

Автоматическая датировка текстов на основе многомерного анализа

Захарова Софья

Темпоральная классификация — автоматическое определение временной принадлежности информации, представленной в документе, с использованием машинного обучения (Niculae & др., 2014).

Цель: разработать методику автоматической датировки текстов с использованием методов многомерного анализа, в частности факторного анализа

Основные шаги:

В рамках многомерного анализа:

- сформировать список языковых признаков
- исследовать ситуационные и языковые характеристики регистров
- выявить основные измерения, определяющие особенности текстов, и проинтерпретировать их с функциональной точки зрения

В рамках машинного обучения:

- обучить модель классификации на основе результатов многомерного анализа
- оценить эффективность разработанного метода
- лингвистически проинтерпретировать полученные результаты

Инструменты:

- Stanza,
- PyMorphu,
- spaCy,
- FactorAnalyzer,
- scikit-learn

Корпус	Вид	Количество
“Пишу тебе”	Почтовые открытки	46 000
“Прожито”	Дневниковые записи	20 000

Методика Д. Байбера

Многомерный анализ — это количественный подход, который позволяет исследовать регистры* по нескольким различным языковым параметрам — “измерениям” (Biber & Conrad, 2009).

Измерение (фактор) — группа совместно встречающихся языковых признаков, выполняющих одну общую функцию (Biber & Conrad, 2009).

Факторный анализ — вид многомерного анализа, методика взаимозависимости, когда исследуются взаимные связи между всеми переменными и причины, лежащие в их основе (Rencher, 2002).

* **Регистр** — языковая подсистема, связанная как с конкретной ситуацией использования, так и с языковыми характеристиками, которые выполняют важные коммуникативные функции в этой ситуации (Biber & Conrad, 2009: 31)

Этапы практической части исследования

1. Подготовка корпуса;
2. Анализ ситуационных характеристик регистров (участники, коммуникативная цель, тематика и т.д.);
3. Выбор языковых признаков и автоматическая разметка. Формирование списка из 101 признака;
4. Проведение факторного анализа и интерпретация полученных факторов;
5. Обучение классификатора на оценках текстов по каждому фактору. Интерпретация модели.

Ситуационные характеристики почтовых открыток

Участники:

- Количество адресатов и адресантов: один или группа
- Взрослый человек, родственник, имеет образование, иногда — титул
- Адресат всегда другой человек

Обстоятельства производства и понимания:

- Есть правила оформления, обязательные элементы, ограничения в объеме, есть возможность исправления текста

Коммуникативные цели:

- Информационная, установление и поддержание отношений

Темы:

- Текущие события, заслуживающие внимания получателя

Разметка языковых признаков

Примеры отобранных признаков:

- Переходные и непереходные глаголы
- Разнообразные падежные формы существительных и прилагательных
- Степени сравнений наречий и прилагательных
- Степень абстрактности существительных и прилагательных

Языковые особенности

1. Персональность & Адресация:

- Открытки: Высокая персональность (↑ я, ты/вы, мы глаголы), диалоговая функция, адресация к получателю.
- Дневники: ↑ он/она, они; повествование о других/событиях.

2. Временная система:

- Открытки: Ориентация на настоящее время.
- Дневники: Ретроспективность (↑ прош. время, ↑ сов. вид), акцент на завершенности.

3. Синтаксическая сложность:

- Открытки: Более длинные предложения, ↑ глубина дерева, ↑ длина глаг. группы → тенденция к формальности/структурированности (но с чертами устности).
- Дневники: ↑ предлогов → аналитичность.

4. Лексические особенности:

- Открытки: ↑ личные имена (персонализация), ↑ абстр. прилаг. (общая оценка), ↑ сравн. степ. прилаг. (пожелания), ↑ повелит. накл. (директивность: *напиши, приезжай*).
- Дневники: ↑ уменьш.-ласк. сущ. (эмоц./субъективность), ↑ наречий (описательность), ↑ непереходных глаголов (состояния, чувства, процессы), ↑ средний залог (фокус на авторе).

