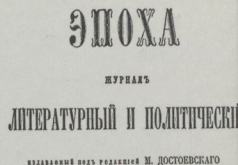


80
178
CA-389248



ПРОБЛЕМА АТРИБУЦИИ В ЖУРНАЛАХ «ВРЕМЯ», «ЭПОХА» И ЕЖЕНЕДЕЛЬНИКЕ «ГРАЖДаниН»

✓

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

А. А. Рогов, Р. В. Абрамов, Д. Д. Бучнева, О. В. Захарова,
К. А. Кулаков, А. А. Лебедев, Н. Д. Москин,
А. В. Отливанчик, Е. Д. Савинов, Ю. В. Сидоров

ca-389248

Проблема атрибуции в журналах «Время», «Эпоха» и еженедельнике «Гражданин»



✓

Издательство «Острова»
Петрозаводск

2021
Государственное бюджетное
учреждение культуры
«Оренбургская областная универсальная
научная библиотека им. Н.К. Крупской»

Введение

Одной из значимых задач в современных исследованиях с применением статистического и формально-грамматического анализа текстов является определение авторства анонимных и псевдонимных статей в журналах «Время», «Эпоха» и еженедельнике «Гражданин». Решением проблемы атрибуции этих статей занимались: В. Л. Комарович, Л. П. Гроссман, О. фон Шульц, В. С. Нечаева, Б. В. Томашевский, В. В. Виноградов, Г. М. Фридлендер, В. А. Туниманов, В. Н. Захаров, В. А. Викторovich, И. Зохраб и др.

Любые используемые методики атрибуции текстов (в том числе основанные на критериях, позаимствованных из сферы точных наук) способны вызвать со стороны специалистов различные возражения и несогласие. Причин этому немало — неправильно сформированные параметры распределения случайных величин; излишняя избирательность и предусмотрительность при выборе анализируемых текстов, позволяющие получить нужный и заранее ожидаемый результат; малый или слишком большой объем привлекаемого лингвистического материала и т. п. В настоящее время бесспорной математической методики, позволяющей однозначно определить авторство текста, не существует.

В предлагаемой монографии продемонстрирована практическая реализация различных математических методов, позволяющих выделить основные особенности авторского стиля в литературно-критических статьях и публицистических материалах, опубликованных в журналах «Время», «Эпоха» и еженедельнике «Гражданин», а также сделать выводы, связанные с принадлежностью анонимных и псевдонимных текстов Ф. М. Достоевскому. В ней комбинируются традиционные филологические и инновационные математические подходы к анализу

следующего анализа, а также демонстрировать разбиванные тесты и результаты анализа широкому кругу пользователей. Благодаря использованию клиент-серверной технологии возможна организация доступа к данным из любой точки мира и, следовательно, организация коллабораций различных групп исследователей. Полученные результаты исследований могут быть независимо подтверждены благодаря доступу к исходным данным.

4.1. Описание метода

Для решения задач атрибуции текстов хорошо зарекомендовали себя следующие математические методы: нейронные сети, метод накопительных сумм QSUM, «Деревья решений», машина опорных векторов (SVG), метод k -средних, байесовский классификатор, марковские цепи, метод главных компонент, дискриминантный анализ, генетические алгоритмы, статистические критерии (хи-квадрат Пирсона, критерии Стьюдента, Колмогорова-Смирнова) и др. Отметим зарубежных авторов, которые применяли математические методы для решения задачи атрибуции текстов: T. C. Mendenhall, A. Q. Morton, J. M. Farrington, B. Efron, R. Thisted, W. J. Teahan, C. E. Chaski, E. Stamatatos, P. Juola, R. D. Peng, T. Joachims, J. J. Diederich, C. Apte, D. Lowe, R. Matthews, F. J. Tweedie, O. de Vel, S. Argamon, S. Levitan, R. Zheng [Calle-Martin], [Farrington], [Батура], [Романов, 2010], [Stamatatos, 2009b], [Baayen]. Среди этих методов интеллектуального анализа данных «Деревья решений» выделяются тем, что они просты в понимании и интерпретации, а также не требуют специальной предварительной обработки данных. Построение «Дерева решений» позволяет однозначно определить, какие из признаков являются наиболее существенными, а потому требующими дополнительной интерпретации в аспекте лингвистического анализа, — они будут вынесены в вершину «Дерева решений». На примере таких «деревьев» можно легко показать, как множество текстов, попавших в некоторый узел, разбивается на два подмножества, в одно из которых попадают объекты, удовлетворяющие некоторому правилу, а в другое — неудовлетворяющие.

