УДК 681.518.5:629.735
МРНТИ 28.17.19, 81.83.20
https://doi.org/10.53364/24138614_2026_41_2_16

Г. Каипбек[1], **А. Савостин**[2*], **К. Кошеков**[1], **Д. Риттер**[2]

[1]Академия гражданской авиации, Алматы, Казахстан
[2]Северо-Казахстанский университет им. М. Козыбаева, Петропавловск, Казахстан

[*]*E-mail: asavostin@ku.edu.kz*

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ АВИАЦИОННОГО ТЕХНИЧЕСКОГО ОБСЛУЖИВАНИЯ НА ОСНОВЕ ПЛОТНОСТНОЙ КЛАСТЕРИЗАЦИИ И БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Аннотация. *Обеспечение безопасности полетов и повышение экономической эффективности технического обслуживания воздушных судов требуют интеллектуального анализа неструктурированных текстовых отчетов. Традиционные методы тематического анализа имеют ограничения, связанные с потерей семантического контекста коротких сообщений и высокой трудоемкостью ручной предобработки текстов. Предметом данного исследования являются методы автоматического извлечения знаний из технических рапортов. Для достижения цели решены задачи формирования оптимального вычислительного конвейера, его программной реализации и экспериментальной проверки на реальных данных.*

Целью исследования является разработка, теоретическое обоснование и экспериментальная верификация комплексного метода автоматического выявления и детальной интерпретации скрытых подгрупп неисправностей.

Разработанный метод базируется на интеграции модели-трансформера (paraphrase-multilingual-MiniLM-L12-v2) для получения контекстных эмбеддингов, алгоритма нелинейного снижения размерности UMAP и иерархической плотностной кластеризации HDBSCAN. Интерпретация тем и автоматическая генерация их человекочитаемых названий реализованы на основе алгоритма c-TF-IDF, семантического отбора KeyBERTInspired и локально развернутой большой языковой модели Qwen3.5-4B-Instruct. Экспериментальная проверка метода проведена на выборке из 1971 текстовой записи категории пассажирского оборудования за 7-летний период эксплуатации девяти воздушных судов. Разработанный метод успешно идентифицировал 7 стабильных содержательных категорий дефектов, изолировав 11,7 % нетипичных записей в шумовой кластер. Сравнение со случайным базовым уровнем подтвердило высокую статистическую значимость результатов. Показатель разнообразия тем составил 0,8286 и соответствует рекомендуемому диапазону.

Разработанный метод превосходит классическую модель LDA по покомпонентной чистоте выделенных тем и исключает этап сложной ручной предобработки данных. Решение рекомендуется для масштабирования на другие разделы стандартов ATA 100 и интеграции в интеллектуальные системы поддержки принятия решений авиапредприятий с целью оптимизации планирования обслуживания и управления запасами компонентов.

Ключевые слова: *техническое обслуживание, авиационный транспорт, интеллектуальный анализ текстов, контекстные эмбеддинги, плотностная кластеризация, тематическое моделирование, большие языковые модели.*

Введение.

Обеспечение безопасности полетов и повышение экономической эффективности процессов технического обслуживания и ремонта (ТОиР) воздушных судов (ВС) являются приоритетными задачами современной гражданской авиации. Внедрение концепций предиктивного обслуживания привело к активному развитию методов диагностического мониторинга на основе анализа структурированных данных, поступающих от бортовых датчиков и систем в режиме реального времени [1]. Тем не менее, существенная часть практически значимой информации о техническом состоянии авиационной техники остается неструктурированной, фиксируясь техническим персоналом и летным составом в виде текстовых отчетов об обнаруженных дефектах и выполненных ремонтных действиях [1]. Систематический интеллектуальный анализ этих текстовых массивов открывает новые возможности для выявления латентных закономерностей отказов и оптимизации процедур обслуживания.

Анализ современных исследований показывает устойчивую тенденцию к переходу от анализа исключительно числовой телеметрии к глубокому семантическому анализу текстовых отчетов ТОиР (программное направление Technical Language Processing, TLP) [2]. Для структурирования этих данных исследователями активно применяются методы классического машинного обучения (SVD, Naive Bayes) и традиционные алгоритмы тематического моделирования [3]. Однако ключевой проблемой существующих решений остается их высокая чувствительность к качеству предобработки текстов и неспособность классических моделей улавливать скрытые контекстуальные связи в ультракоротких сообщениях без построения трудоемких доменных онтологий. Это обуславливает растущий интерес к использованию предобученных моделей-трансформеров и больших языковых моделей (LLM) для автоматизации анализа технических текстов [4].

В предыдущих исследованиях авторами был разработан и апробирован метод автоматической классификации текстовых описаний дефектов по кодам Chapter-Section стандарта ATA 100 (ATA iSpec 2200), что позволяет структурировать поток технических отчетов [5]. Для наиболее массовой и семантически разнородной категории пассажирского оборудования (Chapter-Section 25-21) был проведен первичный анализ внутренней структуры неисправностей с использованием вероятностного тематического моделирования на основе латентного размещения Дирихле (LDA). Однако, хотя этот подход позволил выделить базовые макро-темы (такие как общие проблемы со спинками кресел или столиками), его разрешающая способность оказалась ограниченной. Для оптимизации складских запасов конкретных деталей и эффективного планирования ТОиР требуется переход на качественно новый уровень детализации – к точному покомпонентному анализу неисправностей внутри категории.

На этом уровне детализации классический метод LDA сталкивается с существенными методологическими ограничениями. Короткие технические тексты отчетов ТОиР описывают схожие компоненты очень близкими словами, что в рамках концепции «мешка слов» (bag-of-words) приводит к семантическому перекрытию тем. Кроме того, предыдущий LDA-подход критически зависел от качества предварительной обработки текста, который в этом случае требует трудоемкого ручного построения доменных словарей аббревиатур с привлечением экспертов, лемматизации и жесткой фильтрации, что снижает масштабируемость системы на реальных «сырых» базах данных авиакомпаний.

Для преодоления этих ограничений в данной работе предлагается переход к альтернативному методу выявления скрытых подгрупп неисправностей на основе контекстных векторных представлений и иерархической плотностной кластеризации.

Данный подход базируется на использовании глубоких предобученных нейросетевых моделей-трансформеров для кодирования семантики коротких технических отчетов, что позволяет исключить этап сложной ручной предобработки, сохранив важные контекстуальные маркеры.

На основании этого, целью данного исследования является разработка, теоретическое обоснование и экспериментальная верификация комплексного метода автоматического выявления и детальной интерпретации скрытых подгрупп неисправностей в авиационном ТОиР на основе контекстных эмбеддингов и плотностной кластеризации.

