Method for neural network cyberbullying detection in text content with visual analytic

Iurii Krak[1,2], Olena Sobko[3], Maryna Molchanova[3], Illia Tymofiiev[3], Olexander Mazurets[3] and Olexander Barmak[3]

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

Abstract

The paper proposed the method for neural network cyberbullying detection in text content with visual analytic, designed to explain the neural network's decisions regarding identified types of cyberbullying. Distinctive feature of method, setting it apart from existing approaches, is the use of three alternative visual representations, which enhance the interpretability of decisions made by deep learning models when detecting cyberbullying types in text messages. Using the trained BERT neural network model for multi-label classification, the method identifies various types of cyberbullying in input text samples, along with the percentage representation of each type. The trained model demonstrated high performance across macrometrics, achieving Accuracy of 0.956478, Precision of 0.963677, Recall of 0.956478, and F1-Score of 0.960019. These metrics confirm the model's effectiveness in detecting cyberbullying types within text content. The proposed method provides three modes of interpretation: color-coded visualizations, local word importance diagrams, and overall word importance diagrams. This approach facilitates a clear understanding of the textual features that influenced the AI's decisions regarding cyberbullying detection. The method has potential applications in educational platforms, social media, and content moderation systems to support victims and witnesses of cyberbullying through psychological assistance.

Keywords

cyberbullying, neural networks, interpretation of results, BERT, LIME

1. Introduction

The problem of cyberbullying is becoming more and more relevant over time due to the increase in the number of users of social networks, as well as the decrease in the lower age limit of such users. Thus, there is an increasing demand for systems for detecting cyberbullying in text content, for the implementation of which deep learning models are used [1]. Deep learning models are especially effective in cases of large amounts of data and complex tasks, such as natural language processing (NLP), computer vision and speech recognition. They allow you to automatically train a model to understand complex patterns in the data without the need for manual feature creation [2].

One of the important features of deep learning is its ability to perform multi-level data abstraction, where each layer of the neural network transforms the input data into a set of more abstract features, which allows deep learning models to solve complex tasks, including text analysis, voice recognition, and cyberbullying classification [3]. With the development of natural language processing technologies, in particular transformer-based models such as BERT, it has become possible to develop systems that effectively detect cyberbullying cases and also classify them into different types [4]. However, a high level of efficiency is often accompanied by difficulty in interpreting the results, which calls into question the use of such models in sensitive and socially important contexts such as cyberbullying.

Given the above, we can say that artificial intelligence has become an integral part of the use of modern technologies, which forces us to focus not only on the performance of systems, but also

CS&SE@SW 2024: 7th Workshop for Young Scientists in Computer Science & Software Engineering, December 27, 2024, Kryvyi Rih, Ukraine

yuri.krak@gmail.com (I. Krak); olenasobko.ua@gmail.com (O. Sobko); m.o.molchanova@gmail.com (M. Molchanova); ilia.tumofiev@gmail.com (I. Tymofiiev); exe.chong@gmail.com (O. Mazurets); alexander.barmak@gmail.com (O. Barmak)
0000-0002-8043-0785 (I. Krak); 0000-0001-5371-5788 (O. Sobko); 0000-0001-9810-936X (M. Molchanova); 0009-0006-4610-5889 (I. Tymofiiev); 0000-0002-8900-0650 (O. Mazurets); 0000-0003-0739-9678 (O. Barmak)

on the transparency of algorithms and their ability to provide understandable justifications for their conclusions [5]. That is why Explanatory Artificial Intelligence (XAI) is an important concept in the field of modern technologies, which involves the creation of artificial intelligence systems capable of providing understandable explanations of their decisions to users. Given this, the interpretation of the results of the model for detecting cyberbullying in text content is important to ensure transparency and user trust in the decisions provided by artificial intelligence [6].

Cyberbullying detection software solutions using explanatory AI principles play an important role in ensuring the safety of young people in the online environment. Their implementation will contribute to the achievement of the UNDP Sustainable Development Goals, in particular, such as ensuring healthy lives and promoting well-being for all ages SDG3, ensuring equal access to quality education and promoting lifelong learning opportunities SDG4, achieving gender equality and empowering all women and girls SDG5. In addition, such solutions will contribute to reducing inequalities within and between countries SDG10, as well as ensuring peace, justice and strong institutions SDG16 [7].

The purpose of the paper is to improve the explainability of decisions made by deep learning neural networks regarding types of cyberbullying detected by the neural network in text messages, using visual analytics.

The main contribution of the paper is the proposed method for neural network cyberbullying detection in text content with visual analytic, which will provide visual analytics regarding the decisions of the neural network model regarding types of cyberbullying detection.

2. Related work

The problem of detecting cyberbullying is important because of its significant negative influence on mental health, especially among young people and adolescents. Modern approaches to detecting cyberbullying are based on natural language processing methods that allow analyzing text content to detect and classify different types of cyberbullying [8]. In addition, a number of authors use the principles of explained artificial intelligence in the task of detecting cyberbullying in text content, which allows for the interpretation of the results obtained.

The article by Dobrojevic et al. [9] is devoted to the development of an approach for detecting cyberbullying. The article applies two approaches for converting text into numerical data for ML: TF-IDF and BERT. For classification, the XGBoost model was used, improved using hyperparametric optimization implemented using a modified Coyote Optimization Algorithm (COA). The best model was interpreted using the SHAP (Shapley Additive Explanations) technique, which allowed obtaining important conclusions about the behavioral patterns of users who commit cyberbullying.

Syfullah et al. [10] presents an approach to detecting cyberbullying in Bengali texts. The study combines existing datasets, uses feature engineering techniques, and applies machine and deep learning models. To distinguish between general and specific cyberbullying, two datasets were created: binary and multi-class classification. Text vectors were generated using TF-IDF and Word2Vec, and the models provided an accuracy of 75.33% for the binary set and 93.16% for the multi-class set. To explain the work of the models, LIME was used for local analysis of individual examples and SHAP for global interpretation, which allowed us to understand the contribution of each feature to the overall predictions of the model.

