УДК 930.23
doi 10.17072/2219-3111-2025-2-147-158
EDN: DEDUOH
ASJC 1202
ГРНТИ 03.81.37
Ссылка для цитирования: *Галушко И. Н.* Классификация исторических документов по политическому признаку с помощью BERT: взаимодействие LLM с историческим доменом // Вестник Пермского университета. История. 2025. № 2(69). С. 147–158. DOI: 10.17072/2219-3111-2025-2-147-158. EDN: DEDUOH

КЛАССИФИКАЦИЯ ИСТОРИЧЕСКИХ ДОКУМЕНТОВ ПО ПОЛИТИЧЕСКОМУ ПРИЗНАКУ С ПОМОЩЬЮ BERT: ВЗАИМОДЕЙСТВИЕ LLM С ИСТОРИЧЕСКИМ ДОМЕНОМ[1]

И. Н. Галушко

Московский государственный университет имени М. В. Ломоносова, 119192, Россия, Москва, Ломоносовский пр-т, 27/4, Г-423;
Национальный исследовательский университет «Высшая школа экономики», 109028, Москва, Покровский б-р, 11, корп. S, комн. S938
i.galushko15@gmail.com
ResearcherID: LYP-0220-2024
SPIN-код: 3433-9690

На современном этапе изучения отечественной истории становятся особенно актуальны дискуссии о работе с большими массивами документов по истории конца XIX – начала XXI в. Сегодня активно идет процесс оцифровки архивных коллекций, но в большинстве случаев созданный корпус просто выставляется на сайт, и многие годы с ним никто не работает, так как нередко мы сталкиваемся с трудностями обработки всего массива материалов при обращении к фондам крупного социального института. Оцифрованные фонды могут содержать сотни тысяч листов делопроизводственной документации. Ограниченность временных возможностей одного человека не позволяет даже на уровне беглого чтения охватить все имеющиеся документы. Данную проблему хотя бы частично может решить применение LLM (large language models) для аннотирования или оптимизации текстового поиска. Однако на текущем этапе развития архивного дела специалисты только начинают работать с методами обработки естественного языка. И основной запрос профессионального сообщества состоит в изучении специфики работы моделей искусственного интеллекта и машинного обучения с текстами исторического домена. Данная работа представляет собой предварительное исследование взаимодействия современных LLM с историческими текстами. Для анализа были выбраны одна из наиболее популярных моделей – BERT – и одна из наиболее распространенных NLP-задач – классификация. Важной частью исследования стал анализ весов внимания модели при решении задачи классификации текстов и заполнения пропусков в предложениях. При таком подходе у нас появилась возможность проанализировать, как модель использует семантический контекст для принятия решения.

Ключевые слова: классификация текстов, политическая история, искусственный интеллект, анализ механизмов внимания, машинное обучение, BERT, NLP.

Введение

Сегодня в исследовательском сообществе активно развивается дискуссия о принципах работы с большими коллекциями документов по истории конца XIX – начала XXI в. [*Володин*, 2020]. Активно идет процесс оцифровки имеющихся коллекций, но нередко созданный корпус оказывается просто выставленным на сайт, и многие годы с ним никто не работает. Дело в том, что, изучая один аспект (например, экономические отношения СССР со странами Восточного блока), историк сталкивается с трудностями при обращении к материалам крупного института общесоюзного значения (например, к многочисленным архивным фондам Госплана СССР в

Российском государственном архиве экономики (РГАЭ)). Материалы фондов могут содержать сотни тысяч листов делопроизводственной документации. Ограниченность времени одного исследователя не позволяет даже на уровне беглого чтения охватить все имеющиеся документы. Данную проблему хотя бы частично могло бы решить качественное аннотирование, но для этого требуются большие человеческие ресурсы десятков сотрудников архива, и на данный момент во многих архивах такая работа ведется ограниченно. Здесь естественно возникает идея об использовании генеративных нейронных сетей для автоматической генерации аннотаций к архивным коллекциям: к какому разделу истории относится данный материал? какие *entities* там упоминаются? к какому блоку современных проблематик можно отнести данный документ/данную коллекцию? Однако на текущем этапе развития архивного дела специалисты только начинают работать с методами обработки естественного языка. Профессиональное сообщество проявляет повышенный интерес к исследованию особенностей обработки исторических текстов методами искусственного интеллекта [*Юмашева*, 2022]. Настоящее исследование посвящено анализу взаимодействия крупных языковых моделей (LLM) с текстами исторического домена, что представляет особую сложность из-за отличий в речевой стилистике и специфического временного контекста. В качестве объекта исследования была выбрана популярная модель BERT, демонстрирующая высокие результаты в обработке естественного языка. Для тестирования ее возможностей была взята классическая NLP-задача – классификация текстов, позволяющая оценить способность модели использовать исторический контекст для оптимизации своего вывода.

