Московский государственный университет имени М.В.Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра суперкомпьютеров и квантовой информатики

Акопян Микаэла Тиграновна

Исследование возможностей предиктивного анализа публикационной активности

КУРСОВАЯ РАБОТА

Научный руководитель:
к.т.н.
Григорьева Мария Александровна

Москва, 2025

В условиях стремительного роста объёмов научной информации особую актуальность приобретает задача автоматизированного анализа публикационной активности. Это позволяет исследователям выявлять научные тренды, отслеживать динамику интереса к различным тематикам и строить прогнозы на будущее. В данной работе представлено исследование подходов к тематической кластеризации и предсказанию популярности направлений научных публикаций с использованием данных платформы InspireHEP. Выполнен сбор, обработка и анализ данных с применением методов машинного обучения и моделей временных рядов

Исследование возможностей предиктивного анализа публикационной активности

Акопян Микаэла Тиграновна

Abstract

Содержание

1 Введение

Физика высоких энергий (ФВЭ) – это раздел физики элементарных частиц, который изучает фундаментальные взаимодействия и структуры материи при экстремально высоких энергиях. Эта область считается крайне важной, так как именно она даёт понимание о фундаментальных составляющих материи и энергии, и, следовательно, о строении Вселенной. В данный момент ФВЭ является одной из наиболее динамично развивающихся областей современной науки, требующей постоянного анализа огромных объемов обновляющейся информации.

Как и в любой другой научной области, конференции играют ключевую роль в коммуникации учёных, объединяя исследователей со всех краёв света для представления своих достижений и обсуждения результатов. Важным аспектом таких мероприятий являются публикации, которые отражают актуальность тем и уровень интереса научного сообщества.

Информация о всех публикациях сохраняется на специализированных ресурсах. Одним из таких является InspireHEP — цифровая библиотека в области ФВЭ с открытым доступом. С появлением подобных платформ стало возможным централизованное хранение и доступ к данным о конференциях, публикациях и связанных метаданных. Эти данные открывают новые перспективы для анализа и прогнозирования тенденций в научной деятельности, что особенно актуально в условиях увеличивающегося объема информации. С InspireHEP удобнее всего работать, используя существующий API InspireHEP, который позволяет получить информацию об имеющихся на платформе публикациях, авторах, конференциях, цитированиях. Задачей курсовой работы является изучение данных для выявления закономерностей в научной коммуникации, анализа популярности тем и оценки динамики развития областей знаний.

Именно в силу высокой интенсивности научных исследований и большого объёма публикуемых материалов в области ФВЭ возникает необходимость в автоматизированном анализе публикационной активности. Это позволяет систематизировать знания, отслеживать научные тренды и прогнозировать развитие исследовательских направлений. Цифровая библиотека InspireHEP, являясь специализированной платформой, охватывающей большинство публикаций в этой области, предоставляет открытый и структурированный доступ к релевантным данным, что делает её идеальной основой для проведения такого анализа.

2 Постановка задачи

В данной работе рассматривается задача автоматического тематического анализа публикационной активности в области физики высоких энергий на основе открытых данных, полученных с платформы InspireHEP. Основной целью является выделение скрытых тематических направлений научных публикаций и анализ их распределения и динамики.

Входными данными имеющейся задачи являются метаданные о прошедших в 2014-2024 годах научных конференциях, полученные с помощью API платформы InspireHEP. Данные JSON - файлы в частности содержат в себе

- название конференции;
- даты проведения;
- страну и город проведения;
- описание конференции;
- кодовый номер конференции (cnum);
- число публикаций;

Используя кодовый номер конференции, можно получить информацию о всех публикациях, выставленных на данной конференции, а также о связанных с ними метаданных, среди которых

- название публикации;
- авторы;
- краткая аннотация;
- ссылка на полный текст;

В качестве выходных данных требуется получить предсказание динамики популярности различных тематических направлений в области физики высоких энергий, представленных на конференциях в предыдущие годы.