Gongane et al. [11] proposes a unified BiLSTM-LIME model for multi-class classification of cyberbullying content on the Twitter platform. The authors argue that the LIME technique provides a high level of explanation, highlighting the most relevant tokens that contributed to the model's decision.

The paper by Ashraf et al. [12] addresses the problem of detecting offensive content in Bengali texts. NLP methods were applied along with five different machine learning classifiers: Decision Tree, Random Forest, Multinomial Naïve Bayes, Support Vector Classifier, and Logistic Regression. The main goal of the research was to create an effective algorithm and explain the factors influencing its decision using explanatory artificial intelligence. TF-IDF with n-grams was used for text representation. Optimal hyperparameters for the models were determined using the Grid Search Cross Validation technique.

The Logistic Regression model showed the best results with an accuracy of 85.57%. To ensure the transparency of the model and explain its operation, the LIME method was used.

Aggarwal and Mahajan [13] proposed a new approach to detect and classify cyberbullying in social media texts using an ensemble of BERT and SVM with grid search for multi-class classification. Comparison with other machine learning and deep learning models showed that the proposed model achieves 90% accuracy on test data, outperforming others. The SHAP technique was used to interpret the predictions.

Pawar et al. [14] investigates the problem of cyberbullying on social platforms, in particular on Twitter. The authors propose a model capable of classifying tweets into two categories: bullying or non-bullying. In addition to detecting cyberbullying, the study focuses on ensuring the interpretation of the classification results. For this purpose, the LIME (Local Interpretable Model-agnostic Explanations) method is used, which provides local explanations of the model's operation.

Nuthalapati et al. [15] investigates the problem of cyberbullying detection. Among the tested models such as Random Forest, XgBoost, Naive Bayes, SVM, CNN, RNN and BERT, the BERT model showed the highest results, achieving 88.8% accuracy in the binary classification task and 86.6% accuracy in the multi-label classification task.

Perera and Fernando [6] proposed a new theory for detecting cyberbullying, in which the Support Vector Machine, Naive Bayes and Logistic Regression models were tested in combination with various natural language processing methods. The authors note that the accuracy of detecting cyberbullying is increased by using sentiment analysis, N-gram analysis, as well as non-traditional feature extraction methods such as TF-IDF and profanity detection. The combined approach allows achieving a detection accuracy of 75.17%.

As an interpretative model for multi-label classification, methods such as are often used [16]:

- LIME, which generates local explanations for each prediction, showing which words had the greatest influence on the result [17];
- SHAP, which is based on game theory and calculates the contribution of each word to the prediction, taking into account the interaction between features [18];
- Transformers Interpret, which is an interpretation library specifically designed to work with models based on transformer neural networks, such as BERT, GPT, RoBERTa, and other models from the Hugging Face library [19];
- Attention-based methods that allow analyzing the attention weights of transformers (e.g., in the BERT model) to understand the importance of individual words or phrases in the model's decision-making [20].

In [1] a method for detecting and classifying cyberbullying was developed, which provided a minimum classification accuracy of 96%, but the explainability of the obtained results was not ensured.

Considering the conducted research of recent publications, most authors, who investigate the interpretation of the results of models for detecting cyberbullying, although they use visual analytics tools, are usually limited to superficial analysis or one form of information presentation. Such an approach does not always allow to reveal the full potential of the interpretation tools, since different visualization methods can provide additional context or emphasize different aspects of the models.

Therefore, an approach based on following representations of results interpretation of cyberbullying detection is proposed:

- by color palette;
- by local word importance diagrams;
- by general word importance diagrams.

This approach allows to improve the quality of the analysis, since different visualizations provide an opportunity to understand the behavior of the model and its solution in depth, which is critically important for complex tasks, such as cyberbullying detection.

3. Method for neural network cyberbullying detection in text content with visual analytic

Method for neural network cyberbullying detection in text content with visual analytic involves creating a visual explanation by providing three alternative views of the visual interpretation of the results of detecting types of cyberbullying, namely by color palette, by diagrams of local word importance, interpretation of results by diagrams of general word importance regarding detected types of cyberbullying in text content. The scheme of method for neural network cyberbullying detection in text content with visual analytic is presented in figure 1.

