Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network

Iurii Krak[1,2], Olha Zalutska[3], Maryna Molchanova[3], Olexander Mazurets[3], Ruslan Bahrii[3], Olena Sobko[3] and Olexander Barmak[3]

[1] *Taras Shevchenko National University of Kyiv, Ukraine*
[2] *Glushkov Institute of Cybernetics of NAS of Ukraine, Kyiv, Ukraine*
[3] *Khmelnytskyi National University, Khmelnytskyi, 11, Institutes str., 29016, Ukraine Country*

Abstract

The paper is devoted to the creation and approbation of a method for determining the level of social acceptability of textual Ukrainian-language content, which will be able to determine the level of detection of offensive speech for the Ukrainian language using a recurrent neural network based on the entered textual information. The method of detecting offensive speech involves the use of a combined approach based on the use of a dictionary of offensive words and a neural network approach to determine the sentiment tone of the message. The proposed method will allow to assess the level of social acceptability of Ukrainian-language Internet content for automated moderation of Internet content. Further research can be focused on applied use, which can be a useful tool for assessing the level of social acceptability of digital text content published on social networks and for preventing the spread of harmful or offensive information. It can also help improve the quality of communication and increase the level of social interaction in general.

Keywords

RNN, sentiment analysis, sentiment classification, social acceptance

1. Introduction

The terms and conditions of the vast majority of social networks and messengers clearly define content requirements and acceptable standards of user behavior. For violations of such rules and regulations, social networks usually block either content or accounts. The community standards of Facebook, Instagram, Twitter, YouTube, and other similar platforms prohibit the publication of terrorist content containing hate speech at various levels, prevent copyright infringement, coordinated inauthentic behavior, bullying, and the posting of pornographic materials [1]. Social networks also fight against fake pages, and some platforms even delete inactive accounts. All social media platforms, including Facebook, Twitter, and YouTube, have made huge investments and developed policies to identify and moderate such harmful content [2, 3].

The content that is considered socially acceptable can vary depending on various factors, such as cultural and moral norms, age groups, geographic location, and others. However, in general, socially acceptable content is content that does not violate the rules of a social network and does not cause disgust or outrage among a significant part of the audience [1].

Many published works are devoted to the detection of hate speech and offensive language in comments in English. However, according to the authors' research [4], there are no appropriate

COLINS-2024: 8th International Conference on Computational Linguistics and Intelligent Systems, April 12–13, 2024, Lviv, Ukraine

yuri.krak@gmail.com (I. Krak); zalutska.olha@gmail.com (O. Zalutska); m.o.molchanova@gmail.com (M. Molchanova); exe.chong@gmail.com (O. Mazurets); g.cardinal2009@gmail.com (R. Bahrii); olenasobko.ua@gmail.com (O. Sobko); alexander.barmak@gmail.com (O. Barmak)

0000-0002-8043-0785 (I. Krak); 0000-0003-1242-3548 (O. Zalutska); 0000-0001-9810-936X (M. Molchanova); 0000-0002-8900-0650 (O. Mazurets); 0000-0001-5219-1185 (R. Bahrii); 0000-0001-5371-5788 (O. Sobko); 0000-0003-0739-9678 (O. Barmak)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

methods for automatically detecting obscenities and hate speech in Ukrainian, as the development of such methods is a difficult task for several reasons:

- There are no labeled databases or Ukrainian-language corpora of comments or social media posts with offensive content.
- Due to the peculiarities of word formation in the Ukrainian language, it is almost impossible to determine a definitive list of offensive words.
- Users often invent new, explicitly obscene words by connecting two common stems with an infix or adding prefixes or suffixes.
- The use of surzhik gives rise to a wide variety of swear words.

The paper [5] shows that a combination of machine learning and a lexicon-based approach can achieve higher accuracy than any type of sentiment analysis. The authors used a variety of sentiment analysis, machine learning methods, and dictionary-based sentiment analysis to test and compare the effectiveness of user behavior research.

The study [6] was conducted on sentiment analysis on Twitter. Their proposed method is based on a dictionary and allows identifying sentiments about the AstraZeneca/Oxford, Moderna, and Pfizer/BioNTech COVID-19 vaccines for 4 months. A similar study was conducted [7], but the authors proposed to use TextBlob based on TF-IDF vectorization to assess sentiment. LinearSVC was chosen as the classification model, which resulted in an accuracy of 0.96752 for English-language tweets.

The study [8] was aimed at identifying hate speech, which includes various forms of trolling, intimidation, harassment, and threats directed against specific individuals or groups of athletes. The experiments concern the detection of hate speech in the Serbian language. The proposed BiLSTM deep neural network [9], trained with different parameters, showed high accuracy in detecting hate speech in the sports sphere (96% and 97%) and a fairly low level of memorization.

The authors of paper [10] implemented a special type of deep learning based on a recurrent neural network (RNN) [11] called long-term memory (LSTM) for the automatic identification of hate and offensive content. The authors propose a language-agnostic solution for three Indo-European languages (English, German, and Hindi), the methodology does not use any pre-trained model, which leads to a neutral language solution, and does not require the development of cumbersome features for the proposed model.

According to TIME [12], Facebook removed more than seven million cases of hate speech in the third quarter of 2019, a 59% increase over the previous quarter. More and more of these hateful statements (80%) are now detected not by humans, but automatically by artificial intelligence.

However, the algorithms Facebook currently uses to remove hate speech only work in certain languages. This means that it has become easier for Facebook to curb the spread of racial or religious hatred online in predominantly developed countries and communities dominated by global languages such as English, Spanish, and Chinese.

According to Time, Facebook automatically detects such statements in more than 40 languages around the world. In other languages, Facebook relies on its users and human moderators to control hate speech.

Contrary to the algorithms that Facebook says now automatically detect 80% of hate posts without requiring the user to report them first, these human moderators do not regularly scan the site for hate speech themselves. Instead, their job is to decide whether to remove posts that have already been reported by users [13].

Minority languages are the most affected by this inequality. This means that racially motivated slurs, calls for violence, and targeted insults can spread faster in developing countries than they currently do in the United States, Europe, and elsewhere.

Therefore, although virtual communication through social media platforms is an integral part of human life, there is a downside that comes in the form of harmful online content [14]. Harmful content, whether it is fake news, rumors, hate speech, aggression, or cyberbullying, is a matter of serious concern to society [15]. Such harmful content affects a person's mental health and also leads to losses that cannot be compensated for [16]. Detection and moderation of such content are the primary tasks of information technology. The solution to the problem of determining the level of social acceptability of textual Ukrainian-language Internet content will allow automated moderation of Internet content [17].

Considering the analysis of the studies, we can distinguish approaches that allow automated detection of offensive language in textual Internet content for the Ukrainian language:

- Sentiment analysis. This approach is used to detect and classify the emotional tone of the text, which helps to determine the overall mood of the message. For example, positive, negative.
- Detection of abusive speech. Abusive speech detection tools are used to identify and classify texts that contain rudeness, insults, threats, etc.
- Machine learning. The use of machine learning algorithms allows you to create models that can determine the social acceptability of a text based on training on large data sets. Such models can take into account a wide range of features, such as semantics, syntax, emotional connotation, and other factors that affect the detection of offensive speech [18].

