

Научная статья

УДК 004.421.2: 81'33: 519.178

DOI 10.25205/1818-7900-2023-21-2-29-38

## **Использование платформы ТХМ корпусного анализа для анализа текстов сообществ социальных сетей**

**Алина Игоревна Фокина<sup>1</sup>,  
Александр Андреевич Чеповский<sup>2</sup>,  
Андрей Михайлович Чеповский<sup>3</sup>**

<sup>1-3</sup>Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ)  
Москва, Россия

<sup>3</sup>Российский университет дружбы народов (РУДН)  
Москва, Россия

<sup>1</sup>aifokina@edu.hse.ru

<sup>2</sup>aachepovsky@hse.ru

<sup>3</sup>chepovskiy-am@rudn.ru.

### *Аннотация*

При формировании графов взаимодействующих объектов, построенных при импорте данных из социальных сетей и сетей мгновенного обмена сообщениями, в качестве атрибутов вершин выступают в том числе и текстовые данные. В настоящей работе авторы приводят описание методики исследования текстов, основанной на процедурах корпусного анализа. Целью данной статьи является проверка методологических средств, предоставляемых программным обеспечением ТХМ для сравнительного анализа текстов выделенных сообществ на графе взаимодействующих объектов. Метод предлагается для оценки качества выделения неявных сообществ на графе, полученном при импорте данных из сети каналов мессенджера Telegram.

### *Ключевые слова*

анализ социальных сетей, автоматический анализ текстов, платформа ТХМ

### *Для цитирования*

Фокина А. И., Чеповский А. А., Чеповский А. М. Использование платформы ТХМ корпусного анализа для анализа текстов сообществ социальных сетей // Вестник НГУ. Серия: Информационные технологии. 2023. Т. 21, № 2. С. 29–38. DOI 10.25205/1818-7900-2023-21-2-29-38

## Using TXM Platform of Corpus Analysis for Text Analysis of Social Media

Alina I. Fokina<sup>1</sup>, Aleksandr A. Chepovskiy<sup>2</sup>, Andrey M. Chepovskiy<sup>3</sup>

<sup>1-3</sup>HSE University  
Moscow, Russian Federation

<sup>3</sup>Federal Research Center "Informatics and Management"  
Moscow, Russian Federation

<sup>1</sup>aifokina@edu.hse.ru

<sup>2</sup>aachevovsky@hse.ru

<sup>3</sup>chevovskiy-am@rudn.ru.

### Abstract

When forming graphs of interacting objects built when importing data from social networks and instant messaging networks, text data also act as vertex attributes. In this paper, the authors describe a text research methodology based on corpus analysis procedures. The purpose of this article is to test the methodological tools provided by the TXM software for the comparative analysis of the revealed communities texts on the graph of interacting objects. The method is proposed to assess the quality of the implicit communities revealing on the graph obtained by importing data from the channel network of the Telegram messenger.

### Keywords

social network analysis, automated text analysis, TXM platform

### For citation

Fokina A. I., Chepovskiy A. A., Chepovskiy A. M. Using TXM Platform of Corpus Analysis for Text Analysis of Social Media. *Vestnik NSU. Series: Information Technologies*, 2023, vol. 21, no. 2, pp. 29–38. DOI 10.25205/1818-7900-2023-21-2-29-38

## Введение

Изучение структуры сетевых сообществ актуально для анализа социальных связей и сетей мгновенного обмена сообщениями в контексте средств распространения информации, что имеет большое значение в современном обществе. Исследование возникающих в таких сетях сообществ позволяет определять процессы распространения информации, выделения криминальных групп, корректировать использование каналов маркетинговых коммуникаций [9, 11, 13].

Для решения задач выделения сетевых сообществ пользователей в работе [5] был разработан метод Галактик, который в процессе своего применения обеспечивает выделение пересекающихся неявных сообществ. Под выделением неявных сообществ на графе подразумевается разбиение графа на подграфы такое, что плотность связей внутри этих подграфов намного выше плотности связей между ними. При этом существенно выделение на графе пересекающихся сообществ, подразумевающих наличие общих вершин, принадлежащих сразу двум или более сообществам. Именно такие сообщества и выделяются методом, разработанным в [5].

