УДК 004.421.2: 81’33: 519.178
DOI 10.25205/1818-7900-2021-19-1-5-14

# Характеристики текстов сообществ социальных сетей

**Н. Л. Аванесян** [1], **Ф. Н. Соловьев** [2], **А. А. Чеповский** [1]

[1] *Национальный исследовательский университет «Высшая школа экономики»*
*Москва, Россия*

[2] *Федеральный исследовательский центр «Информатика и управление» РАН*
*Москва, Россия*

*Аннотация*

Приводится описание методики статистического анализа текстов социальных сетей, основанной на сравнении методами корреляционного анализа автоматически сформированных частотных словарей. Рассматриваются психолингвистические характеристики и коэффициенты попарной ранговой корреляции для сравнения частотных характеристик текстов на естественном языке.

*Ключевые слова*

анализ социальных сетей, автоматический анализ текстов, ранговая корреляция, психолингвистические характеристики

*Благодарности*

Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00806

*Для цитирования*

*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

# Characteristics of Texts of Social Networks Communities

**N. L. Avanesyan** [1], **F. N. Solovev** [2], **A. A. Chepovskiy** [1]

[1] *National Research University Higher School of Economics*
*Moscow, Russian Federation*

[2] *Federal Research Center “Informatics and Management”*
*Moscow, Russian Federation*

*Abstract*

In this paper the authors describe the methodology for the statistical analysis of texts in social networks based on comparison of automatically generated frequency dictionaries by methods of correlation analysis. Psycholinguistic characteristics and coefficients of pairwise rank correlation are considered for comparing the frequency characteristics of texts in natural language

*Keywords*

social network analysis, automated text analysis, rank correlation, psycholinguistics characteristics

*Acknowledgements*

The work was carried out with the financial support of the RFBR in the framework of scientific project No. 19-07-00806

*For citation*

Avanesyan N. L., Solovev F. N., Chepovskiy A. A. Characteristics of Texts of Social Networks Communities. *Vestnik NSU. Series: Information Technologies*, 2021, vol. 19, no. 1, p. 5–14. (in Russ.) DOI 10.25205/1818-7900-2021-19-1-5-14

## Введение

Распространение информации в социальных сетях имеет огромное значение в современном обществе. Изучение состава и устройства сетевых сообществ актуально для анализа социальных связей и средств распространения информации. Исследование характеристик различных групп в социальных сетях позволяет определять инструменты консолидации политических и криминальных сообществ, находить каналы маркетинговых коммуникаций.

В рамках одной из актуальных задач выделения сетевых сообществ пользователей в работе [1] предложен метод ядра, который позволяет проводить анализ взвешенных графов социальных сетей и решать задачи выявления неявных сообществ, лидеров мнений и путей распространения информации. Данная работа использует набор данных, полученных при импорте из сети «Твиттер» в рамках этого метода.

В работе [2] разбиение взвешенного графа социальной сети «Твиттер» на неявные сообщества пользователей применяется для оценки субъектности выделенных сетевых сообществ. Выявлена взаимосвязь особенностей графа и показателей субъектности сетевого сообщества. Показано, что характеристики топологии графа взаимодействия и выделенных неявных сообществ значимо коррелируют с частотой определенных дискурсивных маркеров субъектности. Но указанные работы не изучают методами компьютерной лингвистики тексты пользователей.

В [3–5] предложена и опробована методика частотного анализа характеристик естественного языка текстов сети «Интернет». Разработан метод вычисления коэффициента попарной ранговой корреляции для сравнения частотных словарей различных лексических характеристик. На основе сравнительного анализа различных по тематике коллекций текстов показаны возможности использования частотных характеристик для исследования свойств текстов с целью обнаружения противоправных ресурсов. Показаны возможности использования как морфологических характеристик слов и словосочетаний, так и буквосочетаний в качестве дифференцирующих признаков текстов.

В данной работе статистические методы компьютерной лингвистики, основанные на сравнении частотных словарей характеристик текстов [3] и вычислении психолингвистических факторов, применяются к наборам текстов неявных сообществ социальных сетей.

