ВЫЯВЛЕНИЕ ЗНАЧИМЫХ ПРИЗНАКОВ ПРОТИВОПРАВНЫХ ТЕКСТОВ[1]

Аванесян Н.Л.[2], Соловьев Ф.Н.[3], Тихомирова Е.А.[4], Чеповский А.М.[5]

Цель исследования: *разработка методики определения частотными методами лексических характеристик и психолингвистических факторов, как дифференцирующих признаков для задач идентификации тематики противоправных текстов в целях информационной безопасности.*

Метод: *применялся автоматический морфологический и синтаксический анализ, частотные методы, сравнение автоматически сформированных словарей методами корреляционного анализа.*

Полученные результаты: *разработана методика частотного анализа лексики противоправных текстов, которая позволяет по частотным словарям сравнивать различные наборы текстов и выявлять дифференцирующие признаки; приведена методика вычисления коэффициента попарной ранговой корреляции для сравнения частотных словарей различных лексических характеристик; проведен сравнительный анализ различных по тематике коллекций текстов противоправной направленности; показана возможность использования частотных лексических характеристик для исследования свойств текстов с целью обнаружения противоправных ресурсов и сообщений; показаны возможности использования как морфологических характеристик слов и словосочетаний, так и буквосочетаний в качестве дифференцирующих признаков; показана возможность вычисления психолингвистических показателей противоправных текстов, основанных на автоматическом лингвистическом анализе текстов; выделены психолингвистические характеристики, характерные для текстов различных тематик.*

Ключевые слова: *автоматический анализ текстов, именные группы, ранговая корреляция, психолингвистические характеристики, экстремистские тексты.*

DOI: 10.21681/2311-3456-2020-04-76-84

1. Введение

Применение автоматического анализа текстов в задачах обеспечения информационно-психологической безопасности, выделения групп риска по социальным и психологическим показателям, определения экстремистских текстов методами искусственного интеллекта является актуальной задачей информационных технологий. [1, 2].

Для решения задач информационной безопасности необходимы методы вычисления психолингвистических маркеров и статистических показателей психологической характеристики для текстов на русском языке с целью повышения эффективности существующих методов обнаружения экстремизма в социальных сетях.

Разработке на базе реляционно-ситуационного анализа инструментов лингвостатистических исследований психолингвистических показателей посвящен цикл работ [3, 4, 5]. Проведённая в них работа позволила выявить текстовые признаки, наличие которых дает возможность отличить тексты, написанные людьми с различным уровнем личностных особенностей, исследовать проблемы текстовой психодиагностики. Подробные исследования применения психолингвистических характеристик текстов для задач клинической психиатрии приведены в [6, 7], где методами реляционно-ситуационного анализа определялись частотные характеристики текстов, позволяющие оценивать выраженность в конкретной коллекции текстов лексики и маркеров, принадлежащих к различным психологическим состояниям.

В данной работе исследования проводились на корпусе текстов на естественном языке противоправной направленности, насчитывающий почти 3,3 миллиона словоупотреблений. Корпус описывался и использовался нами в работах [8, 9, 10]. Он содержит противоправные тексты семи категорий, а также нейтральные тексты со схожей лексикой:

Подкорпус «агрессия» — тексты агрессивной направленности, с призывами к беспорядкам.

---

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

2 Аванесян Нина Левоновна, студент магистратуры, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: nlavanesyan@edu.hse.ru

3 Соловьев Фёдор Николаевич, научный сотрудник, Институт физико-технической информатики, Москва, Россия. E-mail: the0@yandex.ru

4 Тихомирова Елизавета Алексеевна, доцент Московского государственного технического университета им. Н.Э. Баумана, Москва, Россия. E-mail: elizarti@bmstu.ru

5 Чеповский Андрей Михайлович, доктор технических наук, профессор, профессор Российского университета Дружбы Народов, Российского технологического университета МИРЭА, Национального исследовательского университета «Высшая школа экономики», Москва, Россия. E-mail: achepovskiy@hse.ru

Подкорпус «фашизм» — тексты, распространяющие фашистскую идеологию.

Подкорпус «идеология» — тексты, пропагандирующие идеологическое и религиозное превосходство.

Подкорпус «национализм» — тексты, распространяющие национализм.

Подкорпус «религиозные» — тексты, призывающие к религиозной ненависти.

Подкорпус «сепаратизм» — тексты, распространяющие сепаратизм.

Подкорпус «сепаратизм» — тексты террористической направленности.

Подкорпус «нейтральные» — нейтральные тексты.

Взятый для исследования корпус текстов противоправной направленности подробно анализировался в работах [11, 12, 13], в которых анализ подкорпусов базировался на платформе ТХМ, являющейся программным комплексом, предназначенным для разнообразного анализа текстов на естественном языке.