«Деревом решений» называется ациклический граф, по которому производится классификация объектов (в нашем случае текстов), описанных набором признаков. Каждый узел «дерева» содержит условие

Таблица 41

**Относительные частоты биграммы
в текстах Ф. М. Достоевского**

Номер текста	Общее кол-во биграмм	Количество биграммы «прилагательное + существительное»	Относительная частота биграммы «прилагательное + существительное»
35	1815	100	0,0551
195	5098	216	0,0424
196	4603	292	0,0634
136	6869	270	0,0393
44	4563	185	0,0405
38	2348	111	0,0473
126	4078	197	0,0483
268	929	37	0,0398
40	2510	114	0,0454
25	304	22	0,0724
32	584	30	0,0514
43	1983	64	0,0323
86	1377	55	0,0399
99	7450	337	0,0452
145	1303	44	0,0338
153	4353	166	0,0381
154	3074	115	0,0374
155	3737	179	0,0479
156	4507	144	0,0320
157	2377	105	0,0442
164	3920	209	0,0533
163	3016	135	0,0448
166	2239	103	0,0460
201	1695	77	0,0454

Для эксперимента по методике Г. Хетсо эталонными для Ф. М. Достоевского стали 25 текстов писателя, подписанных им или безусловно принадлежащих ему⁵¹.

Одним из методов атрибуции анонимных и псевдонимных текстов в изданиях братьев Достоевских является «Трансформер»⁵². Исследование с использованием нейросети «Трансформер» потребовало обучения искусственного интеллекта на текстах семи авторов: Ф. М. Достоевский, М. М. Достоевский, Н. Н. Страхов, А. А. Григорьев, Я. П. Полонский и др. При обучении нейросети «Трансформер» тексты Ф. М. Достоевского составляли одну группу, тексты остальных авторов — другую. Для атрибуции анонимных текстов, опубликованных во «Времени» и «Эпохе», использовался алгоритм построения ациклического графа, основанный на выделении биграмм (двух частей речи) — метод «Лес решений», который подробно описан в Главе 4-й.

В результате эксперимента по всем параметрам лингвостатистического анализа по методике Г. Хетсо Ф. М. Достоевскому может быть приписана статья «Письмо Постороннего критика в редакцию нашего журнала по поводу книг г. Панаева и “Нового поэта”» (Время. 1861. № 1. Подп.: «Посторонний критик»). О. фон Шульц атрибутировал этот текст Достоевскому [Schoultz: 7]. В. С. Нечаева вслед за И. Ф. Масановым и Б. В. Томашевским приписала эту статью А. Ф. Писемскому [Нечаева, 1975: 261]. В. А. Туниманов согласился с атрибуцией О. фон Шульца и предложил включить ее в раздел Dubia (ДЗ0, 27: 400). Исследование Г. Хетсо подтвердило атрибуцию О. фон Шульца [Хетсо, 1986: 68]. Сегод-

⁵¹ Список использованных текстов см. в параграфе 3-м Главы 1-й — «Анализ ошибки 2-го рода».

⁵² Об этой нейросети см. подробнее параграф 4-й Главы 5-й — «Методы машинного обучения».

Таблица 60

Результаты тестирования на оригинальность исходных текстов
 batch-32, embed-32, hidden-16, head-2, Embeds-55000, Блоки-1

Точность (на тестовых)	0,91	0,97	0,97	0,97
Точность (обучающих)	0,94	0,93	0,91	0,91

6.1. О возможности применения статистических методов к разным уровням текста в аспекте определения авторского стиля

Выявлять авторство текста можно различными способами; помимо классического привлечения экспертов, способных определить авторство спорного текста с учетом определенных стилистических характеристик и языковых особенностей, данную проблему можно попытаться решить с применением формальных методов, основанных на анализе статистических данных. Благодаря этим методам анализа текста можно с высокой долей достоверности решать вопросы атрибуции анонимных и псевдонимных текстов. Более того, в последние десятилетия появилась возможность применять в данной работе сложные вычисления, которые позволяют проанализировать текст с учетом разных его аспектов (в том числе и тех, которые потенциально могут оказаться «скрытыми» для эксперта, работающего традиционными методами), а также получить полезную информацию для последующей лингвистической и литературоведческой интерпретации текста.

