УДК 550.812.1
DOI: 10.31660/0445-0108-2023-2-35-54

Цифровой керн: нейросетевое распознавание текстовой геолого-геофизической информации

Ю. Е. Катанов*, А. И. Аристов, А. К. Ягафаров, О. Д. Новрузов

Тюменский индустриальный университет, Тюмень, Россия
**katanov-juri@rambler.ru*

Аннотация. Представлен алгоритм аналого-цифрового преобразования первичной геолого-геофизической информации (на примере идентификации литотипов горных пород на базе текстового описания физического керна).

В рамках работы реализовано комплексирование трех видов научных исследований — поисковое, междисциплинарное и прикладное при формировании исходной базы качественных данных.

Описаны распространенные алгоритмы для классификации текстовой информации и механизм предобработки исходных данных с использованием токенизации.

Концепция распознавания текстовых образов реализована с привлечением методов искусственного интеллекта.

Для создания нейросетевой модели распознавания текстовой геолого-геофизической информации использован язык программирования Python в сочетании с технологиями сверточных нейросетей для классификации текста (TextCNN), сетей двунаправленной длительной-кратковременной памяти (BiLSTM) и сетей представлений двунаправленного кодера (BERT).

Стек данных технологий и языка программирования Python, после разработки и апробации базового варианта нейросетевой модели распознавания качественной информации, обеспечили приемлемый уровень работы алгоритма цифровой трансформации текстовых данных.

Наилучший результат (текущая версия нейросетевой модели 1.0; более 3 000 примеров для обучения и тестирования) достигнут при использовании алгоритма распознавания текстовых данных на базе BERT с точностью на валидационном сете (Validation Accuracy) ~0.830173 (25 эпоха), с потерями на валидационном сете (Validation Loss) ~0.244719, с потерями во время обучения (Training Loss) ~0.000984 и вероятностью распознавания исследуемых литотипов горных пород более 95 %.

Определены механизмы модификации кода для дальнейшего улучшения точности текстового прогноза на базе созданной нейросети.

Ключевые слова: распознавание символов, кластеризация текста, контекстная информация, интерпретация, токенизация, нейросеть, выборка

Для цитирования: Катанов, Ю. Е. Цифровой керн: нейросетевое распознавание текстовой геолого-геофизической информации / Ю. Е. Катанов, А. И. Аристов, А. К. Ягафаров, О. Д. Новрузов. – DOI 10.31660/0445-0108-2023-3-35-54 // Известия высших учебных заведений. Нефть и газ. – 2023. – № 3. – С. 35–54.

Digital core: neural network recognition of textual geological and geophysical information

Yuri E. Katanov*, Artyom I. Aristov, Alik K. Yagafarov, Orchan D. Novruzov

Industrial University of Tyumen, Tyumen, Russia
**katanov-juri@rambler.ru*

Abstract. The algorithm of analog-to-digital conversion of primary geological and geophysical information (on the example of identification of rock lithotypes based on the text description of the physical core) is presented.

As part of the work, a combination of three types of scientific research - prospecting, interdisciplinary and applied, in the formation of the initial base of qualitative data is implemented.

Common algorithms for textual information classification and mechanism of initial data preprocessing using tokenization are described.

The concept of text pattern recognition is implemented using artificial intelligence methods.

For creation of the neural network model of textual geological and geophysical information recognition the Python programming language is used in combination with the convolutional neural network technologies for text classification (TextCNN), bi-directional long-short-term memory networks (BiLSTM) and bi-directional coder representation networks (BERT).

The stack of these technologies and the Python programming language, after developing and testing the basic version of the neural network model of qualitative information recognition, provided an acceptable level of performance of the algorithm of digital transformation of text data.

The best result (the current version of neural network model is 1.0; more than 3000 examples for training and testing) was achieved when using the algorithm of text data recognition based on BERT with an accuracy on the validation network (Validation Accuracy) ~0.830173 (25^{th} epoch), with Validation Loss ~0.244719, with Training Loss ~0.000984 and probability of recognition of the studied rock lithotypes more than 95 %.

The mechanisms of code modification for further improvement of textual prediction accuracy based on the created neural network were determined.

Keywords: character recognition, text clustering, content information, interpretation, tokenization, neural network, sampling

For citation: Katanov, Yu. E., Aristov, A. I., Yagafarov, A. K., & Novruzov, O. D. (2023). Digital core: neural network recognition of textual geological and geophysical information. Oil and Gas Studies, (3), pp. 35-54. (In Russian). DOI: 10.31660/0445-0108-2023-3-35-54

Введение

Первичная геолого-геофизическая информация, полученная в процессе испытания скважин, может быть представлена различным образом: в буквенном виде, в символьном и графическом описании, в виде звуковой и видеоинформации.

Поскольку информатизация нефтегазовых месторождений является первоочередным фактором, то возникает необходимость корректной обработки получаемой разнородной информации для формирования комплексных отчетов.

Для существенного увеличения производительности информационно-аналитических работ необходимо привлечение определенных цифровых технологий и методов искусственного интеллекта (ИИ). Это дает возможности повышения информации, как на уровне ретроспективного анализа, так и при минерагенической/прогнозной интерпретации первичных геолого-геофизических данных.

При сопоставлении данных геофизических исследований скважин (ГИС) и данных опробования зачастую наблюдается явная расходимость, даже на уровне лабораторных испытаний. Это возникает при увеличении

объема выборок по петрофизическим характеристикам кернового материала, что в очередной раз подчеркивает нелинейность распределения геологических «помех» и соответствующих им физико-химических неоднородностей.

Лабораторные исследования кернового материала не ограничиваются только лишь классическими методами статистики, геохимии, геофизики и пр. Также они могут включать средства анализа цифровой информации — изображения керна и его текстовое описание; привязка методов ГИС к глубинам пласта и т. п. с привлечением алгоритмов глубокого и машинного обучения.

Стоит отметить, что нельзя избавиться от геологической неоднородности горных пород, как и от неопределенности их структурно-вещественных изменений по глубине/по простиранию пласта — это факт. Но можно разработать серию таких эволюционно-генетических (цифровых) алгоритмов, которые позволят минимизировать случайные высокоэнтропийные изменения петрофизических и литофациальных характеристик горных пород, описанных на качественном и количественном уровнях [1].

Серия цифровых алгоритмов позволит системно идентифицировать и спрогнозировать следующие закономерности: кросс-корреляции петрографических данных при создании литолого-седиментационной модели; качественное описание текстуры матрицы горных пород; предварительную качественную оценку открытой/закрытой пористости, проницаемости пластового флюида (|| и $\perp$ напластованию) [2].

Актуальность исследования состоит в исключении рисков неопределенности при трактовке результатов интерпретации первичной геолого-геофизической информации, что немаловажно в управлении геолого-технологическим проектом на всех стадиях, при должном обеспечении гарантий информационной безопасности.