Для достижения поставленной цели в работе решаются следующие задачи:

1. Формулирование и выбор оптимального стека инструментов для анализа текстовых данных в домене авиационного ТОиР.
2. Разработка подхода к выявлению скрытых подтем неисправностей.
3. Экспериментальная верификация и бенчмаркинг.

Материалы и методы исследования.

В качестве базы исследования использовался деидентифицированный набор данных, содержащий информацию о зарегистрированных дефектах и выполненных ремонтных действиях для девяти коммерческих ВС одного типа за 7-летний период эксплуатации (с 2014 по 2020 годы) [5]. Общий объем базы данных составляет 13204 записи. Каждая запись содержит дату регистрации, условный идентификатор борта (буквы латинского алфавита от A до I), текстовые описания дефекта (Defect) и выполненного действия (Action), а также коды системы и подсистемы согласно стандарту ATA 100 (ATA iSpec 2200) (Chapter и Section).

Для проведения детального микроанализа из общего набора данных была выделена наиболее многочисленная и семантически разнородная категория с кодом Chapter-Section «25-21» (Пассажирское оборудование / Салон ВС). Объем выделенной выборки составил 1971 запись. Текстовые описания дефектов в данной выборке характеризуются лаконичностью (медианная длина сообщения составляет 8 слов), обилием профессионального жаргона и специфических сокращений.

Архитектура предлагаемого метода.

Для выявления скрытых подтем неисправностей в неструктурированных отчетах ТОиР в работе предлагается интегрированный метод плотностной кластеризации и генеративной интерпретации технических текстов. Метод представляет собой единый, воспроизводимый вычислительный конвейер (пайплайн), состоящий из трех последовательных этапов, как показано на структурной схеме рисунка 1.

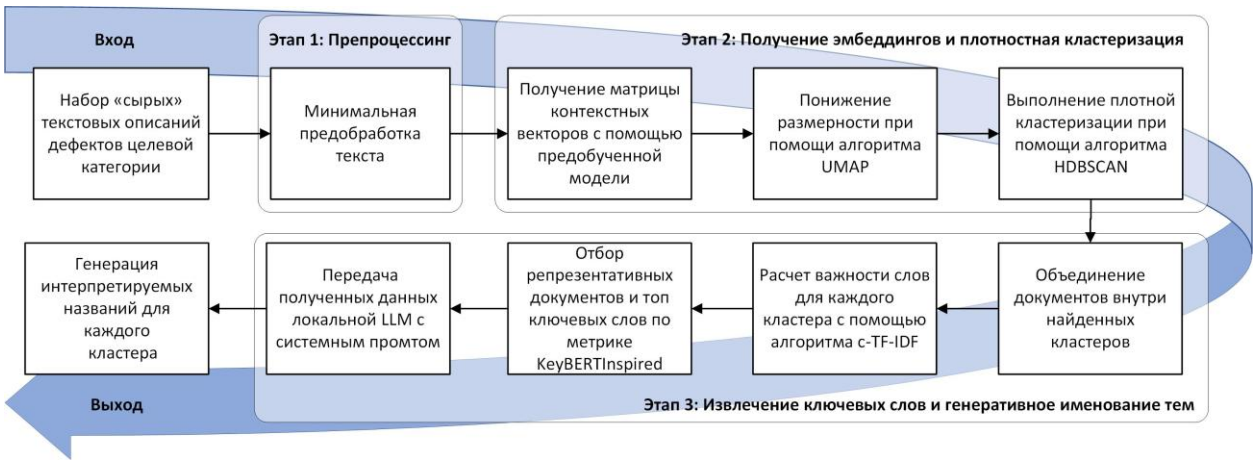

Рисунок 1 – Комплексный метод анализа текстовых данных авиационного ТО

Программная реализация предложенного пайплайна была выполнена на языке Python 3.10 с применением библиотек scikit-learn 1.7.1, sentence-transformers 5.5.1, umap-learn 0.5.12, hdbscan 0.8.43, bertopic 0.7.14 и pandas 2.3.3.

Как следует из рисунка 1, первым этапом предложенного метода является препроцессинг текстовых данных. В отличие от классических вероятностных моделей (например, LDA), требующих глубокой очистки и удаления стоп-слов [1], используемые на последующих этапах современные нейросетевые трансформерные архитектуры опираются на синтаксическую структуру и контекст. Поэтому предобработка документов на данном этапе была минимизирована с целью сохранения контекстуальных связей, необходимых для корректной работы механизма двунаправленного внимания (self-attention).

Процесс препроцессинга включает в себя декодирование XML-артефактов, удаление экранированных управляющих символов, а также приведение текста к нижнему регистру.

В соответствии со схемой рисунка 1, на этапе 2 следующим шагом после предобработки является отображение текстовых описаний дефектов в непрерывное многомерное семантическое пространство. То есть осуществляется преобразование корпуса подготовленных текстов в матрицу контекстных векторов **E** размерности $N \times M$, где N – число описаний дефектов ($N = 1971$), а M – размерность пространства эмбеддингов [7].

Для векторизации используется предобученная модель paraphrase-multilingual-MiniLM-L12-v2, позволяющая сопоставить каждому короткому техническому отчету плотный вектор размерностью $M = 384$. Модель кодирует смысл высказываний, проецируя семантически близкие понятия в соседние координаты векторного пространства, что устраняет ограничения модели «мешка слов». Выбор данной модели обусловлен её доказанной эффективностью на задачах семантического сопоставления коротких текстов и оптимальным балансом между точностью и вычислительными затратами. Будучи моделью из семейства Sentence Transformers, она базируется на архитектуре MiniLM и получена путем дистилляции знаний из более крупных языковых моделей [7], что снижает риск проклятия размерности на последующих этапах.

Далее, как показано на рисунке 1, для снижения размерности эмбеддингов перед кластеризацией используется алгоритм UMAP (Uniform Manifold Approximation and Projection) [8]. Этот шаг необходим для преодоления проклятия размерности, поскольку в пространствах высокой размерности (384) классические метрики расстояний теряют свою эффективность. Метод UMAP нелинейно снижает размерность векторов до 5 компонентов, превосходя PCA и t-SNE по способности сохранять как локальные, так и глобальные семантические связи исходного многообразия. Для обеспечения баланса между локальной структурой микро-кластеров и глобальной топологией данных был выбран параметр *n_neighbors*=15. В качестве метрики сходства использовалось косинусное расстояние (*metric*="cosine"), наиболее точно отражающее семантическую близость контекстных векторов.