Offensive language is characterized by the presence of two aspects:

1. The presence of offensive words that define this content as offensive.
2. The presence of negative sentiment tone of content that actualizes offensive intentions.

Accordingly, in order to detect offensive speech, it is necessary to evaluate the content under study for the presence of each of the manifestations.

The aim of research is to create and validate the method for automated determination of the level of social acceptability of textual Internet content based on a combined approach that includes sentiment analysis using RNNs and verification with a dictionary of offensive words.

The main contributions of this research are follows:

- a method for detecting offensive language in textual Internet content for the Ukrainian language using a recurrent neural network has been developed;
- the developed method allows detecting offensive speech with a given user threshold.

2. Methods and Materials

Given these limitations, there is a necessity to generate experimental data that will satisfy the research objectives.

The method for detecting offensive speech for the Ukrainian language using a recurrent neural network based on a combined approach that includes sentiment analysis and verification with a dictionary of offensive words is proposed. Its main stages of work are also described.

The proposed method has been validated in the format of an application implementation in Python, which clearly demonstrates the effectiveness of this approach. An efficiency study is also performed.

2.1. Datasets

The «Ukrainian Twitter Corpus» (for RNN training) and the «AbusiveLanguageDataset» (for identifying offensive content) will be used as experimental data to implement the method of detecting offensive speech for the Ukrainian language using a recurrent neural network.

«Ukrainian Twitter Corpus» [19] is a corpus of Ukrainian-language tweets collected using the Twitter API. The corpus contains more than 6 million tweets that were collected from December 2015 to May 2017.

Each tweet in the corpus is labeled as positive, negative, or neutral. Annotation was performed using machine learning algorithms and evaluated at the message level. In total, the proposed corpus contains about 3 million positive tweets, 1.7 million negative and 1.6 million neutral tweets. Of that total, 400 thousand tweets are in Ukrainian.

The corpus also contains additional data on the number of likes, retweets, and replies per tweet. The data is available for download on GitHub. As the data in the corpus is collected from the social network Twitter, it may contain informal vocabulary, abbreviations and other speech features that are typical for this source. However, given the topic of the study, it is appropriate to use this corpus to train an RNN that will perform the task of determining the semantic tone of a message.

«AbusiveLanguageDataset» [20] is a dataset that contains Ukrainian-language comments with different levels of offensive content. This dataset consists of 5,000 comments, most of which are negative and contain foul language. The comments were collected from popular Ukrainian online resources, such as: «Ukrainska Pravda», «Tablo ID» and «TSN» [20]. The dataset is presented in many languages, including Ukrainian. Each comment in the selected dataset was pre-evaluated by a human who determined whether it contains offensive content and the level of offensiveness (low, medium or high). In addition, for each comment, the identifier of the source from which the comment was received is indicated.

The dataset will be used as data for a numerical assessment of the level of social acceptability of the textual content of the posts.

The datasets were supplemented with data from previous studies, a dataset of tagged reviews from the hotline consisting of 7656 documents [21]. The peculiarity of the dataset is that it contains russisms, swear words, and some reviews were presented in other languages. It is also supplemented by manually collected and labeled posts and comments from the social network Facebook in the amount of 500 units (250 positive and 250 negative).

Since even the tagged tweets and feedback from the hotline were in different languages, they were cleaned up with the help of the python language using the «langdetect library». The tweets and reviews were also filtered, and tweets consisting of less than 3 words were deleted. After imposing restrictions on tweets and Ukrainian-language reviews from the hotline, their number was as follows: the tweet set consisted of 2400 tweets (1200 positive and 1200 negative), the data set from the hotline consisted of 2000 positive and 2000 negative, the set of posts and comments from the social network Facebook consisted of 500 items (250 positive and 250 negative). The number of offensive words in the dictionary is 959 offensive words (Figure 1).

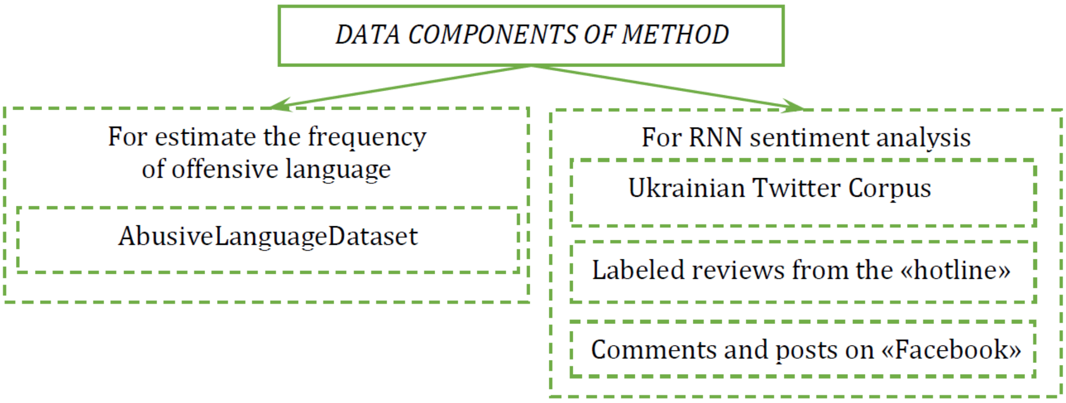

DATA COMPONENTS OF METHOD
For estimate the frequency of offensive language
AbusiveLanguageDataset
For RNN sentiment analysis
Ukrainian Twitter Corpus
Labeled reviews from the «hotline»
Comments and posts on «Facebook»

Figure 1: Data components of the method for detecting offensive speech for the Ukrainian language

The described set of data will be used to implement a method for detecting offensive speech for the Ukrainian language, which will be able to determine the level of acceptability of user-generated Internet content based on the entered textual information according to the established sensitivity threshold.

2.2. Components and stages of the method of detecting offensive speech for the Ukrainian language

In order to identify the level of offensive speech for Ukrainian-language Internet content, based on the analysis of publications, in particular, [22] shows that studies that mainly relied on dictionary tools to extract sentiment from textual data and have a clear advantage in terms of interpretation, obviously lose in accuracy. Therefore, in order to ensure the reliability of the result in determining the level of social acceptability of textual Ukrainian-language Internet content, a neural network will be used to identify emotional content and a dictionary approach will be used to check for elements of unacceptable content.

The input data of the method are: a pre-trained RNN tone analysis model, textual Internet content for analysis, and a dictionary of key offensive expressions. The Internet content for analysis is characterized by some features [23], the main ones being:

- short length (often has a limit on the number of characters, which causes the text to be short and requires a clear and concise expression of thoughts)
- informative nature (usually, information is presented in an informal manner and may contain abbreviations, non-standard vocabulary, or surzhik);
- use of multimedia (texts may include gifs, emojis, etc.);
- dialogic nature (social networks have an extensive structure that facilitates interaction between users and the creation of dialogues).

