УДК 81`322

Шишкин Борис Андреевич
ассистент департамента лингвистики
Северо-Кавказский федеральный университет
г. Ставрополь, Россия

Boris Shishkin
Assistant of the Department of Linguistics
North Caucasus Federal University
Stavropol, Russia
boris-shishkin.work@yandex.ru

ОПЫТ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ И КЛАСТЕРИЗАЦИИ В ИССЛЕДОВАНИИ ВИРТУАЛЬНОГО МЕДИАДИСКУРСА

В данной работе рассматривается контаминированный подход к проведению тематического моделирования и кластеризации текстов сетевого пространства с целью осуществления комплексного исследования виртуального медиадискурса. Базой исследования выступили отобранные методом сплошной выборки публикации из сетевого сообщества Северного Кавказа. Приводится опыт реализации автоматизированной систематизации эмпирического материала при помощи TF-IDF векторизации, кластеризации методом К-средних (k-means) и алгоритма тематического моделирования NMF. Выявляются сильные и слабые стороны данных подходов к анализу эмпирического материала. Отдельно отражается перспективность и эффективность дополнительного использования больших языковых моделей как инструмента автоматического именования сформированных кластеров и тематик. В заключении отмечается высокий потенциал данного подхода к комплексному рассмотрению состава сетевого пространства Северного Кавказа, что также служит основой для будущих исследований в области социолингвистки, лингвостатистики и иных направлений изучения виртуального медиадискурса.

Ключевые слова: тематическое моделирование; кластеризация; сетевое пространство; виртуальность; медиадискурс; К-средних; NMF.

EXPERIENCE IN TOPIC MODELING AND CLUSTERING IN RESEARCH OF VIRTUAL MEDIA DISCOURSE

This paper considers a contaminated approach to topic modeling and clustering of network space texts in order to carry out a comprehensive study of virtual media discourse. The study is based on posts from the North Caucasus online community selected using a continuous sample. The paper presents the experience of implementing automated systematization of empirical material using TF-IDF vectorization, K-means clustering, and the NMF topic modeling algorithm. The paper identifies the strengths and weaknesses of these approaches to analyzing empirical material. The paper indicates the prospects and effectiveness of additional use of large language models as a tool for automatic naming of formed clusters and topics. The conclusion notes the high potential of this approach to a comprehensive consideration of the composition of the North Caucasus network space, which can also be used as a basis for future research in the field of sociolinguistics, linguistic statistics and other areas of studying virtual media discourse.

Key words: topic modeling; cluster analysis; network space; virtual; media discourse; K-means; NMF.

В эпоху стремительно роста цифровых технологий виртуальная реальность, формируемая в сетевом пространстве, стала ключевым каналом коммуникации и источником знаний современного общества. Социальные сети, блоги, форумы, видеохостинги и другие платформы агрегируют колоссальный массив информации, продуцируемой различными пользователями интернета [1]. Данные материалы отражают не только субъективные позиции конкретных индивидов, но и формируют общее пространство дискурсивизации общественных настроений, культурных особенностей и социальных взаимодействий. Однако подобные свойства закономерно являются и препятствиями для исследования виртуального медиадискурса. Обширный объем данных, гетерогенность информации, полимодальность с трудом поддаются анализу при помощи ограниченных человеческих ресурсов.

Данная проблема особенно актуальна при анализе территорий с высокой динамикой коммуникации, где виртуальный медиадискурс отражает сложные социальные процессы, включая межличностные, межнациональные и религиозные отношения, традиции и политические события. Одним из таких субъектов можно назвать Северный Кавказ, являющийся как фронтирным, так и полиэтничным и поликонфессиональным регионом [2]. Потенциальный анализ виртуального медиадискурса подобного региона способен прояснить не только текущие общественные процессы, но и выявить потенциальные точки напряжения, конфликтогенные и синтонные тренды.

Преодолеть упомянутые выше препятствия и всесторонне проанализировать представленный медиадискурс возможно за счет внедрения алгоритмических процедур идентификации различного рода текстов [3]. В частности, продуктивно применение комплексного автоматизированного дискурсивно-модусного подхода к исследованию интернет-ресурсов, одной из главных частей которого является тематическое моделирование и кластеризация сетевого пространства. Подобный подход позволяет сократить временные и трудовые затраты, обеспечивая высокую точность и интерпретируемость результатов, которые далее могут использоваться в иных интересах дискурсивных исследований. Таким образом, цель данной работы состоит в раскрытии сущности тематического моделирования и кластеризации как инструмента комплексного исследования виртуального медиадискурса.

