Тематическая классификация сайтов хостинга «Narod.ru» как часть стратегии по сохранению сайтов раннего интернета

И. В. Асланов, А. С. Козлова, И. В. Бибилов, Е. В. Котельников

Аннотация—**Исследование посвящено сохранению и изучению сайтов хостинга «Narod.ru», активно функционировавшего в 2000–2013 годах. В рамках работы сайты хостинга рассматриваются как исчезающие объекты цифрового наследия, сохранение и анализ которых может быть интересен экспертам из разных предметных областей, в особенности культурологам и исследователям цифрового фольклора раннего интернета.**

В исследовании предложен подход к тематической классификации сайтов с использованием больших языковых моделей. Сначала была проведена ручная разметка выборки из сохраненных сайтов хостинга, в ходе которой главные страницы сайтов были отнесены к тематическим категориям в соответствии с рубрикатором компании Google. Затем, на основе размеченной выборки, выполнена оценка качества 17 проприетарных и открытых больших языковых моделей в задаче многометочной тематической классификации веб-страниц. Наилучший результат продемонстрировала модель gemini-2.5-pro (Samples-F1=0,708). Предложенный подход к тематической классификации сайтов раннего интернета позволит исследователям выявлять и анализировать культурные, социальные и коммуникационные паттерны формирования цифрового общества.

Ключевые слова—**архивирование веб-сайтов, цифровое наследие, Narod.ru, большие языковые модели, многометочная классификация.**

I. Введение

С каждым годом из сети исчезает все больше артефактов раннего Интернета. Четверть всех веб-страниц, существовавших с 2013 по 2023 год, больше не доступны – спустя менее 10 лет после создания [5]. Данные сайты могут быть важны для исследователей из разных областей науки – например, антропологам, социологам и исследователям Интернета.

Проект, посвященный проблеме сохранения и изучения сайтов, в рамках которого проводится настоящее исследование, работает с данными сайтов хостинга «Narod.ru». «Narod.ru» представляет собой хостинговый сервис, основанный компанией «Яндекс» в 2000 году и впоследствии переданный в собственность компании «uCoz» в 2013 году[1]. Сервис давал пользователям возможность бесплатно создавать собственные веб-сайты как самостоятельно, так и с помощью встроенного конструктора. «Narod.ru» пользовался популярностью среди широкого круга пользователей: на платформе создавались как персональные сайты, так и веб-ресурсы сообществ, бизнес-компаний и государственных организаций, включая школы и детские сады. Хостинговый сервис также содержал обширную коллекцию разнообразных документов, включая фотографии, изображения, официальные бумаги и электронные таблицы.

Помимо архивации сайтов существуют и другие задачи, которые могут возникнуть в рамках сохранения культурного наследия сайтов «Narod.ru». Одной из таких задач является классификация сайтов по тематическим категориям. Выполнение данной задачи необходимо для создания удобного рубрикатора в целях фильтрации сайтов по тематикам, в том числе в рамках предоставления возможности углубления в исследования определенной тематики, а также для использования в других задачах проекта, включая фильтрацию нежелательного контента на сайтах «Narod.ru».

В предыдущих исследованиях тематическое моделирование сайтов хостинга было проведено при помощи кластеризации на основе фреймворка BERTopic [10]. Однако данный подход был недостаточно эффективен ввиду несовершенства алгоритмов кластеризации: HDBSCAN определял большое количество сайтов как шум, а метод K-средних добавлял в кластеры сайты, явно не принадлежащие к этой категории [21]. Помимо этого, BERTopic позволял присвоить веб-странице лишь одну категорию, что не соответствует действительности: сайты в Интернете могут одновременно охватывать большое количество разных тематик. В настоящем исследовании

Статья получена 25 октября 2025 г.
Статья подготовлена по итогам выступления на конференции «Интернет и современное общество» (IMS-2025).
Асланов Ильяс Васифович, исследователь Школы вычислительных социальных наук, Европейский университет в Санкт-Петербурге, ORCID 0009-0001-8527-243X (e-mail: iaslanov@eu.spb.ru).
Козлова Анна Сергеевна, заместитель директора программ по направлению «Прикладная информатика» Школы вычислительных социальных наук, Европейский университет в Санкт-Петербурге, ORCID 0009-0008-5482-5952 (email: annakozlova@eu.spb.ru).
Бибилов Иван Владимирович, технический директор программ по направлению «Прикладная информатика» Школы вычислительных социальных наук, Европейский университет в Санкт-Петербурге (email: ibibilov@eu.spb.ru).
Котельников Евгений Вячеславович, доктор технических наук, профессор, директор Школы вычислительных социальных наук, Европейский университет в Санкт-Петербурге, ORCID 0000-0001-9745-1489 (email: e.kotelnikov@eu.spb.ru).

[1] Переезд сайтов narod.ru на платформу uCoz // Официальный блог uCoz. URL: https://blog.ucoz.ru/blog/pereezd_sajtov_narod_ru_na_platformu_ucoz/2013-01-31-255 (дата обращения: 25.10.2025).

предлагается другой подход для тематической категоризации сайтов хостинга «Narod.ru» – многометочная классификация на основе больших языковых моделей.

Вклад настоящей работы заключается в разработке и исследовании методологии многометочной классификации с применением больших языковых моделей, которая в дальнейшем позволит категоризовать по тематикам все сайты хостинга, что напрямую повлияет на развитие проекта по сохранению и изучению наследия «Narod.ru».

Статья имеет следующую структуру. Во втором разделе приведены основные факторы, влияющие на подходы к сохранению сайта как объекта цифрового наследия, описаны детали сохранения сайтов на примере хостинга «Geocities», а также сделан обзор проделанной работы по сохранению данных хостинга «Narod.ru». В третьем разделе описываются исследуемые данные и применяемые модели. В четвертом разделе предлагается методология многометочной классификации на основе больших языковых моделей и способы оценки качества классификации. Пятый раздел содержит анализ полученных результатов и предложения по оптимизации процесса тематической классификации для всех сохраненных сайтов с учетом различных факторов, выявленных в ходе данного исследования.

II. Обзор предыдущих работ

A. Проблема сохранения сайтов как цифрового наследия

Обратимся к определению цифрового наследия для понимания основания, по которому мы включаем сайты хостинга в это понятие. Отправной точкой в формировании определения послужило принятие ЮНЕСКО «Хартии о сохранении цифрового наследия» в 2003 году [6]. В Хартии к цифровому наследию причисляется широкий спектр форматов, одними из которых являются как веб-сайты, так и содержащиеся в них документы, тексты, движущиеся и недвижимые изображения, а также другие типы данных. В Хартии подчеркивается угроза утраты цифрового наследия и необходимость конкретных действий по сохранению и обеспечению доступности сохраняемых цифровых объектов. Также отмечается, что разработка стратегии в области сохранении цифрового наследия должна учитывать степень неотложности и имеющиеся средства, а способствовать решению этой задачи может взаимодействие разных команд от создателей цифрового наследия до правообладателей и других заинтересованных сторон.

Сайты хостинга «Narod.ru» являются полноправной частью цифрового наследия, находятся под угрозой исчезновения и нуждаются в сохранении и изучении. Для этого важно понять контекст проблемы методологии сохранения цифрового наследия и степени сохранности сайтов хостинга «Narod.ru», чему и посвящен данный раздел статьи.

