



УДК 681.518.5:629.735

МРНТИ 28.17.19, 81.83.20

https://doi.org/10.53364/24138614_2026_40_1_3

Г. Каипбек¹, А. Савостин², К. Кошеков¹, Г. Савостина²

¹ Академия гражданской авиации, Алматы, Казахстан

² Северо-Казахстанский университет им. М. Козыбаева,
Петропавловск, Казахстан

*E-mail: kaipbekgulsanat@gmail.com

ИДЕНТИФИКАЦИЯ СКРЫТЫХ ТИПОВ ОТКАЗОВ СИСТЕМЫ ВНЕШНЕГО ОСВЕЩЕНИЯ ВОЗДУШНОГО СУДНА НА ОСНОВЕ ЛАТЕНТНОГО РАЗМЕЩЕНИЯ ДИРИХЛЕ

Аннотация. Непрерывный рост объемов неструктурированных текстовых данных в системах технического обслуживания авиационной техники создает потребность в автоматизированных методах их анализа. Традиционная категоризация дефектов по стандартным кодам часто оказывается недостаточно детальной для понимания истинных причин отказов, смешивая рутинные операции и критические сбои в рамках одной системы. Предметом данного исследования является семантическая структура текстовых описаний неисправностей системы внешнего освещения (АТА 33-40) парка однотипных воздушных судов. Целью публикации является разработка и апробация метода автоматического выявления скрытых эксплуатационных паттернов и режимов отказа без использования размеченных данных. Методологическую основу исследования составляет вероятностное тематическое моделирование с использованием алгоритма латентного размещения Дирихле (LDA). Для повышения качества модели реализован специализированный алгоритм предобработки текста, включающий расшифровку отраслевых аббревиатур и удаление контекстного шума. Оптимальная конфигурация модели определялась на основе количественного анализа метрики когерентности (C_v) и оценки семантической устойчивости тем. Экспериментально установлено, что модель, содержащая шесть тем, обеспечивает наилучшую интерпретируемость данных. Анализ полученных кластеров позволил выявить конструктивно-обусловленные зоны возникновения дефектов, а также классифицировать неисправности по типу проявления. Автоматически выделены скрытые подгруппы, соответствующие отказам электрических цепей и механическим повреждениям элементов конструкции. Предложенный подход позволяет трансформировать неструктурированные записи техперсонала в детализированную диагностическую информацию. Это открывает возможности для совершенствования программ технического обслуживания и перехода к предиктивному управлению надежностью конкретных подсистем воздушного судна.

Ключевые слова: воздушное судно, техническое обслуживание, внешнее освещение, текстовые описания, обработка естественного языка, тематическое моделирование, латентное размещение Дирихле.

Введение.

В современной гражданской авиации процессы поддержания летной годности и технического обслуживания (ТО) характеризуются генерацией больших объемов сопутствующих данных. Несмотря на повсеместное внедрение цифровых систем управления ТО (MRO-систем), значительная часть важной диагностической информации по-прежнему фиксируется инженерно-техническим персоналом в виде неструктурированных текстовых описаний – записей в бортовых технических журналах и картах нарядов. Эти записи содержат уникальные сведения о характере проявления неисправностей, симптомах отказов и предпринятых корректирующих действиях, которые часто теряются при агрегации и нормализации данных в стандартные коды АТА 100 (АТА iSpec 2200) [1].

При этом важной задачей для анализа являются категории неисправностей с высокой частотой возникновения и разнородной физикой отказов. В частности, для воздушных судов (ВС) с высокой интенсивностью эксплуатации категория по АТА 100 (АТА iSpec 2200) Chapter 33, Section 40 – External Lights является одной из лидирующих по количеству записей в бортовых журналах ТО [2]. Однако стандартная кодировка не позволяет автоматически различать плановые работы, механические повреждения элементов конструкции или критические отказы электрических цепей. Отсутствие детализированного понимания структуры дефектов внутри одной АТА-категории ограничивает возможности предиктивной аналитики, планирования складских запасов и выявления скрытых угроз безопасности. В связи с этим, разработка методов автоматического извлечения скрытых паттернов отказов из текстовых массивов авиационного ТО является актуальной научно-технической задачей.