All this makes the processing of short online content specific, different from general methods for working with texts. A generalized diagram of the proposed method is illustrated in Figure 2.

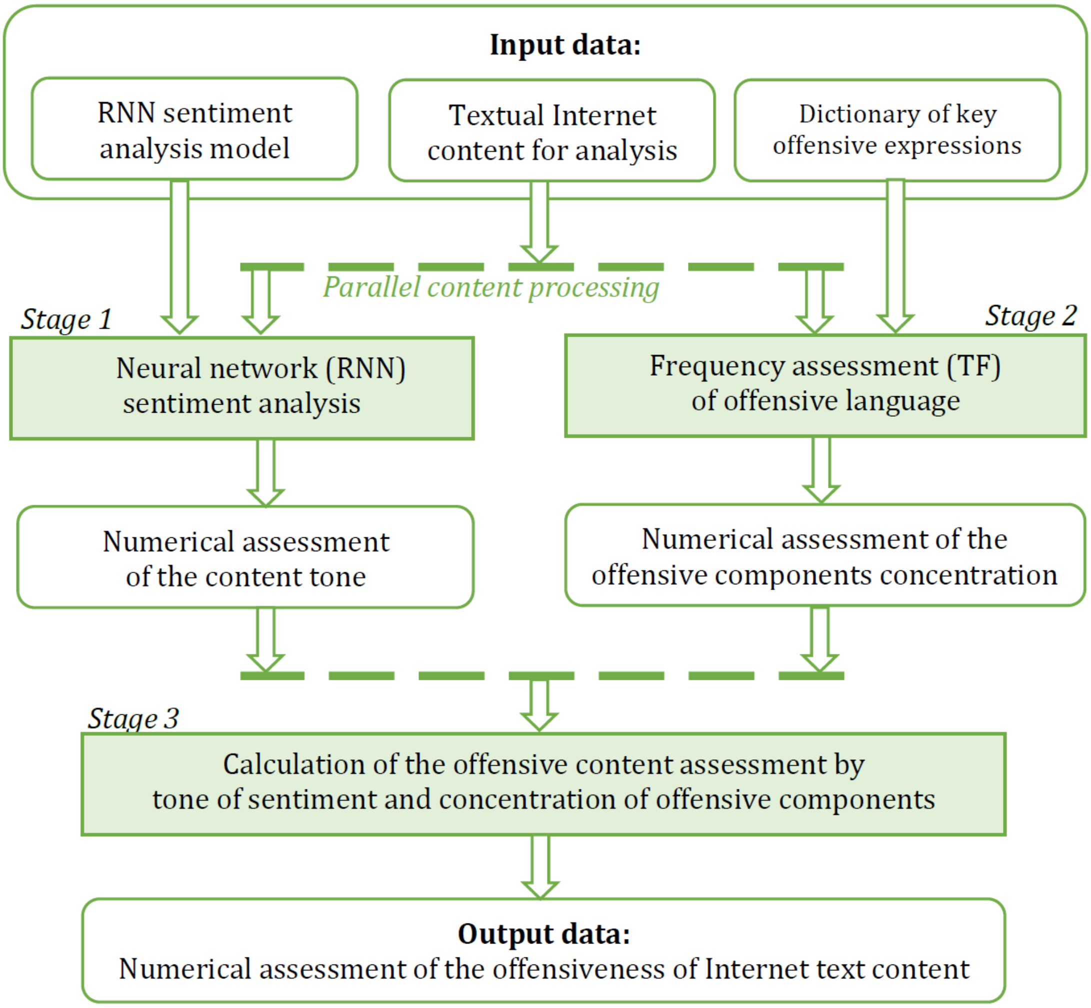

Input data:
RNN sentiment analysis model
Textual Internet content for analysis
Dictionary of key offensive expressions
Parallel content processing
Stage 1
Stage 2
Neural network (RNN) sentiment analysis
Frequency assessment (TF) of offensive language
Numerical assessment of the content tone
Numerical assessment of the offensive components concentration
Stage 3
Calculation of the offensive content assessment by tone of sentiment and concentration of offensive components
Output data:
Numerical assessment of the offensiveness of Internet text content

Figure 2: Scheme of the method for detecting offensive speech for the Ukrainian language using a recurrent neural network

Step 1 is a neural network assessment of the sentiment tone based on a pre-trained RNN model. The result of this step is a numerical assessment of the content's sentiment in the range from 0 to 1, where 0 is completely negative and 1 is completely positive.

Step 2 takes place in parallel with the first step and consists in assessing the frequency (TF) of the occurrence of offensive language, which will be determined by comparing the content of the message with the dictionary of offensive language. The frequency of offensive language will be calculated using the formula [24]:

$$TF = \frac{O_w}{TotalCount}, \tag{1}$$

where O_w is the number of offensive words contained in the dictionary, $TotalCount$ is the total number of words in the analyzed Internet content. The result of this step is a numerical assessment of the concentration of offensive components.

At step 3, the offensiveness score of the textual content is calculated based on the tone of the sentiment and the concentration of offensive components. The numerical assessment of the offensiveness of the textual content will be calculated using the formula:

$$AbusiveLevel = \frac{k \cdot TF + (1-k)(1 - SentimentVal)}{2}, \tag{2}$$

where TF is a numerical estimate of the concentration of offensive components, $SentimentVal$ is a neural network estimate of the sentiment tone of the post, k is the coefficient of the threshold of sensitivity of offensive words to the overall level of detection of offensive speech. It is selected according to the policy of the social service under study.

Accordingly, the output of the method of detecting offensive speech for the Ukrainian language using a recurrent neural network is a numerical assessment of the offensiveness of textual Internet content.

2.3. Formation of a trained RNN sentiment analysis model for detecting offensive speech

Since one of the inputs to the method of detecting offensive speech for the Ukrainian language using RNN is a trained model, it is necessary to obtain it for Ukrainian-language data. The neural network model is trained according to the algorithm shown in Figure 3.