## Исследуемые наборы данных и методы лингвистической обработки

В данной работе исследования проводились на наборе текстов на естественном языке, полученных для графа, построенного на основе импортированных из социальной сети «Твиттер» данных [1]. Вначале были скачаны данные о взаимодействиях пользователей с 8 актуальными разноплановыми постами, включая комментарии, лайки, ретвиты. Эти посты были посвящены действиям властей города Москвы в рамках борьбы с новой коронавирусной инфекцией Covid-19. После импорта данных был сгенерирован взвешенный граф, описывающий взаимодействие пользователей с исходными постами на коротком промежутке времени. Вес ребер между ними был сформирован на основании каждого из произведенных взаимодействий пользователей между собой (подписка, лайк, комментарий, ретвит).

Для полученного графа были выделены неявные сообщества. Из 43 полученных сообществ 8 содержат более 15 вершин. Далее эти сообщества будем обозначать как $S_i$, где $i = 0,\ldots,7$. Именно эти сообщества и были выделены с точки зрения источников распространения информации и наличия лидеров мнений. В частности, было выявлено, что у четырех сообществ: $S_0$, $S_1$, $S_2$, и $S_4$ имеется высокий коэффициент плотности, что является признаком наличия активного взаимодействия внутри данных сообществ.

Для каждого из сообществ $S_i$ были скачаны текстовые сообщения всех пользователей – членов этих сообществ за исследуемый период, связанные с исходными твитами. Получен-

ные данные объединялись в единый для каждого сообщества массив текстов на естественном языке (в данном случае – русском). В первой строке табл. 1 приведены размеры этих массивов текстов. При этом из текстов удалялись специальные имена (имена аккаунтов, почтовые адреса) и рассматривались массивы текстов на естественном языке.

*Таблица 1*

Размеры текстов и частотных словарей текстов сообществ

*Table 1*

Sizes of texts and frequency dictionaries of community texts

| Словари текстов | $S_0$ | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ |
|---|---|---|---|---|---|---|---|---|
| Объем текстов (Кб) | 223 | 292 | 154 | 139 | 241 | 129 | 131 | 100 |
| Словарь существительных | 1359 | 1771 | 1125 | 977 | 1569 | 1073 | 1058 | 847 |
| Словарь глаголов | 826 | 1023 | 584 | 673 | 933 | 641 | 582 | 532 |
| Словарь прилагательных | 464 | 579 | 355 | 337 | 518 | 321 | 319 | 275 |
| Словарь псевдооснов | 3886 | 4984 | 2966 | 2882 | 4461 | 2977 | 2859 | 2388 |
| Словарь именных групп | 3702 | 5203 | 2483 | 2576 | 4507 | 2398 | 2181 | 1898 |
| Словарь глагольных групп | 1017 | 1541 | 655 | 733 | 1362 | 711 | 647 | 534 |

Характеристики текстов определялись процедурами автоматизированной обработки текстов на естественных языках, описанными в [6; 7].

В текстах выделялись словоупотребления, для которых проводился автоматический морфологический анализ словоформ на основе словарной компьютерной морфологии, описанной в [7]. Словоупотребление относится к одному из морфологических классов, каждый из которых имеет набор грамматических характеристик. Определяются канонические (начальные) формы слова, для которых составляются частотные словари.

Процедуры синтаксического анализа применялись для выделения именных и глагольных групп. Выделенные именные и глагольные группы несут информацию о различных аспектах тематического содержания текста и его психолингвистической направленности.

В качестве именной группы рассматривалась группа слов, у которой главное слово существительное, а другие слова связаны с ним подчинительными синтаксическими связями. Методика выделения именных групп основана на рассмотрении всего множества возможных морфологических разборов каждого слова и подразумевает снятие омонимической неопределенности, являющейся следствием множественности результатов морфологического анализа словоупотреблений.

Глагольная группа определялась как словосочетание, главным словом которого является глагол. Для глагола устанавливаются связи с найденными именными группами на основе синтаксического анализа предложения. В основе такого анализа лежит определение глагольного управления как разновидности синтаксической подчинительной связи, на которое накладываются ограничения на употребление зависимого словосочетания в виде набора вариантов допустимых комбинаций грамматических характеристик зависимого словосочетания. Анализ глагольного управления основан на электронном словаре глагольного управления, в который вошли первые две тысячи наиболее частотных глаголов русского языка.

