Юрислингвистика

Legal Linguistics, 2025, 38, 62-73, doi: https://doi.org/10.14258/leglin(2025)3811

ЛИНГВОЭКСПЕРТОЛОГИЯ *УДК 81'33, ББК 81.1, ГРНТИ 16.31.61, Код ВАК 5.9.8*

Статистическое исследование цифровых следов в киберпространстве: языковые маркеры деструктивных сообществ

И. Д. Мамаев[1, 2], М. А. Марусенко[3], В. В. Петров[4]

[1]Балтийский государственный технический университет «Военмех» им. Д. Ф. Устинова
ул. 1-я Красноармейская, 1, 190005, Санкт-Петербург, Россия. E-mail: mamaev_id@voenmeh.ru
Санкт-Петербургский государственный университет
Университетская наб., д. 11, 199034, Санкт-Петербург, Россия. E-mail: [2]i.mamaev@spbu.ru, [3]m.marusenko@spbu.ru
[4]22-я линия Васильевского Острова, 7, 199105, Санкт-Петербург, Россия. E-mail: vadim.petrov@spbu.ru

В связи с повсеместным распространением информационно-коммуникационных технологий растет и тенденция к увеличению агрессивного и деструктивного контента, который представлен в текстовом формате, в связи с чем возникает потребность в разработке новых методов выявления и описания групп лиц, оставляющих подобные «цифровые следы». В настоящей работе описана процедура лингвистического профилирования текстов, основанная на системно-структурном изучении языковых параметров и их количественном исчислении. Для эксперимента собран тестовый набор данных, представленный текстами с платформ, на которых потенциально публикуется деструктивный контент. Корпус кластеризирован методом k-means, были определены тематически ориентированные группы текстов – проекции деструктивных сообществ пользователей. С помощью языка программирования Python реализован алгоритм, который включает предобработку текстов, вычисление статистических связей между языковыми характеристиками и, наконец, определение уровня значимости, который позволяет утверждать о характерных языковых признаках деструктивных сообществ. Установлены значимые зависимости между коэффициентом лексической плотности и частотой употребления глаголов/имен существительных, а также type-token ratio (коэффициент лексического богатства, отношение количества уникальных слов к общему количеству слов в тексте). Приведена стилеметрическая характеристика представленных кластеров.

Ключевые слова: криминалистика, деструктивные группы, экстремистский контент, профилирование автора текста, социальные сети, языковые маркеры.

A Statistical Study of Digital Footprint in Cyberspace: Language Markers of Destructive Communities

I. D. Mamaev[1, 2], M. A. Marusenko[3], V. V. Petrov[4]

[1]Baltic State Technical University "Voenmeh" named after D. F. Ustinov
1 1st Krasnoarmeyskaya, 190005, St. Petersburg, Russia. E-mail: mamaev_id@voenmeh.ru
St. Petersburg State University
11 Universitetskaya Emb., 199034, St. Petersburg, Russia. E-mail: [2]i.mamaev@spbu.ru, [3]m.marusenko@spbu.ru
[4]7 22nd Line V.O., 199105, St. Petersburg, Russia. E-mail: vadim.petrov@spbu.ru

With the widespread use of information and communication technologies, the tendency to spread aggressive and destructive content presented in text format is growing, which creates a need to develop new methods for detecting and describing groups of people who leave such "digital footprint". This paper presents a procedure for linguistic profiling of texts based on a systemic and structural study of language parameters and their quantitative calculation. For the experiment, a test dataset has been collected, it is represented by texts from platforms where potentially destructive content is published. The corpus is clustered using the k-means method, and topically oriented groups of texts, which are the projections of destructive user communities, are detected. Using the Python programming language, the authors implement an algorithm that includes text preprocessing, calculating statistical relationships among language characteristics and, finally, determining the level of significance, which allows asserting the

characteristic language tendencies of destructive communities. Significant dependencies between the coefficient of lexical density and the frequency of use of verbs/nouns, as well as the type-token ratio (the coefficient of lexical richness, the ratio of the number of unique words to the total number of words in the text) are established. The stylistic characteristics of the presented clusters are given.

Key words: forensics, destructive communities, extremist content, authorship profiling, social networks, language markers.

В последние годы проблема противодействия экстремистскому контенту приобрела особую актуальность, поскольку цифровая среда предоставляет преступным сообществам удобные инструменты для распространения идеологических установок. А. Л. Осипенко верно отмечает, что в этой связи «важно организовать прогнозирование изменений оперативной обстановки в киберпространстве на основе ее мониторинга» [Осипенко 2017: 187].

В судебной лингвистике существует сложность однозначной идентификации текстов, которые публикуются членами деструктивных сообществ. Традиционные методы экспертизы зачастую не обеспечивают оперативный анализ больших массивов данных, в связи с чем исследовательский фокус смещен в сторону применения автоматизированных методов лингвистического анализа. Мы вполне согласны со следующим положением, которое высказано В. Д. Пристансковым, А. Г. Харатишвили и Ю. А. Евстратовой: «С ростом числа киберпреступлений по всему миру возрастает необходимость их эффективного и своевременного выявления и расследования... Анализ логов или обычных данных недостаточен в работе с большими данными информации (Big Data). Искусственный интеллект, с его возможностью обрабатывать и анализировать большие объемы информации, может предоставить новые тактико-технические приемы (инструменты) в борьбе с киберпреступностью…» [Пристансков, Харатишвили, Евстратова 2023: 588].

В настоящей статье предлагается алгоритм, который позволяет выявлять маркеры, характерные для интернет-дискурса, создаваемого деструктивными сообществами. Используемые методы направлены на описание языковых закономерностей в текстах деструктивных сообществ. Цель исследования – тестирование алгоритма на корпусных данных.

Социальные сети и другие онлайн-платформы активно используются экстремистскими группами для распространения пропаганды и вербовки сторонников, что активно обсуждается в научных трудах как с позиций поведенческих моделей, так и с позиций языковых параметров. В исследовании [Бакиров, Грязнов, Валиахметов 2023] социально-психологические особенности членов организованных преступных групп сводятся к шести факторам: реалистичность, эгоцентричность, рефлексивность, нормативность, эмпатийность и конформность. На основе этих характеристик выделяются два социально-психологических типа членов организованных преступных групп: ресоциализирующийся (появление новых качеств, которые способствуют быстрой адаптации к обществу) и десоциализирующийся типы (проявление черт социальной дезинтеграции). Анализ этих типов позволяет интерпретировать их как уровни готовности к социальной интеграции и классифицировать как ресоциализирующие и десоциализирующие типы. Работа [Ананьева, Девяткин, Кобозева, Смирнов 2016], напротив, посвящена лингвистической составляющей, а именно – количественному анализу экстремистских текстов с целью выявления характерных лексико-грамматических особенностей. Основное внимание уделено распределению ключевых слов (или выражений), тематических категорий текстов и их последующей обработке методами машинного обучения, однако, по заверению авторов, «тексты не являются линейно-разделимыми с помощью представленной системы маркеров» [Ананьева, Девяткин, Кобозева, Смирнов 2016: 213].