В работах [12, 13] было показано, что указанные выше подкорпуса хорошо разделяются по тематике между собой и противопоставлены нейтральному подкорпусу. Делается вывод, что сформированный корпус может быть использован для машинного обучения в задачах классификации текстов на предмет выявления заданного содержания с целью их углубленного экспертного анализа. Разработаны методы выявления различных дифференцирующих признаков и их комбинаций для тематической классификации подкорпусов текстов, решена задача выделения дифференцирующих признаков с целью применения методов классификации для выявления экстремистских текстов в Интернете. В [14] предложена и опробована методика сравнительного анализа подкорпусов рассматриваемого корпуса текстов, которая позволяет выявлять неявные связи между корпусами разнородных текстов и основана на методиках корпусного анализа.

В данной работе анализ проводится статистическими методами компьютерной лингвистики без использования средств корпусной лингвистики, сравнением частотных словарей лексических единиц. В рамках статистического анализа делается попытка сформировать психолингвистические факторы, позволяющие выявлять направленность текстов по их эмоциональному содержанию.

2. Применяемые методы автоматической обработки текстов

Характеристики текстов определялись процедурами автоматизированной обработки текстов на естественных языках, описанными в [15, 16].

Осуществлялся автоматический морфологический анализ словоформ на основе словарной компьютерной морфологии. Используемая морфологическая модель относит каждое слово к одному из 24 морфологических классов, включающих, помимо частей речи в традиционном понимании, такие разряды, как «неизменяемое слово», «аббревиатура», «топоним». Каждый из этих морфологических классов характеризуется набором грамматических характеристик: род, падеж, число, наклонение и др. Каждая словоформа содержит свои грамматические характеристики и ее каноническую(начальную) форму.

Определялась именная группа, группа слов, у которой главное слово существительное, а другие слова связаны с ним подчинительными синтаксическими связями. При выделении именных групп решалась задача снятия омонимической неопределенности, проистекающей из множественности морфологических разборов отдельных словоупотреблений. Методика выделения именных групп основана на рассмотрении всего множества возможных морфологических разборов каждого слова.

В текстах выделялись глагольные группы, представляющие собой словосочетания, главным словом которых является глагол. Связи найденных именных групп с глаголами строятся на основе синтаксического анализа предложения. Определяется глагольное управления, как разновидность синтаксической подчинительной связи типа управления, в которой главным словом является глагол. При анализе глагольного управления главным словом (глаголом) 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 отличие от отдельных слов, выделенные именные и глагольные группы несут информацию о конкретных отдельных аспектах содержания текста.

В качестве одной из лингвистических характеристик текста используется псевдооснова, под которой понимается часть слова, не содержащая суффиксов и префиксов. Способ автоматического выделения псевдооснов состоит в сопоставлении рассматриваемой словоформы с множеством допустимых в языке структур некорневой части слова [16]. Псевдооснова слова выделяется отбрасыванием всех соответствующих определенной структурной схеме аффиксов, описывающей допустимую в данном языке максимальную комбинации префиксов и суффиксов. Метод псевдооснов позволяет анализировать текстовые конструкции, опираясь не только на точные словоформы.

3. Ранговый анализ частотных словарей

Сравнительный анализ подкорпусов проводился в первую очередь попарным сравнением частотных словарей различных лексических характеристик, составленных для всех исследуемых подкорпусов.

Для оценки близости частотных словарей устанавливаются ранги записей словаря после сортировки по частоте встречаемости занесенной в словарь характеристики. Словари сравниваются вычислением для каждой пары словарей разных подкорпусов коэффициента попарной ранговой корреляции. Записи словарей рассматриваются как случайные величины. Связи между наборами таких элементов разных словарей определяется как Пирсоновская корреляция рангов значений этих случайных величин.

Определим непараметрическую меру ранговой корреляции двух случайных величин **X, Y**.

DOI: 10.21681/2311-3456-2020-04-76-84

Соответствующие этим величинам выборки обозначим $X^n = \{X_i\}_{i=1}^n, Y^n = \{Y_i\}_{i=1}^n$,, а через rgX_i, rgY_i – ранги элементов выборок. Тогда коэффициента попарной ранговой корреляции. для выборок X^n, Y^n есть коэффициент корреляции Пирсона рангов элементов этих выборок и определяется как:

$$r = r(X^n, Y^n) = \rho(rgX^n, rgY^n) = \frac{cov(rgX^n, rgY^n)}{\sigma rgX^n \sigma rgY^n}, \quad (1)$$

Выборочная ковариация определяется следующим образом:

$$cov(X^n, Y^n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X^n})(Y_i - \overline{Y^n}), \quad (2)$$

