Bank transaction text label mining algorithms

A.S. Startseva[1], A.M. Vulfin[1], V.I. Vasilyev[1], A.V. Nikonov[1], A.D. Kirillova[1]

[1]Ufa State Aviation Technical University, K. Marks st. 12, Ufa, Russia, 450008

Abstract. The banking transaction monitoring system implements decision support mechanisms for online payment control procedures for legal entities considering the dynamic risk profile of the client. The system includes a set of algorithms for the intellectual analysis of transaction parameters, including a text label for the purpose of payment, and decision support for an employee of the financial monitoring unit. The development of algorithms for analyzing textual labels for the purpose of payments allows us to clarify the dynamic payment profile of the user and increase the validity of the recommendations of the monitoring system. A block diagram of a system for identifying high-risk banking transactions based on data mining algorithms has been developed. Algorithms for data mining of textual labels of the payment purpose have been developed and the effectiveness of the proposed solution on field data has been evaluated. An algorithm is proposed for the phased analysis of the text label of the payment destination, including the stages of preprocessing, filtering, normalizing and constructing a classifier based on a set of regular expressions and intelligent analysis technologies. The difference between the algorithm is the use of adaptive category dictionaries and the multi-pass application of heterogeneous classifiers, which makes it possible to increase the validity of the decision on whether the transaction belongs to one of the selected classes.

1. Introduction

Banking transaction monitoring system (TMS) [1] should implement decision support mechanisms for online payment control procedures for legal entities taking into account the dynamic customer risk profile. The system should [2] include a set of algorithms for intelligent analysis of transaction parameters, including a text label for the purpose of payment, and decision support for an employee of the financial monitoring unit. Existing developments [3] allow taking into account parameters of the user environment and forming a vector of signs for detecting fraudulent actions, but it is important to improve the payment analysis algorithms for legal entities with dynamic risk profile. Development of the algorithms for analysis text labels of purpose of payment will clarify user dynamic payment profile and increase the validity of monitoring system recommendations.

The goal of the work is increasing the efficiency of the system for identifying high-risk transit operations on the basis of intelligent analysis of the bank transaction text labels.

To achieve this goal, the following tasks were set:

- development of a structural and functional diagram of the system for identifying high-risk banking transactions on the basis of data mining algorithms;
- development of algorithms for identifying high-risk banking transactions based on data mining algorithms in application to the textual labels of payment purposes and evaluating the effectiveness of the proposed solution on the field data.

2. Analysis of existing approaches to the task of identifying high-risk transit operations

Transaction monitoring systems [4] allow to solve applied business tasks of identification and preventing fraudulent transactions; in addition, they serve as a source of primary information about identifying intruders and theft.

A comparative analysis of the most famous antifraud systems is presented in [5-7]. Most Russian antifraud systems are implemented using the set of signature rules "IF-THEN", while fraud scoring is not a popular feature for either Russian or foreign transaction monitoring systems.

User environment analysis is gaining popularity. The "FraudWall" system [8, 9], studied in this work, is a product of Frodex company [6], that is already includes a user environment analysis module.

Decision-making support for a financial monitoring unit employee on procedures of online monitoring payments of legal customers (with taking into account the dynamic risk profile of the client) should be based on the use of a set of algorithms of the intelligent analysis of transaction parameters, including a text label for the purpose of payment. Development of such algorithms will make it possible to clarify the dynamic user payment profile (the totality of data about all user's payment transactions, normalized and sorted by type of activity to which they relate according to "OKVED" – Russian National Classifier of Types of Economic Activity) and increase the validity of the decision made by the monitoring system [10-16].

The main problem of constructing a system for classifying the purpose of payments is weak requirements of text label formalization by the bank, as well as the length of the text label itself. It is necessary to form semantic spaces of each of the categories indicated in regulatory documents in order to improve the performance indicators of classifiers.

3. Development of a structural and functional diagram of a banking transaction monitoring system with an intelligent text label analysis module

Technologies of remote banking services (RBS) for accessing accounts and transactions via a web browser do not require client software installation and are very widespread [17-26] (figure 1).