На следующем шаге (рисунок 1) для выделения тематических групп был применен алгоритм иерархической плотностной кластеризации HDBSCAN [9]. Использование плотностного алгоритма вместо традиционного метода K-Means обосновано спецификой прикладной задачи. HDBSCAN не требует априорного задания количества кластеров, эффективно выделяет нетипичные единичные записи в шумовой кластер (класс -1) и успешно работает с кластерами произвольной формы и объема.

Математической основой HDBSCAN является вычисление метрики взаимной достижимости (mutual reachability distance) между объектами a и b в низкоразмерном пространстве:

$$d_{mrd}(a,b) = \max\left\{core_k(a), core_k(b), d(a,b)\right\}, \quad (1)$$

где $core_k(x)$ – расстояние от объекта x до его k-го ближайшего соседа, а $d(a, b)$ – стандартное евклидово расстояние между объектами.

Параметры алгоритма HDBSCAN были настроены для обеспечения баланса между глубиной детализации выявляемых неисправностей и уровнем фильтрации шума. Основной параметр – это минимальный размер кластера. Он был установлен равным *min_cluster_size*=15. Для анализируемой выборки это позволяет изолировать малочисленные, но практически значимые и технически специфические категории дефектов. Параметр плотности *min_samples* был установлен равным 5, что обеспечивает умеренную консервативность алгоритма и предотвращает избыточное раздувание шумового класса. В качестве метрики расстояния в HDBSCAN использовалось евклидово расстояние, поскольку алгоритм UMAP проецирует косинусоидальные связи исходных эмбеддингов в низкоразмерное пространство, расстояния в котором приобретают евклидов характер.

Целью третьего этапа предложенного метода (рисунок 1) является преобразование абстрактных геометрических кластеров в понятную, структурированную и интерпретируемую информацию. Для исключения субъективизма в работе применен последовательный многоступенчатый конвейер интерпретации, сочетающий статистический, семантический и генеративный подходы.

Для извлечения текстовых признаков документы, отнесенные к каждому из K содержательных кластеров, объединяются в единые мета-документы. Преобразование полученных мета-документов в дискретное признаковое пространство («мешок слов») выполняется с помощью векторизатора CountVectorizer из библиотеки scikit-learn. Настройки векторизатора были адаптированы к авиационной предметной области. Встроенный токенайзер устраняет морфологическую вариативность терминов, использование униграмм и биграмм позволяет выделять устойчивые словосочетания, а стоп-лист сохраняет значимые маркеры состояния систем. Параметр *min_df*=2 обеспечивает удаление редких опечаток.

После векторизации на основе сформированной матрицы частот токенов вычисляется относительная важность каждого термина t для класса (темы) c с помощью алгоритма c-TF-IDF [10] по формуле:

$$\text{c-TF-IDF}_{t,c} = tf_{t,c} \cdot \log\left(1 + \frac{A}{f_t}\right), \tag{2}$$

где $tf_{t,c}$ – частота слова t в классе c; f_t – общая частота слова t во всех классах; A – среднее количество слов в классах. Для улучшения взвешивания в условиях кластеров существенно различного объема использовалась модификация BM25 weighting (*bm25_weighting*=True).

Для семантической оптимизации и исключения неинформативных частых слов была интегрирована модель KeyBERTInspired (рисунок 1) [11]. Данный алгоритм использует исходные плотные эмбеддинги для расчета косинусного сходства между векторами слов-кандидатов и вектором центроида соответствующего кластера в многомерном пространстве [11]. Это гарантирует, что в качестве ключевых слов темы выбираются не просто часто встречающиеся токены, а лексемы, максимально точно отражающие семантическое ядро кластера.

На заключительном шаге метода, в соответствии с рисунком 1, для автоматического формирования интерпретируемых названий тематических кластеров использовалась локально развернутая LLM Qwen3.5-4B-Instruct [12]. Выбор модели обусловлен сочетанием высокой вычислительной эффективности и возможностью локального выполнения инференса без передачи конфиденциальных данных во внешние облачные сервисы. Для повышения стабильности и воспроизводимости результатов использовались

фиксированные параметры генерации: *temperature*=0.2, *top_p*=0.8 при отключенном режиме пошагового логического вывода (reasoning).

Формирование названий кластеров осуществлялось посредством специально разработанного двухуровневого промпта. Системная часть задавала роль LLM как эксперта по анализу авиационных технических отчетов и содержала ограничения на формат ответа:

```
system_prompt = (
"You are an expert in aircraft reliability and maintenance engineering. "
"Your task is to analyze defect examples and topic keywords and generate a concise, human-readable technical topic title."

"FOLLOW THESE RULES:"
"- The title must be a natural English phrase, not a list of keywords;"
"- Combine related keywords into a meaningful technical concept;"
"- Use standard aviation and maintenance terminology whenever possible;"
"- The title must contain 3 to 6 words;"
"- Use only nouns and adjectives;"
"- Correct grammatical issues found in the keywords;"
"- Prefer broader technical concepts over literal keyword repetition;"
"- Do not include explanations, punctuation, or additional text;"
"- DO NOT output reasoning or thinking process;"
"- Return ONLY the final topic title."
```

Количественная оценка и проверка статистической значимости тематической структуры, полученной разработанным методом, производилась по следующим критериям.

1. Когерентность C_{NPMI} (Normalized Pointwise Mutual Information). Данная метрика оценивает взаимную информацию между парами слов с нормированием результата в диапазоне [-1, 1], где 1 соответствует всегда совместной встречаемости, 0 – независимости слов, а -1 – их отсутствию в совместном контексте [13].

2. Когерентность U_{MASS}, которая оценивает совместную встречаемость слов темы непосредственно внутри границ документов, что делает ее более робастной для фрагментированных текстов ТОиР. Значения метрики лежат в диапазоне $[-\infty, 0]$, где близость к 0 указывает на более высокую когерентность [14].

3. Разнообразие тем (*Topic Diversity*) показывает уникальность выделенных тем и оценивает степень их дублирования. Имеет диапазон значений [0, 1], где близость к 1 указывает на отсутствие дублирования [15].

Результаты и их обсуждение.

В ходе экспериментальной проверки предложенного метода была проведена его апробация на корпусе дефектов категории «25-21». На первом этапе разработанный пайплайн на основе алгоритма HDBSCAN автоматически идентифицировал 30 отдельных микро-тем (без учета шумового кластера -1). Данный результат обусловлен эффектом избыточного дробления, характерным для плотных векторных пространств коротких текстов, когда незначительные лексико-синтаксические вариации описания одного дефекта приводят к формированию множества изолированных микро-групп.

