Компьютерный анализ текстов

Исследование характеристик текстов противоправного содержания*

М.И. Ананьева, Д.А. Девяткин, М.В. Кобозева, И.В. Смирнов,
Ф.Н Соловьев, А.М. Чеповский

Аннотация. В работе описаны корпуса текстов для обучения и тестирования методов обнаружения текстов экстремистской направленности. Выполнено исследование характеристик текстов русскоязычного корпуса. Сформирован набор признаков, характерных для материалов противоправного содержания. Эмпирически показана применимость выявленных признаков для решения задачи обнаружения сообщений экстремистского содержания.
Ключевые слова: *противоправные тексты, психолингвистические признаки, дифференцирующие признаки, классификация текстов.*

Введение

Социальные сети и другие интернет-ресурсы часто используются различными экстремистскими организациями для ведения пропаганды, вербовки новых сторонников и координации действий. В связи с этим становится актуальной задача автоматического мониторинга интернет-ресурсов с целью выявления текстовых сообщений с противоправным содержанием. Настоящая работа посвящена исследованию этой задачи, а именно формированию набора признаков, которые могли бы использоваться при выявлении противоправных текстов с помощью методов тематической классификации. Под противоправными мы будем понимать тексты, попадающие под категории экстремизм и терроризм.

Проблема выявления экстремистских текстов может рассматриваться как задача тематической классификации текстов, в которой роль анализируемых объектов будут играть тексты сообщений в социальных сетях, блогах и других ресурсах. Однако такой подход требует наличия размеченного корпуса текстов [1, 2] и заранее определенного набора дифференцирующих признаков [3, 4], определяющих принадлежность текстов к заданным рубрикам. К таким признакам могут относиться, например, результаты полного лингвистического анализа, ключевая лексика, выделенные именованные сущности [5, 6]. Для русского языка подобные корпуса отсутствуют в открытом доступе в первую очередь по причине правовых ограничений.

В настоящей работе на основе сформированных текстовых корпусов эмпирически оценивалась применимость различных групп дифференцирующих признаков для решения задачи выявления экстремистских текстов с помощью методов их тематической классификации на естественных языках.

1. Исследования в области обнаружения и анализа экстремистских сообщений

Ряд основных проблем, связанных с выделением и анализом экстремистских сообщений, представлен в работах [7, 8]. В [7] при выявлении экстремистских текстов результатом являются сообщения противоправного содержания, обнаруженные в общем потоке текстовой информации, генерируемом социальными медиа. А результат анализа противоправных сообщений представляет собой выявленные поведенческие, структурные и языковые особенности определенных экстремистских групп.

* Работа выполнена при поддержке РФФИ, грант №16-29-09546.

В настоящей работе мы, главным образом, будем ориентироваться на выявление признаков, позволяющих решать задачу обнаружения противоправных текстов. При решении этой задачи исследователи сталкиваются с рядом проблем, таких как отсутствие критериев для автоматического выявления противоправных текстов, потребность в создании и постоянной актуализации словарей и баз данных для применения существующих методов, использование упрощенной (бинарной и тернарной) классификации текстов [8].

В целом стоит отметить небольшое количество прикладных исследований для русского языка. Это отчасти связано с тем, что для них необходимы объемные коллекции материалов, тогда как для русского языка такой базы в открытом доступе нет, и создать ее очень сложно, поскольку все экстремистские материалы блокируются по законам РФ. Примером такого закрытого корпуса является коллекция сообщений пермского сегмента социальной сети «ВКонтакте» (vk.com), содержащих признаки этнической агрессии [9]. В этой работе описывается методика создания системы, позволяющей автоматически выявлять этническую агрессию. На основе этого корпуса был составлен словарь терминов, употребление которых в тексте может свидетельствовать о наличии агрессии.

Авторы работы [6] разработали метод автоматической классификации сюжетов на основе поверхностного синтаксического анализа (т.е. определения частей речи и выявления именованных сущностей), а также семантических признаков. Сюжетом в этой работе считается фрагмент текста, в котором описывается некоторое действие, его исполнитель и результат. Алгоритм обучался на корпусе из 16 930 текстов, собранных с сайтов исламистских экстремистов. Тексты отбирались экспертами в данной области. Все тексты предварительно были размечены вручную на «сюжетные» и «несюжетные».

В [10] представлены примеры сбора, анализа и визуализации террористических материалов, находящихся в открытом доступе. Для исследования авторы взяли списки террористических группировок (664 организации) и их сайтов из надежных правительственных источников и загрузили их содержимое (3,6 миллионов веб-страниц) на английском, арабском и испанском языках. В работе также представлены разработанные на собранной коллекции методы анализа связей, контент-анализа и анализа авторства. В [11] изучалось использование СМИ с террористическими целями. Авторы провели комплексный анализ 50 текстов (52 369 слов), собранных с сайтов экстремистских и террористических групп, в которых сектор Газа был главной темой, и имели место призывы к насилию. Тексты были отобраны из другого готового корпуса (264 текста, 500 000 слов), собранного с сайтов экстремистских и террористических групп.

