CEUR-WS.org/Vol-2552/Paper8.pdf

Application of Neural Network Modeling in the Task of Destructive Content Detecting *

Valentin Okhapkin[1]
vpokhapkin@yandex.ru

Anastasia Iskhakova[2,3]
shumskaya.ao@gmail.com

Elena Okhapkina [1]
enaokhapkina@mail.ru

Andrey Iskhakov [2]
iskhakovandrey@gmail.com

[1] Bauman Moscow State Technical University, Moscow
[2] V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow
[3] Tomsk State University of Control Systems and Radioelectronics, Tomsk
Russian Federation

Abstract

The study concerns the problem of identifying text messages containing signs of aggression. The analyzed database of messages was obtained for the period 2015-2016 from the social network “Vkontakte”. The method of vector representation of words and the model of recurrent neural network are used as analysis tools. The result of the simulation is a binary classifier: a message with signs of aggression or neutral content.

Keywords: *deep learning, Bag of Words, neural network, text analysis, social network, aggression, information security.*

1 Introduction

Development of information and communication technologies, combined with steady increase in their availability in the twenty-first century, largely predetermined the rapid transition to the digital method of text and other information transmission. On the one hand, this process has brought advantages in the quantity and speed of information delivery, but on the other hand, these advantages have received a negative connotation in the context of aggression broadcast in virtual space. This fact is reflected in the Doctrine of Information Security of the Russian Federation adopted by the President in December 2016. Among other strategic goals and main directions of ensuring information security, the developers highlight the need to “ensure the protection of citizens from information threats, including through the formation

*Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

of a culture personal information security" [Doctrine, 2016]. However, if the formation of codes and rules of conduct on the Internet is a tool of the future, which can be learned in the most natural way at school age, the purpose of ensuring the protection of citizens in the present is, from the point of view of the authors, methods and software mechanisms capable of identifying unwanted content. The symbiosis of these approaches with solving the problem of protecting the individual from intentional or accidental contacts with information containing signs of aggression, forms a set of conditions for countering this specific threat of virtual space.

At the same time, the subject of the authors' research is difficult to isolate from the general flow of text information, and requires a clear definition of what aggression is in a published text message. Quite often the circulating in the virtual space information has the character of news materials, is a quote or a reference to a source, which in general is not an attempt to create a conflict between two individuals. Research of problems of influence of information influence on various communities is considered in [Gradoselskaya, 2014; Zelinsky, 2008; Parfent'yev, 2009; Sushko, 2017; Khlomov, 2018; Levonevskii et al., 2019]. For this reason, we analyze the space of text messages (posts) of the social network "Vkontakte" for the 2015-2016 period, as a place where in the vast majority of cases, the appeal occurs directly from user to user. It is also understood that a published post for a small or large audience of users again generates a direct dialogue between the author of the publication and the user who responded to it. The analyzed dialogues' data is are open to the entire Internet audience and is not private correspondence. At the time of publication, most of the messages are still publicly available.

2 Text Messages Analysis on Machine Learning Methods

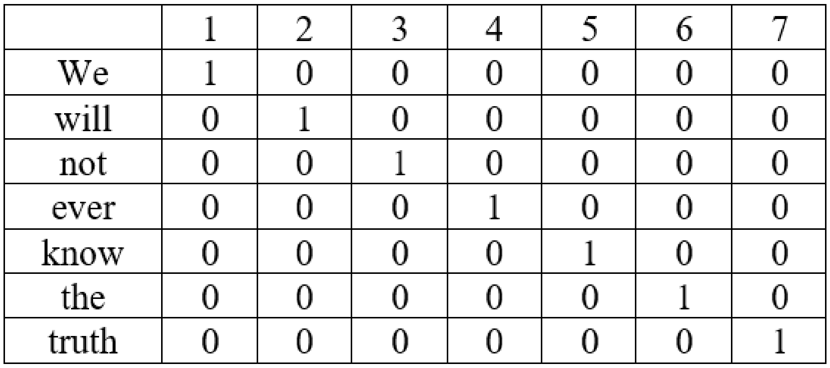
Aggression in the virtual space, expressed in the form of a text, is to be understood as the directed use of vocabulary towards the individual, including obscene lexicon, openly insulting their honor and dignity [Iskhakova, 2018]. The authors in the article here we do not consider the virtual space conflicting parties' attempts to use the references towards news materials, quotations of other social network users or any other sources aiming to offend the dialogue participant.