В качестве одной из лингвистических характеристик текста используется псевдооснова, под которой понимается часть слова без некоторых аффиксов. Алгоритм автоматического выделения псевдооснов базируется на методе структурных схем, описанном в [7]. Псевдоос-

нова слова выделяется отбрасыванием всех аффиксов, соответствующих определенной структурной схеме, описывающей допустимую в данном языке максимальную комбинацию префиксов и суффиксов. Лингвистическая характеристика псевдоосновы позволяет анализировать тексты без использования точных словоформ.

По результатам лингвистического анализа для каждого массива текстов составлялись частотные словари различных характеристик, размеры которых в единицах записей приведены в табл. 1.

**Ранговый анализ частотных словарей**

Сопоставление наборов текстов неявных сообществ социальной сети осуществлялось попарным сравнением частотных словарей различных лингвистических характеристик, составленных для каждого из исследуемых наборов текстов. Для частотных словарей устанавливаются ранги записей словаря по результатам сортировки по частоте встречаемости занесенной в словарь характеристики. Записи словарей рассматриваются как случайные величины. Для каждой пары словарей вычисляются коэффициенты попарной ранговой корреляции, которые являются оценками наличия монотонной связи между случайными величинами. При этом если конкретное значение характеристики встречается только в одном словаре, то в другой словарь вносим эту характеристику, полагая частоту этой характеристики во втором словаре равной 0 и наоборот.

Полагаем, что каждая пара словарей имеет одинаковый размер $n$ записей, который в реальных расчетах ограничивается значением $n \le 10000$. Это обеспечивается отбрасыванием характеристик с низкой частотой использования (как правило, единичным использованием).

Рассмотрим словари как выборки для двух случайных величин $X$, $Y$, которые обозначим $X^n = \{X_i\}_{i=1}^n$, $Y^n = \{Y_i\}_{i=1}^n$. Определим меру зависимости случайных величин $X$, $Y$ через средние значения выборок $\overline{X^n}$ и $\overline{Y^n}$. Ковариация $cov(X^n, Y^n)$ определяется как математическое ожидание центрированных случайных величин:

$$cov(X^n, Y^n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X^n})(Y_i - \overline{Y^n}),$$

а дисперсия этих величин может быть записана в форме:

$$\sigma X^n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X^n})^2}\ .$$

Для отсортированных по частотам словарей будем рассматривать ранги элементов выборок $rgX^n$ и $rgY^n$. Тогда коэффициент попарной ранговой корреляции для рассматриваемых выборок определяется следующим образом [8]:

$$r = r(rgX^n, rgY^n) = \frac{cov(rgX^n, rgY^n)}{\sigma rgX^n \sigma rgY^n}.$$

Для ранжированных словарей размерностью $n$ можем положить среднее рангов

$$\overline{rgX^n} = \frac{n+1}{2}.$$

Получаем конкретное выражение для коэффициента попарной ранговой корреляции через значения рангов записей словарей $rgX_i$ и $rgY_i$:

$$r = \frac{\sum_{i=1}^{n}(rgX_i - \frac{n+1}{2})(rgY_i - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}\left(rgX_i - \frac{n+1}{2}\right)^2}\sqrt{\sum_{i=1}^{n}\left(rgY_i - \frac{n+1}{2}\right)^2}} \quad (1)$$

Формула (1) не накладывает никаких ограничений на порядок элементов, имеющих одинаковые значения. В случае набора $k$ элементов $X_{i_1}, \dots, X_{i_k}$, имеющих одинаковые значения,

они могут быть упорядочены согласно произвольной перестановке, и ранг $i_j$-го элемента можно задать как

$$rgX_{i_j} = R + \pi(j),$$

где

$R$ – ранг элемента, предшествующего по порядку группе элементов $X_{i_1}, \dots, X_{i_k}$,

$\pi(j),\ j = 1, \dots, k$ – произвольная перестановка $j$-го элемента в группе элементов, имеющих одинаковые значения.

Коэффициент попарной ранговой корреляции $r$ из (1) может принимать различные значения в зависимости от выбора перестановки $\pi(j)$. Поэтому для однозначности вычисления коэффициента (1) используется ранг, усредненный по всем перестановкам $\pi(j),\ j = 1, \dots, k$. Равные по значению элементы получают одинаковое значение усредненного ранга, не зависящее от их перестановки.