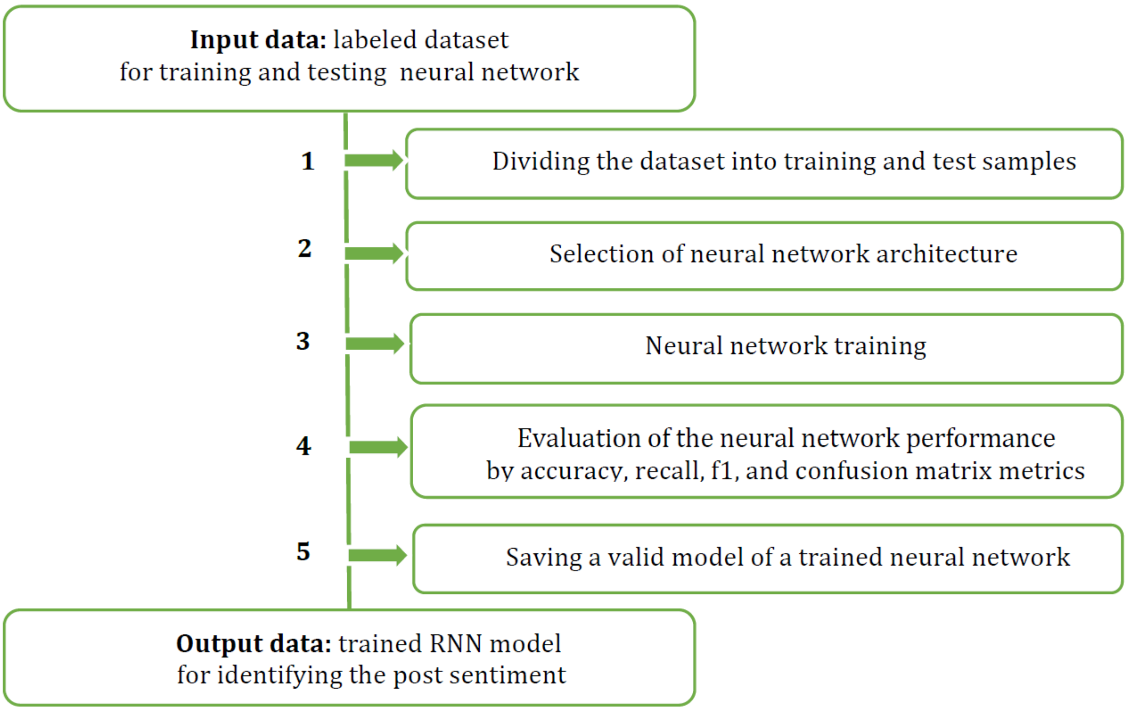

Input data: labeled dataset
for training and testing neural network
1
Dividing the dataset into training and test samples
2
Selection of neural network architecture
3
Neural network training
4
Evaluation of the neural network performance
by accuracy, recall, f1, and confusion matrix metrics
5
Saving a valid model of a trained neural network
Output data: trained RNN model
for identifying the post sentiment

Figure 3: Stages of forming trained RNN sentiment analysis model for detecting offensive speech

Therefore, the input for model selection is the labeled dataset used to train the neural network and evaluate its effectiveness. The sentiment of a post will be determined from a numerical range from 0 to 1, where 0 is a completely negative post, and 1 is a completely positive post.

The first step is to divide the dataset into training and test samples. It was decided to divide the dataset in a 60-40 ratio, where 60% is training sample and 40% is test sample. Accordingly, the number of training examples was 4 140, and the number of test examples was 2 760.

The next step was to select the neural network architecture. It was decided to use a simple three-layer architecture with an Embedding Layer, LSTM Layer, and Dense Layer with a sigmoidal activation function.

The next step was to train the neural network with the above architecture. The training stage was conducted in conjunction with the stage of further model evaluation based on such metrics as accuracy, recall, f1, and confusion matrix [25].

Accuracy is defined as the ratio of the number of correctly classified examples to the total number of examples [26]. *Recall* is defined as the ratio of the number of correctly classified positive examples to the total number of positive examples. *F_1-score* is calculated as a balanced average between accuracy and precision [27].

The following input parameters were analyzed for the training process: Batch size, number of epochs. The number of training epochs shows how many times the model is to be trained. Batch size shows the number of training examples used within one iteration of neural network training. It is very difficult to immediately determine what the perfect batch size is for the needs of a particular task [28], so this parameter will be selected experimentally. The statistics of the metrics for the conducted training are shown in Table 1. Figures 4-7 illustrate the pattern confusions of neural network models. Green is for true positives, red is for true negatives, yellow is for false positives, and blue is for false negatives.