Дисперсия в форме:

$$\sigma X^n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X^n})^2}, \quad (3)$$

Получаем коэффициент корреляции в виде:

$$r = \frac{\sum_{i=1}^{n}(rgX_i - \overline{rgX^n})(rgY_i - \overline{rgY^n})}{\sqrt{\sum_{i=1}^{n}(rgX_i - \overline{rgX^n})^2}\sqrt{\sum_{i=1}^{n}(rgY_i - \overline{rgY^n})^2}}, \quad (4)$$

Положив среднее рангов

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

мы можем переписать (4) следующим образом:

$$r = \frac{\sum_{i=1}^{n}(rgX_i - \frac{n+1}{2})(rgY_i - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}\left(rgX_i - \frac{n+1}{2}\right)^2}\sqrt{\sum_{i=1}^{n}\left(rgY_i - \frac{n+1}{2}\right)^2}}, \quad (5)$$

В случае, если имеется несколько элементов $X_{i_1}, \dots, X_{i_k}$ выборки X^n с одинаковыми значениями $(X_{i_1} = \cdots = X_{i_k})$, и элементы упорядочены от X_{i_1} к X_{i_k}, к , тогда ранг i_j-го элемента есть

$$rgX_{i_j} = R + j, \quad (6)$$

где R – ранг элемента, предшествующего по порядку группе элементов $X_{i_1}, \dots, X_{i_k}$. Заметим, что формула (5) не накладывает никаких ограничений на порядок элементов, имеющих одинаковые значения. Значит, в случае, когда элементы $X_{i_1}, \dots, X_{i_k}$ имеют одинаковые значения, они могут быть упорядочены согласно произвольной перестановке $\pi \in S_k$ и тогда

$$rgX_{i_{\pi(j)}} = R + \pi(j), \quad (7)$$

где R взят из (6).

Коэффициент r из (5) может принимать различные значения, в зависимости от выбора перестановки π.

Для того, чтобы сделать определение коэффициента попарной ранговой корреляции однозначным и независимым от перестановок элементов с одинаковым значением, вместо ранга rgX_{i_1} элемента X_i, совпадающего по значению еще с элементами $X_{i_2}, \dots, X_{i_k}$ выборки X^n используется усредненный по всем перестановкам $\pi \in S_k$ ранг:

$$rg'X_{i_1} = \frac{1}{k!}\sum_{\pi \in S_k} rgX_{i_{\pi(1)}}, \quad (8)$$

Формально задается

$$rg'X_{i_1} = R + \frac{k+1}{2}, \quad (9)$$

и

$$\overline{rg'X^n} = \overline{rgX^n} = \frac{n+1}{2}, \quad (10)$$

Таким образом, равные по значению элементы получают одинаковое значение усредненного ранга, не зависящее от их перестановки. Когда все значения X_i различны, получаем $rg'X_i = rgX_i$.

Пусть имеются два частотных словаря над множеством лексических характеристик $W = \{w_i\}$.

$$D_1 = \{d_{1,i}^m = (w_i, f_i^1) | w_i \in W\}$$
$$D_2 = \{d_{2,i}^m = (w_i, f_i^2) | w_i \in W\}$$

Тогда коэффициент попарной ранговой корреляции для этих двух словарей, с учетом элементов с одинаковыми частотами, может быть вычислен с учетом (10) и подстановками rg' вместо rg и f^1, f^2 и соответственно вместо X, Y по формуле

$$r(D_1, D_2) = \frac{\sum_{w \in W}(rg'f^1 - \frac{n+1}{2})(rg'f^2 - \frac{n+1}{2})}{\sqrt{\sum_{w \in W}\left(rg'f^1 - \frac{n+1}{2}\right)^2}\sqrt{\sum_{w \in W}\left(rg'f^2 - \frac{n+1}{2}\right)^2}} \quad (11)$$

Поскольку размеры словарей D_1 и D_2 могут быть достаточно большими, мы рассматриваем только первые (по убыванию частоты) m элементов каждого из словарей, получая таким образом словари $D_1^m = \{d_{1,i}^m = (w_i, f_i^1) | rgf_i^1 \le m\}$ и $D_2^m \{d_{2,i}^m = (w_i, f_i^2) | rgf_i^2 \le m\}$. В случае, если в D_1^m встречается слово w, не встречающееся в D_2,, мы полагаем его частоту в словаре D_2, при вычислении (11) равной 0 и наоборот. В реальных расчетах полагаем значение m =10000.

