ISSN 2007-9737

Sentiment Analysis for Religious Tweets

Shahislam Abubakir[1], Aruzhan Baimbet[1], Akmaral Kabieva[1], Bauyrzhan Sarsenov[1], Aisha Seidilabek[1], Iskander Akhmetov[1,2], Alexander Gelbukh[3,*]

[1] Kazakh-British Technical University, School of Information Technologies and Engineering, Almaty, Kazakhstan

[2] Insitute of Information and Computational Technologies, Almaty, Kazakhstan

[3] Instituto Politecnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico

i.akhmetov@ipic.kz, gelbukh@cic.ipn.mx

Abstract. Sentiment analysis of Twitter data has received a lot of interest and has shown diverse findings. Because of the brief data used in tweets, it draws the opportunity to get information about public opinion by studying Twitter data and automatically identifying its sentiment polarity. The goal of this study was to classify religious Tweets and extract opinion words about a specific religion. In this paper, we used the Valence Aware Dictionary for Sentiment Reasoner (VADER) to categorize tweets containing the phrase "there is religion." The results indicated that detecting many classes of sentiment analysis was accurate.

Keywords. Sentiment analysis, VADER, religious text, tweets.

1 Introduction

In the century of digitization, social media is extremely essential in everyone's life. Considering all the pros and cons of social networking, it is important to note that it allows us to broaden our minds, share ideas, and communicate with society and organizations. However, the uncertainty in political, religious, and social issues causes extremism among people that are depicted by their sentiments on social media.

Twitter's user base has grown rapidly, and the volumes of messages produced by Twitter every day are vast. Therefore, it is important to interpret the algorithm with high precision and recall, which will identify the religious tweets that wreak havoc around the world. So, most state-of-the-art studies have implemented sentiment analysis to classify Tweets on a variety of topics, like forecasts, reviews, e-commerce, and election.

Many methods now exist for extracting advanced features from text, such as Textblob of Natural Language Toolkit (NLTK). As a result, determining the society's relationship with religion, whether negative, positive, or neutral, is critical, at the very least. For this approach, the research paper suggests implementing the sentiment analysis on religious Tweets using Valence Aware Dictionary for Sentiment Reasoning (VADER).

The remainder of the paper is organized as follows: A brief review of relevant work is presented in Section 2. In Section 3, we describe the data of a model. In Section 4, we present the sentiment analysis processing structure and describe the tool used in this study. Section 5 contains the results of the study.

ISSN 2007-9737

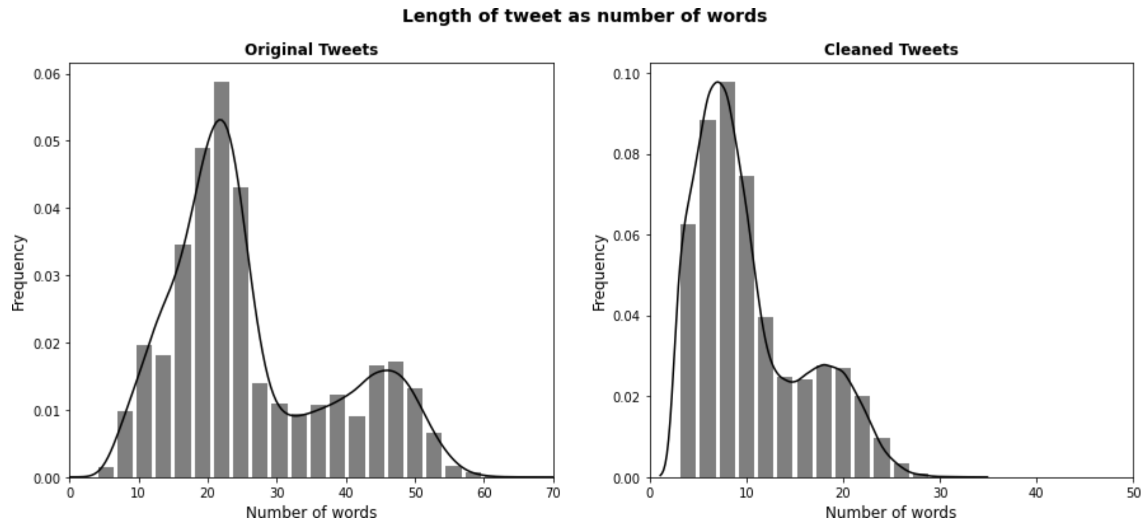

Fig. 1. Length of tweet as number of words

Possible interpretations and future works are drawn in Section 6. Finally, in Section 7, possible final observations are drawn.

