SECTION: INFORMATION TECHNOLOGY AND CYBERSECURITY

INFORMATION SYSTEM FOR DETECTING ABUSIVE SPEECH IN AUDIO CONTENT BY MEANS OF NATURAL LANGUAGE PROCESSING

Viacheslav Nazarov
Grade 9-B student
Khmelnytskyi Lyceum №17,
Khmelnytskyi, Ukraine
uuuvich228@gmail.com
Maryna Molchanova
teacher of Computer Science Department
Khmelnytskyi National University,
momolchanova@gmail.com

Detecting abusive speech in text and audio content is an urgent task, as it not only contributes to the creation of a safe and healthy environment for communication, especially in online platforms, but also contributes to the fight against harassment and discrimination, since audio messages can contain offensive language and can cause harm to listeners [1]. It also allows you to quickly react to the unacceptable behavior of others, increasing the quality of communications and reducing the risk of spreading toxic content. Detecting abusive speech in audio content is important for legal and ethical compliance, as some of it may be not only objectionable, but also illegal.

Abusive content is any type of content that contains offensive, harmful, objectionable or offensive elements that may cause harm to others. In a scientific context, abusive content is often analyzed through the prism of its impact on individuals and society, as well as through the mechanisms of its distribution and detection [2].

To detect abusive speech in Ukrainian-language text and audio content, each component indicating the presence of abusive manifestations should be determined. As already mentioned, abuse is effectively detected by the presence of the following signs in the text:

- abusive speech (use of abusive words);
- negative emotional tone.

The general approach to detecting abusive speech in audio content is shown in Figure 1.

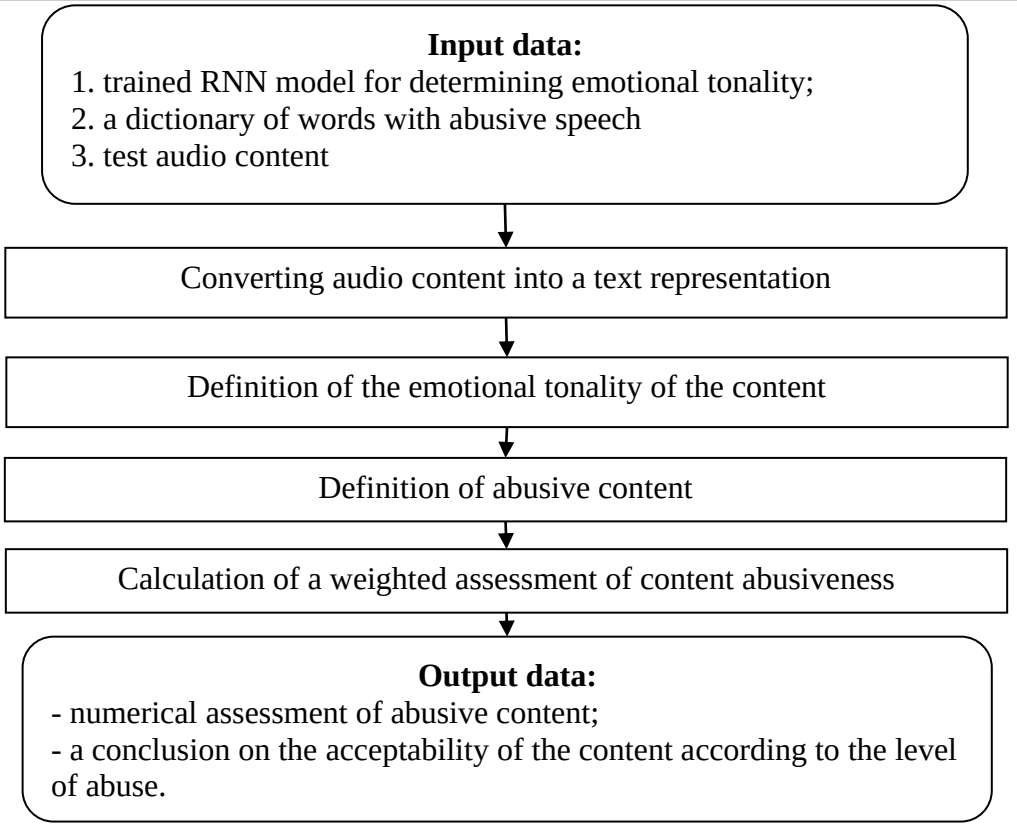

Figure 1. The sequence of actions in determining abusive manifestations in audio content

Based on the above-described approach, an appropriate information technology for the detection of abusive speech is proposed, which is designed to receive input information in the form of a training set of data for a neural network model for determining emotional tonality [3] and test audio content of the output data in the form of a numerical assessment of the abusiveness of the content and a conclusion regarding the acceptability of content according to the level of abuse. The general scheme of information technology for detecting abusive speech in Ukrainian-language audio content is shown in Figure 2.