Отметим, что в случае, когда все частоты внутри каждого из словарей не совпадают (все элементы выборок $\overline{X^n}$ и $\overline{Y^n}$ различны), формула (1) преобразуется в классические формулы ранговой корреляции Спирмана, представленные в [9].

## Результаты анализа текстов сообществ соцсети

Анализировались частотные словари буквосочетаний различной длины. Строились и сравнивались посредством коэффициента попарной ранговой корреляции словари буквосочетаний кириллического алфавита и всех встречаемых символов длиной от одного символа до 6. Примеры сравнения по коэффициенту попарной ранговой корреляции словарей буквосочетаний кириллического алфавита длиной 2 и 5 представлены в табл. 2.

Сравнения по коэффициенту попарной ранговой корреляции частотных словарей буквосочетаний кириллического алфавита показывают сильное совпадение частотных распределений для буквосочетаний длиной от 1 до 3 (пример см. в табл. 2). Данный результат подтверждает утверждение о том, что буквосочетания длиной не более трех характеризуют язык (все наборы текстов на русском языке).

При увеличении длин исследуемых буквосочетаний уменьшается согласованность частотных словарей для наборов текстов разных сообществ (пример см. в табл. 2), что указывает на возможность разделения текстов по содержательной направленности. Аналогичные результаты показывает и сравнение частотных словарей псевдоосновов (см. табл. 2). Но данные характеристики определяют в первую очередь тематику текстов и не показывают разделение текстов для разных неявных сообществ, выделенных из графа общения социальной сети.

Сравнение частотных словарей существительных, прилагательных и глаголов для различных наборов текстов по коэффициенту попарной ранговой корреляции показывают близость всех значений коэффициентов к нулевым значениям. Размеры сравниваемых словарей указанных лексических характеристик приведены в табл. 1. Пример такого сравнения приведен в табл. 2 для существительных. Данные результаты говорят о невозможности принять решение о сравнении текстов разных сообществ и формировании системы дифференцирующих признаков.

Сравнение частотных словарей глагольных групп показывает существенные различия между частотными словарями глагольных групп наборов исследуемых текстов. Результаты сравнения частотных словарей глагольных групп приведены в табл. 2, где все значения коэффициента попарной ранговой корреляции приближаются к –1.

Словари глагольных групп попарно «обратны» по частотам использования словосочетаний в текстах. Это указывает на возможность выделить наиболее часто используемые глагольные группы в наборах текстов разных сообществ и рассматривать глагольные группы как дифференцирующие признаки.