Коэффициент попарной ранговой корреляции (11) принимает значения на интервале [-1,1]. Близкие к 1 значения говорят о монотонной согласованности словарей: если в одном словаре в паре слов одно имеет частоту выше другого, то оно и в другом имеет частоту выше другого. И так для всех пар. Близкие к -1 – об обратном эффекте: если в одном словаре слово с частотой выше второго, то в другом – наоборот, его частота будет ниже. Тоже для всех пар. Если значение близко к 0, то словари несогласованы: соотношение между частотами слов в одном словаре независимо от соотношения с частотами в другом словаре.

4. Результаты анализа лексики

Исследование лексики в плане сравнения словарей различных подкорпусов проводилось для лексических характеристик в начальной форме слова с морфологическими признаками: существительных, глаголов и прилагательных, именных групп и глагольных групп. Анализировались частотные словари псевдоосновов и буквосочетаний различной длины. Размеры сравниваемых словарей указанных лексических характеристик приведены в таблице 1.

Сравнение частотных словарей существительных, прилагательных и глаголов для различных подкорпусов осуществлены по коэффициенту попарной ранговой корреляции. Наблюдаются существенные попарные несогласованности словарей для данных частей речи для различных подкорпусов. Пример такого сравнения приведен в таблице 2 для существительных. Результаты указывают на возможности использовать частей речи в качестве дифференцирующих признаков тематики текстов подкорпусов, но не отличающих их от нейтрального корпуса.

Таблица 1

Размеры частотных словарей для подкорпусов корпуса противоправных текстов

№	Фактор/подкорпус текстов	агрессивности	фашистской идеологии	идеологии превосходства	националистические	нейтральные	Религиозной ненависти	Распространяющие сепаратизм	террористической направленности
1	Существительных	2577	1441	3531	3599	7390	1816	1773	2035
2	Глаголов	1498	761	2802	2334	4483	1281	1116	1403
3	Прилагательных	1208	751	1579	1456	3733	724	855	835
4	Псевдооснов	8050	4401	13417	11585	28445	5642	5629	6595
5	Именных групп	3867	2101	9769	5063	25986	1975	2454	3020
6	Глагольных групп	1425	725	5416	2237	11020	1054	839	1352

Сравнение частотных словарей именных и глагольных групп показывает существенные различия между частотными словарями именных и глагольных групп подкорпусов по частотам использования словосочетаний. Словари попарно «обратны» по частотам использования словосочетаний в текстах. Это указывает на возможность выделения наиболее часто используемых в каждом из подкорпусов словосочетаний и возможности рассматривать словосочетания как дифференцирующие признаки. Более существенное противопоставление по частотам наблюдается на глагольных группах (Таблица 3).

Проведены сравнения по коэффициенту попарной ранговой корреляции частотных словарей буквосочетаний длиной от 3 до 6 для подкорпусов текстов. Было установлено, что частотные словари буквосочетаний длиной 3 практически совпадают при сравнении по коэффициенту попарной ранговой корреляции, что естественно: буквосочетания длиной 3 характеризуют язык, а все подкорпуса на русском литературном языке.

Сравнение частотных словарей буквосочетаний длиной 5 и 6 показывает различие между частотными словарями. Наибольшая несогласованность словарей наблюдалась на словарях буквосочетаний длиной 6 (таблица 4) из рассмотренных нами словарей буквосочетаний. Результаты исследований показывают, что подкорпуса можно разделить по тематической (содержательной) и эмоциональной (психологичской) направленности на основе анализа буквосочетаний длиной более 5, что не подтверждается для задачи отделения их от нейтральных текстов.

5. Выделение психолингвистических характеристик.

Нами рассматривался набор статистических характеристик текстов, как возможные психолингвистические характеристики. Вычисление психолингвистических показателей основывалось на грамматических характеристиках отдельных словоупотреблений. При этом, для каждого словоупотребления рассматриваются все возможные варианты морфологического разбора. Считается, что словоупотребление обладает той или иной грамматической характеристикой, если она встречается хотя бы в одном из вариантов морфологического разбора.

Внутри каждого предложения анализируются словоупотребления, из которых выделяются именные и глагольные группы. В случае, если многозначность морфологического разбора отдельных словоупотреблений приводит к тому, что выделяются несколько именных (глагольных) групп, идентичных по составу, они признаются дубликатами, и из них все, кроме одной, выбираемой произвольно, отбрасываются.