Для анализа структуры полученного тематического пространства были построены карта межтематических расстояний (рисунок 2) и дендрограмма иерархического сходства тем (рисунок 3).

Как иллюстрирует рисунок 2, окружности, представляющие 30 исходных микро-тем, визуально группируются в семь изолированных семантических кластеров. Это наблюдение подтверждается анализом дендрограммы (рисунок 3), согласно которой большинство микро-тем объединяются на высоких уровнях сходства, образуя устойчивые иерархические ветви.

С методологической и практической точек зрения избыточное дробление затрудняет интерпретацию результатов техническим персоналом и снижает разнообразие тем из-за перекрытия их словарей. Объединение семантически близких подтем позволяет сопоставить каждую тему конкретному укрупненному узлу оборудования в соответствии с принципами поддержания летной годности.

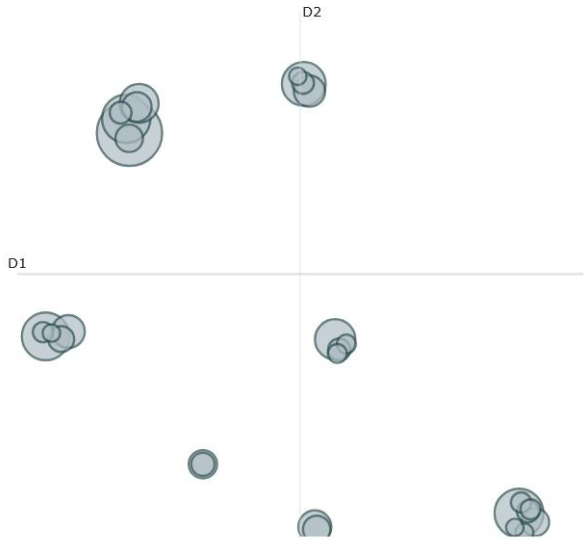

Рисунок 2 – Анализ структуры тематического пространства

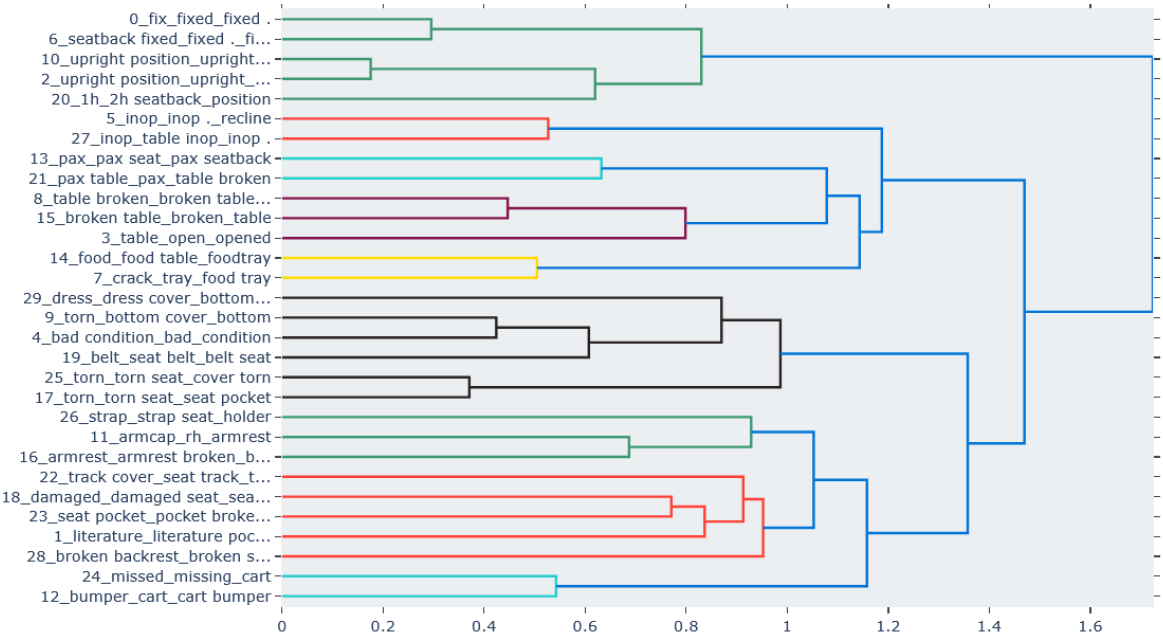

Рисунок 3 – Дендрограмма иерархического сходства

На основе результатов иерархического анализа размерность тематического пространства предложенного метода была сокращена до 7 содержательных категорий с помощью алгоритма reduce_topics (параметр *nr_topics*=8, включая шумовой кластер). Алгоритм итеративно объединил близкие микро-темы на основе косинусного сходства их c-TF-IDF векторов. В результате была получена оптимизированная структура, состоящая из 7 стабильных тем и одного шумового кластера.

Результаты применения разработанного метода для выявления скрытых подгрупп неисправностей в категории пассажирского оборудования ВС представлены в таблице 1.

Наличие пунктуационного шума (например, символа ; в шумовом кластере таблицы 1) среди ключевых слов c-TF-IDF обусловлено отказом от жесткой предобработки исходных текстов. Эксперимент показал, что данный шум не влияет на итоговый результат. Векторизация и кластеризация выполняются в пространстве плотных эмбеддингов, где пунктуация помогает трансформеру точнее кодировать синтаксис. На этапе генерации названий LLM демонстрирует устойчивость к шуму в промпте, успешно игнорируя знаки препинания и формируя чистые, грамматически корректные заголовки тем

Таблица 1 – Оптимизированная тематическая структура дефектов категории «25-21», полученная предложенным методом