На данный момент единственным апробированным в исторической науке NLP-методом для предварительной оценки информационного потенциала коллекции исторических источников является тематическое моделирование. Так, в работе *Topic Modeling in Historical Newspapers* [*Yang, Torget, Mihalcea*, 2011] был предложен концепт использования LDA для первичной проверки, какой теме посвящены конкретные номера коллекции техасских газет за 1829–2008 гг. Все это было представлено в виде общей методики облегчения рутинного труда историка, который теперь мог не вчитываться в каждый номер, а отобрать лишь те документы, темы которых соответствуют его исследовательским задачам. Большие языковые модели в задачах исторического информационного поиска на данный момент не использовались.

Фокус данной работы сосредоточен на прикладном аспекте использования языковых моделей для информационного поиска в коллекции исторических документов. Классификация в данном исследовании рассматривается как потенциальное решение для оптимизации процесса отбора источников на этапе подготовки исторического исследования: вполне возможно, что обученные языковые модели смогут взять на себя черновую работу по поиску документов, относящихся к узкоспециализированной теме. Тогда историку будет оставлена содержательная работа, связанная с источниковедческой критикой и анализом текста в глобальном историческом контексте. Кроме того, само по себе обращение к исследованию специфики взаимодействия языковых моделей с текстами исторического домена представляется интересным и продуктивным: насколько модель, обученная на «Википедии», интернет-контенте и текстах СМИ, окажется способна решать задачи обработки естественного языка применительно к историческим текстам?

Для классификации были отобраны документы эпохи революции 1917 г. и Гражданской войны в России, так как по данному периоду, во-первых, существует большое количество распознанных публикаций, доступных на сайте «Электронной библиотеки исторических документов»; во-вторых, у такой коллекции есть очевидный содержательный критерий для классификации – политическая принадлежность документа, так как рассматриваемый период отличался существенной политической дифференциацией действующих в стране социальных сил. Мы разделили документы на классы (табл. 1).

Таблица 1

Распределение текстов по классам

Группа текстов	Тэг класса	Публикации
Большевики	Bolsheviki	1. Письма во власть. 1917–1927: Заявления, жалобы, доносы, письма в государственные структуры и большевистским вождям / Ин-т гос. управления и соц. исслед. Моск. гос. ун-та им. М.В. Ломоносова, Instituto Italiano Per Gli Studi Filosofici, Ecole Des Hautes Etudes En Sciences Sociales, Maison Des Sciences De L'Homme; сост. А.Я. Лившин, И.Б. Орлов. М.: РОССПЭН, 1998. 664 с. 2. Декреты Советской власти. Т. I. 25 октября 1917 г. – 16 марта 1918 г. / Ин-т марксизма-ленинизма при ЦК КПСС, Ин-т истории акад. наук СССР. М.: Политиздат, 1957. 640 с.
Меньшевики	Mensheviki	1. Меньшевики в большевистской России. 1918–1924. Меньшевики в 1918 году. М.: РОССПЭН, 1999. 799 с. 2. Меньшевики в большевистской России. 1918–1924. Меньшевики в 1919–1920 гг. М.: РОССПЭН, 2000. 936 с. 3. Меньшевики в советской России: сб. док. / сост. В.К. Виноградов, В.И. Крылов, А.Л. Литвин, Я.Ф. Погоний, В.Н. Сафонов. Казань, 1998. 228 с.
Левые эсеры	Left_SRs	Партия левых социалистов-революционеров. Документы и материалы. 1917–1925 гг.: в 3 т. М.: РОССПЭН, 2000.
Кадеты	Kadets	Съезды и конференции конституционно-демократической партии. 1905–1920 гг.: в 3 т. М.: РОССПЭН, 2000.
Монархисты	Rights	Правые партии. 1905–1917 гг. Документы и материалы: в 2 т. Т. 2. 1911–1917 гг. М.: РОССПЭН, 1998. 816 с.
Белая армия	White_army	Журналы заседаний Особого совещания при Главнокомандующем Вооруженными Силами на Юге России А. И. Деникине. Сентябрь 1918-го – декабрь 1919 года. М.: РОССПЭН, 2008. 1003 с.
Рабочая оппозиция большевикам	Work_oppose	1. Независимое рабочее движение в 1918 году: док. и материалы / ред.-сост. и авт. коммент. М. С. Бернштам. Париж: YMCA-PRESS, 1981. 330 с. 2. Рабочее оппозиционное движение в большевистской России. 1918 г. Собрания уполномоченных фабрик и заводов: документы и материалы / сост., авт. вступ. ст. и примеч. Д. Б. Павлов. М.: РОССПЭН, 2006. 656 с.
Дворянское собрание	Nobles	Объединенное дворянство. Съезды уполномоченных губернских дворянских обществ. 1906–1916 гг.: в 3 т. М.: РОССПЭН, 2001.

Данные

Как было сказано выше, в качестве основного источника данных были использованы материалы «Электронной библиотеки исторических документов» Российского исторического общества. Данная коллекция содержит распознанные версии документов, находящихся на хранении в 1592 архивах. Цифра масштабная, но важно уточнить, что многие из представленных архивов – это личные собрания или же небольшие подборки делопроизводственной документации отдельного предприятия (от 1 до 300 ед. хранения). Здесь также представлены и крупные

государственные архивы, вроде РГАЭ (2664 ед. хранения). Для классификации нами были отобраны документы, тематически относящиеся к революционному 1917 г. и Гражданской войне. Всего – 2423 документа, распределенных на восемь классов. Из текстов были удалены номера, ссылки и указания на архивы, в которых хранятся распознанные документы.