На первый взгляд может показаться, что проблема сохранения уже должна быть решена, ведь сохранение артефактов прошлого для изучения и показа – цель, которая в свое время легла в основу определения «музей». За столетия этой осознанной работы такие науки как археология, искусствоведение, музеология выработали сильную методологию по отбору, сохранению, реставрации, хранению и репрезентации памятников и культурных объектов. Созданные и институализированные для этого регламенты основаны на опыте тысяч исследователей на протяжении довольно долгого периода времени. Однако, несколько сложнее ситуация с методологией по сохранению и репрезентации нематериального наследия, и совсем непростой она становится для объектов цифрового наследия. К тому же, проработанность вопроса сохранения цифрового наследия неравномерна для разных регионов, стран и типов наследия и часто зависит от законодательства, финансовой поддержки институций, готовых взять на себя вопрос сохранения, наличия исследовательских и волонтерских объединений и других обстоятельств. Приведем несколько примеров существующих инициатив.

Проект «One Terabyte of Kilobyte Age» [12], посвящен изучению сайтов хостинга «GeoCities», который работал с 1994 по 2009 год. В 2009 году компания «Yahoo!» закрыла хостинг, дав пользователям лишь несколько месяцев на перенос данных. За этот короткий срок «Archive Team» сумела спасти почти терабайт страниц «GeoCities», а спустя год они выложили данные в открытый доступ как «Geocities.archiveteam.torrent» [13]. Позднее, на основе этих данных был создан проект «One Terabyte of Kilobyte Age» [12], который послужил источником исследовательских и художественных работ, включая выставку «Удаленный город» Ричарда Вийгена – проект, визуализирующий сообщества внутри «GeoCities» [17], и книгу «Digital Folklore Reader», ставшую важной отправной точкой для многих инициатив по сохранению цифрового наследия [8].

Как мы можем видеть, существует риск навсегда потерять материал для изучения цифрового фольклора, который также потенциально может содержать уникальные художественные, литературные, культурные и исторические артефакты. Это подтверждает актуальность задачи сохранения и превращения сайтов в открытую базу данных, доступную для исследователей. К материалам раннего Интернета может быть задан широкий круг исследовательских вопросов из самых разных предметных областей – музеологии, цифровой антропологии, литературы, веб-дизайна и так далее. И, наконец, собранные данные могут послужить незаменимым источником для исследователей истории Интернета.

История развития и значимость «GeoCities» и «Narod.ru» похожа для англоязычных и русскоязычных сегментов Интернета соответственно. Для сайтов хостинга «Narod.ru», на сохранении и изучении которых и сосредоточено представленное исследование, до этого момента не было предпринято систематических усилий по сохранению. Прежде чем выстраивать стратегию сохранения сайтов и стратегию развития проекта,

следует проверить существуют ли еще какие-либо институции или команды, которые уже сохранили часть интересующих нас данных.

В первую очередь стоит обратиться к ресурсу «The Wayback Machine», созданному командой «Internet Archive»[2]. Проверить наличие копии сайта в «Internet Archive» можно при помощи API, которое реализовано через добавление интересующего домена в аргументы к URL-запросу. Всего в «Internet Archive» сохранены 71,3% от всех сайтов «Narod.ru», переданных авторам проекта компанией «Яндекс». Необходимо учитывать, что в данном случае речь идет только о снапшотах главных страниц сайтов – в ходе проверки наличия снапшотов для всех сайтов отфильтрованного списка в API подавались лишь адреса главных страниц, поэтому реальное количество сохраненных страниц и файлов для всего списка сайтов будет значительно меньше. Проверка количества архивированных в «Internet Archive» дочерних страниц сайтов на основе 10 000 сайтов, сохраненных в рамках проекта по архивации сайтов хостинга «Narod.ru», выявила, что в «Internet Archive» доступно лишь 22,3% всех веб-страниц

Еще одним источником для проверки гипотезы о том, что сайты хостинга могут быть сохранены другими командами, может стать «Национальный цифровой архив» – инициатива АНО «Информационная культура»[3] по сохранению и архивации данных российского сегмента Интернета, учрежденного И. В. Бегтиным. Однако, проверка показала, что данный архив хранит сохраненные копии лишь 10 уникальных доменов сайтов «Narod.ru».

Таким образом, можно сделать вывод о том, что в настоящий момент сайты хостинга не только находятся в зоне риска полного удаления из веба, но и уже стремительно исчезают и частично разрушаются. Более того, сохраненные нами 10 000 сайтов с большой вероятностью являются самым крупным образцом данных на текущий момент. Однако для сохранения большего объема данных требуется проработанная методология сохранения и предварительный анализ данных, в том числе классификация сайтов по тематикам, что позволит перестать воспринимать сайты хостинга как некоторый черный ящик и проводить не только качественные, но и количественные исследования. В свою очередь более глубокое понимание тематики контента сайтов также поможет привлечь в проект необходимые ресурсы и исследователей.

B. Предыдущие этапы получения и исследования данных

Сохранение и изучение сайтов хостинга «Narod.ru» является целью для разноуровневого многолетнего проекта, который требует привлечения как разработчиков, так и исследователей самых разных областей технических и гуманитарных наук. Однако, первым шагом для разработки стратегии развития такого проекта, как мы уже упоминали ранее, должно стать изучение степени сохранности данных. Это необходимо, так как понимание масштаба утрат и возможных способов и источников пополнения базы данных напрямую влияет на приоритетность задач, поставленных перед командой. Первым источником данных для проекта стал список с доменами третьего уровня сайтов хостинга (например, в исходном списке указано «site» для адреса https://site.narod.ru/), предоставленный компанией «Яндекс» в 2023 году: данный список адресов включал 637 000 доменных имён. Также компанией был предоставлен более полный список на более чем 1 200 000 доменных имен, содержащий все сайты первого списка, а также большое количество некорректно работающих сайтов и сайтов с вредоносным и противоправным содержимым.

Разумеется, полное копирование и сохранение всех сайтов из «большого» списка – важная перспективная цель, однако, это достаточно ресурсозатратно в связи с большим объемом данных. На данном этапе нам удалось сохранить 10 000 сайтов из первого, более короткого списка. Они стали основой для создания прототипа базы данных, на примере которой должны быть разработаны принципы сохранения сайтов. В рамках разработки первого прототипа базы данных были решены следующие задачи: сохранено содержимое 10 000 сайтов; собранные данные были обработаны в целях извлечения из них метаинформации; были созданы базы данных и веб-интерфейс, предоставляющий доступ к сформированному датасету, включая возможности полнотекстового поиска по страницам и сохраненным документам, а также первичный анализ полученных данных.

Для сбора данных использовался веб-краулер, разработанный на языке программирования Python с применением библиотек Selenium и BeautifulSoup. Собранная информация была сохранена в базу данных PostgreSQL. HTML-документы и сопутствующие файлы, загруженные с сайтов, были отдельно сохранены на сервере.

В результате работы веб-краулера удалось обработать 9 956 из 10 000 сайтов. В общей сложности было обработано 532 118 страниц, из которых 401 736 (75,5%) были успешно сохранены в базу данных; остальные страницы содержали ошибки различного рода. Кроме того, было сохранено более 2,2 миллиона файлов общим объемом 84,3 Гбайт, включая 2 130 401 изображение, 14 543 PDF-документа, 18 196 текстовых файлов, 1 934 электронные таблицы и 2 513 презентаций.