Для достижения поставленной цели необходимо выполнить следующие задачи:

- собрать все необходимые метаданные о прошедших в 2014-2024 годах научных конференциях,
- собрать все необходимые метаданные о выставленных на конференциях публикациях,
- провести тематический анализ публикаций с помощью алгоритмов машинного обучения: выделить темы и провести по ним классификацию документов
- провести анализ динамики тематики публикаций по годам, визуализировать полученные результаты
- присвоить каждой теме её название
- разработать модель предсказания динамики популярности тематики публикаций, оценить качество модели на исторических данных
- предсказать динамику популярности тематики публикаций на два года вперёд, визуализировать полученные результаты

3 Обзор используемых методов

3.1 Извлечение данных о конференциях и публикациях с помощью API InspireHEP

Был изучен API и извлечена информация о прошедших за последние пять лет конференциях и публикациях в них. О конференциях были получены: даты проведения, id в системе, краткие названия, публикации на каждой из них. О каждой публикации: названия, id в системе, авторы, ключевые слова, краткие описания.

Наиболее информативными источниками информации о содержаниях публикаций, безусловно, являются аннотации. Однако в InspireHEP аннотации не являются обязательным атрибутом, и в большом количестве работ отсутствуют. Так, только у 30 процентов публикаций есть краткие описания. Аналогично обстоит дело и с ключевыми словами. Поэтому было принято решение работать с названиями статей и извлекать тему из них, так как название является обязательным атрибутом.

3.2 Кластеризация документов по темам

Одним из ключевых этапов анализа публикационной активности является кластеризация документов по темам, позволяющая структурировать корпус текстов и выявить скрытые тематические направления.

Подходящий метод выбирался среди Latent Dirichlet Allocation (LcDA), Non-Negative Matrix Factorization (NMF) и Embedded Topic Model (ETM). LDA – популярный, классический метод, но его ограничение в учёте семантической близости слов стало существенным недостатком при работе с узкоспециализированными текстами. NMF неудобен в интерпретации результатов и требует большое число настроек. После проведённого анализа был выбран ETM, который сочетает вероятностный подход с использованием предобученных эмбеддингов слов. Это позволяет учитывать контекстные связи между словами, что критично для анализа научных публикаций.

3.2.1 Embedded Topic Model (ETM)

ETM — тематическая модель, в которой распределение слов в каждой теме строится не напрямую, а через скалярное произведение вектора темы и векторов слов. Для этого

используются заранее обученные векторные представления слов (эмбеддинги), которые помогают учитывать смысловую близость между слов.
ETM работает следующим образом:

- Каждая тема k представляется вектором $\boldsymbol{\beta}_k \in \mathbb{R}^L$ в эмбеддинговом пространстве размерности L.
- Для каждого документа непосредственно обучается распределение тем $\boldsymbol{\theta}_d = \mathrm{softmax}(\boldsymbol{\eta}_d)$, где $\boldsymbol{\eta}_d$ — вектор логитов, оптимизируемый напрямую градиентным методом.
- Вероятность наблюдения слова w в теме k задаётся через матрицу эмбеддингов $\mathbf{E} = \big[\mathbf{e}_1, \ldots, \mathbf{e}_V\big]^\top \in \mathbb{R}^{V \times L}$:

$$\phi_{k,w} = \big[\mathrm{softmax}\big(\boldsymbol{\beta}_k^\top \mathbf{E}^\top\big)\big]_w.$$

- Каждое слово w_{dn} в документе d порождается как

$$p(w_{dn} = w) = \sum_{k=1}^{K} \theta_{d,k}\, \phi_{k,w}.$$

Все параметры обучаются целиком градиентным методом с помощью оптимизатора Adam(Adaptive Moment Estimation).

Для матрицы частот $\mathbf{X} \in \mathbb{N}^{D \times V}$ максимизируется лог-правдоподобие

$$\mathcal{L} = \sum_{d=1}^{D} \sum_{w=1}^{V} X_{dw}\ \log\Big(\big[\boldsymbol{\theta}_d\ \mathrm{softmax}(\boldsymbol{\beta}\mathbf{E}^\top)\big]_w\Big).$$

Алгоритм оптимизации.