The input data of information technology is a training set of data for a neural network model of emotional tonality determination and test audio content for analysis.

On the basis of the training data set, a recurrent neural network with an LSTM layer is trained according to the selected parameters of the batch size and the parameter of the number of training epochs. Training sets have a text format. The training result is a trained neural network model that will be used for analysis.

The test audio content is converted into a text representation for further analysis. In the future, detection of abusive content will be based on text.

The calculation of the weighted assessment of the abusiveness of the content is carried out according to two parameters, namely the determination of the abusive content content and the determination of the emotional tonality of the content.

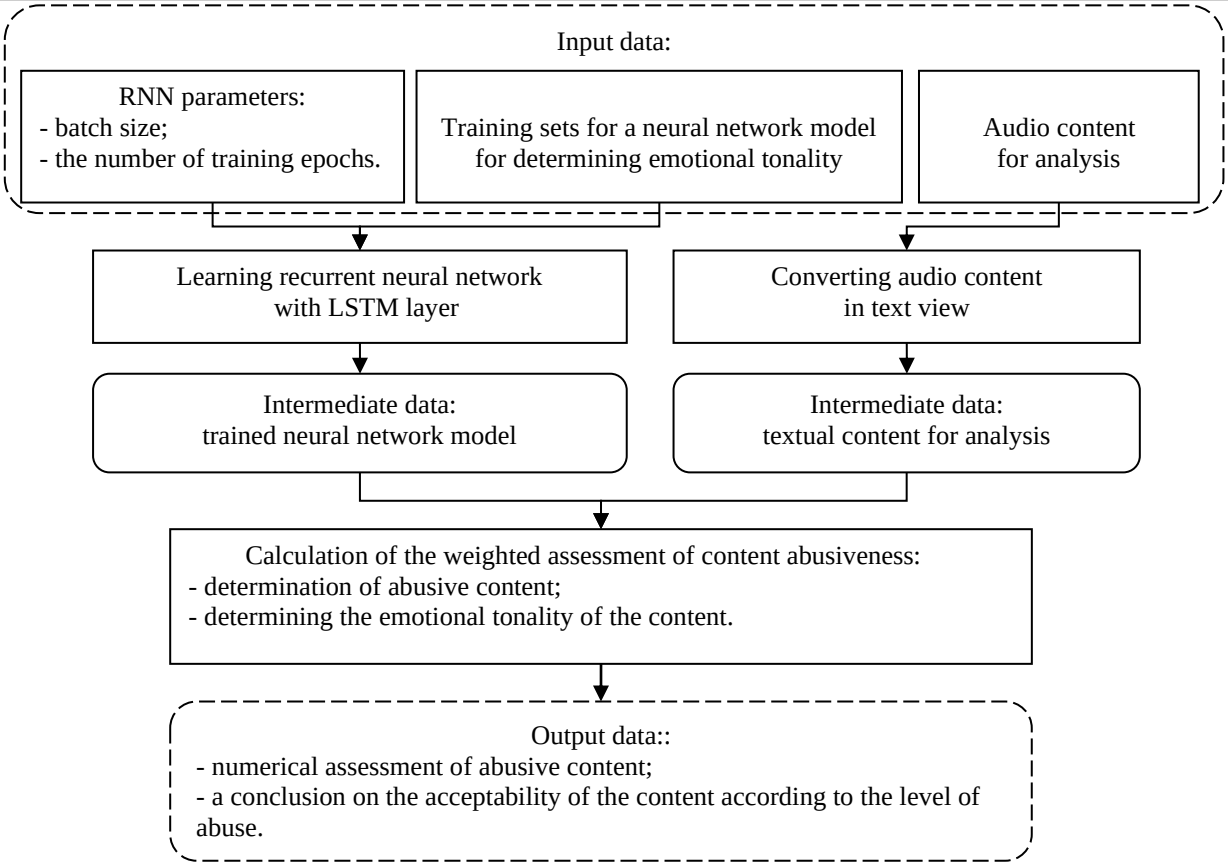

Figure 2. The general scheme of information technology for detecting abusive speech in Ukrainian-language audio content

The output data of the information technology is a numerical assessment of the abusiveness of the content and a conclusion on the acceptability of the content according to the level of abuse.