*Таблица 2*

Сравнение словарей

*Table 2*

Comparison of dictionaries

| | $S_0$ | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ |
|---|---|---|---|---|---|---|---|---|
| Буквосочетания длиной 2 / Letter combinations of length 2 | | | | | | | | |
| $S_0$ | 1 | | | | | | | |
| $S_1$ | 0.97 | 1 | | | | | | |
| $S_2$ | 0.96 | 0.97 | 1 | | | | | |
| $S_3$ | 0.94 | 0.95 | 0.94 | 1 | | | | |
| $S_4$ | 0.95 | 0.97 | 0.95 | 0.94 | 1 | | | |
| $S_5$ | 0.93 | 0.93 | 0.92 | 0.93 | 0.94 | 1 | | |
| $S_6$ | 0.96 | 0.95 | 0.95 | 0.94 | 0.95 | 0.94 | 1 | |
| $S_7$ | 0.94 | 0.94 | 0.93 | 0.93 | 0.94 | 0.93 | 0.93 | 1 |
| Буквосочетания длиной 5 / Letter combinations of length 5 | | | | | | | | |
| $S_0$ | 1 | | | | | | | |
| $S_1$ | 0.47 | 1 | | | | | | |
| $S_2$ | 0.28 | 0.30 | 1 | | | | | |
| $S_3$ | 0.31 | 0.34 | 0.11 | 1 | | | | |
| $S_4$ | 0.35 | 0.42 | 0.17 | 0.31 | 1 | | | |
| $S_5$ | 0.36 | 0.35 | 0.21 | 0.18 | 0.24 | 1 | | |
| $S_6$ | 0.32 | 0.28 | 0.11 | 0.14 | 0.15 | 0.20 | 1 | |
| $S_7$ | 0.27 | 0.30 | 0.11 | 0.21 | 0.25 | 0.22 | 0.10 | 1 |
| Словари псевдооснов / Pseudo-base dictionaries | | | | | | | | |
| $S_0$ | 1 | | | | | | | |
| $S_1$ | 0.36 | 1 | | | | | | |
| $S_2$ | 0.23 | 0.25 | 1 | | | | | |
| $S_3$ | 0.28 | 0.31 | 0.21 | 1 | | | | |
| $S_4$ | 0.31 | 0.31 | 0.21 | 0.32 | 1 | | | |
| $S_5$ | 0.29 | 0.29 | 0.21 | 0.27 | 0.26 | 1 | | |
| $S_6$ | 0.31 | 0.29 | 0.23 | 0.26 | 0.26 | 0.28 | 1 | |
| $S_7$ | 0.35 | 0.32 | 0.25 | 0.32 | 0.33 | 0.28 | 0.27 | 1 |
| Словари именных групп / Dictionaries of nominal groups | | | | | | | | |
| $S_0$ | 1 | | | | | | | |
| $S_1$ | 0.11 | 1 | | | | | | |
| $S_2$ | –0.05 | 0.04 | 1 | | | | | |
| $S_3$ | –0.09 | 0.02 | –0.24 | 1 | | | | |
| $S_4$ | 0.13 | 0.13 | –0.06 | 0.01 | 1 | | | |
| $S_5$ | –0.06 | –0.01 | –0.17 | –0.22 | –0.05 | 1 | | |
| $S_6$ | –0.09 | –0.01 | –0.24 | –0.28 | –0.07 | –0.22 | 1 | |
| $S_7$ | –0.09 | –0.11 | –0.25 | –0.20 | –0.07 | –0.25 | –0.26 | 1 |
| Словари глагольных групп / Dictionaries of verb groups | | | | | | | | |
| $S_0$ | 1 | | | | | | | |
| $S_1$ | –0.92 | 1 | | | | | | |
| $S_2$ | –0.92 | –0.59 | 1 | | | | | |
| $S_3$ | –0.93 | –0.73 | –0.96 | 1 | | | | |
| $S_4$ | –0.95 | –0.89 | –0.79 | –0.85 | 1 | | | |
| $S_5$ | –0.91 | –0.71 | –0.96 | –0.96 | –0.84 | 1 | | |
| $S_6$ | –0.88 | –0.57 | –0.96 | –0.95 | –0.76 | –0.95 | 1 | |
| $S_7$ | –0.75 | –0.16 | –0.96 | –0.89 | –0.46 | –0.94 | –0.94 | 1 |

## Психолингвистические характеристики

Для исследуемых наборов текстов из социальной сети описанными выше методами компьютерной лингвистики вычислялись наборы статистических характеристик текстов как возможные психолингвистические показатели. Рассматривалось 23 показателя, разбитых на три группы разных типов: определяющие общие характеристики текстов (средние количества лексических единиц); показывающие лексические характеристики текста (лексическое разнообразие); указывающие на использование синтаксических связей в словосочетаниях (относительные длины и составы именных и глагольных групп).

По результатам исследований мы выделили из 23 только 4 характеристики, значения которых, на наш взгляд, отличаются для различных исследуемых наборов текстов (табл. 3):

- коэффициент лексического разнообразия 2 – отношение числа уникальных псевдоосновов к числу словоупотреблений;
- коэффициент действия 1 (КД1) – отношение количества глаголов (без причастий и деепричастий) к количеству прилагательных;
- коэффициент действия 2 (КД2) – отношение количества глаголов с причастиями и деепричастиями к количеству прилагательных;
- коэффициент опредмеченности действия (КОД) – отношение количества глаголов к количеству существительных.

С целью сравнения мы приводим в табл. 3 значения этих характеристик, подсчитанных для двух не связанных с соцсетями контрольных наборов текстов:

- nt – «общественные» (политические, новостные, противоправные) тексты, суммарным размером 2,47 Mb;
- lit – литературные тексты (рассказы русских авторов), суммарным размером 4,36 Mb.