DOI: 10.21681/2311-3456-2020-04-76-84

Таблица 2

Сравнение словарей существительных в канонической форме

Фактор/подкорпус текстов	агрессивности	фашистской идеологии	идеологии превосходства	националистические	нейтральные	Религиозной ненависти	Распространяющие сепаратизм	террористической направленности
Агрессивности								
Фашистской идеологии	-0.03							
Идеологии превосходства	-0.01	0.10						
Националистические	0.09	0.09	0.06					
Нейтральные	0.28	0.27	0.28	0.26				
Религиозной ненависти	-0.09	-0.12	0.40	0.06	0.27			
Распространяющие сепаратизм	0.02	-0.10	0.06	0.14	0.35	-0.15		
Террористической направленности	0.01	-0.05	0.38	0.11	0.33	0.29	0.01	

Таблица 3

Сравнение словарей глагольных групп в канонической форме

Фактор/подкорпус текстов	агрессивности	фашистской идеологии	идеологии превосходства	националистические	нейтральные	Религиозной ненависти	Распространяющие сепаратизм	террористической направленности
Агрессивности								
Фашистской идеологии	-0.91							
Идеологии превосходства	-0.81	-0.70						
Националистические	-0.90	-0.84	-0.86					
Нейтральные	-0.86	-0.77	-0.94	-0.90				
Религиозной ненависти	-0.94	-0.99	-0.34	-0.87	-0.82			
Распространяющие сепаратизм	-0.91	-0.97	-0.65	-0.84	-0.79	-0.98		
Террористической направленности	-0.96	-0.98	-0.52	-0.90	-0.85	-0.66	-0.97	

Таблица 4

Сравнение словарей буквосочетания длиной 6

Фактор/подкорпус текстов	агрессивности	фашистской идеологии	идеологии превосходства	националистические	нейтральные	Религиозной ненависти	Распространяющие сепаратизм	террористической направленности
Агрессивности								
Фашистской идеологии	-0.06							
Идеологии превосходства	-0.04	-0.04						

Фактор/подкорпус текстов	агрессивности	фашистской идеологии	идеологии превосходства	националистические	нейтральные	Религиозной ненависти	Распространяющие сепаратизм	террористической направленности
Националистические	0.10	-0.02	0.03					
Нейтральные	0.17	0.02	0.12	0.19				
Религиозной ненависти	-0.07	-0.08	0.27	-0.01	0.07			
Распространяющие сепаратизм	0.09	-0.04	0.01	0.12	0.20	-0.02		
Террористической направленности	0.07	-0.01	0.31	0.12	0.21	0.37	0.08	

Выделялись статистические показатели трех типов, определяющие общие структурные характеристики текстов (тип А), показывающие лексическое разнообразие текстов (тип В) и указывающие на использование синтаксических связей в словосочетаниях (тип С). Всего рассматривалось 22 показателя, разбитых на три группы.

А. *Общие статистические характеристики текста:*

А.1.Средняя длина словоупотреблений в символах.

А.2. Средняя длина предложения в словоупотреблениях.

А.3. Отношение числа знаков препинания к общему количеству словоупотреблений.

В. *Лексические характеристики текста:*

В.1. Коэффициент лексического разнообразия – отношение числа уникальных лексем к числу словоупотреблений

В.2. Коэффициент разнообразия по псевдоосновам – отношение числа уникальных псевдооснов к числу словоупотреблений

В.3. Отношение числа местоимений к числу словоупотреблений.

В.4. Отношение числа наречий к числу словоупотреблений.

В.5. Отношение числа прилагательные к числу словоупотреблений.

В.6. Коэффициент глагольности — отношение количества глаголов и глагольных форм (причастий и деепричастий) к общему количеству всех словоупотреблений.

В.7. Коэффициент действия (КД) — отношение количества глаголов (деепричастия и причастия исключаются) к количеству прилагательных.

В.8. Коэффициент опредмеченности действия (КОД) — соотношение количества глаголов (деепричастия и причастия исключаются) к количеству существительных.

В.9. Коэффициент логической связности — отношение общего количества служебных слов (союзов и предлогов) к общему количеству предложений.

В.10.Коэффициент использования служебных слов — отношение общего количества служебных слов (союзов и предлогов) к общему количеству словоупотреблений.

В.11.Коэффициент связности лексики — отношение числа существительных и глаголов (деепричастия и причастия исключаются) к количеству прилагательных и наречий.

С. *Словосочетательные характеристики текстов*:

С.1. Средняя длина именных групп в словоупотреблениях.

С.2. Отношение числа именных групп к числу словоупотреблений.

С.3. Среднее отношение числа именных групп к длине предложения в словоупотреблениях.

С.4. Среднее количество числа «подгрупп» в одной именной группе.

С.5. Среднее отношение числа глагольных групп к длине предложения в словоупотреблениях.

С.6. Отношение числа глагольных групп к числу словоупотреблений.

С.7.Среднее отношение числа глагольных групп к длине предложения в словоупотреблениях.

С.8. Среднее количество числа «подгрупп» в одной глагольной группе.

По результатам расчетов мы выделили из 22 только 6 характеристик, значения которых отличаются для различных исследуемых подкорпусов. Эти факторы представлены в таблице 4.