Автоматизирование процесса качественной интерпретации разнородных данных позволяет не только сократить время исследований, но и снизить диссипативную характеристику петрофизических особенностей горных пород при соответствующей идентификации их керновых материалов.

Объект и методы исследования

Изучение кернового материала — это дорогой и долгий процесс, требующий большого количества лабораторных экспериментов.

Существенный недостаток лабораторного исследования физического керна состоит в том, что изучение этих образцов, как правило, небезопасно для них самих — они теряют текущие физико-механические свойства, и в дальнейшем не будет возможности воспроизведения их истинных петрофизических, литофациальных и фильтрационных особенностей.

Поэтому формирование цифровой базы разнородных керновых данных, а также создание эволюционно-генетических алгоритмов их обработки и интерпретации является важной основой для многократного проведения экспериментов в виртуальном пространстве [3].

С целью уменьшения влияния человеческого фактора при обработке керновых данных и принятии оптимального решения в условиях информационной неопределенности используются методы глубокого и машинного обучения — технологии компьютерного зрения, сверточные нейросети и т. п.

Согласно исследованию Deloitte, цифровизация керновых данных посредством привлечения искусственного интеллекта осуществляется не только в количественном и графическом описаниях, но и в текстовом виде для повышения производительности и технологической трансформации ИИ [4–6].

При создании модуля нейросетевого распознавания первичной текстовой геолого-геофизической информации была выбрана специализированная среда (IDE) PyCharm, тесно интегрированная с Python, Web и Data Science [7–9].

PyCharm обеспечивает интеллектуальное завершение и инспекцию кода; выделение ошибок в реальном режиме времени и их быстрое исправление; автоматический рефакторинг кода и различные возможности навигации; интеграцию с IPython Notebook; поддержку NumPy, matplotlib и других научных пакетов [10–12].

PyCharm доступен в трех редакциях:

- Community (бесплатная): для интеллектуальной разработки на Python, включая помощь в работе с кодом, рефакторинг, визуальную отладку и интеграцию контроля версий;
- Professional (платная): для профессиональной разработки на Python, Web и Data Science, включая помощь в работе с кодом, рефакторинг, визуальную отладку, интеграцию контроля версий, удаленные конфигурации; развертывание и поддержку популярных веб-фреймворков, таких как Django и Flask; поддержку баз данных и научных инструментов (включая поддержку Jupyter notebook); инструменты для работы с большими данными;
- Edu (бесплатная): для изучения языков программирования и связанных с ними технологий в сочетании с интегрированными образовательными инструментами.

В качестве фреймворка искусственного интеллекта (ИИ) выбрана технология Transformers, представляющая современное машинное и глубокое обучение для PyTorch, TensorFlow и JAX, предоставляющая API легкую загрузку и возможность обучения современных, предварительно обученных моделей. Использование предварительно обученных моделей позволяет снизить затраты на вычисления и сэкономить время на обучение модели «с нуля» [13].

Предобученные модели могут использоваться в различных модальностях: текстовые данные более чем на 100 языках; классификация изображений, обнаружение графических объектов и их сегментация; распознавание речи и классификация аудиоинформации; комплексирование текстовой, символьной и графической информации.

Общая библиотека поддерживает бесшовную интеграцию между тремя наиболее популярными библиотеками глубокого обучения: PyTorch, TensorFlow и JAX.

Каждая архитектура Transformers определена в отдельном модуле Python, поэтому любой ее вариант можно легко настроить для исследований и экспериментов.

В качестве метода проектирования модуля распознавания первичной текстовой геолого-геофизической информации использовался подход нисходящего проектирования.

Процедура кодирования не может быть начата до тех пор, пока не будет достигнут минимально необходимый уровень детализации хотя бы части вычислительной системы в базовом проекте [14].

О методическом подходе исследования текстовой геолого-геофизической информации

Для нейросетевого распознавания первичной геолого-геофизической информации был выбран язык программирования Python по нескольким причинам:

- Python зарекомендовал себя как один из самых эффективных языков программирования для получения решений на базе технологий AI (Artificial Intelligence) и ML (Machine Learning);
- лучшая экосистема библиотек (группы модулей с заранее написанным набором кода) для совместной работы с другими приложениями при сосредоточении на продвижении функциональности разрабатываемого вычислительного модуля;
- возможность сочетания различных стилей программирования благодаря гибкой платформе;
- оптимальный вариант визуализации результатов моделирования (например, библиотека matplotlib);
- платформенная независимость Python на Unix, Linux, macOS, Windows и других операционных системах;
- быстрая разработка системы скриптов и меньшая процедура кодирования при создании прототипов ИИ;
- приемлемая скорость исполнения работы ИИ и машинного обучения.

Результатом использования языка программирования Python было создание единой платформы исследования следующих нейросетевых подходов: сверточная нейронная сеть для классификации текста (TextCNN); сеть двунаправленной длительной-кратковременной памяти (BiLSTM); сеть представлений двунаправленного кодера (BERT).

Идея использования CNN для классификации текста была впервые представлена в работе Юн Ким «Конволюционные нейронные сети для классификации предложений» (Convolutional Neural Networks for Sentence

Classification). Центральная концепция этой идеи заключается в том, чтобы рассматривать документы как изображения.

Изображения также имеют матрицу, отдельные элементы которой являются значениями пикселей. Но вместо пикселей изображения входными данными для задачи являются предложения или документы, представленные в матричном виде с учетом свертки. Каждая строка матрицы соответствует однословному вектору.

TextCNN хорошо работает для классификации текстов, потому что она учитывает слова, находящиеся на близком расстоянии друг к другу. Например, данная нейросеть может «видеть» слово «физико-химический» вместе (слитно). Однако она все еще не может проследить такие сочетания слов во всем контексте, представленном в конкретной текстовой последовательности. Эта нейросеть не будет изучать последовательную структуру данных, где каждое слово зависит от предыдущего слова или слова в предыдущем предложении.

Для решения этой особенности обрабатываемой информации можно использовать рекуррентные нейронные сети (РНС, RNN), которые запоминают предыдущую информацию, используя скрытые состояния, и связывают ее с текущей задачей.

Сети с долговременной памятью (LSTM) — это подкласс RNN, специализирующийся на запоминании информации на длительные периоды времени.

Двунаправленная LSTM сохраняет контекстную информацию в обоих направлениях, что весьма полезно в задачах классификации текстов. Однако она не подходит для задач прогнозирования временных рядов, поскольку в этом случае нет возможности «заглянуть в будущее».