Базой исследования выступил сформированный авторский корпус текстов, отобранных методом сплошной выборки из сообщества ВКонтакте «ЧП Кавказ Дагестан Чечня Ингушетия Осетия КБР» за период с 1 января 2021 г. по 1 июня 2025 г., что охватывает весь срок существования паблика на момент написания данной работы. Выбор данного сообщества обусловлен его широкой аудиторией (86 тыс. подписчиков), регулярностью освещения актуальных тем Северного Кавказа, активностью пользователей и репрезентативностью контента для различных этнокультур. Всего было собранно

3363 публикации, которые содержали новостные заметки, объявления, обсуждения и другие типы контента. Здесь можно заметить, что база исследования представляет собой обширный массив гетерогенных данных, что обуславливает целесообразность применения тематического моделирования и кластеризации в исследовании виртуального медиадискурса. Далее подробнее рассмотрим данные подходы к систематизации исследовательской информации.

Тематическое моделирование представляет собой «способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов» [4]. Оно основано на предположении, что каждый текст представляет собой комбинацию нескольких тем, каждая из которых характеризуется определенным набором слов. При этом кластеризация фокусируется преимущественно на группировке текстов по принципу сходства. Однако здесь важно различать два подхода: жесткую кластеризацию, где каждый документ принадлежит ровно одному кластеру; и мягкую кластеризацию, когда документ может принадлежать нескольким кластерам одновременно, что сближает её с представлением документов в тематическом моделировании.

В данной работе процесс тематического моделирования и кластеризации основывался на применении алгоритма неотрицательного матричного разложения (англ. *NMF*) и метода K-средних (англ. *k-means*) к текстам, прошедши предварительную векторизацию посредством расчета *TF-IDF* меры.

Отдельно поясним, *TF-IDF* (англ. *Term Frequency-Inverse Document Frequency*) – это метод векторизации текстов, который оценивает значимость слов в документах относительно всего корпуса, учитывая частоту слова в тексте (*TF*) и обратную частоту его появления в корпусе (*IDF*). Кластеризация методом K-средних группирует тексты на основе их векторного представления, выявляя тематические кластеры. *NMF*, в свою очередь, разлагает матрицу *TF-IDF* на две неотрицательные матрицы, представляющие темы и их веса в документах, а также слова и их веса в темах, что позволяет выявить мягкие тематические структуры с высокой интерпретируемостью [5, p. 86]. Представленный подход также дополняется использованием больших языковых моделей для автоматического именования тем/кластеров, что повышает точность и скорость анализа результатов.

Стоит отметить, что реализация упомянутых методов осуществлялась на основе программ, написанных на языке *Python* с использованием: специализированных библиотек для машинного обучения, а именно *scikit-learn*; *API* ВКонтакте для сбора первичного эмпирического материала; *API DeepSeek* с целью осуществления прямых запросов к большой языковой модели для автоматического именования тем/кластеров.

В результате выбор подобного подхода к исследованию способствует эффективной автоматизации процессов анализа обширных и специфических

текстов социальных сетей. Общий алгоритм проведения тематического моделирования и кластеризации виртуального медиадискурса можно представить в виде схемы, визуализирующей данный процесс от этапа сбора данных до итогового анализа полученных результатов (Рисунок 1).