1. Инициализация: $\boldsymbol{\eta}_d \sim \mathcal{N}(0, 0.01),\ \boldsymbol{\beta}_k \sim \mathcal{N}(0, 0.01)$.
2. Прямой проход:

$$\begin{aligned} \boldsymbol{\theta}_d &= \mathrm{softmax}(\boldsymbol{\eta}_d), \\ \boldsymbol{\Phi} &= \mathrm{softmax}(\boldsymbol{\beta}\mathbf{E}^\top), \\ \hat{\mathbf{X}}_d &= D_d\, \boldsymbol{\theta}_d \boldsymbol{\Phi}, \end{aligned}$$

где $D_d = \sum_w X_{dw}$ — длина документа d.

3. Обратный проход: вычисление градиента $-\nabla_{\{\eta,\beta\}}\mathcal{L}$ и шаг Adam.

4. Повторение: цикл до заданного числа эпох или пока значение $\mathcal{L}$ не стабилизируется.

После сходимости строки $\mathbf{\Phi}$ задают распределения слов по темам; ключевые слова темы k — это w с максимальными $\phi_{k,w}$; матрица $\mathbf{\Theta}$ (полученная из $\boldsymbol{\eta}$) характеризует распределение тем по документам; компактность параметров $(K \times L)$ упрощает расширение модели на большие словари и ускоряет обучение по сравнению с LDA.

3.2.2 Global Vectors for Word Representation (GloVe)

При выборе подходящих предобученных эмбеддингов слов было принято решение остановиться на GloVe, так как они сами по себе легковесные и простые, но в то же время универсальные, достаточно полно покрывают язык и хорошо работают в большинстве задач.

Global Vectors были обучены на глобальной матрице совместных появлений слов в большом корпусе. В отличие от моделей вроде word2vec, которые обучаются на основе предсказания соседей по контексту (локальная информация), GloVe использует глобальную статистику корпуса: матрицу соотношений совместной встречаемости слов. Идея в том, что отношение частот слов содержит семантическую информацию, которую можно эффективно закодировать в векторном пространстве.

Модель минимизирует функцию потерь, которая старается аппроксимировать логарифм количества совместных появлений двух слов через скалярное произведение их векторов. Это позволяет GloVe учитывать как абсолютную, так и относительную частоту слов, благодаря чему модель сохраняет семантические и синтаксические связи между словами. Вектора, полученные с помощью GloVe, хорошо работают в задачах аналогий (king – man + woman = queen) или (Germany - Berlin = France - Paris), классификации, кластеризации и других NLP-задачах.

3.3 Определение названий выявленных кластеров тем

Для определения названий выявленных кластеров были изучены несколько API больших языковых моделей (LLM). Рассматривались ChatGPT, DeepSeek и Gemini. ChatGPT

показал высокую точность и качество генерации названий кластеров, однако он требует для использования API высокой оплаты и использования VPN. DeepSeek, хоть и не нуждался в этом сервисе, плохо справился с поставленной задачей и имел сравнительно невысокие лимиты использования. Наилучшие результаты показал Gemini. Он, хоть и нуждался в использовании VPN, имел высокие лимиты использования и позволял генерировать названия кластеров с высокой точностью и качеством.
В результате, для определения названий кластеров тем был выбран Gemini.

3.4 Предсказание на временных рядах

Как возможные варианты предсказания на временных рядах были рассмотрены ARIMA, SARIMA и Prophet.

ARIMA (Autoregressive Integrated Moving Average) — это классическая модель временных рядов, которая использует авторегрессию, интеграцию и скользящее среднее для предсказания будущих значений на основе прошлых данных. Она хорошо работает с линейными временными рядами, но может быть сложной в настройке и требует стационарности данных(свойство временного ряда, при котором его статистические характеристики такие, как дисперисия, ковариация, среднее не меняются со временем.).