Для простого объяснения двунаправленной РНС можно представить ячейку РНС как «черный ящик», принимающий на вход скрытое состояние (Hidden State) и вектор слов (Word Vector) и выдающий на выходе вектор (Output) и следующее скрытое состояние (Next Hidden State). Этот «черный ящик» имеет определенные весовые функции, которые необходимо настроить с помощью обратного распространения информационных потерь. Кроме того, одна и та же ячейка применяется ко всем словам, так что веса будут распределены между словами во всем предложении. Это явление называется процедурой распределения весов.

Сеть представлений двунаправленного кодера BERT (Bidirectional Encoder Representations from Transformers) улучшает стандартные трансформеры, устраняя ограничение однонаправленности с помощью цели предварительного обучения модели языка с некоторой маской (MLM).

Маскированная языковая модель случайным образом скрывает некоторые лексемы из входных данных, и задача состоит в том, чтобы предсказать исходный словарный идентификатор замаскированного слова, основываясь только на содержащем его контексте (рис. 1).

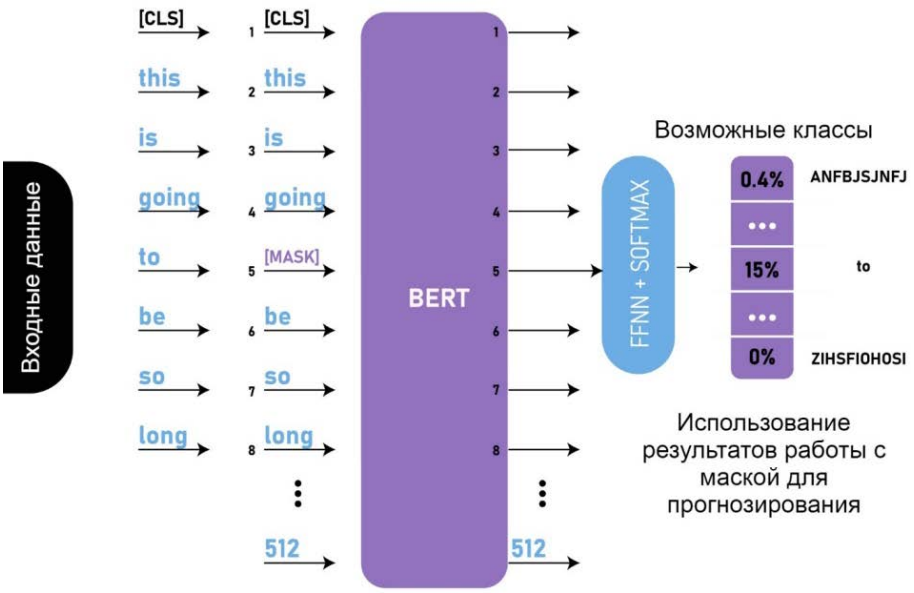


Рис. 1. ***Пример маскированной языковой модели***

В отличие от предварительного обучения лево-правой языковой модели, MLM позволяет объединить левый и правый контекст, что дает возможность предварительного обучения глубокого двунаправленного трансформатора.

Полученные выходные векторы данных пропускаются через ряд плотных слоев и, наконец, слой softmax для построения классификатора текста.

В дополнение к маскированной языковой модели BERT использует задачу предварительного предсказания следующего предложения при совместном обучении представлений различных текстовых пар [15].

BERT достигает двунаправленного обучения посредством использования двух методов: MLM (Masked Language Modelling) и NSP (Next Sentence Prediction).

В заданной текстовой последовательности случайным образом маскируется некоторый процент слов, заменяя их маркером [MASK]. Для данной научной работы было замаскировано 15 % входных слов.

Нейросеть обучается предсказывать эти замаскированные слова, используя контекст оставшихся слов.

Последовательность создания нейросетей была выполнена в соответствии с проработанными действиями [16–24].

1. Импортированы библиотеки Pandas, NumPy, Torch, Tqdm, Transformers.

2. Из библиотек импортировать зависимости BertTokenizer и BertForSequenceClassification.

BertTokenizer — токенизатор, который базируется на алгоритме WordPiece (токенизация на основе субслов). WordPiece используется в таких языковых моделях, как BERT.

BertForSequenceClassification — трансформатор модели BERT для задач классификации/регрессии (наличие линейного слоя поверх объединенного вывода).

Когда происходит работа с текстом, выполняется ряд шагов предварительной обработки, чтобы преобразовать текст в числа. Эти шаги имеют решающее значение в любом процессе разработки модели или даже при анализе текстов. В этом многоэтапном процессе предварительной обработки, одним из важных пунктов является токенизация, которая опять же может быть разных типов.

Токенизация — это процесс получения необработанных текстов и разделения их на токены, которые представляют собой числовые данные для представления слов.

Токенизация на основе субслов — это решение между токенизацией на основе слов и токенизацией на основе символов. Основная идея заключается в решении проблем, с которыми сталкиваются токенизация на основе слов (очень большой объем словаря, большое количество лексем OOV и различное значение очень похожих слов) и токенизация на основе символов (очень длинные последовательности и менее значимые отдельные лексемы).

Алгоритм токенизации на основе подслов разбивает редкие слова на более мелкие значимые подслова. Например, слово «пласт» не разбивается, а слово «пласты» разбивается на «пласт» и «ы». Это помогает модели узнать, что слово «пласты» образовано с помощью слова «пласт» с немного разным значением, но с одним и тем же корневым словом.

3. Импортировать входные данные для обучения и тестирования нейросетевой модели через библиотеку Pandas в формате CSV. Разделителем входных данных является знак «точка с запятой».

4. Извлечение уникальных кластеров из всех меток данных — каждому описанию горной породы присвоена метка «кластер».

5. Обучение и тестирование модели первичной классификации геолого-геофизической информации будут производиться с использованием вычислительных ядер NVIDIA CUDA в графическом процессоре.

6. Создание словаря кластеров в виде связки «идентификатор + кластер».

7. Установление индивидуальной метки кластеров для каждого описания на базе словаря меток.

8. Импортирование из библиотеки sklearn средства разделения входных данных на обучающую и тестовую выборки — train_test_split.

Чтобы разделение выборок было сбалансированным, была произведена стратификация выборок по кластерам.

9. Кодирование входных данных токенизатором BERT, преобразование их в соответствующее числовое представление.

Модель BERT ожидает на вход последовательность лексем (слов).

В каждой последовательности лексем есть две специальные лексемы, которые BERT ожидает получить на вход:

- [CLS]: это первый токен каждой последовательности, который обозначает классификационный токен;
- [SEP]: это отдельный маркер, который позволяет BERT узнать, какой из системных маркеров принадлежит к определенной последовательности. Этот специальный токен в основном важен для задачи предсказания следующего предложения или задачи ответа на вопрос.

Если в наборе данных только одна последовательность, то этот маркер будет добавлен в ее конец.

Допустим, что имеется некоторое предложение.

