УДК 930.2:81’32-13:004.94
DOI:10.17853/2587-6910-2023-09-40-44

ИСПОЛЬЗОВАНИЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ТЕКСТА ДЛЯ ОБНАРУЖЕНИЯ СКРЫТОЙ ИНФОРМАЦИИ В ИСТОРИЧЕСКИХ ИСТОЧНИКАХ

USING TEXT MINING TO UNCOVER HIDDEN INFORMATION IN HISTORICAL SOURCES

Алексей Валерьевич Кузнецов
кандидат исторических наук
historyras@gmail.com
Институт всеобщей истории РАН, Россия, Москва

Alexey Valer’evich Kuznetsov
Candidate of Sciences in History
Institute of World History of Russian Academy of Sciences, Russia, Moscow

Аннотация. *В статье представлен обзор того, как интеллектуальный анализ текста используется для выявления скрытой информации в исторических текстах. Внимание акцентируется на методе тематического моделирования и моделях эмбеддингов слов. Статья иллюстрирует, как эти методы использовались в конкретных исторических исследованиях. Делается вывод о том, что интеллектуальный анализ текста является полезным инструментом для обнаружения скрытой информации в исторических текстах.*

Ключевые слова: *исторические источники; интеллектуальный анализ текст; тематическое моделирование; эмбеддинги слов.*

Abstract. *The article presents an overview of how text mining can be employed to reveal hidden information in historical texts. The attention is focused on the method of thematic modeling and word embedding models. The article illustrates how these techniques have been utilized in historical research. It concludes that text mining is a useful tool for uncovering hidden information in historical.*

Keywords: *historical sources; text mining; topic modeling; word embedding.*

На рубеже 1970–80-х гг. И. Д. Ковальченко сформулировал информационный подход к историческим источникам. В основе этого подхода лежит представление о том, что исторические источники являются не просто записями событий или высказываний, но носителями информации об обществе и культуре, в которых они создавались. Непосредственный создатель исторического источника отражает в нем многообразие взаимосвязей, присущих явлениям окружающего мира, что обусловливает безграничный объем информации как явной, так и скрытой. Анализ сведений, непосредственно выраженных в историческом источнике, позволяет выявлять скрытую информацию. Возможность извлечения скрытой информации лимитируется только познавательными возможностями исследователя и зависит от применяемых им методов [1, с. 121–134]. В последние годы произошли без преувеличения револю-

ционные изменения в области интеллектуального анализа текста (англ. text mining) — направления искусственного интеллекта, целью которого является получение информации из неструктурированных текстовых данных на основе методов машинного обучения и обработки естественного языка. Это оказало большое влияние на исследования в области гуманитарных наук. В исторической науке применение новых методов и технологий позволило выявлять и анализировать как явную, так и скрытую информацию о прошлом. Одним из наиболее часто используемых методов интеллектуального анализа текста стало тематическое моделирование, а передовой технологией кодирования текста — эмбеддинги слов. Цель этой статьи — проанализировать использование тематического моделирования и моделей эмбеддингов слов для обнаружения скрытой информации в исторических текстах, включая их преимущества и ограничения, а также привести примеры их применения в исторических исследованиях.