*Таблица 3*

Психолингвистические факторы для различных наборов текстов

*Table 3*

Psycholinguistic factors for different sets of texts

| Характеристика | $S_0$ | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ | nt | lit |
|---|---|---|---|---|---|---|---|---|---|---|
| Коэффициент лексического разнообразия 2 | 0.35 | 0.34 | 0.40 | 0.36 | 0.34 | 0.40 | 0.40 | 0.42 | 0.16 | 0.12 |
| Коэффициент действия 1 (КД1) | 1.93 | 1.88 | 1.73 | 2.16 | 1.95 | 2.15 | 1.95 | 2.23 | 2.59 | 1.82 |
| Коэффициент действия 2 (КД2) | 2.14 | 2.10 | 1.92 | 2.37 | 2.20 | 2.38 | 2.16 | 2.43 | 3.00 | 2.14 |
| Коэффициент опредмеченности действия (КОД) | 0.59 | 0.58 | 0.49 | 0.64 | 0.56 | 0.64 | 0.55 | 0.68 | 0.61 | 0.69 |

Коэффициенты действия (КД1 и КД2) у текстовых массивов наиболее плотных сообществ с высокой максимальной внутренней степенью $S_0$, $S_1$ $S_2$, $S_4$ отличаются по своим значениям от других сообществ и контрольных наборов. Более низкие значения этих показателей для текстов сообществ $S_0$, $S_1$ $S_2$, $S_4$ указывают на их «созерцательный» характер, не переходящий к «действиям». Это подтверждается близостью значений к аналогичным показателям для литературных текстов (набор lit) и существенным отличием от показателей КД1 и КД2 для общественных текстов (набор nt).

Таким образом, мы выделили лингвистические характеристики, которые можно рассматривать как предполагаемые психолингвистические факторы.

## Заключение

Предложена и опробована методика частотного анализа текстов социальных сетей. Реализована методика вычисления коэффициента попарной ранговой корреляции для сравнения частотных характеристик текстов на естественном языке.

Проведен сравнительный анализ частотных словарей существительных, прилагательных, глаголов, именных и глагольных групп, буквосочетаний для наборов текстов различных сообществ по коэффициенту попарной ранговой корреляции. Результаты анализа показывают близость всех значений коэффициентов к нулевым значениям, т. е. для рассматриваемого набора данных большинство исследованных факторов не являются дифференцирующим для определения активности неявных сообществ. Таким образом, данный пример показывает, что подобная дифференциация не всегда возможна. Тем не менее сама идея формирования набора лингвистических факторов (например, из глагольных групп) для рассмотрения в качестве признаков активности в неявных сообществах требует дальнейшего изучения.

Демонстрируется возможность выделения психолингвистических показателей текстов социальных сетей, которые можно использовать для оценки направленности определенных групп общения. Применимость данных факторов и поиск новых требует дополнительных исследований.

Данная работа вместе с работами [1; 2] показывает общую комплексную методику исследования неявных сообществ социальных сетей.

## Список литературы

1. **Chepovskiy A. A., Leshchev D. A., Khaykova S. P.** Core Method for Community Detection. In: Complex Networks & Their Applications IX. Volume 1: Proceedings of the Ninth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2020. Springer, 2021, p. 38–50. DOI 10.1007/978-3-030-65347-7_4
2. **Воронин А. Н., Ковалева Ю. В., Чеповский А. А.** Взаимосвязь сетевых характеристик и субъектности сетевых сообществ в социальной сети Твиттер // Вопросы кибербезопасности. 2020. № 3(37). С. 40–57. DOI 10.21681/2311-3456-2020-03-40-57
3. **Аванесян Н. Л., Соловьев Ф. Н., Тихомирова Е. А., Чеповский А. М.** Выявление значимых признаков противоправных текстов // Вопросы кибербезопасности. 2020. № 4 (38). С. 76–84. DOI 10.21681/2311-3456-2020-04-76-84
4. **Лаврентьев А. М., Соловьев Ф. Н., Суворова М. И., Фокина А. И., Чеповский А. М.** Новый комплекс инструментов автоматической обработки текста для платформы ТХМ 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. **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
6. **Соловьев Ф. Н.** Автоматическая обработка текстов на основе платформы ТХМ 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
7. **Чеповский А. М.** Информационные модели в задачах обработки текстов на естественных языках. 2-е изд., перераб. М.: Национальный открытый университет «ИНТУИТ», 2015.
8. **Бендат Дж., Пирсол А.** Прикладной анализ случайных данных. М.: Мир, 1989. 540 с.
9. **Деза Е. И., Деза М. М.** Энциклопедический словарь расстояний. М.: Наука, 2008. 444 с.