В качестве первого шага его необходимо преобразовать в последовательность лексем (слов). Несмотря на то, что была проведена токенизация этого входного предложения, необходимо сделать еще один шаг — нужно переформатировать эту последовательность лексем, добавив лексемы [CLS] и [SEP], прежде чем использовать ее в качестве входных данных для модели BERT.

Важно также отметить, что максимальный размер токенов, которые могут быть поданы в модель BERT, составляет 512.

Если токенов в последовательности меньше 512, то можно использовать «прокладку», чтобы заполнить неиспользуемые слоты токенов отдельным токеном [PAD].

Если токены в последовательности длиннее 512, то необходимо выполнить усечение.

Затем модель BERT выдаст вектор встраивания размером 768 в каждую из лексем.

Далее можно использовать эти векторы в качестве входных данных для различных NLP-приложений, будь то классификация текста, предсказание следующего предложения, распознавание именованных существ (NER) или ответы на вопросы.

Для задачи классификации текста в данной научной работе внимание было сосредоточено на векторе встраивания, полученном из специального токена [CLS].

Это означает, что будет использован вектор встраивания размером 768 из токена [CLS] в качестве входа для текущего классификатора, который затем выведет вектор размером с количество классов в данной задаче классификации.

Для токенизации использовалась предобученная модель RuBERT на русской части Википедии и новостных данных (русский язык, cased, 12-слойная, 768–hidden, 12–heads, 180 миллионов параметров) [24].

BERT разработана для предварительного обучения глубоких двунаправленных представлений на основе немаркированного текста путем совместного учета левого и правого контекста во всех слоях. Поэтому предва-

рительно обученная модель BERT может быть точно настроена с помощью всего одного дополнительного выходного слоя для создания самых современных моделей распознавания текстовой информации без существенных изменений архитектуры, специфичных для конкретной задачи.

10. Импортирование предобученной модели RuBERT для дальнейшего переобучения с целью классификации первичной текстовой геолого-геофизической информации.

Каждое геолого-геофизическое описание рассматривается как уникальная последовательность, которая будет отнесена к одному из девяти кластеров (то есть к названию исследуемой горной породы).

11. Для загрузки обучающей и тестовой выборок используются специальные загрузчики данных. DataLoader объединяет набор данных и семплер и предоставляет итерабельную выборку по заданному набору данных. RandomSampler используется для обучения, a SequentialSampler — для проверки.

12. Чтобы сконструировать оптимизатор (optimizer), необходимо передать ему итерабельную переменную, содержащую параметры для оптимизации. Затем можно задать специфические для оптимизатора параметры, такие как скорость обучения, флаг разогрева и т. д.

Алгоритм разогрева (scheduler) создается со скоростью обучения, которая линейно уменьшается от начальной скорости обучения, установленной в оптимизаторе, до 0.

После периода разминки, в течение которого она будет линейно увеличиваться от 0 до начальной скорости обучения, значение которой установлено в оптимизаторе.

13. В качестве показателей эффективности используются «F1-мера» и точность (Accuracy) на каждый кластер.

14. Задаются seed–значения перед процедурой обучения нейросети для подготовки случайных весов.

15. Процесс обучения сети итеративный. Для каждой эпохи на вход сети подаются входные примеры, маски внимания (это двоичные маски для идентификации лексем в виде настоящих слов или просто случайных/произвольных заполнений).

Если лексема содержит [CLS], [SEP] или любое другое реальное слово, то маска будет равна 1; если же лексема представляет собой просто набивку или [PAD], то маска будет равна 0 (эталонные метки); на выходе выдаются целевые метки; вычисляются потери (loss) для тренировочного и тестового наборов данных; пересчитываются параметры оптимизатора (optimizer) и алгоритма разогрева (scheduler); вычисляется взвешенное значение метрики качества прогноза «F1-мера». Каждую эпоху переобученная модель сохраняется на локальный диск в отдельный файл (весом в среднем 600–700 МБ).

16. Для оценки модели на определенной эпохе была определена функция “evaluate”. Таким же образом подаются входные данные тестовой выборки и выводятся значения потерь, а также предсказанные и эталонные метки в связке с входными примерами.

17. Производится оценка (“evaluate”) и выводится точность прогноза на каждый кластер (“accuracy_per_class”).

18. Для визуальной оценки модели классификации выполняется построение графика Confusion Matrix. Он показывает расхождения между предсказанными и фактическими метками. Подавляющее большинство предсказаний должно оказаться на диагонали (предсказанная метка = фактическая метка). Тем не менее может быть несколько ошибочных классификаций.

Результаты и обсуждение

В процессе составления перечня исходных данных и их загрузки в алгоритм классификации нейросетью (при обучении и тестировании) выявлены следующие ключевые особенности, соблюдение которых необходимо для достижения наилучшего результата прогноза: большой объем описания каждого примера; отсутствие цифр в описании каждого примера; отсутствие знаков препинания, диакритических знаков и прочих символов (точки, запятые, точка с запятой, двоеточие, восклицательные и вопросительные знаки, верхние и нижние подчеркивания, знаки вектора).

Выявлено, что на качество результата прогноза не влияет наличие в описании «скобок» и «тире».

В качестве исходных данных для обучения и отладки алгоритма описания текстовой геолого-геофизической информации использовался категориально разделенный перечень описаний (специализированные метрики качества прогноза; более 3 000 примеров для обучения).

Каждому описанию присвоен один из девяти условных кластеров: «Полимиктовый песчаник», «Мономиктовый песчаник», «Олигомиктовый песчаник», «Глинистая порода (глинистый сланец)», «Песчано-глинистая порода (хлоритовый сланец — филлит)», «Глинисто-иловая порода (аргиллит)», «Глинисто-карбонатная порода (мергель)», «Известняк», «Илистая порода (алевролит)».

Формат исходных данных представлен в виде файла с расширением .csv, что соответствует основному формату загрузки текстовой информации в синтаксисе языка Python.

Предварительно задан заголовок в начале содержимого матрицы результатов, отражающий «столбцы» вида «description; cluster». Далее, каждый пример (описание) представлен построчно, с указанием класса в виде «текст; класс» с разделителем «точка с запятой».

Обучение каждого алгоритма выполнялось в 25 итераций. Объем обучающей выборки составил 75 %, тестовой выборки — 25 %. Производительность алгоритмов определялась на тестовой выборке (“validation accuracy”). В качестве метрик оценки классификации использовалась F-мера (“F1–Score”) (рис. 2).

Графики функции потерь алгоритмов для трех нейросетей (TextCNN; BiLSTM; BERT) на тестовой и обучающей выборках представлены на рисунках 3, 4.