Результаты факторного анализа

Определены **9 факторов** и входящие в них языковые признаки.

1. “Контекстуализированное, рефлексивное повествование”;
2. “Динамичность и глагольность”;
3. “Структурная сложность против лаконичности”;
4. “Номинативная плотность и информационная насыщенность”;
5. “Интерактивность и директивность”;
6. “Описательная насыщенность”;
7. “Ритуализация против индивидуализации”;
8. “Формальность и лексико-стилистическая сложность”;
9. “Эмоциональная экспрессивность”.

Фактор “Ритуализация против индивидуализации”

Входящие признаки:

Ритуализация	Индивидуализация
<ul style="list-style-type: none">● местоимение 2 л. мн. числа● координации с существительным● глаголы в форме настоящего времени● глагол в форме 1 л. ед. ч.● глагол в форме 1 л. мн. числа● существительные среднего рода● существительное в Род. п.● прилагательные в Род. п.	<ul style="list-style-type: none">● существительные в Вин. п.● глаголы в форме прошедшего времени● совершенный вид глаголов● индекс абстрактности существительных● местоимение 1 л. ед. ч.

Фактор “Ритуализация против индивидуализации”

Ритуализация в основном соответствует текстам открытых писем.

Примеры: *Бабушка! Поздравляем Вас с праздником! Желаем здоровья, всех земных благ, благополучия. Нина, Аида, Коля; Дорогая моя Люция Алексеевна! Поздравляю Вас с Новым 1994 годом! Поздравляю и желаю: счастья, радости, крепкого здоровья, весёлого праздничного настроения! Ваша ученица Света.*

- **Односоставные определенно-личные полные предложения:** *Поздравляем Вас с праздником. Поздравляю Вас с Новым 1994 годом!*
- Использование **родительного падежа** для **этикетных формул** модели *Желаю(ем) + сущ в Род. п.:* *Желаем здоровья, всех земных благ, благополучия; желаю: счастья, радости, крепкого здоровья, весёлого праздничного настроения!*
- Высокий уровень **абстрактности:** частое употребление в поздравлениях лексем, называющих абстрактные понятия, связанных со сферой нравственно-этических оценок (Куликова, 2024): *здоровье, благо, благополучие, счастье, радость, настроение*

Фактор “Ритуализация против индивидуализации”

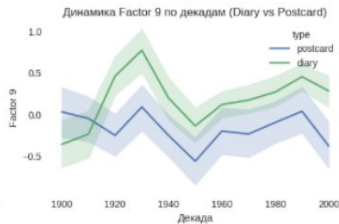
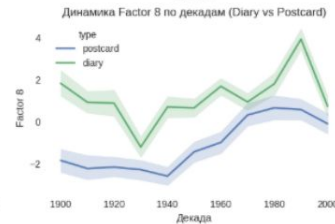
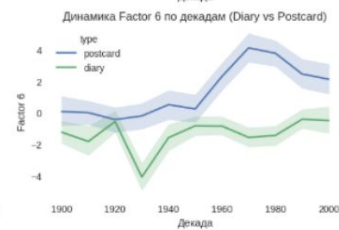
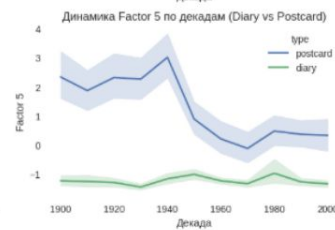
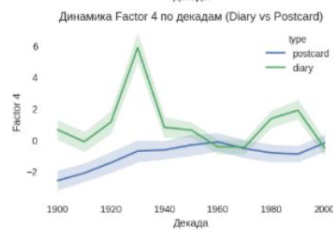
Индивидуализация в основном соответствует текстам дневниковых записей.