Традиционные подходы к анализу текстовых данных в авиации долгое время опирались на ручную категоризацию или поиск, по ключевым словам, что неэффективно в условиях постоянно растущих объемов информации. Современные исследования в области обработки естественного языка (NLP) демонстрируют смещение фокуса в сторону методов машинного обучения [3], [4]. Большинство существующих работ сосредоточено на задачах обучения с учителем в форме автоматического присвоения кодов АТА на основе текста [5], [6], [7]. Такой подход воспроизводит существующую структуру данных, не позволяя выполнить анализ категории и выявить новые знания.

В то же время, методы обучения без учителя, в частности тематическое моделирование, остаются недостаточно изученными применительно к узкоспециализированным авиационным задачам [2]. Существует дефицит исследований, демонстрирующих, как вероятностные модели, такие как латентное размещение Дирихле (LDA), могут быть использованы не просто для группировки текстов, но и для семантической декомпозиции сложных инженерных систем и выявления режимов отказа, специфичных для конкретного типа ВС.

В связи с этим, целью данного исследования является разработка и апробация подхода к выявлению и интерпретации скрытых профилей неисправностей внутри категории АТА (АТА iSpec 2200) 33-40 («Внешнее освещение») на примере парка однотипных воздушных судов (Embraer 190) с использованием метода LDA.

Для достижения поставленной цели в работе решаются следующие задачи.

1. Реализация алгоритма предварительной обработки корпуса реальных записей ТО, направленной на снижение шума и унификацию технического сленга, характерного для обслуживания авиационных систем.
2. Построение тематической модели LDA, обеспечивающей оптимальный баланс между статистической устойчивостью и инженерной интерпретируемостью кластеров.
3. Проведение семантического анализа полученных тем для идентификации ключевых подгрупп дефектов.

Материалы и методы исследования.

Анализ текстового корпуса авиационного ТО.

В исследовании используется текстовый корпус, содержащий данные о зарегистрированных неисправностях в процессе ТО парка из девяти однотипных коммерческих воздушных судов, эксплуатируемых на территории Республики Казахстан с 2014 по 2020 годы. В корпусе содержится 13204 записи из бортовых журналов и систем управления ТО, представленные без предварительной лингвистической обработки, в том числе с сохранением профессионального сленга, аббревиатур, профессиональной и технической терминологии.

Благодаря имеющейся разметке по категориям неисправностей ATA 100 (ATA iSpec 2200), из представленного корпуса были отобраны все записи с меткой Chapter 33, Section 40 – External Lights («Внешнее освещение») в количестве 669 образцов. В представленных записях данная категория отказов является второй по частоте возникновения (после 25-20 – «Passenger Cabin») и представляет особый интерес для анализа по следующим причинам:

- высокая частота возникновения дефектов;
- критическое влияние на ночные операции ВС;
- внутренняя гетерогенность отказов, требующая семантического анализа для разделения рутинных работ (замена ламп) и системных сбоев.

Примеры записей по категории 33-40 приведены в таблице 1. Для соблюдения конфиденциальности реальные регистрационные номера ВС были заменены на условные обозначения (А, В, С и т.д.).

Таблица 1 –Примеры необработанных записей ТО для категории отказов 33-40

№	Date Reported	A/C Reg	Defect
102	2015-08-24	A	RH WING INSP LIGHT IS INOP
639	2016-01-26	B	RH TAXI SIDE LIGHT IS INOP
474	2014-04-30	A	LH WING-ROOT TAXI LIGHT AND LANDING LIGHTARE INOP
6	2016-02-25	C	LH NAV LIGHT (STBY SYS) IS INOP
39	2014-07-26	D	NOSE TAXI LIGHT LAMP FOUND CRACKED

Как следует из таблицы 1, текстовые описания дефектов имеют вид коротких сообщений в техническом домене авиационной отрасли. Наблюдается обилие отраслевых сокращений, аббревиатур и пр. По этой причине данные требуют предварительной обработки перед процессом тематического моделирования.

Предварительная обработка текстовых записей.

Для обеспечения качества тематического моделирования был реализован многоэтапный процесс очистки и нормализации исходных записей с использованием языка Python 3 и библиотеки NLTK. Схема данного процесса показана на рисунке 1.



Рисунок 1 – Схема процесса предварительной обработки текстовых данных

В соответствии со структурной схемой рисунка 1, процедура предобработки включает следующие основные шаги.