В последние годы многие ученые обратились к исследованиям Твиттера. Анализ сообщений из данной социальной сети позволяет предсказывать изменения курса валют на рынке, изучать реакцию общества на социальные и политические события, отношение людей к новым технологиям и многое другое. Авторы [12] проводят автоматическое деление твитов на «экстремистские» и «прочие» на основе лексических признаков (наличия религиозных терминов, оскорбительных и негативно окрашенных слов и др.), с применением SVM и метода ближайших соседей (KNN) для бинарной классификации. Для исследования авторы собрали 45 млн. сообщений социальной сети Твиттер, 10 487 тыс. из которых были размечены и составили обучающую выборку. Готовый корпус экстремистских текстов на английском языке описывается в [2]. Объем корпуса составляет 100 текстов (42 480 словоупотреблений). Все тексты были написаны на арабском языке и позже переведены на английский. Корпус имеет разноплановую разметку (синтаксическую, семантическую, анафорическую), которая проводилась автоматически, а затем проверялась вручную. Кроме того, в текстах размечены временные маркеры и события. В [13] с помощью ряда классификаторов выявлялись сообщения сети Твиттер, в которых высказывается поддержка деятельности экстремистских исламских группировок. Каждое сообщение сети представлялось в виде вектора признаков, в котором позиция каждого признака зависит от частоты его встречаемости в сообщении. Для классификации использовались стилометрические признаки: служебные слова, частотные слова, особенности пунктуации, хештеги, биграммы на символах и словах. Авторы [14] поставили задачу обнаружения рекрутинговых сообщений экстремистских организаций и предложили вручную размеченный корпус из 192 случайно выбранных сообщений социальных сетей на английском языке для тестирования методов решения этой задачи. Был получен высокий результат по показателю качества бинарной классификации AUC [15] и сделан вывод, что предложенная задача является решаемой.

Отметим также отчет RAND Corporation [16], посвященный анализу активности террористической группировки ИГИЛ (запрещенной в РФ) в социальной сети Твиттер. Для выявления и анализа различных сообществ использовался автоматизи-

рованный подход, в рамках которого учитывалась лексика сообщений и структура связей между авторами. С помощью методов выявления сообществ проанализировано 23 миллиона сообщений 771 321 пользователя из 36 различных общин и выделено четыре основных «метасообщества». Затем для выявления основных тем коммуникации внутри этих сообществ, а также их отношения к ИГИЛ, применялись методы анализа лексики сообщений.

В [17, 18] разрабатывались словари экстремистской лексики и методы их применения для тематической классификации текстов. В [17] представлены словари экстремистской лексики, содержащие ключевые слова и словосочетания, характерные для различных видов экстремистской деятельности. Предложен метод, основанный на предположении о том, что наличие ключевой лексики, имеющей отношение к экстремизму, является более важным показателем, чем частота ее встречаемости в сообщении. В [18] разработана методика, на основе которой может быть построена система рубрикации текстов на разных естественных языках. Основная идея методики заключается в использовании специализированных тематических словарей, лексика в которых разбита по нескольким уровням.

Задача статистического анализа текстов также близка к нашему исследованию. К статистическим характеристикам текстов относятся такие показатели как отношение количества глаголов к количеству прилагательных в единице текста, количество местоимений первого лица единственного числа, количество безличных глаголов и другие. В [19] с использованием подобных признаков решается задача установления авторства текста с целью выявления плагиата в научных работах.

Существует также ряд работ по идентификации автора для судебных целей текстовых сообщений из Интернета: сообщений Твиттера [20], электронных писем [21], информационных сообщений [22]. В перечисленных работах использовались разные статистические показатели:

– лексические – средняя длина слов, количество коротких слов (1-3 символа), лексическое разнообразие, частота видоизмененных слов с повторяющимися символами и др.;
– синтаксические – частота знаков препинания, специальных символов, отношение служебных слов к общему количеству слов и др.;
– структурные – общее количество предложений и абзацев, а также предложений, начинающихся с заглавной и прописной буквы и др.

Проведенный обзор показывает, что для выявления сообщений противоправного содержания помимо лексики широко используются статистические, стилометрические, синтаксические и семантические признаки текстов.

2. Формирование корпуса текстов противоправного содержания

Первой задачей данного исследования стало создание корпуса текстов для проведения дальнейших экспериментов по их классификации, а также определения психолингвистических, семантических и других характеристик текстов экстремистской направленности. Для этого в данный корпус помимо противоправных текстов вошла и коллекция сходных по тематике, но нейтральных по стилю текстов, таких как сообщения с оппозиционных и проправительственных политических блогов, разрешенные тексты религиозного содержания, новостные статьи.

Общий объем корпуса на текущий момент составляет 493 текста (650 000 словоупотреблений), из которых 368 текстов относятся к категории экстремистских материалов. Все тексты были собраны вручную.

Отталкиваясь от того, что понятие экстремизма неоднородно по своему содержанию и включает в себя разные виды правонарушений, мы классифицировали собранные тексты в соответствии с их тематикой. В результате корпус делится на следующие семь категорий.

1. Терроризм (27 текстов, 3 296 словоупотреблений): тексты с сайтов, запрещенных в РФ организаций (такие как ИГИЛ, Хизб-ут Тахрир и др.), где пропагандируется их идеология, размещаются обращения авторитетных в этих кругах лиц, реакция на действия властей по преследованию членов организации и пр.
2. Идеологические тексты (26 текстов, 21 131 словоупотребление): тексты, в которых утверждается превосходство некоторой религии над другими, распространяются ложные трактовки священных книг, а также призывы принять другую религию.
3. Религиозная ненависть (55 текстов, 16 697 словоупотреблений): тексты, призывающие к активным жестоким действиям против представителей других религий, формирующие негативный образ других религий, приписывающие опасные намерения лицам другого вероисповедания.
4. Сепаратизм (7 текстов, 852 словоупотреблений): материалы, распространяющие идею отделения некоторых субъектов от РФ, содержащие оскорбления и угрозы в адрес этнических групп, проживающих на территории этих субъектов.

5. Национализм (208 текстов, 19 399 словоупотреблений): тексты, утверждающие изначальную враждебность определенной этнической группы, призывающие к физическому уничтожению ее представителей, требующие вытеснения из различных сфер деятельности лиц определенной национальности, ограничить их права и свободы на территории РФ.
6. Агрессия и призывы к беспорядкам (43 текста, 6757 словоупотреблений): тексты, призывающие к участию в несанкционированных митингах и беспорядках, насильственному свержению власти, содержащие оскорбления в адрес представителей власти и угрозы физического уничтожения их самих и членов их семей.
7. Фашизм (13 текстов, 2 059 словоупотреблений): тексты, оправдывающие или поддерживающие проявления неофашизма и геноцида, отчеты лиц, распространяющих символику и идею фашизма, а также обсуждения запрещенных книг.