Методы стилеметрии, которые изначально применялись к определению авторства неизвестного текста или отдельных фрагментов художественных произведений [Марусенко, Бессонов, Богданова, Аникин, Мясоедова 2001], могут быть адаптированы под задачи исследования участников экстремистских сообществ. Т. А. Литвинова провела серию экспериментов на основе текстов пользователей форума, который внесен в Федеральный список экстремистских материалов РФ. В исследовании [Литвинова 2020] показаны различия в лексическом выборе мужчин и женщин, которые являются участниками экстремистских групп. В целом мужчины чаще обсуждают оружие, Россию и правоохранительные органы, а женщины – литературу и выражение чувств, хотя это наблюдение не всегда применимо к отдельным авторам. Развивая данное направление, Т. А. Литвинова показывает, что тексты, которые созданы участникам экстремистских форумов, близки к текстам лиц, которые обладают низкими показателями на уровне экстраверсии и уровне эмоциональной лабильности [Литвинова 2019].

Для решения задач выявления преступных сообществ разрабатываются и отдельные сетевые ресурсы – лексические онтологии, которые являются формализованным способом представления лексических единиц естественного языка и их семантики. В статье [Resende de Mendonça, Felix de Brito, de Franco Rosa, dos Reis, Bonacin 2020] предложена онтология жаргонных выражений преступного мира. Сформирован корпус более 8 миллионов испаноязычных твитов – коротких текстовых сообщений в сети X (Twitter)[1] – с целью применения онтологии к проблеме обнаружения преступных сообщений и намерений пользователей, которые их опубликовали. Например, для корпусного поста *Quero pó e loló de cafe da manha* (рус. *Я хочу на завтрак кокаиновый порошок и аэрозольный наркотик*) автоматически определяется категория выражения желаний, а также сочетания, указывающие на пристрастие автора к запрещенным веществам.

Таким образом, представленные исследования демонстрируют разнообразие подходов к анализу экстремистских текстов. Настоящая работа дополняет существующие подходы: в ней предлагается применение системно-структурного лингвистического алгоритма для выявления языковых маркеров, характерных для риторики определенных сообществ. Предположение о наличии относительно однообразных лингвистических признаков основано на теории дискурсивных

[1] Социальная сеть X (Twitter) заблокирована на территории Российской Федерации на основании федерального закона «Об информации, информационных технологиях и о защите информации», согласно которому связи с призывами к массовым беспорядкам допустима блокировка.

сообществ, описанной Джоном Свейлзом в 1988 году. В его работе [Swales 1988: 212-213] под дискурсивным сообществом понимается некоторая группа, члены которой, разделяя общий идеологический конструкт, неизбежно воспроизводят сходные риторические стратегии и языковые формулы. Это подтверждается, в частности, рядом исследований повседневной риторики радикально настроенных лиц [Васильева, Майборода, Ясавеев 2017].

Для проведения эксперимента по лингвистическому профилированию социально опасных групп необходимо создать специализированный датасет из открытых источников сети Интернет. На территории РФ (на момент проведения эксперимента – июль 2025 г.) сохраняется доступ к ряду зарубежных веб-корпусов, которые содержат русскоязычные данные, в частности – к Araneum Russicum III Maximum в корпусном менеджере NoSketch Engine [Benko, Zakharov 2016]. Из общего объема корпуса более 15 миллиардов словоупотреблений были отобраны тексты по поиску в поле Simple Query по словам негативного смыслового содержания: «убить», «революция», «ненависть», «насилие», «война», «беспредел». С последующим ручным анализом метаданных был собран тестовый корпус с информационного портала Hatewall общим объемом 36 постов (4824 словоупотреблений). Однородность лингвистического корпуса обеспечивалась двумя факторами – единством источника и единством социально-идеологической направленности (ручная верификация каждого текста на соответствие целевой семантике ненависти и агрессии, заданной поисковым запросом). Некоторая информация с рассматриваемого сайта уже признана экстремистской Министерством Юстиции РФ [Министерство юстиции Российской Федерации. Экстремистские материалы. Hatewall: URL, материалы под номерами 2748 и 3130], поэтому настоящий метод автоматизированного системно-структурного анализа языковых параметров в дальнейшем может послужить базой для усовершенствования существующих методик судебной лингвистики в целях исследования лингвистической экспертизы деструктивных текстов.

С использованием языка программирования Python был разработан специализированный скрипт, решающий эту задачу. На первом этапе установлены необходимые библиотеки, такие как stanza для обработки текста, pandas, numpy для работы с данными и matplotlib для визуализации результатов. На втором этапе загружается собранный корпус в формате Excel в виде матрицы, у которой два столбца: в первом отображены обобщенные метки текстов, а во втором – сами тексты. Далее создана функция, которая анализирует текст и извлекает различные морфосинтаксические и лексические метрики: среднее количество употреблений имен существительных, среднее количество употреблений имен прилагательных, среднее количество употреблений глаголов, среднее количество употреблений наречий, средняя длина предложений, средняя длина предложных конструкций, type-token ratio (коэффициент лексического богатства, отношение количества уникальных слов к общему количеству слов в тексте) и коэффициент лексической плотности. Несмотря на то, что подобные метрики традиционно применяются в задачах атрибуции текста, в данном исследовании вектор направлен не на идентификацию стиля дискурсивного сообщества, а на выявление статистически значимых особенностей, которые характеризуют дискурс сообщества в целом и в дальнейшем потенциально могут применяться в программах автоматической модерации онлайн-сообществ. В скрипте для каждой метрики выполняется тест на нормальность распределения с использованием теста Шапиро-Уилка. Строится матрица корреляций между метриками. В зависимости от результатов теста нормальности и типа данных выбирается соответствующий метод корреляции: для количественных данных с нормальным распределением (далее в работе используется обозначение normal) и линейной связью – корреляция Пирсона, а для случаев нарушений нормальности (далее в работе используется обозначение non-normal) или нелинейных связей – коэффициент ранговой корреляции Спирмена. После статистического анализа выводится тепловая карта, которая отражает системно-структурные связи между исследуемыми метриками и информацию о значимости этих корреляций. В случае значимых корреляций (в эксперименте отмечены три уровня значимости: * для 0.05, ** для 0.01 и *** для 0.001) выводятся корреляционные графики с линией тренда. Значимость коэффициента корреляции определяется через проверку статистической гипотезы о том, что корреляция между переменными отличается от нуля в генеральной совокупности, применяется t-статистика.

Совокупность значимых корреляций понимается как лингвистический профиль сообщества – формализованная модель, количественные показатели морфосинтаксических и лексических параметров которой используются для интерпретации текстов корпуса и прогнозирования дискурсивной активности. Системное представление этих параметров в виде матрицы корреляций и их визуализация позволяют реконструировать диагностические признаки, которые отличают исследуемую группу от других сообществ (подробнее о методологии написано в [Мамаев, Митрофанова, Петров, Марусенко 2025]).