Наиболее интересным представляется класс «Рабочая оппозиция большевикам». В эту группу вошли документы, отражающие позицию рабочих организаций, выступивших против политики большевиков в конце 1917–1918 гг. С точки зрения лексики у данной группы документов есть много пересечений как с самими большевиками, так и с меньшевиками и эсерами, так как авторами этих текстов зачастую выступали образованные рабочие, тяготевшие к социалистическим идеям, но при этом не принявшие последствия большевистских преобразований (жалобы на безработицу, насилие, цензуру). Мы предполагаем, что для этого класса мы должны получить сравнительно более низкое качество модели, так как отделить тексты рабочей оппозиции от меньшевиков гораздо сложнее, и это уже экспертная работа (если сравнивать с сопоставлением текстов белого движения и левых эсеров, тут проходит очевидный лексический водораздел).

Еще одной «подставной» для модели категорией выступает класс *Noble* (Дворянское собрание), в который были включены тексты дворянских собраний, относящиеся к периоду до 1914 г. Мы решили экспериментально добавить в модель тексты другого исторического периода, пересекающиеся по лексике с текстами монархических партий для проверки способности модели отличать их.

Перед тем как приступить к разбору процесса обучения и результатов классификации, предлагаем взглянуть на выгрузку предсказаний еще не обученной модели для анализа ее способности взаимодействовать с историческим контекстом. В одной из недавних статей [*Petroni, Rocktäschel, Lewis* et al., 2019], посвященной изучению принципов работы больших языковых моделей, было выдвинуто предположение, что LLM можно сравнить с базами данных, поскольку пространство весов модели хранит информацию не только о структуре языка, но и о семантических связях между словами. Для характеристики «знаний» модели о рассматриваемом периоде мы используем метод Fill-mask. Поскольку одна из базовых задач обучения BERT-модели – это предсказание пропущенного в предложении слова [*Jacob, Chang, Lee, Toutanova*, 2019], мы можем воспользоваться методом заполнения пропуска ([Mask]) для получения набора наиболее вероятных слов, которые могли бы заполнить эти пропуски в соответствии с семантическим контекстом. В качестве основной модели был выбран классический BERT, дообученный на русскоязычной части «Википедии». Для исследования классификационного потенциала LLM для исторических источников мы приняли решение взять одну из наиболее распространенных моделей. На вход подавались тексты лишь с одним пропуском, поэтому модель видела весь контекст.

Рассмотрим в деталях процесс восстановления предложений на основе контекста с помощью модели BERT (примеры 1-4; табл. 2-5).

Пример № 1

Ю.О. Мартов. Небывалая хозяйственная разруха, катастрофическое положение продовольственного дела в атмосфере непрекращающейся Гражданской войны порождает в отсталых слоях населения угрожающее погромное настроение. При наличности веками слагавшихся **[MASK1]** *предрассудков настроение это под влиянием явно черносотенных, а подчас и примазавшихся к* **[MASK2]** *власти темных элементов выливается по преимуществу в форму* **[MASK3]** *погромов. Все враги* **[MASK4]***, все сторонники прежнего режима прибегают к старому, испытанному оружию антисемитской травле, науськиванию масс на евреев. А ныне стоящая у власти* **[MASK5]** *партия, своей демагогической агитацией развращавшая и развращающая массы, а всей своей правительственной политикой порождающая условия, систематически питающие погромные настроения, абсолютно не в состоянии, каковы бы ни были субъективные настроения ее руководителей, вести хоть сколько-нибудь успешную борьбу с погромной опасностью* [Меньшевики в большевистской России…, 1999, с. 485].

Таблица 2

Восстановление исторического текста с помощью модели BERT. Пример № 1

Номер пропуске в тексте	**Слово в оригинальном документе**	**Предсказывания модели с весами вероятности**
[MASK1]	национальных	**национальных** (0.16)**; еврейских** (0.11)**; народных** (0.07)**; антисемитских** (0.07)**; религиозных** (0.04)
[MASK2]	советской	**советской** (0.19)**; новой** (0.14)**; царской** (0.11)**; Советской** (0.08)**; центральной** (0.03)
[MASK3]	антиеврейских	**массовых** (0.17)**; городских** (0.12)**; революционных** (0.07)**; национальных** (0.02)**; политических** (0.01)
[MASK4]	революции	**народа** (0.28)**; прошлого** (0.11)**; революция** (0.05)**; СССР** (0.04)**; России** (0.03)
[MASK5]	большевистская	**коммунистическая** (0.3)**; политическая** (0.07)**; Коммунистическая** (0.05)**; правящая** (0.05)**; революционная** (0.04)

Пример № 2

Передайте VIII съезду уполномоченных **[MASK1]** *обществ Мою сердечную благодарность за их молитвы, благопожелания и выраженные Государыням* **[MASK2]**, *Мне и Наследнику* **[MASK3]** *чувства; уверен, что верное заветам старины Русское дворянство и впредь всегда будет служить опорой Престола в деле мирного развития великой нашей России* [Объединенное дворянство…, 2001, с. 301].