На данный момент корпус не достаточно объемный и несбалансированный: национализм представлен большим количеством документов, чем все другие категории. Исходя из результатов автоматической классификации текстов данного корпуса в дальнейшем возможно изменение приведенной классификации – объединение близких по тематике категорий или разделение наиболее обширных и неоднородных. Кроме того, планируется расширять корпус новыми материалами, а также применить предложенный метод выявления экстремистских текстов на материале татарского корпуса. На сегодняшний день его объем составляет 276 текстов (148 678 словоупотреблений). На базе этих корпусов (русскоязычного и татарского) были созданы специализированные тематические словари ключевых слов и всех их морфологических форм. [19]. Кроме того, слова русскоязычных словарей делятся на три уровня по степени соответствия тематике:

– первый уровень представлен общеупотребимыми словами, фразами, которые зачастую несут негативный характер и часто встречаются в текстах данной тематики, но не характеризуют ее;
– слова второго уровня встречаются в текстах данной и близких тематик, характеризуют данную тематику, но напрямую к ней не относятся;
– слова третьего уровня наиболее часто встречаются в текстах данной тематики и напрямую связаны с ней.

Такое деление продиктовано многозначностью грамматических морфем в русском языке, что приводит к разнообразию смыслов у разных словосочетаний. Для татарского языка, наоборот, характерна однозначность грамматических морфем. К настоящему времени разработаны русскоязычные словари по следующим тематикам: экстремизм, терроризм, национализм, фашизм, насилие. Татарский язык представлен словарями четырех тематик: экстремизм, терроризм, национализм, сепаратизм.

3. Методика исследования характеристик текстов противоправного содержания

В ходе экспериментов решалась задача выявления лексических, психолингвистических и семантических признаков, позволяющих с помощью методов тематической классификации выявлять тексты противоправного содержания. Для морфологического, синтаксического и семантического анализа текстов корпуса, выделения словосочетаний (именных групп) использовался лингвистический анализатор, созданный в ИСА ФИЦ ИУ РАН [23], и программные комплексы лингвистического анализа [4, 5], включающие частотный анализ. В настоящем исследовании анализировались две основные группы признаков (характеристик) текстов противоправного содержания – лексические, а также психолингвистические и семантические.

Для представления лексических признаков текстов сообщений в ходе различных экспериментов использовалось два подхода. Первым подходом являлся «мешок слов и словосочетаний». Пусть $\{w_1, w_2, .., w_m\}$ – множество мощности m всех слов и словосочетаний, встречающихся в текстах анализируемого корпуса $\mathcal{C}$, $n_i(t)$ – частота признака w_i в некотором тексте t из корпуса $\mathcal{C}$, где i=1,2,..,m. Тогда лексические признаки t_{lex} для каждого текста t можно представить в виде вектора:

$$t_{lex} = (n_1(t), n_2(t), .., n_i(t), .., n_m(t)). \quad (1)$$

Второй подход для представления лексических признаков состоит в следующем. Для русскоязычного корпуса $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$ текстов по исследуемым K тематикам, все тексты $\mathcal{C}_i = \{t_1^i, \dots, t_{|\mathcal{C}_i|}^i\}$ $i = \overline{1,K}$ внутри каждой тематики $\mathcal{C}_i$ $i = \overline{1,K}$ были объединены в один текст T_i $i = \overline{1,K}$ и по каждому такому тексту был построен частотный словарь $D_i = \{(x_1^i, f_1^i), \dots, (x_{|D_i|}^i, f_{|D_i|}^i)\}$, где x_j^i – j-й «признак» – некий текстовый фрагмент, полученный из исходного текста, а f_1^i – «частота», количество таких полученных фрагментов, нормированное на общее число выделенных фрагментов из текста. Множество признаков $\{x_1^i, \dots, x_{|D_i|}^i\}$ словаря D_i обозначим X_i. Выделяемыми признаками могут быть буквенные n-граммы, начальные формы слов или отдельных частей речи, глагольные группы,

именные группы, служебные слова. Каждый извлеченный текстовый фрагмент (признак), может быть отнесен к определенному классу F фрагментов, определяемому способом извлечения признаков: класс начальных форм существительных, класс числительных, класс 4-грамм, и т.п. Набор $\mathcal{F}_l = \{F_{l_1}, \dots, F_{l_{|\mathcal{F}_l|}}\}$ таких классов задает множество выделяемых признаков:

$$\overline{\mathcal{F}_l} = F_{l_1} \sqcup \dots \sqcup F_{|\mathcal{F}_l|}. \quad (2)$$

Были построены следующие наборы признаков: одиночные символы алфавита; последовательности от двух до восьми символов; начальные формы существительных, прилагательных, глаголов, причастий, наречий, количественных числительных, собирательных числительных, местоимений, фамилий, имен, отчеств, топонимов, всех слов; аббревиатуры; союзы; частицы; предлоги; междометия; основы слов; именные, глагольные группы.

Психолингвистические и семантические признаки [24] извлекались из текстов русскоязычного корпуса. Значения психолингвистических признаков оцениваются на основе морфологических признаков лексических единиц анализируемых текстов. Примеры некоторых психолингвистических признаков приведены в табл. 1.

