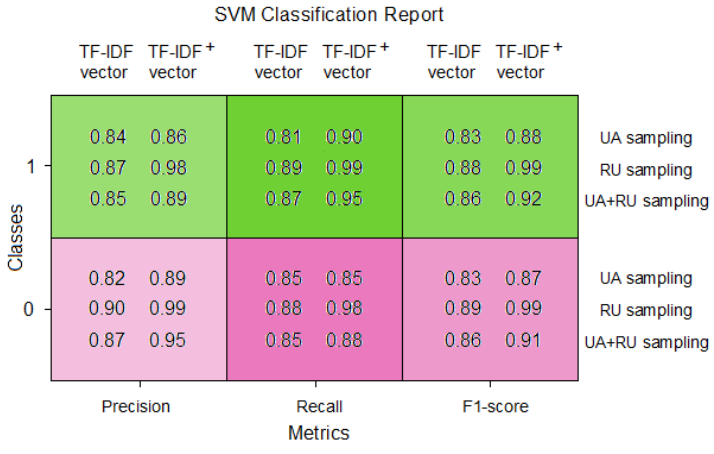
• *F1-score*, as a metric of consistency between precision and recall, which demonstrates how many instances are correctly predicted by the model and how many true instances the model will not miss

$$F1\text{-}score = 2 * (precision * recall) / (precision + recall).$$

F1-score shows a generalized assessment of the model's effectiveness.



a)

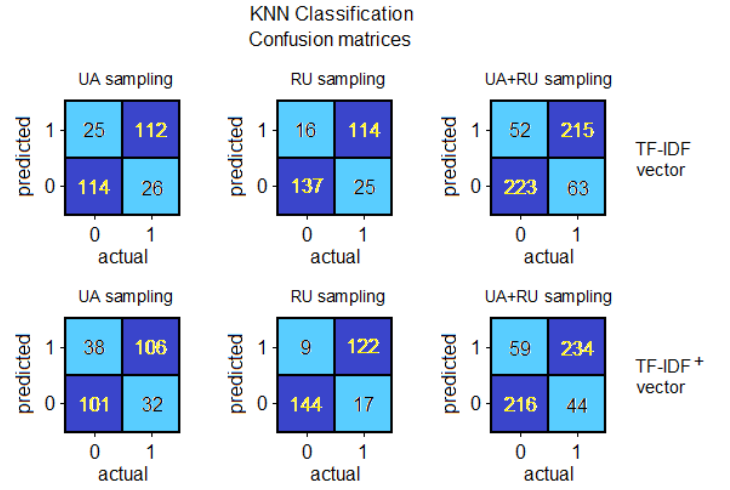


b)

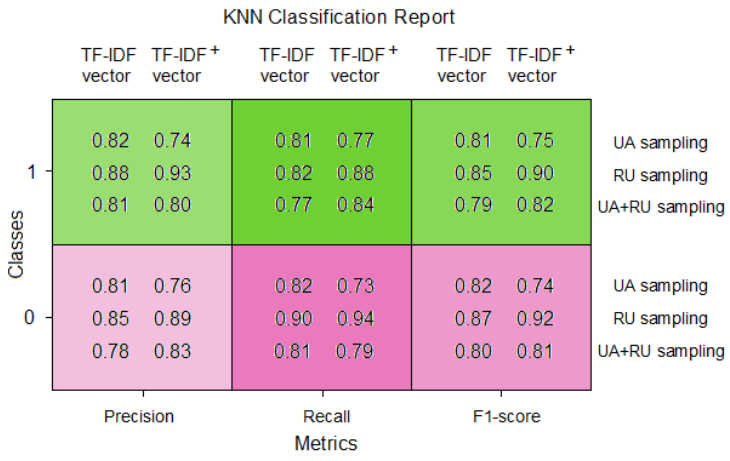
Fig. 3. Confusion matrices (*a*) and fake news detection report (*b*) using SVM classification.