2 Literature Review

Nowadays social networks, media and other online resources process big amount of information on the World Wide Web. All of the data that include different opinions can be used for good purposes in business and other aspects of commercial and scientific industries.

The social network has now emerged as an essential part of an individual's life. It has changed the way of living in the 21st century. Globalization has played an important role by the beginning of the last era of the twentieth century in linking diverse people around the world.

Through online communication, people around the globe have started to understand the norms, culture and traditions of each other. As a result of this, similar-minded people have started working together to achieve a common goal [4].

However, it is likely to be impossible to track and extract convenient information manually from internet space networks. For this purpose, modern world proposed the sentiment analysis, which extracts sentiments from the user opinion.

Sentiments classified into two categories like positive sentiments and negative sentiments, to define the general attitude of the people on different ideas. The main purpose of sentiment analysis is to accurately identify the emotion of opinions.

2.1 Related Works

According to Twitter's latest figures from the fourth quarter of 2020, the platform boasts 192 million daily active users (Twitter, 2021). Half a billion tweets are sent out each day (Mention, 2018). That equates to 5,787 tweets per second [3].

It shows that the Twitter is the main platform to share ideas with the big audience. Therefore it is important to track the activity of users who follow bad intentions. Paper published by Xujuan Zhou, Xiaohui Tao, Jianming Yong and Zhenyu Yang introduces the Tweets Sentiment Analysis Model (TSAM) that can provide early indications of topics and entities for which societal interests are emerging or may emerge, predict the developing trend of an event within a specified time period.

In addition, it provides a fast and less expensive alternative to traditional polls (e.g., telephone poll) for mining public opinion. This research work has demonstrated that building a lexicon-based sentiment analysis intelligent system is doable

Table 1. Dataset with religious tweets

	Tweet Text
0	Islam is a religion full of blessings and good...
1	@ wagnerclaire Religion and pity are antonyms...
2	if the only reason people are on twitter is to...
3	@15MeterClassYas @NotoriousDachi is there such...
4	Religion is needed to provide a moral compass ...

Table 2. Frequency of words

	Frequency
religion	2012
islam	1502
christianity	302
forgiveness	245
deen	227
atheism	214
faith	204
judaism	178

and can be very beneficial. However, in its current form the opinion analysis tool is not yet reached full potential [6]. Domain specificity for sentiment analysis models can also lead to a more accurate classification, as same words can carry different semantics and sentiments across different domains [1].

2.2 Method

The paper work done by Muhammad Asif and Atiab presents the idea of classifying tweets into four categories, including high extreme, low extreme, moderate, and neutral, based on their level of extremism. The algorithm is based on Linear Support Vector Classifier outperforms with an accuracy of 82 percent [3].

C. J. Hutto in his article describes the development, verification, and evaluation of Vader (for the Valence Aware dictionary for reasoning about feelings) [2]. It uses a combination of qualitative and quantitative methods to create and then empirically validate a gold standard sentiment lexicon that is particularly tailored to microblogging-like contexts.

He was able to combine these lexical features with five generalizable rules that embody the grammatical and syntactic conventions that people use when expressing or underlining the intensity of moods, finding that the inclusion of these heuristics increases the accuracy of the sentiment analysis engine in multiple domain contexts (social media text, New York Times editorials, movie reviews, and product reviews).

C. J. Hutto emphasized that Vader's vocabulary works exceptionally well in the field of social media. The correlation coefficient of his work shows that VADER ($r = 0.881$) works in the same way as individual human evaluators ($r = 0.888$) according to the basic truth (aggregated group average of 20 human evaluators for intensity the sentiment of each tweet).

Surprisingly, when he further checked the accuracy of the classification, he saw that VADER ($F1 = 0.96$) actually even outperforms individuals ($F1 = 0.84$) in correctly classifying tweet sentiment into positive, neutral, or negative classes. VADER retains (and even improves on) the advantages of traditional mood lexicons such as LIWC: it is larger, but just as easy to test, understand, and apply quickly (without the need for extensive training/training) and easy to expand.