На первом этапе производится расшифровка сокращений. С помощью разработанного скрипта на языке Python был реализован многоступенчатый алгоритм по извлечению из текстовых описаний дефектов специфичных для авиационного домена аббревиатур и сокращений (рисунок 1). На основе частотного анализа и экспертной валидации был сформирован словарь замен (например, «INOP» → «inoperative», «LH» → «left hand» и др.). На основании полученного словаря проведена расшифровка аббревиатур, за исключением самодостаточных доменных терминов (например, «EICAS»), что позволило унифицировать технический сленг и устранить вариативность написания без избыточного усложнения текста.

На втором этапе выполняется очистка от шума. Записи переводятся в нижний регистр. Из текстов удаляются знаки препинания, за исключением знаков «/» и «-», являющихся частью составных терминов. С помощью регулярных выражений производится фильтрация технических кодов, не несущих семантической нагрузки для LDA (серийные номера деталей, коды инструкций АММ и пр.).

Третий этап реализует лингвистическую обработку. Выполняется токенизация, а для снижения размерности данных применяется лемматизация (приведение к словарной форме).

На заключительном этапе происходит фильтрация стоп-слов. Стандартный список стоп-слов английского языка (из python-модуля NLTK) был расширен за счет частотной общетехнической лексики («hand», «inoperative», «aft», «forward», «left», «right», «rev», «accordance», «ref», «condition», «bad», «fail», «message», «doesnt», «side», «broken») наличие которой может зашумлять результаты тематического моделирования, скрывая специфические признаки неисправностей.

В результате после предварительной обработки текстовые описания дефектов представляются в виде очищенного списка отдельных токенов (рисунок 1).

Построение тематической модели LDA.

LDA представляет собой вероятностную порождающую модель, основанную на байесовском подходе [8]. LDA рассматривает каждый документ d из корпуса D как смесь латентных тем, а каждую тему k из общего набора K тем – как распределение вероятностей на словах из словаря V .

Модель предполагает следующий генеративный процесс для каждого документа из корпуса D . Сначала для каждого документа d генерируется распределение тем θ_d из априорного распределения Дирихле с гиперпараметром α :

$$\theta_d \square \text{Dir}(\alpha). \quad (1)$$

Затем для каждого слова n в документе d выполняются два шага. Сначала генерируется тема $z_{dn} \in \{1, \dots, K\}$ из категориального распределения параметризованного вектором θ_d :

$$z_{dn} = \text{Cat}(\theta_d). \quad (2)$$

Далее выбирается слово w_{dn} из распределения слов, соответствующего теме z_{dn} . Каждая тема $k \in \{1, \dots, K\}$ имеет свое собственное распределение слов ϕ_k , которое также генерируется из априорного распределения Дирихле с гиперпараметром β :

$$\phi_k \square \text{Dir}(\beta). \quad (3)$$

Таким образом, итоговое слово w_{dn} генерируется из распределения, определенного выбранной темой z_{dn} :

$$w_{dn} = \text{Cat}(\phi_{z_{dn}}). \quad (4)$$

Задача тематического моделирования сводится к вычислению апостериорного распределения $p(Z, \Theta, \Phi | V, \alpha, \beta)$, где V – наблюдаемые слова, а Z, Θ, Φ – скрытые переменные (темы слов, распределения тем в документах и распределения слов в темах соответственно). Поскольку точное вычисление этого распределения является неразрешимой задачей из-за высокой размерности и сложной зависимости переменных, в данной работе для его аппроксимации использовался итеративный алгоритм вариационного байесовского вывода (Variational Bayes), реализованный в python-библиотеке scikit-learn 1.7.0 [9].

При создании модели LDA для преобразования токенизированного текста, полученного на этапе предварительной обработки (рисунок 1) в машинно-читаемый формат, использовался метод Bag-of-Words. Суть этого метода заключается в преобразовании корпуса D в разреженную матрицу частот документ-термин X , где элемент x_{ij} соответствует количеству вхождений слова (термина) j в документ i .

Ключевым этапом являлось определение оптимального количества тем K_{opt} , которое заранее неизвестно. Для этого был проведен эксперимент по подбору гиперпараметра, который определяет число тем K в заданном диапазоне значений $\{K_{min}, \dots, K_{max}\}$.

Для оценки качества полученных моделей использовалась метрика когерентности C_v . Данная метрика оценивает степень семантической согласованности слов внутри темы, имитируя человеческое восприятие интерпретируемости [10].