Табл. 1

Примеры психолингвистических признаков

№	Определение
1	Отношение количества глаголов к количеству прилагательных в единице текста
2	Отношение количества глаголов к количеству существительных в единице текста
3	Отношение числа инфинитивов к общему числу глаголов
4	Количество местоимений 1-го лица множественного числа
5	Количество местоимений первого лица единственного числа
6	Отношение числа существительных и глаголов к количеству прилагательных и наречий
7	Количество глаголов прошедшего времени, первого лица, единственного числа
8	Количество местоимений 3-го лица множественного числа
9	Количество безличных глаголов

Значения семантических признаков вычислялись как частоты семантических значений в текстах корпуса. Всего для каждого текста определялись значения 46 семантических признаков.

Для представления значений психолингвистических и семантических признаков текста t из анализируемого корпуса использовался следующий подход. Пусть $\Phi = \{\varphi_1, \varphi_2, .., \varphi_i, .., \varphi_n\}$ – множество мощности n всех семантических и психолингвистических признаков, а $\varphi_i(t)$ – значение признака φ_i в тексте t, где $i=1,..,n$. Тогда значения семантических и психолингвистических признаков t_{psy} для каждого текста t можно представить в виде вектора:

$$t_{psy} = (\varphi_1(t), \varphi_2(t), .., \varphi_i(t), .., \varphi_n(t)). \quad (3)$$

В рамках первого эксперимента выявлялись информативные лексические признаки для каждой из категорий русскоязычного корпуса. Для определения информативности слов и словосочетаний текстов при отнесении их к некоторому тематическому подмножеству σ применялась величина $TIC(w, \mathcal{C}, \sigma)$, вычисляемая следующим образом [25]:

$$idf(w, \tau) = log_2 \frac{|\tau|}{m(w,\tau)}; \quad (4)$$

$$\Delta I(w, \mathcal{C}, \sigma) = idf(w, \mathcal{C} \backslash \sigma) - idf(w, \sigma); \quad (5)$$

$$TIC(w, \mathcal{C}, \sigma) = \Delta I(w, \mathcal{C}, \sigma) H(\Delta I(w, \mathcal{C}, \sigma)); \quad (6)$$

где $m(w, \tau)$ – число текстов в некотором множестве сообщений τ, содержащих слово или словосочетание w, $H(.)$ – функция Хевисайда, $\mathcal{C}$ – корпус текстов.

В рамках второго эксперимента значимость лексических признаков оценивалась с помощью метода *RELIEFF* [26]. Метод RELIEFF применяется для отбора признаков, используемых в задаче бинарной классификации. Классифицируемые объекты представлены признаками. Пусть S – обучающая выборка объектов размера n. F – заданный набор признаков $(f_1, \dots, f_p)$. Объект X задается мерным вектором $(x_1,\dots,x_p)$, где x_j – значение признака f_j. Пусть для каждого из признаков определена функция $diff\,(x_k,y_k)$, принимающая значения на множестве [0,1], где 1 обозначает совпадение значений признаков, а 0 – несовпадение. Метод итеративно аппроксимирует значимость $(r_1,\dots,r_p)$ признаков. Значение r_p лежит в интервале [0,1], где 0 обозначает отсутствие значимости признака, а 1 – высокую значимость. На каждой итерации метод выбирает случайным образом объект X, а также ближайший к нему (по метрике L1) пример $Y^- = (y_1^-, \dots y_p^-)$, из отрицательного и $Y^+ = (y_1^+, \dots, y_p^+)$ положительного класса. Вектор $R = (r_1,\dots,r_p)$ значимости признаков обновляется следующим образом:

$$r_i \leftarrow r_i - ||diff(x_i, y_i^+)||_1 + ||diff(x_i, y_i^-)||_1, \quad (7)$$

После m итераций (m – параметр запуска метода) значимость нормируется в отрезок [0,1]. Таким образом, незначимые признаки получают значение близкое к 0, а значимые – существенно отличное от 0.

В нашей работе мы применяли метод RELIEFF для оценки значимости частотных признаков, отвечающих именным и глагольным группам из текстов обучающей выборки. Значение каждого признака – относительная частота встречаемости в тексте. Лексические признаки текстов в этом эксперименте представлялись с помощью подхода «мешок слов и словосочетаний» в соответствии с формулой (1).

В рамках третьего эксперимента сравнивались частотные словари лексики текстов русскоязычного корпуса (см. формулу (2)), путем вычисления коэффициента корреляции Спирмена [4] между ними. Были проанализированы различные наборы $\mathcal{F}_l$ классов признаков. Для каждого набора $\mathcal{F}_l$ были построены частотные словари $D_1^l, \dots, D_K^l$ по имеющимся текстам $T_1, \dots, T\_K$ .. С целью выделения наиболее удачно разделяющего набора $\mathcal{F}_l$, анализировалась попарная ранговая корреляция (корреляция Спирмена) частотных словарей $\rho^l(D_i^l, D_j^l)\ i,j = \overline{1,K}$. Поскольку вычисление коэффициента корреляции Спирмена предполагает использование совпадающих наборов признаков, коэффициент $\rho^l(D_i^l, D_j^l)$ вычислялся на словарях $\widetilde{D}_i^l = \{(x, f_i) \mid x \in X_i^l \cap X_j^l\}$, $\widetilde{D}_j^l = \{(x, f_j) \mid x \in X_i^l \cap X_j^l\}$.

В четвертом эксперименте для определения потенциальной пригодности лексических признаков при решении задач обнаружения экстремистских сообщений выполнялась оценка качества классификации. Был сформирован тестовый набор данных, в котором для представления текстов также использовался подход «мешок слов и словосочетаний» в соответствии с формулой (1). В этом наборе тексты из русскоязычного корпуса разделялись на две категории – экстремистские и нейтральные. Далее проводилось обучение и тестирование ряда классификаторов на сформированном тестовом наборе. Для оценки качества классификации использовалась F_1-мера, вычисляемая в ходе процедуры 5-кратного перекрестного контроля (cross-validation). В качестве методов классификации использовались мультиномиальный наивный байесовский классификатор, логистическая регрессия, линейный SVM, случайный лес, градиентный бустинг. При проведении этого эксперимента применялась библиотека методов машинного обучения с открытым исходным кодом Scikit-learn, в которой реализованы перечисленные выше методы [27].