При выделении сообществ одной из наиболее сложных проблем является оценка корректности и эффективности работы соответствующих методов [9, 14]. Поэтому проблема оценки качества выделения сообществ на графах ставится как актуальный вопрос информационных технологий [11].

В работах [1, 8] было показано, что для оценки качества исследования сетей Telegram-каналов удобным инструментом является сочетание алгоритмического подхода по выделению сообществ в совокупности с анализом лингвистических характеристик. Это позволяет выделять группы каналов, ведущих активное информационное воздействие, подтверждать корректность

выделения неявных сообществ. В [1, 8] была показана возможность применять сравнение психолингвистических факторов и методику сравнения частотных словарей текстов для подтверждения корректности выделения неявных сообществ в графах, полученных при импорте данных из сетей мгновенного обмена сообщениями.

В данной работе для сравнительного анализа текстов сообществ, выделяемых на графе Telegram-каналов, мы предлагаем использовать методы корпусного анализа на базе платформы ТХМ [12], которые развивались в [2, 3, 4, 7, 13] для выявления дифференцирующих признаков текстов различной природы.

### Формирование корпусов текстов

Исследования проводились на импортированных данных из Telegram-каналов. Импорт данных и формирование графа взаимодействующих объектов осуществлялись согласно (U, M, R)-модели информационного взаимодействия, описывающей распространение информации за импортируемый промежуток времени [6], в которой учитывается количество внешних ссылок в постах (U), количество постов (M) и количество репостов (R) с весовыми коэффициентами.

Первый граф  $G_1$  был получен скачиванием данных посредством обхода графа, начиная с канала @kudago, соответствующего развлекательному сайту <http://kudago.com/msk/>. При импорте данных для этого графа выбирался временной интервал с 03.10.2022 по 17.10.2022. Глубина обхода графа бралась равной 5. Исходный скаченный граф  $G_1$  состоит из 619 вершин и 2973 ребер.

Второй граф  $G_2$  был получен скачиванием данных посредством обхода графа, начиная с канала @ob\_obraz, который посвящен новостям общего и высшего образования РФ. За временной интервал был взят период с 01.12.2022 по 31.01.2023. Глубина обхода графа бралась равной 2. Исходный скаченный граф  $G_2$  состоит из 168 вершин и 697 ребер.

Третий граф  $G_3$  был получен скачиванием данных посредством обхода графа, начиная с нескольких каналов, освещающих ход специальной военной операции Российской Федерации. За временной интервал был взят период с 01.07.2022 по 01.09.2022. Глубина обхода графа бралась равной всего лишь равной 1, ибо в противном случае подключалось много крупных политических каналов, что несколько искажало исследуемую картину. Исходный скаченный граф  $G_3$  состоит из 600 вершин и 18 009 ребер.

К полученным графам  $G_1$ ,  $G_2$  и  $G_3$  был применен метод Галактик [5] и получены разбиения на неявные пересекающиеся сообщества. При обработке графа  $G_1$  в процессе работы алгоритма некоторые вершины и инцидентные им ребра убираются из графа как не участвующие активно во взаимодействии, поэтому после работы метода у графа  $G_1$  осталось 458 вершин. Эти вершины распределились по 8 выделенным пересекающимся сообществам. При обработке графа  $G_2$  аналогичным методом осталось 89 изначальных вершин, из которых были сформированы так же 8 сообществ. По результатам обработки графа  $G_3$  было получено 288 вершин, на которых выделено 18 сообществ. Таким образом, для каждого из трех графов сформированы сообщества  $Community_i$ , где  $i=0, \dots, 7$  для графов  $G_1$  и  $G_2$  и  $i=0, \dots, 17$  для графа  $G_3$ .

Для каждого из сообществ  $Community_i$  были скачаны текстовые сообщения всех каналов – членов этих сообществ за исследуемый период. Все тексты каждого сообщества объединялись в единый для сообщества массив текстов на естественном языке. В таблице приведены размеры этих массивов текстов. Из текстов удалялись специальные имена (имена аккаунтов, почтовые адреса) и далее рассматривались массивы текстов на естественном языке. Полученные массивы текстов сообществ рассматривались как подкорпуса корпуса текстов каждого из трех графов. Поэтому в таблице приведены размеры полученных подкорпусов, измеренные в количестве русскоязычных словоупотреблений как основной единицы корпусного анализа.