№ темы	Ключевые слова c-TF-IDF	Ключевые слова KeyBERT Inspired	**Темы LLM**	Репрезентативные описания дефектов
-1	seat, cover, broken, armrest, station, torn, doe, pen, ;, damaged	seat torn, damaged seat, cover seat, seat track, seat armrest, seatbelt, broken seat, seat, assy seat, seat pocket	**Seat Cover and Armrest Damage**	at seat 2a recline mechanism doesnt work properly, doesnt return in upper position, bottom cover on seat 1a is torn (p/n 4401064-523), 2a seat lh armrest inside plastic cover damaged. (p/n 41401039-11)
0	seatback, fixed, doe, ,, position, upright, seat, upright position, fix, cover	position seatback, seatbacks, fixed seat, seatback, position seat, seatback fixed, fix seat, seat back, condition seat, torn seat	**Seatback Upright Positioning Failure**	32h seat back doesnt fix in upright position, 32c seatback doesnt fix in upright position, seatback 32c doesnt fix in upright position
1	table, ,, broken, food, crack, table seat, tray, food tray, open, food table	table seat, seat table, folding table, small table, table, . table, passenger table, food table, pax table, broken seat	**Seat Food Tray Damage**	seats 33a,38a,40k,41a,43c,44h tables have cracks, food tray tables at seats 4ca,4fd,5d,6a,cd,7ca,7fd,8c,9ca,9f,11a,11fd, 13ca,13d,14ca,14f,16ca,15fd,17f,18a,19f,19 c,21a,21f,23c,23d,24c,24f,25c have cracks., 20c table is broken
2	literature, literature pocket, pocket, spring, pocket seat, pocket spring, condition, bad, ,, bad condition	seat 15h, seat 32c, literature pocket, seat bad, seat 38a, , 36c, 13c, 36k, condition seat, 14k	**Seat Literature Pocket Spring Damage**	literature pocket on seats 11ca, 12ch,13h,13h,44ch are bad condition, 40c,41c,13k seat literature pocket is in bad condition, literature pockets 13h,17c,32a,34c,36ch,41c,42c,43c, 44akch,45c in bad condition
3	inop, inop ., recline, ., recline mechanism, mechanism, mechanism inop, seatback, seat inop, button	inop seat, inop ,, seat inop, inop ., inop (, inop, "inop ", inop pax, mechanism inop, pax seat	**Seat Recline Mechanism Inoperability**	seat 2h recline mechanism is inop., 14h pax recline seat knob is inop., 33k seats recline mechanism is inop.
4	armrest, armcap, rh, lh,), (, bad, condition, bad condition, broken	rh armrest, rh armcap, armrest 2a, armrest), armcap assy, condition 3h, armcap, armrest 31, damaged armrest, armrest	**Seat Armrest and Armcap Damage**	2h rh armrest is in bad condition, armcap on seats 12h(lh) and 12c(rh) are in bad condition, 3h rh armrest is in bad condition
5	strap, strap seat, holder, torn, 1a, kc, holder seat, broken, 2h, 1h	seat broken, broken seat, damage seat, seat torn, seat 3k, seat bad, 2k seat, seat 1k, seat 2h, 1k seat	**KCTV Strap Damage**	kc tv strap on seat 3k is broken., kctv strap on seat 1a is damaged (torn) before seat 2a, kctv strap on seat 2h torn temporary repair is performed
6	inop, table inop, inop ., table, ., . table, h table, glass, pax, table glass	table inop, inop table, pax table, table seat, . table, seat table, h	**Passenger Seat Table Inoperability**	table on 36k is inop., 45c pax seat table is inop., 31 h table is inop.

		table, table, inop pax, small table		

Количественное распределение дефектов по семи оптимизированным категориям и шумовому кластеру (таблица 1) представлено на рисунке 4. В структуре неисправностей доминируют категории *«Seatback Upright Positioning Failure»* (Тема 0 – 51,9 %) и *«Seat Food Tray Damage»* (Тема 1 – 18,6 %). Объем шумового кластера составил 230 текстов – 11,7 %.

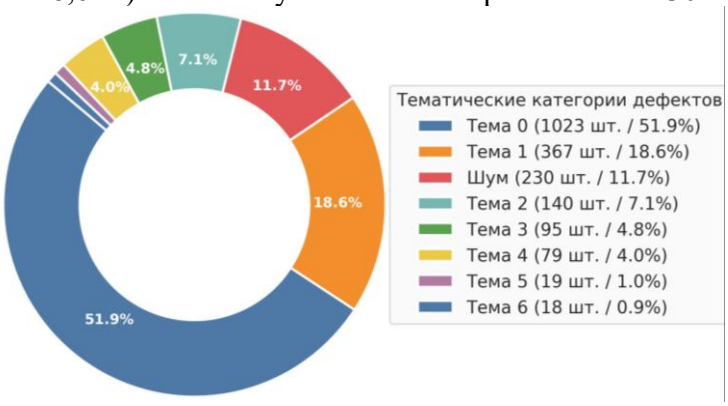

Рисунок 4 – Распределение объема зарегистрированных дефектов по оптимизированным тематическим категориям

Двумерная семантическая проекция документов на рисунке 5 наглядно иллюстрирует высокую степень разделения выделенных категорий на изолированные кластеры.

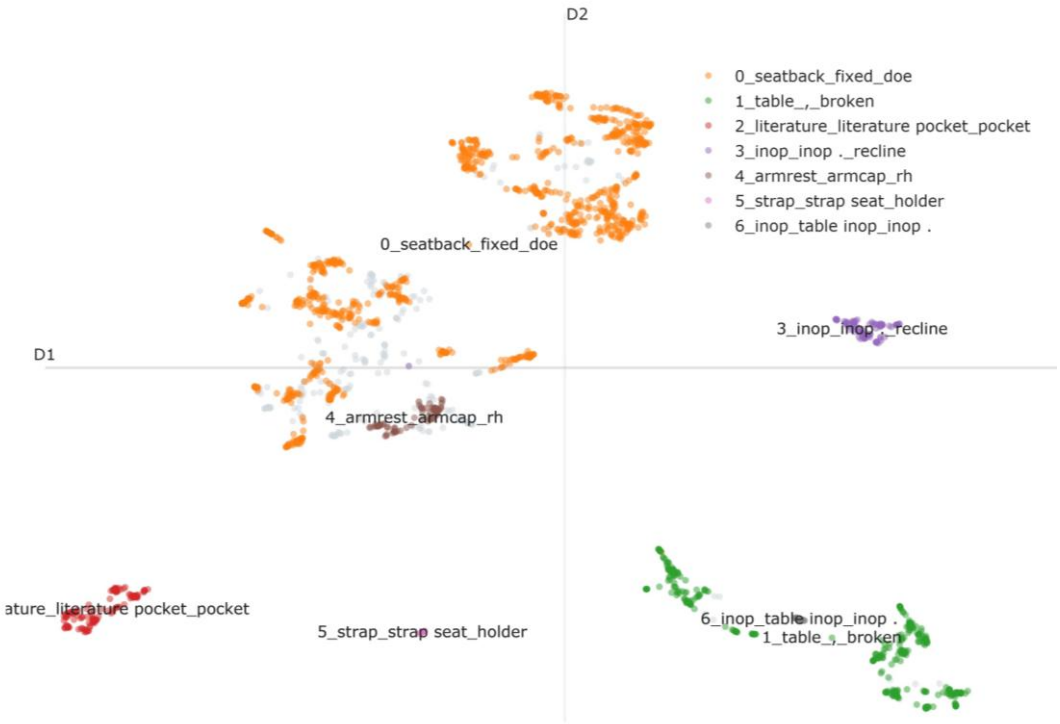

Рисунок 5 – Двумерная проекция корпуса документов и распределение тематических кластеров на плоскости

Пространственная близость Темы 1 и Темы 6 обусловлена их глубоким семантическим сходством (обе описывают дефекты пассажирских столиков). Напротив, содержательно независимые категории (например, Тема 2 – карманы, Тема 3 – механизмы спинок) максимально удалены друг от друга. Серый цвет представляет успешно изолированные алгоритмом нетипичные записи (шумовой кластер).