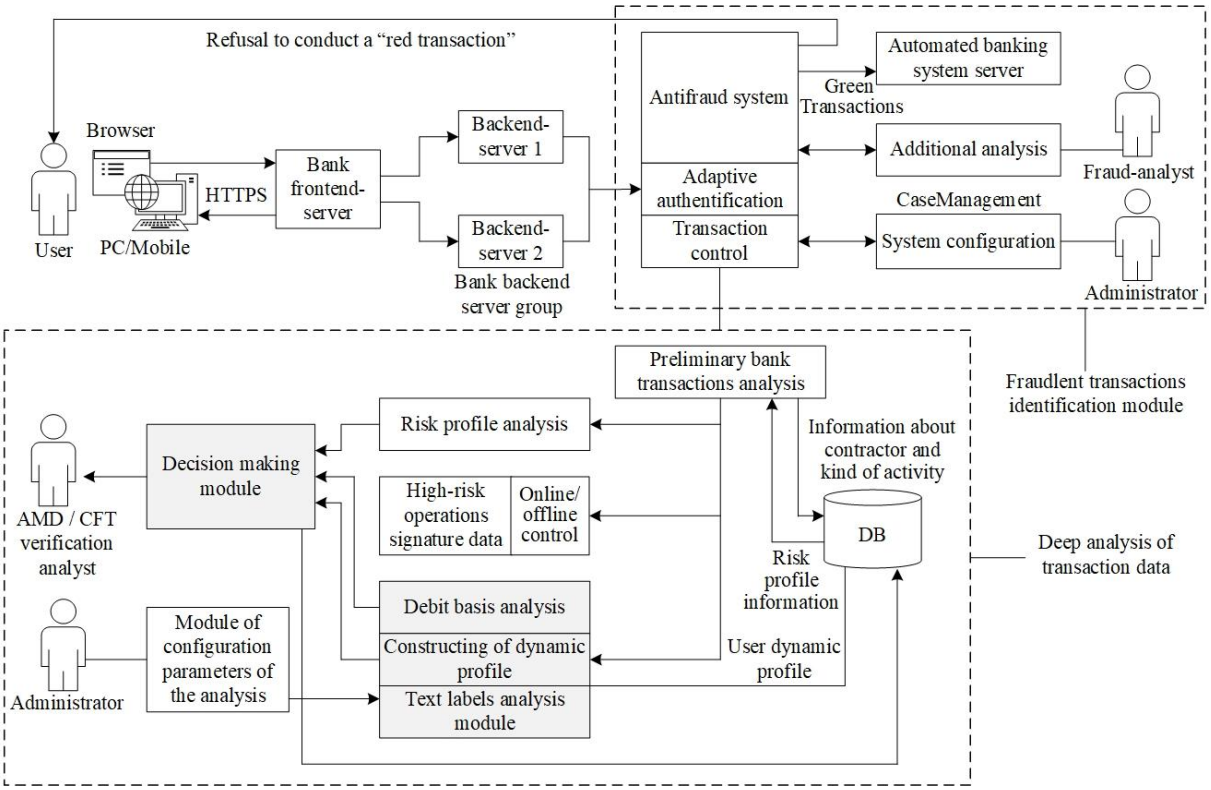

Figure 1. Proposed architecture of a banking transaction monitoring system with a text label mining module.

The standard scenario for working with the RBS system contains several stages [7]. Let us consider the additional processing algorithm in more detail. Transaction data enters the subsystem of deep analysis of transaction data. The first block of this subsystem is the preliminary analytics block of transaction data. The database sends to the preliminary data analytics block of the transaction data already available information about the user's dynamic risk profile. This block divides the data into those that are analyzed for the risk profile, those that will be analyzed by the signature for checking high-risk banking transactions, and those that are necessary to build a dynamic profile of the bank contractor, and the text label of the payment. At the same time data about counterparties and their types of economic activity are received in the database. Of all the blocks in which information came from the preliminary analytics block of transaction data, the processed information is sent to the decision making block. Here the risk profile is analyzed as well as compliance with the conditions for classifying the operation as highly risky, the transaction OKVED code, the dynamic contractor profile obtained after analysis, and makes a decision based on this data.

The decision block sends updated information about the user's dynamic profile to the database. Thus, the database always stores up-to-date information about the user's dynamic profile.

3.1. Development the structure of the text labels intelligent analysis subsystem

Analysis of the text label accompanying banking transactions allows a deeper analysis of customer activity. However, due to the complexity of such analysis, it is not possible to generate new rules for a system based on signature checks, keep these rules up-to-date and constantly add new rules [27].

A classifier is needed that can quickly analyze a text label and classify it in accordance with the OKVED code.

Further work with dictionaries involves the construction of a neural network classifier for complex marking of incoming data of text labels for payment purposes and a classifier based on data mining algorithms.

The developed structure of the subsystem for processing text labels for payment using a neural network analysis block is shown in Figure 2.