## Размеры корпусов текстов сообществ

## Community Text Corpus Sizes

Сообщество	Корпуса текстов графа $G_1$		Корпуса текстов графа $G_2$		Корпуса текстов графа $G_3$	
	Размер (Кб)	Число словоупотреблений	Размер (Кб)	Число словоупотреблений	Размер (Кб)	Число словоупотреблений
Community <sub>0</sub>	335	29293	988	35120	3498	309125
Community <sub>1</sub>	39208	3450208	985	30942	1806	179034
Community <sub>2</sub>	13571	1201393	566	20239	3633	303847
Community <sub>3</sub>	3485	302679	782	26759	1169	100055
Community <sub>4</sub>	1633	138912	1021	39702	2975	260058
Community <sub>5</sub>	37897	3272627	8914	322660	750	67982
Community <sub>6</sub>	2925	244706	11881	437942	5730	446270
Community <sub>7</sub>	10794	952898	1981	69321	703	59385
Community <sub>8</sub>	—	—	—	—	4134	422629
Community <sub>9</sub>	—	—	—	—	11489	463161
Community <sub>10</sub>	—	—	—	—	3659	308019
Community <sub>11</sub>	—	—	—	—	9298	441104
Community <sub>12</sub>	—	—	—	—	11350	452082
Community <sub>13</sub>	—	—	—	—	5398	463593
Community <sub>14</sub>	—	—	—	—	71538	455354
Community <sub>15</sub>	—	—	—	—	17063	443049
Community <sub>16</sub>	—	—	—	—	4243	358384
Community <sub>17</sub>	—	—	—	—	6660	448108

## Методы лингвистического анализа

Сформированные, как описано выше, массивы текстов сообществ исследовались методами корпусного анализа на базе платформы ТХМ [12]. Платформа ТХМ является эффективным средством, позволяющим проводить комплексный анализ корпусов текстов процедурами анализа соответствий, кластеризации, построения лексических таблиц, поиска сложных лексических конструкций.

В текстах выделялись словоупотребления, для которых проводился автоматический морфологический анализ словоформ. В настоящей работе используется программный пакет TreeTagger [15], предоставляющий возможность совместного морфологического анализа слов предложения на основе статистической модели. По результатам работы TreeTagger определяются канонические (начальные) формы слова. Преимуществом пакета является однозначность морфологического анализа словоупотреблений.

Наборы текстов с вычисленными характеристиками импортируются в пакет ТХМ для последующего анализа. В рамках платформы выделяются подкорпуса, представляющие собой объединенный текст публикаций членов каждого выделенного в графе неявного сообщества.

Для исследований корпусов текстов сообществ мы используем анализ соответствий [10], который включен как один из инструментов в платформу ТХМ. Анализ соответствий является методом исследования корпуса текстов, который разделен на подкорпуса. При делении текстов

на подкорпуса есть возможность интерпретировать близость между значениями характеристик подкорпусов как оценку, указывающую на сходство или различие между этими подкорпусами.

Данный метод состоит в анализе частот совместного появления значений переменных в таблице частот лингвистических характеристик. В основе этого лежит изучение симметричной матрицы, сопоставляющей признаки друг другу. Такой подход позволяет увидеть взаимосвязи между признаками, а также дает возможность интерпретировать выделенные факторы как совокупность некоторого набора выделенных из текстов характеристик.

В процессе решения задачи разделения подкорпусов методом анализа соответствий исследуется частота совместного появления значений определенных переменных, в качестве которых рассматриваются выделенные при формировании корпусов лингвистические характеристики.

Результаты работы метода наглядно представляются в графической интерпретации анализа соответствий (рис. 1–3), которая иллюстрирует пространственное расположение подкорпусов в зависимости от частот совместного появления значений исследуемых переменных (лингвистических характеристик).

### Результаты анализа корпусов текстов сообществ

Описанные выше корпуса для графов (см. табл.) были проанализированы с использованием описанной выше функциональности ТХМ «анализ соответствий». Детально были исследованы следующие лексические объекты: словоформы; начальные формы слов с морфологическими характеристиками, полученные с помощью TreeTagger; начальные формы слов, отнесенные к различным частям речи.