Важно отметить, что на момент сохранения 10 000 сайтов в первый прототип базы данных весной 2024 года, полученные 9 956 сайтов были доступны в Интернете, однако сейчас значительная часть этих сайтов недоступна. И тем не менее даже два года назад их доступность была относительна, так как степень видимости этих сайтов для поисковых систем попадает в так называемых феномен «скрытого веба» [4]. Такие сайты не индексируются, и, соответственно, не появляются в поисковой выдаче. Дело не только в том,

[2] The Wayback Machine // Internet Archive. URL: https://archive.org/ (дата обращения: 25.10.2025).
[3] Национальный цифровой архив России. URL: https://ruarxive.org/ (дата обращения: 25.10.2025).

что эти сайты находятся в «скрытом вебе», но и в том, что их функционирование при нулевом посещении не приносит прибыли владельцам хостинга, что может привести к удалению сайтов, как это было ранее упомянуто на примере хостинга «GeoCities» и компании «Yahoo!». Также сайты могут перестать работать из-за «цифрового распада» – утраты элементов или архитектуры сайтов из-за устаревания кода [11]. Возможно, точные причины утрат сайтов установить не удастся, однако, скорость потерь сайтов стремительна. Например, из небольшой выборки, которую мы использовали для ручной разметки за полтора года, которые прошли от сбора данных до настоящего момента, доступными осталось 215 из 400 сайтов. Этот факт еще раз подтверждает необходимость срочного сохранения сайтов хостинга «Narod.ru» в связи с большим риском утраты сайтов.

Полученные нами данные стали источником для развития двух студенческих проектов под руководством авторов статьи и послужили материалом для магистерских диссертаций, выступлений и статей, что позволило ввести данные о хостинге «Narod.ru» в научный оборот.

III. Данные и модели

C. Данные

В исследовании были использованы материалы проекта по сохранению сайтов хостинга «Narod.ru» как объектов цифрового культурного наследия[4]. В ходе этого проекта было заархивировано свыше 400 000 страниц сайтов хостинга. Для настоящего исследования отбирались главные страницы сохраненных сайтов. Для анализа использовались тексты длиной более 50 символов, что позволило исключить случаи, где определение тематик могло быть затруднено из-за малого объема текста. Итоговый датасет составил 9 466 главных страниц из базы данных проекта по сохранению сайтов «Narod.ru».

A. Модели

В экспериментах были использованы 17 современных больших языковых моделей от разных компаний: 7 проприетарных и 10 открытых моделей.

Anthropic Claude:
- Claude-sonnet-4.5 – проприетарная модель, с возможностью включения режима рассуждения [2].

Google Gemini:
- Gemini-2.5-pro – проприетарная рассуждающая модель [7];
- Gemini-2.5-flash – проприетарная модель, оптимизированная для стоимости и качества ответов [7].

OpenAI GPT:
- Gpt-5 – проприетарная рассуждающая модель [15];
- Gpt-oss-120b – Mixture-of-Experts (MoE) архитектура, открытая рассуждающая модель, 117 млрд. параметров, 5,1 млрд. активных параметров [1];
- Gpt-oss-20b – MoE-архитектура, открытая рассуждающая модель, 21 млрд. параметров, 3,6 млрд. активных параметров [1].

Alibaba Qwen:
- Qwen3-max – проприетарная модель, более 1 трлн параметров[5];
- Qwen3-next-80b-a3b-instruct – MoE-архитектура, открытая модель, 80 млрд параметров, 3,9 млрд активных параметров[6];
- Qwen3-next-80b-a3b-thinking – MoE-архитектура, открытая рассуждающая модель, 80 млрд параметров, 3,9 млрд активных параметров[7];
- Qwen3-235b-a22b-thinking-2507 – MoE-архитектура, открытая модель, 235 млрд параметров, 22 млрд активных параметров [18];
- Qwen3-235b-a22b-2507 – MoE-архитектура, открытая рассуждающая модель, 235 млрд параметров, 22 млрд активных параметров [18].

xAI Grok:
- Grok-4-fast – проприетарная рассуждающая модель [20];
- Grok-4 – проприетарная рассуждающая модель, оптимизированная для стоимости и качества ответов [19].

Deepseek:
- Deepseek-r1-0528 – MoE-архитектура, открытая рассуждающая модель [9];
- Deepseek-chat-v3.1 – MoE-архитектура, открытая модель с возможностями рассуждения, 671 млрд параметров, 37 млрд активных параметров [14];
- Deepseek-v3.1-terminus – MoE-архитектура, открытая модель с возможностями рассуждения, 671 млрд параметров, 37 млрд активных параметров[8].

MoonshotAI Kimi:
- Kimi-k2-0905 – MoE-архитектура, открытая модель, более 1 трлн параметров, 32 млрд активных параметров [3].

IV. Методология

B. Тематический рубрикатор

Для назначения тематик веб-страницам хостинга «Narod.ru» был использован подход многометочной (multi-label) классификации на основе больших языковых моделей.

В качестве множества классов для задачи многометочной классификации использовались 27 тематик верхнего уровня рубрикатора веб-страниц компании Google (см. Приложение 1). Каждая тематическая категория рубрикатора сопровождалась подкатегориями, которые в данном исследовании использовались как примеры контекста категорий для больших языковых моделей. Для анализа сайтов «Narod.ru» было решено добавить к рубрикатору

[4] Народ и цифровое наследие. URL: https://projects.pandan.eusp.org/narod (дата обращения: 25.10.2025).

[5] Qwen3-Max: Just Scale it. URL: https://qwen.ai/blog?id=241398b9cd6353de490b0f82806c7848c5d2777d&from=research.latest-advancements-list (дата обращения: 25.10.2025).

[6] Qwen3-Next: Towards Ultimate Training & Inference Efficiency. URL: https://qwen.ai/blog?id=4074cca80393150c248e508aa62983f9cb7d27cd&from=research.latest-advancements-list (дата обращения: 25.10.2025).

[7] Ibid.

[8] DeepSeek-V3.1-Terminus. URL: https://api-docs.deepseek.com/news/news250922 (дата обращения: 25.10.2025).

дополнительную категорию, к которой относились следующие веб-страницы:

– страницы с некорректными кодировками;
– страницы, которые открывались, но внутри текста сообщали об ошибках;
– пустые страницы с технической информацией текущего хостинга;
– страницы с недостаточной информацией для определения категории.

C. Разметка данных

Для определения качества тематической классификации на основе больших языковых моделей была проведена ручная разметка сохраненных сайтов хостинга. Для разметки была отобрана случайная выборка из 400 сайтов исходного датасета, собранного в рамках проекта. При общем размере в 9 466 сайтов такой объем обеспечивает допустимую погрешность в пределах 5% при доверительном интервале 95%.