The analysis of the obtained results allows us to conclude that classical methods of machine learning recognize fake messages in Ukrainian and Russian languages with satisfactory

effectiveness. However, quantitative and qualitative indicators of training text data sets are extremely important for machine and deep learning models. This is probably the reason for the higher effectiveness of detecting fakes in the Russian language than in Ukrainian by all applied metrics for most of the used methods. In particular, the biggest difference between F1-score values was for the naive Bayes classifier. The least influence of the training data language on the effectiveness of the detection of fake messages was observed for the logistic regression method.



a)

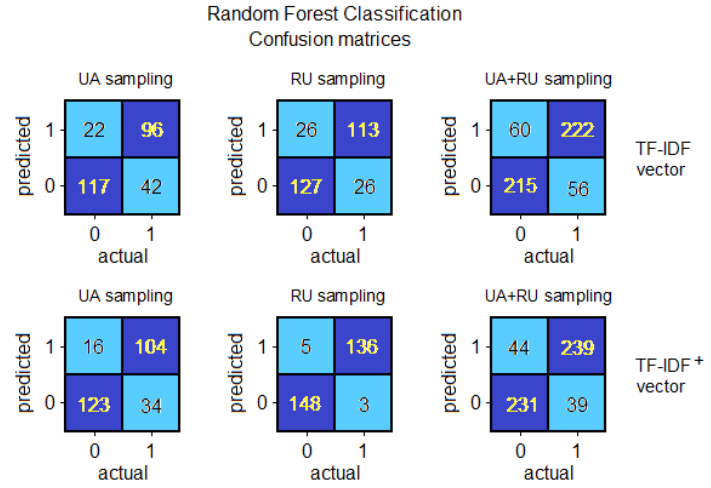


b)

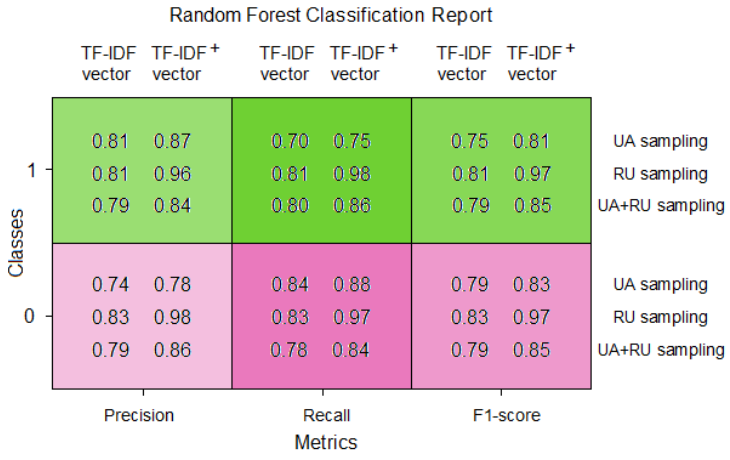
Fig. 4. Confusion matrices (*a*) and fake news detection report (*b*) using KNN classification.

Combining training data sets in Ukrainian and Russian languages does not significantly affect the effectiveness of recognition of fake news. The values of the F1-score almost do not differ from those for the Ukrainian-language models. On the other hand, combining the data

sets makes it possible to simplify the process of automated classification of textual information because it does not require an additional procedure for recognizing the language of the news message.



a)

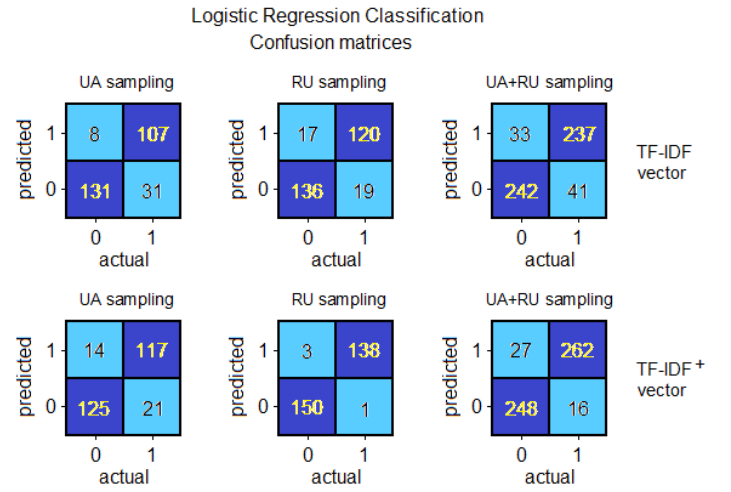


b)

Fig. 5. Confusion matrices (*a*) and fake news detection report (*b*) using random forest classification.