Примеры: *Я начал курс. Встретили молча, провожали шумным одобрением. Дичились друг друга — я их, они меня; Делали покупки: кресла, стулья. Привезли в Чегем. Потом поехали в ресторан «Чегемские водопады». Зашли в магазин, купили Кайсыну и Расулу туфли. Купили электрокамин.*

- Описание событий **от первого лица**: *я начал курс, я их [дичился]*
- Прямое воздействие на объекты (**Вин. п.**): *купили туфли, купили электрокамин*
- **Прошедшие события**, связанные с личным опытом: *делали покупки, поехали в ресторан, встретили молча, зашли в магазин*
- **Конкретные существительные**: *кресла, стулья, магазин, туфли, электрокамин*

Динамика факторов

- Фактор 1 — Контекстуализированное, рефлексивное повествование,
- Фактор 2 — Динамичность и глагольность,
- Фактор 3 — Структурная сложность против лаконичности,
- Фактор 4 — Номинативная плотность и информационная насыщенность,
- Фактор 5 — Интерактивность и директивность,
- Фактор 6 — Описательная насыщенность,
- Фактор 7 — Ритуализация против индивидуализации,
- Фактор 8 — Формальность и лексико-стилистическая сложность,
- Фактор 9 — Эмоциональная экспрессивность



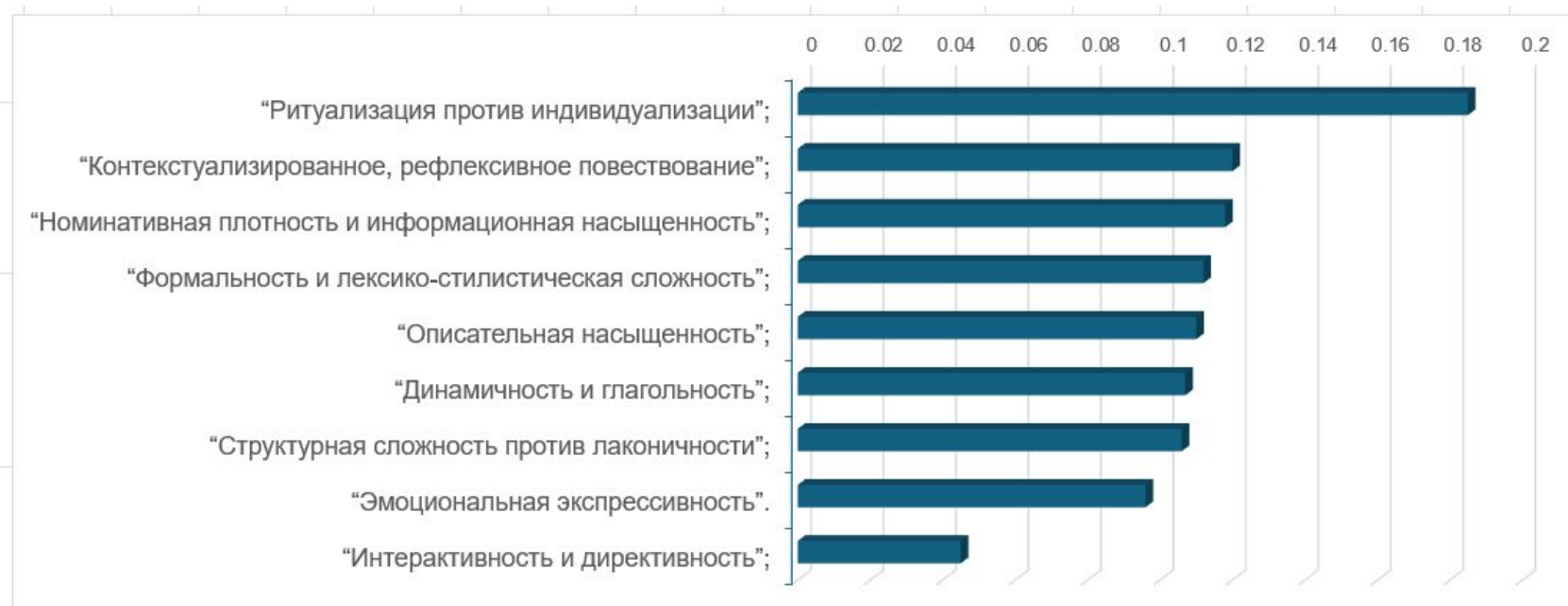
Результаты обучения модели на объединенном корпусе