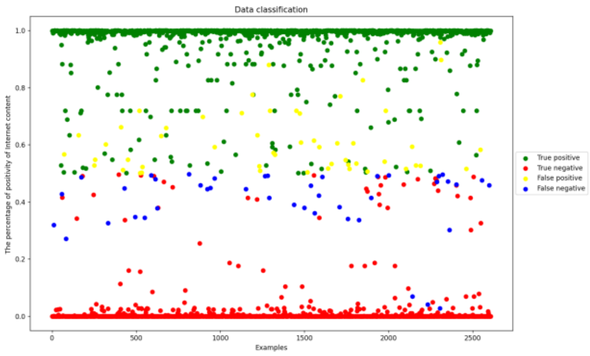


Figure 4: Classification of reviews by the V1 model

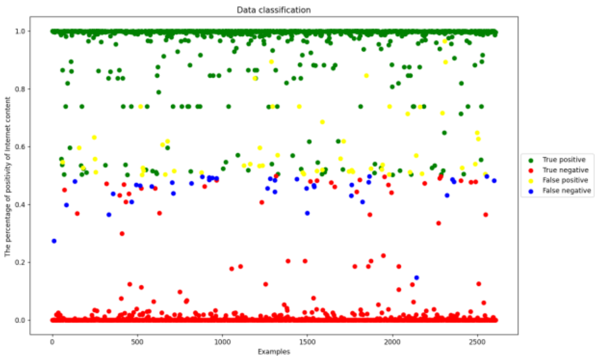


Figure 5: Classification of responses by the V2 model

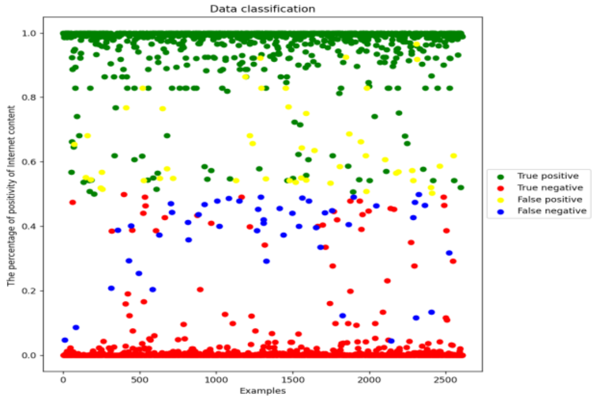


Figure 6: Classification of responses by the V5 model

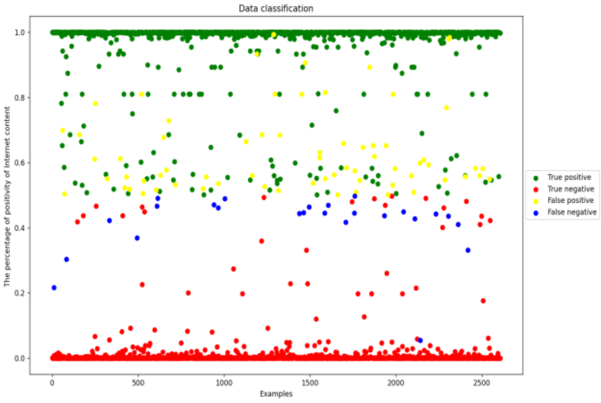


Figure 7: Classification of responses by the V7 model

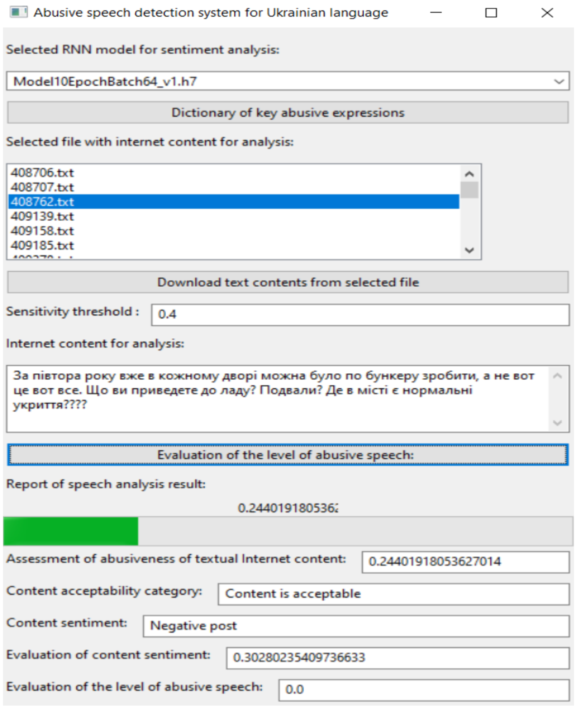
Table 1
RNN training parameters and results

Parameters	V1	V2	V3	V4	V5	V6	V7
Number of learning epochs	20	20	20	20	10	10	10
Batch size	128	64	32	16	64	32	16
Results							
Training time (sec)	257	343	503	921	183	255	442
Accuracy	0.951	0.951	0.957	0.949	**0,96**	0.956	0.947
Recall	0.963	0.968	0.948	0.936	0.959	0.943	**0.978**
F_1	0.959	0.961	0.956	0.95	0.957	0.956	**0.960**
True positive	0.96	0.97	0.95	0.94	0.96	0.94	0.98
True negative	0.96	0.96	0.97	0.97	0.97	0.98	0.95
False positive	0.036	0.037	0.028	0.026	0.035	0.02	0.047
False negative	0.037	0.032	0.05	0.06	0.04	0.06	0.022

As can be seen from Figures 4-7, in general, all models cope with the task, but given the purpose of the study, it was decided to use the *V5* model, which has the highest *Accuracy*. Although the *V7* model also has fairly high *Recall* and *F₁*values, it is more important to identify negative samples more accurately.

2.4. Study of the efficiency of the proposed method

To study the efficiency of the method of detecting offensive speech for the Ukrainian language using a recurrent neural network, a corresponding software implementation was created. Python tools were used for development, and the «wx» library was used for the user interface [29]. The Sklearn library was used for training and further use of the neural network [30]. An example of content identification is illustrated in Figure 8 and Figure 9.


Abusive speech detection system for Ukrainian language
Dictionary of key abusive expressions
Download text contents from selected file
Evaluation of the level of abusive speech:

Selected RNN model for sentiment analysis: Model10EpochBatch64_v1.h7

Selected file with internet content for analysis:
408706.txt
408707.txt
408762.txt
409139.txt
409158.txt
409185.txt

Sensitivity threshold : 0.4

Internet content for analysis:
За півтора року вже в кожному дворі можна було по бункеру зробити, а не вот це вот все. Що ви приведете до ладу? Подвали? Де в місті є нормальни укриття????

Report of speech analysis result:
0.244019180536:

Assessment of abusiveness of textual Internet content: 0.24401918053627014

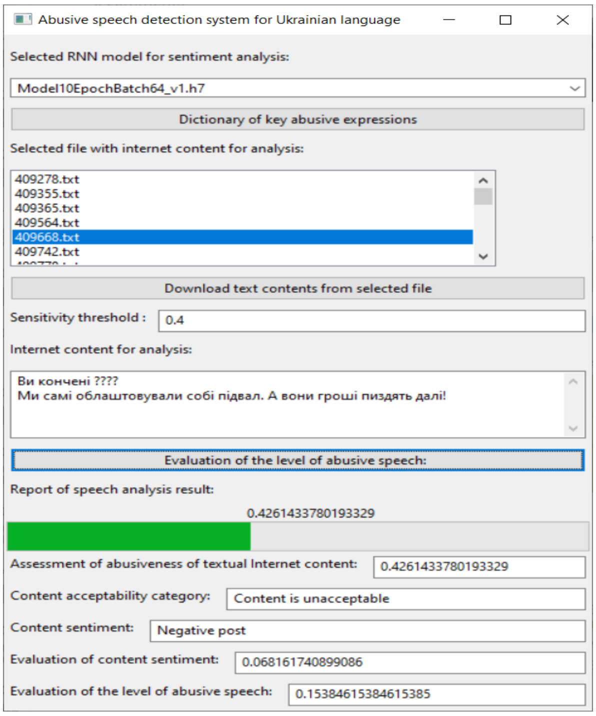
Content acceptability category: Content is acceptable

Content sentiment: Negative post

Evaluation of content sentiment: 0.30280235409736633

Evaluation of the level of abusive speech: 0.0

Figure 8: Acceptable-negative feedback


Abusive speech detection system for Ukrainian language
Dictionary of key abusive expressions
Download text contents from selected file
Evaluation of the level of abusive speech:

Selected RNN model for sentiment analysis: Model10EpochBatch64_v1.h7

Selected file with internet content for analysis:
409278.txt
409355.txt
409365.txt
409564.txt
409668.txt
409742.txt

Sensitivity threshold : 0.4

Internet content for analysis:
Ви кончені ????
Ми самі облаштовували собі підвал. А вони гроші пиздять далі!

Report of speech analysis result:
0.42614337801933329

Assessment of abusiveness of textual Internet content: 0.42614337801933329

Content acceptability category: Content is unacceptable

Content sentiment: Negative post

Evaluation of content sentiment: 0.068161740899086

Evaluation of the level of abusive speech: 0.15384615384615385

Figure 9: Unacceptable negative feedback

In the illustrated examples, 2 comments from Facebook were taken as test content. Comment 1: *«**Original:** За півтора року в дворі можна було по бункеру зробити, а не вот це вот все. Що ви приведете до ладу? Подвали? Де в місті є нормальні укриття? / **English translate:** In a year and a half, you could have built a bunker in the yard, not this crap. What will you improve? Basements? Where are there normal shelters in the city?».* This comment was classified by the neural network as negative with a score of 0.303. However, despite the negative content, the comment does not contain direct abuse or offensive words, so the overall offensiveness score is 0.244 with a sensitivity threshold of 0.4. Therefore, this comment is acceptable. Comment 2: *«**Original:** Ви кончені ???? Ми самі облаштовували собі підвал. А вони гроші пиздять далі! / **English translate:** You're fucking nuts????We were setting up our basement ourselves. And they keep fucking with the money!»*. As for this comment, it already contains a direct abusive content, is negative with a score of 0.068 (where 0 is completely negative content) and with a content abusiveness score of 0.15, its overall offensive speech score is 0.43, with a threshold value of 0.4, so this review is unacceptable.