Таблица 3

Восстановление исторического текста с помощью модели BERT. Пример № 2

Номер пропуска в тексте	**Слово в оригинальном документе**	**Предсказывания модели с весами вероятности**
[MASK1]	дворянских	**дворянских** (0.53)**; акционерных** (0.10)**; сельских** (0.05)**; русских** (0.02)**; вольных** (0.02)
[MASK2]	Императрицам	**Вам** (0.34)**; Вас** (0.07)**; Мне** (0.03)**; Вы** (0.03)**; Тебе** (0.03)
[MASK3]	Цесаревичу	**государства** (0.11)**; народа** (0.09)**; России** (0.06)**; общества** (0.04)**; Церкви** (0.02)

Пример № 3

Петроградский **[MASK1]** *послал в эти дни к вам, московским* **[MASK2]**, *своих делегатов для того, чтобы связать свое движение с вашим, для того чтобы рассказать вам о нуждах и планах петроградских пролетариев, и обратно привезти ваше сочувствие и вашу поддержку. Петроградский пролетариат оказался в чрезвычайно тяжелом положении. Испытавший на себе наиболее остро все последствия дикой* **[MASK3]** *политики, потерявший в результате этой политики силу и независимость своих организаций, он, фактически безоружный и беззащитный, столкнулся лицом к лицу с последними бедствиями, надвинувшимися на* **[MASK4]**. *Обреченный город начал эвакуироваться. Совет Народных* **[MASK5]** *покинул Петроград. Начались расчеты на заводах. Во всех вопросах — эвакуации,* **[MASK6]**, **[MASK7]**, *охраны безопасности населения — рабочий* **[MASK8]** *воочию увидел одну разруху, полную дезорганизацию. Перевыборы в* **[MASK9]** *фактически были запрещены и оторвавшись совершенно от рабочей массы, эти большевистские канцелярии бездействовали, или сами усиливали дезорганизацию. Деятельность профессиональных союзов, заводских* **[MASK10]**, *была извращена и разрушена, кооперативы — оттерты от широкой работы* [Рабочее оппозиционное движение…, 2006, с. 54].

Таблица 4

Восстановление исторического текста с помощью модели BERT. Пример № 3

Номер пропуска в тексте	**Слово в оригинальном документе**	**Предсказывания модели с весами вероятности**
[MASK1]	пролетариат	**Совет** (0.74)**; совет** (0.16)**; ВРК** (0.03)**; комитет** (0.01)**; градоначальник** (0.01)
[MASK2]	рабочим	**рабочим** (0.20)**; властям** (0.1)**; большевикам** (0.10)**; людям** (0.06)**; городам** (0.02)
[MASK3]	большевистской	**большевистской** (0.27)**; революционной (0.07); внутренней** (0.04)**; внешней** (0.04)**; буржуазной** (0.03)
[MASK4]	Петроград	**Петроград** (0.40)**; город** (0.06)**; рабочих** (0.06)**; Россию** (0.06)**; большевиков** (0.03)
[MASK5]	Комиссаров	**Комиссаров** (0.97)**; комиссаров** (0.02)**; рабочих** (0.01)**; депутатов** (0.01)**; представителей** (0.01)
[MASK6]	безработицы	**транспорта** (0.14)**; торговли** (0.05)**; снабжения** (0.04)**; труда** (0.04)**; отопления** (0.04)
[MASK7]	продовольствия	**труда** (0.07)**; мобилизации** (0.05)**; забастовки** (0.03)**; питания** (0.03)**; торговли** (0.02)
[MASK8]	класс	**класс** (0.44)**; Совет** (0.07)**; народ** (0.07)**; Петроград** (0.04)**; совет** (0.02)
[MASK9]	Совете	**городах** (0.24)**; Петрограде** (0.11)**; губернии** (0.07)**; Советы** (0.07)**; городе** (0.05)
[MASK10]	комитетов	**рабочих** (0.25)**; профсоюзов** (0.09)**; организаций** (0.07)**; комитетов** (0.07)**; союзов** (0.06)

Пример № 4

В дополнение утвержденного Главнокомандующим 18 мая **[MASK1]** *года постановления Особого* **[MASK2]** *о предоставлении учреждениям гражданского ведомства права* **[MASK3]** *по железным дорогам чинов, командируемых по службе, и грузов для надобностей Вооруженных Сил на* **[MASK4]** *России по воинским предложениям предоставить указанным учреждениям право* **[MASK3]** *по воинским предложениям тех чинов и грузов по водным путям с соблюдением правил, установленных означенным постановлением для перевозок по железным дорогам, с тем чтобы перевозки водой за счет казны по воинским предложениям производились лишь в тех случаях, когда эти перевозки будут обходиться дешевле, чем перевозки между теми же пунктами по железным дорогам, или когда между соответствующими пунктами вовсе нет железных дорог* [Журналы заседаний Особого совещания…, 2008, с. 892].