Тематическое моделирование — это метод машинного обучения, который позволяет проанализировать большую текстовую коллекцию и определить, к каким темам относится каждый её документ и какие слова составляют каждую тему. Под темой понимается перечень слов, часто совместно встречающихся в отдельном документе. В настоящее время предложено множество разнообразных алгоритмов тематического моделирования [4, с. 63], но самым популярным, благодаря большому количеству хорошо задокументированных инструментов, остается латентное размещение Дирихле, предложенное ещё в 2003 году [5]. Тематическое моделирование стало полезным инструментом в самых разных областях исследований [7]. Ключевым преимуществом метода является возможность анализировать скрытую тематическую структуру огромного объема текстовых документов и отслеживать эволюцию этих тем с течением времени. Подчеркнем, что исследователь изначально не знает, какие темы и в каком объеме присутствуют в анализируемых текстах, именно поэтому при использовании тематического моделирования речь идет о выявлении скрытой, неявной тематической структуры текстовой коллекции. Первым академическим историческим исследованием, использующим это метод, стала статья Д. Ньюмана и Ш. Блок 2006 года «Вероятностная тематическая декомпозиция американской газеты восемнадцатого века», посвященная анализу тематики газеты Pennsylvania Gazette в период с 1728 по 1800 год [16]. Авторы проанализировали тексты общим объемом 25 миллионов слов в статьях и рекламных объявлениях, отражающих повседневную жизнь нескольких поколений до, во время и после основания Соединенных Штатов Америки. Другим классическим примером стало использование К. Блевинсом тематического моделирования для анализа дневника акушерки Марты Баллард (1735–1812), которая делала записи более 27 лет, включая эпоху Войны за независимость в США [6]. К. Блевинс сосредоточился на выявлении взаимосвязи между появлением тем в дневнике и течением времени. Это позволило ему выявить интересные закономерности в повседневной жизни и коммуникации акушерки.

С 2010 года наметился рост интереса к тематическому моделированию со стороны гуманитариев [21, p. 2]. Объектом анализа помимо периодических изданий и дневников чаще всего становятся коллекции писем [15], хроники [2], записи парламентских дебатов [11] и судебные решения [10].

Необходимо принимать во внимание, что результат тематического моделирования носит вероятностный характер. Большинство алгоритмов являются разновидностью методов машинного обучения без учителя. Итог их применения сильно зависит от того, как будет предварительно обработан текстовый корпус и выбранны параметры моделирования. В настоящее время не выработано единого мнения ни по предобработке текста, ни по подбору параметров моделирования [3, с. 10]. Одна из самых больших проблем — выбор оптимально количества тем. Тематическая модель, построенная несколько раз с одинаковыми настройками, на одних и тех же данных вполне может дать разное распределение тем по документам и слов по темам [17, p. 437]. Такая неустойчивость результатов моделирования составляет основное ограничение метода. В гуманитарных исследованиях тематическое моделирование часто характеризуют как пример «дальнего чтения» (англ. distant reading) — подхода литературоведа Ф. Моретти, базирующегося на количественном

анализе объемных текстовых коллекций, и противопоставляют его привычному «пристальному чтению» (англ. close reading) [4]. Однако для получения интерпретируемых результатов от исследователя требуется не только понимание технических аспектов работы алгоритма, но и детальное знакомство с контекстом, в котором был создан исторический текст. По этой причине в исторической науке тематическое моделирование в настоящее время не используется как самостоятельный метод исследования, а лишь в сочетании с традиционными методами внимательного чтения. Такой подход сочетает в себе сильные стороны количественных и качественных методов, обеспечивая более глубокий уровень анализа [3, с. 11].

Значимой тенденцией в области компьютерного анализа исторических текстов является все более широкое использование для кодирования текстов распределенных векторных представлений слов, известных как эмбеддинги слов. Модели эмбеддингов слов кодируют семантику слов и семантические отношения между ними на основе контекста, представляя каждое слово как вектор в плотном векторном пространстве. Под контекстом в данном случае понимается несколько слов, окружающих целевое. Слова, которые встречаются в сходных контекстах, расположены в векторном пространстве близко друг к другу, а слова, встречающиеся в разных контекстах, находятся сравнительно дальше друг от друга [14, p. 136–137]. Для измерения семантической близости слов чаще всего используется мера косинусного сходства — косинус угла между векторами слов. Модели эмбеддингов слов создаются путем обучения нейронной сети на объемном текстовом корпусе, а затем могут использоваться для различных задач интеллектуального анализа текстов, таких как семантический анализ слов, морфологический анализ, синтаксический анализ, машинный перевод, анализ тональности текста, определении авторства и другие. Высказываются предположения, что использование эмбеддингов слов в гуманитарных исследованиях в ближайшие годы значительно расширится [12, p. 448].