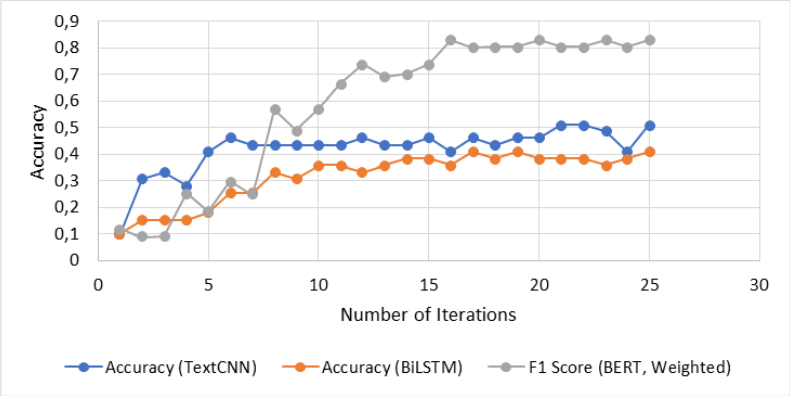


Рис. 2. ***Графики точности прогноза***

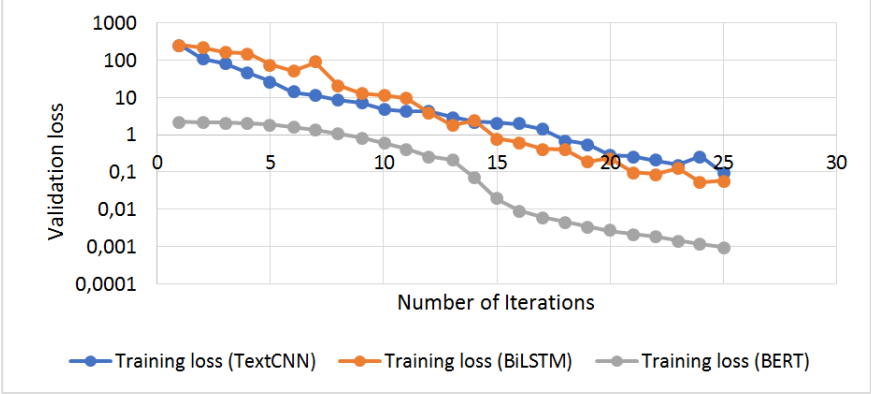


Рис. 3. ***График функции потерь на обучающей выборке***

Из рисунков 2, 3 видна достаточно хорошая тенденция к обучению для алгоритма на базе нейросети BERT-потери, начиная с 10 эпохи, стремительно уменьшаются почти к 0.

Сети TextCNN и BiLSTM, скорее всего, чрезмерно подгоняются нейросетью, поскольку доля потерь возрастает — возможно, они изучают закономерности с учетом ошибок первого рода — то есть те закономерности, которые случайно оказываются верными в обучающих данных, но не имеют под собой оснований в реальности и поэтому неверны в тестовых данных.

В нейросети BERT подобный эффект не наблюдается, поскольку она является уже предобученной на большом объеме текстовой информации и имеет базовую высокую точность.

Несмотря на вышеперечисленные особенности алгоритмов TextCNN и BiLSTM, общая тенденция к обучению/тестированию свидетельствует об улучшении результатов прогноза (необходимо еще увеличить количество примеров для обучения).

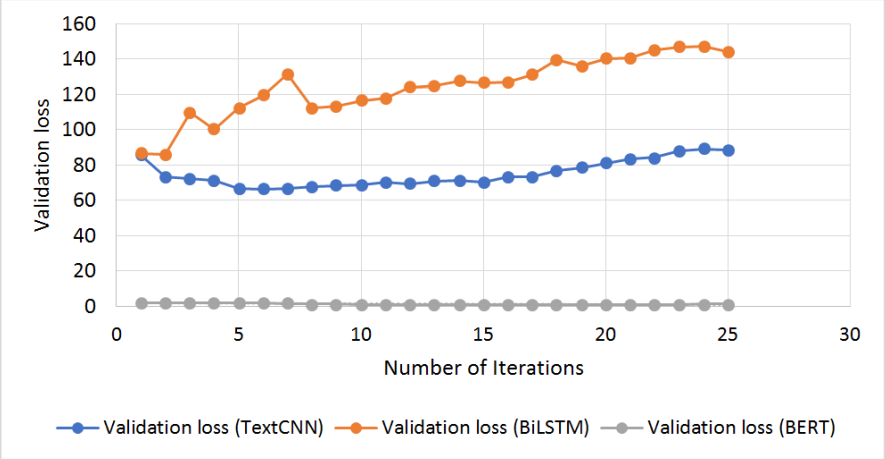


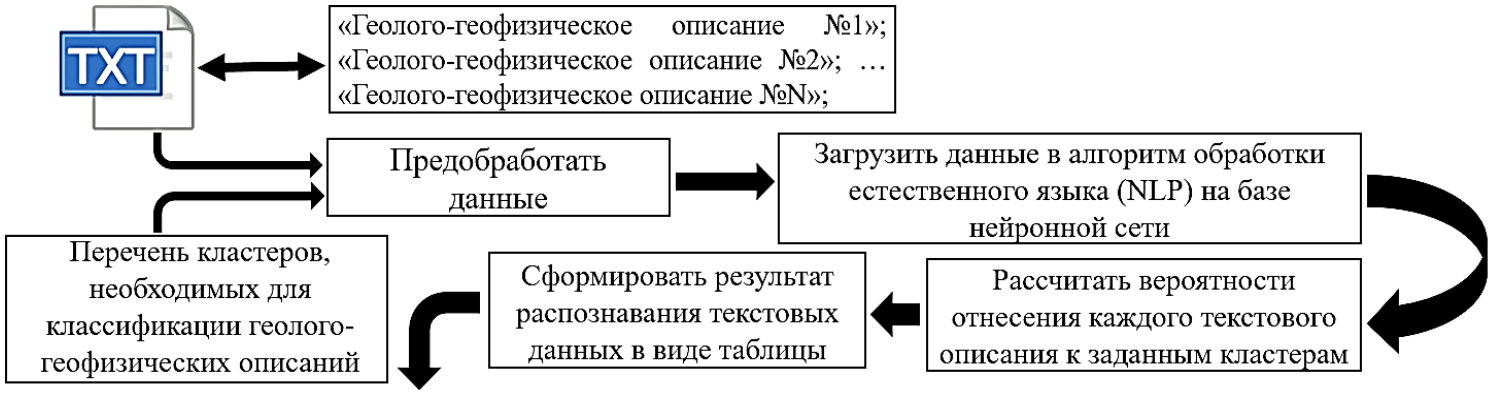
Рис. 4. ***График функции потерь на тестовой выборке***

Подробное разделение точности прогноза наборов геолого-геофизического описания по классам (на тестовой выборке) приведено в таблице. Относительная точность прогноза для каждого класса вычисляется как отношение числа правильно спрогнозированных случаев на их общее количество в пределах тестовой выборки.

