УДК 658.71.08,519.87
DOI: 10.26467/2079-0619-2024-27-4-34-49

Метод выявления актуальных тем тренажерной подготовки пилотов на основе кластеризации отчетов по безопасности полетов

З.Р. Заббаров[1,2], А.К. Волков[2]
[1]ПАО «Аэрофлот – российские авиалинии», г. Москва, Россия
[2]Ульяновский институт гражданской авиации имени Главного маршала авиации Б.П. Бугаева, г. Ульяновск, Россия

Аннотация: Технологии обработки естественного языка (natural language processing – NLP) в одном из своих применений обеспечивают эффективное исследование закономерностей и тенденций в больших наборах текстовых данных. Текстовые данные по безопасности полетов, представленные в виде отчетов по расследованию авиационных происшествий, являются перспективным объектом для извлечения новой полезной информации, которую можно использовать как при управлении безопасностью полетов, так и в рамках тренажерной подготовки. В данной работе рассматриваются вопросы применения технологий NLP для исследования корпуса отчетов по безопасности полетов ПАО «Аэрофлот – российские авиалинии». Целью исследования является разработка метода выявления актуальных тем тренажерной подготовки пилотов. Представлен анализ существующих зарубежных исследований в области интеллектуального анализа текстовой информации в гражданской авиации. Выявлено, что за рубежом активно применяют технологии NLP для изучения отчетов по безопасности полетов. В статье представлена схема метода выявления актуальных тем тренажерной подготовки пилотов, основанного на кластеризации отчетов по безопасности полетов. Описаны процедуры предварительной обработки текста и построение его векторного пространства. Научной новизной подхода является то, что в отличие от предыдущих работ предлагается использовать полное векторное представление отчетов по безопасности полетов, которое строится объединением матриц тематических и семантических векторов. Проведена апробация предложенного метода. Анализируемый корпус текстов составил 1080 отчетов. В результате применения алгоритма кластеризации были идентифицированы 36 кластеров, которые затем были визуализированы с помощью алгоритма t-распределенного стохастического эмбеддинга соседей (t-distributed Stochastic Neighbor Embedding – t-SNE). Практическая значимость результатов исследования заключается в том, что подход, основанный на кластеризации отчетов, позволит проводить более глубокий анализ отчетов по безопасности полетов, что может упростить и ускорить работу как специалистов по управлению безопасностью полетов, так и инструкторов по тренажерной подготовке пилотов.

Ключевые слова: безопасность полетов, тренажерная подготовка, отчет, кластеризация, обработка естественного языка, тематическое моделирование, модель Doc2Vec.

Для цитирования: Заббаров З.Р., Волков А.К. Метод выявления актуальных тем тренажерной подготовки пилотов на основе кластеризации отчетов по безопасности полетов // Научный Вестник МГТУ ГА. 2024. Т. 27, № 4. С. 34–49. DOI: 10.26467/2079-0619-2024-27-4-34-49.

A method for identifying relevant topics of pilot simulator training based on clustering of flight safety reports

Z.R. Zabbarov[1,2], A.K. Volkov[2]
[1]Public Joint Stock Company "Aeroflot – Russian Airlines", Moscow, Russia
[2]Ulyanovsk Civil Aviation Institute named after Air Chief Marshal B.P. Bugaev, Ulyanovsk, Russia

Abstract: Natural language processing (NLP) technologies, in one of their applications, provide effective research of patterns and trends in large sets of textual data. Textual safety data presented in the form of accident investigation reports is a promising object for extracting new useful information that can be used both in flight safety management and in the framework of simulator training.

This paper discusses the application of NLP technologies for the study of the body of flight safety reports of PJSC Aeroflot – Russian Airlines. The aim of the work is to develop a method for identifying relevant topics of simulator training for pilots. The paper presents an analysis of existing foreign works in the field of intellectual analysis of textual information in civil aviation. It has been revealed that NLP technologies are actively used abroad to study flight safety reports. The paper presents a scheme of a method for identifying relevant topics of pilot simulator training based on clustering of flight safety reports. The procedures of text preprocessing and the construction of its vector space are described. The scientific novelty of the approach is that, unlike previous works, it is proposed to use a full vector representation of flight safety reports, which is built by combining matrices of thematic and semantic vectors. The proposed method has been tested. The analyzed corpus of texts amounted to 1080 reports. As a result of the clustering algorithm, 36 clusters were identified, which were then visualized using the algorithms t-distributed stochastic embedding of neighbors (t-SNE). The practical significance of the research results lies in the fact that the approach based on clustering of reports will allow for a more in-depth analysis of flight safety reports, which can simplify and speed up the work of both safety management specialists and flight simulator instructors.

Key words: flight safety, simulator training, report, clustering, natural language processing, thematic modeling, Doc2Vec model.

For citation: Zabbarov, Z.R., Volkov, A.K. (2024). A method for identifying relevant topics of pilot simulator training based on clustering of flight safety reports. Civil Aviation High Technologies, vol. 27, no. 4, pp. 34–49. DOI: 10.26467/2079-0619-2024-27-4-34-49

Введение

С появлением больших объемов данных и постоянно растущей компьютерной мощности области применения технологий искусственного интеллекта за последние годы значительно расширились. Применение данных технологий в основном связано с разработкой моделей глубокого и машинного обучения для обнаружения объектов и классификации изображений, обработки текстовых, голосовых и видеоданных. В настоящее время на воздушном транспорте генерируется много информации на естественном языке (к примеру, окончательные или промежуточные отчеты по расследованию авиационных происшествий, добровольные сообщения и др.). При этом ручной анализ таких текстовых данных обычно сложен и требует значительных вложений в человеческие ресурсы. В связи с этим регулирующие органы, к примеру Европейское агентство безопасности полетов[1], обращают внимание на потенциал технологий искусственного интеллекта в развитии гражданской авиации. Также Национальный совет по безопасности на транспорте Соединенных Штатов Америки (National Transportation Safety Board – NTSB) видит прикладную ценность технологий NLP в решении задач в области управления безопасностью полетов [1]. Используя технологии NLP, заинтересованные стороны могут получить ценную информацию о причинах человеческих ошибок в авиации и разработать инновационные подходы для решения этих проблем. Применение технологий NLP также поможет дополнить существующие инструменты организации тренажерной подготовки пилотов, тем более что ИКАО рекомендует внедрять подход к подготовке на основе фактических данных[2] (Evidence-Based Training – EBT). Такой подход предполагает определение, развитие и оценку компетенций, необходимых пилотам для выполнения своей деятельности на основе анализа фактических данных, собранных в ходе их обучения и эксплуатации парка воздушных судов. В отличие от существующих программ тренажерной подготовки, предполагающих наличие формализованных сценариев подготовки, программа EBT предполагает повышение уровня индивидуализации данных сценариев. В связи с этим существует потребность во внедрении моделей машинного и глубокого обучения, основанных на технологиях NLP, которые помогут реализовать программу EBT в авиакомпаниях.

[1] EASA-AI-Roadmap [Электронный ресурс] // EASA. 2020. 33 p. URL: https://www.lboro.ac.uk/media/wwwlboroacuk/content/avrrc/downloads/EASA-AI-Roadmap-v1.0.pdf (дата обращения: 20.01.2024).