SARIMA (Seasonal Autoregressive Integrated Moving Average) — это расширение ARIMA, которое учитывает сезонные компоненты временных рядов. Она добавляет сезонные параметры к модели ARIMA, что полезно, однако, SARIMA может быть сложной в настройке и требует больше вычислительных ресурсов. Кроме того, она обладает такими же недостатками, что и предшественница, и так же, как она требует стационарности данных и ручной настройки параметров.

Prophet — это библиотека от Facebook, которая предназначена для предсказания временных рядов с учетом сезонности и праздников. Она проста в использовании и позволяет быстро получать результаты.

Prophet автоматически обрабатывает пропуски в данных и может работать с нестационарными временными рядами. Она не требует ручной настройки параметров и слабо зависит от предположений о распределении данных. Кроме этого, большим преимуществом данной библиотеки являестся поддержка редких и коротких временных рядов,что актуально для полученных данных о публикациях.

4 Описание практической части

Код был написан на языке Python с использованием библиотек pandas, numpy, sklearn, matplotlib. Все эксперименты проводились на графических процессорах NVIDIA Tesla V100 32 Gb с использованием фреймворка PyTorch. Программная реализация в полном виде выложена в репозитории на GitHub.

4.1 Извлечение данных о конференциях и публикациях с помощью API InspireHEP

Инструкция по использованию API InspireHEP была взята с официального github-репозитория. Кроме того, была использована расширенная документация к API. Благодаря запросам к API InspireHEP были получены JSON-файлы с данными о прошедших конференциях и публикациях в них. Примеры полученных JSON-файлов: описание конференции, описание всех публикации данной конференции, описание одной из статей, представленных на данной конференции.

В таблице ниже приведены примеры публикаций, полученные с использованием API InspireHEP. Для каждой статьи указано название, год публикации, ключевые слова и аннотация. Эти метаданные используются в последующем тематическом моделировании и анализе динамики научных направлений. Особое внимание уделяется ключевым словам и аннотациям — они являются основным источником информации о тематике публикации. Однако, как уже отмечалось ранее, аннотации и ключевые слова доступны не для всех документов, что делает наличие даже частичных данных ценным источником для анализа.

Название	Год	Ключевые слова	Аннотация
Beam Commissioning of PAL-XFEL	2016	gun; undulator; laser; linac; cathode	The Pohang Accelerator Laboratory X-ray Free electron Laser (PAL-XFEL) project aims at the generation of X-ray FEL radiation...
Commissioning of the MAX IV Light Source	2016	storage-ring; emittance; injection; lattice; vacuum	This presentation reports on the beam commissioning status of MAX IV, experience gained and lessons learned...
Limits and Possibilities of Laser Wakefield Accelerators	2016	electron; laser; plasma; coupling; focusing	This presentation provides an outlook into the future of laser-driven plasma wakefield accelerators. What has been achieved...
Review of Linear Optics Measurements and Corrections in Accelerators	2016	optics; coupling; quadrupole; collider; betatron	The measurement and correction of optics parameters has been a major concern since the advent of strong focusing synchrotron accelerators...
Design and Optimization Strategies of Nonlinear Dynamics in Diffraction-limited Synchrotron Light Sources	2016	sextupole; lattice; optics; emittance; resonance	This talk introduces the most recent achievements in the control of nonlinear dynamics in electron synchrotron light sources...

Таблица 1: Пример публикаций с ключевыми метаданными

4.2 Кластеризация документов по темам

Как уже упоминалось, одним из ключевых этапов анализа публикационной активности является кластеризация документов по темам, позволяющая структурировать корпус текстов и выявить скрытые тематические направления. Был написан программный модуль, который позволяет проводить тематическое моделирование с использованием Embedded Topic Model (ETM) и предобученных эмбеддингов GloVe (точная версия: glove.6B.300d.txt).