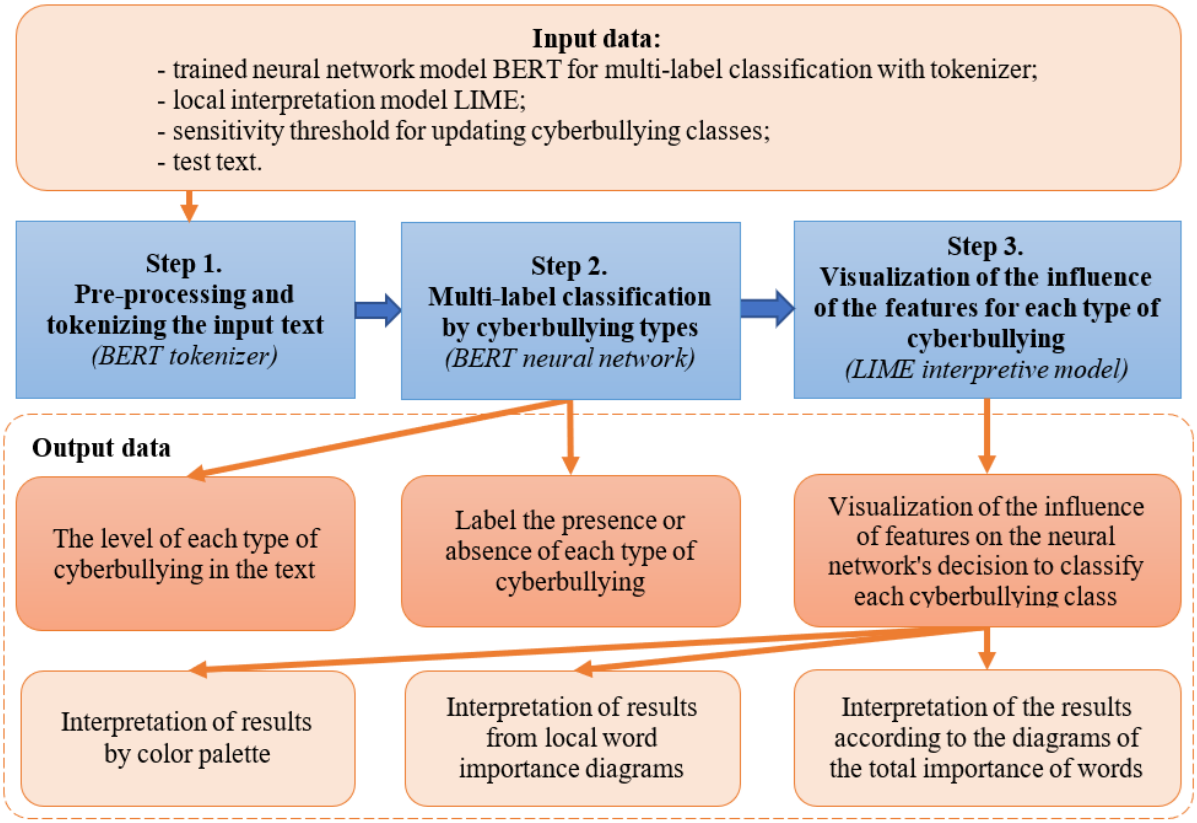

Figure 1: Schema of method for neural network cyberbullying detection in text content with visual analytic.

The input data of the scheme of method for neural network cyberbullying detection in text content with visual analytic shown in figure 1 is a trained model for multi-label classification, which is able to recognize different types of cyberbullying, such as age, ethnicity, gender, religion and a separate generalized type containing other types of cyberbullying [1]. An interpretive model is used, which allows explaining the influence of individual words or phrases on the classification result. The list of cyberbullying classes includes the types of cyberbullying, according to which the model classifies and interprets the results of the interpretive model. Also, the input data includes text, which is analyzed for signs of different types of cyberbullying and the results are interpreted.

In the first step, the input text is transformed into a sequence of tokens using a tokenizer, which breaks the text into individual elements (words or parts of words). It is proposed to use BERT as an input model for multi-label classification. For it and similar transformers, the tokenizer transforms the text input into numerical sequences that the model can work with [21].

In the second step, the BERT model trained on multi-label classification predicts the probability of the text belonging to each of the possible cyberbullying classes. Thus, the model determines whether the text contains signs of certain types of cyberbullying (age, ethnicity, gender, religion, and other types of cyberbullying), providing a percentage probability of the presence of each type of cyberbullying.

In the third step, the classification results are explained and visualized. Using an interpretive model, it is shown which words or phrases had the greatest influence on the classification of the text as a specific type of cyberbullying, which helps to understand which parts of the text contributed to the identification of signs of a specific type of cyberbullying [22]. Also at this step, three alternative representations of the visual interpretation of the results of detecting types of cyberbullying are formed, namely: by the color palette, by diagrams of local word importance, interpretation of the results by diagrams of the general importance of words regarding the detected types of cyberbullying in text content. The developed method proposes to use the LIME machine learning model of interpretation of predictions for visual interpretation of the results of detecting types of cyberbullying, since this model is designed to explain local predictions of complex neural network models. In particular, LIME allows you to understand which parts of the input data influenced the decision-making of the neural network.

The output data is the level of manifestation of each type of cyberbullying in the text, which is determined in the form of a neural network probabilistic assessment. The method also provides visualization of the influence of features on the decision to attribute the text to a specific class of cyberbullying by graphically representing the text in three alternative representations: a color palette, where important words are highlighted according to their significance for each class; diagrams of local word importance; interpretation of the results in diagrams of the overall importance of words in relation to the detected types of cyberbullying in the text content.

Therefore, the given method for neural network cyberbullying detection in text content with visual analytic will allow to obtain not only the results regarding the detected types of cyberbullying in the text sample and the level of each type of cyberbullying in the text, but also will provide alternative visual representations of the interpretation of the results of cyberbullying detection.

4. Experiment, results and discussion

To train the BERT model, which is used in step 2 of the method for neural network cyberbullying detection in text content with visual analytic (figure 1), the “Cyberbullying Classification” dataset [23] was used, which contains text messages with labels indicating whether each message belongs to one of the classes: Age, Ethnicity, Gender, Religion, Other type of cyberbullying, Not cyberbullying. Detailed statistics on the number of records are shown in figure 2.

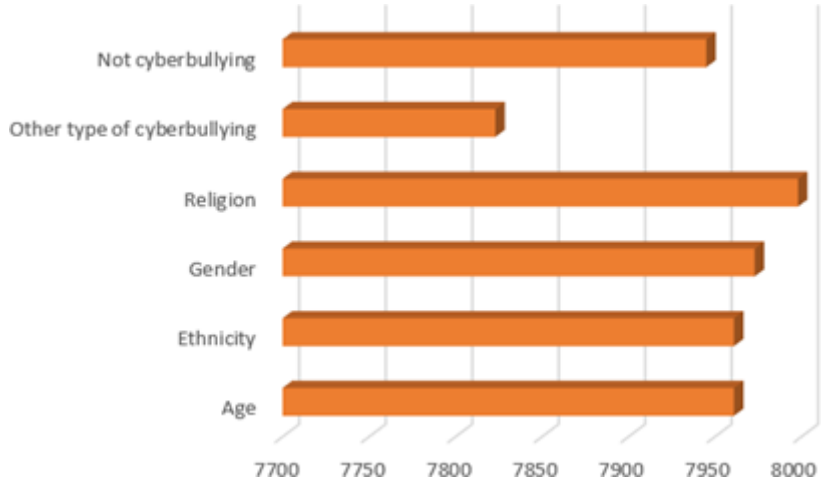

Figure 2: Statistics of records number in dataset classes for detecting cyberbullying.

