

Information System for the Intellectual Assessment Customers Text Reviews Tonality Based on Artificial Neural Networks

Nicolay Rudnichenko¹ [0000-0002-7343-8076], Svetlana Antoshchuk¹ [0000-0002-9346-145X],
Vladimir Vychuzhanin¹ [0000-0002-6302-1832], Andrii Ben² [0000-0002-9029-3489],
Igor Petrov³ [0000-0002-8740-6198]

¹Odessa National Polytechnic University, Odessa, Ukraine
nickolay.rud@gmail.com, vint532@yandex.ua, asgonpu@gmail.com

²Kherson State Maritime Academy, Kherson, Ukraine
a_ben@i.ua

³National University "Odessa Maritime Academy", Odessa, Ukraine
firmness@list.ru

Abstract. This article presents the results of the concept development and software information system for assessing text data tonality implementation by users based on artificial neural networks. The main problems in this topic are identified, the features of using deep machine learning for the text data mining problems are presented. An information system project has been developed, the preprocessing procedure and data filtering algorithms have been described, the specifics of data normalization for formalizing artificial neural network models are formalized. The options for using the information system, the block structure, the interface prototype and the procedure for user interaction with the software application are developed. The training effectiveness study results and the use of an artificial neural network model to solve the tasks are presented, the most suitable values of hyperparameters that have a primary impact on the model quality are identified and selected.

Keywords: machine learning, big data, data mining, data science, neural networks, deep learning, nature language processing

1 Introduction

Currently, in the Internet there is a rapid increase in the volume of heterogeneous data, which is associated with the development and dissemination of social networks, online stores, thematic blogs and information web systems, which significantly affects the electronic commerce various areas activity and trade in various electronic goods (EG) in particular [1].

In connection with the regular appearance and active development of new commercial and information resources, modern consumers of virtual and physical goods

and services are increasingly experiencing difficulties in choosing companies, organizations, manufacturers of technical gadgets and tools specific models [2].

This creates the need for additional information about the actual functionality and features of the EG operation from other users and experts. Additional difficulties are introduced by the need for filtering and analysis of marketing activities of competing manufacturing companies to identify the most suitable goods and services for the specific user's needs, which requires a large number of data computational operations [3].

In order to obtain competitive advantages and for better understanding customer's needs vendors also have to obtain the most reliable and relevant data extracted from large amounts of information based on user opinions analysis [4,5].

A partial solution to the identified problems is represented by existing systems and information resources, aggregating text reviews, comments and comparative video reviews of the characteristics and specifics of using EG in different conditions and modes [6].

However, these information platforms do not always have a flexible, convenient and informative interface, a thin search system and visualization of summary statistics with the formation of aggregated and crosstab reports [7,8]. The analysis of the data posted on such information resources is often difficult due to the need to view interesting reviews and comments on products in manual mode, which is associated with large time costs, i.e. analyzing user-generated opinions on the goods and services offered is a relevant and time-consuming process [9,10].

In this regard, it is advisable to automate the evaluation process suitable for the user EG, according to his individual preferences, by searching and analyzing the collected data characterizing various products on the basis of solving the classifying problem with semantic content into relevant groups.

To solve this task in practice, natural language processing (NLP) existing approaches are used, in particular, methods for analyzing the text's tonality, morphological analysis of its constituent entities, and evaluating expressions emotional coloring [11]. Sentiment analysis refers to the use of computational linguistics to identify and extract subjective information in source materials [12].

Existing approaches to the analysis of text's tonality are divided into the following main categories: definition of keywords, lexical similarity, statistical and conceptual methods [13].

2 Description of Problem

In general terms, the task of user reviews types determining for purchased goods is not fully clear and unambiguous, therefore it is realized by classifying them into separate groups in a linguistic form.

In various works on the classification of user reviews for various modern products on existing information resources, both standard text classification methods and modified methods are often used, which take into account the possible inversion of the

valuation word values, the syntactic structure of sentences, the dependencies between words [14].

The specificity and main difficulty of applying the classic NLP methods for different sets of user reviews is the need to collect enough adequate data to train the selected classifier model, to perform a number of laborious preparatory procedures for data preprocessing and cleaning to ensure an acceptable level of accuracy and speed of use. In this regard, it is advisable to analyze modern promising approaches to the classification of texts.

Currently, in practice, 2 approaches are used to solve the problem: methods based on logical rules and machine learning [15].

According to the results of a comparative analysis of the algorithms [16-19], the ANN method was chosen as one of the most used in practice and promising in implementation. An additional advantage of this method is the high functionality of existing libraries for the neural network models implementation from Google, their constant support and updating, which will provide opportunities for improving the system in the future.

Existing solutions in the text content analysis market have significant limitations in the amount of input data for processing, do not provide flexible settings for collecting and processing text in different languages, and do not allow evaluating the accuracy of reviews taking into account semantic topics [20-26].

In this regard, the urgent task is to develop our own information system (IS) that implements the functionality for evaluating user feedback on EG.

The purpose of the work is to study the possibilities of using the apparatus of artificial neural networks to assess user preferences for groups of acquired goods by automating their opinions analyzing process based on the classification problem solution.

The task of classifying text information is defined as follows. Let a document description exist $d \in X$, where X - vector document space, and a fixed set of classes $C = \{c_1, c_2, \dots, c_m\}$. From the training set (many documents with previously known classes) $D = \{\langle d, c \rangle \mid \langle d, c \rangle \in X \times C\}$ using the learning method G it is necessary to obtain a classification function $G(D) = \gamma$, which maps documents to classes $\gamma: X \rightarrow C$.

3 Information system development

3.1 System concept

The concept of the developed system is based on a combination of statistical methods of intelligence analysis and data preprocessing, as well as the artificial neural networks (ANN) theory [27].

The classification problem specificity under consideration is to carry out the following procedures for the text data preprocessing:

- Bringing all characters found in the text to lowercase in order to reduce the total unique number of terms in the dictionary.
- Exclusion of non-literal characters from the text. Such a procedure significantly reduces the number of unique terms in the dictionary, in cases where the text is characterized by an abundance of punctuation that does not carry a fundamental semantic load. In the considered problem, this can significantly reduce the amount of computational operations.
- Duplicate characters exclusion. This allows us to replace existing in the text sequences of identical characters to reduce the dictionary size.
- Isolation of the word base from a input text data set (stemming).

The listed actions are performed before the text classification process in order to increase the speed and reduce the iterative and logical complexity of data processing.

A formal description of the proposed classification concept in a schematic form of decomposition is shown in Fig. 1.

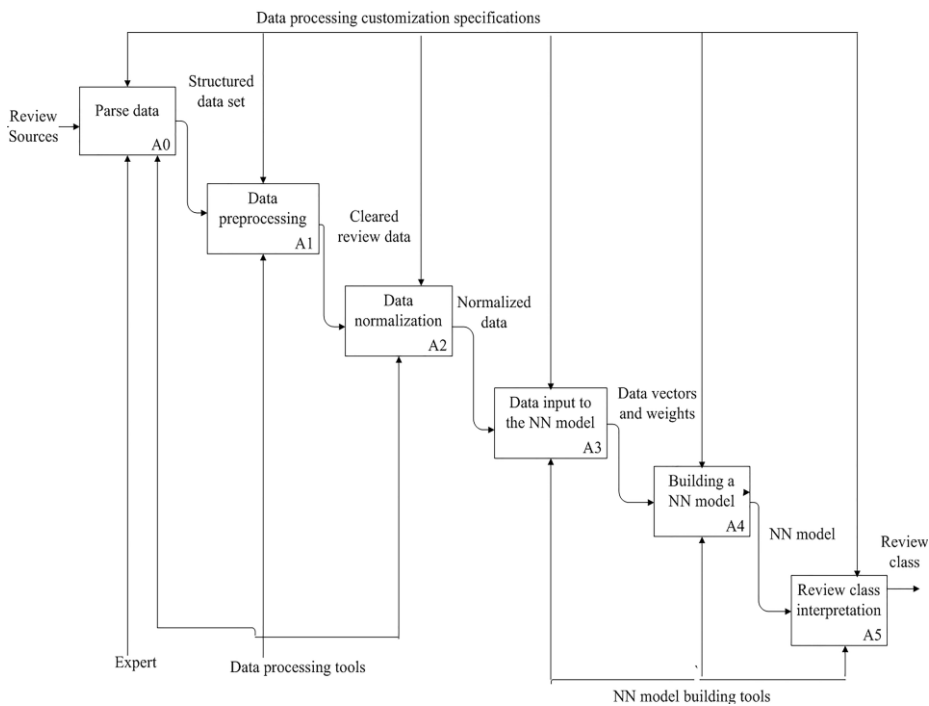


Fig. 1. Formalized system concept

The first of the system concept indicated stages consists in parsing data from the specified local or remote sources, on the basis of which a training sample is generated for the ANN model.

The second stage consists in filtering data by language and cleaning out extraneous characters that do not carry a semantic load in the recall (punctuation marks, unions,

special characters). As a result, by means of vector semantics operations, a vector representation of given dimension text feedback words is formed for further use by the ANN model. In general terms, the second stage of the proposed method is shown in Fig. 2.

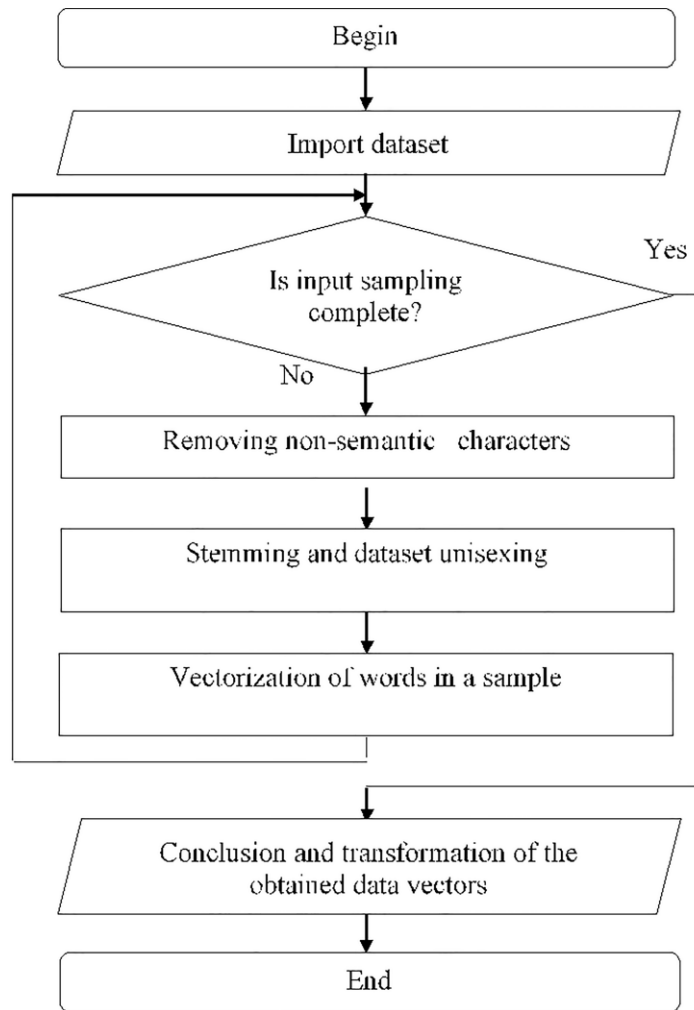


Fig.2. Data filtering algorithm

The third stage is to bring all the numerical values of the text' vector representation to the same area of change, whereby they are reduced to a single set of training data for the neural network model for classifying reviews.

Actual, the procedure for normalizing input data is being implemented to convert all elements of the input data set into binary code, which is acceptable for further processing by an artificial neural network.

A generalized algorithm for normalizing data when creating a neural network is shown in Fig. 3.

The minimax function performs a linear transformation after determining function's minimum and maximum values so that the obtained values are in the desired range from -1 to 1.

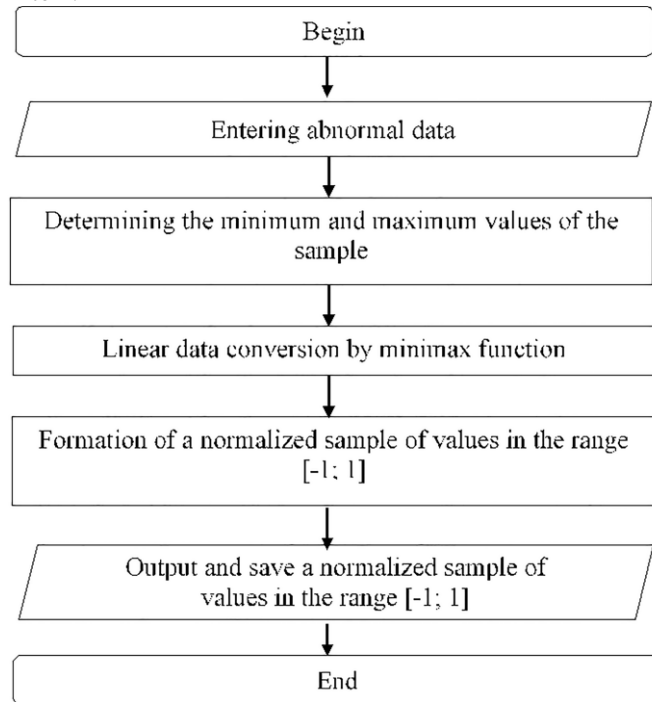


Fig.3. Generalized data normalization algorithm

The fourth step is to break down the processed data sets into separate blocks for model training, testing and validating, taking into account the nature of the data in a given ratio.

The fifth step is to initialize the ANN model to classify text reviews into three different classes (positive, neutral, and negative). Initialization of a neural network model is the process of creating a neural network object, loading a normalized data set, initializing the learning process and model saving, which is based on the recursive ANN models usage.

The sixth step is to numerically evaluate the accuracy of the created ANN model to solve the text tonality assessing problem. Conducting calculations, on the basis of the supplied text string by the user, created model analyzes the input data and classifies them according to the available classes. It is a test of the ANN model operation on a test sample. At the same time, the result of the classification is converted, the obtained values are translated into a text view that is understandable for the user. This stage is based on the use of a reliability metric to determine the proportion of correctly classi-

fied text reviews and the loss function to assess the dependence of training accuracy on the weight matrix coefficients.

To ensure convenient and efficient operation IS implements the proposed concept, it is necessary to introduce a number of restrictions. Due to the fact that text reviews are of different sizes and carry different semantic load, and processing too large text fragments can be time-consuming and expensive in computational resources terms, it is advisable to limit their volume. In particular, the program should support the ability to analyze the text in Russian, Ukrainian and English, the total text should be up to 2000 characters, the analysis should not exceed 10 seconds.

The IS input receives text data of user comments and reviews, as a result of processing, a text classes table is formed, estimation accuracy level (classification error by the ANN model), a summary statistics form, and a file with output classification results in * .xls format are calculated.

The main stages of the project are as follows:

- Development of a parser software module for searching, receiving, and collecting a data set to form a ANN training samples.
- Filtering data by language and cleaning extraneous characters that do not carry a semantic load in the recall.
- Export of the obtained sample to the *.csv format for import into the neural network structure.
- Creation and configuration ANN structure, the selection of training algorithms and its work evaluation.
- IS graphical user interface development that includes the functions of entering a text commentary and viewing the classification result.
- Text evaluation in one of the possible recall classes.

The stage of creating and configuring a neural network in a more detailed form is divided into a number of the following tasks:

- Getting the input string (array of strings) is the process of writing a input text data set into a variable.
- Input data normalization, for converting all data set elements into binary code, which is acceptable for further processing by an ANN.
- ANN model initialization is the process of creating a neural network object, loading a normalized data set and initializing the learning process and saving the model.
- Conducting calculations, based on the user-supplied text strings of feedback, the ANN model analyzes the input data and classifies them according to the available classes.
- Transformation of the result of the classification (denormalization), translation of the obtained values into a text form, understandable for the user.
- The output of the obtained value during the execution of this stage in the user interface displays the classification result.

As the development language we used Python 3.7, which is expanded by the following data structure processing libraries: Numpy, to support the use of multidimensional data arrays and implement the necessary mathematical functions number for their processing; Pandas, for the implementation of modeling and analysis functions during data processing and normalization.

To normalize and denormalize the data, create, configure and train the ANN model, the keras library and its components are used (tokenizer, TensorBoard, LSTM modules).

The PyQt library and the QtDesigner module were used to create a graphical user interface, layout the necessary widgets and elements of the program form.

3.2 System project implementation

When forming requirements for the created IS, a use-case diagram was developed (Fig. 4), in accordance with which the requirements for user roles are formalized (represented by a typical user and system administrator).

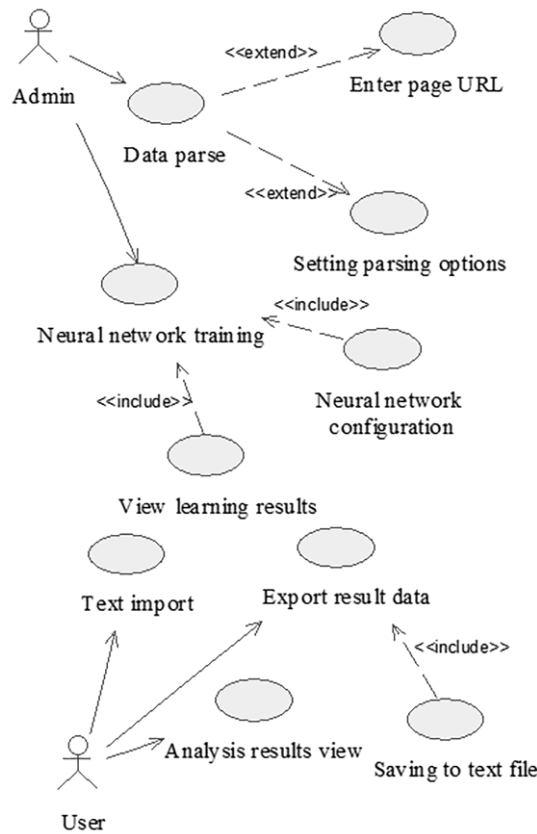


Fig. 4. System's use case diagram

The user should have the following options for interacting with IS through a graphical interface (form):

- entering and editing the corresponding text review within the corresponding text field;
- viewing the result of the review class analysis (positive, negative, neutral);
- exporting the result to a text file.

The administrator has the ability to parse data from the specified page URL and set additional parameters for parsing, as well as configure and train the ANN with viewing the results. Based on the analysis and determination of IS requirements, a block structure has been developed (Fig. 5).

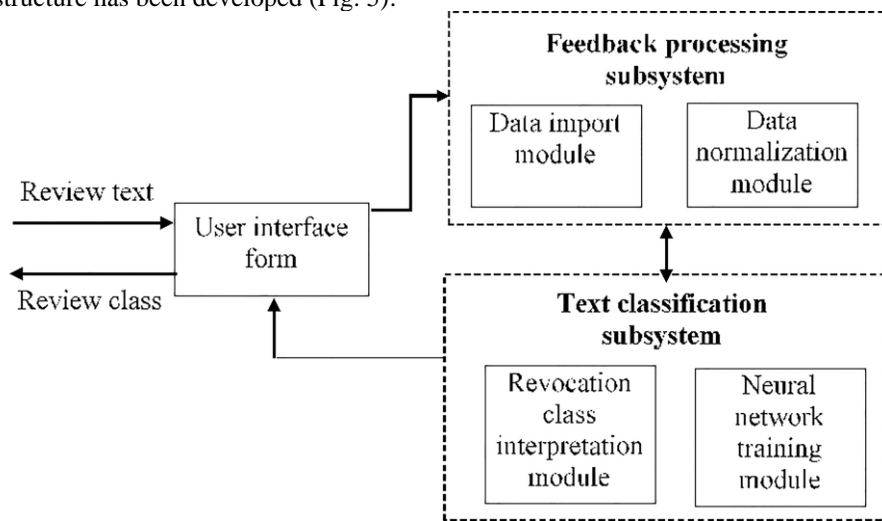


Fig.5. IS block structure

The designed IS includes the following components:

- Subsystem for processing text reviews (data import module and normalized module for imported data).
- Subsystem for classifying user reviews text (neural network training module and module for interpreting the recall class).
- The form of the graphical user interface.

For convenient user working process with IS, the arrangement of widgets on the form is done in an adaptive style, when resizing the working window, their location is scaled in proportion to the screen resolution. IS main form graphical user interface is shown in Fig. 6. The upper part of the form displays informational messages about the application process, which are automatically saved as an event log in a *.txt file if it is necessary to track errors or incorrect data processing by the system. The classification results are displayed in tabular form, for a detailed view of the review text, user must select the appropriate line.

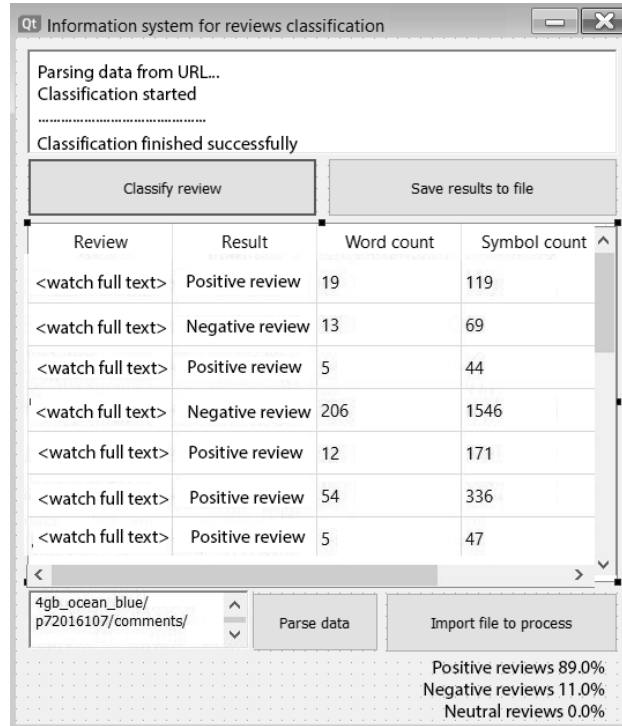


Fig. 6. Software GUI form

At the bottom of the form is the input field for the source web page URL, as well as text labels that display the percentage of positive, negative and neutral reviews.

4 Experiments and results analysis

To carry out a study created IS functioning specifics on the use of artificial neural networks, a test texts selection for EG from a number of popular online stores was prepared and aggregated: 120,000 texts (30,000 texts for each of the possible classes).

The sample was obtained through the development of a specialized data parser that performs filtering and data cleaning. The assignment of class types for each record was carried out manually. The entire volume of the text reviews obtained sample was divided into training, test and validation sets (60%, 20% and 20%, respectively) in order to evaluate the quality of the model. As part of the IS research process, classification accuracy was assessed, i.e. the number of correctly classified text user reviews. As the numerical characteristics of the performance assessment IS used:

- ACCURACY is a confidence metric that allows us to evaluate the classification accuracy, i.e. determine the proportion of correctly classified texts.

- LOSS is a function of losses during neural network operation, this indicator illustrates the dependence of training accuracy on the weight matrix coefficients.

To conduct numerical studies of the created neural network model use framework of the developed information system and obtained results graphic the Tensor Board data analysis tool was deployed. The dependence of the value of assessing the reliability of the neural network (ordinate axis) by the passed training eras (abscissa axis) is shown in Fig. 7.

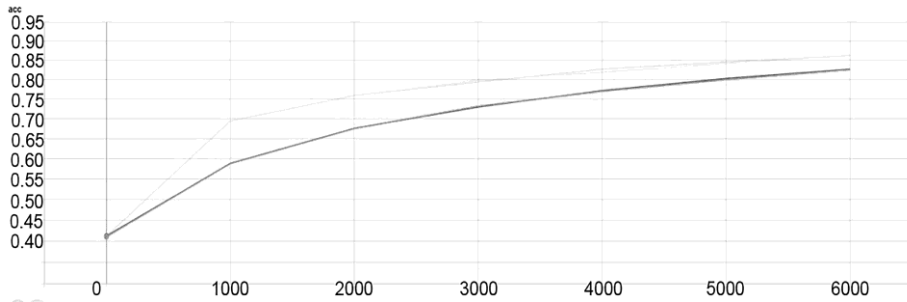


Fig. 7. Neural network reliability assessing value

A thin line marks the results of a training sample of reviews, and a thick line shows the results of using a neural network in a test sample. The overall accuracy of the created neural network was about 89%. The dependence of the values of the loss function on the epoch of neural network training is shown in Fig. 8.

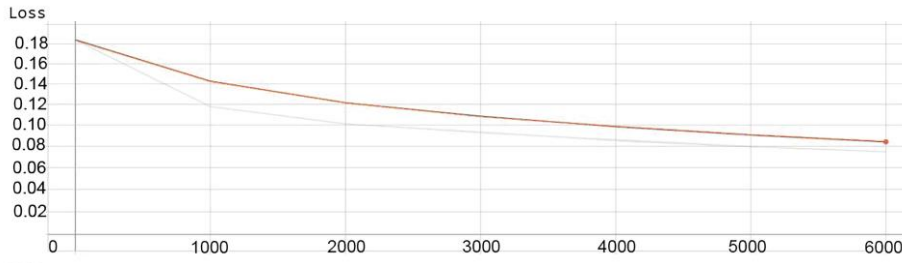


Fig. 8. Neural network training loss function dependence

In order to study the possibility of improving the quality of the solution to the classification problem created by an artificial neural network (text feedback submitted to it at the input), it is advisable to evaluate the performance of the developed neural network model for various values of a number of its parameters. As variable parameters were used: max_features, maxlen and batch_size. The results of the model assessment for various parameters are given in table 1. The best result of the Accuracy value (0.92) was obtained with the following parameter values: max_features - 7000; maxlen - 100; batch_size - 64. Based on the analysis of the ANN model characteristics with various parameter values, the dependence of the neural network operation accuracy metric value and the max_features model parameter was studied (Fig.9).

Table 1. The results of the model assessment

Test Number	Accuracy	max_features	maxlen	batch_size
1	0.72371	3000	30	8
2	0.76229	3500	40	16
3	0.79847	4000	50	24
4	0.81451	4500	60	32
5	0.85012	5000	70	64
6	0.84832	5500	80	128
7	0.86233	6000	50	256
8	0.87431	6500	60	8
9	0.87985	7000	70	16
10	0.88615	7500	80	24
11	0.89132	8000	50	32
12	0.88434	8500	60	64
13	0.88934	9000	70	128
14	0.89091	9500	80	256
15	0.88347	10000	85	64
16	0.90831	5000	90	128
17	0.91217	6000	95	256
18	0.92331	7000	100	64

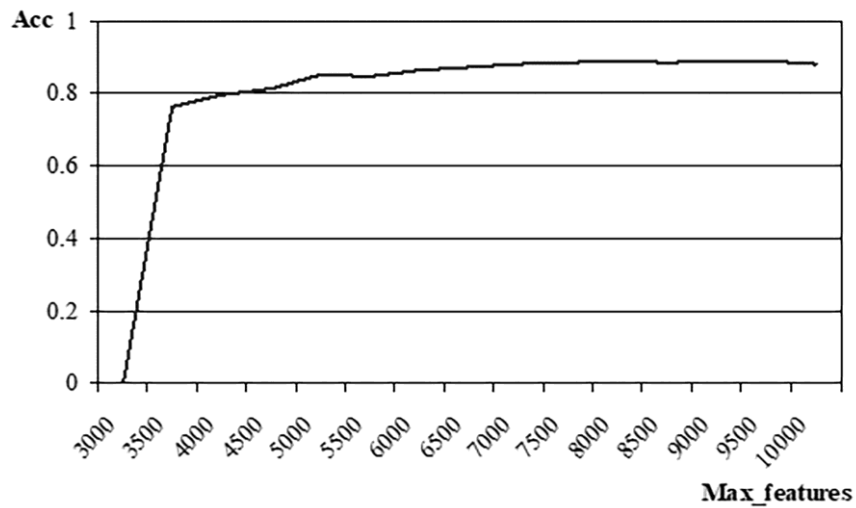


Fig. 9. ANN accuracy and max_features dependence

It should be noted that the classification confidence level increases with the increase in max_features; the peak is reached in the range from 5500 to 8000. As a result of a IS operation study based on a neural network (a selected recurrent architecture of the LSTM type), classification accuracy of about 92% was achieved.

This allows us to conclude that for text reviews of the specifics examined in the EG field, the most significant ANN parameters from the point of view of influence on classification accuracy are the weight matrix rewriting border size and the number of words in the reviews text sample, the maximum length of one review is less important. With batch_size = 64, the highest accuracy is achieved.

5 Conclusion

The developed information system implements the proposed concept of assessing the tonality of electronic goods reviews and is a cross-platform solution providing a fairly high classification accuracy of more than 90%, which indicates the reliability of the solution to the problem.

Based on the results of the user reviews classification, it becomes possible to form an aggregated integrated indicator for evaluating the relevant goods, which can be used to prioritize the customers preferences in a ranked form in order to support and facilitate decision-making processes for choosing and buying.

Large trading floors can use the results of evaluating user opinions to analyze and select the most reputable and reliable vendors for further cooperation or stopping purchases from suppliers whose products are regularly criticized by customers.

The subsequent logical development of the proposed approach to the classification of user reviews is the integration of analysis mechanisms for the reliability of data sampling in order to cut off noise and non-informative data, expanding class types and implementing a number of quantitative indicators corresponding to them to clarify the estimates formed.

Reference

1. Rudnichenko, N., Vychuzhanin, V., Shybaieva, N., Shybaiev, D., Otradskaya, T., Petrov, I.: The use of machine learning methods to automate the classification of text data arrays large amounts. Information management systems and technologies. Problems and solutions. Ecology, Odessa, pp.31-46 (2019)
2. Rudnichenko, N., Vychuzhanin, V., Shybaieva, N., Shybaiev, D.: Big data intellectual analysis in the diagnosis of the transportation systems technical condition. Systems and means of transport. Problems of operation and diagnostics. KSMA, Kherson, pp.57-69 (2019)
3. Rudnichenko, M.D., Gezha, N.I., Belyaev, K.O., Kuzmin, A.D.: Performance analysis of machine learning model ensembles. In III All-Ukrainian scientific-practical conference of young scientists, students and cadets "Information protection in information and communication systems". Lviv. pp.259-260 (2019)

4. Adaskina, Yu. V., Panicheva P.V., Popov, A. M. : Sentimental analysis of tweets based on syntactic links. In computer linguistics and intellectual technologies: based on the materials of the annual international conference Dialogue. Moscow, pp.25–35 (2015)
5. Vasiliev, V.G., Khudyakova, M.V., Davydov, S.: Classification of user reviews using fragment rules. In Computational Linguistics and Intellectual Technologies: Based on the materials of the annual Dialog International Conference. Moscow, pp.66-78 (2012)
6. Garshina, V.V., Kalabukhov, K.S., Stepansov, V.A., Smotrov, S.V.: Development of a system for analyzing the tonality of textual information. Gerald of VSU, series: system analysis and information technology. vol. 3, pp.185-194 (2017)
7. Lysenko, V.D.: Text sentiment analysis for forecasting stock market prices. Young scientist. vol. 22, pp.420-423 (2018)
8. Pavlov, Yu.N., Maystruk, K.A. : Comparison of text tonality assessment methods. Young scientist. vol. 12, pp.59-64 (2016)
9. Loukachevitch, N., Kotelnikov, E., Rubtsova, Y.: SentiRuEval: testing object-oriented sentiment analysis systems in Russia. In proceedings of International Conference Dialogue-2015, Moscow, pp. 313 (2015)
10. Rubova, Y.V.: Building a body of texts for tuning the tone classifier. Software Products and Systems. vol. 109, pp.72–78 (2015)
11. Menshikov, I.L., Kudryavtsev, A.G.: A review of systems for analyzing the tonality of a text in Russian. Young scientist. vol.12, pp.140-143 (2012)
12. Kotelnikov, E.V., Klekovkina, M.V.: Automatic analysis of tonality of texts based on machine learning methods. In Computational Linguistics and Intellectual Technologies: Based on the materials of the annual International Conference “Dialogue”. Moscow, pp.15–21 (2012)
13. Sboev, A.G., Voronina, I.E., Gudovskikh D.V., Selivanov, A.A.: Advanced neural network models for solving the problem of determining tonality. Bulletin of the Voronezh State University. System Analysis and Information Technology. vol. 4, pp.178–183 (2016)
14. Gorban, A.N.: Training of neural networks. Moscow: ParaGraph (2010)
15. Shybaiev, D.S., Otradsкая, T.V., Stepanchuk, M.V., Shybaieva, N.O., Rudnichenko, N.D.: Predicting system for the estimated cost of real estate objects development using neural networks. ZhSTU Herald. Technical science. vol.83, pp.154-160 (2019)
16. Silge, J., Robinson, D.: Text Mining with R: A Tidy Approach. O'Reilly Media (2017)
17. Chaudhuri, A.: Visual and Text Sentiment Analysis through Hierarchical Deep Learning Networks. Springer (2019)
18. Aggarwal, C.C.: Machine Learning for Text. Springer (2018)
19. Rudy P., Thelwall, M., Sentiment analysis: A combined approach. Journal of Informetrics. vol. 3, pp.143-157 (2009)
20. Asad, A., Siti, M. S., Shafaatunnur H., Jalil P.: Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment. Expert Systems with Applications. vol. 109, pp.66-85 (2018)
21. Yi, C, Qingbao, H., Zejun, L., Jingyun, X, Zhenhong, C., Qing, L.: Recurrent neural network with pooling operation and attention mechanism for sentiment analysis: A multi-task learning approach. Knowledge-Based Systems (2020)
22. Saerom, P., Jaewook L., Kyoungoo K.: Semi-supervised distributed representations of documents for sentiment analysis. Neural Networks. vol.119, pp.139-151 (2019)
23. Wang, J., Tao, Q.: Machine Learning: The State of the Art. IEEE Intelligent Systems. vol.23, pp. 49-55 (2008)
24. Rahul, A., Surabhi, M.: NLP based Machine Learning Approaches for Text Summarization. pp.535-538. (2020)

25. Hung, C.C., Song, E., Lan, Y.: Foundation of Deep Machine Learning in Neural Networks (2019)
26. Wu, Z., Ding, X., Xu, X., Ju, C.: ECG arrhythmias classification based on deep learning approach. ICIC Express Letters, Part B: Applications. pp.843-850 (2017)
27. Miikkulainen, R. Topology of a Neural Network (2011)