Like LIWC (but unlike some other lexicons or models machine learning), VADER 's sentiment lexicon is a gold standard quality and has been validated by humans. VADER differs from LIWC in that it is more sensitive to expressions of feelings in social media contexts, and also generalizes more favorably to other areas.

3 Description of the Data

For our research, we used data from GitHub [5], these are extracted tweets containing the phrase “there is religion”. To increase accuracy, we have used the balanced dataset. Dataset gathered by retrieved 1000 tweets per month from January 2015 to October 2019. This results in roughly 57,351 tweets, which are then loaded into a dataframe ready for preprocessing. In overall dataset includes 1416651 words. Mean number of words per tweet: 24.69 words.

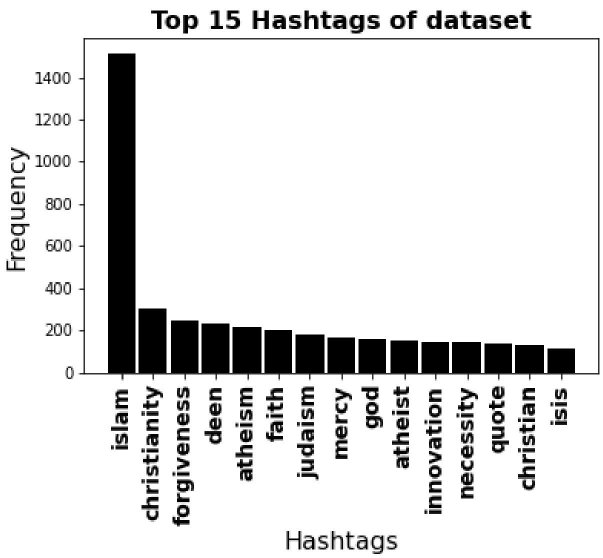

Fig. 2. Top 15 hashtags of dataset

Total length of the dataset is: 8671277 characters. Mean Length of a tweet is: 151.0 characters. Most frequently used words in the dataset are religion, Islam and Christianity. To study the hashtags among 53.939 Tweets, 15557 Hashtags were used.

4 Description of the Method

Any data analysis workflow starts with loading the data. Next, we need to pass them through the preprocessing pipeline:

– Tokenize text-split text into sentences, words, and other units.

– Delete stop words.

– Bring the words to their normal form.

– Vectorize texts – make numeric representations of texts for further processing by the classifier.

All these steps serve to reduce the noise inherent in any normal text and improve the accuracy of the classifier results. In the Figure 1, we described the architecture of our model. First, tweets were converted to strings for checking the duplicates. The next step is the removing the Twitter Handles to identify the most frequently mentioned word.

People use the hashtag symbol (#) before a relevant keyword or phrase in their Tweet to categorize those Tweets and help them show more easily in Twitter search. Clicking or tapping on a hash tagged word in any message shows you other Tweets that include that hashtag. Hashtags can be included anywhere in a Tweet. There is the visualization of top 15 hashtags in the dataset.

4.1 Tokenization

Tokenization is the process of breaking text into smaller parts. The spaCy library already has a built-in pipeline, which starts its work on text processing with tokenization. In this guide, we will divide the text into individual words. The captured tokens include punctuation marks and other non-word strings. This is normal behavior when using the default pipeline.

4.2 Deleting Stop Words

Stop words are words that may be important in human communication, but do not make sense to machines. The space library comes with a default stop-word list (it can be customized). With a single line of Python code, we can filter out stop words from tokenized text using the token attribute .is_stop. After deleting stop words, the list became shorter, pronouns and service words disappeared: articles, conjunctions, prepositions, and post positions.

4.3 Return to Normal Form

In the process of normalization, all forms of the word are reduced to a single representation. There are two main approaches to normalization: stemming and lemmatization. In the case of stemming, the base of the word is allocated, adding which you can get the descendant words.

However, this is a naive approach – stemming just truncates the string, discarding the ending. Lemmatization seeks to solve this problem by using a data structure in which all forms of a word are associated with its simplest form, the lemma. Lemmatization is usually more useful than stemming, and is therefore the only normalization strategy offered by spaCy.