На рисунках представлены пространственные расположения подкорпусов на основе слов для подкорпусов текстов сообществ графа  $G_1$  (рис. 1), на основе нормальных форм слов для подкорпусов текстов сообществ графа  $G_2$  (рис. 2), на основе существительных для подкорпусов текстов сообществ графа  $G_3$  (рис. 3).

В названии осей указывается процент вариации по корпусу характеристик, входящих в выделенный процедурой анализа соответствий фактор. По осям откладывается характеристика степени отклонения набора признаков от указанной в названии оси процента вариации для конкретного подкорпуса. Таким образом, чем дальше от пересечения осей координат расположен подкорпус по осям координат, тем сильнее в нем отличаются наборы признаков (частота встречаемости) от признаков по корпусу.

На рис. 1 представлены пространственные расположения подкорпусов по результатам анализа соответствий для частотных таблиц слов для корпуса текстов сообществ графа  $G_1$ . Точки подкорпусов сообществ *Community*<sub>0</sub> и *Community*<sub>2</sub> лежат во втором квадранте, сообществ *Community*<sub>4</sub> и *Community*<sub>5</sub> лежат в третьем квадранте, сообществ *Community*<sub>3</sub> и *Community*<sub>7</sub> лежат в четвертом квадранте, а *Community*<sub>6</sub> лежит в первом квадранте. На диаграмме подкорпуса разнесены в презентационном пространстве результатов анализа соответствий.

На рис. 2 представлены пространственные расположения подкорпусов по результатам анализа соответствий для частотных таблиц нормальных форм слов для корпуса текстов сообществ графа  $G_2$ . Точка подкорпуса сообщества *Community*<sub>6</sub> лежит на оси координаты фактора 1, *Community*<sub>5</sub> лежит в первом квадранте, точки подкорпусов сообществ *Community*<sub>1</sub>, *Community*<sub>2</sub>, *Community*<sub>3</sub> и *Community*<sub>7</sub> лежат во втором квадранте, сообществ *Community*<sub>0</sub> и *Community*<sub>4</sub> лежат в четвертом квадранте. Таким образом, как и в первом примере, подкорпуса разнесены в презентационном пространстве результатов анализа соответствий.