В таблице 2 представлены результаты, полученные при проведении тематического моделирования на базе LDA для тех же данных категории «25-21».

Сопоставление качественных результатов тематического моделирования, представленных в таблицах 1 и 2, демонстрирует существенные различия в разрешающей способности и семантической чистоте двух методов.

Разработанный метод (таблица 1) обеспечивает высокую покомпонентную чистоту и гранулярность. Полученные темы хорошо соответствуют изолированным конструктивным узлам пассажирского кресла. Метод не только успешно разделил дефекты спинок кресел (тема 0), откидных столиков (тема 1) и карманов (тема 2), но и выделил специфические малочисленные категории (рисунок 4), такие как неисправности механизмов реклинирования (тема 3), подлокотников (тема 4) и ремешков KCTV (тема 5, всего 19 записей). Важным преимуществом является изоляция нетипичных отчетов в шумовой кластер (-1), что позволило сохранить семантическую однородность остальных категорий.

Таблица 2 – Тематическая структура дефектов категории «25-21», полученная при помощи LDA [2]

№	Кол-во	Темы LDA (сформулированы экспертом)	Ключевые слова LDA
1	582	Pocket and belt defects	seat, pocket, back, literature, cover, spring, belt
2	584	Seat mechanism malfunctions	fix, seatback, position, upright, properly, recline, seat
3	366	Damage to folding tables	table, seat, crack, pax, food, open, tray
4	275	Damage to upholstery and seat covers	seat, cover, backrest, bottom, torn, button, strap
5	164	Armrest damage and minor defects	seat, armrest, passenger, damage, armcap, small, pen

Базовая модель (baseline) на основе LDA (таблица 2) предлагает более обобщенное представление данных, однако характеризуется выраженным семантическим перекрытием признаков. Так, тема 1 объединяет дефекты карманов и ремней безопасности (pocket и belt), тема 2 смешивает неисправности фиксации спинки и механизмов реклинирования, а тема 5 объединяет подлокотники с неспецифическими мелкими дефектами. Из-за отсутствия встроенного механизма фильтрации шума LDA вынужден распределять уникальные аномалии по основным классам, что снижает их интерпретируемость и приводит к потере редких категорий отказов (например, ремешков KCTV).

Кроме того, предложенный метод демонстрирует превосходство в интерпретируемости результатов. Интеграция c-TF-IDF, KeyBERT Inspired и генеративной LLM позволила полностью автоматически сформировать точные, человекочитаемые названия тем. В то же время интерпретация тем LDA требовала ручного экспертного анализа и сложной предобработки текстов.

Для количественной оценки качества разработанного метода и проверки статистической значимости полученных им результатов были выбраны три метрики, описанные ранее.

Для доказательства содержательности выделенной тематической структуры было проведено сравнение полученных метрик со случайным базовым уровнем (baseline), рассчитанным на основе 30 независимых испытаний со случайной генерацией тем из словаря корпуса. Результаты сравнения представлены в таблице 3.

Таблица 3 – Результаты статистической валидации разработанного метода (n = 30 испытаний)

Метрика	Предложенный метод	Baseline		Разница, Δ
		Среднее значение. $\bar{x}$	Среднеквадратическое отклонение, s	
C_{NPMI}	0,1133	-0,4345	0,0101	0,5478
U_{Mass}	-4,7235	-20,328	0,1346	15,5046

Topic Diversity	0,8286	0,9976	0,0053	-0,169

Представленные в таблице 3 результаты позволяют дать объективную оценку качества разработанного метода.

По метрике U_{Mass} разработанный метод показал результат -4,7235 против -20,328 у случайного уровня. Разница составляет 15,5 единицы U_{Mass} и значительно превышает вариативность результатов базового уровня.

Аналогично, по метрике C_{NPMI} разработанный метод превосходит базовый уровень на 54 стандартных отклонения**,** демонстрируя положительное значение 0,1133 (при baseline -0,4345). Полученные результаты доказывают наличие выраженной внутренней физической и лингвистической логики в сформированных темах. Вероятность случайного формирования подобной структуры крайне мала.

Показатель разнообразия тем разработанного метода составляет 0,8286. Результат, находящийся у верхней границы рекомендуемого диапазона [0,60; 0,85], свидетельствует о том, что темы являются уникальными, практически не дублируют друг друга и не содержат избыточных пересекающихся ключевых слов.

При этом случайный базовый уровень ожидаемо демонстрирует показатель, близкий к единице (0,9976), поскольку при случайной выборке слов из обширного словаря вероятность совпадения терминов между темами стремится к нулю. Отрицательная разница (Δ = -0,169) отражает естественный процесс концентрации специфической доменной лексики вокруг реальных физических дефектов.

Заключение.

В настоящем исследовании был разработан, теоретически обоснован и экспериментально верифицирован комплексный метод автоматического выявления и детальной интерпретации скрытых подгрупп неисправностей в авиационном ТО на основе контекстных эмбеддингов и плотностной кластеризации.

В исследовании был сформирован оптимальный стек инструментов (Sentence-Transformers, UMAP, HDBSCAN, c-TF-IDF) в виде единого пайплайна, способного выявлять скрытые подтемы на основе «сырых» текстов без применения сложной ручной предобработки. В отличие от baseline-модели на основе LDA, предложенный метод не требует глубокой очистки текста, лемматизации, расшифровки аббревиатур, фильтрации и удаления стоп-слов, сохраняя синтаксический контекст. Интеграция локальной LLM позволила полностью автоматизировать процесс генерации точных легко интерпретируемых названий тем.