Чтобы исключить зависимость оценки от произвольного выбора глубины анализа (топ-N слов), был применен метод оценки семантической устойчивости. Для каждой обученной модели при фиксированном K расчет C_v производился многократно для изменяющегося количества наиболее вероятных слов $W = \{2, \dots, 10\}$. Для каждого числа K рассчитывалось среднее значение когерентности C_v по всему диапазону W и доверительный интервал.

Результаты эксперимента по определению оптимального числа тем в виде усредненных по всем значениям топ-N слов W для каждого K с 95% доверительным интервалом, представлены на рисунке 2.

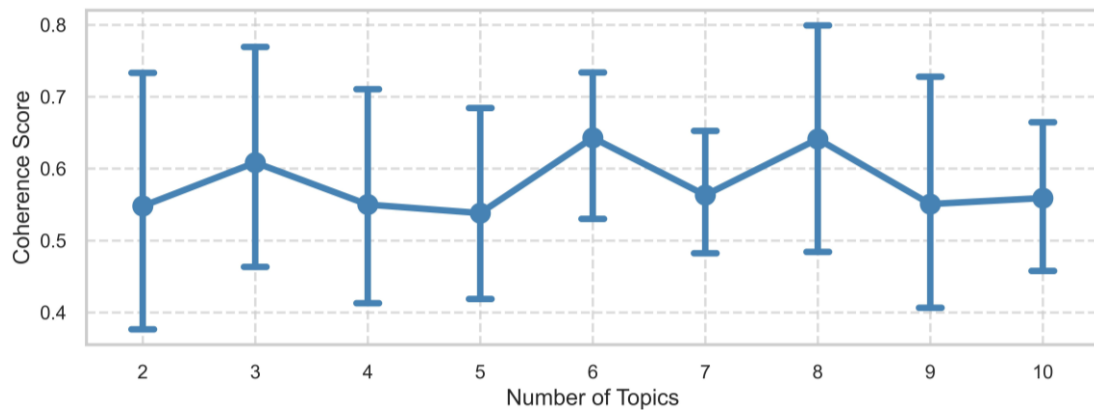


Рисунок 2 – Зависимость средней когерентности C_v от количества тем для категории отказов 33-40 – External Lights

Анализ зависимости метрики когерентности от количества тем выявляет локальные максимумы при $K = 3$, $K = 6$ и $K = 8$ (рисунка 2). Несмотря на то, что модель с 8 темами достигает высокого среднего значения, она демонстрирует неустойчивость семантического ядра, подтверждаемую широким доверительным интервалом. В противовес этому, конфигурация с $K = 6$ показывает сопоставимо высокую среднюю когерентность ($C_v \approx 0,64$) при существенно меньшем разбросе значений, что свидетельствует о формировании устойчивых и интерпретируемых кластеров. На основании этого для финального анализа было выбрано 6 тем, что обеспечивает необходимую глубину детализации отказов, избегая как чрезмерного обобщения ($K = 3$), так и семантической фрагментации ($K = 8$).

Таким образом, финальная модель LDA_{opt} была обучена на полном наборе данных с параметром $K_{opt} = 6$. Смысловая валидация полученных кластеров проводилась экспертным путем на основе анализа распределения вероятностей слов в темах, показанной на рисунке 3.