6.2. О возможности применения статистических методов в аспекте определения авторского стиля

Т. В. Батура в своей работе «Формальные методы определения авторства текстов» выделяет пять уровней текста, на которых можно использовать методы атрибуции:

- пунктуационный уровень (особенности использования знаков препинания и типичные ошибки);

Оглавление

Введение	3
Глава 1. Проблемы атрибуции в изучении Достоевского	7
1.1. Литературная подпись Достоевского.....	8
1.2. Какие тексты без имени и под псевдонимами принадлежат Достоевскому?	24
Глава 2. Использование методики атрибуции Г. Хетсо	83
2.1. Описание методики	84
2.2. Критические исследования методики.....	89
2.3. Анализ методики на наличие ошибки второго рода	92
Глава 3. Система «СМАЛТ» и ее возможности	103
3.1. Общее описание системы	104
3.2. Размеченный корпус текстов XIX века в дореволюционной и современной графике	114
3.3. Программный модуль ИС СМАЛТ для формально- грамматической и синтаксической разметки текстов в дореволюционной и современной графике	125
3.4. Программный модуль ИС СМАЛТ для статистической обработки текстов.....	134
3.5. Программный модуль ИС СМАЛТ для визуализации результатов обработки текстов	138
Глава 4. Использование «Деревьев решений» и случайных «Лесов решений» для задачи атрибуции	141
4.1. Описание метода.....	142
4.2. Числовые характеристики метода	144
4.3. Задача атрибуции.....	174
4.4. Выявление сильных позиций в тексте	179
4.5. Поиск неоднородных фрагментов в тексте.....	182
4.6. Проблема несбалансированности эталонных текстов при атрибуции с помощью биграмм.....	190
4.7. Применение «Леса решений»	196
Глава 5. Применение других методов для атрибуции текстов, включая методы машинного обучения (искусственного интеллекта)	223
5.1. Описание методики оценки парной связи грамматических классов.....	224

5.2. Проведенные на базе методики оценки парной связи грамматических классов исследования	230
5.3. Анализ методики оценки парной связи грамматических классов	234
5.4. Методы машинного обучения	242
Глава 6. Статистические модели стиля писателя	249
6.1. О возможности применения статистических методов к разным уровням текста в аспекте определения авторского стиля	250
6.2. О возможности применения статистических методов в аспекте определения авторского стиля	250
6.3. Использование статистических методов, основанных на анализе комбинаций частей речи	252
6.4. Использование статистических методов в решении задач атрибуции анонимных и псевдонимных статей	254
6.5. Использование статистических методов, основанных на последовательностях частей речи, в определении стиля Ф. М. Достоевского	261
Глава 7. Проблема атрибуции анонимных и псевдонимных текстов в журналах «Время» и «Эпоха»	273
Глава 8. Проблема атрибуции анонимных и псевдонимных текстов в еженедельнике «Гражданин»	291
8.1. История и современное состояние исследований	292
8.2. Исследование текстов «Гражданина» с использованием статистических методов	304
Заключение	346
Приложения	347
Приложение 1	348
Приложение 2	351
Приложение 3	353
Приложение 4	355
Приложение 5	356
Приложение 6	358
Приложение 7	359
Приложение 8	364
Приложение 9	365
Приложение 10	367
Список литературы	372

Конец ознакомительного фрагмента

Уважаемый читатель!

**Размещение полного текста данного
произведения невозможно в связи с ограничениями
по IV части ГР РФ.**

**Эту книгу вы можете почитать в Оренбургской
областной универсальной научной библиотеке
им. Н. К. Крупской по адресу: г. Оренбург,
ул. Советская, 20; тел. для справок: (3532) 60-61-28**