The class “Not cyberbullying” was not used to train the BERT model for multi-label classification, so it was removed from the dataset before training. And the class “Other type of cyberbullying” was augmented with synthetic samples using the SMOTE-balancing technique [24]. By pre-processing the “Cyberbullying Classification” dataset, a balanced training sample was obtained, which was used to train the BERT model for the task of multi-label classification of cyberbullying types in text content.

To study the effectiveness of the method for neural network cyberbullying detection in text content with visual analytic, software was developed in the form of a web application, and the Google Colab environment was used to train the BERT model [25]. The BERT neural network model was trained on such types of cyberbullying as age, gender, religion, ethnicity, and a separate type – other cyberbullying. Figure 3 shows the confusion matrices for each type of cyberbullying.

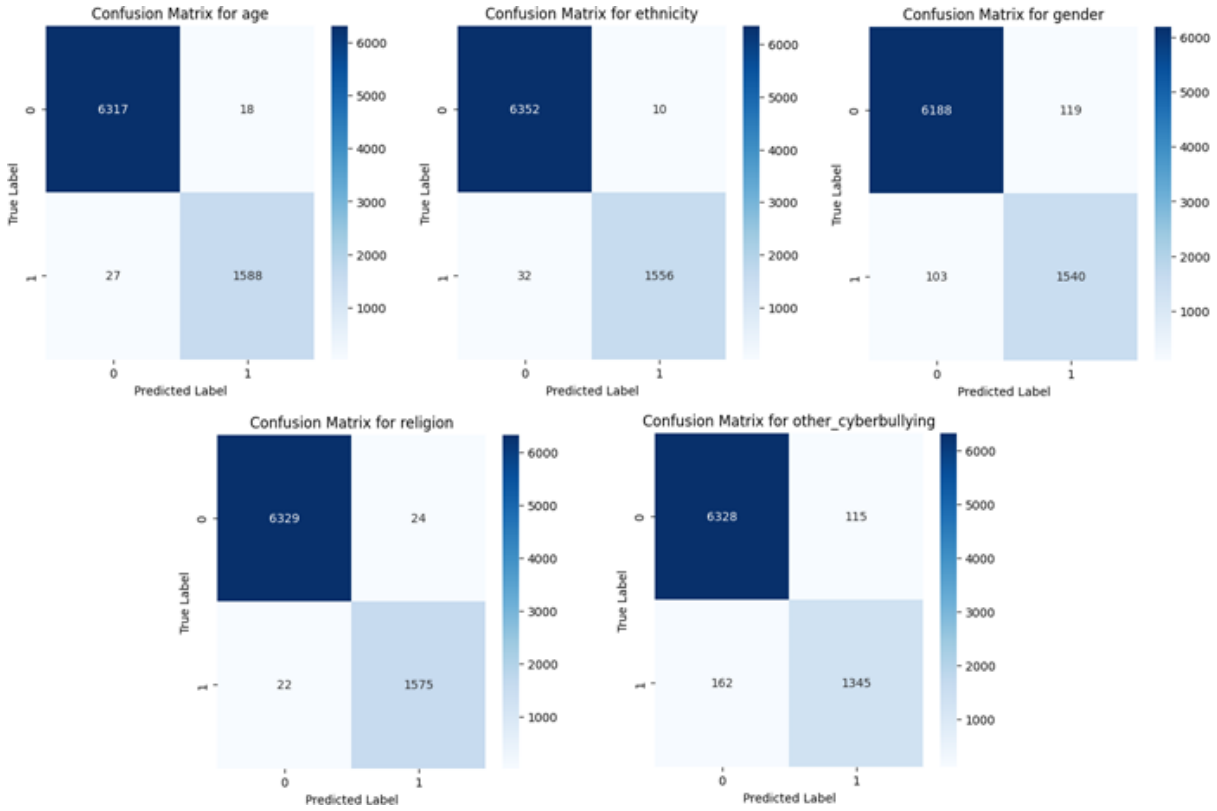

Figure 3: Confusion matrices for cyberbullying types.

The macrometric indicators of the trained BERT model for multi-label classification of cyberbullying types received the values of Accuracy 0.956478, Precision 0.963677, Recall 0.956478, F1 Score 0.960019, which indicates the high ability of the model to detect cyberbullying types in text content. Compared with known studies, the proposed method for neural network cyberbullying detection in text content with visual analytic showed a higher accuracy rate. Compared to [10], where 93.16% accuracy was obtained for multi-class set, the performance is improved by 2.49%. Compared to [13], where 90% accuracy was obtained for multi-class classification, the performance is improved by 5.65%. Compared to [15] for multi-label classification, where 86.6% accuracy was obtained, the performance is improved by 9.05%.

For the study of method for neural network cyberbullying detection in text content with visual analytic, the following text sample was used: *"Your God has no place here. Stick to your country and stop dragging your outdated traditions and religions into ours"*. The BERT model identified the following levels of manifestation of each type of cyberbullying, expressed in probabilities:

- age cyberbullying: 0.06%;
- ethnic cyber bullying: 0.08%;
- gender cyberbullying: 0.10%;
- religious cyberbullying: 99.86%.

To interpret the results of the BERT model for multi-label classification of cyberbullying types in a text sample, the LIME model was applied and three alternative representations of the visual interpretation of the results of detecting cyberbullying types were obtained, namely by color palette, by diagrams

of local word importance, and interpretation of the results by diagrams of general word importance regarding the detected types of cyberbullying in text content.

The results of the visual interpretation of the detected types of cyberbullying according to the color palette using the absolute value of the weights and for the interpretation of the results of the detection of the types of cyberbullying taking into account the positive and negative influences are presented in figure 4.

