DOI: 10.37791/2687-0649-2021-16-5-66-81

# Система классификации документов с маркшейдерскими данными

***В. В. Еремеев***[1]***, М. С. Цыганова***[1*]***, А. Г. Ивашко***[1]
[1] *Тюменский государственный университет, Тюмень, Россия*
** m.s.cyganova@utmn.ru*

**Аннотация.** Все предприятия, осуществляющие геологоразведочные работы на территории РФ, сталкиваются с необходимостью формирования задач для маркшейдерской службы и контроля выполнения поставленных задач. Это отражается в процессах документооборота предприятий. В данной связи существует проблема организации эффективной обработки документов в системах электронного документооборота – своевременного выявления документов, содержащих маркшейдерские данные. В статье представлено возможное решение указанной проблемы – автоматизированная система классификации документов в СЭД в виде рекомендательной надстройки над системой 1С:Документооборот. В рамках создания системы классификации был разработан и реализован сценарий предварительной обработки первичных текстов документов, включающий очистку, лемматизацию и удаление стоп-слов, а также подготовку входных признаков для классификатора. Исследована применимость различных алгоритмов машинного обучения к решению рассматриваемой задачи классификации, определены значения гиперпараметров, обеспечивающие наибольшее значение метрики ROC AUC. Выполнена оценка качества всех полученных моделей с использованием метрик Precision, Recall и F-меры, исследована устойчивость качества классификации к изменению входных данных. Выявленная проблема нестабильности результатов классификации решалась путем построения модели машинного обучения в виде ансамбля классификаторов. Обученная модель (ансамбль классификаторов) тестировалась на наборе реальных документов ООО «Газпром недра»; качество классификации на тестовой выборке по метрике ROC AUC составило 0,91. Кроме собственно модуля классификации разработанная система включает базу данных хранения результатов обучения, библиотеку функций для организации работы с базой данных, а также API-интерфейсы, позволяющие обрабатывать запросы на классификацию, приходящие из внешних систем. В API-интерфейсах, в частности, реализованы возможности загрузки сохраненных обученных моделей, валидации данных, приходящих из внешних систем, предварительной обработки входных текстовых документов, обучения новых моделей и оценки их качества, сохранение как обученных моделей, так и результатов их тестирования. Реализована возможность дообучения сохраненных моделей на новых данных.

**Ключевые слова:** система классификации, классификация документов, маркшейдерские данные, предварительная обработка текста, машинное обучение, ансамбль классификаторов

**Для цитирования:** *Еремеев В. В., Цыганова М. С., Ивашко А. Г.* Система классификации документов с маркшейдерскими данными // Прикладная информатика. 2021. Т. 16. № 5. С. 66–81. DOI: 10.37791/2687-0649-2021-16-5-66-81

# Classification system for documents with mine surveying data

*V. Eremeev*[1], *M. Tsyganova*[1*], *A. Ivashko*[1]
[1] *University of Tyumen, Tyumen, Russia*
** m.s.cyganova@utmn.ru*

**Abstract.** All enterprises engaged in exploration activities on the territory of the Russian Federation, are facing the need to formulate tasks for the mine surveyor service and control their execution. It affects enterprise's workflow process. Due to it, a problem of organization of efficient document processing in electronic document management systems (timely identification of documents containing mine surveying data) takes place. The article presents possible solution of this problem – automated document classification system into EDMS in the form of optional add-on for 1C:Document Management. Within the classification system creation a preprocessing script for primary document texts, including cleaning, lemmatization, stop words removing, as well as preparation of input features for the classifier were developed and implemented. Applicability of different machine learning algorithms to solution of considering classification problem was studied, the values of hyperparameters providing the highest value of the ROC AUC metric were determined. The quality of all obtained models was assessed using metrics Precision, Recall and F-measures, the stability of the classification quality to changes in the input data was investigated. The identified problem of instability of classification results was solved by building and implementing a machine learning model in the form of ensemble of classifiers. Classification model (an ensemble of clusters) was tested on the set of real documents of Gazprom nedra Ltd; classiffication quality on the test sample by ROC AUC metric was 0,91. Except the classification module itself, developed system contains the storage database for learning outcomes, function library for organization of work with the database and API interfaces allowing to process classification requests, coming from external systems. These API interfaces, in particular, implement the ability to load saved trained models, validate data coming from external systems, preprocess input text documents, train new models and assess their quality, save both trained models and the results of their testing. Also the possibility of the additional training of the models on a new data was realized.

**Keywords:** classification system, document classification, mine surveying data, text preprocessing, machine learning, ensemble of classifiers

**For citation:** Eremeev V., Tsyganova M., Ivashko A. Classification system for documents with surveying data. *Prikladnaya informatika*=Journal of Applied Informatics, 2021, vol.16, no.5, pp.66-81 (in Russian). DOI: 10.37791/2687-0649-2021-16-5-66-81

## Введение

В соответствии с требованиями статьи 24 Закона «О недрах» [1] выполнение комплекса маркшейдерских работ является обязательной частью геологоразведочных и горных работ на территории РФ. Маркшейдерское сопровождение необходимо для обеспечения безопасности работ, связанных с пользованием недрами, а также для организации множества подготовительных работ (подъезда техники, создания разведочной площадки, поиска источников и путей подвода воды). Поэтому соответствующие службы всех предприятий, осуществляющих геологоразведочные работы на территории РФ, должны формировать задачи для маркшейдерской службы, а также контролировать выполнение поставленных задач.