## References


1. **Chepovskiy A. A., Leshchev D. A., Khaykova S. P.** Core Method for Community Detection. In: Complex Networks & Their Applications IX. Volume 1: Proceedings of the Ninth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2020. Springer, 2021, p. 38–50. DOI 10.1007/978-3-030-65347-7_4
2. **Voronin A. N., Kovaleva J. B., Chepovskiy A. A.** Interconnection of network characteristics and subjectivity of network communities in the social network twitter. *Voprosy kiberbezopasnosti*, 2020, no. 3(37), p. 40–57. (in Russ.) DOI 10.21681/2311-3456-2020-03-40-57
3. **Avanesyan N. L., Solovev F. N., Tikhomirova E. A., Chepovskiy A. M.** Identifying the significant features in illegal texts. *Voprosy kiberbezopasnosti*, 2020, no. 4 (38), p. 76–84. (in Russ.) DOI 10.21681/2311-3456-2020-04-76-84
4. **Lavrentyev A. M., Solovev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M.** Novyy kompleks instrumentov avtomaticheskoy obrabotki teksta dlya platformy TXM i yego aprobatsiya na korpuse dlya analiza ekstremistskih tekstov. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2018, vol. 16, no. 3, p. 19–31. (in Russ.) DOI 10.25205/1818-7935-2018-16-3-19-31
5. **Lavrentiev A. M., Smirnov I. V., Solovev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M.** Analiz korpusov tekstov terroristicheskoi i antipravovoy napravlennosti. *Voprosy kiberbezopasnosti*, 2019, no. 4 (32), p. 54–60. (in Russ.) DOI 10.21681/2311-3456-2019-4-54-60
6. **Solovev F. N.** Embedding Additional Natural Language Processing Tools into the TXM Platform. *Vestnik NSU. Series: Information Technologies*, 2020, vol. 18, no. 1, p. 74–82. (in Russ.) DOI 10.25205/1818-7900-2020-18-1-74-82
7. **Chepovskiy A. M.** Informatsionnyye modeli v zadachakh obrabotki tekstov na estestvennykh yazykakh. 2nd ed. Moscow, "INTUIT" Press, 2015. (in Russ.)
8. **Bendat J., Piersol A.** Prikladnoy analiz sluchainikh dannikh. Moscow, Mir, 1989, 540 p. (in Russ.)
9. **Deza E. I., Deza M. M.** Enciclopedicheskiy slovar rasstoayniy. Moscow, Nauka, 2008,444 p. (in Russ.)


*Материал поступил в редколлегию*
*Received*
*25.12.2020*

## Сведения об авторах

**Аванесян Нина Левоновна**, студент магистратуры, Национальный исследовательский университет «Высшая школа экономики» (Москва, Россия)

nlavanesyan@edu.hse.ru

**Соловьев Федор Николаевич**, младший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» РАН (Москва, Россия)

the0@yandex.ru

**Чеповский Александр Андреевич**, кандидат физико-математических наук, доцент, Национальный исследовательский университет «Высшая школа экономики» (Москва, Россия)

aachepovsky@hse.ru

**Information about the Authors**

**Nina L. Avanesyan**, Master's Student, National Research University "Higher School of Economics" (Moscow, Russian Federation)

nlavanesyan@edu.hse.ru

**Fedor N. Solovev**, Junior Researcher, Federal Research Center "Informatics and Management" of the Russian Academy of Sciences (Moscow, Russian Federation)
the0@yandex.ru

**Alexander A. Chepovskiy**, PhD (Mathematics), Associate Professor, National Research University Higher School of Economics (Moscow, Russian Federation)

aachepovsky@hse.ru.