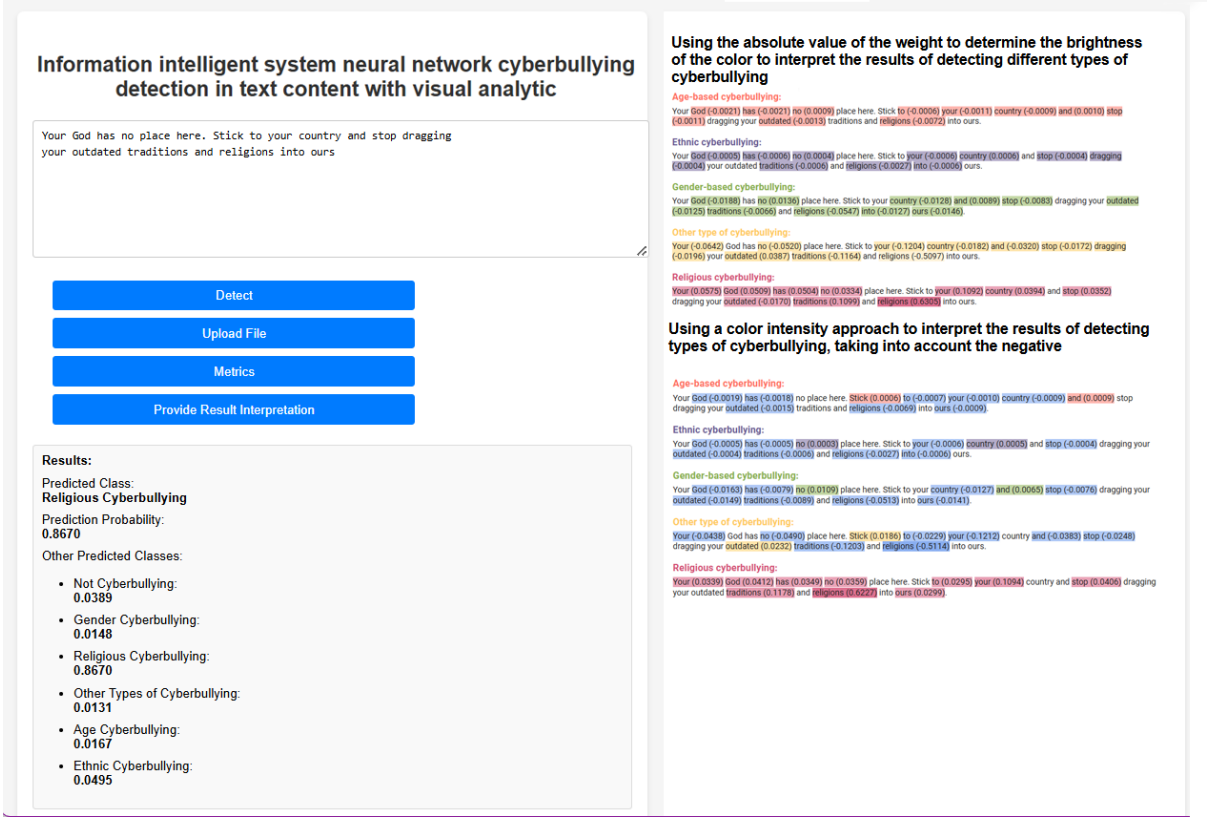

Figure 4: Results of visual interpretation of detected cyberbullying types by color palette.

To interpret the decisions made by the BERT model using the absolute value of the weights, words are highlighted in colors – the brightest color means the highest value of the word's weight, i.e. this word had the greatest influence, the lightest – the least.

As can be seen from figure 4, words with positive and negative values are highlighted in the same bright color. In this type of visual interpretation, the absolute value of the weight is used to determine the brightness of the color, due to which negative and positive values have the same brightness. Negative values of the weights indicate that the word decreases the probability of a particular class, while positive values increase the probability of this class and have the same effect on the decision made by the model. And the weight value, without taking into account the sign next to it, indicates how strong it was.

In the case of LIME, it is important to show not only how strong the influence of a word is, but also whether this influence is positive (increases the probability) or negative (decreases the probability). Therefore, an approach to changing the brightness is implemented so that negative values are less bright and uses a different color shade for negative and positive values.

The choice of separate color palettes for positive and negative values in the visualization of LIME interpretations is appropriate for several reasons that derive from the principles of information perception and analysis of the results of machine learning models. Negative weights, by their nature, indicate a decrease in the probability of a certain class, while positive ones indicate an increase in it, so using the same visual characteristics for these two types of influence can lead to erroneous interpretation of the results if additional types of visual interpretation are not provided, because values of the same

intensity but opposite sign can appear equally important, although their role is fundamentally different.

Additionally, diagrams were created for graphical interpretation of the influence of individual words of the text on the probability of attributing this text to specific type of cyberbullying (figure 5).