It is obvious that computer processing of the text in any problems' statement requires representation of the source data in the form, which makes possible its loading into the computer memory, application of analysis algorithms and meaningful interpretation of results. Machine deep learning methods are no exception. This approach does not mean that with the help of software and mathematical algorithms the computer will get the tool for text understanding in the human sense but its application is able to solve the problems of classification, text emotional component assessment, the target audience of the written. The recurrent neural network model is to be used in order to assess the aggressiveness of the analyzed messages. As for any other architectures of neural networks, a set of numerical parameters must be submitted to the input of the recurrent network (in the literature devoted to machine learning, one can meet the naming "tensors"). In this regard, it is necessary to resolve the issue of representation of the text message in numerical form. This representation is possible by converting each character of a sentence or each word into a vector. There is a more complex approach involved in creating n-grams from characters or words, which are overlapping sets of characters or words. For instance, text message: "We will not ever know the truth" represented as a bigram will be written as: "We", "We will", "will", "will not", "not", "not ever", "ever", "ever know", "know", "know the", "the", "the truth". The creation of such sets was called the "bag of words". Despite the more complex semantic load while "bag of words" elements creation, the approach associated with the transformation of each word into a vector will be used in the study. This is because the application of the recurrent neural network does not require the direct use of such an approach, and the network itself is able to receive groups of words (or symbols) without their explicit definition.

3 Vector Representation of Words and Dictionary of Terms Creation

Direct encoding of words or vector representation can solve representations of message's words as a numeric vector. The first approach is implemented by assigning a unique index to each word of the message, which is then converted into a binary vector, the dimension of which coincides with the dictionary size. After this conversion, the vector will consist of zeros except for a unique index, which will be assigned a value of "1". For the "bag of words" example, table 1 shows the direct encoding result[1] .

Table 1: Vector representation of the message~~'s~~ words after a direct encoding



	1	2	3	4	5	6	7
We	1	0	0	0	0	0	0
will	0	1	0	0	0	0	0
not	0	0	1	0	0	0	0
ever	0	0	0	1	0	0	0
know	0	0	0	0	1	0	0
the	0	0	0	0	0	1	0
truth	0	0	0	0	0	0	1

The disadvantage of direct encoding is the large size of the stored data. With this approach to dictionary of terms creation, large messages receive greater dimension, and numerical values are sparse. The vector representation of words, which condenses the vector and thereby makes it small, allows to get rid of this disadvantage. In short, the words vector representation allows fitting a large amount of information in a smaller number of dimensions [Chollet, 2018].

It should be noted that the vector representation of words forms a space where words synonyms can get distantce from each other in a geometric sense indices. Deep neural network

[1] The program code of the directed encoding of the message's words in R language to be found in Appendix 1.

does not take this fact into account. In this sense, the result of the assessment of the network for the presence or absence of aggression in the message will depend on how structurally correct are the vector presentation of words with similar meanings. It is also true that the use of the word vectorization method depends on the scope of application even within one language: detection of aggression in text messages in business correspondence, interviews, and social networks puts the task of constructing vector spaces for specific tasks of the studied area.

The practical material for the vector space of words construction will be a database of messages of the social network “Vkontakte”, aggregated for the period 2015-2019 in the Russian-speaking network segment. The research method includes the following algorithm:

1) create the mostly used words dictionary, including those with aggressive connotation;
2) assign the unique indices to each word, used in the messages;
3) encode messages in the terms of the unique indexes;
4) construct the recurrent neural network to classify a message as neutral or aggressive.