[2] DOC 9995: AN/497 Руководство по подготовке персонала на основе анализа фактических данных. 1-е изд. // ИКАО, 2013. 170 с.

Целью настоящей статьи является создание подхода, основанного на технологиях NLP, для кластеризации текстовых описаний отчетов по безопасности полетов, чтобы разработать метод выявления актуальных тем тренажерной подготовки пилотов.

Обзор существующих научных работ

Общая проблема извлечения полезной информации из текстовых данных не нова и широко отражается в различных областях исследований, таких как здравоохранение, IT-сфера, сфера услуг и т. д. С этой точки зрения проанализируем исследования, в которых предпринимались попытки извлечь информацию из отчетов по безопасности полетов с использованием технологий NLP.

В работе [2] реализована кластеризация 931 текстового фрагмента отчетов, содержащих системные сбои и неисправности, собранные гражданской авиацией Китайской Народной Республики в 2017 году. Построение векторного пространства текста осуществлялось с использованием метода «частота документов, обратная частоте терминов» (Term Frequency – Inverse Document Frequency – TF-IDF). В качестве алгоритма кластеризации авторы использовали метод *k*-средних. Визуализация полученных кластеров проводилась с использованием метода многомерного шкалирования.

В статье [3] технологии обработки естественного языка применяются к отчетам по безопасности полетов авиакомпании Air France. Авторы предложили оригинальный подход, основанный на применении лингвистического анализа текста для кодирования текстовой информации с последующей классификацией отчетов. В результате предложен метод автоматического анализа текста с целью вычисления оценки сходства между любыми двумя отчетами в коллекции путем сравнения их описательных частей. В сочетании с хронологической информацией данный метод позволяет выявлять новые риски или необычную частоту известных рисков безопасности полетов.

В работе [4] рассматривается задача кластеризации базы данных системы отчетности по безопасности полетов (Aviation Safety Reporting System – ASRS), разработанной и поддерживаемой Национальным управлением по аэронавтике и исследованию космического пространства Соединенных Штатов Америки. ASRS представляет собой большую базу данных добровольно сообщаемых описаний событий, связанных с безопасностью полетов, и обеспечивает средство анализа того, как различные условия полета влияют на его характер и исход. К настоящему времени включает в себя более миллиона деидентифицированных добровольно представленных отчетов, описывающих инциденты на коммерческих воздушных рейсах. Задача кластеризации и визуализации отчетов решалась с помощью комбинации метода *k*-средних и алгоритма t-SNE соответственно. В результате применения метода было выявлено 10 основных кластеров и в общей сложности 31 подкластер.

В статье [5] представлена методология, которая может использовать авиационные текстовые данные для выявления высокоуровневых причин задержек и отмен рейсов, используя задержки в качестве показателя операционной неэффективности. Набор данных извлекается из ASRS (4 195 отчетов). Векторное пространство признаков текста строится на основе использования модели «мешка» слов (Bag-of-Words – BoW) и метода TF-IDF. Задача кластеризации и визуализации отчетов решается аналогично [2], с использованием метода *k*-средних и алгоритма t-SNE. В результате выявлено семь основных кластеров и в общей сложности 23 подкластера.

В работе [6] предложена система анализа и классификации человеческого фактора (Human Factor Analysis and Classification System, HFACS). Для векторного представления документов использовались два подхода: метод TF-IDF и технология Doc2Vec. Для решения задачи классификации авторы использовали метод полууправляемого присваивание меток (Semi-supervised Label Spreading – LS) и метод опорных векторов (Support

Vector Machine – SVM). В работе[3] представлен подход к автоматической классификации отчетов по безопасности полетов с применением машинного обучения. Предложенная модель способна классифицировать отчеты по семи классам. Набор данных обучения модели составил 19 815 отчетов. В качестве классификатора использовался LightGBM, реализующий деревья принятия решений с градиентным бустингом. Векторное представление отчетов строилось с использованием методов частоты терминов (Term Frequency – TF) и TF-IDF.

В статье [7] описывается применение структурного тематического моделирования (Structural Topic Modeling – STM) к данным ASRS, охватывающим инциденты, произошедшие в период с января 2010 года по апрель 2015 года. STM – это форма тематического моделирования, предполагающая вероятностный способ описания документов в терминах тем [7]. STM является обобщением более часто используемых моделей скрытого распределения Дирихле (Latent Dirichlet Allocation – LDA) и коррелированных тематических моделей.

В работе [8] показана возможность анализа корпуса ASRS в поисках первичных индикаторов проблем безопасности полетов. Для этого авторы адаптировали подход анализа тональности текста. При этом делается предпосылка, что субъективные слова (несколько слов или фраз, используемых в негативном контексте) в отчетах либо являются прямыми причинами, либо связаны с первопричиной инцидента.

В исследовании [9] авторы проанализировали отчеты об инцидентах ASRS на предмет извлечения информации о положительных действиях авиационного персонала в чрезвычайных ситуациях. В последующем предполагалась классификация отчетов по безопасности по выявленному человеческому фактору.

[3] Automatic aviation safety reports classification [Электронный ресурс] // Essay. 2019. 60 p. URL: https://essay.utwente.nl/79286/1/torrescano_MA_ewi.pdf (дата обращения: 23.01.2024).

Таким образом, в настоящее время за рубежом активно применяют технологии NLP для анализа текстовой информации в гражданской авиации (в частности, отчетов по безопасности полетов). С точки зрения российских научных публикаций авторам статьи не удалось найти каких-нибудь работ по этой теме. Это в свою очередь дополнительно подтверждает актуальность данной статьи. Разработка и внедрение инструментов анализа текстовых данных в гражданской авиации на основе технологий NLP позволит решить широкий круг задач, таких как автоматическая классификация и кластеризация отчетов по безопасности полетов, разработка прогнозных моделей, которые позволят выявлять причины авиационных событий и предвидеть потенциальные человеческие ошибки, а также выявлять факторы риска безопасности полетов.

Методы и методология исследования

Рассмотрим методы, которые лежат в основе предлагаемого подхода к выявлению актуальных тем тренажерной подготовки пилотов.

Методы векторного представления слов (текста)

Для решения различных задач, связанных с обработкой текста (классификация, кластеризация и др.), необходимо первоначально построить векторное пространство анализируемого корпуса текстов. Иными словами, каждому текстовому фрагменту требуется сопоставить свой вектор чисел, который и будет подаваться на вход различных алгоритмов машинного и глубокого обучения. Рассмотрим два основных подхода для векторного представления текста:

1) тематическое моделирование текста;

2) модели векторного представления слов, основанные на дистрибутивной семантике.

Тематическое моделирование методом LDA

Тематическое моделирование представляет собой метод идентификации тем в текстах, которые представлены набором слов. Предполагается, что некоторые слова, упомянутые в тексте, представляют отдельную тему или, возможно, некоторую часть от общего дискурса. Среди методов тематического моделирования можно выделить: латентно-семантический анализ (Probabilistic Latent Semantic Analysis – PLSA) и латентное размещение Дирихле (Latent Dirichlet Allocation – LDA).