Лучший результат удалось получить с помощью метода случайного леса (F1-мера после подбора гиперпараметров — **0.334**).

Период	F1-мера
1900–1919	0.432
1920-1939	0.458
1940–1959	0.376
1960-1979	0.387
1980-1999	0.408
2000-2019	0.088
2020-2025	0.189
ассигасу	0.392
среднее значение F1-меры	0.334

Отчет о классификации на валидационной выборке

Интерпретация результатов обучения: важность факторов



Оценка важности признаков с помощью метода среднего уменьшения примесей (Mean Decrease Impurity, MDI)

Интерпретация результатов обучения: динамика фактора “Формальность и лексико-стилистическая сложность”

Пример: *Вчера вынесен военно-окружным судом приговор по возмутительному делу братьев Ковалевских. Старший, — ранивший выстрелами из револьвера четырех человек, в том числе городского — приговорен к трем месяцам гауптвахты, без ограничения каких-либо прав; младший — посвящавший кулаком другого городского в рыцари — оправдан*

Период: 1900–1919

- Перенос **акцента** с субъекта на **действие** (*вынесен приговор, старший приговорен, младший оправдан*)
- **Причастия** для характеристики (*ранивший выстрелами, посвящавший в рыцари кулаком*)
- **Уточняющие конструкции** и причастные обороты (*в том числе городского, посвящавший кулаком другого городского в рыцари*)
- **Термины** (*гауптвахта, военноокружной суд, ограничение прав*)

Пример: *Затем нас перебросили и мы шли с боем на город Шапрон.*

Период: 1940–1959

- Отсутствие осложнений
- Короткие предложения

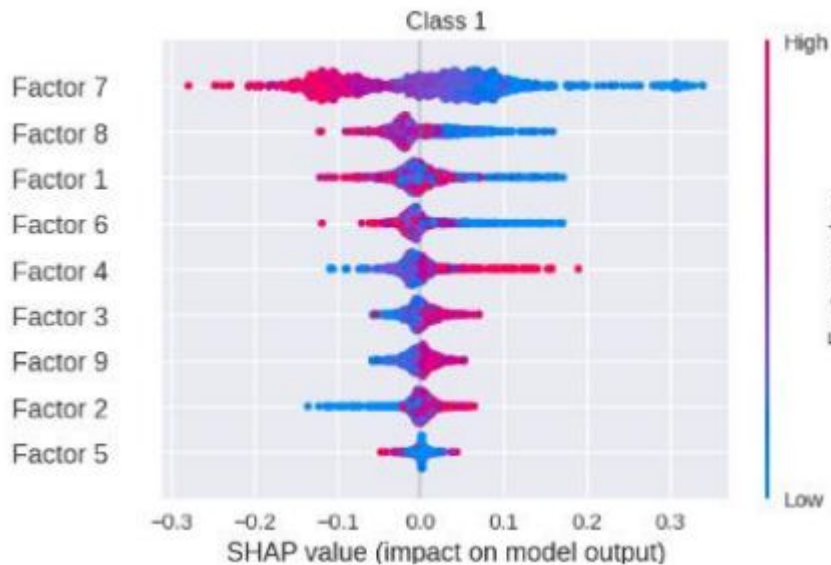
Интерпретация результатов обучения: моделирование языковых изменений

SHAP (Аддитивные объяснения Шепли, SHapley Additive exPlanations) присваивает каждому признаку значение важности для конкретного предсказания, основываясь на теории кооперативных игр Шепли (Lundberg & Lee, 2017).