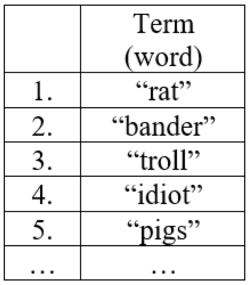
Note that real text messages contain a number of slang and openly violate the grammar of the Russian language expressions and words, which obviously imposes additional restrictions on the accuracy of the proposed technique. In this work the implementation of filters helps to get rid of such words and expressions. For instance, for the message “Not like Evryone Else” words “not like” and “everyone” can be brought to the normal language, but the type of messages like “ugotit” requires isolation of separate tokens and correction of spelling errors. The solution of this problem involves the creation of a dictionary of all possible distortions for the studied database and the development of an algorithm for rewriting the original text messages in a corrected form. This aim widens the frames of authors’ research.

To assess the aggressiveness of certain words and terms in the analyzed database, the semantic thesaurus WordNet-Affect is used. The markup of this dictionary in addition to the main emotional labels of the emotion, mood, cognitive state, physical state, emotional response and other types contains important emotional labels that express the general context of the analyzed text: positive, negative, ambiguous and neutral. Note that the selected thesaurus is difficult to structure in terms of detailed classification of types and subtypes of labels. Without setting a goal to identify all possible emotional labels in existing messages, only a limited number of them are used: emotion, mood, feature, hedonic signal, and attitude-position. This restriction was introduced intentionally in order to avoid uncertainty in the classification of social network messages containing exclusively aggressiveness. In addition, in the created dictionary based on the thesaurus WordNet-Affect words and terms that occur at least 100 times in 100 messages are included. The original database contains 1048576 messages. With the use of the noted criteria, a dictionary of terms and words has a size of 16000 units. Some examples from the created dictionary are given in table 2.

Note that the criterion chosen makes it possible in a limited number of cases to include in the dictionary terms and words used in messages less than 100 times. Critically, this fact does not affect the qualitative content of the dictionary. Most often, the number of occurrences is close to the criteria parameters. Table 3 shows one of the classified messages.

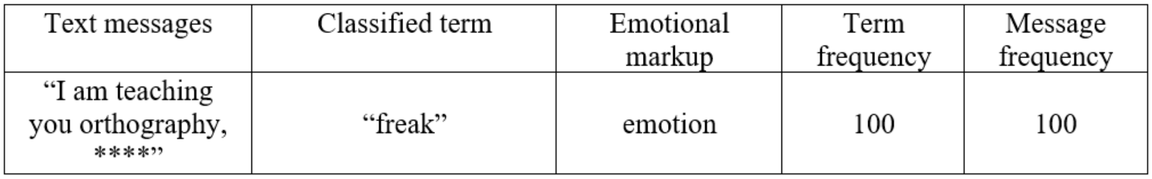
As it was mentioned before the vector representation method implies the unique indexes assignment to the words. For the table 3 example the vectorization of words will take the following form: [32 11 143 28 345]. The resulting word vectors do not contain coding for punctuation marks as symbols that do not have a significant impact on emotional coloring, including the context of identifying aggression. Besides, users often consciously and unconsciously violate the rules of punctuation, which in the case of accounting the vector representation of the

Table 2: Short listing of the terms and words included in the dictionary



	Term (word)
1.	"rat"
2.	"bander"
3.	"troll"
4.	"idiot"
5.	"pigs"
…	…

Table 3: Messages with the signs of aggression



Text messages	Classified term	Emotional markup	Term frequency	Message frequency
"I am teaching you orthography, ****"	"freak"	emotion	100	100

message imposes additional difficulties.

4 Recurrent neural network construction

The recurrent type of neural network to detect messages with signs of aggression is chosen due to the fact that classical neural networks do not have memory. Each input is handled independently, with no state between the networks [Chollet, 2018]. The use of fully connected neural networks or convolution based networks involves the processing of the entire sequence as a single data packet. However, human understanding of the text is carried out consistently – word by word. A recurrent neural network reproduces this principle in a simplified form: preserving the states of the neural network by processing the subsequent word index. The network state is reset after processing one sequence of two words. A simplified recurrent neural network can be represented in figure 1.