Разметка выполнялась двумя аннотаторами независимо. Аннотаторы опирались на текст страницы и присваивали ей произвольное число категорий рубрикатора. Согласованность аннотаторов измерялась с помощью альфы Криппендорфа с поправкой на дистанцию согласованности множеств (MASI – Measuring Agreement on Set-valued Items) [16] и коэффициента Жаккара. Значения после первого этапа разметки составили: альфа Криппендорфа – 0,35; коэффициент Жаккара – 0,51. Низкий уровень согласованности можно объяснить следующими факторами: большое количество категорий (28) для разметки, многозначность текстов сайтов, а также уровень строгости дистанции MASI, которая в большей степени, чем коэффициент Жаккара, штрафует за единичные различия в множествах меток между аннотаторами. В связи с низким уровнем согласованности противоречия были разрешены в ходе совместного обсуждения между аннотаторами, в результате чего для выборки была достигнута полная согласованность.

D. Многометочная классификация на основе больших языковых моделей

Тематическая классификация веб-страниц выборки включала следующие этапы. Тексты главных страниц случайным образом объединялись в батчи по 10 страниц. Каждой из моделей передавался промпт (см. Приложение 2), в котором модель просили определить все релевантные категории для каждой страницы и вывести результат в формате JSON. Вместе с промптом в JSON-формате передавались сам рубрикатор с подкатегориями и батч с текстами веб-страниц.

Итоговый промпт отправлялся большим языковым моделям через API[9]. Процесс отправки был распараллелен для ускорения получения ответов моделей. В параметрах вызова задавалась температура, равная 0, для максимальной детерминированности. Количество выходных токенов ограничивалось 2 500 с целью снижения стоимости обращений к рассуждающим моделям.

Полученные ответы моделей проверялись на корректность структуры JSON и соответствие числа ответов количеству переданных документов. В случае ошибок модели, включая выход за рамки лимита выходных токенов и «галлюцинации», проблемные батчи отправлялись на повторную итерацию запроса к большой языковой модели. Для такой повторной итерации максимум выходных токенов увеличивался на 1 000. При повторных ошибках выполнялись еще две дополнительные итерации на необработанных батчах, для каждой из которых лимит увеличивался еще на 1 000. Если после трех повторных проходов оставались нерешенные батчи, соответствующие тексты фиксировались как отсутствующие предсказания для данной модели.

Полученные результаты моделей оценивались при помощи следующих метрик: точность, полнота, F1-мера, мера Samples-F1 – подсчет F1-меры для каждого объекта по множеству его меток, Subset-точность – точность по полному совпадению множеств меток объекта, доля ложных меток – меток, отсутствующих в рубрикаторе, количество повторных итераций для модели (от 0 до 3), доля отсутствующих текстов сайтов во всей выборке.

V. Результаты

E. Разнообразие тем веб-сайтов

Рис. 1 демонстрирует распределение 400 главных страниц размеченной выборки по категориям рубрикатора на основе ручной разметки. До роста популярности социальных сетей бесплатные хостинги закрывали потребность пользователей в онлайн-общении. Личные сайты и блоги лидируют среди ресурсов «Narod.ru», что подтверждается первой позицией категории «Онлайн-сообщества». С небольшим отрывом на втором месте расположена рубрика «Бизнес и промышленность»: простота создания сайтов на хостинге «Narod.ru» способствовала появлению множества страниц компаний и продавцов услуг – от сборки и доставки мебели до продажи щенков породистых собак. Категория «Работа и образование» объединяет большое число сайтов школ и детских садов, а также персональные страницы преподавателей с материалами по обучению и портфолио для поиска работы. Популярность рубрики «Источники и справочные материалы» обусловлена большим количеством ресурсов с инструкциями и руководствами (Do It Yourself) и ресурсов справочного, энциклопедического характера, а также площадок об изучении иностранных языков. Тематика «Искусство и развлечение» в основном представлена сервисами для скачивания музыки, фильмов или сериалов и фан-клубами.

[9] OpenRouter. URL: https://openrouter.ai. (дата обращения: 25.10.2025).

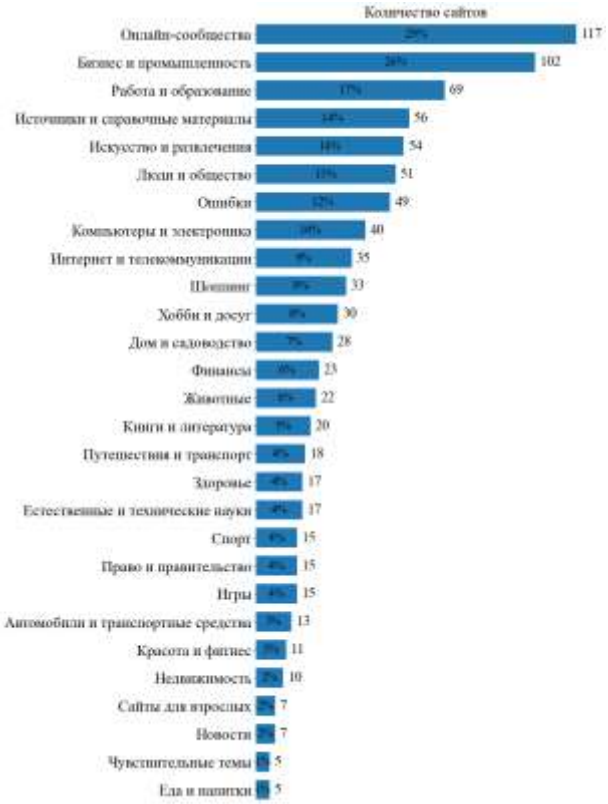

Рисунок 1. Результаты ручной разметки выборки.

Отдельно отметим значительную долю (12%) страниц, отнесенных аннотаторами к сайтам с ошибками и к случаям с недостаточной информацией для определения категории. Большинство подобных сайтов либо не содержат ничего, кроме стандартной схемы сайта хостинга, либо являются нечитаемыми. Показательно, что за последние полтора года хостинг «uCoz» убрал из доступа 32 из 49 таких страниц.

На рис. 2 показано распределение количества категорий на сайт. Большинство ресурсов отнесено к одной, двум или трем категориям. Среди ресурсов с единственной тематикой большинство (49/111) составляют сайты с ошибками, далее следуют «Онлайн-сообщества» с личными дневниками и «Бизнес и промышленность» с акцентом на услуги в сельскохозяйственной сфере и тяжелой промышленности. Для сайтов с двумя категориями наиболее характерны пары [«Бизнес и промышленность» и «Дом и садоводство»] – продажа мебели и техники для дома, [«Искусство и развлечения» и «Онлайн-сообщества»], где пользователям предлагают скачать фильмы, сериалы или музыку, а также страницы государственных образовательных учреждений с метками [«Работа и образование» и «Люди и общество»]. Среди сайтов с тремя категориями выделяются персональные сайты преподавателей технических дисциплин, хранилища программ и руководств по программированию, а также интернет-магазины косметических изделий.