3. Result and Discussion

Method for detecting offensive speech for the Ukrainian language was developed using a recurrent neural network, which includes a combined approach: an RNN network for determining the numerical assessment of the sentiment tone of the content and an approach for numerically assessing the concentration of offensive components based on the «AbusiveLanguageDataset». The values of metrics for trained versions of the neural networks at 20 epochs and different batch sizes are shown in Figure 10. The values of metrics for trained versions of neural networks at 10 epochs and different batch sizes are shown in Figure 11.

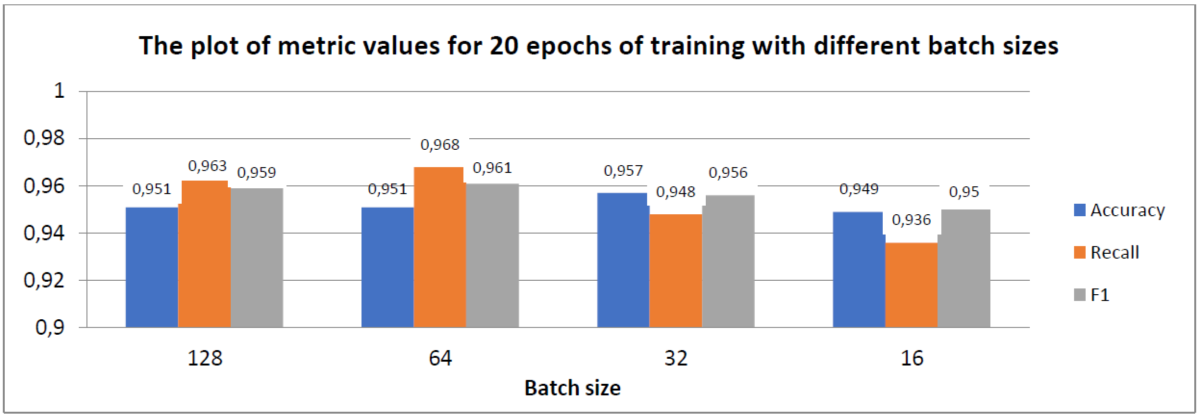


Figure 10: The value of metrics for determining the sentiment of RNN content for 20 epochs

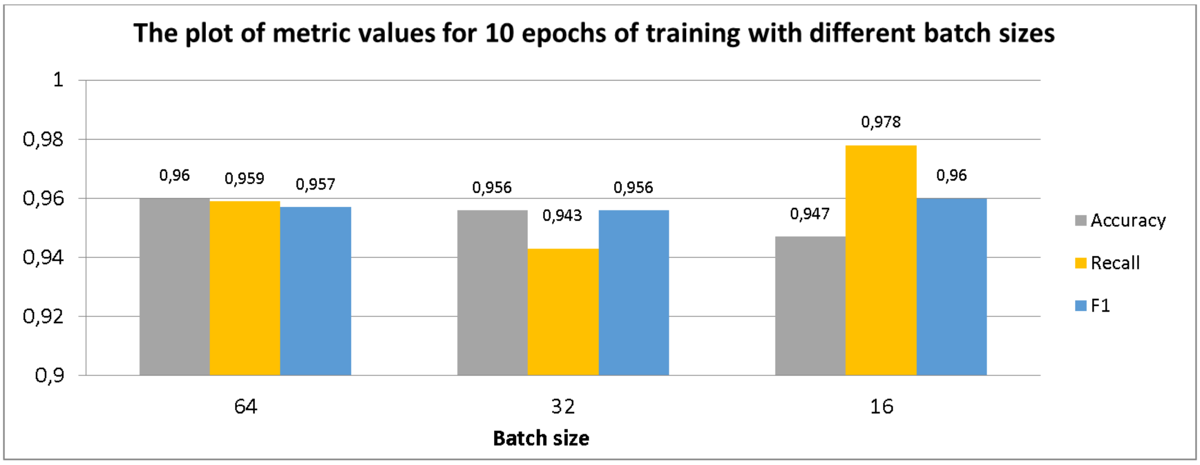


Figure 11: The value of metrics for determining the sentiment of RNN content for 10 epochs

As can be seen from the diagrams above, the metrics do not decrease below 94%, so the neural network shows high performance in all models for classifying the sentiment tone of text content.

As part of the study, we collected 50 comments that are not included in the training and test data from Facebook and Instagram, which were evaluated by experts into 2 categories: «acceptable» and «unacceptable». According to the expert evaluation, 32 comments were labeled as «unacceptable» and 18 as «acceptable». Based on the testing conducted using the developed method, 30 comments were unacceptable and 20 were acceptable. The data are illustrated in Table 2.

Table 2
Results of testing the method of detecting offensive speech for the Ukrainian language, pcs.

	Acceptable	Unacceptable
Acceptable	18	0
Unacceptable	2	30

However, the controversial comments that were characterized by the previous expert as unacceptable, and by the developed method as acceptable, were proposed to be evaluated by 3 more experts, and in the first case, 2 out of 3 also classified them as acceptable. The text of the comment was as follows: *«**Original:** Та я як молодий інженер електрик змушений роботами нижчого рівня перебиватися і по крупинкам практику збирати, бо всілякі там рішали хочуть на своїх підприємствах бачити спеціалістів за максимум 12к гривень. Думаю, я не один такий. / **English translate:** But as a young electrical engineer, I have to do lower-level jobs and collect my practice bit by bit, because all sorts of bastards want specialists for a maximum of 12 thousand hryvnias at their enterprises. I think I'm not the only one.»*. The score of this comment by the method of detecting offensive speech for the Ukrainian language was 0.386, with a threshold value of 0.4.

The second comment was as follows: *«**Original:** Це ж виходить можна безкарно гасити всіх в кого не має родичів. Нема родичів, нема кому звинувачивати. / **English translate:** That means you can put out everyone who has no relatives with impunity. No relatives, no one to blame.»*. The content acceptability score was 0.39, and the post was identified as negative with a score of 0.06, where 0 is completely negative content. For this comment, 2 out of three experts gave the rating «unacceptable».

The results with examples of how the method was used are also shown in Figure 12. For comparison, the same comments were submitted for the GPT chat evaluation, where he gave the scores shown in Figure 13.