Figure 2 shows a simple recurrent neural network deployed in time. Each time interval is the result of a cycle at time t. In the output tensor each time interval t corresponds to information about time intervals from 0 to t in the input sequence – about the entire past. Therefore, it is not necessary to have the entire sequence of results in many cases; it's enough to get the last result (the outputt value at the end of the loop), because it already contains information about the entire sequence [Chollet, 2018].

The easiest and the most optimal convenient tool to construct this type of neural networks is Keras library. The algorithm for processing a sequence of words followed by a reset of the RNN state is associated with the launch of the layer embedding and layer simple rnn. The first layer allows creating a dictionary with integer indexes, and the second layer – to implement

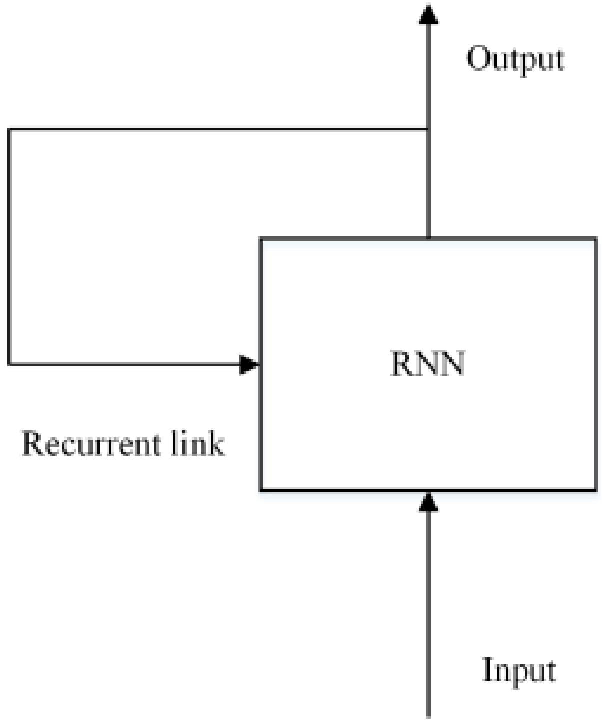


Figure 1: Recurrent neural network

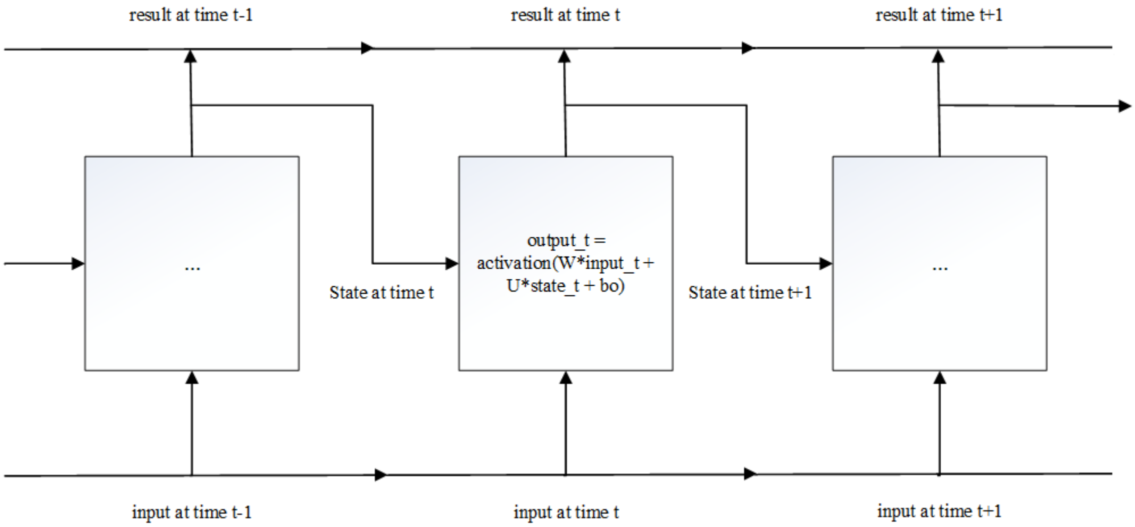


Figure 2: RNN deployed in time

the principle of recursion in time: preserving the previous state of the neural network and using it in the processing of the current data packet. Figure 3 shows the accuracy and loss during the training and verification phases of the text message analysis model.