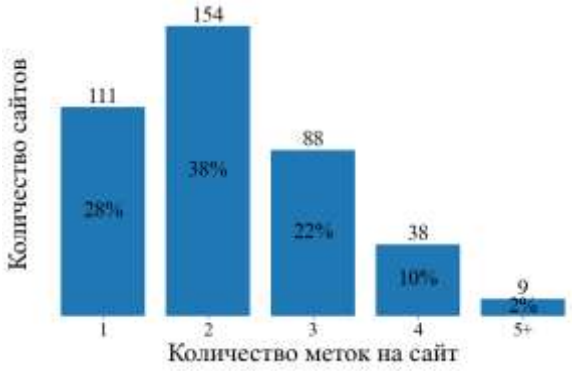

Рисунок 2. Распределение количества категорий на сайт

F. Сравнительный анализ моделей

Таблица 1 представляет оценки качества тематической классификации датасета большими языковыми моделями. Значения отсортированы по метрике Samples-F1. Наилучший результат показала модель gemini-2.5-pro (Samples-F1=0,708).

Таблица 1. Результаты тематической классификации

Модель	subset_acc	macro_f1	samples_f1
gemini-2.5-pro	**0,338**	**0,683**	**0,708**
gpt-5	0,288	0,632	0,653
grok-4	0,282	0,638	0,645
qwen3-max	0,272	0,593	0,625
gemini-2.5-flash	0,262	0,632	0,621
grok-4-fast	0,228	0,601	0,606
claude-sonnet-4.5	0,265	0,589	0,605
deepseek-r1-0528	0,248	0,588	0,601
deepseek-chat-v3.1	0,252	0,570	0,589
deepseek-v3.1	0,255	0,558	0,589
qwen3-235b-a22b	0,275	0,538	0,581
gpt-oss-120b	0,222	0,549	0,574
kimi-k2-0905	0,238	0,538	0,571
qwen3-235b-a22b-th	0,202	0,550	0,569
gpt-oss-20b	0,200	0,552	0,542
qwen3-next-in	0,205	0,480	0,506
qwen3-next-th	0,148	0,401	0,407

В Приложении 3 приведена полная информация по метрикам качества моделей. Наиболее низкие показатели наблюдаются по полноте (recall), что указывает на недостаточное покрытие всех категорий веб-сайтов. Распределение gemini-2.5-pro, в отличие от остальных моделей, близко к статистике аннотаторов на рис. 2. При этом все модели в большинстве случаев присваивают сайту только одну категорию, что отличается от размеченного вручную датасета (см. Приложение 4). Лишь три модели – claude-sonnet-4.5, gemini-2.5-pro и qwen3-max – присваивают две категории более чем для 100 сайтов (для сравнения, при ручной разметке 2 метки были присвоены в 154 случаях).

Также фиксируется заметный разрыв между проприетарными моделями и моделями с открытым

исходным кодом. Проприетарные модели превосходят открытые модели по всем метрикам.

G. Вызовы и будущие исследования

Низкая эффективность моделей с открытым исходным кодом для задачи классификации сайтов хостинга «Narod.ru» потенциально ставит исследователей в зависимость от проприетарных решений. Главным фаворитом среди протестированных моделей является gemini-2.5-pro. Существенным ограничением этой модели остается стоимость: 1,25 долл. США за 1 млн входных токенов и 10 долл. США за 1 млн выходных токенов. По оценкам для 1,2 млн главных страниц «Narod.ru» количество входных токенов составляет около 2,34 млрд, выходных – около 339 млн. Соответственно, обработка только главных страниц всех сайтов обойдется более чем в 500 тыс. рублей, что является определенным ограничением для исследований.

При необходимости, потенциальной альтернативой проприетарным моделям могут стать открытые модели, однако требуется иная методика работы: качество открытых моделей при обработке документов батчами недостаточно высокое. В данном случае можно попытаться повысить качество моделей, отправляя в каждом промпте один документ. Однако такой подход может существенно увеличить время обработки всего датасета. Другой возможной альтернативой может быть схема из N независимых бинарных классификаторов на базе открытой модели, где N совпадает с числом тематик рубрикатора. В таком подходе каждая модель определяет принадлежность сайта к одному классу, что потенциально может понизить сложность исходной задачи многометочной классификации.

Следует отметить, что проект сталкивается с некоторыми другими ограничениями. Наиболее серьезным является невозможность предоставления собранного датасета в открытый доступ из-за того, что часть сохраненных ресурсов может содержать нежелательный контент или контент, запрещенный законодательством Российской Федерации. Более точная разметка данных в категориях «Сайты для взрослых» и «Чувствительные темы» мог бы помочь с выявлением сайтов, содержащих потенциально проблемные материалы.

Тем не менее, тематическая классификация всех сохраненных сайтов позволит точнее описать их содержание и привлечь исследователей из разных областей для совместной работы над проектом, что в свою очередь в будущем может помочь сохранению большего числа сайтов и созданию на их базе удобного поискового исследовательского инструмента для публичного доступа.

VI. Заключение

В данной работе предложен подход к тематической классификации сайтов хостинга «Narod.ru» на основе больших языковых моделей. Выполнена ручная разметка выборки главных страниц сайтов, сохраненных в рамках проекта по архивации цифрового культурного наследия хостинга.

При обработке выборки из 400 главных страниц хостинга наилучшие результаты показала модель gemini-2.5-pro (Samples-F1=0,708). В среднем модели продемонстрировали относительно низкий уровень метрики полноты на исследуемой выборке: они присваивали сайтам меньшее количество тематических меток по сравнению с ручной разметкой. Проприетарные модели справились с задачей тематической классификации лучше открытых: наилучшая открытая модель deepseek-r1 уступила по качеству наихудшей из проприетарных – claude-sonnet-4.5.

Классификация всего корпуса сайтов хостинга наилучшей моделью может потребовать существенных ресурсов. Оценочная стоимость классификации более 1,2 млн сайтов «Narod.ru» при помощи gemini-2.5-pro превышает 500 000 рублей.

При необходимости снижения затрат, в качестве альтернативы проприетарным моделям, можно пересмотреть методологию тематической классификации для применения открытых моделей. Один из вариантов – сократить число документов, подаваемых в промпте, до одного, однако это может негативно сказаться на скорости обработки всего корпуса. Другой вариант – упростить задачу за счет использования нескольких открытых моделей в роли бинарных классификаторов.

В целом, предложенный подход к тематической классификации сайтов раннего интернета позволит исследователям выявлять ключевые направления развития ранней сетевой культуры, интересы пользователей и динамику общественных дискурсов. Это создаст основу для анализа процессов формирования цифровой идентичности, онлайн-сообществ и эволюции коммуникационных практик в зарождающемся интернет-пространстве начала XXI века.

Библиография

[1] Agarwal S. et al. Gpt-oss-120b & gpt-oss-20b Model Card // arXiv, 2025. URL: https://arxiv.org/abs/2508.10925 (дата обращения: 25.10.2025).

[2] Anthropic. Claude Sonnet 4.5 System Card // Anthropic. 2025. URL: https://www.anthropic.com/claude-sonnet-4-5-system-card (дата обращения: 25.10.2025)

[3] Bai Y. [et al.] Kimi K2: Open Agentic Intelligence // arXiv, 2025. URL: https://arxiv.org/abs/2507.20534 (дата обращения: 25.10.2025)

[4] Bergman M. The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing, 2001.

[5] Chapekis A. et al. When Online Content Disappears // Pew Research Center. 17 May. 2024. URL: https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/ (дата обращения: 25.10.2025).

[6] Charter on the Preservation of Digital Heritage. // UNESCO. 15 October 2003. URL: https://www.unesco.org/en/legal-affairs/charter-preservation-digital-heritage (дата обращения: 25.10.2025).