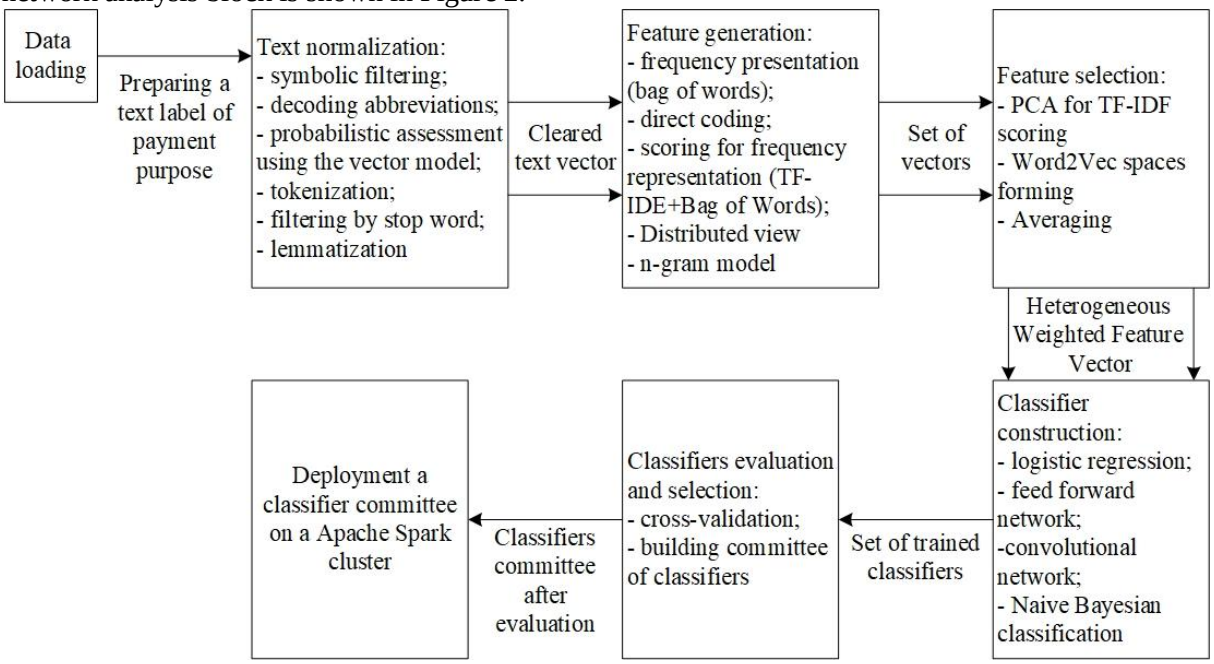

Figure 2. The structure of the subsystem for processing text labels for the purpose of payments using a neural network analysis unit (TF-IDF - a statistical measure to assess the importance of a word in the context of a document; Bag of Words (BOW) is a model of an unordered set of words that appear in text; Word2Vec - methods for constructing a compressed space of word vectors using a neural network; PCA – principal component analysis; N-gram-word/char – N-grams based on consecutive words/characters).

The data of bank transactions from the MySQL database are submitted to the input of the classifier, including the data of the payment purpose text label. The text label data is filtered out and transferred to the next block, then the classifier works only directly with the label.

In the text label text normalization block, irrelevant symbolic filtering, decryption of abbreviations, etc. occurs, that is, at the output of the normalization module, we have a cleaned text vector, which is then passed to the feature generation module.

In the module of feature generation, further processing of the cleaned vector occurs, and the set of vectors is obtained at the output of the module.

This set of vectors is fed to the input of the PCA feature selection module for TF-IDF scoring and space formation using the Word2Vec and FastText algorithms, then averaging is performed. The output of this module is a heterogeneous weighted feature vector.

Next, the heterogeneous weighted feature vector obtained at the last step is fed into the module for constructing various classifiers. The output of this module is a set of trained classifiers.

This set of trained classifiers is fed into the block of assessment and selection of classifiers. It cross-checks and builds a committee of classifiers.

The classifier committee resulting from the output of this module after evaluation is fed to the input of the classifier committee deployment module, where, in fact, the deployment to the Apache Spark cluster takes place.

The developed text data processing [27] pipeline using mining methods is shown in Figure 3.