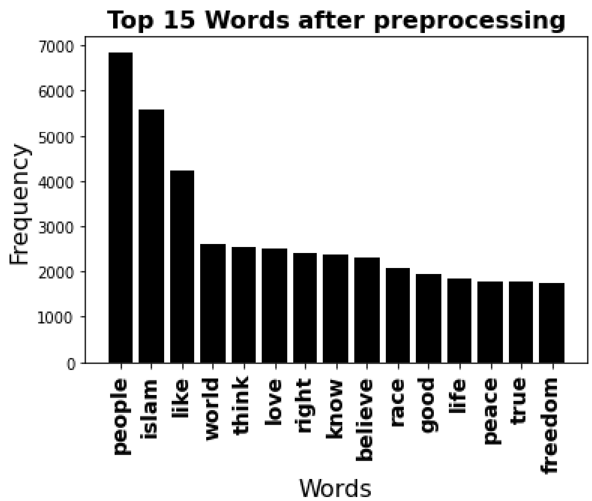

Fig. 3. Top 15 words after preprocessing

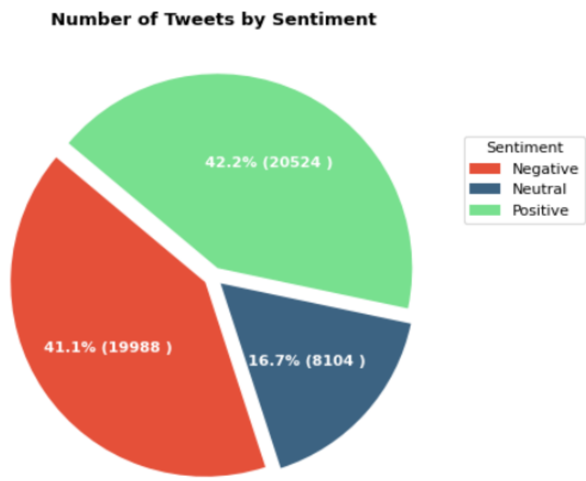

Fig. 4. Number of Tweets by sentiment

Within the NLP pipeline, lemmatization occurs automatically. The lemma for each token is stored in the .lemma_

4.4 Text Vectorization

Vectorization is the conversion of a token into a numeric array that represents its properties. In the context of the task, the vector is unique for each token. Vector representations of tokens are used to evaluate word similarity, classify texts, and so on. In spaCy, tokens are vectorized as dense arrays in which non-zero values are defined for each position.

This distinguishes the approach used from the earlier methods, which used sparse arrays for the same purposes and most positions were filled with zeros. There is the visualization of Top 15 Words after preprocessing.

4.5 Applying VADER Sentiment Analyzer

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. It is available in the NLTK package and can be applied directly to unlabeled text data. VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text:

– We divide the data into training and test samples (data sets).
– Choosing the model architecture.
– We use the training data to configure the model parameters (this process is called training).
– We use test data to evaluate the quality of the model training.
– We use the trained model on the new, previously not considered input data to create forecasts.

Machine learning specialists usually divide a data set into three components:

– Data for training.
– Data for validation.
– Data for the test.

IN VADER model we divide the dataset to positive, negative, neutral and compound:

– Positive: Compound score $>=$ 0.05
– Neutral: Compound score between -0.05 and 0.05
– Negative: Compound score $<=$ -0.05