Результаты выгрузки предсказаний модели показывают, что даже базовая модель, обученная на русскоязычных текстах (в первую очередь – на «Википедии») способна угадывать исторический контекст и подбирать оригинальные (пример 1, строка 1; пример 2, строка 1; пример 3, строка 3) или близкие к оригинальным слова (пример 4, строка 3). Также возможна ситуация, в которой модель в целом верно определяет исторический период, к которому относится текст, но ошибается с конкретным годом (пример 4, строка 1). В то же время следует отметить, что ряд достаточно очевидных позиций не был угадан моделью (пример 1, строка 4; пример 3, строка 10). Довольно любопытны случаи, в которых модель предлагает верный вариант, но ставит его не на первые места по вероятности (пример 4, строка 4).

Таблица 5

Восстановление исторического текста с помощью модели BERT. Пример № 4

Номер пропуска в тексте	**Слово в оригинальном документе**	**Предсказывания модели с весами вероятности**
[MASK1]	1919	**1916** (0.08)**, 1917 (0.07); 1912 (0.06); 1915 (0.05); 1918 (0.04)**
[MASK2]	совещания	**совещания (0.72); отдела (0.04); Комитета (0.03); комитета (0.02); особ (0.01)**
[MASK3]	перевозок	**перевозки (0.66), перевозок (0.25), перевозить (0.02), перевозку (0.01), перевозка (0.01)**
[MASK4]	Юге	**территории (0.35); Юге (0.26); территорию (0.031); юге (0.03); границах (0.08)**

Подводя краткий итог сказанному, хочется отметить, что модель однозначно способна понимать семантику текста и угадывать по ней исторический контекст, однако делает это нестабильно. На фоне рассмотренной выгрузки представляется вполне допустимым, что, решая задачу классификации текстов, модель BERT в процессе обучения будет охватывать значительно больший информационный контекст; операция сравнения текстов будет носить куда более сложный характер, нежели простой подсчет ключевых слов, маркирующих принадлежность текста к каждому из классов. В этом контексте нам представляется интересным сравнить результаты работы модели с другими подходами, основанными на частотном анализе слов (наивный байесовский классификатор), линейной регрессии с использованием эмбеддингов и градиентном бустинге (gradient boosting) на основе деревьев решений.

Сравнение результатов предсказания модели с другими классификаторами

Нами был использован подход дообучения большой языковой модели (fine-tuning) на классификацию текстов доменного датасета. На тренировочную часть было выделено 80 % данных (1938 документов), на тестирование – 20 % (485 документов).

Результаты работы модели на тестовой выборке (документы, которые модель не видела в процессе обучения) представлены в табл. 6. Отдельно отметим, что качество классификации оказалось чуть ниже на группе документов «Рабочая оппозиция большевикам», а эти документы, как мы указывали, по содержанию и лексике очень похожи на документы эсеров и меньшевиков. Мы можем констатировать, что модель действует согласованно с гипотезой, сформулированной на общих знаниях об исторической эпохе.

Таблица 6

Сравнение результатов классификации BERT с альтернативными классификаторами. F1-score

Название класса	NB	Logistic Regression	XGBoost	**BERT-Classifier**	Количество текстов
Bolsheviki	0.68	0.77	0.82	**0.99**	52
Mensheviki	0.76	0.67	0.73	**0.94**	62
Left_SRs	0.84	0.78	0.80	**0.99**	70
Kadets	0.79	0.83	0.85	**0.99**	37
Rights	0.71	0.81	0.79	**0.99**	58
White_army	0.83	0.86	0.86	**0.96**	56
Work_oppose	0.42	0.64	0.81	**0.95**	42
Nobles	0.69	0.82	0.88	**0.98**	108

Далее сравним результаты работы модели с альтернативными классификаторами, основанными на принципах классического машинного обучения. Хотя применяемые классификаторы показали в целом неплохую точность классификации, LLM значительно превосходит их. Это позволяет нам рекомендовать использовать именно LLM в задачах классификации текстов, так как современные пакеты для работы с языковыми моделями (например, Transformers [*Wolf, Debut, Sanh* et al., 2020]) ориентированы на пользователей, обладающих базовым уровнем программирования, что дает возможность широким кругам исследователей обращаться к рассматриваемым методам. В целом сегодня использование LLM для прикладных задач не выглядит намного сложнее, чем запуск базовых моделей из пакетов для машинного обучения (Scikit-learn).

Анализ весов предсказаний классификатора

Нейросети давно перестали быть непроницаемым «черным ящиком», как о них периодически высказываются. В действительности исследователи обладают большим набором методов, позволяющих анализировать поведение моделей. Безусловно, часть процесса обучения и работы механизмов внимания все еще остается скрыта. Но сегодня мы можем проверять, насколько корректны зависимости, которые модели приписывают данным. В частности, для этого может использоваться библиотека SHAP [*Lundberg, Scott, Lee*, 2017], оценивающая вклад каждого признака в итоговое предсказание модели. В случае с LLM мы можем получить подробную карту вкладов каждого из токенов (слов) в решении модели приписать тексту тот или иной класс. Причем SHAP также дает нам информацию о том, какие слова уменьшают вероятность принадлежности текста к одному из классов, показанных модели на этапе обучения. Так, если мы проанализируем два небольших отрывка из табл. 2 с помощью SHAP (рис. 1), то обнаружим, что наша модель явно согласуется с контекстом рассматриваемой эпохи. Например, предписывая первому тексту класс «Меньшевики», модель обращает особое внимание на сочетание токенов «Март», «ов» (см. рис. 1). Любопытно, что последнее предложение модель относит к классу «Кадеты» (рис. 2). Вероятно, в ходе обучения модель выявила особое значение риторики, связанной с погромами, в кадетских текстах. Слово «главнокомандующий» воспринимается ею как явный маркер класса «Белая армия» (рис. 3), и она занижает вероятность для всех других классов (как, например, для класса «Меньшевики» (рис. 4). Мы привели лишь несколько примеров. Сегодня любой LLM-классификатор может подвергнуться детальному анализу, проверяющему правильность изученных моделью зависимостей.