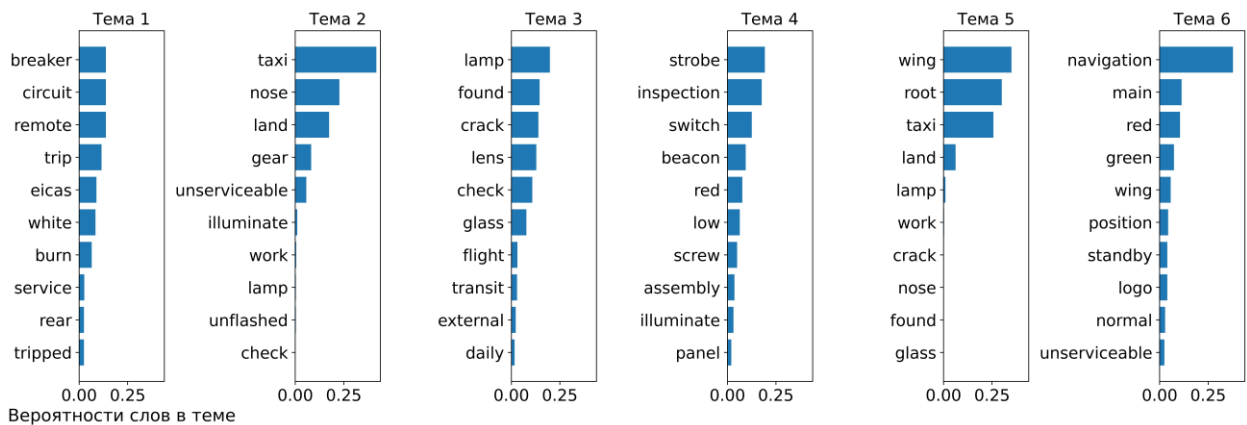


Рисунок 3 – Распределение вероятностей слов в темах по результатам LDA моделирования

На рисунке 4 показано распределение числа документов по темам в результате тематического моделирования с помощью, предложенной LDA модели.

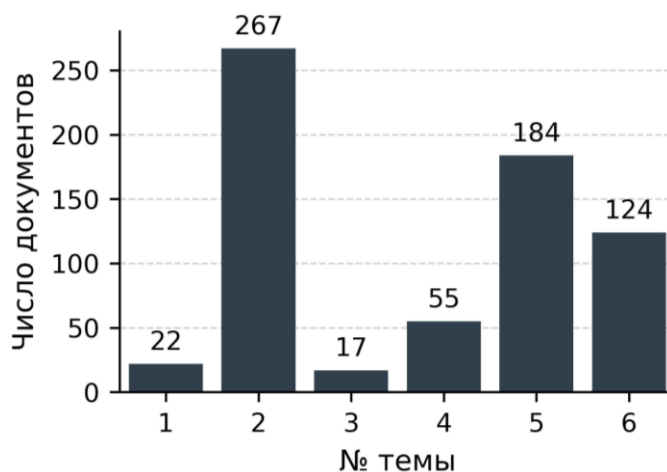


Рисунок 4 – Распределение документов по выделенным темам

Результаты и их обсуждение.

Анализ распределения ключевых слов и их весов (рисунок 3) позволил интерпретировать выделенные латентные темы как устойчивые кластеры неисправностей, классифицируемые по двум принципам: локализация оборудования и физика отказа

Характеристика тем, выявленных в ходе моделирования, представлена ниже.

Тема 1 указывает на отказы электрической защиты. Ключевые слова, такие как «breaker» (0,13), «circuit» (0,13), «remote» (0,13), «trip» (0,11), «eicas» (0,09) имеют достаточно равномерное распределение высоких весов, что указывает на семантическую плотность темы. Модель LDA смогла отделить системные электрические сбои от проблем с лампами. Данный кластер имеет высокую когерентность и описывает срабатывание автоматов защиты сети и появление соответствующих сообщений на дисплее EICAS.

Тема 2 отражает проблемы с освещением носовой стойки шасси. Ключевыми словами являются «taxi» (0,42), «nose» (0,23), «land» (0,17), «gear» (0,08). Доминирование слова «taxi» с высоким весом в связке с «nose» однозначно локализует неисправности фар руления и посадки на передней стойке шасси.

В Теме 3 объединяются симптомы физических повреждений с ключевыми словами «lamp» (0,20), «found» (0,15), «crack» (0,14), «lens» (0,13), «check» (0,11), «glass» (0,08). Данный кластер сформирован моделью не по системному, а по симптоматическому признаку. В него входят случаи обнаружения трещин линз и стекло, обусловленные, вероятнее всего, ударами посторонних предметов или термическими нагрузками.

Тема 4 описывает отказы импульсных/инспекционных огней, а также механические компоненты. К ключевым словам относятся «strobe» (0,19), «inspection» (0,18), «switch» (0,13), «beacon» (0,09), «screw» (0,08), «assembly» (0,06). Тема охватывает импульсные маяки и инспекционные фары, но, в отличие от навигационных огней, указывает на вероятные проблемы с коммутацией или креплением данных узлов (термины «switch» и «screw»).

Тема 5 локализует проблемы с освещением в корневой части крыла. Ключевые слова – это «wing» (0,35), «root» (0,30), «taxi» (0,26), «land» (0,06). В данном случае модель выделила специфичную для данного типа ВС зону установки фар в наплыве крыла («root»), которые используются для руления («taxi») и посадки.