Рассмотрим модель LDA, которая реализует вероятностную генеративную модель. В данной модели вероятность того, что в документе d встретится слово w, описывается формулой [10]

$$p(d,w)=\sum_{t\in T} p(d)p(w|t)p(t|d),$$

при этом также формулируются следующие предположения:

– векторы документов $\theta_d=(p(t|d):t\in T)$ порождаются одним и тем же вероятностным распределением на нормированных $|T|$-мерных векторах; это распределение удобно взять из параметрического семейства распределений Дирихле $Dir(\theta,\alpha),\alpha\in R^{|T|}$;

– векторы тем $\phi_t=(p(w|t):w\in W)$ порождаются одним и тем же вероятностным распределением на нормированных векторах размерности $|W|$; это распределение удобно взять из параметрического семейства распределений Дирихле $Dir(\theta,\beta),\beta\in R^{|W|}$.

Для подробного ознакомления можно обратиться к оригинальной работе [10].

По результатам тематического моделирования каждому тексту из общего корпуса ставится в соответствие вектор, равный числу выбранных тем, в котором содержатся вероятности принадлежности текста к конкретным темам. Оценка качества тематического моделирования проводится по таким мерам, как сложность (perplexity) модели и ее согласованность (coherence). Для поиска оптимального количества тем можно применить следующий подход: построить множество LDA-моделей с разными значениями количества тем (k) и выбрать ту, которая дает наибольшее значение когерентности.

Векторное представление слов (word embeddings)

Исходно текстовые документы описывались в терминах частотного словаря, однако ему на смену пришли векторные вложения слов (word embeddings). При представлении слов частотным словарем соответствующие им бинарные векторы слишком разрежены. Проблема такого бинарного представления заключается в том, что семантически похожие слова могут иметь сильно различающиеся векторные представления. Эффективный способ решения этой проблемы – использование вложения слов (word embeddings). В данном случае каждый разреженный унитарный вектор, представляющий отдельное слово в документе, отображается в низкоразмерный плотный вектор, так что семантически похожие слова оказываются рядом [11]. Это может существенно снизить разреженность данных. Под «семантически похожими» словами понимаются слова, которые встречаются в похожих контекстах.

Самой популярной моделью вложений слов является модель Word2Vec [11]. Это небольшая однослойная нейронная сеть для предсказания слова по его контексту. Существует два варианта реализации данной модели. Первая называется «непрерывный мешок слов» (continuous bag-of-words – CBOW), вторая – Skip-gram. В реализации Skip-gram контекст слов (выходные слова) предсказывается на основе целевого (входного) слова. В реализации CBOW целевое (выходное) слово предсказывается по близлежащим (вход-

ным) словам (его контексту). В Word2Vec само решение задачи предсказания слов не главное – это средство получения семантических векторов. После обучения модели на большом корпусе текстов матрица весов от входного слоя к скрытому слою нейронов и интерпретируется как векторы для слов. Размерность latent vector l обычно составляет от 100 до 500, в зависимости от масштабов информации в использовавшемся для их обучения корпусе.

Дальнейшим развитием метода Word2Vec является нейросетевая архитектура Doc2Vec, которая будет использоваться в предлагаемом методе и позволяет генерировать семантические векторы для целых предложений и параграфов.

Методы кластерного анализа данных

Кластерный анализ представляет собой процедуру разделения набора входных данных на относительно однородные группы (кластеры) по схожести каких-либо признаков. Для решения задачи кластеризации отчетов по безопасности полетов предлагается использовать алгоритм «Иерархическая основанная на плотности пространственная кластеризация для приложений с шумами» (Hierarchical Density-Based Spatial Clustering of Applications with Noise –HDBSCAN) [12]. К основным преимуществам HDBSCAN относятся: обработка больших объемов данных, которые к тому же могут иметь много шумов и выбросов; отсутствие необходимости задания исходного количества кластеров. Обязательным параметром HDBSCAN является только минимальный размер кластера.

Общий принцип работы HDBSCAN заключается в следующем. Первым шагом является преобразование пространства данных в соответствии с плотностью. Вместо использования единого порога плотности алгоритм строит иерархию кластеров на основе различной плотности. HDBSCAN использует расстояние взаимной достижимости, которое гарантирует, что точки в более плотных областях будут расположены ближе друг к другу в преобразованном пространстве. На шаге построения иерархии кластеров алгоритм начинает с обработки каждой точки данных как отдельного кластера. Затем эти точки объединяются в кластеры на основе расстояния их взаимной достижимости. Высота слияния в дереве указывает расстояние, на котором кластеры объединились, иллюстрируя ландшафт плотности данных. Затем дерево уплотняется путем обрезки ветвей, которые не соответствуют критерию минимального размера кластера.

Качество кластеризации оценивается по таким параметрам, как доля событий, не попавших ни в один кластер n, и коэффициент силуэта s (мера усредненной обособленности кластеров).

Методы визуализации результатов кластерного анализа

Основная идея метода SNE заключается в том, чтобы преобразовать евклидовы расстояния в пространстве высокой размерности в условные вероятности, представляющие сходство [13]. Недостатком SNE является то, что он пытается сблизить в пространстве погружения (обычно двумерном) точки, отстоящие довольно далеко друг от друга в пространстве высокой размерности. Одно из возможных решений – использовать в латентном пространстве распределение вероятностей с более тяжелыми хвостами, устранив тем самым нежелательные силы притяжения между точками, отстоящими далеко друг от друга в пространстве высокой размерности [14]. Для этого применяют t-распределение Стьюдента. Суть работы алгоритма t-SNE заключается в минимизации расхождения функции Кульбака – Лейблера между двумя распределениями вероятностей, что позволяет получить проекцию данных в сниженное пространство [15].

В настоящее время существуют несколько обобщений t-SNE, в которых делается попытка улучшить быстродействие, качество