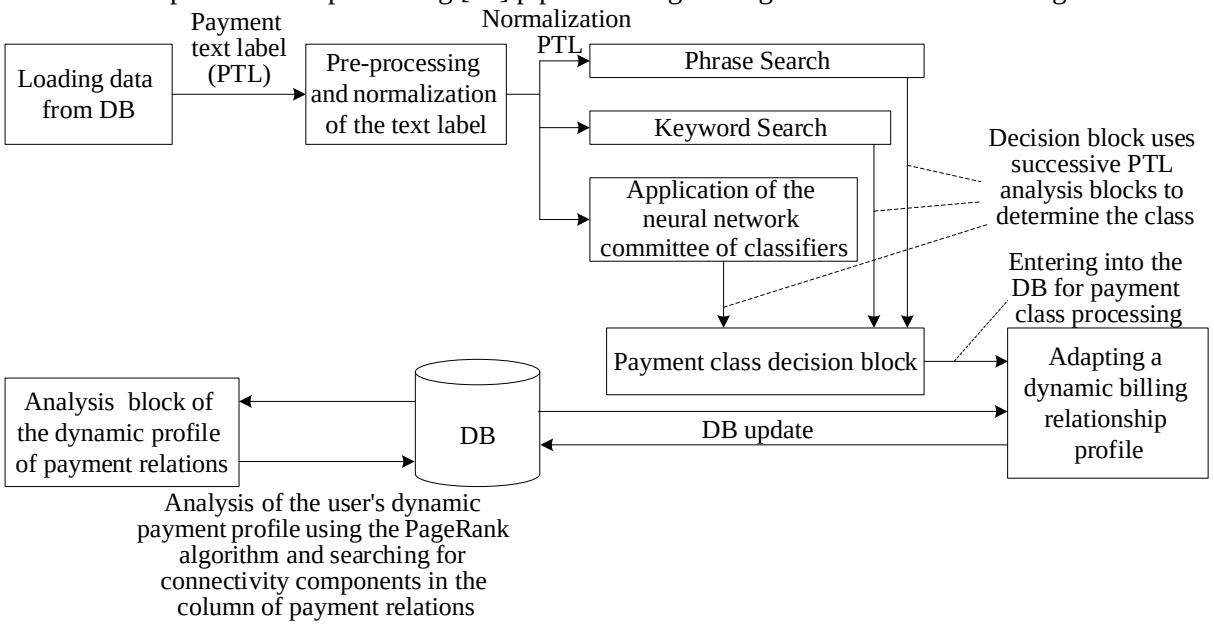

Figure 3. Text data processing pipeline using data mining methods as part of the TMS module.

Consider the operation algorithm of the text data processing pipeline using data mining methods.

From the database, the classifier receives data from banking transactions. Fields containing the text label of the payment purpose are transferred to the next module (pre-processing and normalization of the text label), where symbol filtering, lemmatization, expansion of abbreviations, etc. take place, so at the output of the module we get the normalized text label of the payment purpose.

The normalized text label alternately (until an answer with the required degree of confidence is received) enters the text label analysis blocks by searching for key phrases and keywords in it, and then to the block using the neural network committee of classifiers.

The decision block classifies the label based on the modules passed (or the module, if the required confidence is achieved when passing through the key word search block).

The classified text label is sent to the dynamic profile adaptation block of the payment relations in order to update the dynamic profile taking into account the payment, which is accompanied by the processed text label. The changed profile of the payment relations is sent to the database for updating

and further transferring to the analysis unit for the dynamic profile of the payment relations, where a deep analysis of the dynamic profile of the bank's counterparty takes place, taking into account the data on the type of economic activity according to the OKVED of the counterparty and the results of assigning a text label to each transaction made (or attempt to conduct a payment transaction) to any class by type of economic activity.

The analyzed profile arrives again in the database for storage and is stored in it on demand – the commission of new transit operations by the same counterparty.

4. Development an algorithm for the intelligent analysis of text labels in a banking transaction monitoring system

The algorithm for the intelligent analysis of the text label (figure 4) accompanying banking transactions, proposed in this work consist of many stages. Among them, 3 groups can be distinguished:

- work with dictionaries;
- text label preprocessing;
- text label classification.

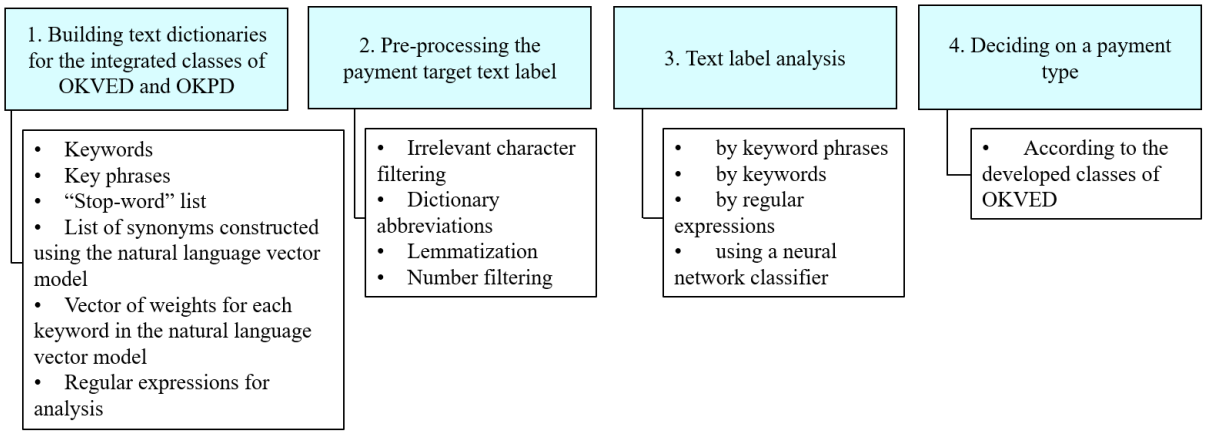

Figure 4. Payment label analysis algorithm.