Существенным свойством моделей эмбеддингов слов, используемым в исторических исследованиях, является то, что они позволяют проследить изменения с течением времени значения слов, а также идей и понятий, передаваемых словами. Обучение и анализ моделей эмбеддингов слов стало фундаментальной инновацией для исторической семантики. Обнаружение семантических изменений дает ценную информацию о социальных и культурных изменениях в обществе [13]. Другим направлением использования эмбеддингов слов стало выявление и изучение эволюции гендерных, этнических и социальных стереотипов [8; 9], проявление которых традиционными способами фактически не фиксируется.

Характерным примером и образцом использования диахронических моделей эмбеддингов слов для выявления семантических изменений служит недавняя статья Н. Педраццини и Б. Макгилливрей «Машины в СМИ: семантическое изменение лексики механизации в британских газетах XIX века» [18]. В статье впервые предпринят масштабный анализ семантических изменений на протяжении XIX века, терминов английского языка, относящихся к сфере механизации (traffic, trade, train, coach, wheel, railway, matches, bulb, gear, stamp). Анализ опирается на корпус британских газет XIX–начала XX веков объемом 4,6 миллиарда слов. Авторы обучили 12 моделей эмбеддингов слов отдельно для каждого десятилетия с 1800 по 1910 годы. На их основе они смогли проследить изменения в значении слов с течением времени, определить поворотные моменты, а полученные результаты сравнили с предыдущими лингвистическими исследованиями, использующими традиционные методы. В итоге авторы пришли к выводу, что применение моделей эмбеддингов слов для обнаружения сематических изменений дало результаты, совпадающие с наблюдениями, сделанными традиционными методами. Причем в некоторых случаях им удалось уловить семантические изменения, не идентифицированные в предыдущих исследованиях.

К настоящему времени был предложен целый ряд алгоритмов построения моделей эмбеддингов слов, таких как word2vec, FastText или GloVe, но для реализации любого из них необходим большой текстовый корпус. Объем такого корпуса обычно составляет несколько миллионов слов. Для большинства историков такое количество оцифрованного материала недоступно. В этом состоит основное ограничение

для использования эмбеддингов слов в исторических исследованиях. Кроме того, необходимым условием использования эмбеддингов слов является дополнение исследования этапом оценки построенной модели, что составляет отдельную проблему в интеллектуальном анализе текста [22, p. 235–236]. Наконец, из-за зависимости моделей эмбеддингов слов от алгоритмов, которые не всегда могут давать согласованные результаты в разных наборах данных или контекстах, интерпретация полученных результатов требует тщательного анализа при принятии выводов на их основе. Несмотря на сложности, в последние годы наблюдается всплеск интереса к использованию моделей эмбеддингов слов в исторических исследованиях, в связи с растущей доступностью исторических корпусов в цифровой форме. В частности, появились модели для таких языков как латинский [20] и древнегреческий [19].

Достижения в области интеллектуального анализа текстов продолжают открывать новые возможности и направления исследований исторических источников. Метод тематического моделирования и модели эмбеддингов слов являются популярными и полезными инструментами анализа больших объемов текстовых данных, помогающими выявить в них неявную информацию. В то же время их применение требует осторожности и понимания технических аспектов работы алгоритмов машинного обучения.