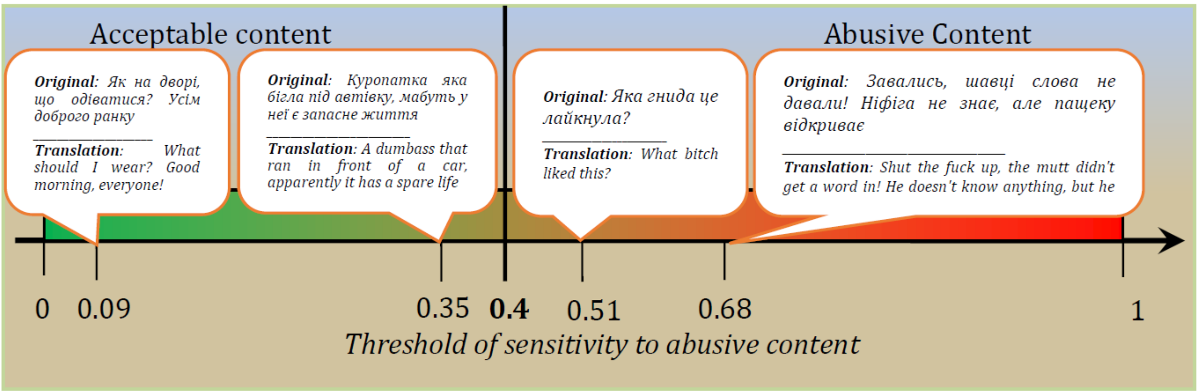

Acceptable content
Abusive Content
Original: Як на дворі, що одіватися? Усім доброго ранку
Translation: What should I wear? Good morning, everyone!
Original: Куропатка яка бігла під автівку, мабуть у неї є запасне життя
Translation: A dumbass that ran in front of a car, apparently it has a spare life
Original: Яка гнида це лайкнула?
Translation: What bitch liked this?
Original: Завались, шавці слова не давали! Ніфіга не знає, але пащеку відкриває
Translation: Shut the fuck up, the mutt didn't get a word in! He doesn't know anything, but he
0
0.09
0.35
0.4
0.51
0.68
1
Threshold of sensitivity to abusive content

Figure 12: The result of the applied use of the method

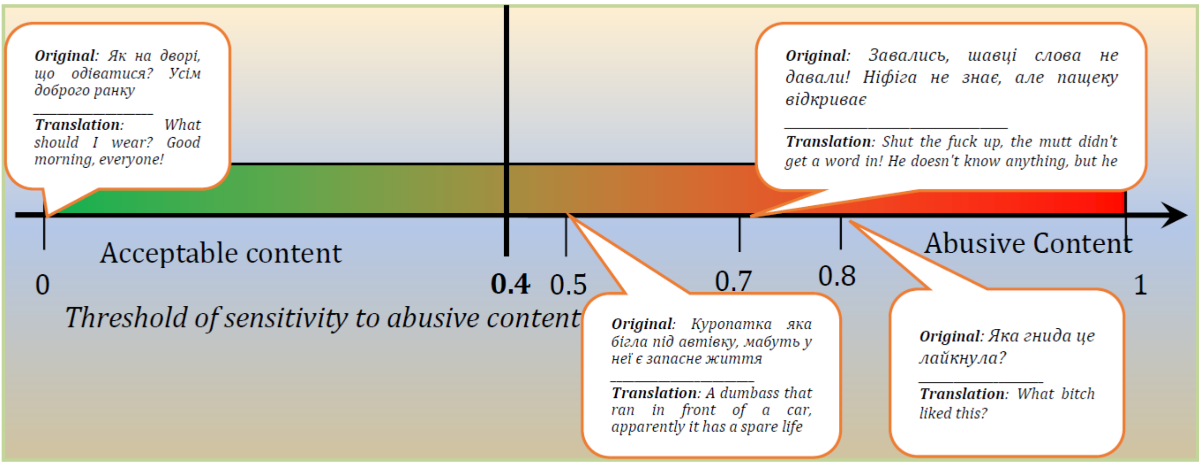

Original: Як на дворі, що одіватися? Усім доброго ранку
Translation: What should I wear? Good morning, everyone!
Original: Завались, шавці слова не давали! Ніфіга не знає, але пащеку відкриває
Translation: Shut the fuck up, the mutt didn't get a word in! He doesn't know anything, but he
Acceptable content
Abusive Content
0
0.4
0.5
0.7
0.8
1
Threshold of sensitivity to abusive content
Original: Куропатка яка бігла під автівку, мабуть у неї є запасне життя
Translation: A dumbass that ran in front of a car, apparently it has a spare life
Original: Яка гнида це лайкнула?
Translation: What bitch liked this?

Figure 13: Evaluating comments with GPT chat

Comment: «***Original***: *Куропатка яка бігла під автівку, мабуть у неї є запасне життя. /* ***Translation***: *A dumbass that ran in front of a car, apparently it has a spare life»* is rated 0.5 by GPT chat, while the proposed method is rated 0.35. However, the reason for this assessment by GPT chat is the presence of an unfavorable or cruel situation, such as an animal running in front of a car. This can cause negative emotions or seem offensive to those who care about animal welfare. Therefore, this assessment is not entirely correct. The rest of the comments are not in conflict and are determined to be unacceptable by both the proposed method and the GPT chat.

4. Conclusion

The paper considers the current state of the field of offensive speech detection, which is one of the key areas in working with texts for many languages. According to the analysis, the main approaches to solving the problem of detecting offensive speech and identifying offensive content were identified, including: sentiment analysis, detection of abusive speech (dictionary approach) and the use of machine learning. It was decided to use a combined approach based on sentiment tone analysis and abusive speech detection (dictionary approach). Also, taking into account the specifics of the application for the Ukrainian language, it was necessary to create an appropriate dataset consisting of tweets from the «Ukrainian Twitter Corpus», marked responses from the «Abusive Language Dataset» and «hotline»

The tweets and reviews were filtered, and tweets with less than 3 words were removed. After applying restrictions to the tweets and Ukrainian-language reviews from the hotline, their number was as follows: the tweet set was 2400 tweets (1200 positive and 1200 negative), the data set from the hotline was 2000 positive and 2000 negative, the set of posts and comments from the «Facebook» social network was 500 units (250 positive and 250 negative). The number of offensive words in the dictionary was 959 offensive words. In total, the sample amounted to 6900 Ukrainian-language texts.

To train the RNN, it was decided to divide the dataset in a 60:40 ratio, where 60% is the training sample and 40% is the test sample. Accordingly, the number of training examples was 4140, and the number of test examples was 2760.

The trained RNN model, which was subsequently used for sentiment analysis, had an accuracy of 0.96, while *Recall* and F_1 had scores of 0.959 and 0.957, respectively.

Abusive speech detection method for the Ukrainian language using a recurrent neural network was tested on the developed software, and the study shows that the method has a high efficiency of detecting abusive Ukrainian-language content. According to the research, the method has an estimated identification accuracy of more than 90%, however, the assessment of abusiveness can be subjective, and the perception of content can vary from person to person. However, to improve the result, it is necessary to supplement the dictionary of abusive expressions, as the Ukrainian language is rich in surzhik and other foreign language twists, which are not currently fully represented in the dictionary. The proposed method also has a number of limitations: it works with text content of 3 words or less and no more than 500 words, and it works only in the Ukrainian language.