Относительная точность прогноза по классам (BERT, 25 эпоха, тестовая выборка)

Кластер	Относительная точность прогноза
Полимиктовый песчаник	4/4
Мономиктовый песчаник	1/2
Олигомиктовый песчаник	1/2
Глинистая порода (глинистый сланец)	1/2
Песчано-глинистая порода ряда (хлоритовый сланец — филлит)	4/4
Глинисто-иловая порода (аргиллит)	2/3
Глинисто-карбонатная порода (мергель)	3/3
Известняк	6/7
Илистая порода (алевролит)	3/3

В области вывода исходных данных, после их обработки, выводится матрица данных с результатами присвоения каждому описанию соответствующего кластера.



Геолого - геофизическое описание	**Вероятность распознавания Р события А (Р(А), %) / Итоговый кластер**	**Полимиктовый песчаник (Р(А) данного кластера, %)**	**Мономиктовый песчаник (Р(А) данного кластера, %)**	**Олигомиктовый песчаник (Р(А) данного кластера, %)**	**Глинистая порода (глинистый сланец) (Р(А) данного кластера, %)**	**Песчано-глинистая порода ряда (хлоритовый сланец – филлит) (Р(А) данного кластера, %)**	**Глинисто-иловая порода (аргиллит) (Р(А) данного кластера, %)**	**Глинисто-карбонатная порода (мергель) (Р(А) данного кластера, %)**	**Известняк (Р(А) данного кластера, %)**	**Илистая порода (алевролит) (Р(А) данного кластера, %)**
По своему минеральному составу, обломочный материал представленной породы имеет единую минеральную основу …	99.965537 / Мономиктовый песчаник	30.343953	99.965537	32.259771	24.753813	32.071167	26.086401	29.811808	27.763560	30.344942

Рис. 5. ***Структура разработанной нейросетевой модели распознавания текстовой геолого-геофизической информации***

Наилучший результат достигнут при использовании алгоритма распознавания текстовых данных на базе BERT с точностью на валидационном сете (Validation Accuracy) ~0.830173 (25 эпоха), с потерями на валидационном сете (Validation Loss) ~0.244719, с потерями во время обучения (Training Loss) ~0.000984.

Вывод данных возможен в визуальном формате, а также в виде отдельных документов с расширениями .txt или .csv.

Область ввода исходных данных позволяет ввести входную информацию как вручную, так и с помощью средства загрузки файлов (рис. 5).

Данная матрица включает в себя следующие столбцы для вывода результатов распознавания:

- «геолого-геофизическое описание» — новые наборы геолого-геофизической информации для обработки;
- «вероятность распознавания, % / Итоговый кластер» — идентифицированный литотип (согласно данным из столбца «Описание») с соответствующей наибольшей вероятностью распознавания;
- ряд столбцов по каждому из исходных кластеров (с соответствующими вероятностями распознавания), в каждом из которых представлены частные вероятности отнесения новых текстовых наборов (эти значения вероятностей формируются при совпадении каких-то отдельных текстовых фрагментов исходных наборов при обучении и новых наборов при тестировании).

Каждое описание выделяется отдельно (построчно) по нижней границе с использованием соответствующего стиля оформления — чем темнее цвет ячеек при отнесении нового текстового набора к исходным девяти кластерам, тем вероятнее результат распознавания.

Выводы

Определен стек технологий для разработки модуля распознавания текстовой геолого-геофизической информации, в том числе описано ключевое ядро разработки.

Выявлены основные закономерности, по которым необходимо производить подготовку входных данных для использования разработанного пилотного варианта модуля (версия 1.0).

Получен нейросетевой классификатор (вычислительный модуль) первичной текстовой геолого-геофизической информации в условиях информационно-логической неопределенности, схема функционирования которого представлена на рисунке 5.

Для проверки нейросетевой модели на новых данных были использованы три новых текстовых набора, каждый из которых был идентифицирован корректно с вероятностью более 99 %.

Результаты разработки алгоритмической части вычислительного модуля и его апробация показывают, что необходимо дальнейшее совершенствование качества нейросетевого прогнозирования при решении задачи

классификации первичной текстовой геолого-геофизической информации с целью достижения планируемого показателя точности на валидационном сете (Validation Accuracy) ~0.9-1.0.

Список источников

1. Катанов, Ю. Е. Исследование влияния капиллярных явлений при фильтрации двухфазных несмешивающихся жидкостей в пористых средах / Ю. Е. Катанов, А. К. Ягафаров, И. И. Клещенко [и др.]. – DOI 10.31660/0445-0108-2020-1-19-29. – Текст : непосредственный // Известия высших учебных заведений. Нефть и газ. – 2020. – № 1. – С. 19–29.
2. Katanov, Yu. E. A probabilistic and statistical model of rock deformation / Yu. E. Katanov. – Text : electronic // E3S Web of Conferences. – 2021. – Vol. 266. – URL: https://doi.org/10.1051/e3sconf/202126603011. – Published: June, 04, 2021.
3. Katanov, Yu. E. Geological and mathematical description of the rocks strain during behavior of the producing solid mass in compression (Tension) / Yu. E. Katanov, Yu. V. Vaganov, M. V. Listak. – DOI 10.33271/mining15.04.091. – Direct text // Journal of Mines, Metals & Fuels. – 2020. – Vol. 68, Issue 9. – P. 285–293.
4. Ломов, П. А. Аугментация обучающего набора при обучении нейросетевой языковой модели для наполнения онтологии / П. А Ломов, М. Л. Малоземова.– DOI 10.37614/2307-5252.2021.5.12.002 – Текст : непосредственный // Труды Кольского научного центра РАН. Информационные технологии. – 2021. – Вып. 12. – Т. 12, № 5. – С. 22–34.
5. Сайгин, А. А. Векторизация нормативно-справочной информации с помощью модели нейронной сети BERT / А. А. Сайгин, Н. П. Плотникова. – Текст : электронный // Информационные технологии и математическое моделирование в управлении сложными системами : электронный журнал. – 2021. – № 2. – С. 52–59. – URL: https://doi.org/10.26731/2658-3704.2021.2(10).52-59.
6. Соломин, А. А. Современные подходы к мультиклассовой классификации интентов на основе предобученных трансформеров / А. А. Соломин, Ю. А. Иванова. – DOI 10.17586/2226-1494-2020-20-4-532-538. – Текст : непосредственный // Научно-технический вестник информационных технологий, механики и оптики. – 2020. – Т. 20, № 4. – С. 532–538.
7. Text classification models for the automatic detection of nonmedical prescription medication use from social media / M. A. Al-Garadi, Y. C. Yang, H. Cai [et al.]. – Text : elctronic // BMC medical informatics and Decision Making. – 2021. – Vol. 21. – URL: https://doi.org/10.1186/s12911-021-01394-0. Published: January, 26, 2021.
8. Comparison of Pre-Trained Language Models for Multi-Class Text Classification in the Financial Domain / Y. Arslan, K. Allix, L. Veiber [et al.]. – DOI 10.1145/3442442.3451375. – Direct text // Companion Proceedings of the Web Conference. – 2021. – P. 260–268.
9. Çelikten, A. Turkish Medical Text Classification Using BERT / A. Çelikten, H. Bulut. – Text : electronic // 2021 29th Signal Processing and Communications Applications Conference (SIU). IEEE. – 2021. – URL: https://doi.org/10.1109/SIU53274.2021.9477847.