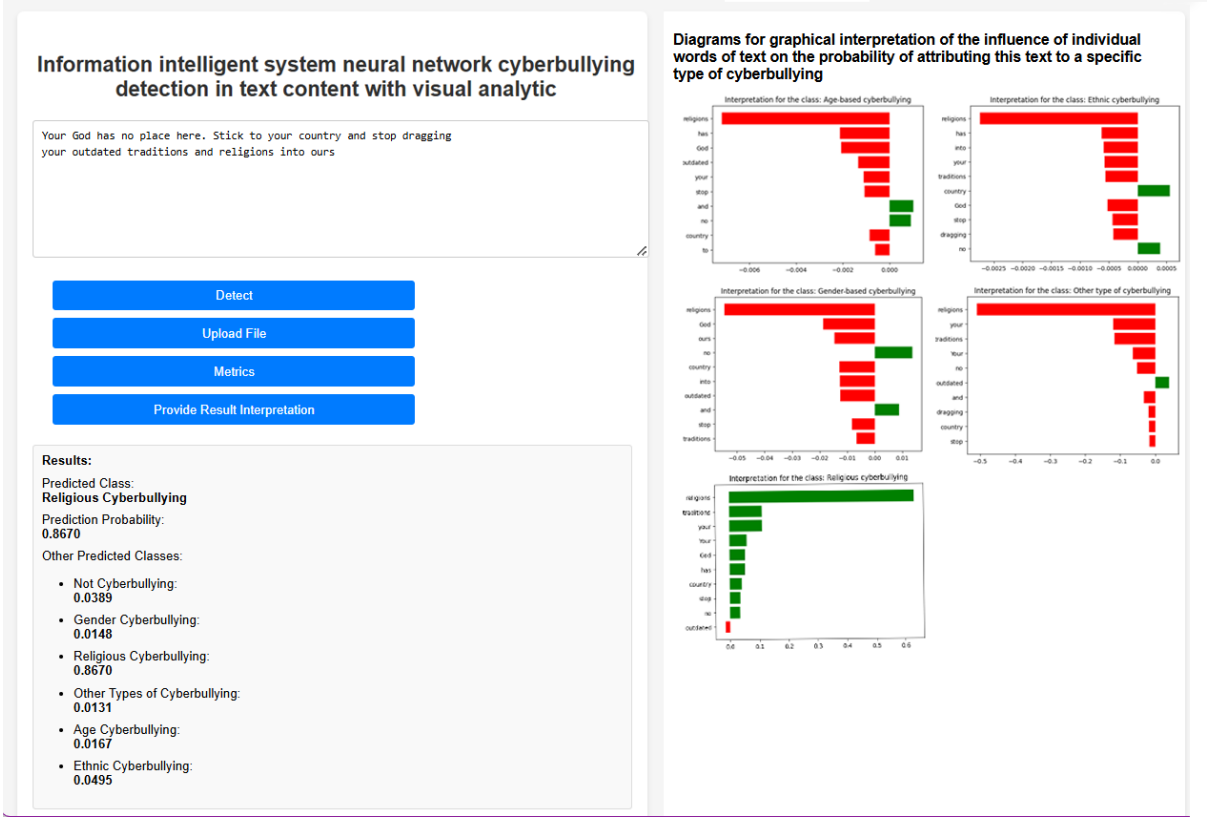

Figure 5: Diagrams for influence graphical interpretation of individual words of text on the probability of attributing this text to specific cyberbullying type.

The diagrams illustrate how the model estimates the weight of each word in the text, depending on its contribution to the decision made. The influence of words is represented as horizontal bars, the length of which corresponds to the magnitude of the influence (weight), and the color – the direction of this influence. Red bars reflect the negative influence of words, i.e. reducing the probability of assigning the text to the selected class, while green bars reflect the positive influence, increasing this probability. The magnitude of the influence is measured in numbers, and these values are represented on the horizontal axis of the diagrams.

The average importance value of each word for all classes was also calculated, which gives an idea of the overall influence of each word regardless of the specific type of cyberbullying. The calculated values are visualized through the corresponding diagram (figure 6).

Calculating the overall influence of words on the model results for all types of cyberbullying is also important for interpreting the model's performance and understanding its decisions. The analysis is performed by aggregating the word weights that the model estimates for each class [26]. The weight modulus is used, i.e. an absolute value that indicates the intensity of the word's influence regardless of its positive or negative contribution. This approach allows us to identify words that the model considers important regardless of the specific type of cyberbullying. For example, words that reflect different types of cyberbullying may have high weights for several classes. If a word has a high overall influence, this may indicate its universal role in the context of cyberbullying. For example, words that indicate ethnicity or religion may have high influences for several classes, such as "ethnic cyberbullying" and "religious cyberbullying", which may indicate potential cross-modality of the features that the model uses to make decisions. If words influence only one class, this emphasizes their specificity, which may indicate unique speech patterns for this type of cyberbullying.