В законодательстве РФ понятие «деструктивное сообщество» напрямую не прописано, однако в Федеральном законе от 25 июля 2002 года № 114-ФЗ «О противодействии экстремистской деятельности» [Федеральный закон от 25.07.2002 №114-ФЗ «О противодействии экстремистской деятельности» URL], Федеральном законе от 6 марта 2006 года № 35-ФЗ «О противодействии терроризму» [Федеральный закон от 6 марта 2006 года № 35-ФЗ «О противодействии терроризму» URL], статьях 280 и 282 Уголовного кодекса РФ [Уголовный кодекс Российской Федерации от 13.06.1996 N 63-ФЗ URL] и ряде других официальных документов указано, что все положения, которые касаются групп лиц, пропагандирующих насилие, экстремизм, терроризм или оказывающих психологическое воздействие на участников, подлежат регуляции через соответствующие законы. С точки зрения прикладной лингвистики настоящий корпус текстов деструктивного сообщества был разделен на тематические группы, которые затем подвергались процедуре лингвистического профилирования. Для этого был разработан скрипт кластеризации корпуса методом k-means, который оптимально подходит для анализа корпусов различного объема и менее ресурсоемок, чем альтернативные алгоритмы [Ahmed, Tiun, Omar, Sani 2022]. В качестве признаков для кластеризации использовались векторные представления текстов, которые построены на основе частотности лексем (bag-of-words) с нормализацией TF-IDF. Таким образом, тексты объединялись в кластеры по максимальному сходству их лексического состава: чем ближе распределение лексем и их весов в двух текстах, тем меньше расстояние между ними в многомерном пространстве признаков. Оптимальное количество кластеров определялось с помощью метода «локтя»: для разных значений числа кластеров вычислялась сумма внутрикластерных расстояний (инерция) и строился график зависимости этой величины от количества кластеров. В точке, где уменьшение инерции переставало быть значительным,

образуя характерный «излом», фиксировалось оптимальное число кластеров. И для исходных, и для предобработанных текстов эта точка приходилась на значение k = 4, что указывает на наличие четырех устойчивых тематических сообществ внутри корпуса (см. рис. 1 и 2). Для наглядности распределения текстов по кластерам их многомерные векторы были спроецированы в двумерное пространство методом главных компонент (PCA), что позволило визуализировать границы и плотность каждого кластера.

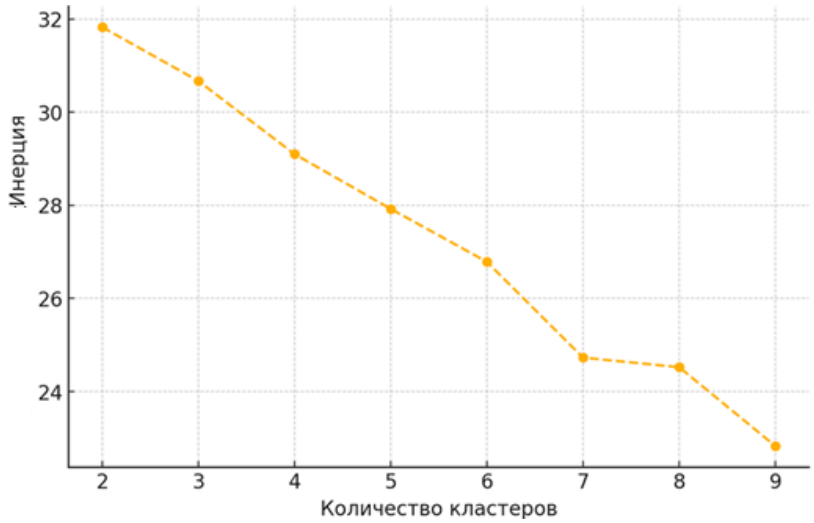

Рис. 1. Визуализация «метода локтя» для определения количества кластеров в корпусе необработанных текстов

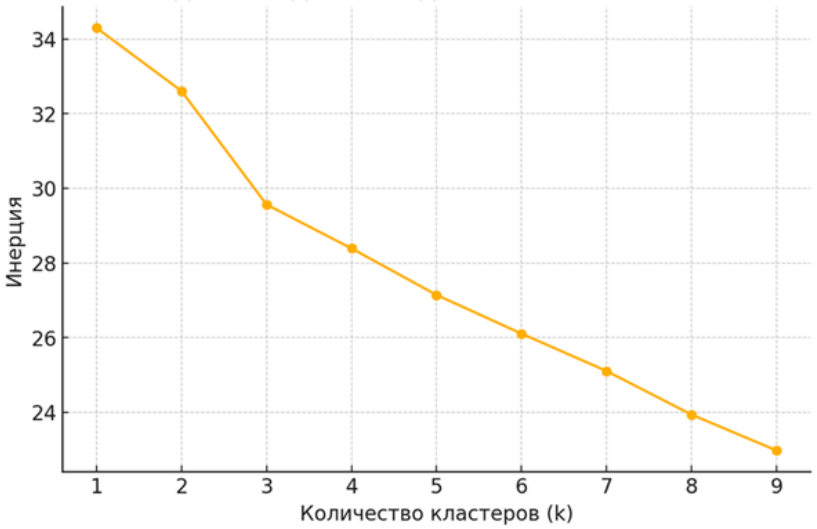

Рис. 2. Визуализация «метода локтя» для определения количества кластеров в корпусе обработанных текстов

Результаты кластерного анализа (рис. 3 и 4) показывают, что в «сырых» данных кластер 4 представлен наибольшим количеством вхождений – 12 текстов, за ним идет кластер 2 – 11 текстов, кластер 3 – 8 текстов, кластер 1 – 5 текстов. После обработки данных распределение меняется. Кластер 4 продолжает занимать лидирующую позицию: в нем оказывается 17 текстов. В кластере 1 количество текстов увеличилось до 8. Кластер 2 теряет часть текстов, общий объем сокращается до 7. Наконец, в кластере 3 оказалось 4 текста. Доля постов, которые остались в своем кластере после проведения процедур обработки, составляет 38.9%, что указывает на относительную чувствительность корпуса к деструктуризации текстов, при этом кластер 4 сохранил наибольшее количество текстов – 7, а кластер 3 оказался полностью перераспределенным. Общее распределение текстов по кластерам представлено в таблице 1 (RAW – необработанные тексты, PROCESSED – обработанные тексты).