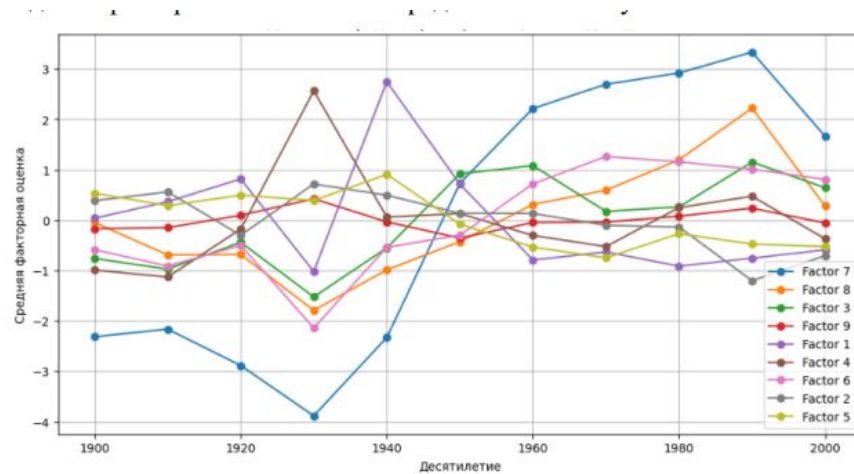
SHAP-графики позволяют оценить способность модели связывать языковые изменения и временной период.

В целом модель определила верную логику для классификации текстов по периодам. Неточность обнаруживается для периодов с 1940 по 1959 и с 1960 по 1979 (это же отражает классификационный отчет)

Интерпретация результатов обучения: моделирование языковых изменений



SHAP-график для Класса 1 — период 1920–1939



Динамика средних факторных оценок на сбалансированном датасете по десятилетиям

Создание регистровых классификаторов

- В ходе исследования были выявлены регистровые различия
- Были построены классификаторы для каждого регистра
- Модель, обученная на дневниковых записях, демонстрирует более высокие показатели по всем основным метрикам

Материал	F1-мера
объединенный корпус	0.334
открытые письма	0.254
дневниковые записи	0.417

F1-меры для трех классификаторов

Заключение

Успешно разработана и применена методика автоматической темпоральной классификации русскоязычных текстов, интегрирующая многомерный анализ языковых признаков и алгоритмы машинного обучения.

Создан адаптированный набор языковых признаков для многомерного анализа, написана программа на Python для автоматической разметки.

С помощью факторного анализа выявлены 9 функционально-коммуникативных факторов.

Модель продемонстрировала несколько ограниченную, но интерпретируемую предсказательную способность.

Эксперимент выявил существенные различия в чувствительности регистров к темпоральным изменениям.

Список литературы

1. Виноградов И. А. Новые датировки пятидесяти писем Н. В. Гоголя // Литературный факт. 2018. №10
2. Вяткина, С. В. (ред.). Синтаксис современного русского языка: учебник для высших учебных заведений РФ: под ред. С. В. Вяткиной. СПбГУ, 2009
3. Гловинская М. Я. Активные процессы в грамматике (на материале инноваций и массовых языковых ошибок) // Русский язык конца XX столетия (1985-1995) / Отв. ред. Е. А. Земская, М., 1996. С. 237-304
4. Золотова Г. А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка. М.: 2004 — 544 с
5. Кожина М. Н. Стилистика русского языка: учебник / М.Н. Кожина, Л.Р. Дускаева, В.А. Салимовский. — М. : ФЛИНТА, 2016.
6. Колобов В. В. Об эволюции содержания и стиля писательского дневника А. В. Жигулина // Филология и человек. 2017. №2.
7. Колокольникова О. Д. Эволюция стиля роберта бриджеса сквозь призму метафоры // Вестник Северного (Арктического) федерального университета. Серия: Гуманитарные и социальные науки. 2021. №3.