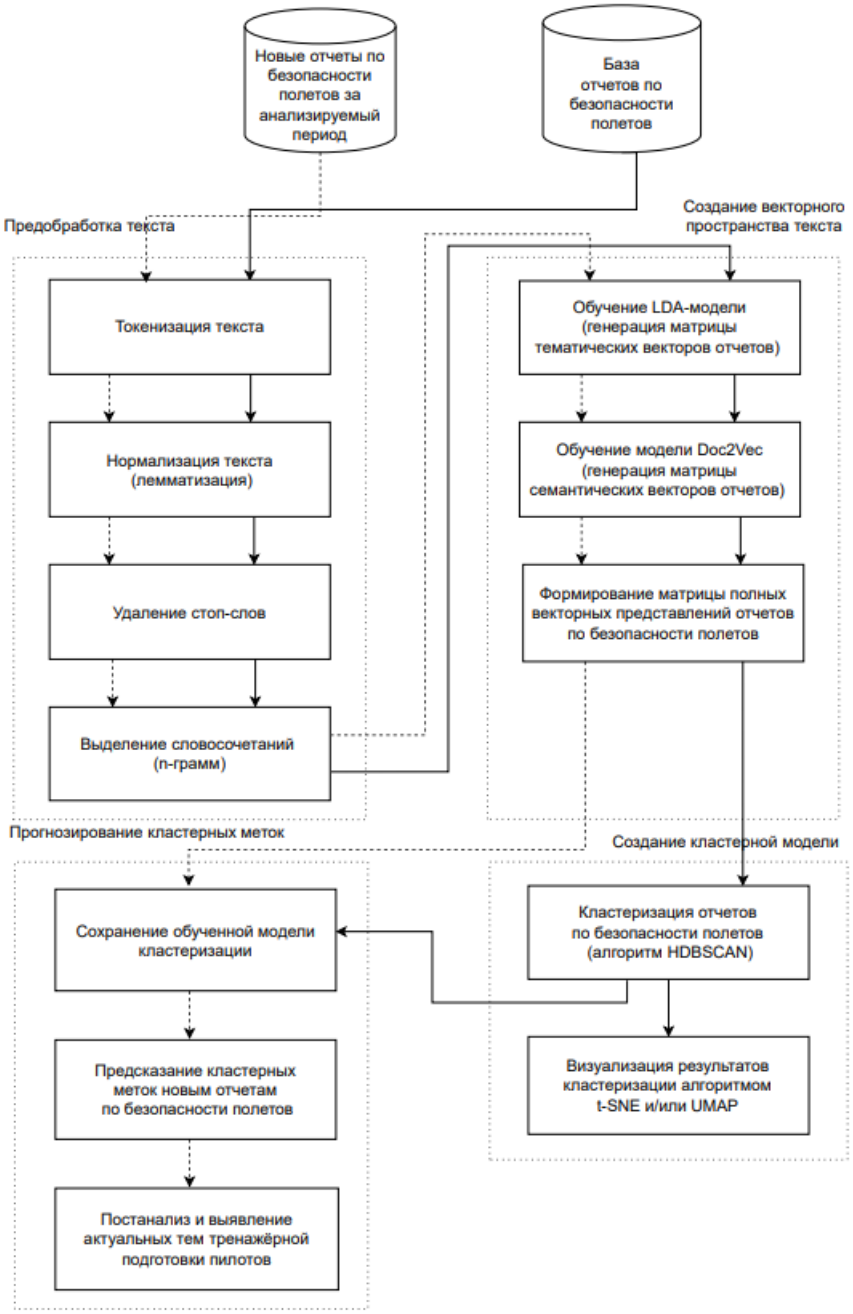

Рис. 1. Схема метода выявления актуальных тем тренажерной подготовки пилотов
Fig. 1. The scheme of the method for identifying relevant topics of simulator training for pilots

пространства погружений или способность к погружению в пространство размерности больше 2. Одним из них является UMAP [11]. На базовом уровне оно похоже на t-SNE, но обычно лучше сохраняет глобальную структуру данных и гораздо быстрее работает.

Метод выявления актуальных тем тренажерной подготовки пилотов на основе кластеризации отчетов по безопасности полетов

Предложенный метод выявления актуальных тем тренажерной подготовки пилотов представлен на рис. 1.

В качестве базы отчетов подразумевается набор материалов расследований, зарегистрированных в какой-либо информационной системе, к примеру в автоматизированной системе обеспечения безопасности полетов авиакомпании. Также в качестве данной базы может использоваться Архив материалов расследований инцидентов и производственных происшествий Росавиации (АМРИПП Росавиации) либо объединенный набор материалов из различных информационных систем. Под новыми отчетами подразумеваются те отчеты, которые не использовались в процессе обучения модели кластеризации и для которых необходимо спрогнозировать кластерные метки.

Рассмотрим конвейер предобработки текстовых данных для дальнейшего обучения моделей, включающий такие методы, как токенизация текста, нормализация (стемминг или лемматизация), удаление стоп-слов, а также выделение словосочетаний (n-грамм). По итогам данного этапа составляется словарь токенов для анализируемого корпуса текстов.

Токенизация в NLP представляет собой процедуру разбиения документов на отдельные сущности (токены), которые несут конкретное информационное содержание. В качестве таких сущностей (токенов) могут выступать абзацы, предложения, фразы, отдельные слова и знаки препинания.

Нормализация – это приведение слов к базовой морфологической форме. Одним из самых распространенных методов нормализации является стемминг, заключающийся в поиске общей основы различных форм слова (как пример, «летаю» – «лета»). Данный метод предполагает нивелирование небольших смысловых различий в окончаниях слов [11]. В предлагаемом методе для нормализации слов используется другая процедура – лемматизация, так как она учитывает значение слова и поэтому является более точной, чем стемминг. Для этого при лемматизации применяется база знаний синонимов и окончаний слов. Это позволяет объединять в один токен только близкие по смыслу слова (как пример, «летаю» – «летать»).

Стоп-слова – это наиболее часто встречающиеся слова в каком-то языке, но при этом несущие меньшую полезную информационную нагрузку о смысле фразы [11]. С точки зрения русского языка к стоп-словам можно отнести союзы (и, а, но, да, если и др.), местоимения (я, ты, твой, его и др.) и т. д. Ввиду особой специфики корпуса отчетов по безопасности полетов к базовым стоп-словам на русском языке был добавлен ряд других, к примеру: пилот, командир, экипаж, КВС и т. п.

N-грамма – это последовательность, содержащая до n элементов, которые были извлечены из последовательности этих элементов, обычно строки [11]. N-граммы могут представлять собой как отдельные буквы, слоги, слова, так и отдельные символы. N-граммы могут генерироваться либо на базе сложных слов, либо по причине привлечения внимания счетчика токенов, если слова встречаются достаточно часто вместе.

После предварительной обработки отчетов обучается LDA-модель и генерируется матрица тематических векторов отчетов по безопасности полетов с количеством столбцов, равным числу тем k. Затем обучается модель Doc2Vec и генерируется матрица семантических векторов отчетов с количеством столбцов, равным величине латентного вектора l. При объединении полученных матриц формируется матрица полных векторных представлений коллекции отчетов по безопасности полетов.

Полная матрица векторных представлений отчетов размером $N \times L$, где N – число отчетов по безопасности полетов, L – размер полного векторного представления (равный $k + l$), подается на вход алгоритма кластеризации HDBSCAN.

Затем происходит визуализация полученной кластерной структуры коллекции отчетов алгоритмами t-SNE и (или) UMAP.