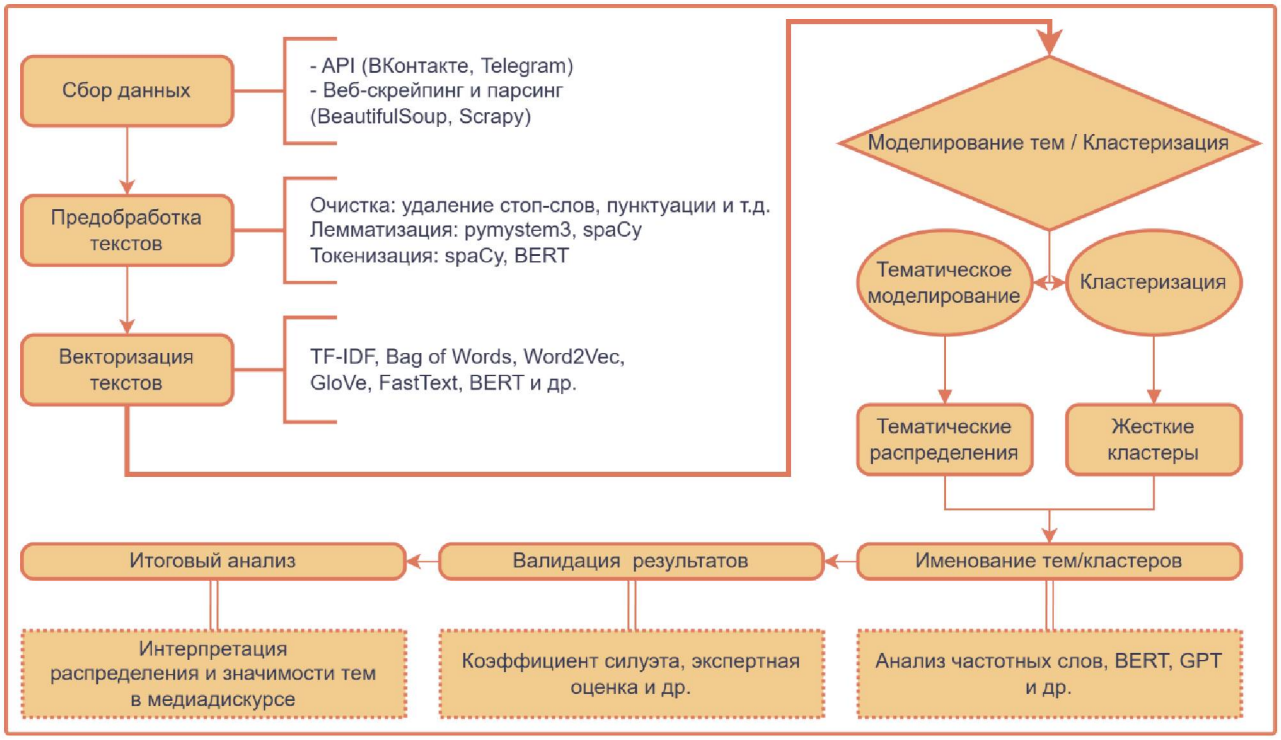

Рис 1. Алгоритм проведения тематического моделирования и кластеризации виртуального медиадискурса

В рамках исследования на основе метода К-средних с дополнительным применением коэффициента силуэта было выявлено 8 кластеров, соответствующих основным темам, представленным в отобранных публикациях социальных сетей. При помощи алгоритма *NMF* также было выделено аналогичное количество тематик анализируемого эмпирического материала. Далее рассмотрим конкретные выявленные кластеры и их наполнение:

1) ЧП и происшествия (310 записей). Ключевые слова по *TF-IDF* мере: пожар, махачкала, писать, мы, пострадать, гореть;

2) Дети и семья (348 записей). Ключевые слова по *TF-IDF* мере: ребёнок, девочка, мальчик, дом, летний, год;

3) Реклама чата (33 записи). Ключевые слова по *TF-IDF* мере: репост, чат, телеграм, chat, заходить, спать;

4) Криминальные происшествия (163 записи). Ключевые слова по *TF-IDF* мере: девушка, её, парень, который, летний, махачкала;

5) Разные инциденты (646 записей). Ключевые слова по *TF-IDF* мере: мужчина, летний, полиция, полицейский, парень, сотрудник;

6) Новости Кавказа (1257 записей). Ключевые слова по *TF-IDF* мере: дагестан, видео, район, наш, махачкала, свой;

7) Дорожные происшествия (309 записей). Ключевые слова по *TF-IDF* мере: дтп, погибнуть, водитель, авария, пострадать, автомобиль;

8) Преступления и инциденты (295 записей). Ключевые слова по *TF-IDF* мере: летний, уголовный, дело, житель, подозревать, мужчина, задержать.

Аналогичным образом представим обнаруженные тематики алгоритмом *NMF:*

1) Преступления и происшествия (491 запись). Ключевые слова: летний, житель, мужчина, задержать, дело, уголовный, подозревать, убийство, возбудить;

2) ДТП и аварии (423 записи). Ключевые слова: дтп, водитель, погибнуть, автомобиль, авария, результат, пассажир, происшествие, пострадать;

3) Приглашения в чаты (53 записи). Ключевые слова: репост, чат, телеграм, chat, заходить, спать, телеграм, беседа, максимальный;

4) Криминальные происшествия (1040 записей). Ключевые слова: девушка, парень, видео, деньга, её, день, человек, женщина, полиция;

5) Чрезвычайные ситуации (295 записей). Ключевые слова: пожар, пострадать, произойти, информация, гореть, дом, взрыв, пожарный, мчс;

6) Криминал и происшествия (276 записей). Ключевые слова: махачкала, писать, акушинский, сегодня, проспект, драка, хасавюрт, кизилюрт, вчера;

7) События в регионах (416 записей). Ключевые слова: дагестан, район, село, уточняться, писать, кизилюртовский, цунтинский, казбековский, утонуть;

8) Трагедии и дети (367 записей). Ключевые слова: ребёнок, мальчик, девочка, мать, дом, родитель, больница, мама, отец.