Таблица 1

Принадлежность текстов к кластерам в зависимости от типа входных данных

Файл	Кластер (RAW)	Кластер (PROCESSED)
extremist_post_1.txt	2	2
extremist_post_2.txt	1	1
extremist_post_3.txt	4	4
extremist_post_4.txt	2	2
extremist_post_5.txt	2	2
extremist_post_6.txt	1	1
extremist_post_7.txt	1	1

extremist_post_8.txt	3	1
extremist_post_9.txt	2	1
extremist_post_10.txt	4	4
extremist_post_11.txt	4	4
extremist_post_12.txt	2	3
extremist_post_13.txt	2	3
extremist_post_14.txt	2	3
extremist_post_15.txt	2	3
extremist_post_16.txt	3	4
extremist_post_17.txt	4	4
extremist_post_18.txt	2	1
extremist_post_19.txt	4	2
extremist_post_20.txt	4	4
extremist_post_21.txt	4	2
extremist_post_22.txt	3	4
extremist_post_23.txt	4	4
extremist_post_24.txt	3	4
extremist_post_25.txt	1	4
extremist_post_26.txt	2	4
extremist_post_27.txt	4	1
extremist_post_28.txt	4	1
extremist_post_29.txt	3	4
extremist_post_30.txt	1	4
extremist_post_31.txt	2	2
extremist_post_32.txt	4	2
extremist_post_33.txt	3	4
extremist_post_34.txt	3	4
extremist_post_35.txt	4	4
extremist_post_36.txt	3	4