При появлении новых отчетов по безопасности полетов, которые появляются за анализируемый период, полученная модель кластеризации используется, чтобы прогнозировать кластерные метки. Затем специалист проводит постанализ и описывает актуальные темы тренажерной подготовки.

```
25682 диспетчер_давать_указание 4
4654  диспетчер_дать_указание 9
14242 диспетчер_диспетчерский 3
20668 диспетчер_диспетчерский_пункт_круг  5
38413 диспетчер_доложить  1
11907 диспетчер_дп_вышка  2
21010 диспетчер_дпк 16
21011 диспетчер_дпк_опознать  2
13266 диспетчер_дпп 8
9351  диспетчер_дпр 4
45965 диспетчер_дцуп  2
32535 диспетчер_ин  1
11908 диспетчер_кдец  1
27504 диспетчер_кдп 4
34866 диспетчер_командный 1
39889 диспетчер_контролёр 1
29662 диспетчер_красноярск  1
44427 диспетчер_красноярский  1
23630 диспетчер_круг  5
25683 диспетчер_круг_аэродром_якутск  4
23859 диспетчер_л_вадный_расшифровка  2
23293 диспетчер_мдп 2
3523  диспетчер_мур 2
23631 диспетчер_набрать 1
23632 диспетчер_невозможность 3
6696  диспетчер_овд 16
```

```
21997 эскадрилья  1
43948 эстоп 1
43949 эстоп_ао  3
42776 эсу 2
42777 эсу_зт  1
46168 этаж  1
29220 этажерка  1
14128 эталонный 1
10829 этап  37
8995  этап_взлёт  4
24418 этап_возникнуть 1
15346 этап_выравнивание 3
25470 этап_заход  2
2717  этап_заход_посадка  14
25533 этап_набор  4
20078 этап_отвлечение 2
46081 этап_постановка 2
4878  этап_пробег 7
38609 этап_пробег_включение_реверс  3
2190  этап_производство 6
30323 этап_разбег 3
8261  этап_руление  6
10993 этап_снижение 9
9108  этап_снижение_заход_посадка 6
18245 этап_столкновение 4
14207 этап_уход 1
```

Рис. 2. Словарь токенов
Fig. 2. The Token dictionary

Результаты исследования

Результаты апробации предложенного метода

В качестве базы отчетов по безопасности полетов (обучающего набора данных) использовался набор материалов расследований, зарегистрированных в автоматизированной системе обеспечения безопасности полетов ПАО «Аэрофлот – российские авиалинии» (1 080 отчетов за 2019–2020 годы). Отчеты представляют собой текстовый документ, содержащий следующие основные разделы:

– описание события;
– заключение о причинах события;
– типы событий, этапы эксплуатации, причины/факторы;
– рекомендации.

В настоящей работе для анализа был использован только раздел «Описание события», для чего первоначально была проведена предобработка, предполагающая выделение и сохранение в отдельном файле только данного раздела отчетов.

Для предварительной обработки коллекции отчетов использовались библиотеки Natural Language Toolkit, Gensim, а также морфологический анализатор для русского языка pymorphy2. Для выделения биграмм и триграмм использовался метод, который предоставляет модуль Gensim.

На рис. 2 показан фрагмент полученного словаря токенов (в представленном формате первая позиция характеризует идентификатор токена, вторая позиция – сам токен и последняя позиция отображает, сколько раз токен встречается в коллекции текстов).

Для реализации LDA-модели и Doc2Vec также использовался специальный модуль библиотеки Gensim. В качестве количества тем в LDA-модели принято значение равное 20. Данное значение выбрано исходя из анализа зависимости между значением количества тем (k) и значением когерентности LDA-моделей, представленной на рис. 3.

Для оценки качества LDA-модели используются следующие метрики: перплексия (perplexity) и когерентность (coherence) тематик.

Перплексия рассчитывается по следующей формуле [16, 17]:

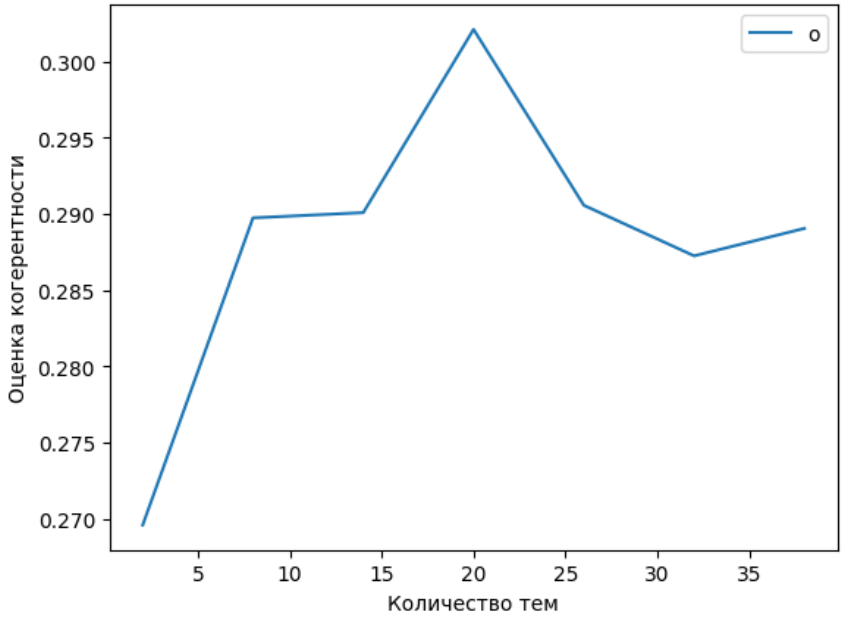

Рис. 3. Зависимость между значениями количества тем (k) и значением когерентности LDA-моделей
Fig. 3. The relationship between the values of the number of topics (k) and the coherence value of LDA models

$$\text{Perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d \mid d)}{\sum_{d=1}^{M} N_d}\right\},$$

где D_{test} – тестовый набор данных; M – количество документов; $p(w_d \mid d)$ – вероятностная модель порождения данных; N_d – количество слов в документе d.

Перпелексию можно интерпретировать как меру несоответствия модели $p(w_d \mid d)$ терминам w, наблюдаемым в документах d коллекции D. Чем ниже значение перплексии, тем лучше модель описывает данные.

Когерентность тематик характеризует их согласованность, значение которой тем выше, чем выше среднее значение когерентности. Когерентность рассчитывается по следующей формуле [18, 19]:

$$C = \sum_{j=2}^{k}\sum_{i=1}^{j-1}\log\frac{D(w_j, w_i)+1}{D(w_i)},$$

где C – когерентность темы для документов, $D(w_i)$ – количество документов, где встречается терм w_i, $D(w_i, w_j)$ – количество документов, где встречаются w_i и w_j одновременно, w_i – i-й термин в порядке убывания веса в матрице «термы – темы» для темы t, k – количество термов в теме t (постоянная 1 включена для возможности вычисления логарифма 0).

Для разработанной LDA-модели значение перплексии составило −10,165, а значение согласованности – 0,287.

Исходя из обучающего объема коллекции отчетов, размерность скрытого слоя (вектора) модели Doc2Vec была выбрана равной 100.

В табл. 1 представлен фрагмент матрицы полных векторных представлений для первых пяти отчетов (L = 120).