Список литературы

8. Лазаренко О. М. Лексические данные как основа относительной датировки отдельных частей Септуагинты // Индоевропейское языкознание и классическая филология. 2008
9. Тучкова Н. А. Проблема исторической датировки фольклорного материала // Вестн. Том. гос. ун-та. 2018. №432.
10. Abubakar M., Dr. Javed A.K., Dr. Shahzad Q., A Diachronic Study of Political Press Reportage in Pakistani Print Media: A Multidimensional Analysis. Migration Letters, 21(S11), 1529–1539, 2024
11. Ali M., Dr. Bashir A., Ali S., Aleem M. Studying multidimensional patterns of change overtime in writing letter to editor. PalArch's Journal of Archaeology of Egypt / Egyptology, 18(7), 2610–2621, 2021
12. Biber D. Variation across Speech and Writing. Cambridge: Cambridge University Press, 1988
13. Biber D., Conrad S. Register, Genre and Style. Cambridge, 2009.
14. Boldsen S. & Wahlberg F. Survey and reproduction of computational approaches to dating of historical texts. In Nordic Conference on Computational Linguistics (NoDaLiDa), pages 145–156. Linköping University Electronic Press, Sweden, 2021
16. Dalli A. & Wilks Y. Automatic Dating of Documents and Temporal Text Classification. In Proceedings of the Workshop on Annotating and Reasoning about Time and Events, pages 17–22, Sydney, Australia. Association for Computational Linguistics, 2006

Список литературы

17. Degaetano-Ortlieb, St. and Teich, E. 2016. Information-based Modeling of Diachronic Linguistic Change: from Typicality to Productivity. In Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 165–173, Berlin, Germany. Association for Computational Linguistics.
18. Halliday, M. A. K. (2013). Meaning as choice. In L. Fontaine, T. Bartlett, & G. O’Grady (Eds.), *Systemic Functional Linguistics: Exploring Choice* (pp. 15– 36). chapter, Cambridge: Cambridge University Press.
19. Maksimowicz E. Русский язык на рубеже XX–XXI веков. Лексико-семантические изменения. Białystok: Wydawnictwo Prymat, 2016
20. Niculae V., Zampieri M., Dinu L., & Ciobanu A. M.. 2014. Temporal Text Ranking and Automatic Dating of Texts. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, pages 17–21, Gothenburg, Sweden. Association for Computational Linguistics.
21. Ren H., Wang H., Zhao Y., Ren Y. Time-Aware Language Modeling for Historical Text Dating // Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, p. 13646–13656
22. Rencher, A. *Methods of Multivariate Analysis*, 2nd Edition, Wiley-Interscience, New York, 2002.



Введение

Актуальность: автоматическая датировка текста важна для многих областей, но для русскоязычных текстов разработана недостаточно. Кроме того, алгоритм темпоральной классификации текстов способен облегчить процесс их датировки

Новизна: адаптация и применение многомерного анализа для реализации темпоральной классификации текстов.

Объект: диахроническая вариативность русскоязычных письменных регистров (эволюция языковых норм, лексики, грамматики в почтовых открытках и дневниковых записях)

Предмет: автоматическая датировки текстов, основанная на анализе диахронической вариативности письменных регистров.

Этапы практической части исследования

1. Подготовка корпуса;
2. Анализ ситуационных характеристик регистров (участники, коммуникативная цель, тематика и т.д.);
3. Выбор языковых признаков и автоматическая разметка. Формирование списка из 101 признака;
4. Проведение факторного анализа и интерпретация полученных факторов;
5. Обучение классификатора на оценках текстов по каждому фактору. Интерпретация модели.