The developed text label mining algorithm involves the following steps:

1) Formation of a list of words and phrases from the categories of OKVED defined as suspicious in [28] and the corresponding OKPD (All-Russian Classifier of Products by Activity). Rearrangement of sixteen available classes into eleven more related topics for the convenience and unambiguity of the text label classification. The resulting classes received their numbering (101, 102, etc.), which does not repeat the numbering of the original classes from the Central Bank document. At the same time, the original classes with their numbering are saved. Thus, a text label can be classified simultaneously according to the old and new class systems. The resulting classes, as well as the correspondence of the new division into classes to the original, are shown in Table 1.

2) A list of synonyms and hyperonyms was compiled and added to each word from the list of words of each of the new classes [29]. Such lists were generated using the pre-trained Word2Vec and FastText models. Also, phrases suitable for its meaning were added to each class. Figure 5 shows a graph of semantic proximity of keywords for different classes: 1 (food) and 8 (medicine). The figure shows that the keywords of the classes are semantically close to each other, while the classes themselves are far from each other.

3) Further, advanced dictionaries were processed to highlight keywords and phrases. Moreover, it is necessary to observe the rule that key phrases should not include keywords.

All phrases include lemmas of words.

Examples of keywords and key phrases for the resulting classes after processing are shown in table 1.

Text label processing involves the implementation of many steps. Let's consider them in detail.

1) Irrelevant symbolic filtering.

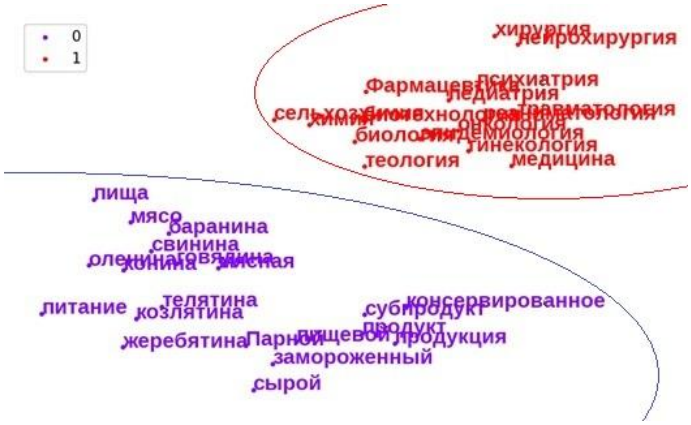

Figure 5. Graph of semantic affinity of keywords 1 and 8 classes.

Table 1. Examples of keywords and phrases from the resulting classes.

Class No.	Class name	Keyword examples	Key phrases examples
101	Food	pork, sorbet, sugar	mineral water, crude product, vegetable oil
102	Agriculture	butterfly, compound feed, fisheries	agriculture, annual culture, late variety
103	Transport and auto parts	car, steering wheel, overpass	passenger car, braking distance, spark plug
104	Appliances	fridge, iron, coffee machine	microwave, home cinema, music center
105	Business Support Services	legal, certification, accounting	financial service, foreign currency, assistance
106	Tobacco products	nicotine, tobacco, cigarette case	all words on this topic are the key
107	Property	property, hotel, rental	plot of land, capital investment, weaving land
108	Medicine	medicinal, orthopedic, medical	ambulance, emergency care
109	Fuel, metals, etc.	fuel oil, gasoline, metal	crude oil, lubricant
110	Clothing, shoes and accessories	bag, shoes, clothes	military uniform, uniform, headgear
111	Other trade	furniture, construction, dishes	sanitary equipment, software, kitchen

It implies the replacement of “extra” characters – punctuation marks, brackets, dates, numbers, numbers, etc. on a space in the text label.

2) Deployment of abbreviations and abbreviations found in text labels. Abbreviations are searched in the dictionary, the probabilities of collocations are evaluated based on the Word2Vec database and the correct decryption is assessed by the probability of using the expanded word in the general context of the sentence.

3) Lemmatization. By this term is meant the reduction of a word to a lemma – the initial form of a word. At the lemmatization step, the label of a part of speech and the label of a named entity are added to each word.

4) Filtering. It implies the removal of “stop-words” labels from the text, that is, words that do not exactly refer to any of the classes highlighted by the Central Bank of Russia.

5) Classification. At this stage, the processed text labels are searched using regular expressions for key phrases and keywords highlighted at the stage of creating dictionaries, and, accordingly, the class is assigned a text label.

5. Evaluation of the effectiveness of the proposed solutions on field data

For evaluating the effectiveness of the proposed algorithms for the operation of the mining module, the base of examples containing 1000 examples in 10 classes was used.

Experiment 1. Using the Bag of Words method for source data without using TF-IDF [30], a classifier based on logistic regression is constructed.

Experiment 2. Dictionary created and TF-IDF tags built. The size of the original feature vector: (10000, 165698), where 10000 is the number of examples of text labels, and 165698 is the length of the feature vector of one text label.