The application of feature engineering, namely the addition of new features related to the sentiment analysis of the news message to the TF–IDF vector, makes it possible to improve the effectiveness of fake detection by various machine learning models. An increase in the values of precision, recall, F1-score and accuracy was observed for almost all the proposed models. The results of assessing the effectiveness of recognition of fake news according to the accuracy

metric are shown in Table 1. However, the model trained on the UA sampling that uses the KNN algorithm and TF–IDF^{+} vector demonstrates a decrease in the accuracy of the message classification. The detected decrease in the effectiveness of recognizing fakes is probably due to the small size of the UA sampling on the one hand and the poorly chosen distance metric on the other hand.



a)

b)

Fig. 6. Confusion matrices (*a*) and fake news detection report (*b*) using logistic regression classification.

The most significant progress in the automated classification of text messages due to the proposed extension of the feature vector was observed for the RU sampling (see Table 1). We associate this with a stronger correlation between fake messages and aggressive rhetoric in

Russian-language news than in Ukrainian-language ones. In particular, the values of the linear correlation coefficient between misinformation and aggression were 0.6 and 0.8 for the UA and RU samplings, respectively [11].

Table 1. Accuracy of machine learning models for fake news detection.

Method	Accuracy, %					
	UA sampling		RU sampling		UA+RU sampling	
	TF–IDF	TF–IDF$^+$	TF–IDF	TF–IDF$^+$	TF–IDF	TF–IDF$^+$
Naive Bayes	81	84	91	96	81	83
SVM	83	87	89	99	86	92
KNN	82	75	86	91	79	81
Random Forest	77	82	82	97	79	85
Logistic regres-sion	86	87	88	99	87	92

In general, the SVM and logistic regression methods demonstrate the best classification results of news messages in Ukrainian and Russian languages related to Russian aggression. After taking into account additional features, in particular, the number of positive and negative words, the tone of the text, and the presence of aggressive rhetoric in the message, the accuracy of the classification of the UA+RU sampling is 92%.

4. Conclusions

The analytical system proposed in the work can process Ukrainian and Russian texts and automatically detect fake news. The analysis of news messages in electronic media and social networks related to the full-scale Russian invasion of Ukraine was carried out using naive Bayes classifier, SVM, KNN, logistic regression and random forest methods. The effectiveness of various machine learning models in the classification of information materials was compared.

It has been established that feature engineering makes it possible to increase the accuracy of detecting fake news by 2–6 % for machine learning models trained on the UA+RU sampling. In particular, after adding to the feature vector the word number in the message number of positive and negative words, the text tone, and the aggression presence, the classification accuracy can reach 92%. The SVM and logistic regression methods demonstrate the best results of text message classification. The lowest effectiveness of detecting fake news in the Ukrainian and Russian languages was in the KNN method.

References

1. *Zhang X., Ghorbani A.A.* An overview of online fake news: Characterization, detection, and discussion // Inf. Process. Manag. – 2020. – Vol. 57, no. 2. – P. 1–26.
2. *Aimeur E., Amri S., Brassard G.* Fake news, disinformation and misinformation in social media: a review // Social Network Analysis and Mining. – 2023. – Vol. 13. – 30. https://doi.org/10.1007/s13278-023-01028-5.