Результатом работы данного модуля является набор из 25 кластеров, каждый из которых представляет собой группу публикаций, связанных общей тематикой.

Для наглядной демонстрации тематической структуры корпуса была построена двумерная проекция документов с использованием алгоритма t-SNE на основе усреднённых эмбеддингов GloVe. Из всех тем, выделенных в процессе тематического моделирования,

были выбраны пять, максимально различающихся по семантическому содержанию (расстояние между центроидами векторных представлений).

Следует отметить, что проекция, представленная на рисунке, основана на алгоритме t-SNE, который отображает данные из многомерного пространства (в данном случае — 300-мерного пространства GloVe-векторов) в двумерное. Поскольку алгоритм стремится сохранить локальные отношения между точками, но не глобальные расстояния, существует частичное наложение кластеров.

Тем не менее, даже в проекции можно наблюдать тенденцию к формированию отдельных плотных облаков, что подтверждает наличие тематической структуры в исходных данных.

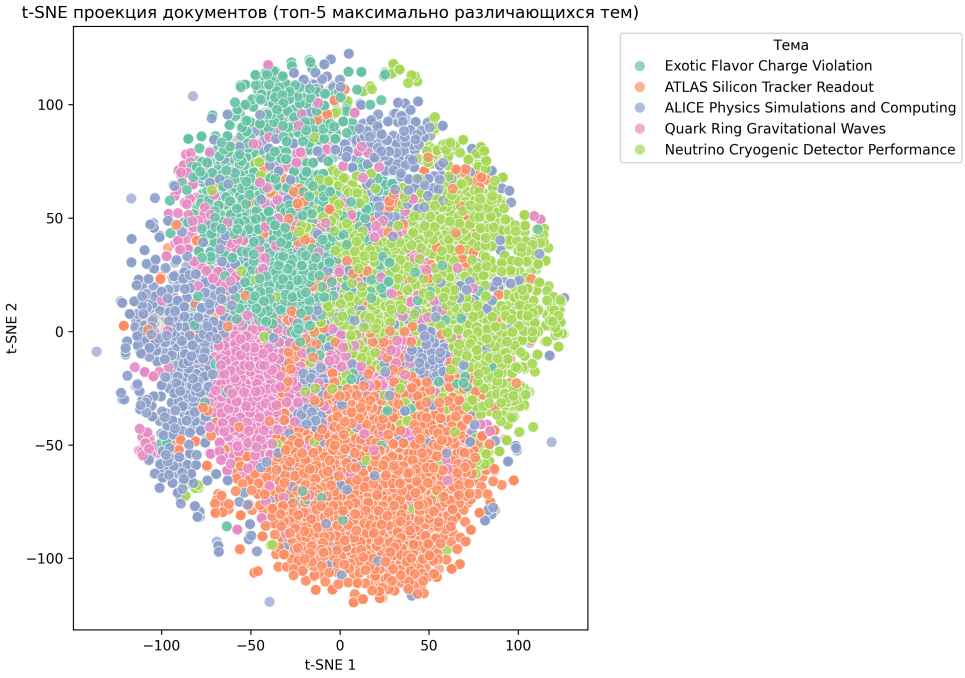

Рис. 1: Результаты кластеризации публикаций(проекция t-SNE)

4.3 Присваивание общих кластерам публикаций

В результате кластеризации публикаций были выделены 25 тем, которые можно использовать для дальнейшего анализа. Однако данные результаты были бы бессмысленными без привязки к реальным научным направлениям. С помощью вызовов к API Gemini (конекретнее, к модели gemini-2.0-flash) было получено 25 меток, которые были присво-

ены кластерам.

Пример результатов работы модели приведен в таблице ниже.