Как можно заметить, обнаруженные кластеры и темы в целом совпадают друг с другом, отражая схожие семантические характеристики анализируемых публикаций. Однако алгоритм *NMF* закономерно показал лучшие результаты в распределении записей по конкретным тематикам. Например, короткая рекламная запись: *НАША БЕСЕДА В ТЕЛЕГРАММЕ* [6], которая относится к кластеру *Новости Кавказа*, была помечена темой *Приглашения в чаты*. В целом различия в распределении публикаций на темы и кластеры возможно проследить на представленной ниже тепловой карте (Рисунок 2).

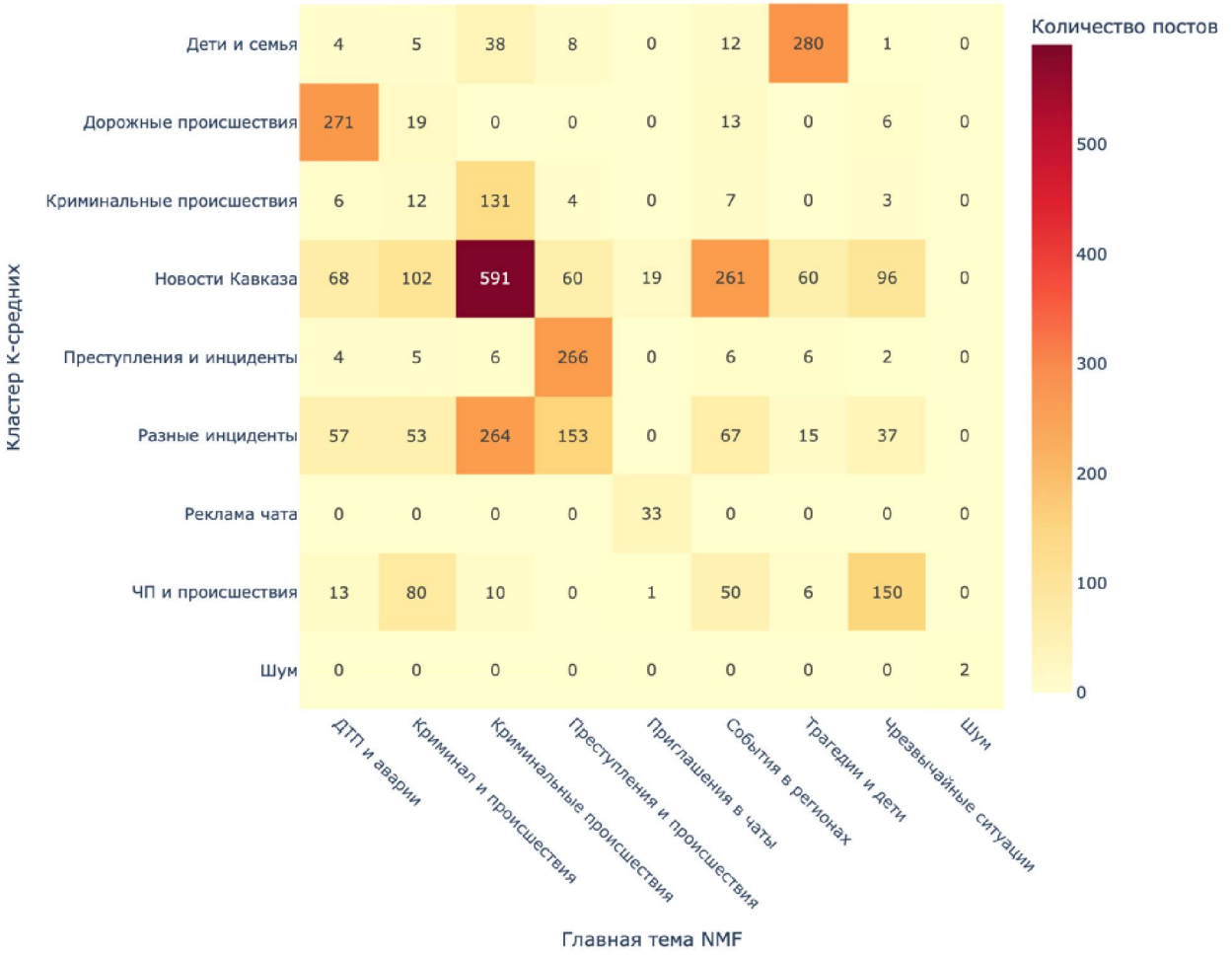

Рис 2. Распределение постов по кластерам К-средних
и главным тематикам *NMF*

Алгоритм NMF, как уже было упомянуто ранее, способен также выделять и дополнительные темы в конкретных текстах. Здесь в пример можно привести следующую публикацию:

Поисково-спасательные работы идут в Бежтинском участке, где 11 мая пропал мальчик 2021 года рождения, сообщает МЧС по Дагестану. Предполагают, что ребёнок мог упасть с моста в реку. Нашли его велосипед, висящий на опоре моста. В поисках участвуют 21 спасатель, в том числе кинолог с собакой, а также 50 сотрудников МВД и местные жители [7].

Данная запись ожидаемо была отнесена к кластеру 2 *Дети и семья* с главной темой 8 *Трагедии и дети*. Однако в этой публикации также были обнаружены элементы и других тем, а именно 1 *Преступления и происшествия*, 7 *События в регионах*, 5 *Чрезвычайные ситуации*.

Отдельно следует отметить, и метод К-средних, и алгоритм NMF имеют как свои слабые, так и сильные стороны. К-средних достаточно прост в реализации, имеет высокую скорость и предоставляет четкое разделение текстов на группы. Однако данный подход игнорирует семантические связи,

а также чувствителен к шуму. NMF более гибок, предоставляет мягкое тематическое распределения с высокой интерпретируемостью, но сложнее в настройке и также чувствителен к шуму. В итоге выбор между этими методами сводится к поставленным задачам исследования, а их сочетание способствует комплексному анализу виртуального медиадискурса.