Тема 6 включает неисправности навигационных огней. Ключевые слова здесь «navigation» (0,38), «main» (0,11), «red» (0,11), «green» (0,07), «wing» (0,06). Это четко очерченный кластер сигнальных огней, характеризующийся цветовой дифференциацией («red», «green») и упоминанием основных ламп («main»).

Распределение дефектов по выделенным темам (рисунок 4) демонстрирует значительную неравномерность, характерную для эксплуатационных данных.

Доминирующими категориями являются Тема 2 (39.9%) и Тема 5 (27.5%), которые в сумме охватывают 67.4% всех записей. Это свидетельствует о том, что основная масса работ по ТО приходится на обслуживание фар высокой интенсивности (посадочные/рулежные). Высокая доля Темы 2 коррелирует с уязвимостью расположения фар на носовой стойке к вибрациям и попаданию посторонних предметов с взлетно-посадочной полосы (ВПП).

Миноритарные кластеры, такие как Тема 1 (3,3%) и Тема 3 (2,5%), несмотря на малый объем, представляют высокую диагностическую ценность. Их выделение в отдельные группы свидетельствует о способности модели LDA фокусироваться на редких, но критических инцидентах (короткие замыкания, разрушение конструкций), не растворяя их в общей массе рутинных замен ламп.

Особого внимания заслуживает способность модели LDA к устранению неоднозначности общих терминов. Наглядным примером служит распределение термина «taxi». В отличие от традиционных методов кластеризации, LDA не сгруппировал все записи с этим словом в один класс. Вместо этого модель выявила бимодальное распределение термина. В Теме 2 он ассоциирован с контекстом «nose» (фары на передней стойке), а в Теме 5 – с контекстом «root»/«wing» (фары в корне крыла). Это подтверждает, что алгоритм корректно реконструировал архитектуру системы освещения данного типа ВС, где функция рулежного света реализуется двумя пространственно разнесенными подсистемами, имеющими разные профили повреждаемости.

Также примечательна дифференциация между сигнальными системами: Тема 6 (Навигационные огни) характеризуется преимущественно отказами источников света (ламп), в то время как Тема 4 (Стробы и Маяки) содержит маркеры отказов коммутационной аппаратуры («switch») и монтажных элементов («screw», «assembly»), что предполагает различные стратегии технического обслуживания для этих подсистем.

Таким образом, применение тематического моделирования позволяет выявить ряд скрытых эксплуатационных паттернов, недоступных при стандартном анализе кодов АТА 100 (АТА iSpec 2200). Для инженеров по надежности кластер Темы 1 может быть маркером состояния электросети ВС, который требует иных мер профилактики по сравнению с простой заменой перегоревших ламп. Аналогично, Тема 3 выделяет группу дефектов (трещины линз), которая акцентирует внимание на условиях эксплуатации (например, состояние покрытия ВПП), а не на электрических компонентах.

Заключение.

В настоящем исследовании продемонстрирована эффективность применения тематического моделирования (LDA) для семантического анализа текстовых записей ТО ВС одной модели. Ключевым преимуществом предложенного подхода является способность автоматически извлекать интерпретируемые профили неисправностей из неструктурированных данных без использования трудоемкой ручной разметки.

В рамках анализа категории 33-40 («Внешнее освещение») выявлено шесть устойчивых кластеров, отражающих конструктивные особенности данного типа ВС. Метод позволил не только локализовать дефекты (разделяя зоны носовой стойки и корня крыла), но и классифицировать их по физической природе, выделив скрытые категории критических электрических сбоев и механических повреждений, недоступные при стандартном кодировании.

К ограничениям метода следует отнести чувствительность LDA к качеству предобработки словаря и дисбалансу классов в малых выборках, что требует тщательного подбора гиперпараметров. Направления дальнейшей работы включают масштабирование подхода на другие функциональные системы ВС при накоплении достаточного объема текстовых данных, а также исследование гибридных моделей, сочетающих вероятностный подход с векторными представлениями слов, для повышения точности классификации и интеграции полученных знаний в системы предиктивного обслуживания.