Then the PCA is applied to reduce the dimension of the feature space to (10000, 100) – now the length of the feature vector is 100.

The sample is divided into training and test in the ratio of 70% to 30%. Testing is conducted for a group of classifiers. Next, two classifiers are selected – SGD (Stochastic gradient descent) [31] and RF (Random Forest) [32], and for them the selection of hyperparameters is performed.

Figure 6 shows the ROC curves obtained during classification by the SGD classifier.

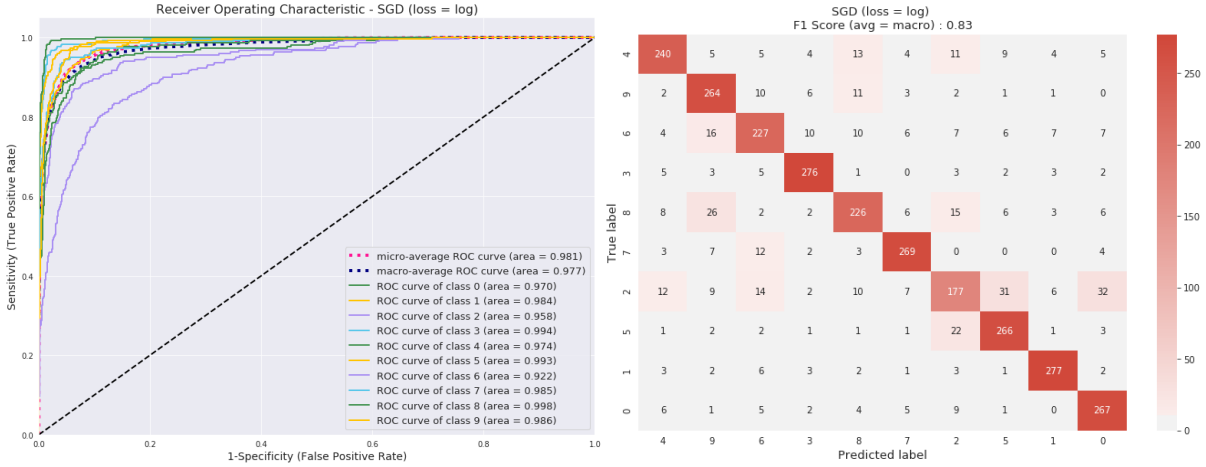

Figure 6. ROC – curves and inaccuracy matrix obtained during the work of the SGD classifier.

Similarly, the selection of hyperparameters for the RF (Random Forest) classifier is performed.

Figure 7 shows the ROC curves obtained during classification by the RF classifier.

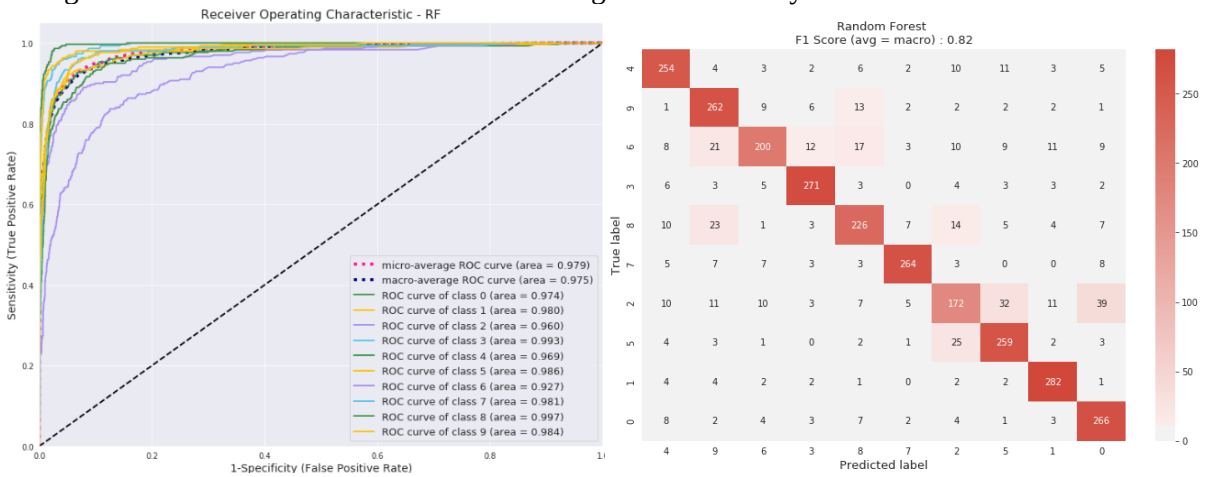

Figure 7. ROC – curves and inaccuracy matrix obtained during the work of the RF classifier.

Experiment 3. Next, apply the Word2Vec classifier. Classifier parameters are given in table 2.

When using the BOW + Xgboost Model [34], 20% was deferred to the validation sample. The model showed an accuracy of 0.9970 on the training set, and 0.8229 on the test set.