Таким образом, проведенное исследование демонстрирует эффективность тематического моделирования и кластеризации при анализе обширных массивов данных виртуального медиадискурса. При этом использование больших языковых моделей для автоматического именования кластеров и тематик повышали интерпретируемость результатов. В целом данный подход способствовал комплексному рассмотрению состава сетевого пространства Северного Кавказа, что также может послужить основой для будущих исследований в области социолингвистки, лингвостатистики и иных направлений.

ЛИТЕРАТУРА

1. Шишкин Б. А. Вариативность нарратива гармонизации сетевой коммуникации в медиадискурсе [Электронный ресурс] // Актуальные проблемы филологии и педагогической лингвистики. 2024. № 3. С. 135–150. DOI: https://doi.org/10.29025/2079-6021-2024-3-135-150.
2. Социокультурные основы Российской идентичности молодежи Северного Кавказа: ценностные трансформации и сетевые репрезентации / А. М. Ерохин, Е. А. Авдеев, С. Н. Бредихин, С. М. Воробьев, Б. А. Шишкин. Ставрополь : ООО «Бюро новостей», 2024. 324 с.
3. Каменский М. В., Бредихин С. Н. Алгоритмические процедуры идентификации рекламных текстов в дискурсивном пространстве средств массовой информации // Вестник Волгоградского государственного университета. Серия 2: Языкознание. 2025. Т. 24, № 1. С. 64–78. DOI: https://doi.org/10.15688/jvolsu2.2025.1.6.
4. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. 2012. Т. 23. С. 215–44.
5. Aggarwal C.C., Zhai C. A Survey of Text Clustering Algorithms [Electronic resource] // Mining Text Data. Boston : Springer, 2012. P 77–128. DOI: https://doi.org/10.1007/978-1-4614-3223-4_4.
6. Ночной чат // ЧП Кавказ Дагестан Чечня Ингушетия Осетия КБР. URL: https://vk.com/wall-201560969_256536 (дата обращения: 20.06.2025).
7. Поисково-спасательные работы идут в Бежтинском участке // ЧП Кавказ Дагестан Чечня Ингушетия Осетия КБР. URL: https://vk.com/wall-201560969_338817 (дата обращения: 20.06.2025).