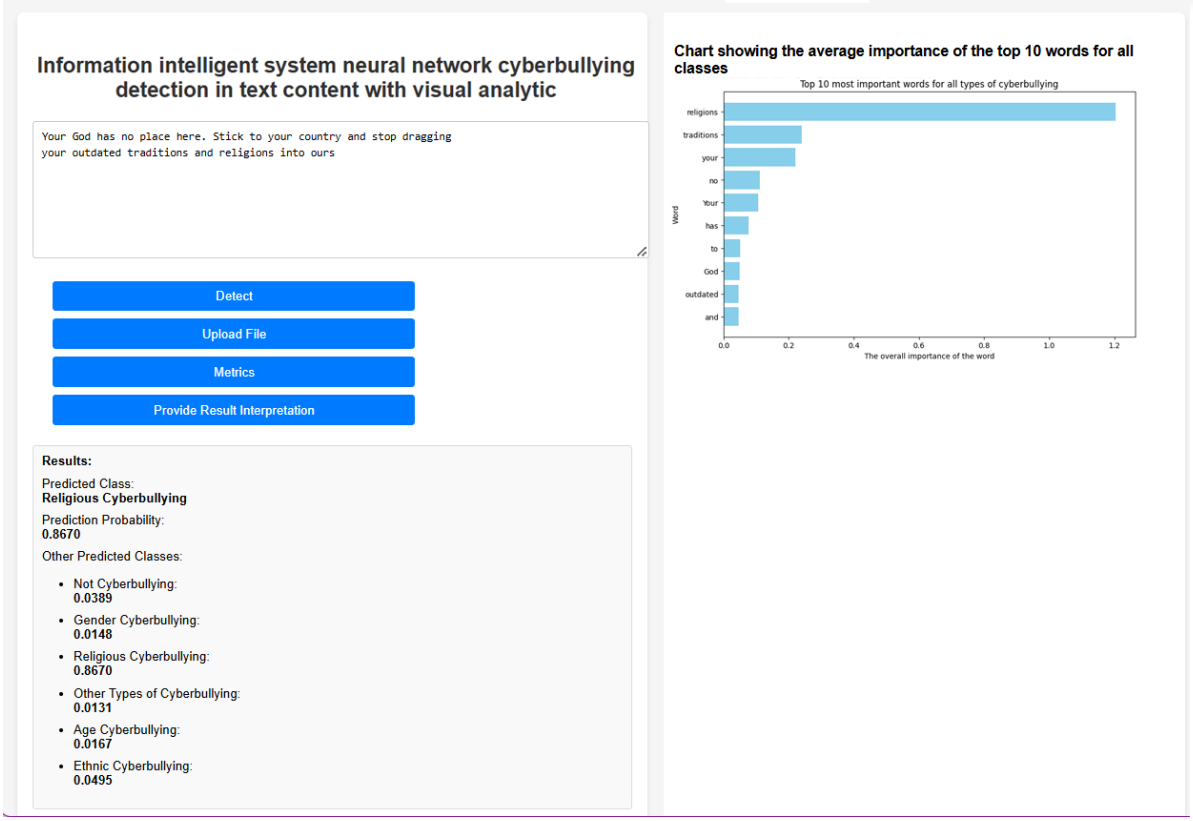

Figure 6: Diagram showing the average importance of the top 10 words for all classes.

Thus, the proposed visual interpretations of the results of detecting cyberbullying in text content allow a person to comprehensively assess whether the model uses relevant features for decision-making, or whether its behavior may be due to random or irrelevant factors. For example, if the text contains words that have no semantic connection with age-related cyberbullying, but have a high influence, this may indicate an error or bias in the model.

5. Conclusion

The article proposes a method for neural network cyberbullying detection in text content with visual analytic, which is designed to explain the decisions of the neural network model regarding the types of cyberbullying identified in text content. The method is original in that it performs visual analytics of the results for each detected type of cyberbullying separately, which is achieved by using a multi-label classifier of the transformer neural network architecture and a local machine learning interpretive model. A distinctive feature of the method from analogues is that it implements visual analytics for decisions made by the neural network in three alternative representations, which provides an increase in the level of explainability of decisions made by deep learning neural networks regarding the types of cyberbullying in text messages detected by the neural network.

By using the trained BERT neural network model for multi-label classification of cyberbullying types in the input text sample, different types of cyberbullying are detected with the percentage of each of them. The trained model demonstrated high results in macrometrics: Accuracy 0.956478, Precision 0.963677, Recall 0.956478, F1-measure 0.960019. These indicators indicate the effectiveness of the model in identifying types of cyberbullying in text content.

According to the developed method, for the visual interpretation of the results of cyberbullying detection, an approach based on the use of a machine learning model for the local interpretability of LIME models was used, which allows visualizing the influence of the use of individual words on the

model's decision regarding whether the text belongs to different types of cyberbullying. The developed method provides three ways of interpreting the results of cyberbullying detection: interpretation of results by color palette, interpretation of results by diagrams of local word importance, interpretation of results by diagrams of general word importance. Interpreting the results using a color palette involves using the absolute value of the weights to determine the brightness of the color, where the brightest color indicates the greatest influence of the word on the model's decision, and the least bright color indicates the least influence, regardless of whether this influence was positive or negative. Interpretation of results from diagrams of local word importance is provided by constructing diagrams of the influence of individual words of the text on the probability of attributing this text to a specific type of cyberbullying, which allows to see how the model assesses the weight of each word in the text depending on its contribution to the model's decision. Interpretation of the results from the overall word importance diagrams is provided by forming a set of 10 words that the model considers important regardless of the specific type of cyberbullying.

The results of the experiments showed that the created method provides interpretation of decisions regarding the results of neural network detection of cyberbullying at a level sufficient for a human to understand the text features that influenced the decision-making of artificial intelligence regarding the detection of types of cyberbullying. The proposed method for interpreting the results of cyberbullying detection in text content belongs to the category of visual analytics tools for artificial intelligence solutions, the creation of which is a practical necessity to ensure ethics, transparency, and trust in such artificial intelligence systems in society, especially regarding such sensitive topics as cyberbullying detection. Accordingly, the study emphasizes the importance of not only the accuracy of models, but also their explainability, which is a key factor in building trust in artificial intelligence systems.

Developed method for visual interpretation of cyberbullying detection results complies with goals SDG3, SDG4, SDG5, SDG10 and SDG16, and can be implemented in educational platforms, social media platforms, in moderation systems to provide psychological assistance to victims or witnesses of cyberbullying.

Prospects for proposed method further research include adapting the developed method to work with texts in other languages, conducting experiments with users to assess the impact of visual analytics on human decision-making, in particular, in the work of moderators or psychologists, and creating additional ways of presenting interpretation and testing alternative interpretative models.

Declaration on Generative AI: The authors have not employed any Generative AI tools.

References

[1] I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, R. Bahrii, O. Sobko, O. Barmak, Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network, in: V. Vysotska, Y. Burov (Eds.), Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems. Volume III: Intelligent Systems Workshop, Lviv, Ukraine, April 12-13, 2024, volume 3688 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 16–28. URL: https://ceur-ws.org/Vol-3688/paper2.pdf. doi:10.31110/COLINS/2024-3/002.

[2] T. Talaei Khoei, H. Ould Slimane, N. Kaabouch, Deep learning: Systematic review, models, challenges, and research directions, Neural Computing and Applications 35 (2023) 23103–23124. doi:10.1007/s00521-023-08957-4.

[3] N. Shlezinger, J. Whang, Y. C. Eldar, A. G. Dimakis, Model-based deep learning, Proceedings of the IEEE 111 (2023) 465–499. doi:10.1109/JPROC.2023.3247480.

[4] M. Sen, J. Masih, R. Rajasekaran, From Tweets to Insights: BERT-Enhanced Models for Cyberbullying Detection, in: 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), IEEE, 2024, pp. 1289–1293. doi:10.1109/ICETSIS61505.2024.10459672.