Коэффициент действия (КД) значительно понижается на подкорпусах текстов распространяющих фашистскую идеологию и сепаратизм и повышается на подкорпусах текстов, пропагандирующих идеологическое превосходство, религиозную ненависть. Интересно, что проявляются различия на двух специфичных показателях структурного многообразия, полученных при анализе синтаксических связей словосочетаний (показатели С4 и С8). Среднее количество числа «подгрупп» в словосочетаниях характерно изменяются для подкорпуса текстов агрессивной направленности (повышается) и подкорпуса текстов, пропагандирующих религиозную ненависть (понижается).

Таким образом, мы выделили лингвистические характеристики, которые можно рассматривать как предполагаемые психолингвистические факторы, характеризующие тексты противоправной направленности.

DOI: 10.21681/2311-3456-2020-04-76-84

Таблица 5

Психолингвистические факторы для различных подкорпусов противоправных текстов

№	Характеристика / подкорпус текстов	агрессивности	фашистской идеологии	идеологии превосходства	националистические	нейтральные	Религиозной ненависти	Распространяющие сепаратизм	террористической направленности
1	B.1. Коэффициент лексического разнообразия.	0.159	0.208	0.074	0.124	0.078	0.184	0.165	0.166
2	B.7. Коэффициент действия (КД)	1.043	0.766	1.941	1.348	1.226	1.849	0.894	1.525
3	B.9. Коэффициент логической связности.	2.42	2.55	2.94	1.96	2.35	3.19	3.0	2.80
4	B.11.Коэффициент связности лексики	3.06	2.92	4.36	3.33	3.39	4.13	2.84	4.14
5	C.4. Среднее количество числа «подгрупп» в одной именной группе.	5.87	4.54	3.90	4.39	4.47	3.45	5.07	4.56
6	C.8. Среднее количество числа «подгрупп» в одной глагольной группе.	7.21	7.15	5.41	5.21	6.04	4.69	6.59	5.97

6. Выводы

Предложена и опробована методика частотного анализа лексики противоправных текстов. Разработан метод вычисления коэффициента попарной ранговой корреляции для сравнения частотных словарей различных лексических характеристик.

На основе сравнительного анализа различных по тематике коллекций текстов противоправной направленности показана возможность использования частотных лексических характеристик для исследования свойств текстов с целью обнаружения противоправных ресурсов.. Показаны возможности использования как морфологических характеристик слов и словосочетаний, так и буквосочетаний в качестве дифференцируюших признаков для разделения текстов по «узкой» тематической направленности.

Показана возможность выделения психолингвистических показателей противоправных текстов, основанных на автоматическом лингвистическом анализе текстов. Применимость данных факторов и поиск новых требует дополнительных исследований методами компьютерной лингвистики.

Данная работа вместе с работами [12, 13, 14] формирует единую методику исследования свойств экстремистских текстов с целью обнаружения противоправных ресурсов и сообщений в Интернете.

Литература

1. Hawkins, R. C. II, & Boyd, R. L. Such stuff as dreams are made on: Dream language, LIWC norms, and personality correlates. Dreaming, 2017, 27(2), 102-121.
2. Latov Y., Grishchenko L., Gaponenko V., Vasiliev F. Mechanisms of Countering the Dissemination of Extremist Materials on the Internet // Big Data-driven World: Legislation Issues and Control Technologies. – Springer, 2019. – P. 145-162.
3. Ковалёв А.К., Кузнецова Ю.М., Минин А.Н., Пенкина М.Ю., Смирнов И.В., Станкевич М.А., Чудова Н.В. Методы выявления по тексту психологических характеристик автора (на примере агрессивности) // Вопросы кибербезопасности. 2019. № 4(32). С. 72-79. DOI: 10.21681/2311-3456-2019-4-72-79.
4. Кузнецова, Ю. М., Смирнов, И. В., Станкевич, М. А., Чудова, Н. В. Создание инструмента автоматического анализа текста в интересах социо-гуманитарных исследований. Часть 2. Машина РСА и опыт ее использования //Искусственный интеллект и принятие решений. – 2019. – №. 3. – С. 40-51.
5. Смирнов И.В., Шелманов А.О., Кузнецова Е.С., Храмоин И.В. Семантико-синтаксический анализ естественных языков. Часть II. Метод семантико-синтаксического анализа текстов // Искусственный интеллект и принятие решений. М.: ИСА РАН – 2014. – №1 – С. 11-24.