Model Word level TF-IDF + Xgboost Model showed the following results: on the training sample 0.9994, and on the test 0.8141.

Table 2. Word2Vec classifier options.

Vector size	Window	Corpus	Vocabulary size	Algorithm	Lemmatization
300	20	Russian National Corpus	189193	Gensim Continuous Bag-of-Words [33]	True

The N – gram Level TF-IDF model showed the following results: on the training set, the accuracy was 0.9428, on the test set – 0.7178. Table 3 shows the parameters when using the N – gram Level TF-IDF model.

Table 3. Parameters when using the N-gram Level TF-IDF model.

Parameter	Value
ngram_range	2,3
max_features	5000

Experiment 4. When using a convolutional neural network.
Summary parameters of classifiers are given in table 4.

Table 4. Results obtained using various classifiers.

Method for creating a feature space	Classifier	The value of the metrics in the test sample			
		Accuracy	Precision	Recall	F_1 measure
Bag of Word (BOW)	Logistic regression	0.878	0.878	0.878	0.877
BOW + XGboost	XGBoost	–	0.82	0.82	0.82
N-gramm+TF-IDF	XGBoost	–	0.72	0.72	0.72
TF-IDF+PCA	k-nearest neighbors algorithm	0.80	0.80	0.80	0.80
	Stochastic Gradient Descent	0.80	0.80	0.80	0.80
	Random Forest	0.77	0.77	0.77	0.77
	AdaBoost	0.71	0.70	0.71	0.70
	Gaussian Naive Bayes	0.70	0.71	0.70	0.70
	Decsision Tree	0.69	0.69	0.69	0.69
Word2Vec	Logistic regression	0.768	0.769	0.768	0.768
	Convolutional neural network	0.84	–	–	0.89

6. Conclusion

A structural diagram of a banking transaction monitoring system was developed, which includes a module for the intelligent analysis of text labels for payment purpose. Integration of this module allows to draw conclusions about whether the transit operation belongs to any of the classes of the proposed classification, which allows us to build a dynamic profile of the bank's contractor.

An algorithm is proposed for the phased analysis of the text label of the payment purpose, including the stages of preprocessing, filtering, normalizing and constructing a classifier based on a set of regular expressions and intelligent analysis technologies. The algorithm difference is in the use of adaptive category dictionaries and the multi-pass application of heterogeneous classifiers, which allows to increase the validity of the decision on whether the transaction belongs to one of the selected classes.

The evaluation results showed that using the classifiers composition, the classification accuracy of 81% was achieved (on the test set during cross-validation), while the basic version of the algorithm (using regular expressions) made it possible to achieve classification with an accuracy of about 60%.

However, the main problem in constructing a system for automatic classification of payment purposes is the weak requirements for formalization of a text label by the bank and the length of the

text label itself. It is necessary to improve the semantic spaces of each of the categories indicated in regulatory documents in order to improve the performance indicators of classifiers.