[5] Y. V. Krak, O. V. Barmak, O. V. Mazurets, The practice investigation of the information technology

efficiency for automated definition of terms in the semantic content of educational materials, Problems in programming (2016) 237–245. doi:10.15407/pp2016.02-03.237.

[6] A. Perera, P. Fernando, Cyberbullying detection system on social media using supervised machine learning, Procedia Computer Science 239 (2024) 506–516. doi:10.1016/j.procs.2024.06.200.

[7] R. Russell-Bennett, M. S. Rosenbaum, R. P. Fisk, M. M. Raciti, SDG editorial: improving life on planet earth – a call to action for service research to achieve the sustainable development goals (SDGs), Journal of Services Marketing 38 (2024) 145–152. doi:10.1108/JSM-11-2023-0425.

[8] M. F. Gan, H. N. Chua, M. B. Jasser, R. T. Wong, Categorization of Cyberbullying based on Intentional Dimension, in: 2024 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), IEEE, 2024, pp. 285–290. doi:10.1109/I2CACIS61270.2024.10649619.

[9] M. Dobrojevic, L. Jovanovic, L. Babic, M. Cajic, T. Zivkovic, M. Zivkovic, S. Muthusamy, M. Antonijevic, N. Bacanin, Cyberbullying Sexism Harassment Identification by Metaheurustics-Tuned eXtreme Gradient Boosting, Computers, Materials & Continua 80 (2024). doi:10.32604/cmc.2024.054459.

[10] M. K. Syfullah, R. Amin, M. Manirujjaman, M. M. Hasan, M. Ragib Abid, S. Ahmed, Detecting and Understanding Cyberbullying in Bengali: An Explainable AI Approach, in: 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS), 2024, pp. 1–7. doi:10.1109/COMPAS60761.2024.10797170.

[11] V. U. Gongane, M. V. Munot, A. Anuse, Explainable AI for Reliable Detection of Cyberbullying, in: 2023 IEEE Pune Section International Conference (PuneCon), 2023, pp. 1–6. doi:10.1109/PuneCon58714.2023.10450132.

[12] K. Ashraf, M. H. Hosen, S. Asgar, M. T. Islam, S. Nawar, Analyzing Abusive Bangla Comments on Social Media: NLP & Explainable AI, in: 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS), 2024, pp. 1–6. doi:10.1109/iCACCESS61735.2024.10499547.

[13] P. Aggarwal, R. Mahajan, Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification, Journal of Information Systems and Informatics 6 (2024) 607–623. doi:10.51519/journalisi.v6i2.692.

[14] V. Pawar, D. V. Jose, A. Patil, Explainable AI Method for Cyber bullying Detection, in: 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), 2022, pp. 1–4. doi:10.1109/ICMNWC56175.2022.10031652.

[15] P. Nuthalapati, S. A. Abbaraju, G. H. Varma, S. Biswas, Cyberbullying Detection: A Comparative Study of Classification Algorithms, 2024. doi:10.22541/au.170664263.38254624/v1.

[16] S. Kiefer, CaSE: Explaining Text Classifications by Fusion of Local Surrogate Explanation Models with Contextual and Semantic Knowledge, Information Fusion 77 (2022) 184–195. doi:10.1016/j.inffus.2021.07.014.

[17] R. S. Tiwari, Hate speech detection using LSTM and explanation by LIME (local interpretable model-agnostic explanations), in: Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications, Elsevier, 2024, pp. 93–110. doi:10.1016/B978-0-443-22009-8.00005-7.

[18] X. Men, V. Y. Mariano, Explainable Fake News Detection Based on BERT and SHAP Applied to COVID-19, International Journal of Modern Education and Computer Science (IJMECS) 16 (2024) 11–22. doi:10.5815/ijmecs.2024.01.02.

[19] A. Bennetot, I. Donadello, A. El Qadi El Haouari, M. Dragoni, T. Frossard, B. Wagner, A. Sarranti, S. Tulli, M. Trocan, R. Chatila, A. Holzinger, A. Davila Garcez, N. Díaz-Rodríguez, A Practical Tutorial on Explainable AI Techniques, ACM Comput. Surv. 57 (2024). doi:10.1145/3670685.

[20] J. J. van Dieten, Attention Mechanisms in Natural Language Processing, B.S. thesis, University of Twente, 2024. URL: http://essay.utwente.nl/98223/.

[21] S. Alissa, M. Wald, Text simplification using transformer and BERT, Computers, Materials & Continua 75 (2023) 3479–3495. doi:10.32604/cmc.2023.033647.

[22] O. Kovalchuk, V. Slobodzian, O. Sobko, M. Molchanova, O. Mazurets, O. Barmak, I. Krak, N. Savina, Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets, in:

S. Babichev, V. Lytvynenko (Eds.), International Scientific Conference "Intellectual Systems of Decision Making and Problem of Computational Intelligence", volume 149 of *Lecture Notes on Data Engineering and Communications Technologies*, Springer, 2022, pp. 591–607. doi:10.1007/978-3-031-16203-9_33.

[23] Kaggle, Google: Cyberbullying Classification Dataset, 2024. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification.

[24] E. Manziuk, I. Krak, O. Barmak, O. Mazurets, V. A. Kuznetsov, O. Pylypiak, Structural Alignment Method of Conceptual Categories of Ontology and Formalized Domain, in: D. Chumachenko, O. Sokolov, N. Kryvinska, S. Yakovlev (Eds.), Proceedings of the International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2021) 2021, Kharkiv, Ukraine, September 20-21, 2021, volume 3003 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 11–22. URL: https://ceur-ws.org/Vol-3003/paper2.pdf.

[25] Colab, Google: Colab, 2025. URL: https://colab.research.google.com/.

[26] O. V. Barmak, O. V. Mazurets, I. V. Krak, A. I. Kulias, L. E. Azarova, K. Gromaszek, S. Smailova, Information technology for creation of semantic structure of educational materials, Proceedings of SPIE - The International Society for Optical Engineering 11176 (2019) 1117623. doi:10.1117/12.2537064.