В качестве параметров алгоритма HDBSCAN были выбраны следующие (минимальный размер кластера – 10, минимальное количество выборок в окрестности точки, которая будет рассматриваться как основная точка, – 1). По результатам работы алгоритма было выделено 43 кластера с параметрами n = 0,147 и s = 0,06. Оценка качества кластеризации по данным показателям, говорит о том, что есть необходимость в дальнейшем

Таблица 1
Table 1

Матрица полных векторных представлений отчетов
Matrix of full vector representations of reports

№ п/п	Полное векторное представление отчета													
	d_1	d_2	d_3	d_4	d_5	d_6	…	d_{114}	d_{115}	d_{116}	d_{117}	d_{118}	d_{119}	d_{120}
1	0	0	0	0	0	0	…	0,066	1,081	1,111	0,211	0,411	0,039	0,211
2	0	0	0	0	0	0	…	0,078	1,674	1,734	0,271	0,726	0,0491	0,441
3	0	0	0	0	0	0,369	…	0,115	1,627	1,664	0,275	0,662	0,070	0,348
4	0	0	0	0	0	0	…	0,121	1,476	1,497	0,245	0,589	0,065	0,305
5	0	0	0	0	0	0	…	0,081	0,784	0,808	0,137	0,295	0,031	0,169

Таблица 2
Table 2

Пример кластеров
Example of clusters

Номер кластера	Описание отчета в виде токенизированного документа
16	«предполётный_подготовка», «запуск_руление_взлёт_набор», «маршрут», «отклонение», «снижение», «пролёт_опрс», «эшелон_110», «заметить», «уменьшение», «гидрожидкость», «прибор», «литр», «бортмеханик», «пассажирский_салон», «указание», «след_подтекание», «гидрожидкость», «задний», «гондола», «дальнейший_снижение», «подлёт», «разворот», «количество_гидрожидкость_уменьшиться», «литр», «выпуск_шасси», «проводиться», «механически», «бортмеханик», «контроль», «выпустить»
16	«разгерметизация_гидросистема», «кадр», «состояние», «гс_утечка», «жидкость», «мнемокадр», «гидросистема_гс», «уровень_литр», «бортовой», «инженер», «отключить», «гидронасос», «насосный_станция», «рлэ_а1», «следующий», «появиться_надпись», «гс_уровень», «мина», «фактический», «количество_жидкость»
16	«докладная_записка», «горизонтальный_эшелон», «ес_появиться», «сообщение_падение», «уровень_гидрожидкость_жёлтый_гидросистема», «проверить_уровень», «жидкость», «жёлтый_гидросистема», «жёлтый», «изолировать», «зелёный_гидросистема», «отключить», «проанализировать», «обстановка», «погода», «рассчитать_посадочный», «дистанция_фактический», «состояние», «принять_решение», «назначение», «процедура», «заход», «подключить», «насос_жёлтый_гидросистема»

улучшении модели, например варьированием минимальным размером кластера, который является основным инструментом подстройки модели кластеризации под решение конкретной задачи.

В табл. 2 в качестве примера представлен 16-й кластер, выделенный моделью кластеризации, с описанием трех отчетов, попавших в данный кластер, в виде фрагментов токенизированных документов. Всего в него вошли 50 отчетов.

На рис. 4 представлена визуализация полученных кластеров алгоритмом t-SNE в 2D-проекции.

В качестве значений параметров алгоритма t-SNE выбраны следующие: количество измерений – 2; перплексия – 30; метрика близости – евклидово расстояние. Параметр перплексия в алгоритме t-SNE характеризует количество соседей, которое будет учитываться в расчете условных вероятностей сходства.

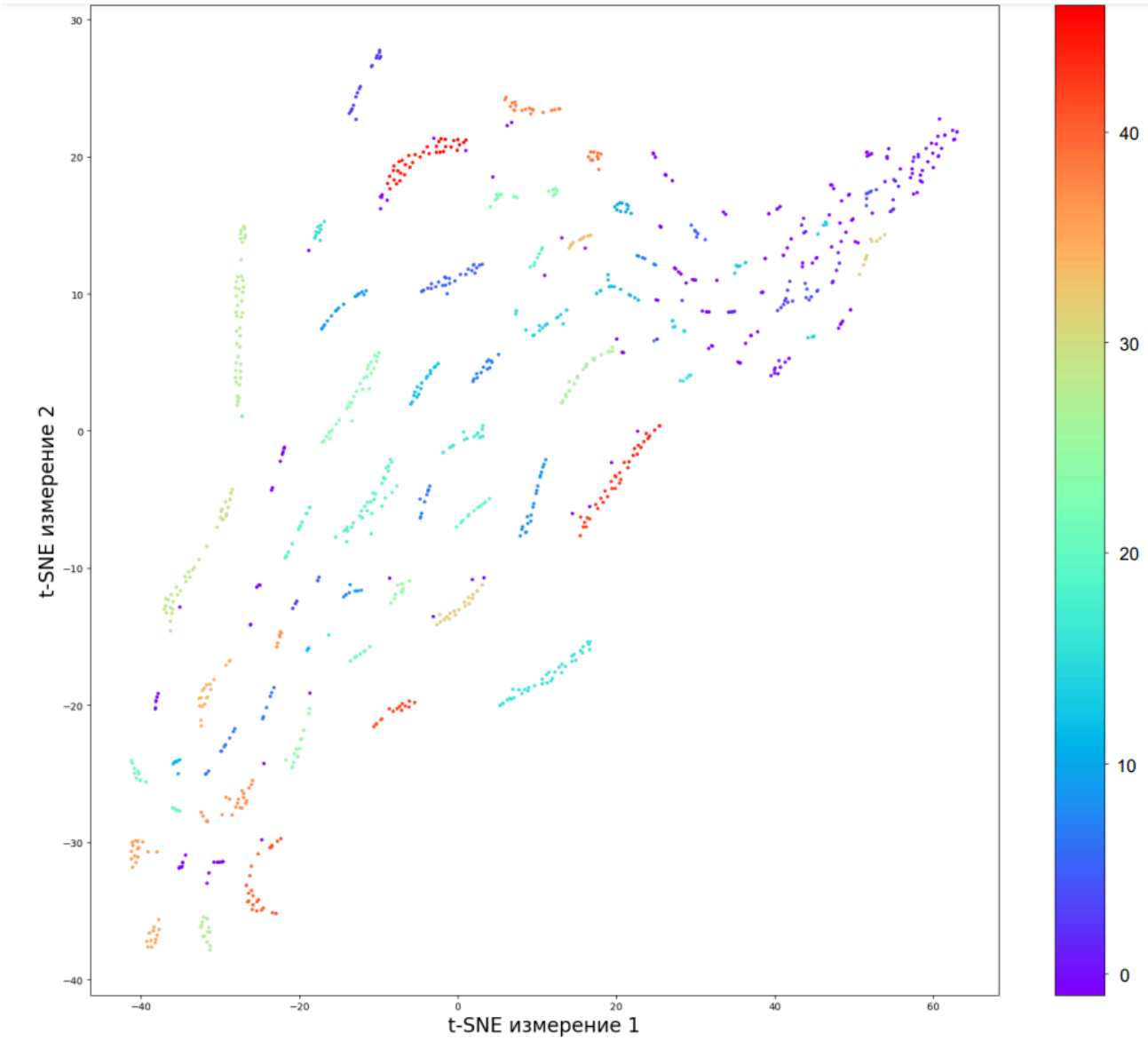

Рис. 4. Визуализация кластеров с помощью алгоритма t-SNE
Fig. 4. Visualization of clusters using the algorithm t-SNE

Шкала справа на рис. 4 отражает номера кластеров, которые также характеризуются и определенным цветом. Точки, имеющие фиолетовый цвет, – это события с меткой кластера равной −1, т. е. не отнесенные ни к одному кластеру.