В работе была выполнена экспериментальная апробация метода на данных категории «25-21» (Пассажирское оборудование / Салон ВС), которая доказала его превосходство над LDA. Качественный анализ подтвердил высокую покомпонентную чистоту метода. Темы строго соответствуют физическим узлам оборудования (спинки кресел, столики, подлокотники и т.д.), тогда как LDA демонстрирует семантическое размытие. Статистическая значимость результатов подтверждена количественно: метрика когерентности. U_{Mass} превысила случайный baseline на 15,5 единиц, а C_{NPMI} на 54 стандартных отклонения. Показатель разнообразия тем составил высокий показатель 0,8286 при успешной изоляции шума (11,7 % выборки).

Ограничения метода связаны с объемами данных для редких классов, вычислительными требованиями локальной LLM, а также прямой зависимостью качества генерации названий тем от параметрической емкости используемой языковой модели.

Направления будущих исследований включают масштабирование метода на другие разделы ATA 100 и автоматическое прогнозирование ремонтных воздействий (Action) по описаниям дефектов (Defect). Использование мультиязычной модели эмбеддингов

формирует технологический задел для масштабирования метода на анализ многоязычных баз данных ТОиР без необходимости предварительного перевода исходных текстов.

Список литературы

1. Sathyananda Swamy, H. V., Manoj, B. N., Zaiba, N. & Pandey, M. (2024). A Study of Artificial Intelligence in Aviation Management. QTanalytics Publication (Books), pp. 108–114. https://doi.org/10.48001/978-81-966500-8-7-11.
2. Sundaram, S. & Zeid, A. (2025). Technical language processing for Prognostics and Health Management: applying text similarity and topic modeling to maintenance work orders. J. Intell. Manuf., 2025, vol. 36, pp. 1637–1657. https://doi.org/10.1007/s10845-024-02323-4.
3. Sarica, S., & Luo, J. (2020). Stopwords in technical language processing. arXiv. https://doi.org/10.48550/arXiv.2006.02633.
4. Aziida Nanyonga, Keith Joiner, Ugur Turhan and Graham Wild (2025). Applications of Natural Language Processing in Aviation Safety: A Review and Qualitative Analysis. AIAA 2025-2153 Session: AI/ML and Autonomy Software Engineering Practices. https://doi.org/10.2514/6.2025-2153.
5. Liya Wang, Jason Chou, David Rouck, Alex Tien, Diane M. (2023). Baumgartner Adapting Sentence Transformers for the Aviation Domain. https://doi.org/10.48550/arXiv.2305.09556.
6. Savostin, A., Kaipbek, G., Koshekov, K. & Savostina, G. K. (2025). Wardle Comprehensive analysis of aviation maintenance text reports using natural language processing methods. Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu, vol. (6): pp. 157 – 167. https://doi.org/10.33271/nvngu/2025-6/157.
7. Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan & Wei Wang (2022). Language-agnostic BERT Sentence Embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
8. Wang, W., Wei, F., Dong, L., Bao, H., Yang,N. & Zhou, M.(2020). "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre- N. Trained Transformers," Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 5776–5788.
9. McInnes, L., Healy, J. & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv. https://doi.org/10.48550/arXiv.1802.03426.
10. Campello, R. J. G. B., Moulavi, D., Zimek, A. & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Transactions on Knowledge Discovery from Data, 10(1), 1–51. https://doi.org/10.1145/2733381.
11. Maarten Grootendorst. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv. https://doi.org/10.48550/arXiv.2203.05794.
12. Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT [Computer software]. GitHub. https://github.com/MaartenGr/KeyBERT.
13. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B. et al. (2025). Qwen3 technical report. arXiv. https://doi.org/10.48550/arXiv.2505.09388.
14. Röder, M., Both, A. & Hinneburg, A. (2015). Exploring the space of topic coherence measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15), 399–408. https://doi.org/10.1145/2684822.2685324
15. Mimno, D., Wallach, H. M., Talley, E., Leenders, M. & McCallum, A. (2011). Optimizing semantic coherence in topic models. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), 262–272. Association for Computational Linguistics. https://aclanthology.org/D11-1024/.
16. Adji B. Dieng, Francisco J. R. Ruiz & David M. Blei (2020). Topic Modeling in Embedding Spaces. Transactions of the Association for Computational Linguistics, 8 439–453. doi: https://doi.org/10.1162/tacl_a_00325.

АВИАЦИЯЛЫҚ ТЕХНИКАЛЫҚ ҚЫЗМЕТ КӨРСЕТУДІҢ ҚҰРЫЛЫМДАЛМАҒАН ДЕРЕКТЕРІН ТЫҒЫЗДЫҚТЫ КЛАСТЕРЛЕУ ЖӘНЕ ҮЛКЕН ТІЛДІК МОДЕЛЬДЕР НЕГІЗІНДЕ ЗИЯТКЕРЛІК ТАЛДАУ

Аңдатпа. *Ұшу қауіпсіздігін қамтамасыз ету және әуе кемелеріне техникалық қызмет көрсетудің экономикалық тиімділігін арттыру құрылымдалмаған мәтіндік есептерді зияткерлік талдауды талап етеді. Тақырыптық талдаудың дәстүрлі әдістері қысқа хабарламалардың семантикалық контекстінің жоғалуымен және мәтіндерді қолмен алдын ала өңдеудің жоғары еңбек сыйымдылығымен байланысты шектеулерге ие. Осы зерттеудің пәні техникалық рапорттардан білімді автоматты түрде алу әдістері болып табылады. Қойылған мақсатқа жету үшін оңтайлы есептеу конвейерін қалыптастыру, оны бағдарламалық іске асыру және нақты деректер негізінде эксперименттік тексеру міндеттері шешілді.*

Зерттеудің мақсаты – ақаулардың жасырын ішкі топтарын автоматты түрде анықтау мен оларды егжей-тегжейлі интерпретациялаудың кешенді әдісін әзірлеу, теориялық тұрғыдан негіздеу және эксперименттік түрде верификациялау.

Әзірленген әдіс контекстік эмбеддингтерді алу үшін трансформер-модельді (paraphrase-multilingual-MiniLM-L12-v2), өлшемділікті сызықтық емес төмендетудің UMAP алгоритмін және HDBSCAN иерархиялық тығыздықтық кластерлеу алгоритмін біріктіруге негізделген. Тақырыптарды интерпретациялау және олардың адамға түсінікті атауларын автоматты түрде генерациялау c-TF-IDF алгоритмі, KeyBERTInspired семантикалық іріктеу әдісі және жергілікті түрде орналастырылған Qwen3.5-4B-Instruct үлкен тілдік моделі негізінде жүзеге асырылды. Әдістің эксперименттік тексеруі тоғыз әуе кемесінің 7 жылдық пайдалану кезеңіндегі жолаушылар жабдығы санатына жататын 1971 мәтіндік жазбадан тұратын іріктемеде жүргізілді. Әзірленген әдіс ақаулардың мазмұндық тұрғыдан тұрақты 7 санатын сәтті сәйкестендіріп, типтік емес жазбалардың 11,7 %-ын шу кластеріне оқшаулады. Кездейсоқ базалық деңгеймен салыстыру нәтижелердің жоғары статистикалық маңыздылығын растады. Тақырыптардың әртүрлілік көрсеткіші 0,8286 мәнін құрады және ұсынылатын диапазонға сәйкес келеді.