Название	**Год**	**Ключевые слова**	**Аннотация**	**Тема**
Beam Commissioning of PAL-XFEL	2016	gun; undulator; laser; linac; cathode	The Pohang Accelerator Laboratory X-ray Free electron Laser (PAL-XFEL) project aims at the generation of X-ray FEL radiation...	High Energy Cosmic Rays
Commissioning of the MAX IV Light Source	2016	storage-ring; emittance; injection; lattice; vacuum	This presentation reports on the beam commissioning status of MAX IV, experience gained and lessons learned...	Particle Astrophysics and Calorimetry
Limits and Possibilities of Laser Wakefield Accelerators	2016	electron; laser; plasma; coupling; focusing	This presentation provides an outlook into the future of laser-driven plasma wakefield accelerators...	Neutron Electromagnetic Interaction Spectroscopy
Review of Linear Optics Measurements and Corrections in Accelerators	2016	optics; coupling; quadrupole; collider; betatron	The measurement and correction of optics parameters has been a major concern since the advent of strong focusing synchrotron accelerators...	Particle Physics Power Control Development
Design and Optimization Strategies of Nonlinear Dynamics in Diffraction-limited Synchrotron Light Sources	2016	sextupole; lattice; optics; emittance; resonance	This talk introduces the most recent achievements in the control of nonlinear dynamics in electron synchrotron light sources...	ATLAS Silicon Tracker Readout

Таблица 2: Примеры публикаций с ключевыми метаданными и назначенными темами

4.4 Предсказание на временных рядах

Для предсказания на временных рядах использовалась модель Prophet. Было принято решение не делать предсказание всех тем, а только тех, которые содержат в себе более 30 публикаций и были представлены на конференциях не менее 4 лет подряд. Из них были отобраны 10 тем, которые входили в датасет для предсказания.

4.4.1 Подготовка данных для предсказания

При построении временных рядов каждая тема представлялась как совокупность публикаций, отнесённых к соответствующему кластеру. Для каждой темы формировался временной ряд, в котором каждому году сопоставлялась доля публикаций, относящихся к данной теме, относительно общего числа публикаций за этот год. То есть значения ряда отражали не абсолютное количество публикаций, а нормированную популярность темы во времени. Для сглаживания колебаний применялось скользящее среднее с окном в два года. Такой способ агрегации позволил корректно учитывать рост общего числа публикаций и обеспечил сравнимость значений между годами.

4.4.2 Оценка качества модели

Оценка качества модели проводилась с помощию разделения выборки на test и train. Модель Prophet обучалась на train выборке, представляющей из себя собранные до 2023 года данные о популярности выявленных тем. Предсказание проводилось на test выборке, которая состояла из данных о популярности тем в 2024 году.

В качестве метрики для оценки качества предсказания использовались следующие метрики:

- MAE (Mean Absolute Error) — средняя абсолютная ошибка;
- MAPE (Mean Absolute Percentage Error) — средняя абсолютная процентная ошибка;
- SMAPE (Symmetric Mean Absolute Percentage Error) — симметричная средняя абсолютная процентная ошибка;
- RMSE (Root Mean Square Error) — корень из средней квадратичной ошибки;

Результаты приведены в таблице ниже.

MAE	**MAPE, %**	**RMSE**	**SMAPE, %**
0.02	16.35	0.02	17.82

Таблица 3: Средние метрики

На основе полученных значений метрик можно сделать вывод о достаточно высоком качестве предсказания модели Prophet. Значение MAE = 0.02 (в диапазоне от 0 до 1) указывает на низкую абсолютную ошибку. RMSE = 0.02 (также в диапазоне от 0 до 1), практически равный MAE, говорит о отсутствии крупных выбросов: ошибки распределены равномерно, без сильных отклонений. MAPE = 16.35% и SMAPE = 17.82% — это приемлемый уровень относительной ошибки для задач предсказания временных рядов.

Таким образом, модель демонстрирует устойчивое поведение и хорошую способность к обобщению на новых данных, собранных после 2023 года.

На рисунке ниже приведены результаты предсказания на временных рядах. Точки отображают предсказания на 2024 год, а линии — фактические данные о популярности тем.