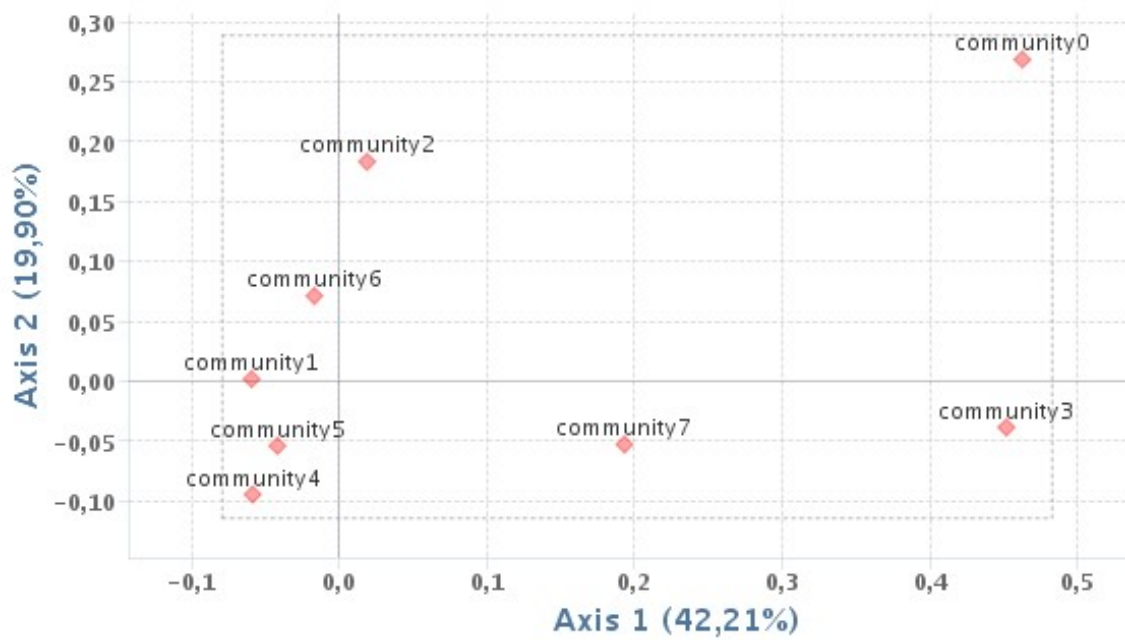


Рис. 1. Анализ соответствий для корпуса сообществ графа  $G_1$  для слов  
 Fig. 1. Words Correspondence Analysis for the corpus of communities of graph  $G_1$

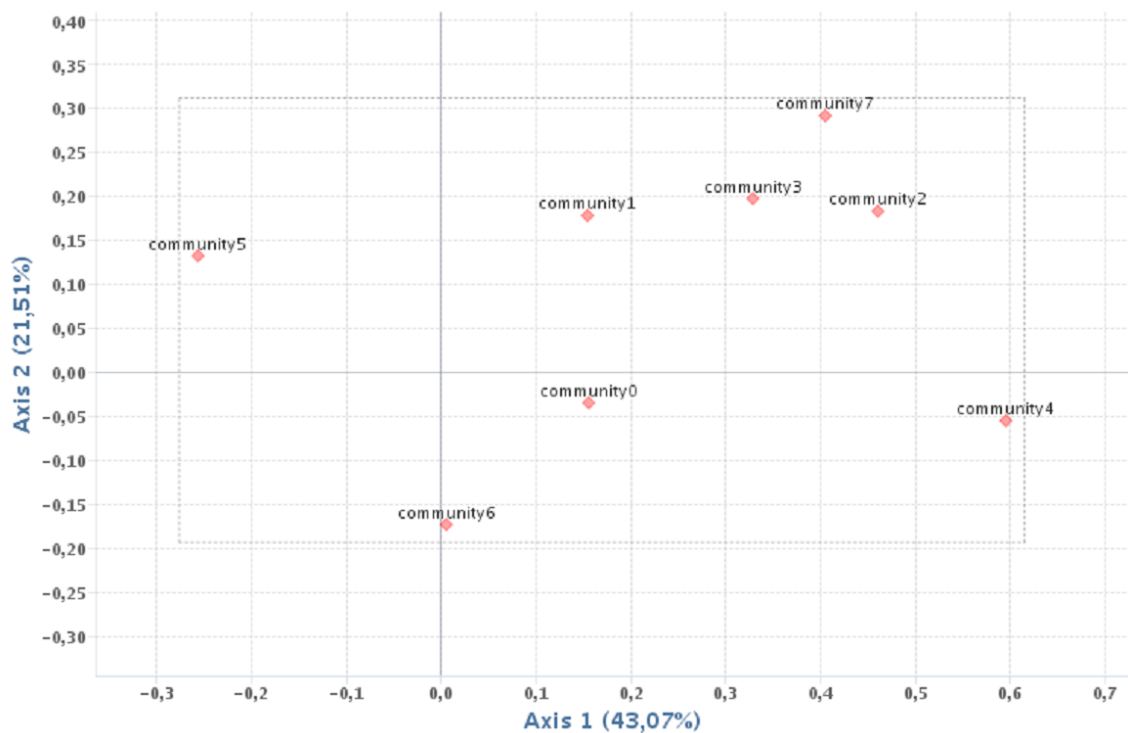


Рис. 2. Анализ соответствий для корпуса сообществ графа  $G_2$  для нормальных форм слов  
 Fig. 2. Lemmas Correspondence Analysis for the corpus of communities of graph  $G_2$

На рис. 3 представлены пространственные расположения подкорпусов по результатам анализа соответствий для частотных таблиц существительных для корпуса текстов сообществ

графа  $G_3$ . Видно, что по аналогии с первым и вторым примерами подкорпуса разнесены в презентационном пространстве результатов анализа соответствий.