6. Ениколопов С. Н., Кузнецова Ю. М., Смирнов И. В., Станкевич М. А., Чудова Н. В. Создание инструмента автоматического анализа текста в интересах социо-гуманитарных исследований. Часть 1. Методические и методологические аспекты // Искусственный интеллект и принятие решений. – 2019. – №. 2. – С. 28-38.
7. Ениколопов С.Н., Медведева Т.И., Воронцова О.Ю. Лингвистические характеристики текстов при депрессии и шизофрении // Медицинская психология в России: электрон. науч. журн. – 2019. – Т . 11, № 5(58) [Электронный ресурс]. – URL: http://mprj.ru (дата обращения: 25.06.2020).
8. Ананьева М. И., Кобозева М. В., Соловьев Ф. Н., Поляков И. В., Чеповский А. М. О проблеме выявления экстремистской направленности в текстах // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2016. Т. 14. № 4. С. 5-13.
9. Ананьева М. И., Девяткин Д. А., Кобозева М. В., Смирнов И. В., Соловьев Ф. Н., Чеповский А. М. Исследование характеристик текстов противоправного содержания // Труды Института системного анализа Российской академии наук. 2017 Т. 67 № 3 С. 86-97.
10. Chepovskiy A., Devyatkin D., Smirnov I., Ananyeva M., Kobozeva M., Solovyev F. Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts), in: 2017 IEEE International Conference on Intelligence and Security Informatics: Security and Big Data, ISI 2017 Institute of Electrical and Electronics Engineers Inc., 2017 P. 188-190.
11. Лаврентьев А. М., Смирнов И. В., Соловьев Ф. Н., Суворова М. И., Фокина А. И., Чеповский А. М. Создание специальных корпусов текстов на основе расширенной платформы ТХМ // Системы высокой доступности. 2018. Т. 14. № 3. С. 76-81.
12. 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 Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2018 Т. 16 № 3 С. 19-31.
13. 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
14. 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: 10.681/2311-3456-2020-03-58-65
15. 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.
16. Чеповский А. М. Информационные модели в задачах обработки текстов на естественных языках. Второе издание, переработанное. М.: Национальный открытый университет «ИНТУИТ», 2015.

Рецензент: *Баранов Александр Павлович, доктор физико-математических наук, академик Академии криптографии России, заместитель Генерального директора Главного научного инновационного внедренческого центра, г. Москва, Россия. E-mail:baranov.ap@yandex.ru.*

IDENTIFYING THE SIGNIFICANT FEATURES IN ILLEGAL TEXTS

Avanesyan N.L.[6], Solovev F.N.[7], Tikhomirova E.A.[8], Chepovskiy A.M.[9]

The purpose of the study: *development of a technique for determining lexical characteristics and psycholinguistic factors as discriminative features for identifying the topics of illegal texts by frequency methods for information security purposes.*

Method: *automatic morphological and syntactic analysis, frequency methods, comparison of auto-generated dictionaries by correlation analysis methods.*

Results: *a technique of frequency analysis of the illegal texts vocabulary has been developed, which allows to compare different sets of texts using frequency dictionaries and identify discriminative features; a technique of calculating pairwise rank correlation coefficient for comparison of frequency dictionaries of various lexical characteristics has been presented; a comparative analysis of different illegal texts collections has been carried out; the possibility of using frequency lexical characteristics to study the properties of texts in order to detect illegal resources and messages has*

6 Nina Avanesyan, master's student, National Research University Higher School of Economics, Moscow, Russia. E-mail: nlavanesyan@edu.hse.ru

7 Fedor Solovev, postgraduate, Federal Research Center "Informatics and Management", Moscow, Russia. E-mail: daha-r@yandex.ru

8 Elizaveta Tikhomirova, associate professor of Bauman Moscow State Technical University, Moscow, Russia. E-mail: elizarti@bmstu.ru

9 Andrey Chepovskiy, Dr. Sc. (Eng.), professor, Peoples Friendship University of Russia (RUDN University), Moscow, Russia, Russian Technological University (RTU MIREA), Moscow, Russia, National Research University Higher School of Economics, Moscow, Russia. E-mail: achepovskiy@hse.ru

been shown; the possibilities of using both morphological characteristics of words and word combinations and letter combinations as discriminative features have been shown; the possibility of calculating the psycholinguistic indicators of illegal texts based on automatic linguistic text analysis has been shown; the psycholinguistic characteristics for texts of various topics have been highlighted.