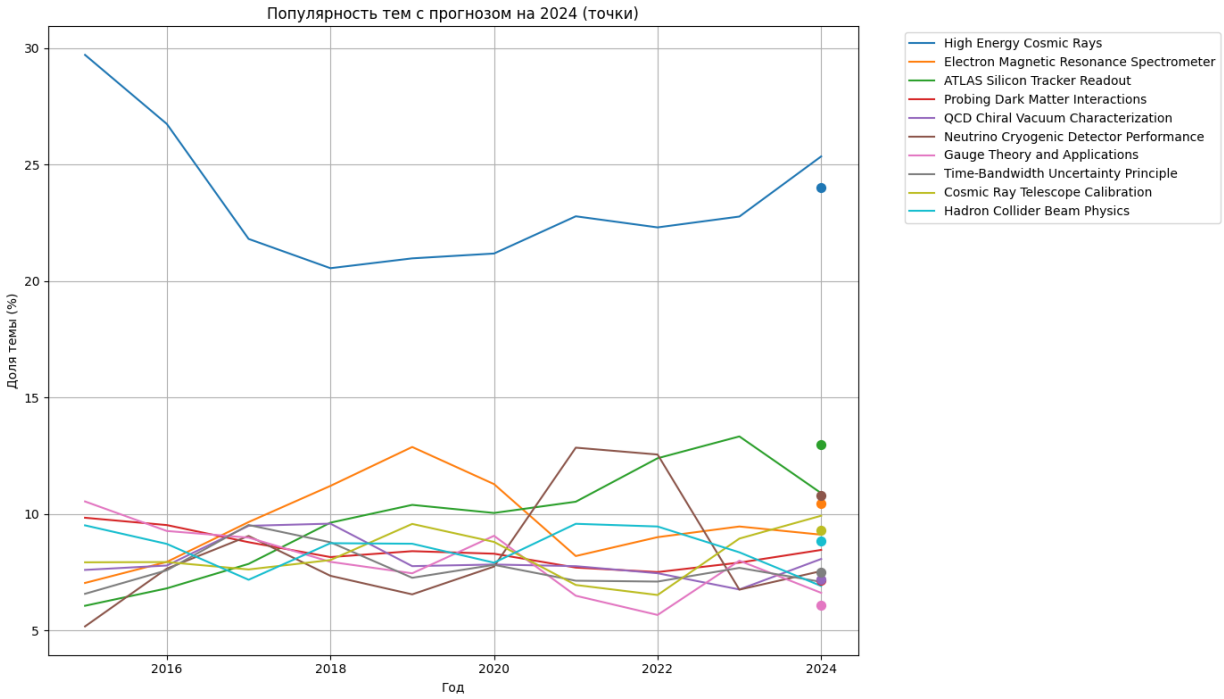

Рис. 2: Результаты предсказания на временных рядах

4.4.3 Предсказание на 2025-2026 год

В связи с удовлетворительными рещультатами предсказания на тестовой выборке, модель была использована для предсказания на 2025-2026 год. Результаты предсказания приведены на рисунке ниже.