Представленные на рис. 1, 2 и 3 результаты анализа соответствий для всех корпусов трех исследуемых графов явно показывают разделение подкорпусов текстов выделенных неявных пересекающихся сообществ по лингвистическим характеристикам. Это, по нашему мнению, можно рассматривать как свойство сообществ графов взаимодействующих объектов, которое может подтверждать корректность выделения сообществ.

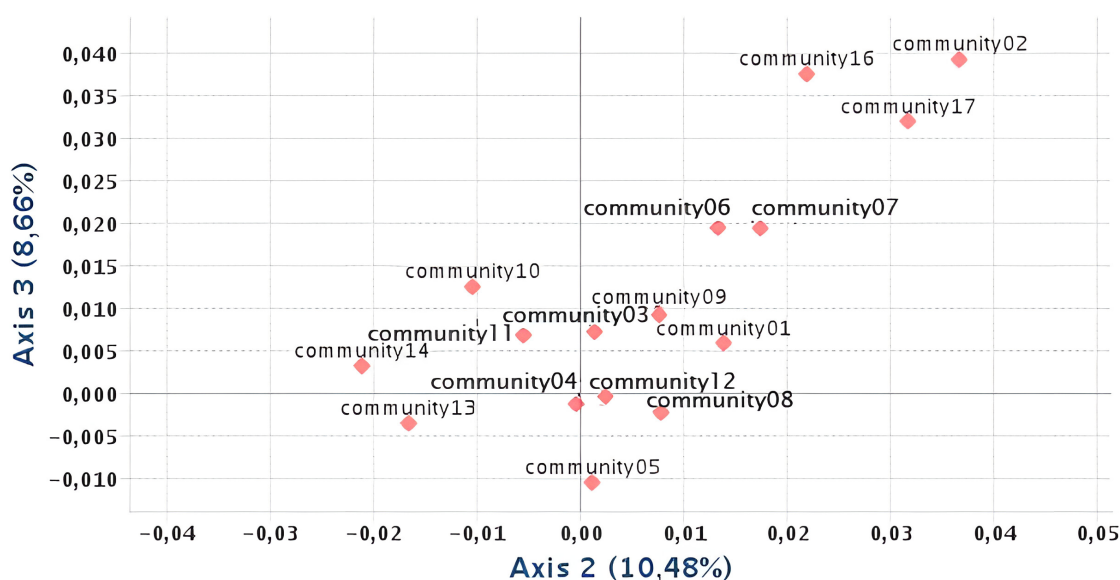


Рис. 3. Анализ соответствий для корпуса сообществ графа  $G_3$  для существительных

Fig. 3. Noun Correspondence Analysis for the corpus of communities of graph  $G_3$

### Заключение

Предложена и опробована методика для исследования методами корпусного анализа текстов, сформированных из атрибутивных данных групп вершин графов взаимодействующих объектов. Для различных групп общения проведены исследования текстов выявленных неявных сообществ таких графов, а именно графов информационного взаимодействия сети Telegram-каналов.

В рамках работы к наборам текстов выделенных сообществ были успешно применены средства анализа корпусной платформы ТХМ. Проведен анализ соответствий для различных лингвистических характеристик корпусов текстов сообществ (слов, начальных форм, частей речи).

На основе исследования реальных данных показана возможность оценки корректности выделения пересекающихся сообществ на графе информационного взаимодействия. Данный подход основан на анализе методами компьютерной лингвистики объединенных корпусов текстов, составленных по публикациям каналов, входящих в выделенные сообщества.

Данная работа вместе с работами [1, 8] формулирует общую комплексную методику оценки корректности выделения пересекающихся сообществ на графах взаимодействующих объектов.