On the one hand, the results obtained demonstrate a good level in the "quality of learning – loss in learning network" model, but it should be noted that they could be higher with a high dictionary processing degree and elimination of punctuation marks, so-called smileys, special characters that replace the spelling of the letters of the Russian alphabet.

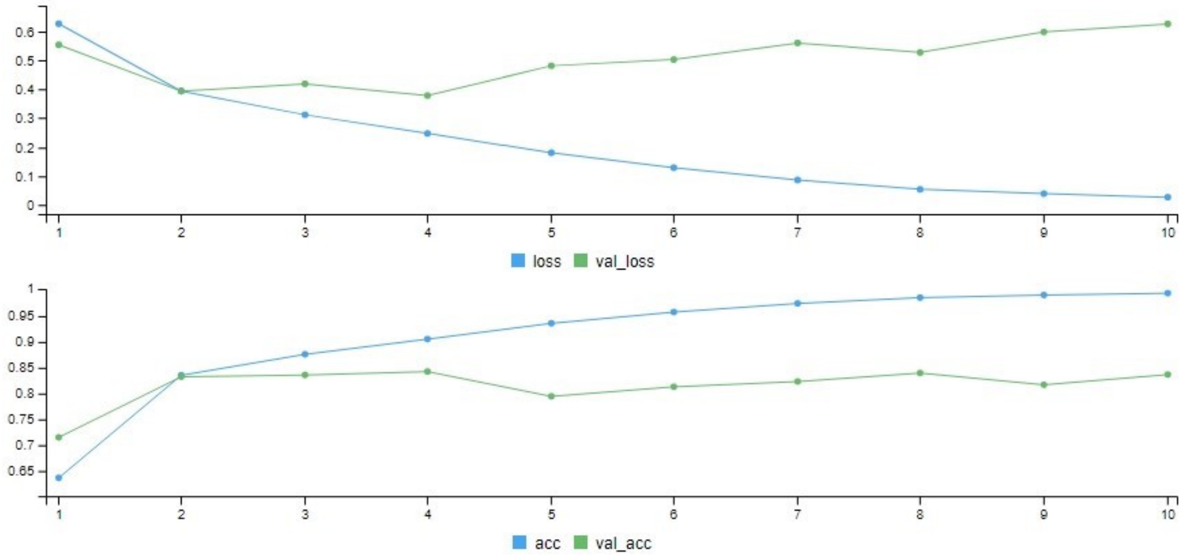


Figure 3: Loss and accuracy of RNN for the database of messages of the social network "Vkontakte"

5 Discussion

The study requires continuation and solving the problem of message texts deep processing. Having more than one million records in the database will obviously require automation and development of semantic analysis algorithms. The study development is also forced by use/application of more complex layers in the recurrent neural network: layer lstm and layer gru. The use of these layers will eliminate the problem of storing information about the input data in previous iterations of the network. Often neural networks, in which it is necessary to keep significant amounts of information about previous states in memory at any given time, are not amenable to training. The layerlstm allows to store a sequence of information blocks until further use, including network signals with delayed time intervals. In fact, the task of this layer is to re-use the information about the previous states of the RNN at the necessary time, thereby preventing the problem of growing gradient damping. The use of the layer gru allows for the reduction of the problem dimension when designing RNN by one parameter.