10. Das, S. Identification of Cognitive Learning Complexity of Assessment Questions Using Multi-class Text Classification / S. Das, S. K. D. Mandal, A. Basu. – DOI 10.30935/cedtech/8341. – Text : electronic // Contemporary Educational Technology. – 2020. – Vol. 12, Issue 2. – URL: https://doi.org/10.30935/cedtech/8341.

11. Auto-labelling entities in low-resource text : a geological case study / M. Enkhsaikhan, W. Liu, E. J. Holden, P. Duurin. – DOI: 10.1007/s10115-020-01532-6. – Direct text // Knowledge and Information Systems. – 2021. – Vol. 63. – P. 695–715.

12. Gao, X. Named entity recognition in material field based on Bert-BILSTM-Attention-CRF / X. Gao, Q. Li. – DOI 10.1109/TOCS53301.2021.9688665. – Direct text // 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS). – 2021. – P. 955–958.

13. Glazkova, A. A Comparative study of Feature Types for Age-Based Text Classification / A. Glazkova, Yu. Egorov, M. Glazkov. – DOI 10.1007/978-3-030-72610-2_9. – Direct text // International Conference on Analysis of Images, Social Networks and Texts. – 2020. – P. 120–134.

14. Evaluating Transformer-Based Multilingual Text Classification / S. Groenwold, S. Honnavalli, L. Ou [et al.]. – Text : electronic // arXiv:2004.13939v2 [cs.CL]. – 2020. – URL: https://doi.org/10.48550/arXiv.2004.13939.

15. Research on a geological entity relation extraction model for gold mine based on BERT / X. Huang, Y. Zhu, L. Fu [et al.]. – DOI 10.12090/j.issn.1006-6616.2021.27.03.035. – Direct text // Journal of Geomechanics. – 2021. – Vol. 27, Issue 3. – P. 391–399.

16. BERT for Russian news clustering / A. S. Kabaev, S. V. Khaustov, N. E. Gorlova, A. V. Kalmykov. – Text : electronic // Computational Linguistics and Intellectual Technologies. – 2021. – URL: https://doi.org/10.28995/2075-7182-2021-20-385-390.

17. Chinese named entity recognition in the geoscience domain based on BERT / X. Lv, Z. Xie, D. Xu [et al.]. – Text : electronic // Earth and Space Science. – 2022. – Vol. 9, Issue 3. – URL: https://doi.org/10.1029/2021EA002166. – Published: February, 14, 2022.

18. What is this article about? Generative summarization with the BERT model in the geosciences domain / K. Ma, M. Tian, Y. Tan [et al.]. – DOI 10.1007/s12145-021-00695-2. – Direct text // Earth Science Informatics. – 2022. – Vol. 15. – P. 21–36.

19. Piao, G. Scholarly Text Classification with Sentence BERT and Entity Embeddings / G. Piao. – DOI 10.1007/978-3-030-75015-2_8. – Direct text // PAKDD 2021 : Trends and Applications in Knowledge Discovery and Data Mining. – 2021. – P. 79–87.

20. Prabhu, S. Multi-class Text Classification using BERT-based Active Learning / S. Prabhu, M. Mohamed, H. Misra. – Text : electronic // arXiv:2104.14289v2 [cs.IR]. – 2021. – URL: https://doi.org/10.48550/arXiv.2104.14289.

21. A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification / R. Qasim, W. H. Bangyal, M. A. Alqarni, A. Ali Almazroi. – Text : electronic // Journal of Healthcare Engineering. – 2022. – URL: https://doi.org/10.1155/2022/3498123. – Published: January, 07, 2022.

22. Text classification on software requirements specifications using transformer models / D. Kici, A. Bozanta, M. Cevik. [et al.]. – DOI 10.5555/3507788.3507811. – Direct text // Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering. – 2021. – P. 163–172.

23. Lun, C. H. Extracting Knowledge with NLP from Massive Geological Documents / C. H. Lun, T. Hewitt, S. Hou // 82nd EAGE Annual Conference & Exhibition. European Association of Geoscientists & Engineers. – 2021. – URL: https://doi.org/ 10.3997/2214-4609.202112807.

24. Smetanin, S. I. Toxic comments detection in Russian / S. I. Smetanin. – DOI 10.28995/2075-7182-2020-19-1149-1159. – Direct text // Computational Linguistics and Intellectual Technologies. – 2020. – P. 1149–1159.

References

1. Katanov, Yu. E., Yagafarov, A. K., Kleshchenko, I. I. Savina, M. E., Shlein, G. A., & Yagafarov, A. K. (2020). Studying the influence of capillary phenomena in two-phase filtration of immiscible fluids in porous media. Oil and Gas Studies, (1), pp. 19-29. (In Russian). DOI: 10.31660/0445-0108-2020-1-19-29

2. Katanov, Yu. E. (2021). A probabilistic and statistical model of rock deformation. E3S Web of Conferences, 266. (In English). Available at: https://doi.org/10.1051/e3sconf/202126603011

3. Katanov, Yu. E., Vaganov, Yu. V., & Listak, M. V. (2020). Geological and mathematical description of the rocks strain during behavior of the producing solid mass in compression (Tension). Journal of Mines, Metals & Fuels, 68(9), pp. 285-293. (In English). DOI: 10.33271/mining15.04.091

4. Lomov, P. A., & Malozemova, M. L. (2021). Training set augmentation in training neural-network language model for ontology population. Transactions of the Kola Science Centre. Information technologies. Series 12, 12(5), pp. 22-34. (In Russian). DOI: 10.37614/2307-5252.2021.5.12.002

5. Saygin, A. A., & Plotnikova, N. P. (2021). Vectorization of regulatory-reference information using the BERT neural network. Information technology and mathematical modeling in the management of complex systems, (2), pp. 52-59. (In Russian). Available at: https://doi.org/10.26731/2658-3704.2021.2(10).52-59

6. Solomin, A. A., & Ivanova, Yu. A. (2020). Modern approaches to multiclass intent classification based on pre-trained transformers. Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 20(4), pp. 532-538. (In Russian). DOI: 10.17586/2226-1494-2020-20-4-532-538