В пятом эксперименте исследовалась потенциальная пригодность системы психолингвистических и семантических признаков для решения задачи разделения текстов на экстремистские и нейтральные. Был сформирован экспериментальный набор данных, в котором каждый текст корпуса представлен в виде вектора, состоящего из значений этих признаков (см. формулу (3)), и помечен меткой «нейтральный», либо «экстремистский». Лексические признаки в этот набор не включались. Значимость предложенной системы признаков была подтверждена непосредственно путем обучения и тестирования ряда классификаторов на сформированном наборе данных. Для оценки качества классификации использовался такой же подход, как и в предыдущем эксперименте.

4. Результаты экспериментов

В рамках первого эксперимента выявлялась ключевая лексика для каждой из категорий русскоязычного корпуса. Примеры выявленных слов и словосочетаний для всех категорий корпуса представлены в табл. 2.

Во втором эксперименте применение метода *RELIFF* [26] показало, что большая часть лексических признаков является значимой для решения задачи выявления экстремистских сообщений. Таким образом, эта методика не позволила на имеющемся корпусе текстов разделить признаки по их значимости.

В третьем эксперименте была вычислена ранговая корреляция Спирмена [4] частотных словарей тематик экстремистских текстов русскоязычного корпуса. Наиболее низкая корреляция между парами словарей была достигнута при выборе двух групп признаков: именных и глагольных. Однако даже в этом случае корреляция больше нуля и не опускается ниже значения 0,26 (см. табл. 3). То есть классы либо скоррелированы, либо не разделимы с помощью рассматриваемых наборов признаков. Таким образом, задача отбора групп признаков не находит решения при анализе групп признаков с применением коэффициента корреляции Спирмена.

В четвертом эксперименте производилось выявление экстремистских сообщений на текстах русскоязычного корпуса с учетом только лексических признаков. Для этого тексты русскоязычного корпуса анализировались с помощью нескольких методов классификации, таких как мультиномиальный наивный байесовский классификатор, логистическая регрессия, линейный SVM, случайный лес, гради-

Табл. 2

Примеры ключевых слов (разделены по частям речи)

Категория	**Ключевые слова**		
	Существительные	**Прилагательные**	**Глаголы**
Терроризм	Подрыв, взрыв, перестрелка, мощность, заложник, гранатомет.	Вооруженный, взрывной, верховный, проезжий.	Взорваться, застрелить, подъехать, обстрелять, попадаться, размещаться.
Идеологические тексты	Халифат, хизб-ут-тахрир, аллах, посланник, халяль, агент, шазада, конференция, мусульманин.	Славянский, царский, международный.	Разнести, умереть, укрепить, церемониться, убеждаться, забивать, рождаться, добиться, торопиться.
Религиозная ненависть	Хизб-ут-тахрир, агент, хаджалмах, православие, уммиюм, вилайат.	Расстрельный, православный, населенный, исламский.	Разнести, прервать, пытать, увезти, проявиться, плодиться, значиться, придерживаться.
Сепаратизм	Привлечение, народ, единица, вклад, фланг, зверство, копье, эффективность.	Электронный, ядерный, четкий, ведущий, мусульманский, многочисленный, чеченский.	Рассылать, привлекать, изнасиловать, ослабить, расстреливать, исключить.
Национализм	Аул, мигрант, славянин, азиат, ничтожество, кацап, дармоед, дитя гор, горец, шваль, отстрел.	Аульский, носатый, типичный, подлый, цивилизованный, азиатский, кавказский.	Рулить, драться, возмущаться, уничтожаться, сдохнуть, резать, грабить.
Агрессия, призывы к беспорядкам	Анархист, каратель, полицай, оккупация, концлагерь, свержение, кучка, шайка,протест.	Чекистский, ультраправый, воровской, неминуемый, националистический.	Четвертовать, ломать, сжигать, свергать, взрывать, превращать, мечтать.
Фашизм	Мерка, бандеровец, нацист, бюрократия, Гитлер.	Немецкий, генетический, арийский, славянский, непримиримый, павший, неизбежный.	Сформироваться, разоблачать, вмешиваться, бросаться, погнать, присягнуть, тушить.

Табл. 3

Корреляция Спирмена по лексике между различными тематиками экстремистских сообщений

Категория	Идеологические тексты	Религиозная ненависть	Сепаратизм	Национализм	Агрессия, призывы к беспорядкам	Фашизм	Терроризм	Нейтральные тексты
Идеологические тексты	1	0.87	0.37	0.44	0.40	0.38	0.34	0.39
Религиозная ненависть	0.87	1	0.41	0.50	0.51	0.50	0.56	0.36
Сепаратизм	0.38	0.41	1	0.37	0.49	0.28	0.67	0.12
Национализм	0.44	0.50	0.37	1	0.43	0.53	0.26	0.38
Агрессия, призывы к беспорядкам	0.40	0.51	0.49	0.43	1	0.56	0.36	0.30
Фашизм	0.38	0.50	0.28	0.53	0.56	1	0.36	0.21
Терроризм	0.34	0.56	0.67	0.26	0.49	0.28	1	0.20
Нейтральные тексты	0.39	0.36	0.12	0.38	0.30	0.21	0.20	1

ентный бустинг. Оценки качества классификации этими методами (F_1-мера и ее среднеквадратичное отклонение σ) представлены в табл. 4. По итогам проведения эксперимента выявлено, что возможно добиться удовлетворительного качества выявления экстремистских текстов на имеющемся русскоязычном корпусе с использованием извлеченных лексических признаков.