[7] Comanici G. [et al.] Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next-Generation Agentic Capabilities // arXiv, 2025. URL: https://arxiv.org/abs/2507.06261 (дата обращения: 25.10.2025)

[8] Digital Folklore: to computer users, with love and respect / eds. O. Lialina et al. Stuttgart: Merz & Solitude, 2009. 286 p.

[9] Guo D. et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning // arXiv, 2025. URL: https://arxiv.org/abs/2501.12948 (дата обращения: 25.10.2025).

[10] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv, 2022. URL: https://arxiv.org/abs/2203.05794 (дата обращения: 25.10.2025).
[11] Haslop, Blaire, Marc Aurel Schnabel, and Serdar Aydin. “Digital Decay.” Parallelism in Architecture, Environment And Computing Techniques (PACT), 2016.
[12] Lialina O., Espenschied D. One terabyte of kilobyte age. // Tumblr. URL: https://oneterabyteofkilobyteage.tumblr.com/ (дата обращения: 25.10.2025).
[13] Lialina O. Ruins and Templates of Geocities // Still there. – URL: https://contemporary-home-computing.org/still-there/geocities.html (дата обращения: 25.10.2025).
[14] Liu A. et al. DeepSeek-V3 Technical Report // arXiv, 2024. URL: https://arxiv.org/abs/2412.19437 (дата обращения: 25.10.2025).
[15] OpenAI. GPT-5 System Card // OpenAI. 2025. URL: https://cdn.openai.com/gpt-5-system-card.pdf (дата обращения: 25.10.2025).
[16] Passonneau R. Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation // Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06) Genoa, Italy. 2006. URL: https://aclanthology.org/L06-1392/ (дата обращения: 25.10.2025).
[17] Vijgen R. The Deleted City: A Digital Archaeology // Parsons Journal for Information Mapping. URL: http://piim.newschool.edu/journal/issues/2013/02/pdfs/ParsonsJournalForInformationMapping_Vijgen_Richard.pdf (дата обращения: 25.10.2025).
[18] Yang A. et al. Qwen3 Technical Report // arXiv, 2025. URL: https://arxiv.org/abs/2505.09388 (дата обращения: 25.10.2025)
[19] xAI. Grok 4 Model Card // xAI. 2025. URL: https://data.x.ai/2025-08-20-grok-4-model-card.pdf (дата обращения: 25.10.2025).
[20] xAI. Grok 4 Fast Model Card // xAI. 2025. URL: https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf (дата обращения: 25.10.2025).
[21] Козлова А.С., Асланов И.В., Бибилов, И.В. Котельников Е.В. Сохранение сайтов раннего интернета для междисциплинарных исследований на примере сайтов хостинга «Narod.ru» (2000–2013) // Информационное общество: образование, наука, культура и технологии будущего. Вып. 9. Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23–25 июня 2025 г. СПб: Университет ИТМО, 2025 (в печати).

Приложения

Приложение 1. Рубрикатор тематических категорий

1. Сайты для взрослых (Adult): ресурсы с сексуальным контентом, включая эротику, порнографию, NSFW-индустрию, сайты с товарами для взрослых и предложениями интим-услуг.
2. Искусство и развлечения (Arts & Entertainment): ресурсы о кино и ТВ, музыке и подкастах, онлайн-медиа и стримингах, сценических и изящных искусствах, комиксах и анимации, юморе.
3. Автомобили и транспортные средства (Autos & Vehicles): все виды транспорта – от велосипедов и автомобилей до лодок и мотоциклов; покупку и сравнение транспортных средств, обслуживание и запчасти, тюнинг, ПДД.
4. Красота и фитнес (Beauty & Fitness): услуги красоты и спа, косметика и парфюмерия, уход за кожей и волосами, мода, фитнес, тренировки и оборудование для них, программы для похудения.
5. Книги и литература (Books & Literature): книги в разных форматах (бумажные, электронные, аудио), жанры от классики и поэзии до детской литературы и фанфиков, книжные магазины и ресурсы для писателей.
6. Бизнес и промышленность (Business & Industrial): бизнес и B2B-сервисы, маркетинг и финансы, производство и строительство, энергетика и химическая промышленность, логистика и ритейл, сельское хозяйство и промышленное оборудование.
7. Компьютеры и электроника (Computers & Electronics): компьютеры и периферия, потребительская электроника, корпоративные ИТ и сети, информационная безопасность, программирование и ПО.
8. Финансы (Finance): банковские услуги и платежи, кредиты и страхование, инвестирование, бухгалтерские услуги и финансовое планирование.
9. Еда и напитки (Food & Drink): еда и напитки (от рецептов и кухонь мира до здорового питания), розничная торговля и доставка продуктов, рестораны, фастфуд, кейтеринг.
10. Игры (Games): видеоигры и их жанры, настольные и карточные игры, азартные игры (покер, ставки, лотереи), киберспорт, разработка и магазины игр.
11. Здоровье (Health): сведения о заболеваниях и лечении, медицинские услуги и оборудование, здоровье мужчин/женщин, психическое здоровье, лекарства и общественное здравоохранении.
12. Хобби и досуг (Hobbies & Leisure): увлечения и досуг – сообщества и организации, рукоделие и моделизм, активный отдых на природе, праздники, свадьбы и другие особые события.
13. Дом и садоводство (Home & Garden): товары и услуги для дома и участка – интерьер и мебель, техника и уборка, ремонт и безопасность, системы кондиционирования и сантехника, сад, ландшафт и борьба с садовыми вредителями.
14. Интернет и телекоммуникации (Internet & Telecom): связь и Интернет-услуги, провайдеры телекоммуникационных услуг, мобильные устройства и приложения, электронная почта и мессенджеры, поисковые системы, а также веб-сервисы– облачные сервисы, SEO/маркетинг, веб-разработка.
15. Работа и образование (Jobs & Education): обучение всех форматов и уровней – курсы, школы, вузы, материалы и сертификаты; трудоустройство – вакансии, стажировки, ресурсы для карьеры, резюме, портфолио.
16. Право и правительство (Law & Government): государственные институты и политика, право и судебная система, оборона, правоохранительные органы, общественная безопасность и социальные службы.
17. Новости (News): средства массовой информации – мировые, политические, бизнес-, технологические, спортивные и локальные новости.
18. Онлайн-сообщества (Online Communities): платформы для общения и самовыражения онлайн, блоги и персональные сайты, социальные сети, дейтинг, обмен фото, видео и файлами.
19. Люди и общество (People & Society): семья и отношения, религия и саморазвитие, социальные проблемы и активизм, права человека, экология, социальные науки, субкультуры.
20. Животные (Pets & Animals): домашние животные и дикая природа, уход за животными, ветеринарная помощь, корма и товары для животных, ресурсы по видам животных, защита животных.
21. Недвижимость (Real Estate): рынки и услуги недвижимости, покупка/продажа и аренда жилой и коммерческой недвижимости, инспекция и оценка недвижимости, управление объектами недвижимости, агентские и эскроу-сервисы.
22. Источники и справочные материалы (Reference): справочные ресурсы и инструменты, словари и энциклопедии, карты и календари, образовательные и технические справочники, Do It Yourself культура, шаблоны, калькуляторы, гуманитарные науки (история, философия), языковые ресурсы.
23. Естественные и технические науки (Science): фундаментальные и прикладные науки – астрономия, биология, химия, физика, математика и статистика, информатика и компьютерные науки, робототехника.
24. Чувствительные темы (Sensitive Subjects): сложные и потенциально травмирующие темы – аварии и трагедии, смерть, огнестрельное оружие, наркотики, self-harm и суицид, насилие и военные конфликты.
25. Шоппинг (Shopping): онлайн- и офлайн-покупки, одежда и обувь, подарки и цветы, игрушки, товары класса люкс, обзоры и сравнение цен, скидки и купоны, фотосъемка, торговые порталы.
26. Спорт (Sports): командные и индивидуальные спортивные дисциплины, соревнования, инвентарь, тренировки, фан-атрибутика и сувенирная продукция, водные, зимние и экстремальные виды спорта.
27. Путешествия и транспорт (Travel & Transportation): туры, путевки и размещение путешественников, направления путешествий, экскурсии, бронирование и сервисы для путешествий, транспорт – авиаперелеты, поезда и автобусы, прокат автомобилей и велосипедов, такси и городской транспорт.
28. Ошибки (категория добавлена авторами исследования): страницы с ошибками, некорректной кодировкой, недостатком данных, из-за чего их нельзя корректно классифицировать, страницы вне рамок рубрикатора.