Выводы

В ходе нашего исследования были подтверждено, что LLM могут успешно использоваться для решения задач классификации исторических текстов по политическому признаку. Представляется, что стратегия информационного поиска, опирающаяся на дообучение языковой модели на небольшой выборке текстов (около 200–250 единиц), может применяться для оптимизации отбора исторических документов, относящихся к определенной историографической проблеме (в нашем случае – к политической истории революционного 1917 г.). Другим важным результатом исследования мы считаем высокое качество классификации, полученное с помощью классического BERT. Нами были использованы токенайзер, обученный на современных текстах на русском языке, и только первые 512 слов наших текстов в качестве токенов. Вероятно, лексическая близость начала XX в. к современному русскому языку позволяет получать высокое качество классификации даже без специализированной доменной адаптации к историческому периоду на этапе первичного обучения (pre-train). Это может оказаться ценным результатом, поскольку коллекции распознанных исторических документов все еще очень далеки по объему от текстовых массивов, на которых обучаются современные LLM. Другим возможным объяснением высокого качества работы модели может быть изначальное обучение BERT на текстах русскоязычной «Википедии». Интернет-энциклопедия содержит очень подробные статьи по истории революций 1917 г. и Гражданской войны. Если это предположение правдиво, это открывает дальнейшие перспективы изучения специфики влияния источников обучения на качество работы больших языков моделей в рамках сложной доменной адаптации.

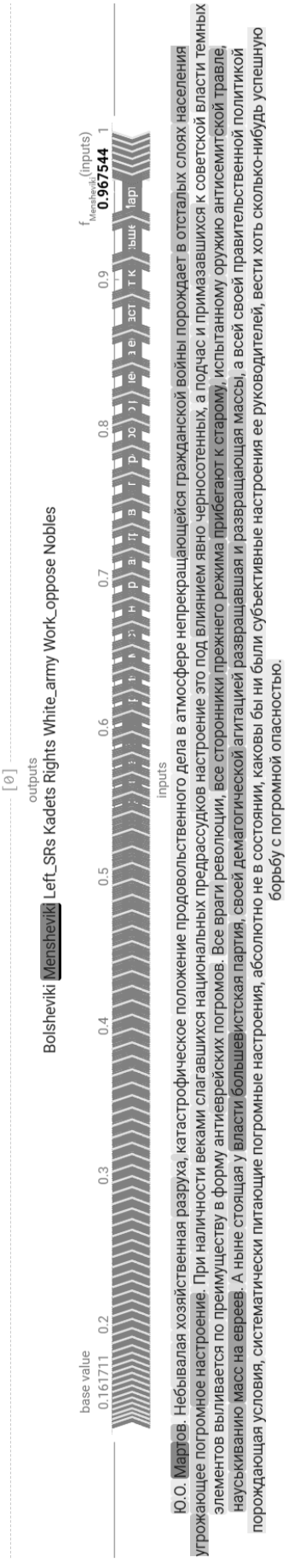

Рис. 1. Анализ вероятности принадлежности случайного текста к классу «Меньшевики» с помощью SHAP

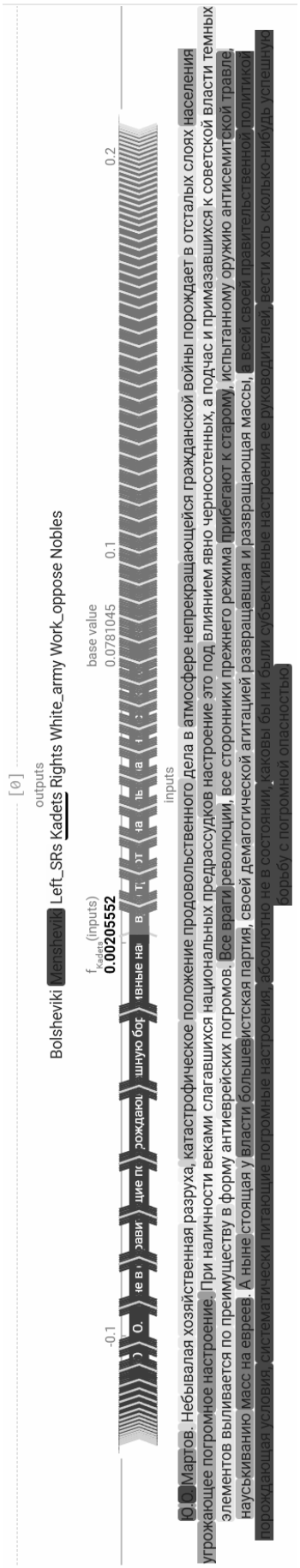

Рис. 2. Анализ вероятности принадлежности случайного текста к классу «Кадеты» с помощью SHAP