Табл. 4

Результаты выявления экстремистских сообщений с использованием лексических признаков

Метод классификации	F_1-мера	σ
Логистическая регрессия	0.90	0.03
Мультиномиальный наивный байесовский классификатор	0.92	0.01
Случайный лес	0.80	0.03
Градиентный бустинг	0.83	0.03

В рамках пятого эксперимента значения психолингвистических и семантических признаков для русскоязычного корпуса использовались для классификации текстов на экстремистские и нейтральные без учета лексических признаков (см. табл. 5).

Табл. 5

Результаты выявления экстремистских сообщений с использованием психолингвистических и семантических признаков

Метод классификации	F_1-мера	σ
Мультиномиальный наивный байесовский классификатор	0.41	0.01
Логистическая регрессия	0,55	0,06
Линейный SVM	0,55	0,03
Случайный лес	0,76	0,03
Градиентный бустинг	0,76	0,03

На основе результатов этого эксперимента можно сделать вывод, что линейные методы классификации (линейный SVM и логистическая регрессия) не позволяют выявлять тексты противоправного содержания на основе значений психолингвистических и семантических признаков (55% по F_1-мере), однако с помощью более сложных методов на основе ансамблей деревьев решений можно добиться удовлетворительного качества классификации без использования лексических признаков (76% по F_1-мере).

Заключение

В ходе исследования сформированы закрытые экспериментальные корпуса текстов на русском и татарском языках для обучения и тестирования методов выявления экстремистских сообщений. Построена специфическая лексика, релевантная отдельным категориям противоправных текстов. Исследовалась система психолингвистических и семантических признаков текстов, используемых ранее для определения уровня эмоциональной напряженности текстов.

Показано, что, используя только психолингвистические и семантические характеристики текстов и методы тематической классификации экстремистских сообщений на основе ансамблей деревьев решений, можно добиться удовлетворительного качества обнаружения экстремистских сообщений. При этом размерность пространства психолингвистических и семантических признаков значительно меньше размерности пространства лексических признаков, поэтому их использование снижает требования к размеру обучающих корпусов и повышает производительность. Для большей надежности и эффективности обнаружения текстов экстремистской направленности целесообразно использовать совместно лексические, семантические и психолингвистические признаки.

В дальнейшем планируется пополнение представленных корпусов текстов на русском и татарском языках и проведение дополнительных экспериментальных исследований по выявлению текстов экстремистской направленности на основе совместного использования лексических, семантических и психолингвистических признаков. Кроме того, планируется провести эксперименты по анализу сообщений сети Твиттер.

Литература

1. *Cohen K., Johansson F., Kaati L. and Mork J.C.* Detecting linguistic markers for radical violence in social media // Terrorism and Political Violence 2014. Vol. 26, No 1. pp. 256–256.
2. *Finlayson M. A., Halverson J. R., Corman S. R.* The N2 corpus: A semantically annotated collection of Islamist extremist stories //LREC. – 2014. – p. 896-902.
3. *Осипов Г.С.* Методы искусственного интеллекта. – М.: ФИЗМАТЛИТ. – 2011. – 296 с.
4. *Чеповский А. М.* Информационные модели в задачах обработки текстов на естественных языках. Второе издание, переработанное. М.: Национальный открытый университет «ИНТУИТ», 2015. – 276 с.
5. *Поляков И.В., Соколова Т.В., Чеповский А.А., Чеповский А.М.* Проблема классификации текстов и дифференцирующие признаки // Вестник

Новосибирского государственного университета. Серия: Информационные технологии. – 2015. – Т. 13. – № 2. – С. 55 – 63.
6. *Ceran B. et al.* A semantic triplet based story classifier //Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). – IEEE Computer Society, 2012. – p. 573-580.
7. *Correa D., Sureka A.* Solutions to detect and analyze online radicalization: a survey //arXiv preprint arXiv:1301.4916. – 2013.
8. *Ананьева М.И, Кобозева М.В., Соловьев Ф.Н., Поляков И.В., Чеповский А.М.* О проблеме выявления экстремистской направленности в текстах. // Вестник НГУ. – 2016. – Т. 14. – №. 4. – С. 5-13.
9. *Жданова С. Ю. и др.* Особенности репрезентации этнической агрессии в корпусе сообщений пермского сегмента социальной сети «Вконтакте»(Vk. com) //Вектор науки Тольяттинского государственного университета. Серия: Педагогика, психология. – 2012. – №. 4 (11).
10. *Chen H.* Exploring extremism and terrorism on the web: the dark web project //Pacific-Asia Workshop on Intelligence and Security Informatics. – Springer Berlin Heidelberg, 2007. – С. 1-20.
11. *Prentice S. et al.* Analyzing the semantic content and persuasive composition of extremist media: A case study of texts produced during the Gaza conflict //Information Systems Frontiers. – 2011. – Vol. 13(1). – pp. 61-73
12. *Agarwal S., Sureka A.* Using KNN and SVM based one-class classifier for detecting online radicalization on twitter //International Conference on Distributed Computing and Internet Technology. – Springer International Publishing, 2015. – pp. 431-442.
13. *Ashcroft M. et al.* Detecting jihadist messages on twitter //Intelligence and Security Informatics Conference (EISIC), 2015 European. – IEEE, 2015. – pp. 161-164.
14. *Scanlon J. R., Gerber M. S.* Automatic detection of cyber-recruitment by violent extremists // Security Informatics. – 2014. – Vol. 3 (1). – p. 1.
15. *Huang J., Ling C. X.* Using AUC and accuracy in evaluating learning algorithms //IEEE Transactions on knowledge and Data Engineering. – 2005. – Vol. 17(3). – pp. 299-310
16. *Bodine-Baron E. et al.* Examining ISIS Support and Opposition Networks on Twitter //RAND Corporation. – 2016. – pp. 29-30.
17. *Wadhwa P., Bhatia M. P. S.* Classification of radical messages in Twitter using security associations // Case studies in secure computing: Achievements and trends. – 2014. – pp. 273-294.
18. *Михайлов А.С., Соколова Т.В., Чеповский А.А., Чеповский А.М.* Выявление тематической направленности текстов на естественных языках // Искусственный интеллект и принятие решений. 2016. – № 1. – С. 9 – 17.
19. *Zurini M.* Stylometry Metrics Selection for Creating a Model for Evaluating the Writing Style of Authors According to Their Cultural Orientation //Informatica Economica. – 2015. – Vol. 19 (3). – pp. 107.
20. *Bhargava M., Mehndiratta P., Asawa K.* Stylometric analysis for authorship attribution on twitter //International Conference on Big Data Analytics. – Springer International Publishing, 2013. – pp. 37-47.
21. *Brocardo M. L., Traore I., Woungang I.* Toward a framework for continuous authentication using stylometry //Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on. – IEEE, 2014. – С. 106-115.
22. *Nirkhi S. M., Dharaskar R. V., Thakare V. M.* Authorship Attribution of online messages using Stylometry: An Exploratory Study //International Conference on Advances in Engineering and Technology (ICAET'2014). – 2014.
23. *Osipov G. et al.* Relational-situational method for intelligent search and analysis of scientific publications //Proceedings of the Integrating IR Technologies for Professional Search Workshop. – 2013. – pp. 57-64.
24. *Vybornova O. et al.* Social tension detection and intention recognition using natural language semantic analysis: On the material of Russian-speaking social networks and Web forums // Intelligence and Security Informatics Conference (EISIC), 2011 European. – IEEE, 2011. – pp. 277-281.
25. *Драль А. А., Соченков И. В., Мбайкоджи Э.* Метод автоматической классификации коротких текстовых сообщений //Информационные технологии и вычислительные системы. – 2012. – С. 93-102.
26. *Kira K., Rendell L.A.* The feature selection problem: Traditional methods and a new algorithm // AAAI. – 1992. – Т. 2. – С. 129 – 134.
27. *Pedregosa F. et al.* Scikit-learn: Machine learning in Python //Journal of Machine Learning Research. – 2011. – Vol. 12. – No. Oct. – pp. 2825-2830