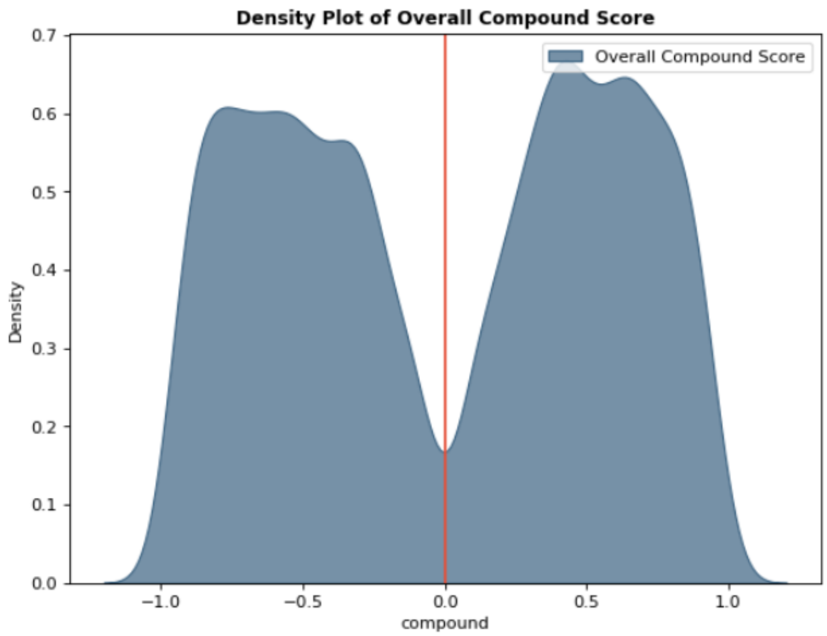

Fig. 5. Density plot of overall compound score

```
neg_tweets = df.sentiment.value_counts()[−1]
neu_tweets = df.sentiment.value_counts()[0]
pos_tweets = df.sentiment.value_counts()[1]
```

Listing 1. Applying the scores to sentiments

Model shows that 41,2 percent is negative sentiment, 42,2 percent is positive and other 16,7 percent is neutral.

4.6 Accuracy of the Algorithm

Building on the results of experiment, we calculate the precision, recall and F-score in order to describe the accuracy of the algorithm:

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{1}$$

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{2}$$

Precision means proportion of objects called positive by the classifier and at the same time really positive.

Recall shows proportion of objects of a positive class from all objects of a positive class the algorithm found. F-Score is the average of Precision and Recall and it takes takes both false positives and false negatives into account:

$$\text{Recall} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{3}$$

In next step we remove the neutral compound scores to compare the negative and positive tweets.

5 Results

As a result the VADER sentiment analysis classified the dataset of tweets into positive, negative and neutral sentiments. Following Compound Score Distribution diagram shows the compound of positive and negative tweets.

In addition algorithm identified the 10 most negative and positive tweets. VADER analyzer shows the sentiment of overall dataset and the percentage distribute of whole tweets. Overall sentiment is: ‘neg’: 0.34, ‘neu’: 0.437, ‘pos’: 0.223, ‘compound’: -0.5267:

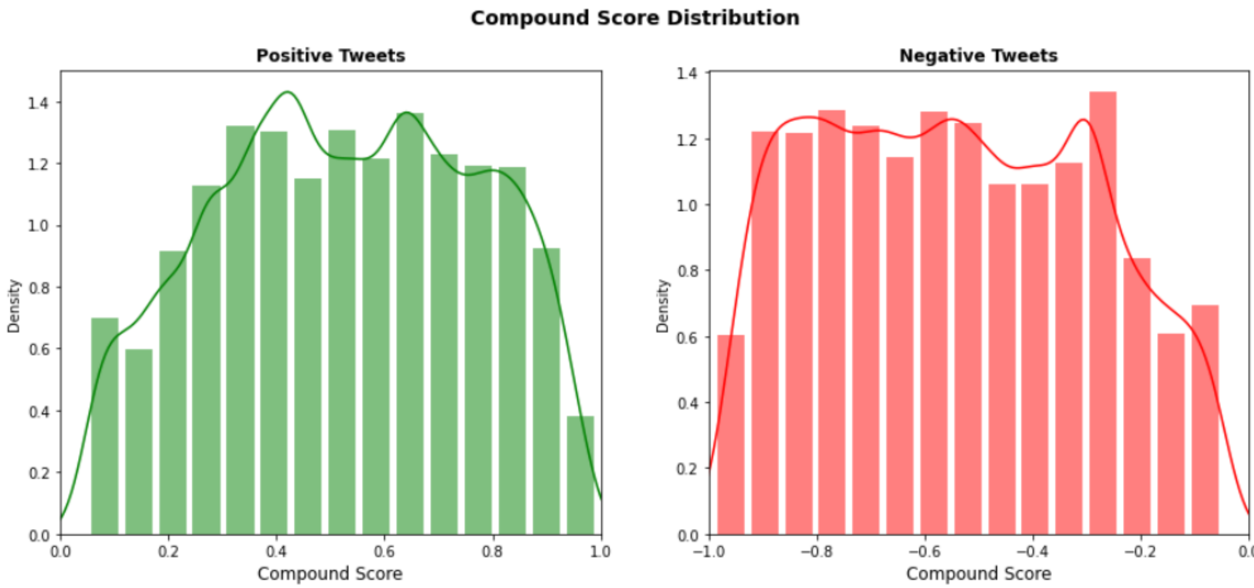

Fig. 6. Diagram of compound score distribution

– Sentence was rated as 34.0.

– Sentence was rated as 43.7.

– Sentence was rated as 22.3.

– Sentence overall rated as negative.

6 Discussion

Religion is a sphere in which public opinion is divided, which requires observing a very fine line, so in virtual world it is impossible to refute the presence of hate motivation and offensive expressions between their discussions.

Thus, social media platforms should make more efforts to prevent such offensive behavior of their users in their online discussions. Uncovering the relationship of the society towards religion, whether it is negative, positive, or neutral is crucial, at least at requires attention.

Despite the fact that the data is not from a variety of sources and not only in text form, it is impossible to deny the fact that religion occupies a very large share of importance in the life of society and opening new horizons of this field of activity is very important for science.

7 Conclusion

The sentiment analysis for religious text in social media comments is introduced in this paper. The paper also goes through how to collect data, how to filter them, how to preprocess text by removing unnecessary data, removing some unrelated words and text, and how to normalize the text. Displaying the results in numbers sentiment analysis showed that the public is divided when it comes to their sentiment towards religion, 42.2 % of tweets are classified as positive.

41.1 % were classified as negative with the rest classified as neutral. The indicator shows that the opinions of the public are very different and studying this area more and more and collecting all the results of this activity will lead to a society that is more loyal to a particular religion and other national conflicts regarding religion or react in time to more radical interpretations about religion.

Plans include new improvements to the algorithm. The current version of the model works only with text. It means that the algorithm does not understand emoticons, tags, etc. In addition, the civilization is constantly developing and we should not limit human attitude towards the religious text in negative modality only.

ISSN 2007-9737

Further research will include these aspects and prevent biased understanding among people. Furthermore, to improve the accuracy of our experiments, further research and experiments will be conducted using various text feature extraction techniques as well as other Deep Neural networks.

Acknowledgments

This research is conducted within the framework of the grant num. AP09058174 “Development of language-independent unsupervised methods of semantic analysis of large amounts of text data”.

References

1. **Akhmetov, I., Gelbukh, A., Mussabayev, R. (2022).** Topic-aware sentiment analysis of news articles. Computación y Sistemas, Vol. 26, No. 1, pp. 423–439. DOI: 10.13053/cys-26-1-4179.
2. **Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H., Shah, J. (2020).** Sentiment analysis of extremism in social media from textual information. Telematics and Informatics, Vol. 48, pp. 101345. DOI: 10.1016/j.tele.2020.101345.
3. **Lin, Y. (2021).** 10 Twitter statistics every marketer should know in 2021. www.oberlo.com/blog/twitter-statistics.
4. **Nagaraj, B., Francis, A. (2015).** International conference on innovations in information, embedded and communication systems. Karpagam College of Engineering.
5. **Noka, H. (2020).** Religion of Twitter – Topic modeling, sentiment analysis and hate speech detection (dataset). github.com/HurmetNoka/religion_on_twitter/blob/master/full_data.csv.
6. **Shen, W., Ochoa, S. F., James, A., Li, X., Li, X., Schneider, D., Luo, J., Iqbal, R., de-Souza, J., Gao, L., Lu, T., Liu, T., Tao, X., Laing, R., Bian, J., Meseguer, R., Zhang, J., Jardim-Goncalves, R., Pino, J. A., Li, W., et al. (2013).** Proceedings of the IEEE 17th international conference on computer supported cooperative work in design. IEEE. DOI: 10.1109/cscwd30294.2013.

Article received on 11/10/2024; accepted on 24/11/2024.
**Corresponding author is Alexander Gelbukh.*