It would be wise to expand the list of emotional responses used in the compilation of the dictionary. The WordNet-Affect thesaurus has a fairly branched tree that classifies large groups of responses and defines the subtle shades of the messages being analyzed. In particular, from the example in table 3, the emotional response can be divided into responses according to the scheme "negative emotions – general disgust – hatred – hostility – aggression".

Taking into account the frequent events of physical violence on the part of the educational institutions parolees, the design and development of specialized advanced algorithms for text mining becomes actually important. A significant number of tragic events that occurred in our country began as a reaction to aggression, demonstrated in virtual space. It is safe to say that the timely detection of aggressiveness in texts and manifestos published in youth communities, and not only there, allowed the relevant services to attract the attention of the relevant services and prevent possible victims.

Acknowledgements

The reported study is funded by RFBR according to the research project No. 18-29-22104 “Development of socio-cyberphysical system of monitoring of diverse Internet-content for counteraction to manifestation of aggression, pressure and other forms of destructive impact on individual and group consciousness of users”.

References

[Doctrine, 2016] Information security doctrine of the Russian Federation (2016), December, 5, 2016, Moscow, Russia. 17 p. (In Rus.) = Doktrina informacionnoj bezopasnosti Rossijskoj Federacii, Moskva, 2016. 17 p. Avaible at http://www.kremlin.ru/acts/bank/41460

[Gradoselskaya, 2014] Gradoselskaya G.V. (2014) Grouping politically active communities in facebook by grain clustering (in Rus) = Gruppirovka politicheski aktivnykh soobshchestv v Facebook metodom zernovoy klasterizatsii. 56 p. Avaible at http://wciom.ru/fileadmin/file/nauka/grusha2015/s2$_2$/$gradoselskaya.pdf$

[Zelinsky, 2008] Zelinsky S.A. (2008) Analysis of mass manipulation in Russia. Analysis of mass management manipulative methods in the case of the modern era destructive based on the example of Russia. Psychoanalytic approach. SPb, “Scythia”. 280 p. (In Rus.) = Zelinskiy S.A. (2008) Analiz massovykh manipulyatsiy v Rossii. Analiz zadeystvovaniya manipulyativnykh metodik upravleniya massami v issledovanii destruktivnosti sovremennoy epokhi na primere Rossii. Psikhoanaliticheskiy podkhod. SPb. Skifiya. 280 p.

[Zelinsky, 2008] Zelinsky S.A. (2008) Individual and mass manipulation. Authority manipulative technologies in the attack on individual and mass subconscious. SPb.: Publishing and Trading House “SCYTHIA”. 240 p. (In Rus.) = Zelinskiy S.A. (2008) Manipulirovaniye lichnost’yu i massami. Manipulyativnyye tekhnologii vlasti pri atake na podsoznaniye individa i mass. SPb.: Izdatel’sko-Torgovyy Dom “SKIFIYA”. 240 p.

[Parfent’yev, 2009] Parfent’yev U. (2009) Cyber-aggressors // Children in the information society. 2. Pp. 66-67. (In Rus.) = Parfent’yev U. (2009) Kiber-agressory // Deti v informatsionnom obshchestve. 2. Pp. 66-67.

[Sushko, 2017] Sushko V.A. (2017) The method of sociometry and analysis of social networks: Textbook. M .: “KDU”, “University Book”. 310 p. (In Rus.) = Sushko V.A. (2017) Metod sotsiometrii i analiz sotsial’nykh setey: Uchebnoye posobiye. M.: “KDU”, “Universitetskaya kniga”. 310 p.

[Khlomov, 2018] Khlomov K. (2018) On the types of cyberbullying, their influence on the psyche and new roles of the victim and the aggressor. (In Rus.) = Khlomov K. (2018) O vidakh kiberbullinga, ikh vliyanii na psikhiku i novykh rolyakh zhertvy i agressora. Avaible at https://postnauka.ru/longreads/86459

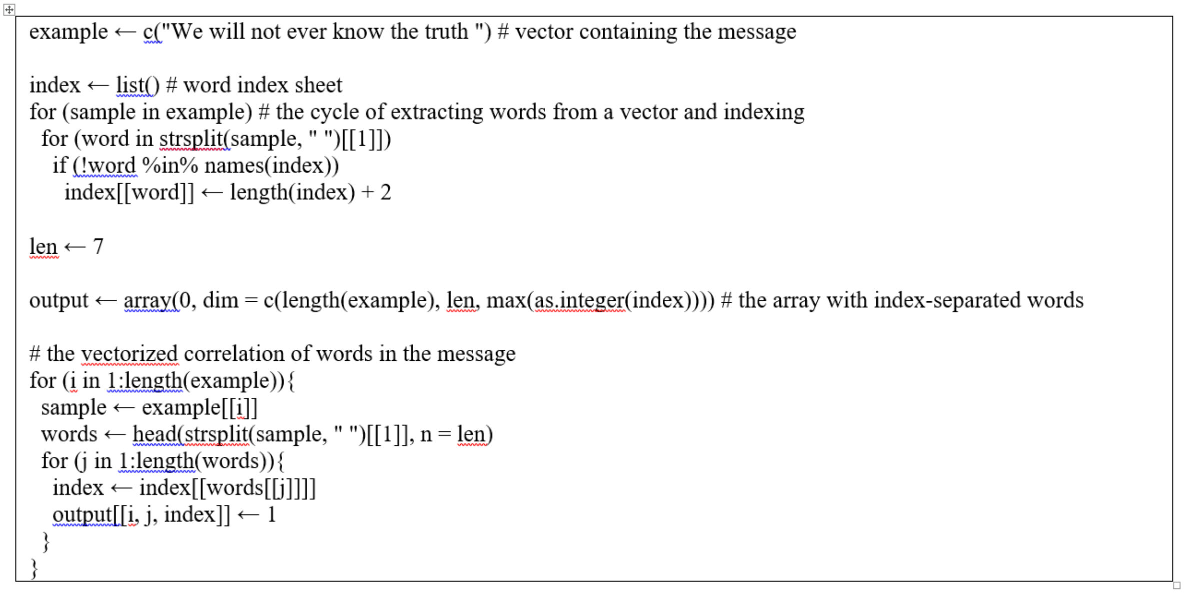
[Levonevskii et al., 2019] Levonevskii D., Shumskaya O., Velichko A., Uzdiaev M., and Malov D. (2019) Methods for Determination of Psychophysiological Condition of User Within Smart Environment Based on Complex Analysis of Heterogeneous Data // Proceedings of

14th International Conference on Electromechanics and Robotics "Zavalishin's Readings". Smart Innovation, Systems and Technologies, vol 154. Pp. 511-523.

[Iskhakova, 2018] Iskhakova A., Iskhakov A., and Meshcheryakov R. (2019) Research of the estimated emotional components for the content analysis // Proceedings of the International Conference "Applied Mathematics, Computational Science and Mechanics: Current Problems" (AMCSM 2018, Voronezh). Voronezh: Institute of Physics Publishing. Vol. 1203, issue 1. Article no. 012065.

[Chollet, 2018] Chollet F., and Allaire J.J. (2018) Deep learning with R. MANNING Shelter Island. 400 p. (In Rus) = Glubokoe obuchenie na R. SPb.: Izd-vo Piter. 400 p.

Appendix 1. Direct encoding of words



```
example ← c("We will not ever know the truth ") # vector containing the message

index ← list() # word index sheet
for (sample in example) # the cycle of extracting words from a vector and indexing
  for (word in strsplit(sample, " ")[[1]])
    if (!word %in% names(index))
      index[[word]] ← length(index) + 2

len ← 7

output ← array(0, dim = c(length(example), len, max(as.integer(index)))) # the array with index-separated words

# the vectorized correlation of words in the message
for (i in 1:length(example)){
  sample ← example[[i]]
  words ← head(strsplit(sample, " ")[[1]], n = len)
  for (j in 1:length(words)){
    index ← index[[words[[j]]]]
    output[[i, j, index]] ← 1
  }
}
```