В качестве примера спрогнозируем кластерную метку для реального инцидента, который произошел 28.12.2020 (исходя из того, что информация об инциденте носит корпоративный характер, конкретные сведения о воздушном судне, бортовом номере, полете опущены). Получено полное векторное представление и токенизированное представление отчета, фрагмент которого показан в табл. 3.

В результате работы предложенного метода, спрогнозирована кластерная метка для данного отчета (16-й кластер, табл. 2) и сделан вывод, что одной из актуальных тем при организации тренажерной подготовки является «Отработка действий при отказах одной или двух гидросистем на различных этапах полета».

Таблица 3
Table 3

Фрагмент токенизированного отчета
A fragment of a tokenized report

Описание отчета в виде токенизированного документа
«набор_маршрут», «особенность», «подготовка», «снижение_заход», «провести_полный_объём», «рпп», «начало_снижение», «эшелон_fl350», «зарегистрировать_прохождение», «сигнализация_master», «caution_сообщение», «низкий_уровень_гидрожидкость», «hyd_rsvr_lo_lvl», «уровень_гидрожидкость_гидросистема_blue», «составить_q», «литр», «сообщение_ec_hyd_rsvr», «выключить», «электрогидравлический_насос», «гидросистема_blue», «blue_elec_pump», «вследствие_отключение», «электрогидравлический_насос», «сработать_сигнализация_низкий_давление», «гидросистема_hyd», «sys_lo», «pr_давление», «bl', 'sys», «снизиться», «3002psi», «76psi», «уровень_гидрожидкость», «бак», «отключение_насос», «повыситься», «q_литр», «объяснительный_записка», «запись_cvr», «гидросистема», «доложить_диспетчер_орвд»

Заключение

В статье предложен подход к выявлению актуальных тем тренажерной подготовки пилотов на основе кластеризации отчетов по безопасности полетов. Научной новизной подхода является то, что в отличие от предыдущих работ предлагается использовать полное векторное представление отчетов по безопасности полетов, которое строится объединением матриц тематических и семантических векторов.

Практическая значимость результатов исследования заключается в том, что подход, основанный на кластеризации, позволяет проводить более глубокий анализ отчетов по безопасности полетов, что может упростить и ускорить работу специалистов по управлению безопасностью полетов.

Разработанный метод показывает перспективные результаты в выявлении общих закономерностей в отчетах по безопасности полетов, которые не очевидны при их ручном анализе, и предлагает новый инструмент для получения информации из текстовых данных о безопасности полетов, который дополнит существующие методы информационной поддержки инструкторов в рамках ЕВТ.

Дальнейшие направления исследований связаны:

– с увеличением коллекции отчетов по безопасности полетов;

– исследованием влияния предварительного понижения размерности полного векторного представления, к примеру методом главных компонент, на качество кластеризации;

– исследованием применения других методов формирования векторного пространства отчетов по безопасности полетов и алгоритмов кластеризации.

Список литературы

1. **Groff L.** Applying natural language processing tools to occurrence reports [Электронный ресурс] // ICAO, 2018. 20 p. URL: https://www.icao.int/safety/iStars/Documents/IUG%20Meeting%201/Presentations/Applying%20Natural%20Language%20Processing%20Tools%20to%20Occurrence%20Reports%20-%20Loren%20Groff.pdf (дата обращения: 20.01.2024).

2. **Junjie L., Huijuan Y., Yinlan D.** Application of text analysis technology in aviation safety information analysis [Электронный ресурс] // Journal of Physics: Conference Series. 2020. Vol. 1624, no. 3. Pp. 032033. DOI: 10.1088/1742-6596/1624/3/032033 (дата обращения: 20.01.2024).

3. **Pimm C.** Natural Language Processing (NLP) tools for the analysis of incident and accident reports / C. Pimm, C. Raynal, N. Tulechki, E. Hermann, G. Caudy [Электронный ресурс] // International Conference on Human-

Computer Interaction in Aerospace (HCI-Aero). Belgium, Brussels, 2012. Pp. 1–7. URL: https://core.ac.uk/download/pdf/50536379.pdf (дата обращения: 20.01.2024).

4. Rose R.L., Puranik T.G., Mavris D.N. Natural language processing based method for clustering and analysis of aviation safety narratives [Электронный ресурс] // Aerospace. 2020. Vol. 7, no. 10. ID: 143. DOI: 10.3390/aerospace7100143 (дата обращения: 20.01.2024).

5. Miyamoto A., Bendarkar M.V., Mavris D.N. Natural language processing of aviation safety reports to identify inefficient operational patterns [Электронный ресурс] // Aerospace. 2022. Vol. 9, no. 8. ID: 450. DOI: 10.3390/aerospace9080450 (дата обращения: 20.01.2024).

6. Madeira T. Machine learning and natural language processing for prediction of human factors in aviation incident reports / T. Madeira, R. Melício, D. Valério, L. Santos [Электронный ресурс] // Aerospace. 2021. Vol. 8, no. 2. ID: 47. DOI: 10.3390/aerospace8020047 (дата обращения: 20.01.2024).

7. Kuhn K.D. Using structural topic modeling to identify latent topics and trends in aviation incident reports // Transportation Research Part C-emerging Technologies. 2018. Vol. 87. Pp. 105–122. DOI: 10.1016/j.trc.2017.12.018

8. Switzer J., Khan L., Muhaya F.B. Subjectivity classification and analysis of the ASRS corpus [Электронный ресурс] // 2011 IEEE International Conference on Information Reuse & Integration. USA, Las Vegas, NV, 2011. Pp. 160–165. DOI: 10.1109/IRI.2011.6009539 (дата обращения: 22.01.2024).

9. Ono M., Nakanishi M. Analysis of human factors and resilience competences in asrs data using natural language processing // Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. HCII 2023. Lecture Notes in Computer Science. 2023. Vol. 14029. Pp. 548–561. DOI: 10.1007/978-3-031-35748-0_37

10. Blei D.M., Ng A.Y., Jordan M.I. Latent dirichlet allocation // The Journal of Machine Learning Research. 2003. Vol. 3. Pp. 993–1022.

11. Мэрфи К.П. Вероятностное машинное обучение: введение / Пер. с англ. А.А. Слинкина. М.: ДМК Пресс, 2022. 990 с.

12. Ester M. A Density-based algorithm for discovering clusters in large spatial databases with Noise / M. Ester, H.P. Kriegel, J. Sander, X. Xu [Электронный ресурс] // Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), USA, Washington, DC, 1996. Pp. 1–6. URL: https://cs.fit.edu/~pkc/classes/ml-internet/papers/ester96kdd-dbscan.pdf (дата обращения: 23.01.2024).

13. Nils H., Gampfer F., Buchkremer R. Latent dirichlet allocation and t-distributed stochastic neighbor embedding enhance scientific reading comprehension of articles related to enterprise architecture // AI. 2021. Vol. 2, no. 2. Pp. 179–194. DOI: 10.3390/ai2020011

14. Van Der Maaten L. Accelerating t-SNE using tree-based algorithms // Journal of Machine Learning Research. 2015. Vol. 15, Pp. 3221–3245.