Список литературы

1. Air Transport Association of America, 2021, iSpec 2200: Information Standards for Aviation Maintenance, Harvard Dataverse <https://doi.org/10.7910/DVN/G1DSMX>.
2. Savostin A., Kaipbek G., Koshekov K., Savostina G., Wardle K. Comprehensive analysis of aviation maintenance text reports using natural language processing methods. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*. 2025, v. 6, p. 157 – 167. <https://doi.org/10.33271/nvngu/2025-6/157>.
3. Aziida Nanyonga, Keith Joiner, Ugur Turhan and Graham Wild 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>.
4. Liya Wang, Jason Chou, David Rouck, Alex Tien, Diane M. Baumgartner Adapting Sentence Transformers for the Aviation Domain. 2023. <https://doi.org/10.48550/arXiv.2305.09556>.
5. Akhbardeh, F., Desell, T., Zampieri, M. MaintNet: A collaborative open-source library for predictive maintenance language resources. In: M. Ptaszynski & B. Ziolko (Eds.), Proceedings of the 28th international conference on computational linguistics: System demonstrations. International Committee on Computational Linguistics (ICCL), pp. 7-11. <https://doi.org/10.18653/v1/2020.coling-demos.2>.
6. Zhou, S., Chen, B., Zhang, Y., Liu, H., Xiao, Y., Pan, X. A Feature Extraction Method Based on Feature Fusion and its Application in the Text-Driven Failure Diagnosis Field. *Int. J. Interact. Multimed. Artif. Intell.*, 2020, vol. 6, pp. 121–130. <https://doi.org/10.9781/ijimai.2020.11.006>.
7. Xu, Z., Chen, B., Zhou, S., Chang, W., Ji, X., Wei, C., Hou, W. A Text-Driven Aircraft Fault Diagnosis Model Based on a Word2vec and Priori-Knowledge Convolutional Neural Network. *Aerospace*, 2021, vol. 8, article no. 112. <https://doi.org/10.3390/aerospace8040112>.
8. David M. Blei, Andrew Y. Ng, and Michael I. Jordan Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 2003, vol. 3, pp. 993–1022.
9. Scikit-learn developers. Scikit-learn: Machine Learning in Python, version 1.7.0. Available at: <https://scikit-learn.org/stable/> (accessed: July 23, 2025).
10. Jia Peng Lim, Hady W. Lauw; Aligning Human and Computational Coherence Evaluations. *Computational Linguistics* 2024; 50 (3): 893–952. doi: https://doi.org/10.1162/coli_a_00518.

ДИРИХЛЕНІН ЛАТЕНТТІ ОРНАЛАСТЫРУЫ НЕГІЗІНДЕ ӘУЕ КЕМЕСІНІҢ СЫРТҚЫ ЖАРЫҚТАНДЫРУ ЖҮЙЕСІНДЕГІ ЖАСЫРЫН ІСТЕН ШЫҒУ ТҮРЛЕРІН АНЫҚТАУ

Андапна. Авиациялық техниканы техникалық қызмет көрсету жүйелерінде құрылымданбаған мәтіндік деректер көлемінің тұрақты өсуі оларды талдаудың автоматтандырылған әдістеріне деген сұранысты арттырып отыр. Ақауларды стандартты кодтар бойынша дәстүрлі жіктеу көбіне істен шығулардың нақты себептерін толық ашуға жеткіліксіз, себебі ол бір жүйе аясында жоспарлы жұмыстар мен сыни ақауларды ажыратпай қарастырады. Осы зерттеудің нысаны бір типті әуе кемелері паркінің сыртқы жарықтандыру жүйесіне (АТА 33–40) қатысты ақаулардың мәтіндік сипаттамаларының семантикалық құрылымы болып табылады. Жарияланымның мақсаты алдын ала таңбаланған деректерді қолданбай, жасырын эксплуатациялық үлгілер мен істен шығу режимдерін автоматты түрде анықтау әдісін әзірлеу және сынақтан өткізу болып табылады. Зерттеудің әдіснамалық негізі Дирихленің латентті орналастыру алгоритміне (Latent Dirichlet Allocation, LDA) негізделген ықтималдық тақырыптық модельдеуге сүйенеді. Модельдің сапасын арттыру үшін