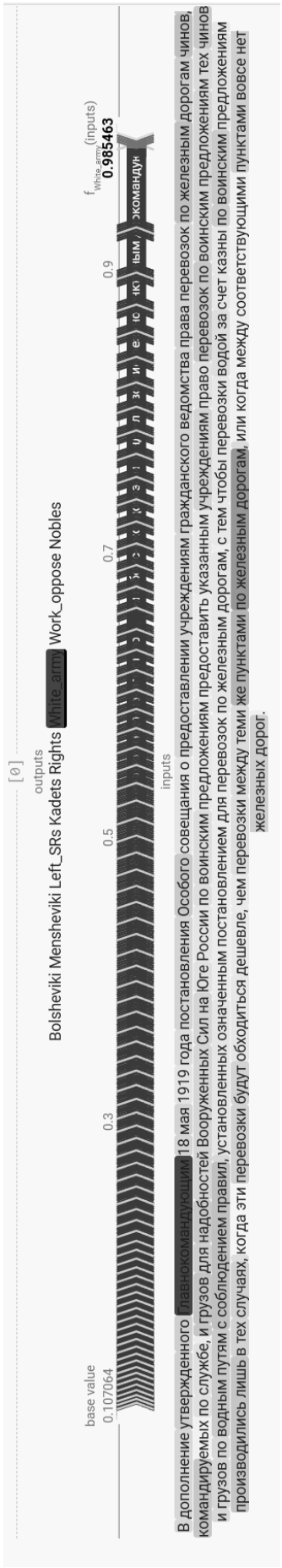

Рис. 3. Анализ вероятности принадлежности случайного текста к классу «Белая армия» с помощью SHAP

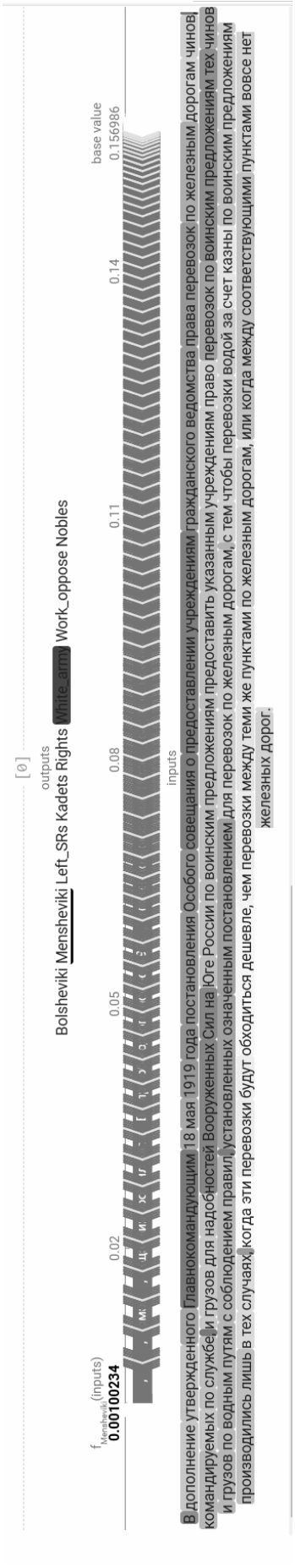

Рис. 4. Анализ вероятности принадлежности случайного текста к классу «Меньшевики» с помощью SHAP

Примечания

[1] Исследование выполнено при финансовой поддержке некоммерческого фонда развития науки и образования «Интеллект» (Non-commercial Foundation for the Advancement of Science and Education INTELLECT). Весь программный код исследования доступен в этом репозитории: https://github.com/GalushkoIlya/Hist_Bert_classifier.

Библиографический список

Володин А.Ю. Цифровая трансформация истории? Данные, стандарты, подходы [Электронный ресурс] // История. 2020. Т. 11, вып. 3 (89). URL: https://history.jes.su/s207987840009746-9-1/ (дата обращения: 10.08.2024). DOI: 10.18254/S207987840009746-Электронная библиотека исторических документов (проект РИО): http://docs.historyrussia.org/ru/nodes/1-glavnaya. EDN: JLAZBU.

Журналы заседаний Особого совещания при Главнокомандующем Вооруженными Силами на Юге России А.И. Деникине. Сентябрь 1918-го – декабрь 1919 года. М.: РОССПЭН, 2008. 1003 с.

Меньшевики в большевистской России. 1918—1924. Меньшевики в 1918 году. М.: РОССПЭН, 1999. 799 с.

Объединенное дворянство. Съезды уполномоченных губернских дворянских обществ. 1906–1916 гг.: а 3 т. Т. 2. 1909–1912 гг. Кн. 2. 1911–1912 гг. М.: РОССПЭН, 2001. 608 с.

Рабочее оппозиционное движение в большевистской России. 1918 г. Собрания уполномоченных фабрик и заводов. Документы и материалы. М.: РОССПЭН, 2006. 656 с.