7. Acknowledgments
The reported study was funded by RFBR, project number 20-08-00668.

8. References

[1] Van Vlasselaer, V. APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions / V. Van Vlasselaer, C. Bravo, O. Caelen., T. Eliassi-Rad, L. Akoglu, M. Snoeck, B. Baesens // Decision Support Systems. – 2015. – Vol. 75. – P. 38-48.
[2] Trelewicz, J.Q. Big data and big money: The role of data in the financial sector // IT Professional. – 2017. – Vol. 19(3). – P. 8-10.
[3] Overview of antifraud systems: how to detect fraud and save a budget [Electronic resource]. – Access mode: https://blog.admobispy.com/poleznoe/obzor-antifrod-sistem-kak-vyiyavit-moshennichestvo-i-sohranit-byudzhet (26.12.2019).
[4] Khanuja, H.K. Forensic analysis for monitoring database transactions / H.K. Khanuja, D.S. Adane, // International Symposium on Security in Computing and Communication – Berlin, Heidelberg: Springer, 2014. – P. 201-210.
[5] Sapozhnikova, M.U. Anti-fraud system on the basis of data mining technologies / M.U. Sapozhnikova, A.V. Nikonov, A.M. Vulfin, M.M. Gayanova, K.V. Mironov, D.V. Kurennov // IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). – IEEE, 2017. – P. 243-248.
[6] Sapozhnikova, M.U. Distributed infrastructure for big data processing in the transaction monitoring systems / M.U. Sapozhnikova, M.M. Gayanova, A.M. Vulfin, A.V. Nikonov, A.V. Chuykov // CEUR Workshop Proceedings. – 2018. – Vol. 2212. – P. 228-235.
[7] Sapozhnikova, M.U. Intrusion detection system based on data mining technics for industrial networks / M.U. Sapozhnikova, A.V. Nikonov, A.M. Vulfin // International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), 2018. – P. 1-5.
[8] Mironov, K.V. Data mining technologies in the problem of designing the bank transaction monitoring system / K.V. Mironov, M.U. Sapozhnikova, M.M. Gayanova, A.M. Vulfin, A.V. Nikonov // Computer Science and Information Technologies (CSIT), 2017. – P. 45-55.
[9] Lutskovich, A.I. Antifraud for banks is becoming increasingly relevant. – 2013. – Vol. 213(11).
[10] Federal Law of August 07, 2001 No. 115-FZ “On Counteracting the Legalization (Laundering) of Criminally Obtained Incomes and the Financing of Terrorism”.
[11] Bank of Russia Regulation dated March 02, 2012 No. 375-P “On requirements for the rules of internal control of a credit institution in order to counter the legalization (laundering) of proceeds from crime and the financing of terrorism”.
[12] Rosinformonitoring Public Report “National assessment of the risks of legalization (laundering) of criminal proceeds 2017-2018”.
[13] Rosinformonitoring Public Report “National Terrorism Financing Risk Assessment 2017-2018”.
[14] Federal Law of June 27, 2018 No. 167-FZ “On Amending Certain Legislative Acts of the Russian Federation with Respect to Countering Money Theft”.
[15] Bank of Russia Regulation dated June 9, 2012 No. 382-P “On requirements for ensuring the protection of information when making money transfers and on the procedure for the Bank of Russia to monitor compliance with the requirements for ensuring information protection when making money transfers”.
[16] Federal Law of June 27, 2011 No. 161-FZ “On the National Payment System”.
[17] Universal antifraud – solution for the ecommerce industry [Electronic resource]. – Access mode: http://payler.com/assets/files/antifroud.pdf (26.12.2019).
[18] SmartVista Solutions [Electronic resource]. – Access mode: https://www.bpcbt.com/ru/smartvista–solutions/ (26.12.2019).

[19] Meet this fraudwall [Electronic resource]. – Access mode: http://www.fraudwall.ru/ (26.12.2019).
[20] FRAUD – Analysis [Electronic resource]. – Access mode: http://www.bssys.com/ solutions/financial–institutions/fraud/ (26.12.2019).
[21] Kaspersky Fraud prevention [Electronic resource]. – Access mode: https://kfp.kaspersky.com /ru/ (26.12.2019) – (in Russian).
[22] Signs of the transfer of funds without the consent of the client (approved by the order of the Bank of Russia dated September 27, 2018 No. OD-2525).
[23] Features of the organization of compliance control in Russian banks [Electronic resource]. – Access mode: https://bankir.ru/publikacii/20090814/osobennosti-organizacii-komplaens-kontrolya-v-rossiiskih-bankah-2303295/ (26.12.2019).
[24] Abbad, M. The Development of E-Banking in Developing Countries in the Middle East / M. Abbad, J.M. Abed, M. Abbad // Journal Financial Account Managemaent. – 2012. – Vol. 2. – P. 107-123.
[25] Jarrett, J.E. Internet Banking Development // J. Entrep. Organ. Manag. – 2016. – Vol. 5. – P. 2-5.
[26] Ogwueleka, F.N. Neural network and classification approach in identifying customer behavior in the banking sector: A case study of an international bank / F.N. Ogwueleka, S. Misra, R. Colomo-Palacios, L. Fernandez // Human factors and ergonomics in manufacturing & service industries. – 2015. – Vol. 25(1). – P. 28-42.
[27] Applied text analysis with Python. Machine learning and building natural language processing applicationsтер / B. Bengfort, R. Bilbro // Izd. "Piter" – 2019. – Vol. 2. – P. 104-106.
[28] Bank of Russia Methodological Recommendations No. 18–MP dated July 21, 2017 "On Approaches to the Management of Credit Institutions by the Risk of Legalization (Laundering) of Criminally Obtained Incomes and the Financing of Terrorism".
[29] Nayak, N. In learning hyperonyms over word embeddings // arXiv, 2015.
[30] Ramos, J. Using TF-IDF to determine word relevance in document queries // Proceedings of the first instructional conference on machine learning. – 2003. – Vol. 242. – P. 133-142.
[31] Bordes, A. SGD-QN: Careful quasi-Newton stochastic gradient descent / A. Bordes, L. Bottou, P. Gallinari // Journal of Machine Learning Research. – 2009. – Vol. 10. – P. 1737-1754.
[32] Liaw, A. Classification and regression by random Forest / A. Liaw, M. Wiener // R news. – 2002. – Vol. 2(3). – P. 18-22.
[33] Ma, L. Using Word2Vec to process big text data / L. Ma, Y. Zhang // IEEE International Conference on Big Data (Big Data), 2015. – P. 2895-2897.
[34] Chen, T. Xgboost: A scalable tree boosting system / T. Chen, C. Guestrin // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. – ACM, 2016. – P. 785-794.