Ананьева Маргарита Игоревна. Младший научный сотрудник ИСА ФИЦ ИУ РАН. Окончила в 2013 г. Московский государственный лингвистический университет. Количество печатных работ: 8. Область научных интересов: методы лингвистического анализа текстов, корпусная лингвистика, дискурсивный анализ текстов. E-mail: ananyeva@isa.ru

Девяткин Дмитрий Алексеевич. Младший научный сотрудник ИСА ФИЦ ИУ РАН. Окончил в 2011 г. Рыбинскую государственную авиационную технологическую академию. Количество печатных работ: 21. Область научных интересов: машинное обучение, классификация и кластеризация текстов, методы обработки больших данных, методы анализа патентной и наукометрической информации.
E-mail: devyatkin@isa.ru

Кобозева Мария Вадимовна. Младший научный сотрудник ИСА ФИЦ ИУ РАН. Окончила в 2014 г. МГУ им. М.В. Ломоносова, в 2016 г. – Магистратуру РГГУ. Количество печатных работ: 5. Область научных интересов: компьютерная лингвистика, автоматическая обработка естественного языка, дискурсивная структура текста. E-mail: kobozeva@isa.ru

Смирнов Иван Валентинович. Доцент. Заведующий лабораторией ИСА ФИЦ ИУ РАН. Окончил в 2003 г. РУДН. Кандидат физико-математических наук. Количество печатных работ: 52. Область научных интересов: обработка естественного языка, интеллектуальный анализ информации. E-mail: ivs@isa.ru

Соловьев Федор Николаевич. Младший научный сотрудник Института Физико-технической информатики (г. Протвино). Окончил в 2014 г. МФТИ. Область научных интересов: компьютерная лингвистика, автоматическая обработка текстов, распознавание образов. E-mail: the0@yandex.ru

Чеповский Андрей Михайлович. Профессор Национального исследовательского университета «Высшая школа экономики» и Московского политехнического университета. Окончил Московский энергетический институт в 1979 г. Доктор технических наук. Количество печатных работ: более 100. Область научных интересов: автоматическая обработка текста, информационный поиск, кибернетика, информационная безопасность. E-mail: achepovskiy@hse.ru

The study of extremist texts features

Ananyeva M., Devyatkin D., Kobozeva M., Smirnov I., Solovyev F., Chepovskiy A.

Abstract. This article presents methods for identifying the extremist activities of violent groups and individuals within the Internet. We describe our training and testing datasets in Russian and Tatar, as well as research of Russian extremist text characteristics. This resulted in a formation of a feature set for the extremist texts. The applicability of these features for detection of extremist messages was empirically showed.
Keywords: *extremist texts, psycholinguistic features, separating features, text classification.*