Список литературы

1. Ковальченко И. Д. Методы исторического исследования. 2-е изд., доп. М.: Наука, 2003. 486 с.
2. Кузнецов А. В. Компьютерный анализ текстов на латинском языке: тематическое моделирование «Истории готов, вандалов и свевов» Исидора Севильского // Историческая информатика. 2020. № 2. С. 202–217. https://doi.org/10.7256/2585-7797.2020.2.32961.
3. Кузнецов А. В., Ямщиков С. В. Тематическое моделирование в исторической науке // Социосфера. 2022. № 4. С. 9–12. URL: http://sociosphera.com/files/conference/2022/SF-4-22/9-12_Kuznetsov.pdf.
4. Милкова М. А. Тематические модели как инструмент «дальнего чтения» // Цифровая экономика. 2019. №. 1 (5). С. 57–70. https://doi.org/10.34706/DE-2019-01-06 6.
5. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. Vol. 3. P. 993–1022. URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.
6. Blevins C. Topic Modeling Martha Ballard's Diary. Posted on April 1, 2010. URL: http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary.
7. Boyd-Graber J., Hu Y., Mimno D. Applications of Topic Models // Foundations and Trends® in Information Retrieval. 2017. Vol. 11, no. 2–3. P. 143–296. http://dx.doi.org/10.1561/1500000030/.
8. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words / Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., Banaji, M. R. // Psychological Science. 2021. Vol. 32, no. 2. P. 218–240. https://doi.org/10.1177/0956797620963619.
9. Charlesworth T. E. S., Caliskan A., Banaji M. R. Historical representations of social groups across 200 years of word embeddings from Google Books // Proceedings of the National Academy of Sciences. 2022. Vol. 119, №. 28. P. e2121798119-e2121798119. https://doi.org/10.1073/pnas.2121798119.
10. Grajzl P., Murrell P. A machine-learning history of English caselaw and legal ideas prior to the Industrial Revolution I: generating and interpreting the estimates // Journal of Institutional Economics. 2021. Vol. 17, iss. 1. P. 1–19. https://doi.org/10.1017/S1744137420000326.
11. Guldi J. Parliament's debates about infrastructure: an exercise in using dynamic topic models to synthesize historical change // Technology and Culture. 2019. Vol. 60, iss. 1. P. 1–33. https://doi.org/10.1353/tech.2019.0000.

12. Indukaev A. Studying ideational change in Russian politics with topic models and word embeddings // The Palgrave Handbook of Digital Russia Studies. Cham: Palgrave Macmillan, 2021. P. 443–464. https://doi.org/10.1007/978-3-030-42855-6_25.

13. Kozlowski A. C., Taddy M., Evans J. A. The geometry of culture: Analyzing the meanings of class through word embeddings // American Sociological Review. 2019. Vol. 84, iss. 5. P. 905–949. https://doi.org/10.1177/0003122419877135.

14. Kutuzov A., Andreev I. Texts in, meaning out: neural language models in semantic similarity task for Russian // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). Вып. 14 (21). Т. 2: Доклады специальных секций. М.: РГГУ, 2015. С. 133–144.

15. McGillivray B., Buning B., Hengchen S. Topic Modelling: Hartlib's Correspondence before and after 1650 // Reassembling the Republic of Letters in the Digital Age. Göttingen: Göttingen University Press, 2019. P. 426–428. https://doi.org/10.17875/gup2019-1146.

16. Newman D. J., Block Sh. Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper // Journal of the American Society for Information Science and Technology. 2006. Vol. 57, iss. 6. P. 753–767.

17. Oiva M. Topic Modeling Russian History // The Palgrave Handbook of Digital Russia Studies. Cham: Palgrave Macmillan, 2021. P. 427–442.

18. Pedrazzini N., McGillivray B. Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers // Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities. Association for Computational Linguistics, 2022. P. 85–95. URL: https://www.researchgate.net/publication/365670282_Machines_in_the_media_semantic_change_in_the_lexicon_of_mechanization_in_19th-century_British_newspapers.

19. Rodda M., Probert P., McGillivray B. Vector space models of Ancient Greek word meaning, and a case study on Homer // Traitement Automatique des Langues. 2019. Vol. 60, iss. 3. P. 63–87. URL: https://aclanthology.org/2019.tal-3.4.pdf.

20. Sprugnoli R., Passarotti M., Moretti G. Vir is to Moderatus as Mulier is to Intemperans. Lemma Embeddings for Latin // Sixth Italian Conference on Computational Linguistics. Accademia University Press, 2019. P. 1–7. https://doi.org/10.5281/zenodo.3565572.

21. Weingart S. B., Meeks E. The Digital Humanities Contribution to Topic Modeling // The Journal of Digital Humanities. 2012. Vol. 2, no. 1. P. 2–6.

22. Wevers M., Koolen M. Digital begriffsgeschichte: Tracing semantic change using word embeddings // Historical Methods: A Journal of Quantitative and Interdisciplinary History. 2020. Vol. 53, iss. 4. P. 226–243. https://doi.org/10.1080/01615440.2020.1760157.