Приложение 2. Промпт для обработки главных страниц сайтов большими языковыми моделями

```
## Роль и цель

Ты — строгий классификатор веб-сайтов. Твоя задача — выполнить МНОГОМЕТОЧНУЮ (multi-label) классификацию текстов вебсайтов по заданному тематическому словарю.
Выводишь *только JSON*, соответствующий требуемой схеме. Никаких комментариев, объяснений или дополнительного текста.

## Входные данные
1. topics — JSON-словарь тематик.
   Каждая тема содержит:
   - id — строковый идентификатор темы (его нужно вернуть в labels);
   - subcategories — список подтематик (используются только для понимания смысла, не возвращаются).
2. input_texts — список объектов вида:
   {{
     "text_1": "text",
     "text_2": "text",
     ...
   }}

Гарантируется наличие служебной темы NonApplicablePages с id: "non_applicable" (описана в конце словаря).

## Жёсткие ограничения
- Возвращай только id из topics. Новых меток не придумывай.
- Работай в режиме строгого JSON-ответа (никаких, объяснений, YAML, Markdown и т.д.).
- Порядок результатов в results должен строго соответствовать порядку во входном input_texts.
- В labels — от 1 до нескольких идентификаторов. Если ни одна тема не подходит или текст непригоден, верни единственную метку "non_applicable".

## Предобработка (обязательна перед классификацией)
При классификации игнорируй следующий шум:
- Игнорируй дефолтные заголовки бесплатных хостингов вроде «Персональный сайт — Главная» — не используй их для определения «личного сайта».
```

- Игнорируй рекламу и вставки про uCoz, если они не относятся к тематике сайта. В частности, не учитывай фразы вида:
«Сайт создан в системе uCoz», «Бесплатный конструктор сайтов — uCoz», «Сайт управляется системой uCoz», «Хостинг от uCoz», «Powered by uCoz», «uCoz templates» и подобные.
- Если текст состоит почти целиком из такого шума — считай, что тематической информации нет.

Критерии непригодности текста (non_applicable)
Верни "non_applicable", если выполняется любой из пунктов:
- Текст нечитабелен (битая кодировка/«кракозябры») или крайне мал.
- Текст целиком про хостинг/движок (uCoz и пр.) и не даёт тематической информации о сайте.
- Ни одна тема из словаря обоснованно не подходит.
В случае non_applicable не добавляй другие метки.

Алгоритм выбора меток (multi-label)
- Сопоставление по смыслу: соотнеси главный контент текста с описаниями/примером области в subcategories каждой темы.
- Специфичность > общность: выбирай темы, смысл которых прямо отражён в тексте.
- Доказательность: добавляй тему, если есть индикаторы в пользу темы или термин/фраза, относящие к теме (пример: «афиша концертов», «интернет-магазин запчастей», «рецепты выпечки»).
- Множественность: если текст охватывает несколько тематик, добавь все релевантные темы. Убедись, что ты добавил все подходящие темы.
- Возвращай только id тем из корневого списка topics (подтемы из subcategories служат лишь для ориентира).

Формат ответа (строго)
Верни строго JSON следующего вида и ничего больше:
```
{{
  "text_1": ["<topic-id-1>", "<topic-id-2>"], // лейблы для text_1
  "text_2": ["<topic-id-1>"], // лейблы для text_2
}}
```

Требования к выводу:
* Ключи (text_id) в выходе совпадают с ключами (text_id) из input_texts JSON.
* Количество записей в выходе равно количеству записей из input_texts JSON во входе.
* Валидный JSON: двойные кавычки вокруг ключей и значений, без лишних запятых и текста до/после.

Мини-пример
Вход:
```
{{
  42: "Афиша концертов в Москве, обзоры альбомов и интервью с музыкантами.",
  87: "Персональный сайт — Главная. Сайт создан в системе uCoz. Хостинг от uCoz."
}}
```

Возможный выход:
```
{{
  42: ["arts_entertainment"],
  87: ["non_applicable"]
}}
```

topics:
{TOPICS_JSON}

input_texts:
{INPUT_TEXTS_JSON}

ПРИЛОЖЕНИЕ 3. МЕТРИКИ КАЧЕСТВА БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ЗАДАЧИ МНОГОМЕТОЧНОЙ КЛАССИФИКАЦИИ

Модель	subset accuracy	micro			macro			weighted			samples_f1	unknown prediction rate	retries	missing docs
		precision	recall	f1	precision	recall	f1	precision	recall	f1				
gemini-2.5-pro	0,338	0,745	0,657	0,698	0,711	0,679	0,683	0,754	0,657	0,694	0,708	0	1	0
gpt-5	0,288	0,834	0,494	0,621	0,812	0,550	0,632	0,834	0,494	0,594	0,653	0	0	0
grok-4	0,282	0,829	0,494	0,619	0,835	0,548	0,638	0,832	0,494	0,601	0,645	0	0	0
qwen3-max	0,272	0,761	0,490	0,596	0,744	0,526	0,593	0,797	0,490	0,573	0,625	0,007	0	0
gemini-2.5-flash	0,262	0,765	0,493	0,600	0,798	0,554	0,632	0,799	0,493	0,585	0,621	0	1	0
grok-4-fast	0,228	0,848	0,437	0,577	0,857	0,510	0,601	0,870	0,437	0,533	0,606	0	0	0
claude-sonnet-4.5	0,265	0,788	0,467	0,587	0,792	0,508	0,589	0,810	0,467	0,563	0,605	0	3	0,025
deepseek-r1-0528	0,248	0,796	0,441	0,568	0,805	0,498	0,588	0,817	0,441	0,539	0,601	0	2	0
deepseek-chat-v3.1	0,252	0,783	0,429	0,554	0,837	0,479	0,570	0,844	0,429	0,524	0,589	0	0	0
deepseek-v3.1	0,255	0,776	0,430	0,553	0,789	0,471	0,558	0,816	0,430	0,526	0,589	0	0	0
qwen3-235b-a22b	0,275	0,760	0,426	0,546	0,781	0,458	0,538	0,827	0,426	0,511	0,581	0,002	0	0
gpt-oss-120b	0,222	0,787	0,422	0,549	0,726	0,479	0,549	0,783	0,422	0,514	0,574	0,004	1	0
kimi-k2-0905	0,238	0,779	0,399	0,528	0,793	0,449	0,538	0,829	0,399	0,493	0,571	0,002	3	0,002
qwen3-235b-a22b-th	0,202	0,839	0,394	0,536	0,816	0,455	0,550	0,843	0,394	0,497	0,569	0	3	0,025
gpt-oss-20b	0,200	0,783	0,380	0,512	0,804	0,445	0,552	0,772	0,380	0,488	0,542	0	2	0
qwen3-next-in	0,205	0,700	0,351	0,467	0,746	0,405	0,480	0,774	0,351	0,424	0,506	0,009	2	0
qwen3-next-th	0,148	0,818	0,281	0,418	0,756	0,293	0,401	0,822	0,281	0,395	0,407	0,003	3	0,3