7. Al-Garadi, M. A., Yang, Y. C., Cai, H., Ruan, Y., O'Connor, K., Graciela, G. H., & Sarker, A. (2021). Text classification models for the automatic detection of nonmedical prescription medication use from social media. BMC medical informatics and Decision Making, 21. (In English). Available at: https://doi.org/10.1186/s12911-021-01394-0

8. Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T. F., Klein, J., & Goujon, A. (2021). Comparison of Pre-Trained Language Models for Multi-Class Text Classification in the Financial Domain. Companion Proceedings of the Web Conference. pp. 260-268. (In English). DOI: 10.1145/3442442.3451375

9. Çelikten, A., & Bulut, H. Turkish Medical Text Classification Using BERT. (2021). 2021 29th Signal Processing and Communications Applications Conference (SIU). IEEE. (In English). Available at: https://doi.org/10.1109/SIU53274.2021.9477847

10. Das, S., Mandal, S. K. D., & Basu, A. (2020). Identification of Cognitive Learning Complexity of Assessment Questions Using Multi-class Text Classification. Contemporary Educational Technology, 12(2). (In English). Available at: https://doi.org/10.30935/cedtech/8341

11. Enkhsaikhan, M., Liu, W., Holden, E. J., & Duuring, P. (2021). Auto-labelling entities in low-resource text: a geological case study. Knowledge and Information Systems, 63, pp. 695-715. (In English). DOI: 10.1007/s10115-020-01532-6

12. Gao, X., & Li, Q. (2021). Named entity recognition in material field based on Bert-BILSTM-Attention-CRF. 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), pp. 955-958. (In English). DOI: 10.1109/TOCS53301.2021.9688665

13. Glazkova, A., Egorov, Y., & Glazkov, M. (2020). A Comparative study of Feature Types for Age-Based Text Classification. International Conference on Analysis of Images, Social Networks and Texts, pp. 120-134. (In English). DOI: 10.1007/978-3-030-72610-2_9

14. Groenwold, S., Honnavalli, S., Ou, L., Parekh, A., Levy, S., Mirza, D., & Wang, W. Y. (2021). Evaluating Transformer-Based Multilingual Text Classification. arXiv:2004.13939v2 [cs.CL]. (In English). Available at: https://doi.org/10.48550/arXiv.2004.13939

15. Huang, X., Zhu, Y., Fu, L., Liu, Y., Tang, K., & Li, J. (2021). Research on a geological entity relation extraction model for gold mine based on BERT. Journal of Geomechanics, 27(3), pp. 391-399. (In English). DOI: 10.12090/j.issn.1006-6616.2021.27.03.035

16. Kabaev, A. S., Khaustov, S. V., Gorlova, N. E., & Kalmykov, A. V. (2021). BERT for Russian news clustering. (In English). Available at: https://doi.org/10.28995/2075-7182-2021-20-385-390

17. Lv, X., Xie, Z., Xu, D., Jin, X., Ma, K., Tao, L., Qiu, Q., & Pan, Y. (2022). Chinese named entity recognition in the geoscience domain based on BERT. Earth and Space Science, 9(3). (In English). Available at: https://doi.org/10.1029/2021EA002166

18. Ma, K., Tian, M., Tan, Y., Xie, X., & Qiu, Q. (2022). What is this article about? Generative summarization with the BERT model in the geosciences domain. Earth Science Informatics, (15) pp. 21-36. (In English). DOI: 10.1007/s12145-021-00695-2

19. Piao, G. (2021). Scholarly Text Classification with Sentence BERT and Entity Embeddings. PAKDD 2021: Trends and Applications in Knowledge Discovery and Data Mining, pp. 79-87. (In English). DOI: 10.1007/978-3-030-75015-2_8

20. Prabhu, S., Mohamed, M., & Misra, H. (2021). Multi-class Text Classification using BERT-based Active Learning. arXiv:2104.14289v2 [cs.IR]. (In English). Available at: https://doi.org/10.48550/arXiv.2104.14289

21. Qasim, R., Bangyal, W. H., Alqarni, M. A., & Ali Almazroi, A. (2022). A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification. Journal of Healthcare Engineering. (In English). Available at: https://doi.org/10.1155/2022/3498123

22. Kici, D., Bozanta, A., Cevik, M., Parikh, D., & Başar, A. (2021). Text classification on software requirements specifications using transformer models. Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering, pp. 163-172. (In English). DOI: 10.5555/3507788.3507811

23. Lun, C. H., Hewitt, T., & Hou, S. (2021). Extracting Knowledge with NLP from Massive Geological Documents. 82nd EAGE Annual Conference & Exhibition. European Association of Geoscientists & Engineers. (In English). Available at: https://doi.org/10.3997/2214-4609.202112807

24. Smetanin, S. I. (2020). Toxic comments detection in Russian. Computational Linguistics and Intellectual Technologies, pp. 1149-1159. (In English). DOI: 10.28995/2075-7182-2020-19-1149-1159

Информация об авторах

Катанов Юрий Евгеньевич, *кандидат геолого-минералогических наук, доцент кафедры прикладной геофизики, ведущий научный сотрудник лаборатории технологий капитального ремонта скважин и интенсификации притока, Тюменский индустриальный университет, г. Тюмень, katanov-juri@rambler.ru, ORCID: https://orcid.org/0000-0001-5983-4040*

Ягафаров Алик Каюмович, *доктор геолого-минералогических наук, профессор, Тюменский индустриальный университет, г. Тюмень*

Аристов Артем Игоревич, *лаборант лаборатории цифровых исследований в нефтегазовой отрасли, Тюменский индустриальный университет, г. Тюмень*

Новрузов Орхан Джанполад оглы, *лаборант лаборатории цифровых исследований в нефтегазовой отрасли, Тюменский индустриальный университет, г. Тюмень*

Information about the authors

Yuri E. Katanov, *Candidate of Geology and Mineralogy, Associate Professor at the Department of Applied Geophysics, Leading Researcher at Well Workover Technology and Production Stimulation Laboratory, Industrial University of Tyumen, katanov-juri@rambler.ru, ORCID: https://orcid.org/0000-0001-5983-4040*

Alik K. Yagafarov, *Doctor of Geology and Mineralogy, Professor, Industrial University of Tyumen*

Artyom I. Aristov, *Assistant at the Laboratory of Digital Research in the Oil and Gas Industry, Industrial University of Tyumen*

Orchan D. Novruzov, *Assistant at the Laboratory of Digital Research in the Oil and Gas Industry, Industrial University of Tyumen*

Статья поступила в редакцию 10.03.2023; одобрена после рецензирования 16.04.2023; принята к публикации 21.04.2023.
The article was submitted 10.03.2023; approved after reviewing 16.04.2023; accepted for publication 21.04.2023.