## Список литературы

1. **Аванесян Н. Л., Соловьев Ф. Н., Чеповский А. А.** Характеристики текстов сообществ социальных сетей // Вестник НГУ. Серия: Информационные технологии. 2021. Т. 19, № 1. С. 5–14. DOI 10.25205/1818-7900-2021-19-1-5-14.
2. **Лаврентьев А. М., Рябова Д. М., Тихомирова Е. А., Фокина А. И., Чеповский А. М., Шерстинова Т. Ю.** Сравнительный анализ специальных корпусов текстов для задач безопасности // Вопросы кибербезопасности. 2020. № 3(37). С. 58–65. DOI: 10.21681/2311-3456-2020-03-58-65.
3. **Лаврентьев А. М., Смирнов И. В., Соловьев Ф. Н., Суворова М. И., Фокина А. И., Чеповский А. М.** Анализ корпусов текстов террористической и антиправовой направленности // Вопросы кибербезопасности. 2019. № 4(32). С. 54–60. DOI: 10.21681/2311-3456-2019-4-54-60.
4. **Лаврентьев А. М., Соловьев Ф. Н., Суворова М. И., Фокина А. И., Чеповский А. М.** Новый комплекс инструментов автоматической обработки текста для платформы TXM и его апробация на корпусе для анализа экстремистских текстов // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2018 Т. 16 № 3 С. 19–31. DOI 10.25205/1818-7935-2018-16-3-19-31.
5. **Попов В. А., Чеповский А. А.** Выделение неявных пересекающихся сообществ на графе взаимодействия Telegram-каналов с помощью «метода Галактик» // Труды ИСА РАН. Т. 72. 4/2022. С. 39–50. DOI: 10.14357/20790279220405.
6. **Попов В. А., Чеповский А. А.** Модели импорта данных из мессенджера Telegram // Вестник НГУ. Серия: Информационные технологии. 2022. Т. 20. № 2. С. 60–71. DOI: 10.25205/1818-7900-2022-20-2-60-71.
7. **Соловьев Ф. Н.** Автоматическая обработка текстов на основе платформы TXM с учетом анализа структурных единиц текста // Вестник НГУ. Серия: Информационные технологии. 2020. Т. 18. №1. С. 74–82. DOI 10.25205/1818-7900-2020-18-1-74-82.
8. **Чеповский А. А.** Об особенностях построения и анализа графов взаимодействующих объектов в сети Telegram.-каналов. Вопросы кибербезопасности. 2023; 1(53):75-81. DOI:10.21681/2311-3456-2023-1-75-81.
9. **Чеповский А. А.** О неявных сообществах на графе взаимодействующих объектов. Успехи кибернетики. 2023;4(1):56–64. DOI: 10.51790/2712-9942-2023-4-1-08.
10. **Benzécri J.-P.** L'analyse des données: l'analyse des correspondances. 2nd ed. Paris: Dunod, 1979. Vol. 2.
11. **Fortunato, S., Newman, M. E. J.** 20 years of network community detection. Nat. Phys. 2022; 18:848–850.
12. **Heiden, S.** The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In: Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation. Sendai, Japan. P. 389–398.
13. **Lavrentiev A., Sherstinova T., Chepovskiy A., Pincemin B.** Using TXM Platform for Research on Language Changes over Time: The Dynamics of Vocabulary and Punctuation in Russian Literary Texts // Vestnik Tomskogo Gosudarstvennogo Universiteta, Filologiya. 2021. Vol. 70. P. 69-89. DOI: 10.17223/19986645/70/5.
14. **Newman M. E. J.** Networks: An Introduction. Oxford University Press, 2010. 784 p.
15. **Schmid H.** Probabilistic Part-of-Speech Tagging Using Decision Trees // Proc. of International Conference on New Methods in Language Processing. Manchester, UK. 1994. URL = <http://www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf>. (дата обращения: 30.05.2023).