Subset accuracy – точность по полному совпадению множеств меток объекта.
Samples F1 – F1-мера для каждого объекта по множеству его меток (усреднено по всем объектам).
Unknown prediction rate – доля меток, отсутствующих в рубрикаторе, от общего числа меток.
Retries – количество повторных проходов модели по датасету (от 0 до 3).
Missing docs – доля объектов, для которых модель не дала ответ, от всех объектов выборки.

ПРИЛОЖЕНИЕ 4. КОЛИЧЕСТВО МЕТОК НА СТРАНИЦУ

Модели	0	1	2	3	4	5+
claude-sonnet-4.5	10	264	120	5	0	1
gemini-2.5-pro	0	149	148	84	12	7
gemini-2.5-flash	0	271	98	23	7	1
deepseek-r1-0528	0	318	76	4	2	0
deepseek-chat-v3.1	0	319	78	3	0	0
deepseek-v3.1	0	316	78	6	0	0
gpt-5	0	293	94	9	4	0
gpt-oss-120b	0	333	63	3	0	1
gpt-oss-20b	0	375	22	2	1	0
qwen3-next-in	0	363	33	3	0	1
qwen3-next-th	120	264	13	2	0	1
qwen3-235b-a22b-th	10	367	21	2	0	0
qwen3-235b-a22b	0	310	84	6	0	0
qwen3-max	0	255	127	13	4	1
grok-4-fast	0	351	44	4	1	0
grok-4	0	291	95	11	2	1
kimi-k2-0905	1	348	48	3	0	0

Thematic classification of Narod.ru hosting websites as part of the strategy for preserving early Internet sites

Ilias Aslanov, Anna Kozlova, Ivan Bibilov, Evgeny Kotelnikov

***Abstract*—The study focuses on the preservation and analysis of websites hosted on "Narod.ru," an active web-hosting platform during 2000–2013. Within this work, the hosted websites are considered as disappearing objects of digital heritage, whose preservation and examination may be of interest to experts from various fields, particularly cultural scholars and researchers of early internet digital folklore. The study proposes an approach to thematic classification of websites using large language models. First, a manual annotation of a sample of archived hosting websites was conducted, where main website pages were assigned thematic categories according to Google's taxonomy. Then, based on the annotated sample, the performance of 17 proprietary and open large language models was evaluated for the task of multi-label thematic classification of web pages. The best result was achieved by the *gemini-2.5-pro* model (Samples F1 = 0.708). The proposed approach to thematic classification of early internet websites enables researchers to identify and analyze cultural, social, and communicative patterns in the formation of digital society.**

***Keywords*—web archiving, digital heritage, Narod.ru, large language models, multi-label classification.**

REFERENECES

[1] S. Agarwal et al., "Gpt-oss-120b & gpt-oss-20b Model Card," *arXiv*, 2025, Available: https://arxiv.org/abs/2508.10925.

[2] "Anthropic. Claude Sonnet 4.5 System Card," *Anthropic*, 2025. Available: https://www.anthropic.com/claude-sonnet-4-5-system-card

[3] Y. Bai et al., "Kimi K2: Open Agentic Intelligence," *arXiv*, 2025. Available: https://arxiv.org/abs/2507.20534

[4] M. Bergman, "The Deep Web: Surfacing Hidden Value," *Journal of Electronic Publishing*, 2001.

[5] A Chapekis. et al., "When Online Content Disappears," *Pew Research Center*, 17 May. 2024. Available: https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/

[6] *Charter on the Preservation of Digital Heritage*, UNESCO, 15 October 2003. Available: https://www.unesco.org/en/legal-affairs/charter-preservation-digital-heritage

[7] G. Comanic. [et al.], "Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next-Generation Agentic Capabilities," *arXiv*, 2025. Available: https://arxiv.org/abs/2507.06261

[8] *Digital Folklore: to computer users, with love and respect*, eds. O. Lialina et al., Stuttgar, Merz & Solitude, 286 p., 2009.

[9] D. Guo et al., "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," *arXiv*, 2025. Available: https://arxiv.org/abs/2501.12948

[10] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv, 2022. Available: https://arxiv.org/abs/2203.05794

[11] B. Haslop, M. A. Schnabel and S. Aydin, *"Digital Decay." Parallelism in Architecture, Environment And Computing Techniques (PACT)*, 2016.

[12] O. Lialina and D. Espenschied, "One terabyte of kilobyte age," *Tumblr*, Available: https://oneterabyteofkilobyteage.tumblr.com/

[13] O. Lialina, "Ruins and Templates of Geocities," *Still there*, Available: https://contemporary-home-computing.org/still-there/geocities.html

[14] A. Liu et al., "DeepSeek-V3 Technical Report," *arXiv*, 2024. Available: https://arxiv.org/abs/2412.19437

[15] OpenAI. GPT-5 System Card, *OpenAI*, 2025. Available: https://cdn.openai.com/gpt-5-system-card.pdf

[16] R. Passonneau, "Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation," In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, 2006. Available: https://aclanthology.org/L06-1392/

[17] R. Vijgen, "The Deleted City: A Digital Archaeology," *Parsons Journal for Information Mapping*. Available: http://piim.newschool.edu/journal/issues/2013/02/pdfs/ParsonsJournalForInformationMapping_Vijgen_Richard.pdf

[18] A. Yang et al., "Qen3 Technical Report," *arXiv*, 2025. Available: https://arxiv.org/abs/2505.09388

[19] xAI. Grok 4 Model Card, *xAI*, 2025. Available: https://data.x.ai/2025-08-20-grok-4-model-card.pdf

[20] xAI. Grok 4 Fast Model Card, *xAI*, 2025. Available: https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf

[21] A. Kozlova, I. Aslanov, I. Bibilov and E Kotelnikov, "Preserving Early Internet Websites for Interdisciplinary Research: A Case Study of the "Narod.ru" Hosting Platform (2000–2013)," In *Information Society: Education, Science, Culture, and Technologies of the Future. Issue 9. Proceedings of the XXVIII International Joint Scientific Conference "Internet and Modern Society,"* IMS-2025, St. Petersburg, June 23–25, 2025, St. Petersburg, ITMO University, 2025 (in print).