Әзірленген әдіс бөлінген тақырыптардың компоненттік тазалығы бойынша классикалық LDA моделінен асып түседі және деректерді күрделі қолмен алдын ала өңдеу кезеңін жояды. Бұл шешімді ATA 100 стандарттарының басқа бөлімдеріне масштабтау және қызмет көрсетуді жоспарлауды әрі құрамдас бөліктер қорын басқаруды оңтайландыру мақсатында авиациялық кәсіпорындардың шешім қабылдауды қолдаудың зияткерлік жүйелеріне интеграциялау ұсынылады.

Түйін сөздер: *техникалық қызмет көрсету, авиациялық көлік, мәтіндерді зияткерлік талдау, контекстік эмбеддингтер, тығыздықты кластерлеу, тақырыптық модельдеу, үлкен тілдік модельдер.*

INTELLIGENT ANALYSIS OF UNSTRUCTURED AVIATION MAINTENANCE DATA BASED ON DENSITY-BASED CLUSTERING AND LARGE LANGUAGE MODELS

Abstract. *Ensuring flight safety and improving the economic efficiency of aircraft maintenance require intelligent analysis of unstructured textual reports. Traditional topic modelling methods are limited by the loss of semantic context in short messages and the high labour intensity of manual text preprocessing. The subject of this study is methods for the automated extraction of knowledge from technical reports. To achieve this objective, the tasks of designing an optimal computational pipeline, implementing it in software, and conducting experimental validation using real-world data were addressed.*

The aim of the study is to develop, theoretically substantiate, and experimentally verify a comprehensive method for the automatic identification and detailed interpretation of latent fault subgroups.

The proposed method is based on the integration of a transformer model (paraphrase-multilingual-MiniLM-L12-v2) for generating contextual embeddings, the UMAP nonlinear dimensionality reduction algorithm, and HDBSCAN hierarchical density-based clustering. Topic interpretation and the automatic generation of human-readable topic labels are implemented using the c-TF-IDF algorithm, KeyBERTInspired semantic selection, and a locally deployed large language model, Qwen3.5-4B-Instruct. Experimental validation of the method was performed on a dataset comprising 1,971 textual records related to passenger equipment collected over a seven-year operational period from nine aircraft. The proposed method successfully identified seven stable and meaningful defect categories while isolating 11.7% of atypical records into a noise cluster. Comparison with a random baseline confirmed the high statistical significance of the results. The topic diversity score reached 0.8286, which falls within the recommended range.

The proposed method outperforms the classical LDA model in terms of the component-wise purity of the extracted topics and eliminates the need for complex manual data preprocessing. The solution is recommended for scaling to other sections of the ATA 100 standard and for integration into intelligent decision-support systems of aviation enterprises to optimise maintenance planning and component inventory management.

Keywords: *maintenance, air transport, text mining, contextual embeddings, density-based clustering, topic modeling, large language models.*

Сведение об авторах

Каипбек Гульсанат Мэлскызы	Докторант, Академия Гражданской Авиации, Алматы, КазахстанE-mail: kaipbekgulsanat@gmail.com
Савостин Алексей Александрович	Кандидат технических наук, ассоциированный профессор, профессор кафедры «Энергетика и радиоэлектроника» Северо-Казахстанского университета им. М. Козыбаева, Петропавловск, Казахстан. E-mail: asavostin@ku.edu.kz
Кошеков Кайрат Темирбаевич	Доктор технических наук, профессор, проректор по научной деятельности, Академия Гражданской Авиации, Алматы, Казахстан, E-mail: kkoshekov@mail.ru
Риттер Дмитрий Викторович	Кандидат технических наук, ассоциированный профессор, профессор кафедры «Энергетика и радиоэлектроника» Северо-Казахстанского университета им. М. Козыбаева, Петропавловск, Казахстан. E-mail: dritter@ku.edu.kz

Авторлар туралы мәлімет

Каипбек Гүлсанат Мэлсқызы	Азаматтық авиация академиясының докторанты, Алматы, ҚазақстанE-mail: kaipbekgulsanat@gmail.com
Савостин Алексей Александрович	Техника ғылымдарының кандидаты, қауымдастырылған профессор, Солтүстік Қазақстан университетінің "Энергетика және радиоэлектроника" кафедрасының профессоры. М. Қозыбаева, Петропавл, Қазақстан. E-mail: asavostin@ku.edu.kz
Көшеков Қайрат Темірбаевич	Техника ғылымдары докторы, профессор, Азаматтық авиация академиясының ғылыми жұмыстар жөніндегі проректоры, Алматы, Қазақстан. E-mail: kkoshekov@mail.ru
Риттер Дмитрий Викоторвич	Техника ғылымдарының кандидаты, қауымдастырылған профессор, Солтүстік Қазақстан университетінің "Энергетика және радиоэлектроника" кафедрасының профессоры. М. Қозыбаева, Петропавл, Қазақстан. E-mail: dritter@ku.edu.kz

Information about the authors

Gulsanat Kaipbek	Doctoral researcher, Civil Aviation Academy, Almaty, Kazakhstan, E-mail: kaipbekgulsanat@gmail.com
Alexey Savostin	Candidate of Technical Sciences, Associate Professor, Professor of the Department of "Power Engineering and Radio Electronics" of the M. Kozybayev North Kazakhstan University, Petropavlovsk, Kazakhstan, E-mail: asavostin@ku.edu.kz

Kayrat Koshekov	Doctor of Technical Sciences, Professor, Vice-Rector for Scientific Activities, Civil Aviation Academy, Almaty, Kazakhstan E-mail: kkoshekov@mail.ru
Ritter Dmitriy	Candidate of Technical Sciences, Associate Professor, Professor of the Department of "Power Engineering and Radio Electronics" of the M. Kozybayev North Kazakhstan University, Petropavlovsk, Kazakhstan, E-mail: asavostin@ku.edu.kz