Keywords: *automated text analysis, noun phrases, rank correlation, psycholinguistics characteristics, extremist texts*

References:

1. Hawkins, R. C. II, & Boyd, R. L. Such stuff as dreams are made on: Dream language, LIWC norms, and personality correlates. Dreaming, 2017, 27(2), 102-121.
2. Latov Y., Grishchenko L., Gaponenko V., Vasiliev F. Mechanisms of Countering the Dissemination of Extremist Materials on the Internet // Big Data-driven World: Legislation Issues and Control Technologies. – Springer, 2019. – P. 145-162.
3. Kovalev A.K., Kuznetsova Y.M., Minin A.N., Penkina M.Y., Smirnov I.V., Stankevich M.A., Chudova N.V. Metodi viayvleniay po tekstu psikhologicheskikh kharakteristik avtora (na primere agressivnosti) // Voprosi kiberbezopasnosti.. 2019. № 4(32). С. 72-79. DOI: 10.21681/2311-3456-2019-4-72-79. (in Russian).
4. Kuzneczova, Yu. M., Smirnov, I. V., Stankevich, M. A., Chudova, N. V. Sozdanie instrumenta avtomaticheskogo analiza teksta v interesax socio-gumanitarny`x issledovanij. Chast` 2. Mashina RSA i opy`t ee ispol`zovaniya //Iskusstvenny`j intellekt i prinyatie reshenij. – 2019. – №. 3. – S. 40-51. (in Russian).
5. Smirnov I.V., Shelmanov A.O., Kuzneczova E.S., Xramoin I.V. Semantiko-sintaksicheskij analiz estestvenny`x yazy`kov. Chast` II. Metod semantiko-sintaksicheskogo analiza tekstov // Iskusstvenny`j intellekt i prinyatie reshenij. M.: ISA RAN – 2014. – №1 – S. 11-24. (in Russian).
6. Enikolopov S. N., Kuzneczova Yu. M., Smirnov I. V., Stankevich M. A., Chudova N. V. Sozdanie instrumenta avtomaticheskogo analiza teksta v interesax socio-gumanitarny`x issledovanij. Chast` 1. Metodicheskie i metodologicheskie aspekty` // Iskusstvenny`j intellekt i prinyatie reshenij. – 2019. – №. 2. – S. 28-38.
7. Enikolopov S.N. , Medvedeva T .I. , Vorontsova O.Y u. Linguistic text characteristics in depression and schizophrenia. Med. psihol. Ross., 2019, vol. 11, no. 5 (in Russian). Available at: http://mprj.ru
8. Anan'yeva M. I., Kobozeva M. V., Solov'yev F. N., Polyakov I. V., Chepovskiy A. M.. The problem of detection of extremist texts // Vestnik NSU. Series: Information Technologies. 2016. Vol. 14. № 4. S. 5-13. (in Russian).
9. Anan'yeva M. I., Devyatkin D. A., Kobozeva M. V., Smirnov I. V., Solov'yev F. N., Chepovskiy A. M. Issledovaniye harakteristik tekstov protivopravnogo soderzhaniya // Trudy Instituta sistemnogo analiza Rossiyskoy akademii nauk. 2017 T. 67 № 3 S. 86-97. (in Russian).
10. Chepovskiy A., Devyatkin D., Smirnov I., Ananyeva M., Kobozeva M., Solovyev F. Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts), in: 2017 IEEE International Conference on Intelligence and Security Informatics: Security and Big Data, ISI 2017 Institute of Electrical and Electronics Engineers Inc., 2017 P. 188-190.
11. Lavrent'ev A. M., Smirnov I. V., Solovyev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M. Sozdaniye spetsial'nyh korpusov tekstov na osnove rasshirennoy platformy TXM // Sistemy vysokoy dostupnosti. 2018. T. 14. № 3. S. 76-81. (in Russian).
12. Lavrentyev A. M., Solovyev 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 Novosibirskogo gosudarstvennogo universiteta. Seriya: Lingvistika i mezhkul'turnaya kommunikatsiya. 2018 T. 16 № 3 S. 19-31. (in Russian).
13. Lavrent'ev A. M., Smirnov I. V., Solovyev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M. Analis korpusov tekstov terroristicheskoi i antipravovoy napravlennosti // Voprosi kiberbezopasnosti. 2019. № 4(32). S. 54-60. DOI: 10.21681/2311-3456-2019-4-54-60 (in Russian).
14. Lavrentyev A. M., Raybova D.M., Tikhomirova E.A., Fokina A. I., Chepovskiy A. M., Sherstinova T.Yu. Sravnitelniy analis specialnikh korpusov tekstov dlay zadach bezopasnosty // Voprosi kiberbezopasnosti.. 2020. № 3(37). С. 58-65. DOI: 10.681/2311-3456-2020-03-58-65. (in Russian).
15. Soloviev 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 Russian)
16. Chepovskiy A. M. Informatsionnyye modeli v zadachah obrabotki tekstov na yestestvennyh yazykah. Vtoroye izdaniye, pererabotannoye. M.: Natsional›nyy otkrytyy niversitet "INTUIT", 2015. (in Russian).на основе следующих параметров: