Machine learning application for support for automated control systems users

S K Khromov\textsuperscript{1}, M A Kulagin\textsuperscript{1} and V G Sidorenko\textsuperscript{1,2}

\textsuperscript{1} Federal State Institution of Higher Education «Russian University of Transport» (RUT – MIIT), Moscow, Russian Federation
\textsuperscript{2} Higher School of Economics, Moscow, Russian Federation

Abstract. The article presents the results of the analysis of determining the possibility of Machine Learning (ML) using for solving the problems of incident classification of users on the example of enterprise resource planning (ERP) systems of JSC Russian Railways and choosing a rational method for solving this problem. The presented problem is a special case of the Natural Language Processing Problem (NLP) with the use of neural networks. The article covers findings of classification of user queries: binary and multi class, based on semantic analysis. In 2018, 1.4 million incidents were reported from users of ERP systems, only 25% of them are suitable for ML applications. Resolving other incidents requires the participation of a support specialist. Implementation of the software will reduce the waiting time for incident processing by 64% of the average. This will increase staff productivity by 5-7%. This paper explicitly proves that using convolutional neural networks for user incident classification problem has high potential in JSC Russian Railways.

1. Introduction

Interest of the President of the Russian Federation and the Government in the field of information technology (IT), strongly promotes the introduction of innovative techniques like Machine learning (ML) in Russian state-owned companies, including JSC "Russian Railways". Machine learning is already widely used in driving solution for the following problems:

- ad targeting and sorting articles by category,
- telecommunications network management,
- information security monitoring,
- spam filtering,
- classification of user incidents [1].

The solution suggested in the article [1] is a full-fledged service desk platform with ML elements. Some companies already operate their own service desk systems. Russian Railways is one of these companies. Because of high price of the IT solution for a full-fledged service desk platform with ML elements, the We suggests implementing their own software. As an additional remark it is worth mentioning that the software mentioned above can be integrated with other systems.

2. Task

In this paper, the authors analyze the possibility of using ML to solve the problem of classifying user incidents of automated control systems of JSC "Russian Railways". This is necessary to provide users with responses to incidents. For example, in 2018, users created 5.6 million incidents. A quarter of them are incidents for ERP systems and these hits make the datasets for further analysis. Automation process with the help of machine learning techniques helps to improve cost and time efficiency.

In a sample of 36 thousand queries, the existing model was able to classify positively 80% of incidents with a probability higher than 75%. The implementation of the developed software will reduce the waiting time for the solution of the considered incidents of users of ERP systems of JSC "Russian Railways" by an average of 10.1 hours or 64%. This will save in total 60 thousand hours.
3. **Approach**

As already mentioned, the system includes four groups of automated business processes. It is proposed to define business processes for each of the groups of business processes. The paper deals with three of groups.

The problem of multi-classification is as follows: it’s necessary to defined a set of pairs "object, class" $X \times Y$ based on knowledge of the finite training sample of observations $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. It is required to construct an algorithm $\alpha: X \rightarrow Y$, which is able to classify an arbitrary object $x \in X$. In the technological problems being solved, the "object, class" pair is "text, class", where the class is determined based on what type the text belongs to in terms of semantic content. In other words, the classical problem from the field of Natural Language Processing (NLP) is solved [2, 3, 4]. To solve the set technological tasks, it was decided to use deep neural networks, in particular, convolutional and recurrent. This solution is based on the success of these algorithms to solve text classification problems [5-8]. A separate neural network was trained for each task.

The algorithm for training and selecting a neural network model is shown in figure 1.

![Diagram of the learning process and model selection.](image)

1. Tokenization and normalization
2. Lemmatization and deletion of stop words
3. Data pre-processing
4. Models training using cross-validation
5. Evaluating model performance
6. Choosing the best model

In block 1, texts are divided into words and other components (tokenization) and punctuation and other symbols are eliminated (normalization) [9].

In block 2, data is cleared from words that do not carry any semantic load (stop words, in Russian-these are interjections, conjunctions, etc.) using a ready – made library [10]. At this stage, words are lemmatized using a ready-made library [11], i.e., words are reduced to the semantic and canonical form of the word. For example, "prices – price".
In block 3, data is pre-processed to convert it to numeric attributes. The following methods were used to convert text to a number: "word bag" (BOW) [12], TF-IDF [13,14], N-Gramm (N-gramm) [15], and Embedding. The first three methods of converting text to numbers were used for classical machine learning algorithms, and the last one was used for deep neural networks.

In block 4, algorithms are trained using various optimization methods. The amount of training data is small for NLP tasks. Therefore, the training was performed using cross-validation. The main models used were:

- logistic regression (LogReg),
- nearest neighbor method (kNN),
- single-layer neural network (NN),
- convolutional neural network (CNN),
- recurrent neural network (RNN).

In block 5, the quality of each of the machine learning algorithms is evaluated. Quality control is performed on a validation sample that did not participate in the training. Metrics are used for evaluation:

- Accuracy (percentage of correct algorithm response),
- Precision (the proportion of the really correct choice of a class to the total number of definitions of the same class by the algorithm in the test sample),
- Recall (the percentage of objects with the specified class found by the classifier among all objects with the specified class in the test sample),
- F-measure (the harmonic mean between Precision and Recall).

In block 6, the analysis of each algorithm in combination with the text processing method was performed.

The following data sets were used for each task:

- in Task 1, we needed to build a classifier that divides the sample into 16 classes. The class distribution is shown in Figure 2. 45831 records were used for deep neural network training, and 19642 records were used for testing.
- in Task 2, we needed to build a classifier that divides the sample into 13 classes. 45778 records were used for deep neural network training, and 1,920 records were used for testing.
- in Task 3, we needed to build a classifier that divides the sample into 16 classes. 137496 records were used for deep neural network training, and 58927 records were used for testing.

The Embedding method is used for preprocessing data and converting it to numeric attributes. To limit the feature vector used when converting text to a number using Embedding, a limit of 31, 41, and 36 words was selected for each task, respectively. The restriction was chosen by the authors based on the fact that the quantile at the level $\alpha = 0.95$ for the distribution of record lengths is provided for the following numbers of words $X^{(1)}_\alpha = 31$, $X^{(2)}_\alpha = 41$, $X^{(3)}_\alpha = 36$. The distribution of samples by class and the distribution of the length of rows used for training for task 1 are shown in figure 2.

4. Results

The results of training on the test set are presented in Tables 1 and 2. The main models used were:

- convolutional neural network (CNN);
- recurrent neural network (RNN) [25].

In Table 1, the quality of the algorithm is measured using the Accuracy metric [8]. There is a problem of unbalanced classes in the sample. Therefore, it is necessary to use a more objective quality metric-F-measure [5]. The results of the algorithms are presented in Table 2.
Figure 2. Distribution of samples by class for Task 1.

Figure 3. Distribution of the length of rows that were used for training for Task 1.

Table 1. The results of neural networks (Accuracy metric)

| The algorithm of neural network | The type of record | Task 1  | Task 2  | Task 3  |
|--------------------------------|--------------------|--------|--------|--------|
| 1                              | Convolutional      | Unbalanced | 0.8679 | 0.8104 | 0.7479 |
| 2                              | Convolutional      | Balanced  | 0.8661 | 0.8919 | 0.7435 |
| 3                              | Recurrent          | Unbalanced | 0.8537 | 0.8001 | 0.7403 |
| 4                              | Recurrent          | Balanced  | 0.8531 | 0.8897 | 0.7453 |

Table 2. The result of the best neural network (f-measure metric)

| The number of class | Task 1 | Task 2 | Task 3 |
|---------------------|--------|--------|--------|
| 1                   | 0.8591 | 0.8597 | 0.4936 |
| 2                   | 0.0    | 0.2586 | 0.7815 |
| 3                   | 0.6563 | 0.5903 | 0.6047 |
| 4                   | 0.8987 | 0.9123 | 0.6989 |
| 5                   | 0.7997 | 0.7245 | 0.8361 |
| 6                   | 0.8664 | 0.8543 | 0.8267 |
| 7                   | 0.7622 | 0.5229 | 0.7548 |
| 8                   | 0.1818 | 0.0    | 0.6978 |
| 9                   | 0.8440 | 0.2975 | 0.6296 |
| 10                  | 0.8507 | 0.8298 | 0.7827 |
| 11                  | 0.9069 | 0.2882 | 0.7143 |
| 12                  | 0.9346 | 0.1031 | 0.5461 |
| 13                  | 0.2227 | 0.3681 | 0.5873 |
| 14                  | 0.8784 | -      | 0.8659 |
| 15                  | 0.8952 | -      | 0.7766 |
| 16                  | 0.1756 | -      | 0.4    |
For classes determination in Task 1 and Task 2, convolutional neural networks, trained on an unbalanced sample, showed the best results from the point of view of the F-measure. For classes determination in Task 3, the best one was a convolutional network trained on a balanced sample. As a result of research on the data obtained, the following conclusion was made: to improve the value of quality indicators, additional data collection on minor classes is required.

The results of the classification will be used to determine type of guidance materials which need users experiencing difficulties. Depending on the percentage of users who can use the materials to solve their own issue, the effect will be different. However, not all business processes were defined with sufficient accuracy (BP with accuracy below 50% is excluded from the calculation). An example of effect calculating for Task 1 is shown in Table 3 based on statistics on the input flow of user incidents for 2018. Similarly, the effect is defined for the other two Tasks (2, 3) and summed.

| Business process | Percentage of correctly defined incidents | The number of incidents | The average labor costs (man hours) | The average waiting time (hours) | 5% | 10% |
|------------------|--------------------------------------|------------------------|----------------------------------|-------------------------------|----|-----|
| 15               | 93%                                  | 18305                  | 0:29                             | 18:02                         | 851| 332:23| 1702| 664:47| 16307:39|
| 14               | 84%                                  | 1188                   | 0:30                             | 16:38                         | 50 | 20:35| 36:45| 100 | 41:10| 73:30|
| 12               | 91%                                  | 4269                   | 0:42                             | 19:17                         | 194 | 120:03| 930:12| 388 | 240:06| 1860:24|
| 11               | 95%                                  | 11563                  | 0:45                             | 19:27                         | 549 | 352:12| 7867:49| 1098| 704:25| 15735:39|
| 10               | 84%                                  | 1234                   | 0:34                             | 16:56                         | 52 | 24:40| 50:56| 104 | 49:20| 101:52|
| 9                | 80%                                  | 551                    | 0:39                             | 17:13                         | 22 | 12:50| 13:21| 44 | 25:41| 26:43|
| 7                | 69%                                  | 158                    | 0:39                             | 22:36                         | 5  | 4:50 | 1:05 | 11 | 9:40 | 2:11|
| 6                | 81%                                  | 492                    | 0:29                             | 18:57                         | 20 | 9:23 | 5:53 | 40 | 18:46| 11:46|
| 5                | 78%                                  | 1207                   | 0:31                             | 17:59                         | 47 | 23:21| 40:30| 94 | 46:43| 81:01|
| 4                | 89%                                  | 10140                  | 0:30                             | 12:57                         | 451 | 138:09| 2725:09| 902| 276:19| 5450:19|
| 3                | 58%                                  | 1034                   | 0:38                             | 12:50                         | 30 | 17:30| 44:40| 60 | 35:01| 89:20|
| 1                | 87%                                  | 1828                   | 0:36                             | 23:02                         | 80 | 52:37| 125:18| 159| 105:15| 250:37|

For ease of interpretation, the effect is presented in easy-to-understand units in table 4. Where the conditional technologist is 1445 people * 1970 hours of labor per year (in 2018).

5. Conclusion
Classifying user incidents and based on this rapid providing of guidance materials to users will increase the productivity of support staff by 5-7%.

As a result of the presented work, neural networks were developed for classifying incidents from users of the ERP. It is revealed that the use of convolutional neural networks for classifying user incidents has high prospects in JSC Russian Railways. The subject of further research is to improve
the accuracy of existing models and expand the range of algorithm of machine learning implementation.

| The effect value | The labor saving (man hours) | The value of saving waiting time (hours) | The effect value (technologist) | The reduction in waiting time (man years) |
|------------------|-------------------------------|------------------------------------------|--------------------------------|------------------------------------------|
| 5%               | 6864.2                        | 162071.6                                 | 4.8                            | 82.3                                     |
| 10%              | 13728.4                       | 324143.2                                 | 9.5                            | 164.5                                    |

6. References
[1] Artificial intelligence for classifying incidents in Service Now: https://it-guild.com/info/blog/iskusstvennyy-intellekt-dlya-klassifikatsii-zaprosov-v-servicenow/
[2] Powers D M 2011 Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation
[3] Kotsiantis S B, Zaharakis I D and Pintelas P E 2007 Supervised machine learning: A review of classification techniques Emerging artificial intelligence applications in computer engineering 160 pp 3–24
[4] Manning C and Schütze H 1999 Foundations of statistical natural language processing (MIT press)
[5] Ramos J 2003 Using tf-idf to determine word relevance in document queries Proc. 1st instructional Conf. on machine learning 242
[6] Matthias E, Vogel S and Waibel A 2005 Low cost portability for statistical machine translation based on n-gram frequency and TF-IDF Int. Workshop on Spoken Language Translation (IWSLT)
[7] Kohavi R 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection Ijcai 14 No 2
[8] Goodfellow I et al 2016 Deep learning (Cambridge: MIT press)
[9] Aw A, Zhang M, Xiao J and Su J 2006 A phrase-based statistical model for SMS text normalization Proc. COLING/ACL on Main conference poster sessions (Sydney: Association for Computational Linguistics) pp 33–40
[10] Manning C et al 2014 The Stanford CoreNLP natural language processing toolkit Proc. 52nd annual meeting of the association for computational linguistics: system demonstrations
[11] NLTK 3.4 documentation: https://www.nltk.org/
[12] Pymystem3 package: https://pythonhosted.org/pymystem3/pymystem3.html
[13] Zhang Y, Rong J and Zhi-Hua Zh 2010 Understanding bag-of-words model: a statistical framework Int. Journal of Machine Learning and Cybernetics 1.1-4 pp 43–52
[14] Ramos J 2003 Using tf-idf to determine word relevance in document queries Proc. of the first instructional conference on machine learning vol 242
[15] Matthias E, Vogel S and Waibel A 2005 Low cost portability for statistical machine translation based on n-gram frequency and TF-IDF Int. Workshop on Spoken Language Translation (IWSLT)