салалық аббревиатураларды тарқатуды және мағыналық жүктемесі төмен сөздерді жоюды қамтитын мәтінді алдын ала өңдеудің арнайы алгоритмі қолданылды. Модельдің оңтайлы параметрлері когеренттілік метрикасын (C_v) сандық талдау және тақырыптардың семантикалық тұрақтылығын бағалау арқылы анықталды. Эксперимент нәтижелері алты тақырыптан тұратын модельдің деректерді түсіндіру тұрғысынан ең тиімді екенін көрсетті. Алынған кластерлерді талдау ақаулардың конструкциялық ерекшеліктермен байланысты пайда болу аймақтарын анықтауға, сондай-ақ істен шығуларды көріну сипаты бойынша жіктеуге мүмкіндік берді. Электр тізбектерінің істен шығуына және конструкция элементтерінің механикалық зақымдалуына сәйкес келетін жасырын ішкі топтар автоматты түрде айқындалды. Ұсынылған тәсіл техникалық персоналдың құрылымданбаған жазбаларын егжей-тегжейлі диагностикалық ақпаратқа айналдыруға мүмкіндік береді. Бұл техникалық қызмет көрсету бағдарламаларын жетілдіруге және әуе кемесінің жекелеген қосалқы жүйелерінің сенімділігін предиктивті басқаруға көшуге негіз қалайды.

Түйін сөздер: әуе кемесі, техникалық қызмет көрсету, сыртқы жарықтандыру, мәтіндік сипаттамалар, табиғи тілді өңдеу, тақырыптық модельдеу, Дирихленің жасырын орналастыруы.

IDENTIFICATION OF LATENT FAILURE TYPES IN AN AIRCRAFT EXTERIOR LIGHTING SYSTEM BASED ON LATENT DIRICHLET ALLOCATION

Abstract. *The continuous growth of unstructured textual data volumes in aircraft maintenance systems creates a demand for automated analysis methods. Traditional defect categorization using standard codes is often insufficiently detailed to reveal the true root causes of failures, as routine operations and critical malfunctions are frequently combined within a single system category. The subject of this study is the semantic structure of textual fault descriptions related to the exterior lighting system (ATA 33–40) of an aircraft fleet of a single model. The objective of the study is to develop and validate a method for the automatic identification of latent operational patterns and failure modes without the use of labeled data. The methodological foundation of the research is probabilistic topic modeling based on the Latent Dirichlet Allocation (LDA) algorithm. To improve model quality, a specialized text preprocessing procedure was implemented, including the expansion of industry-specific abbreviations and the removal of contextual noise. The optimal model configuration was determined through quantitative analysis of the topic coherence metric (C_v) and an assessment of topic semantic stability. Experimental results show that a six-topic model provides the highest level of interpretability. Analysis of the resulting clusters made it possible to identify design-related defect occurrence zones and to classify failures according to their manifestation type. Latent subgroups corresponding to electrical circuit failures and mechanical damage to structural components were automatically identified. The proposed approach enables the transformation of unstructured maintenance personnel records into detailed diagnostic information, thereby creating opportunities to improve maintenance programs and to transition toward predictive reliability management of specific aircraft subsystems.*

Keywords: *aircraft, maintenance, external lighting, textual descriptions, natural language processing, topic modeling, Latent Dirichlet Allocation.*

Сведение об авторах

Каипбек Гульсанат Мэлскызы	Докторант, Академия Гражданской Авиации, Алматы, Казахстан E-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
Кошекков Кайрат Темирбаевич	Доктор технических наук, профессор, проректор по научной деятельности, Академия Гражданской Авиации, Алматы, Казахстан, E-mail: kkoshekov@mail.ru
Савостина Галина Владимировна	PhD, доцент кафедры «Энергетика и радиоэлектроника» Северо-Казахстанского университета им. М. Козыбаева, Петропавловск, Казахстан, E-mail: gvsavostina@ku.edu.kz

Авторлар туралы мәлімет

Каипбек Гүлсанат Мэлскызы	Азаматтық авиация академиясының докторанты, Алматы, Қазақстан E-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
Көшекков Кайрат Темирбаевич	Техника ғылымдары докторы, профессор, , Азаматтық авиация академиясының ғылыми жұмыстар жөніндегі проректоры, Алматы, Қазақстан E-mail: kkoshekov@mail.ru
Савостина Галина Владимировна	PhD, Солтүстік Қазақстан университетінің "Энергетика және радиоэлектроника" кафедрасының доценті. М. Қозыбаева, Петропавл, Қазақстан E-mail: gvsavostina@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
Galina Savostina	PhD, Associate Professor of the Department of "Power Engineering and Radio Electronics" of the M. Kozybayev North Kazakhstan University, Petropavlovsk, Kazakhstan E-mail: gvsavostina@ku.edu.kz