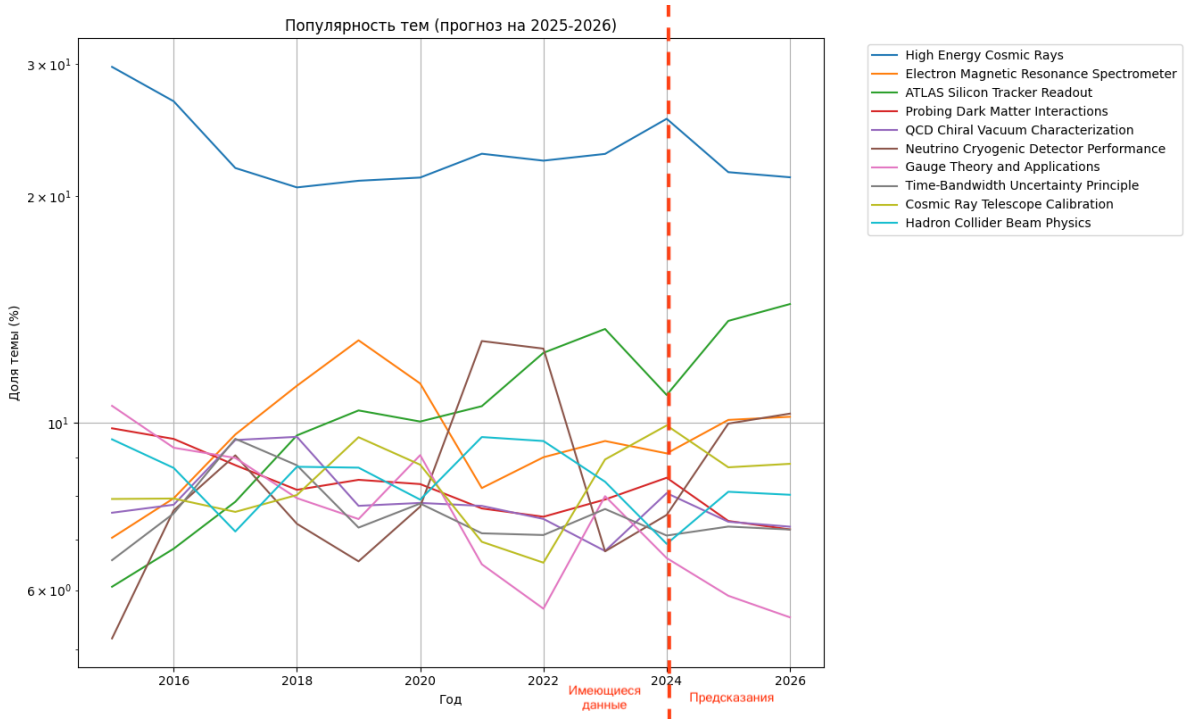

Рис. 3: Результаты предсказания на 2025-2026 год

5 Заключение

В данной работе была разработана и реализована система анализа и прогнозирования публикационной активности в области физики высоких энергий на основе данных платформы InspireHEP. В рамках исследования:

- Был ыполнен сбор метаданных о научных конференциях и публикациях, проходивших в 2014–2024 годах, с использованием API InspireHEP.
- Была роведена кластеризация научных публикаций по темам с помощью модели Embedded Topic Model (ETM), использующей предобученные эмбеддинги GloVe. В результате было выделено 25 устойчивых тематических кластеров.
- Для каждого тематического кластера были сгенерированы осмысленные названия с использованием API Gemini, что позволило интерпретировать результаты тематического моделирования.
- Был выполнен анализ динамики публикационной активности по темам, построены временные ряды популярности каждой темы.
- Была азработана и обучена модель предсказания на временных рядах на базе библиотеки Prophet. Модель показала хорошее качество, что свидетельствует об адекватности предсказаний.

Результаты показали, что предложенный подход способен автоматически выявлять актуальные темы в области физики высоких энергий и предсказывать их динамику. Метод доказал свою применимость для анализа научных данных и может быть расширен для других предметных областей.

6 Список литературы

1. *Dieng, Adji B.* Topic Modeling in Embedding Spaces. — 2019. https://arxiv.org/abs/1907.04907.

2. *Pennington, Jeffrey.* GloVe: Global Vectors for Word Representation / Jeffrey Pennington, Richard Socher, Christopher Manning // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) / Ed. by Alessandro Moschitti, Bo Pang, Walter Daelemans. — Doha, Qatar: Association for Computational Linguistics, 2014. — . — Pp. 1532–1543. https://aclanthology.org/D14-1162.

3. *Kingma, Diederik P.* Adam: A Method for Stochastic Optimization. — 2017. https://arxiv.org/abs/1412.6980.

4. *MacQueen, James.* Some methods for classification and analysis of multivariate observations / James MacQueen et al. // Proceedings of the fifth Berkeley symposium on mathematical statistics and probability / Oakland, CA, USA. — Vol. 1. — 1967. — Pp. 281–297.