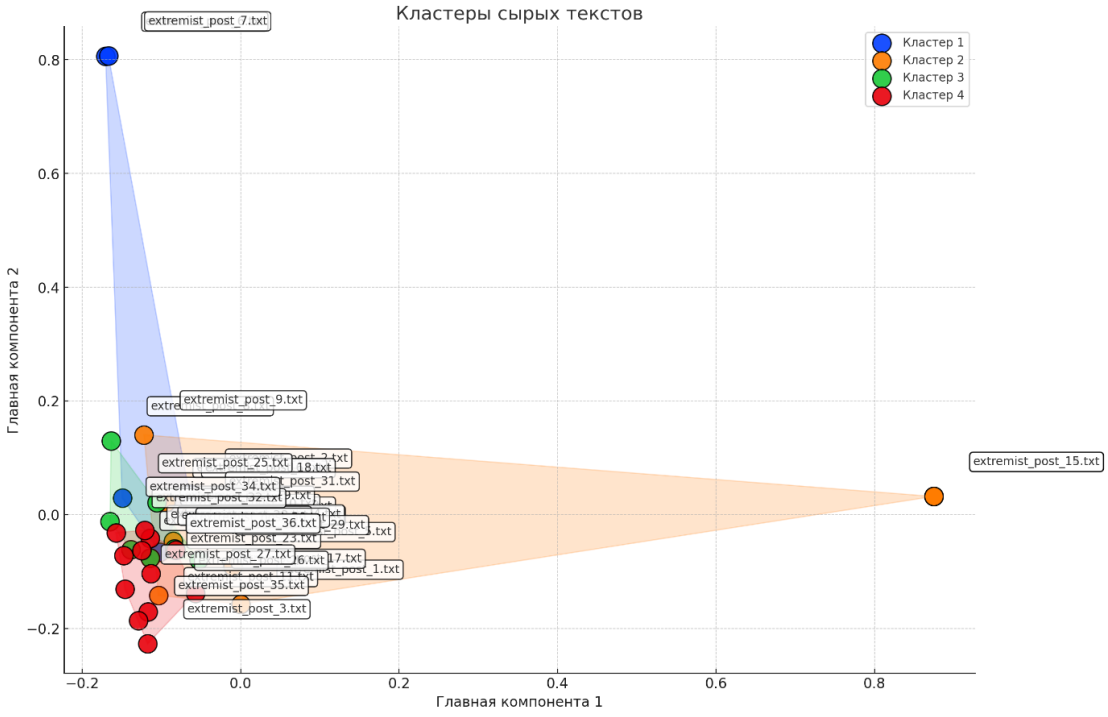

Рис. 3. Кластеры необработанных текстов корпуса

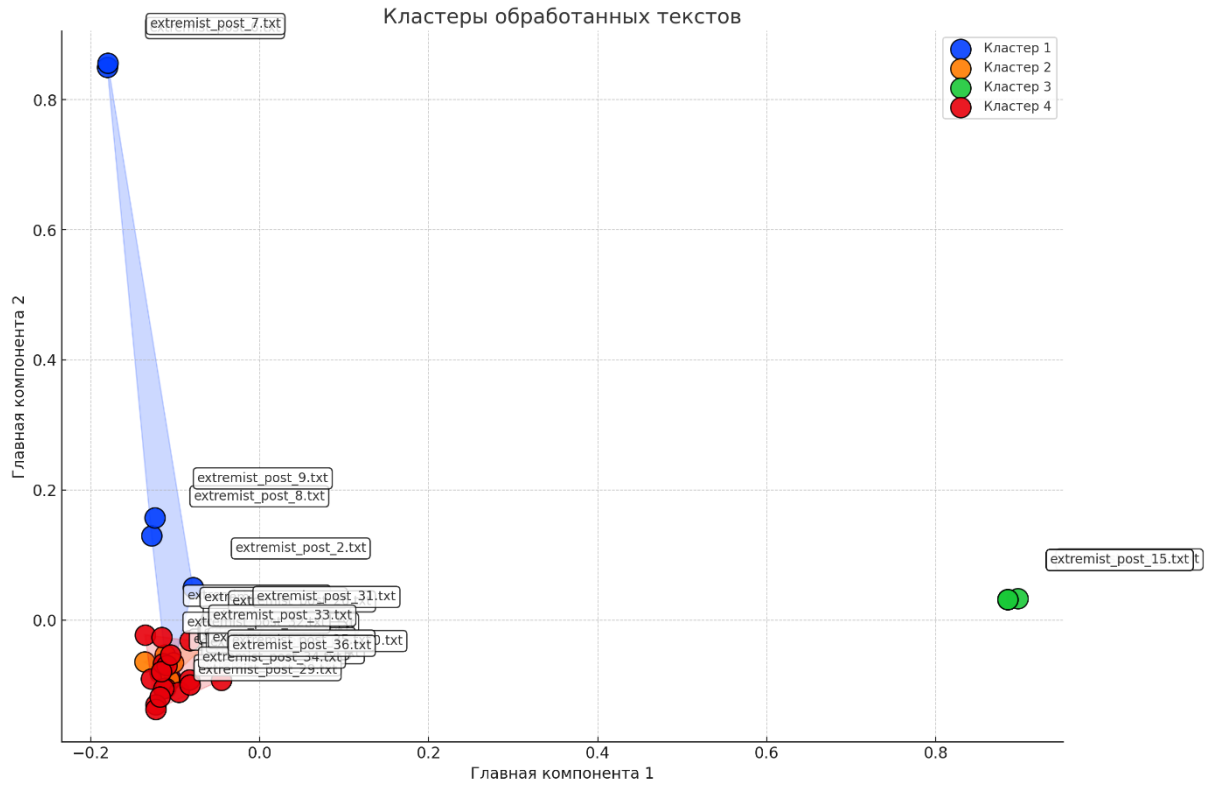

Рис. 4. Кластеры обработанных текстов корпуса

Выявленные сообщества могут быть описаны с двух точек зрения: уголовно-наказуемой и лингвистической. С уголовно-наказуемой позиции тексты кластеров 1, 2 и 3 (RAW и PROCESSED) потенциально связаны со статьей 280 УК РФ (экстремизм): *«…у вас один выход и это революция!…»*[2] (extremist_post_6.txt), *«Да я уверен, что если бы началась война вы бы, патриоты, забились бы по углам и сидели стонали, а не пошли бы в бой…»* (extremist_post_1.txt), *«…Сейчас у нас намечается чуть ли ни революция, буряты хотят добиться отставки главы республики……»* (extremist_post_2.txt). 18 текстов (50% всего корпуса) как раз можно рассматривать с позиции упомянутой статьи. Напротив, кластер 4 (RAW и PROCESSED) наиболее разнородный: тексты потенциально можно рассматривать с позиции статьей 105 УК РФ (убийство) – *«…Хотела убить её за это, потому, что это мой кот…»* (extremist_post_23.txt), 135 УК РФ (нарушение норм сексуальной морали) – *«…мне всего 45, но ненависть к мужским *** нас обьединяет…»* (extremist_post_15.txt), а также со статьями 280 и 282 (разжигание ненависти) УК РФ – *«…Я ненавижу жирных баб! Куда вы столько жрете??…»* (extremist_post_19.txt). При этом к общим собственно языковым признакам постов относятся политически маркированные слова и сочетания *(«революция», «война», «отставка главы республики»)*, негативно окрашенные единицы *(«ненавижу», «жирных баб», «патриоты… забились бы по углам»)* и применение ряда риторико-стилистических приемов: гиперболы (*«чуть ли ни революция»*) и сарказма (*«Куда вы столько жрете??»*).

На рисунках 5 и 6 на материале необработанных текстов отмечается несколько фрагментов, которые могут дать представление о системно-структурных лингвистических особенностях (значимых языковых корреляциях) текстов отдельных деструктивных сообществ.

[2] Прим. авторов: в настоящей статье примеры публикаций приводятся с сохранением оригинальной орфографии и пунктуации, обсценная лексика замаскирована.

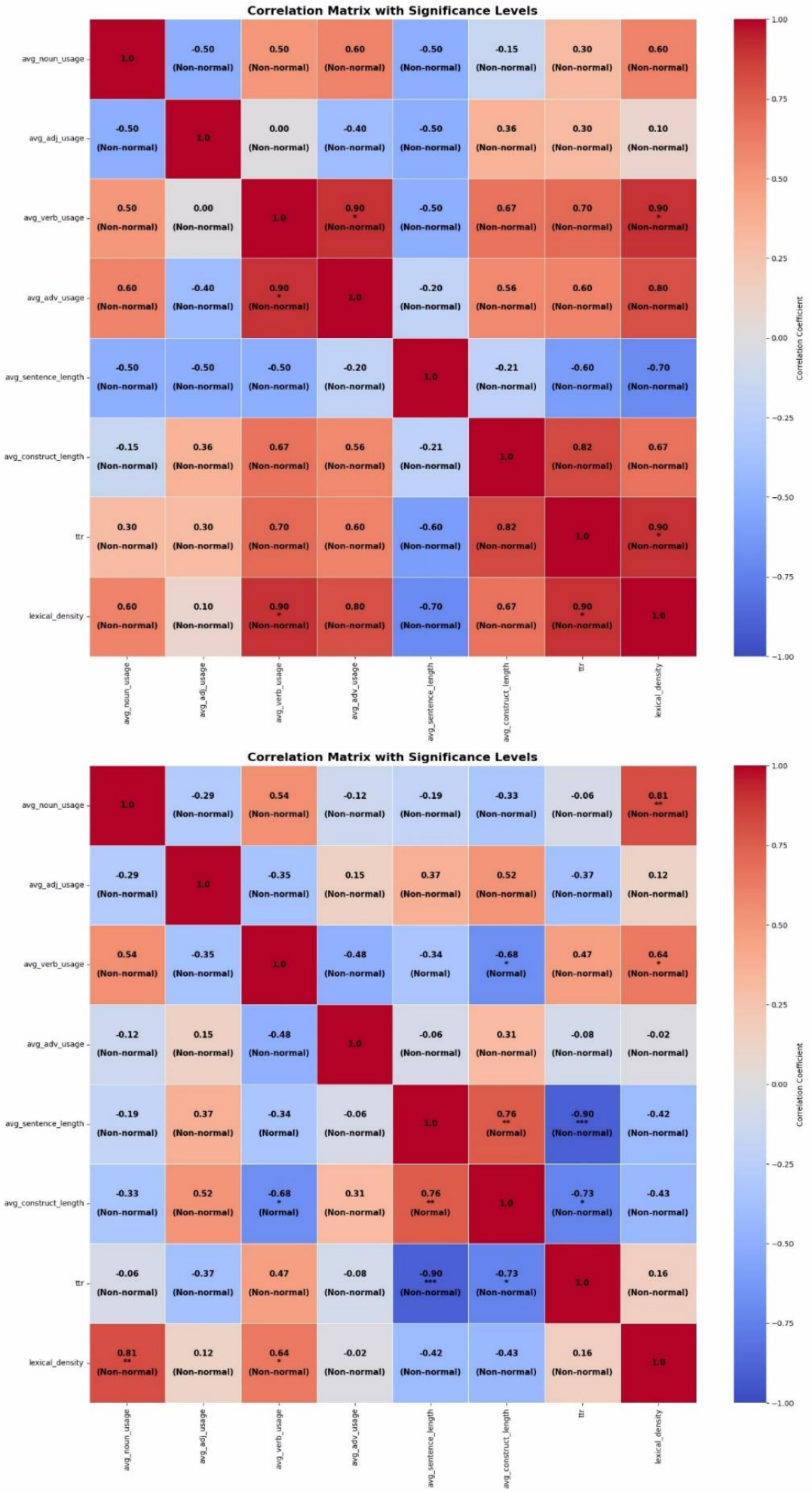

Рис. 5. Лингвистические профили кластеров (сообществ) 1 и 2, полученных на основе необработанных текстов корпуса