3. *Rubin V.* On deception and deception detection: Content analysis of computer-mediated stated beliefs // Proceedings of the American Society for Information Science and Technology. - 2010. https://doi.org/10.1002/meet.14504701124
4. *Zhou Z., Guan H., Bhat M.M., Hsu J.* Fake News Detection via NLP is Vulnerable to Adversarial Attacks // 11th International Conference on Agents and Artificial Intelligence. – 2019. https://doi.org/10.5220/0007566307940800
5. *Villela H.F., Correa F., Ribeiro J.S. de A.N., Rabelo A., Carvalho D.B.F.* Fake news detection: a systematic literature review of machine learning algorithms and datasets // Journal on Interactive Systems. - 2023. - Vol. 14. - P. 47 - 58. https://doi.org/10.5753/jis.2023.3020.
6. *Khanam Z., Alwasel B.N., Sirafi H., Rashid M.* Fake News Detection Using Machine Learning Approaches // IOP Conf. Series: Materials Science and Engineering. – 2021. – Vol. 1099. – 012040. https://doi.org/10.1088/1757-899X/1099/1/012040.
7. *Umer M., Imtiaz Z., Ullah S., Mehmood A., Choi G.S., On B.W.* Fake news stance detection using deep learning architecture (CNN-LSTM) // IEEE Access. – 2020. – Vol. 8. – P. 156695–156706. https://doi.org/10.1109/ACCESS.2020.3019735.
8. *Zhang G., Giachanou A., Rosso P.* SceneFND: Multimodal fake news detection by modelling scene context information // Journal of Information Science. – 2022. – P. 1–13. https://doi.org/10.1177/01655515221087683
9. *Cao J., Qi P., Sheng Q., Yang T., Guo J., Li J.* Exploring the role of visual content in fake news detection // In book: Disinformation, Misinformation, and Fake News in Social Media. – 2020. – P. 141–161. https://doi.org/10.1007/978-3-030-42699-6_8
10. *Song C., Ning N., Zhang Y., Wu B.* A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks // Information Processing and Management. – 2021. – Vol. 58. – P. 1–14. https://doi.org/10.1016/j.ipm.2020.102437
11. *Prytula M., Olenych I.* Detection of aggressive rhetoric in text using machine learning algorithms // Electronics and information technologies. – 2023. – Issue 22. – P. 34–45. https://doi.org/10.30970/eli.22.4.
12. *Thelwall M., Buckley K., Paltoglou G., Kappas A., Cai D.* Sentiment strength detection in short informal text // Journal of the American Society for Information Science and Technology. – 2010. – No. 61. – P. 2544–2558.
13. *Robertson S.* Understanding Inverse Document Frequency: On Theoretical Arguments for IDF // Journal of Documentation. – 2004. – Vol. 60, No. 5. – P. 503–520.
14. Ukrainian tonal dictionary [Electronic resource]. - Mode of access: https://github.com/lang-uk/tone-dict-uk/blob/master/tone-dict-uk.tsv.
15. Ukrainian tonal dictionary [Electronic resource]. - Mode of access: https://github.com/lang-uk/tone-dict-uk/blob/master/tone-dict-uk-manual.tsv.
16. Russian tonal dictionary [Electronic resource]. - Mode of access: https://github.com/dkulagin/kartaslov.
17. *Vijayarani S., Nithya M.N.* Efficient machine learning classifiers for automatic information classification // Int. J. Mod. Trends Eng. Res. – 2015. – Vol. 2. – P. 685–694.

ВИКОРИСТАННЯ ІНЖЕНЕРІЇ ОЗНАК У МОДЕЛЯХ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН

І. Оленич, М. Притула, Я. Бойко, О. Сінькевич, О. Хамар

Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, 79005 м. Львів, Україна
igor.olenych@lnu.edu.ua

Стрімке збільшення обсягу інформаційних потоків, інтернет-новин і повідомлень у соціальних мережах створює нові виклики для суспільства і потребує сучасних інструментів для структуризації та аналізу інформації в режимі реального часу, а також визначення її достовірності. Важливість протидії дезінформації та забезпечення інформаційної безпеки значно зросла з початком повномасштабного російського вторгнення в Україну. Тому розробка нових, пристосованих до сучасних умов технічних засобів виявлення джерел фейкових новин та агресивної риторики є необхідними заходами для нейтралізації цифрових загроз.

У роботі запропоновано моделі класифікації текстової інформації українською та російською мовами для виявлення фейкових повідомлень. Для навчання та тестування розроблених моделей машинного навчання було використано набір із понад 4000 новин в електронних засобах масової інформації та соціальних мережах, пов'язаних із повномасштабною російською агресією. На основі аналізу новинних повідомлень за допомогою наївного класифікатора Байєса та методів опорних векторів, *k*-найближчих сусідів, логістичної регресії та випадкового лісу порівняно ефективність класифікації інформаційних матеріалів різними моделями машинного навчання. Особливу увагу зосереджено на шляхах підвищення ефективності класифікації новинних матеріалів. Враховуючи виявлену кореляцію між фейковими та агресивними повідомленнями, запропоновано розширити вектор ознак класифікаційних моделей інформацією емоційного характеру. Зокрема, вектор ознак, одержаний за допомогою статистичного показника Term Frequency – Inverse Document Frequency (TF–IDF), був доповнений даними про наявність агресивної риторики у повідомленні та його обсяг, значенням тональності тексту та кількістю позитивних і негативних слів.

Встановлено, що інженерія ознак дає змогу підвищити точність виявлення фейкових новин на 2–6 % для моделей машинного навчання, навчених на вибірці повідомлень українською та російською мовами. Методи опорних векторів і логістичної регресії демонструють найкращі результати класифікації текстової інформації.

Ключові слова: комп'ютерний аналіз тексту, виявлення фейків, машинне навчання, інженерія ознак.

Стаття надійшла до редколегії 29.10.2023
Прийнята до друку 24.11.2023