## References

1. **Avanesyan N. L., Solovev F. N., Chepovskiy A. A.** Characteristics of Texts of Social Networks Communities // Vestnik NSU. Series: Information Technologies. 2021. Vol. 19(1). Pp. 5–14. (in Russ.) DOI 10.25205/1818-7900-2021-19-1-5-14
2. **Lavrentiev A. M., Raybova D. M., Tikhomirova E. A., Fokina A. I., Chepovskiy A. M., Sherstinova T. Yu.** Comparative analysis of special text corpora for security-related tasks // Voprosi kiberbezopasnosti. 2020. № 3(37). Pp. 58–65. (in Russ.) DOI 10.21681/2311-3456-2020-03-58-65
3. **Lavrentiev A. M., Smirnov I. V., Solovev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M.** Analiz korpusov tekstov terroristicheskoi i antipravovoy napravlenosti // Voprosi kiberbezopasnosti. 2019. № 4(32). Pp. 54–60. (in Russ.) DOI 10.21681/2311-3456-2019-4-54-60
4. **Lavrentiev A. M., Solovev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M.** A New Toolkit for Natural Text Processing with the TXM Platform and its Application to a Corpus for Analysis of Texts Propagating Extremist Views // Vestnik NSU. Series: Linguistics and Intercultural Communication. 2018. Vol. 16, № 3. Pp. 19–31. (in Russ.). DOI 10.25205/1818-7935-2018-16-3-19-31
5. **Popov V. A., Chepovskiy A. A.** Vydelenie neyavnyh peresekayushchihsya soobshchestv na grafe vzaimodeystviya Telegram-kanalov s pomoshch'yu «metoda Galaktik» // Trudy ISA RAN. 2022. Vol. 72. № 4. Pp. 39–50. (in Russ.). DOI 10.14357/20790279220405
6. **Popov V. A., Chepovskiy A. A.** Telegram Messenger Data Import Models // Vestnik NSU. Series: Information Technologies. 2022. Vol. 20, № 2. Pp. 60–71. (in Russ.). DOI 10.25205/1818-7900-2022-20-2-60-71
7. **Solovev F. N.** Embedding Additional Natural Language Processing Tools into the TXM Platform // Vestnik NSU. Series: Information Technologies. 2020. Vol. 18, no. 1. Pp. 74–82. (in Russ.). DOI 10.25205/1818-7900-2020-18-1-74-82
8. **Chepovskiy A. A.** On the construction and analysis of graphs of interacting objects in the Telegram-channels network // Voprosy kiberbezopasnosti. 2023. Vol. 1(53). Pp. 75–81. (in Russ.). DOI 10.21681/2311-3456-2023-1-75-81
9. **Chepovskiy A. A.** Implicit Communities Defined on the Graph for Interacting Objects // Russian Journal of Cybernetics. 2023. Vol. 4(1). Pp. 56–64. (in Russ.). DOI 10.51790/2712-9942-2023-4-1-08
10. **Benzécri J.-P.** L'analyse des données: l'analyse des correspondances. 2nd ed. Paris: Dunod, 1979. Vol. 2.
11. **Fortunato S., Newman M. E. J.** 20 years of network community detection // Nat. Phys. 2022. Vol. 18. Pp. 848–850.
12. **Heiden S.** The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme // Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation. Sendai, Japan. Pp. 389–398.
13. **Lavrentiev A., Sherstinova T., Chepovskiy A., Pincemin B.** Using TXM Platform for Research on Language Changes over Time: The Dynamics of Vocabulary and Punctuation in Russian Literary Texts // Vestnik Tomskogo Gosudarstvennogo Universiteta, Filologiya. 2021. Vol. 70. Pp. 69–89. DOI 10.17223/19986645/70/5
14. **Newman M. E. J.** Networks: An Introduction. Oxford University Press, 2010. 784 p.
15. **Schmid H.** Probabilistic Part-of-Speech Tagging Using Decision Trees [Online] // Proc. of International Conference on New Methods in Language Processing. Manchester, UK. 1994. URL: <http://www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf> (accessed on: 30.05.2023).

### Информация об авторах

**Фокина Алина Игоревна**, аспирант Национального исследовательского университета «Высшая школа экономики»

**Чеповский Александр Андреевич**, кандидат физико-математических наук, доцент Национального исследовательского университета «Высшая школа экономики»

**Чеповский Андрей Михайлович**, доктор технических наук, профессор Российского университета дружбы народов (РУДН)

### Information about the Authors

**Alina I. Fokina**, postgraduate student, National Research University Higher School of Economics Moscow, Russia

**Alexander A. Chepovskiy**, Ph.D. (mathematics), Associate Professor, National Research University Higher School of Economics Moscow, Russia

**Andrey M. Chepovskiy**, Dr Sc. (Eng), Peoples Friendship University of Russia (RUDN University)

*Статья поступила в редакцию 02.06.2023;  
одобрена после рецензирования 27.06.2023; принята к публикации 27.06.2023*

*The article was submitted 02.06.2023;  
approved after reviewing 27.06.2023; accepted for publication 27.06.2023*