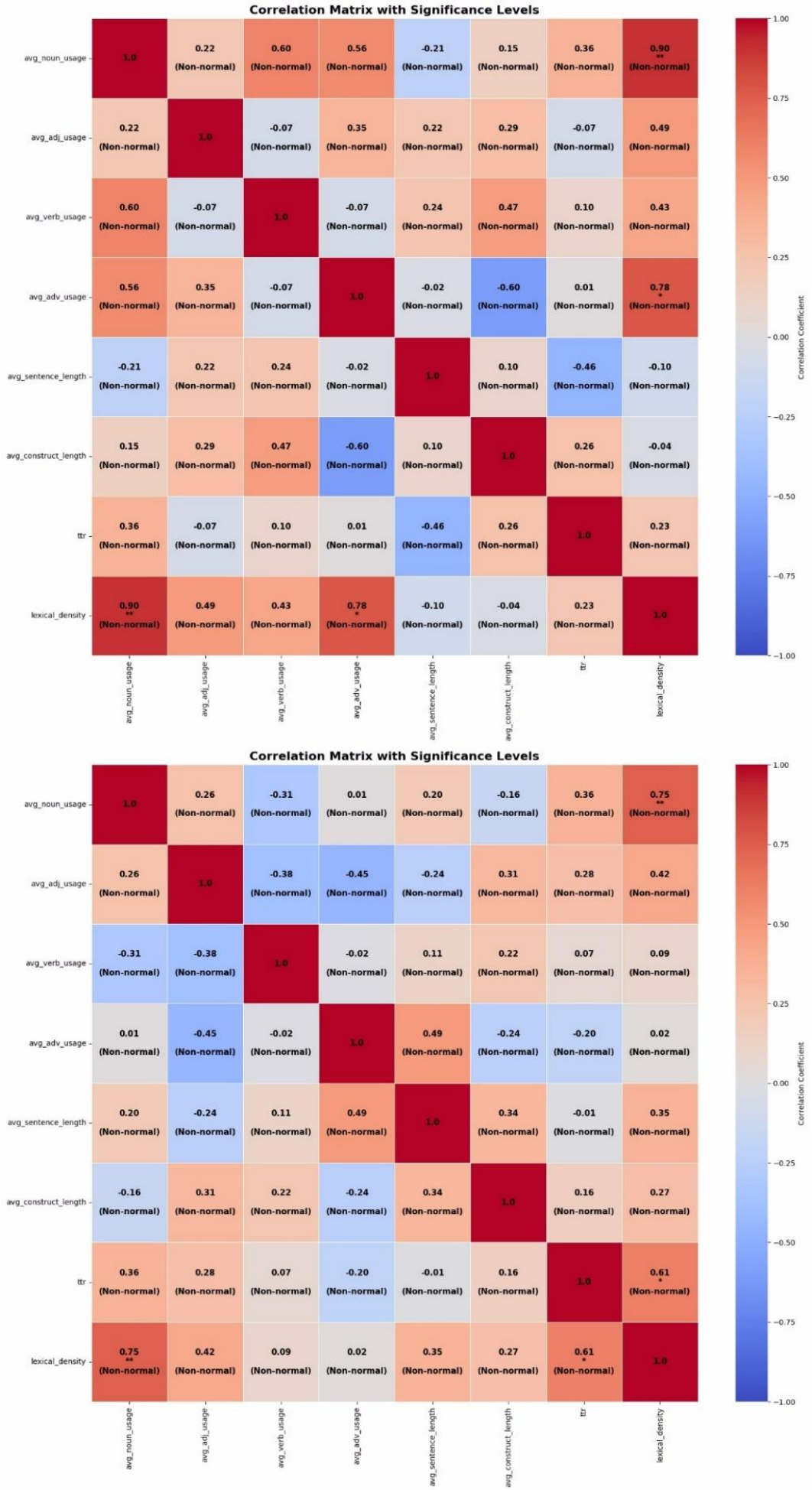

Рис. 6. Лингвистические профили кластеров (сообществ) 3 и 4, полученных на основе необработанных текстов корпуса

В первом кластере наблюдается высокая положительная корреляция между коэффициентом лексической плотности и средним употреблением глаголов ($k = 0.90$, $\alpha = 0.05$), что указывает на аргументативный характер создаваемых пользователями платформы сообщений: *«…убейте *** и всех его приспешников и тогда США и Евросоюз будут к вам благосклонны…»* (extremist_post_6.txt). Во втором кластере эта же корреляция обладает средней силой связи ($k = 0.64, \alpha = 0.05$), а корреляция лексической плотности и среднего количества употребления существительных выражена сильнее ($k = 0.81, \alpha = 0.01$), что указывает на тенденцию пользователей создавать более номинализированными тексты без акцента на действия: *«А на этот вопрос: "Почему ты не любишь Россию?" есть только один типичный ответ: "Да там же грязь, насилие и беспредел."»* (extremist_post_18.txt). В кластерах 3 и 4 коэффициент корреляции лексической плотности и среднего количества употребления существительных соизмерим с показателями в кластере 2 ($k = 0.90, \alpha = 0.01$ и $k = 0.75, \alpha = 0.01$ соответственно), корреляции лексической плотности и среднего количества употребления глаголов не значима. Еще одной значимой корреляцией, которая прослеживается кластерах 1 и 4, является связь между type-token ratio и лексической плотностью ($k = 0.90, \alpha = 0.05$ и $k = 0.61, \alpha = 0.01$ соответственно), график корреляции для указанного коэффициента в кластере 4 представлен на рисунке 7. В кластере 4 наблюдается превалирование коэффициентов корреляции, которые стремятся к 0 или к отрицательным значениям, что связано с тематической неоднородностью полученной группы.

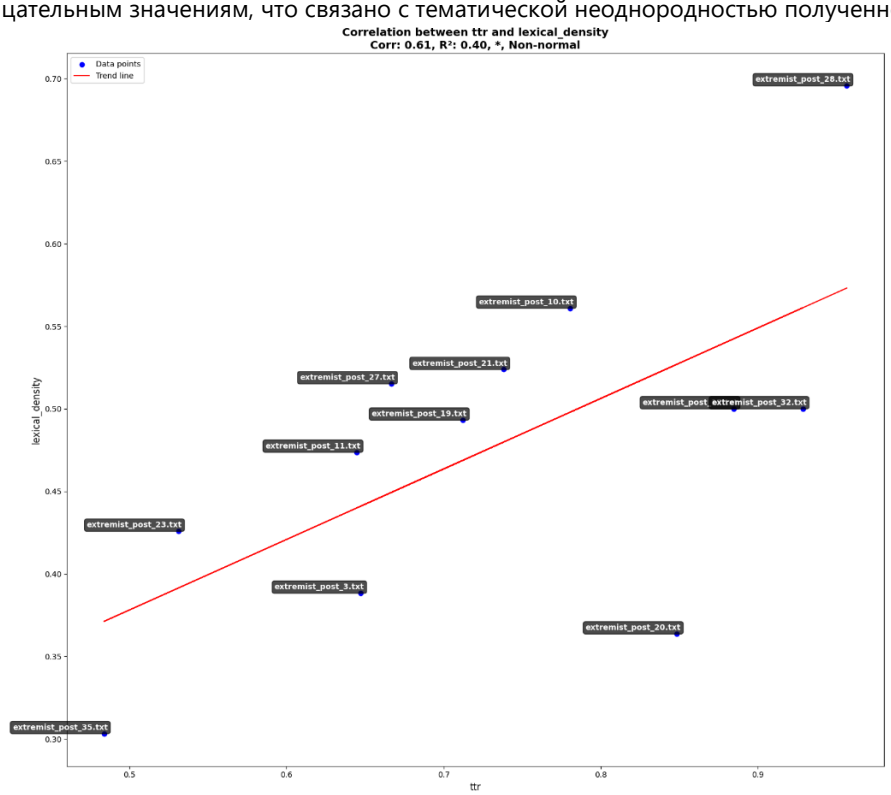

Рис. 7. График корреляции между лексической плотность и type-token ratio в кластере 4

Общая тенденция сохранения профилей с незначительными изменениями уровней значимости и колебаниями коэффициентов корреляции сохраняется в лингвистических профилях, которые получены для сообществ на основе кластерного анализа обработанных текстов. Кластер 3 не был подвергнут процедуре лингвистического профилирования, поскольку выборка постов для сообщества значительно мала (4 текста).