To study the effectiveness of information technology for detecting abusive speech in Ukrainian-language audio content, appropriate software was developed. The Python programming language and the PyCharm programming environment were used to implement the intelligent system. The developed software includes a software module for training recurrent neural network models and further saving of trained instances, and a software module for detecting abusive manifestations in Ukrainian-language audio content using trained neural network models.

The module for learning recurrent neural network models does not have a graphical interface and is intended exclusively as an auxiliary module for an intelligent system that uses a trained neural network.

The interface of the information system for detecting abusive speech is shown in Figure 3. The abusive content detection module allows you to detect abusive audio content based on the input data of the threshold of sensitivity to abusive content, the selected model of the trained neural network and the test audio content, which can be either dictated or downloaded from a file.

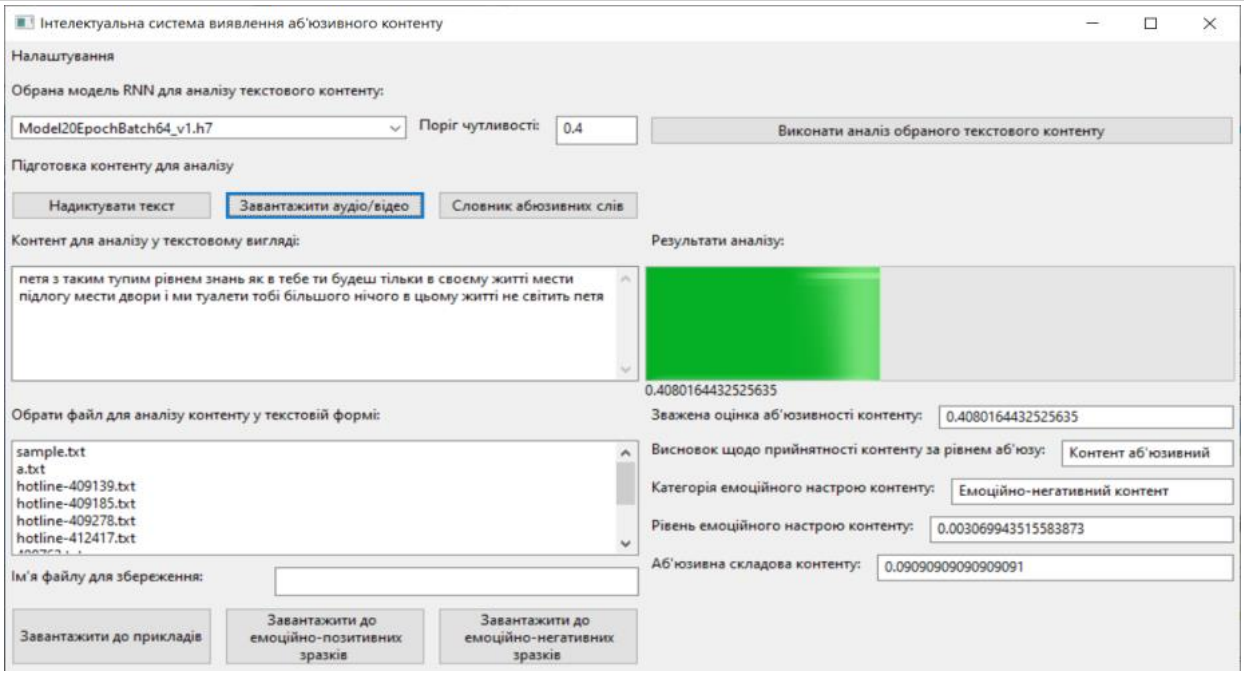

Figure 3. The interface of the intelligent system for detecting abusive content

The developed test software modules make it possible to conduct an applied study of the information technology of detecting abusive speech.

The developed information system has limitations. Since the recurrent neural network is trained on a short text data set (up to 200 words, the average length of the test examples is 17 words), the system is less efficient at identifying texts that are longer than 200 words. To increase efficiency in such cases, it is necessary to expand the training sets with longer text data and retrain the neural network.

References

1. Amrit Ch. Identifying child abuse through text mining and machine learning. Expert systems with applications 88, 2017. P 402-418. URL: https://www.sciencedirect.com/science/article/abs/pii/S0957417417304529

2. Nobata Ch. Abusive language detection in online user content. Proceedings of the 25th international conference on world wide web, 2016. URL: https://dl.acm.org/doi/10.1145/2872427.2883062

3. Молчанова М.О., Залуцька О.О., Бармак О.В. Метод інтелектуального аналізу тональності текстів. Матеріали XII Всеукраїнської науково-практичної конференції «Глушковські читання». Київ – 2023. с. 113-116.