Further research can be directed towards practical applications, which can be a useful tool for assessing the level of social acceptability of digital textual content published on social networks and for preventing the spread of harmful or offensive information.

References

[1] Meta, Facebook Community Standards. 2024. URL: https://transparency.fb.com/policies/community-standards/

[2] Gongane, V.U., Munot, M.V., Anuse, A.D., Detection and moderation of detrimental content on social media platforms: current status and future directions, Soc. Netw. Anal. Min. 12, 129 (2022). doi: 10.1007/s13278-022-00951-3

[3] J. J. Van Bavel, C. E. Robertson, K. Rosario, J. Rasmussen, S. Rathje, Social Media and Morality, Annual Review of Psychology, Volume 75, 2024 , Van Bavel, pp 311-340. doi: Annual Review of Psychology Volume 75, 2024 Van Bavel, pp 311-340

[4] Kovalchuk O., Slobodzian V., Sobko O., Molchanova M., Mazurets O., Barmak O., Krak I., Savina N. Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets. Book Chapter. Lecture Notes on Data Engineering and Communications Technologies. 2023. Vol. 149. pp. 591–607. URL: https://link.springer.com/chapter/10.1007/978-3-031-16203-9_33. doi: 10.1007/978-3-031-16203-9_33

[5] H. Li, Qi Chen, Zh. Zhong, R. Gong, G. Han, E-word of mouth sentiment analysis for user behavior studies, Information Processing & Management. 2022. doi: 10.1016/j.ipm.2021.102784

[6] R. Marcec, R. Likic, Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines, Postgraduate Medical Journal, Volume 98, Issue 1161. 2022. pp. 544–550. doi: 10.1136/postgradmedj-2021-140685

[7] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, P. Cotae, Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Systems with Applications. 2023. doi: 10.1016/j.eswa.2022.118715

[8] V. Stanković, S., Mladenović, M., An approach to automatic classification of hate speech in sports domain on social media, J Big Data 10, 109. 2023. doi: 10.1186/s40537-023-00766-9

[9] Anfilets, S., Bezobrazov, S., Golovko, V., Sachenko, A., Komar, M., Dolny, R., Kasyanik, V., Bykovyy, P., Mikhno, E., & Osolinskyi, O. Deep Multilayer Neural Network for Predicting the Winner of Football Matches. International Journal of Computing, 2020, 19(1), 70-77. doi: 10.47839/ijc.19.1.1695

[10] Saha, B.N., Senapati, A. CIT Kokrajhar Team: LSTM based Deep RNN Architecture for Hate Speech and Offensive Content (HASOC) Identification in Indo-European Languages, FIRE, 2019, 12-15 December, Kolkata, India. 2019. URL: https://ceur-ws.org/Vol-2517/T3-24.pdf

[11] Roy, K. S., Islam, S. M. R. An RNN-based Hybrid Model for Classification of Electrooculogram Signal for HCI. International Journal of Computing, 2023, 22(3), 335-344. doi: 10.47839/ijc.22.3.3228

[12] Time, Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch. 2019. URL: https://time.com/5739688/facebook-hate-speech-languages/

[13] Gibson, A. D., What Teams Do: Exploring Volunteer Content Moderation Team Labor on Facebook. Social Media + Society, 9(3), 2023. doi:10.1177/20563051231186109

[14] A. Y. Lee, J. T. Hancock, Social media mindsets: a new approach to understanding social media use and psychological well-being, Journal of Computer-Mediated Communication, Volume 29, Issue 1, January, 2024. URL:

https://academic.oup.com/jcmc/article/29/1/zmad048/7612379. doi: 10.1093/jcmc/zmad048

[15] Yesmen N., Arnab N., Sumona Sh., Hate crimes in social media: A criminological review. International Journal of Social Science, 3(1), 2023, pp. 23–28. doi:10.53625/ijss.v3i1.5607

[16] D. Neumann, N. Rhodes, Morality in social media: A scoping review, New Media & Society, 26(2), pp. 1096-1126, 2024. URL: https://journals.sagepub.com/doi/full/10.1177/14614448231166056. doi: 10.1177/14614448231166056

[17] M. Katsaros, J. Kim, T. Tyler, Online Content Moderation: Does Justice Need a Human Face?, International Journal of Human–Computer Interaction, 40:1, pp. 66-77, 2024. URL: https://www.tandfonline.com/doi/abs/10.1080/10447318.2023.2210879. doi: 10.1080/10447318.2023.2210879

[18] Norval, M., Wang, Z. Speech Emotion Recognition using Hybrid Architectures. International Journal of Computing, 2024, 23(1), 1-10. doi: 10.47839/ijc.23.1.3430

[19] GitHub, Ukr-twi-corpus. 2019. URL: https://github.com/saganoren/ukr-twi-corpus

[20] Hatespeechdata, Hate Speech Dataset Catalogue. 2020. URL: https://hatespeechdata.com/

[21] Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I., Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network, CEUR Workshop Proceedings, 2023, vol. 3387, pp. 561–571. URL: https://ceur-ws.org/Vol-3387/paper26.pdf

[22] J. Hartmann, M. Heitmann, Ch. Siebert, Ch. Schamp, More than a Feeling: Accuracy and Application of Sentiment Analysis, International Journal of Research in Marketing. 2022. doi: 10.1016/j.ijresmar.2022.05.005

[23] D. C. Gkikas, K. Tzafilkou, P. K. Theodoridis, A. Garmpis, M. C Gkikas, How do text characteristics impact user engagement in social media posts: Modeling content readability, length, and hashtags number in Facebook, International Journal of Information Management Data Insights.тVolume 2, Issue 1. 2022. doi:10.1016/j.jjimei.2022.100067

[24] M. M. Abedi, E. Sacchi, A machine learning tool for collecting and analyzing subjective road safety data from Twitter, Expert Systems with Applications, Volume 240, 2024. URL: https://www.sciencedirect.com/science/article/abs/pii/S0957417423030841. doi: 10.1016/j.eswa.2023.122582

[25] Y. Krak, O. Barmak, O. Mazurets, The practice implementation of the information technology for automated definition of semantic terms sets in the content of educational materials, CEUR Workshop Proceedings, 2018, vol. 2139, pp. 245-254. doi: 10.15407/pp2018.02.245

[26] J. Hartmann, M.Heitmann, Ch. Siebert, Ch. Schamp, More than a Feeling: Accuracy and Application of Sentiment Analysis, International Journal of Research in Marketing, Volume 40, Issue 1, 2023, Pages 75-87, doi: 10.1016/j.ijresmar.2022.05.005

[27] Scikit-learn, sklearn.metrics.f1_score. 2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[28] Medium, How does Batch Size impact your model learning. 2022. URL: https://medium.com/geekculture/how-does-batch-size-impact-your-model-learning-2dd34d9fb1fa

[29] wxPython. 2023. URL: https://wxpython.org/index.html

[30] Scikit-learn. Machine Learning in Python. 2023. URL: https://scikit-learn.org/stable/