Итоговым этапом эксперимента стал стилеметрический анализ полученных кластеров с помощью программы Stylo в среде R. В качестве базовых единиц анализа были выбраны слова (words). В связи с небольшими размерами корпуса частотный диапазон слов (Most Frequent Words, MFW) был установлен в пределах от 16 до 60 слов с шагом 1. Метод culling использовался для устранения редких и слишком частотных слов, что позволяло сосредоточиться на стабильных стилеметрических признаках. Минимальный порог был установлен на уровне 0%, максимальный – 20%, шаг – 4. При визуализации результатов в виде консенсусного дерева (рис. 8) оказалось, что наиболее единообразным в стилистическом плане является кластер 3 (синий цвет), так как ветви расположены близко друг к другу, а соединительные линии минимальной длины, что свидетельствует об использовании близких языковых конструкций участниками данной преступной группы. В кластере 2 (зеленый цвет) также наблюдается достаточно высокая однородность элементов. Разреженная структура характерна для кластеров 1 и 4, при этом, как уже отмечалось, для кластера 4 характерна гетерогенная тематическая наполненность, а также практически полное отсутствие значимых корреляций.

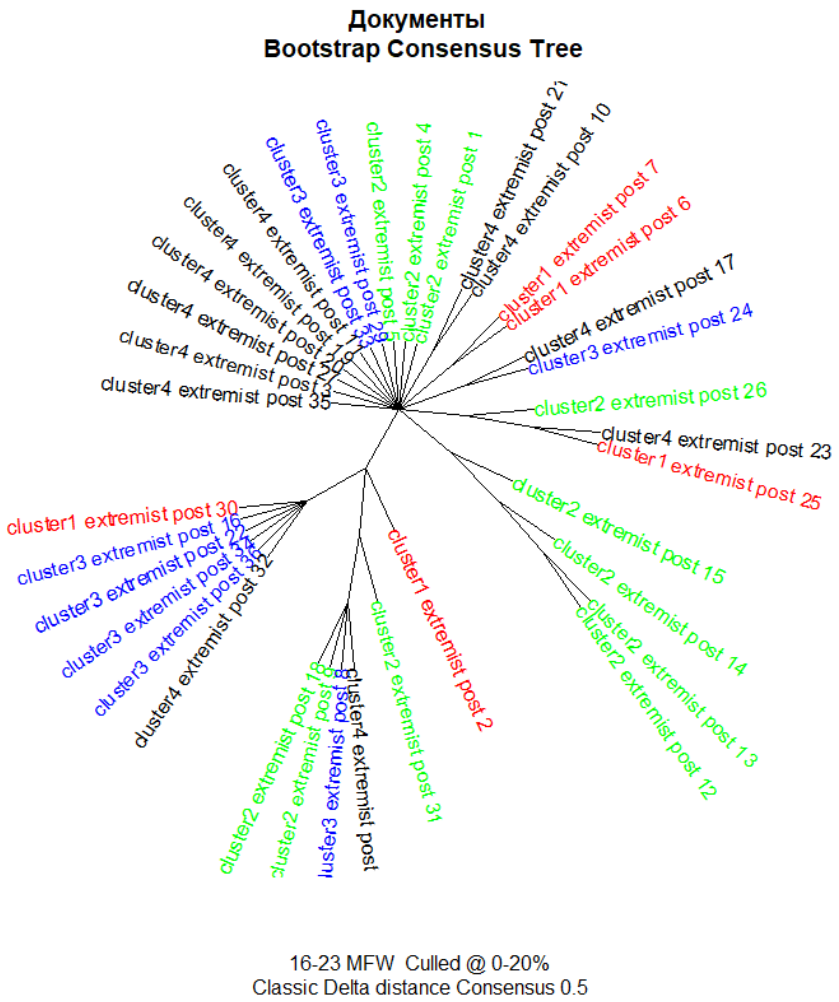

Рис. 8. Результаты стилеметрического анализа корпуса

В процессе исследования с применением процедур кластеризации тексты корпуса были сгруппированы по тематикам, каждая из которых представляет отдельный аспект деструктивной деятельности. К каждому сообществу применен алгоритм лингвистического профилирования: статистически значимые корреляции показывают, что дискурс сообщества тяготеет к высокой лексической плотности при повышенной частоте глаголов либо существительных, что указывает на преобладание аргументативных или номинализированных стратегий выражения агрессии. Результаты исследования пригодны для автоматизации процедур судебно-лингвистических экспертиз и могут служить основой систем автоматического мониторинга сообществ. Дальнейшие шаги предполагают расширение корпуса и обсуждение этико-правовых рамок применения цифровых технологий.

Особое внимание в будущих исследованиях следует уделить этическим и правовым аспектам внедрения цифровых лингвистических технологий в судебную практику. Автоматизированные методы анализа текстов должны использоваться с учетом соблюдения принципов конфиденциальности, объективности и законности, а итоговое слово остается за экспертом [Володин 2025].

Литература

Ананьева М. И., Девяткин Д. А., Кобозева М. В., Смирнов И. В. Лингвостатистический анализ текстов экстремистской направленности / Ситуационные центры и информационно-аналитические системы класса 4i для задач мониторинга и безопасности: материалы Международной конференции (SCVRT2015–16). – 2016. – С. 210-213.

Бакиров Р. Р., Грязнов А. Н., Валиахметов И. Р. Социально-психологическая типология членов ОПГ с позиции их готовности к социальной интеграции / Общество, государство, личность: применение научных знаний и технологий в решении социально-экономических задач региона. – 2023. – С. 101-111.

Васильева Н. В, Майборода А. В., Ясавеев И. Г. «Почему уходят в ИГИЛ[3]?»: дискурс-анализ нарративов молодых дагестанцев / Социологическое обозрение. – 2017. – Т. 16. – №. 2. – С. 54-74.

Володин Е. А. Особенности внедрения искусственного интеллекта в судебные процессы: автоматизация и цифровизация правоприменения / Юридическая наука. – 2025. – №. 4. – С. 85-89.

[3] ИГИЛ — «Исламское государство Ирака и Леванта», организация, деятельность которой запрещена на территории Российской Федерации.