Юмашева Ю.Ю. Историческая наука, архивы, библиотеки, музеи и искусственный интеллект: год спустя // Документ. Архив. История. Современность: сб. науч. тр. Екатеринбург: Изд-во Урал. ун-та, 2022. Вып. 22. С. 217–241. EDN: BPPUWY.

Devlin J., Ming-Wei C., Kenton L., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // NAACL-HLT 2019. 2019. P. 4171–4186. DOI: 10.18653/v1/N19-1423.

Kuratov, Y., Arkhipov, M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // arXiv preprint arXiv:1905.07213. 2019. Available at: https://arxiv.org/abs/1905.07213 (accessed: 04.08.2024).

Lundberg S. M., Lee S. I. A unified approach to interpreting model predictions // Advances in neural information processing systems. 2017. Vol. 30.

Petroni F., Rocktäschel T., Patrick L., Bakhtin A., Wu Yu., Miller A.H., Riedel S. Language models as knowledge bases? // arXiv preprint arXiv:1909.01066. 2019. Available at: https://arxiv.org/abs/1909.01066 (accessed: 04.08.2024).

Wolf T., Debut L., Sanh V. [et al.]. Transformers: State-of-the-Art Natural Language Processing // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020. P. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

Yang Tze-I, Torget A., Mihalcea R. Topic Modeling on Historical Newspapers // Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Portland, OR, USA: Association for Computational Linguistics, 2011. P. 96–104.

Дата поступления рукописи в редакцию 29.08.2024

HISTORICAL DOCUMENTS CLASSIFICATION USING BERT: LLM AND HISTORICAL DOMAIN

I. N. Galushko

Moscow State University, Lomonosovskiy ave., 27-4, Moscow, 119192, Russia
National Research University Higher School of Economics, Pokrovsky blvd. 11, Moscow, 109028, Russia
i.galushko15@gmail.com
ResearcherID: LYP-0220-2024
SPIN: 3433-9690

At the present stage of studying Russian history, discussions about processing large collections of historical documents are becoming especially relevant. Today, the process of digitizing archival collections is actively underway, but in most cases, the created corpus is simply posted on the site and remains unused for years. This is because we

often encounter difficulties in processing an entire collection when accessing the funds of a large social institution; digitized funds can contain hundreds of thousands of pages of documentation. Limited time does not allow even a quick reading to cover all the available documents. This problem could be at least partially solved by using LLMs for annotation or text search optimization. However, at the current stage of archival development, specialists are just beginning to work with natural language processing methods. The main request of the professional community is to study the specifics of the work of artificial intelligence models and machine learning on historical domain texts. This article is a preliminary study of modern LLMs' interaction with historical texts. For the analysis, we chose one of the most popular models – BERT – and one of the most common NLP tasks – classification.

Key words: text classification, political history, artificial intelligence, attention mechanism analysis, machine learning, BERT, NLP.

Acknowledgments

[1] The reported study was supported by the Non-commercial Foundation for the Advancement of Science and Education INTELLECT. The entire program code of the study is available in this repository: https://github.com/GalushkoIlya/Hist_Bert_classifier

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Kuratov, Y., & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. *arXiv*. https://doi.org/10.48550/arXiv.1905.07213

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Mensheviki v bol'shevistskoy Rossii. 1918–1924. Mensheviki v 1918 godu [Mensheviks in Bolshevik Russia. 1918–1924: The Mensheviks in 1918]. (1999). ROSSPEN.

Ob"edinennoe dvoryanstvo. S"ezdy upolnomochennykh gubernskikh dvoryanskikh obshchestv. 1906–1916 gg. V 3 t. T. 2. 1909–1912 gg. Kn. 2. 1911–1912 gg. [United Nobility: Congresses of Authorized Representatives of Provincial Noble Societies. 1906–1916. In 3 vols. Vol. 2: 1909–1912. Book 2: 1911–1912]. (2001). ROSSPEN.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases? *arXiv*. https://arxiv.org/abs/1909.01066

Rabochee oppozitsionnoe dvizhenie v bol'shevistskoy Rossii. 1918 g. Sobraniya upolnomochennykh fabrik i zavodov. Dokumenty i materialy [The Workers' Opposition Movement in Bolshevik Russia, 1918: Meetings of Factory and Plant Delegates. Documents and Materials]. (2006). ROSSPEN.

Volodin, A. (2020). Digital transformation of history? Data, standards, approaches. *ISTORIYA, 11*(3). https://doi.org/10.18254/S207987840009746-9

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yang, T.-I., Torget, A., & Mihalcea, R. (2011). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 96–104). Association for Computational Linguistics.

Yumasheva, Y. Y. (2022). Historical science, archives, libraries, museums, and artificial intelligence: A year later. *Dokument. Arkhiv. Istoriya. Sovremennost'*, 22, 217–241.

Zhurnaly zasedaniy Osobogo soveshchaniya pri Glavnokomanduyushchem Vooruzhennymi Silami na Yuge Rossii A. I. Denikine [Minutes of the Meetings of the Special Council under the Commander-in-Chief of the Armed Forces of South Russia, A. I. Denikin]. (2008). ROSSPEN.