15. Van der Maaten L., Hinton G.E. Visualizing high-dimensional data using t-SNE // Journal of Machine Learning Research. 2008. Vol. 9. Pp. 2579–2605.

16. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. 2012. № 23. С. 215–242. DOI: 10.15514/ISPRAS-2012-23-13

17. Slutsky A., Hu X., An Y. Tree labeled LDA: a hierarchical model for web summaries [Электронный ресурс] // Proceedings of the 2013 IEEE International Conference on Big Data. USA, Silicon Valley, CA, 2013. Pp. 134–140. DOI: 10.1109/BigData.2013.6691745 (дата обращения: 28.01.2024).

18. Краснов Ф.В., Баскакова Е.Н., Смазневич И.С. Оценка прикладного качества тематических моделей для задач кластеризации // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2021. № 56. С. 100–111. DOI: 10.17223/19988605/56/11

19. Shaheen S., Marco R.S. Full-text or abstract? Examining topic coherence scores using Latent Dirichlet Allocation [Электронный ресурс] // 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Japan, Tokyo, 2017. Pp. 165–174.

DOI: 10.1109/DSAA.2017.61 (дата обращения: 28.01.2024).

References

1. Groff, L. (2018). Applying natural language processing tools to occurrence reports. *ICAO*, 20 p. Available at: https://www.icao.int/safety/iStars/Documents/IUG%20Meeting%201/Presentations/Applying%20Natural%20Language%20Processing%20Tools%20to%20Occurrence%20Reports%20-%20Loren%20Groff.pdf (accessed: 20.01.2024).

2. Junjie, L., Huijuan, Y., Yinlan, D. (2020). Application of text analysis technology in aviation safety information analysis. *Journal of Physics: Conference Series*, vol. 1624. no. 3, pp. 032033. DOI: 10.1088/1742-6596/1624/3/032033 (accessed: 20.01.2024).

3. Pimm, C., Raynal, C., Tulechki, N., Hermann, E., Caudy, G. (2012). Natural Language Processing (NLP) tools for the analysis of incident and accident reports. *In: International Conference on Human-Computer Interaction in Aerospace (HCI-Aero)*. Belgium, Brussels, pp. 1–7. Available at: https://core.ac.uk/download/pdf/50536379.pdf (accessed: 20.01.2024).

4. Rose, R.L., Puranik, T.G., Mavris, D.N. (2020). Natural language processing based method for clustering and analysis of aviation safety narratives. *Aerospace*, vol. 7, no. 10, ID: 143. DOI: 10.3390/aerospace7100143 (accessed: 20.01.2024).

5. Miyamoto, A., Bendarkar, M.V., Mavris, D.N. (2022). Natural language processing of aviation safety reports to identify inefficient operational patterns. *Aerospace*, vol. 9, no. 8. ID: 450. DOI: 10.3390/aerospace9080450 (accessed: 20.01.2024).

6. Madeira, T., Rui, M., Duarte, V., Luis, S. (2021). Machine learning and natural language processing for prediction of human factors in aviation incident reports. *Aerospace*, vol. 8, no. 2. ID: 47. DOI: 10.3390/aerospace8020047 (accessed: 20.01.2024).

7. Kuhn, K.D. (2018). Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C-emerging Technologies*, vol. 87, pp. 105–122. DOI: 10.1016/j.trc.2017.12.018

8. Switzer, J., Khan, L., Muhaya, F.B. (2011). Subjectivity classification and analysis of the ASRS corpus. *In: 2011 IEEE International Conference on Information Reuse & Integration*, USA, Las Vegas, NV, pp. 160–165. DOI: 10.1109/IRI.2011.6009539 (accessed: 22.01.2024).

9. Ono, M., Nakanishi, M. (2023). Analysis of human factors and resilience competences in asrs data using natural language processing. *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. HCII 2023. Lecture Notes in Computer Science*, vol. 14029, pp. 548–561. DOI: 10.1007/978-3-031-35748-0_37

10. Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022.

11. Murphy, K.P. (2022). Probabilistic machine learning: An introduction. Boston: MIT Press, 864 p.

12. Ester, M., Kriegel, H.P., Sander, J., Xu, X. (1996). A Density-based algorithm for discovering clusters in large spatial databases with Noise. *In: KDD-96*, USA, Washington, DC, pp. 1–6. Available at: https://cs.fit.edu/~pkc/classes/ml-internet/papers/ester96kdd-dbscan.pdf (accessed: 23.01.2024).

13. Nils, H., Gampfer, F., Buchkremer, R. (2021). Latent dirichlet allocation and t-distributed stochastic neighbor embedding enhance scientific reading comprehension of articles related to enterprise architecture. *AI*, vol. 2, no. 2, pp. 179–194. DOI: 10.3390/ai2020011

14. Van Der Maaten, L. (2015). Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, vol. 15, pp. 3221–3245.

15. Van der Maaten, L., Hinton, G.E. (2008). Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605.

16. Korshunov, A., Gomzin, A. (2012). Topic modeling in natural language texts. *Trudy Instituta sistemnogo programmirovanija RAN*,

no. 23, pp. 215–242. DOI: 10.15514/ISPRAS-2012-23-13 (in Russian)

17. Slutsky, A., Hu, X., An, Y. (2013). Tree labeled LDA: a hierarchical model for web summaries. *In: Proceedings of the 2013 IEEE International Conference on Big Data*, USA, Silicon Valley, CA, pp. 134–140. DOI: 10.1109/BigData.2013.6691745 (accessed: 28.01.2024).

18. Krasnov, F.V., Baskakova, E.N., Smaznevich I.S. (2020). Assessment of the applied quality of topic models for clustering problems. *Tomsk State University Journal of Control and Computer Science*, no. 56, pp. 100–111. DOI: 10.17223/19988605/56/11 (in Russian)

19. Shaheen, S., Marco, R.S. (2017). Full-text or abstract? Examining topic coherence scores using Latent Dirichlet Allocation. *In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Japan, Tokyo, pp. 165–174. DOI: 10.1109/DSAA.2017.61 (accessed: 28.01.2024).

Сведения об авторах

Заббаров Зульфат Рифкатович, аспирант, Ульяновский институт гражданской авиации; пилот ПАО «Аэрофлот – российские авиалинии», zabbarovz@gmail.com.

Волков Александр Константинович, кандидат технических наук, доцент, доцент кафедры обеспечения авиационной безопасности ФГБОУ ВО «Ульяновский институт гражданской авиации», volkovalex8@rambler.ru.

Information about the authors

Zulfat R. Zabbarov, Postgraduate Student, Ulyanovsk Civil Aviation Institute named after Air Chief Marshal B.P. Bugaev; Pilot, PJSC «Aeroflot – Russian Airlines», zabbarovz@gmail.com.

Alexander K. Volkov, Candidate of Technical Sciences, Assistant Professor, Assistant Professor at the Department of Aviation Security of Ulyanovsk Civil Aviation Institute named after Air Chief Marshal B.P. Bugaev, volkovalex8@rambler.ru.

Поступила в редакцию	16.03.2024	Received	16.03.2024
Одобрена после рецензирования	27.05.2024	Approved after reviewing	27.05.2024
Принята в печать	25.07.2024	Accepted for publication	25.07.2024