Литвинова Т. А. Компаративное исследование текстов участников экстремистского форума и лиц с известными психологическими характеристиками с использованием методов стилеметрического анализа / Известия Воронежского государственного педагогического университета. – 2020. – №. 1. – С. 168-175.
Литвинова Т. А. Стилеметрическое исследование текстов участников экстремистского форума: гендерный аспект / Известия Воронежского государственного педагогического университета. – 2019. – №. 4. – С. 227-236.
Мамаев И. Д., Митрофанова О. А., Петров В. В., Марусенко М. А. Методы автоматического выявления и анализа дискурса сообщества иностранных агентов в цифровой среде (лингвокриминалистический аспект) / Филологические науки. Вопросы теории и практики. – 2025. – Т. 18. – №. 7. – С. 3106-3115.
Марусенко М. А., Бессонов Б. Л., Богданова Л. М., Аникин М. А., Мясоедова Н. Е. В поисках потерянного автора: Этюды атрибуции. СПб., 2001.
Министерство юстиции Российской Федерации. Экстремистские материалы. Hatewall. URL: https://www.minjust.gov.ru/ru/extremist-materials/?q=hatewall
Осипенко А. Л. Организованная преступная деятельность в киберпространстве: тенденции и противодействие / Юридическая наука и практика: Вестник Нижегородской академии МВД России. – 2017. – №. 4 (40). – С. 181-188.
Пристанскров В. Д., Харатишвили А. Г., Евстратова Ю. А. Искусственный интеллект – новая форма использования специальных знаний в расследовании и раскрытии киберпреступлений / Всероссийский криминологический журнал. – 2023. – Т. 17, № 6. – С. 586-596.
Уголовный кодекс Российской Федерации от 13.06.1996 N 63-ФЗ (ред. от 28.12.2024) (с изм. и доп., вступ. в силу с 08.01.2025). URL: https://www.consultant.ru/document/cons_doc_LAW_10699/
Федеральный закон от 25.07.2002 №114-ФЗ «О противодействии экстремистской деятельности». URL: https://www.consultant.ru/document/cons_doc_LAW_37867/
Федеральный закон от 6 марта 2006 года № 35-ФЗ «О противодействии терроризму». URL: https://www.consultant.ru/document/cons_doc_LAW_58840/
Ahmed M. H., Tiun S., Omar N., Sani N.S. Short text clustering algorithms, application and challenges: A survey / Applied Sciences. – 2022. – Vol. 13. – No. 1. – Pp. 1-38.
Benko V., Zakharov V. P. Very Large Russian Corpora: New Opportunities and New Challenges / Computational Linguistics and Intellectual Technologies. – Российский государственный гуманитарный университет. – 2016. – Pp. 79-93.
Resende de Mendonça R., Felix de Brito D., de Franco Rosa F., dos Reis J. C., Bonacin R. A framework for detecting intentions of criminal acts in social media: A case study on twitter / Information. – 2020. – Vol. 11. – No. 3. – Pp. 1-40.
Swales J. Discourse communities, genres and English as an international language / World Englishes. – 1988. – Vol. 7. – No. 2. – Pp. 211-220.

References

Artanyeva, M. I., Devyatkin, D. A., Kobozeva, M. V., Smirnov, I. V. (2016). Linguostatistical analysis of extremist texts. In Situational centers and information-analytical systems of the 4i class for monitoring and security tasks: proceedings of the International Conference (SCVRT2015–16), 210-213 (in Russian).
Bakirov, R. R., Gryaznov, A. N., Valiahmetov, I. R. (2023). Socio-psychological typology of organized crime group members in terms of their readiness for social integration. In Society, State, Individual: application of scientific knowledge and technologies in solving regional socio-economic problems, 101-111 (in Russian).
Vasilyeva, N. V., Maiboroda, A. V., Yasaveev, I. G. (2017). "Why do they go to ISIL?": a discourse analysis of young Dagestanians' narratives. Russian Sociological Review, 16(2), 54-74 (in Russian).
Volodin, E. A. (2025). Features of the introduction of artificial intelligence in lawsuits: automation and digitalization of law enforcement. Legal Science, 4, 85-89 (in Russian).
Litvinova, T. A. (2020). A comparative study of texts of participants of extremist forum and persons with known psychological characteristics using methods of stylometric analysis. Izvestiya of Voronezh State Pedagogical University, 1, 168-175 (in Russian).
Litvinova, T. A. (2019). A stylometric study of the extremist forum posts: gender dimension. Izvestiya of Voronezh State Pedagogical University, 4, 227-236 (in Russian).
Mamaev, I. D., Mitrofanova, O. A., Petrov, V. V., Marusenko, M. A. (2025). Methods of automatic detection and analysis of the discourse of the foreign agent community in the digital environment (a forensic linguistic aspect). Philology. Theory & Practice, 18(7), 3106-3115 (in Russian).
Marusenko, M. A., Bessonov, B. L., Bogdanova, L. M., Anikin, M. A., Myasoedova, N. E. (2001). In search of the lost author: studies of attribution. St. Petersburg (in Russian).
Ministry of Justice of the Russian Federation. Extremist materials. Hatewall. Available from: https://www.minjust.gov.ru/ru/extremist-materials/?q=hatewall (in Russian).
Osipenko, A. L. (2017). Organized criminal activities in cyberspace: trends and fighting. Legal Science and Practice: Journal of Nizhny Novgorod Academy of the Ministry of Internal Affairs of Russia, 4(40), 181-188 (in Russian).
Pristanskov, V. D., Kharatishvili, A. G., Evstratova, Yu. A. (2023). Artificial intelligence – a new form of using special knowledge in investigatingand solving cybercrimes. Russian Journal of Criminology, 17(6), 586-596 (in Russian).
Criminal Code of the Russian Federation dated 13.06.1996 No. 63-FZ (as amended on 28.12.2024, effective from 08.01.2025). Available from: https://www.consultant.ru/document/cons_doc_LAW_10699/ (in Russian).
Federal Law dated 25.07.2002 No. 114-FZ "On Counteracting Extremist Activity". Available from: https://www.consultant.ru/document/cons_doc_LAW_37867/ (in Russian).

Federal Law dated March 6, 2006 No. 35-FZ "On Counteracting Terrorism". Available from: https://www.consultant.ru/document/cons_doc_LAW_58840/ (in Russian).
Ahmed, M. H., Tiun, S., Omar, N., Sani, N. S. (2022). Short text clustering algorithms, application and challenges: a survey. Applied Sciences, 13(1), 1-38.
Benko, V., Zakharov, V. P. (2016). Very large Russian corpora: new opportunities and new challenges. Computational Linguistics and Intellectual Technologies. Russian State University for the Humanities, 79-93 (in Russian).
Resende de Mendonça, R., Felix de Brito, D., de Franco Rosa, F., dos Reis, J. C., Bonacin, R. (2020). A framework for detecting intentions of criminal acts in social media: a case study on Twitter. Information, 11(3), 1-40.
Swales, J. (1988). Discourse communities, genres and English as an international language. World Englishes, 7(2), 211-220.

Citation:
Мамаев И. Д., Марусенко М. А., Петров В. В. Статистическое исследование цифровых следов в киберпространстве: языковые маркеры деструктивных сообществ // Юрислингвистика. – 2025 – 38. – С. 62-73.
Mamaev I. D., Marusenko M. A., Petrov V. V. (2025) A Statistical Study of Digital Footprint in Cyberspace: Language Markers of Destructive Communities. Legal Linguistics, 38, 62-73.