References

1. *Cohen K., Johansson F., Kaati L. and Mork J.C.* Detecting linguistic markers for radical violence in social media // Terrorism and Political Violence 2014. Vol. 26, No 1. pp. 256–256.
2. *Finlayson M. A., Halverson J. R., Corman S. R.* The N2 corpus: A semantically annotated collection of Islamist extremist stories //LREC. – 2014. – p. 896-902.
3. *Osipov G.* 2011. Metody iskusstvennogo intellekta [Methods for artificial intelligence]. Moscow: Fizmatlit. 296 p.
4. *Chepovskiy A. M.* 2015. Informatsionnyye modeli v zadachakh obrabotki tekstov na yestestvennykh yazykakh. Vtoroye izdaniye [Information models for text processing. Second edition]. Moscow: The National Open University "INTUIT". 276 p.
5. *Polyakov I.V., Sokolova T.V., Chepovskiy A.A., Chepovskiy A.M.* 2015. Problema klassifikatsii tekstov i differentsiruyushchiye priznaki [The problem of text classification and separating features] Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnyye tekhnologii [Bulletin of the Novosibirsk State University. Series: Information Technology] 13. № 2:55–63.
6. *Ceran B. et al.* A semantic triplet based story classifier //Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). – IEEE Computer Society, 2012. – p. 573-580.
7. *Correa D., Sureka A.* Solutions to detect and analyze online radicalization: a survey //arXiv preprint arXiv:1301.4916. – 2013.
8. *Ananyeva M.I., Kobozeva M.I., Solovyev F.N., Polyakov I.V., Chepovskiy A.M.* 2016. O probleme vyyavleniya ekstremistskoy napravlennosti v tekstakh [About the problem of extremist texts identification]. Vestnik Novosibirskogo gosudarstvennogo universiteta [Bulletin of the Novosibirsk State University]. 14. №. 4:5-13.
9. *Zhdanova S.Yu. et al.* 2012. Osobennosti reprezentatsii etnicheskoy agressii v korpuse soobshcheniy permskogo segmenta sotsial'noy seti «Vkontakte» (Vk. Com) [Features of representation of ethnic aggression in the Permian segment of the social network "Vkontakte" (vk.com)]. Vektor nauki Tol'yattinskogo gosudarstvennogo universiteta. Seriya: Pedagogika, psikhologiy [Vector of Science. Togliatti State University: Pedagogy, Psychology]. 4 (11).
10. *Chen H.* Exploring extremism and terrorism on the web: the dark web project //Pacific-Asia Workshop on Intelligence and Security Informatics. – Springer Berlin Heidelberg, 2007. – C. 1-20.
11. *Prentice S. et al.* Analyzing the semantic content and persuasive composition of extremist media: A case study of texts produced during the Gaza conflict //Information Systems Frontiers. – 2011. – Vol. 13(1). – pp. 61-73
12. *Agarwal S., Sureka A.* Using KNN and SVM based one-class classifier for detecting online radicalization on twitter //International Conference on Distributed Computing and Internet Technology. – Springer International Publishing, 2015. – pp. 431-442.
13. *Ashcroft M. et al.* Detecting jihadist messages on twitter //Intelligence and Security Informatics Conference (EISIC), 2015 European. – IEEE, 2015. – pp. 161-164.
14. *Scanlon J. R., Gerber M. S.* Automatic detection of cyber-recruitment by violent extremists // Security Informatics. – 2014. – Vol. 3 (1). – p. 1.
15. *Huang J., Ling C. X.* Using AUC and accuracy in evaluating learning algorithms //IEEE Transactions on knowledge and Data Engineering. – 2005. – Vol. 17(3). – pp. 299-310
16. *Bodine-Baron E. et al.* Examining ISIS Support and Opposition Networks on Twitter //RAND Corporation. – 2016. – pp. 29-30.
17. *Wadhwa P., Bhatia M. P. S.* Classification of radical messages in Twitter using security associations // Case studies in secure computing: Achievements and trends. – 2014. – pp. 273-294.
18. *Mikhaylova A.S., Sokolova T.V., Chepovskiy A.A., Chepovskiy A.M.* 2016. Vyyavleniye tematicheskoy napravlennosti tekstov na yestestvennykh yazykakh [Identification of thematic focus of texts]. Iskusstvennyy intellekt i prinyatiye resheniy [Artificial intelligence and decision-making].1:9–17.

19. *Zurini M.* Stylometry Metrics Selection for Creating a Model for Evaluating the Writing Style of Authors According to Their Cultural Orientation //Informatica Economica. – 2015. – Vol. 19 (3). – pp. 107.
20. *Bhargava M., Mehndiratta P., Asawa K.* Stylometric analysis for authorship attribution on twitter //International Conference on Big Data Analytics. – Springer International Publishing, 2013. – pp. 37-47.
21. *Brocardo M. L., Traore I., Woungang I.* Toward a framework for continuous authentication using stylometry //Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on. – IEEE, 2014. – C. 106-115.
22. *Nirkhi S. M., Dharaskar R. V., Thakare V. M.* Authorship Attribution of online messages using Stylometry: An Exploratory Study //International Conference on Advances in Engineering and Technology (ICAET'2014). – 2014.
23. *Osipov G. et al.* Relational-situational method for intelligent search and analysis of scientific publications //Proceedings of the Integrating IR Technologies for Professional Search Workshop. – 2013. – pp. 57-64.
24. *Výbornova O. et al.* Social tension detection and intention recognition using natural language semantic analysis: On the material of Russian-speaking social networks and Web forums // Intelligence and Security Informatics Conference (EISIC), 2011 European. – IEEE, 2011. – pp. 277-281.
25. *Dral A.A., Sochenkov I.V., Mbaykodzi E.* 2012. Metod avtomaticheskoy klassifikatsii korotkikh tekstovykh soobshcheniy [The method of automatic classification of short text messages]. Informatsionnyye tekhnologii i vychislitel'nyye sistemy [Information technology and computer systems]. 93-102p.
26. *Kira K., Rendell L.A.* The feature selection problem: Traditional methods and a new algorithm // AAAI. – 1992. – T. 2. – C. 129 – 134.
27. *Pedregosa F. et al.* Scikit-learn: Machine learning in Python //Journal of Machine Learning Research. – 2011. – Vol. 12. – No. Oct. – pp. 2825-2830

Margarita I. **Ananyeva.** Junior research fellow, ISA FRC CSC RAS. Graduated from Moscow State Linguistic University in 2013. Author of 9 scientific papers. Research interests: methods for linguistic analysis, information technology, discourse analysis. E-mail: ananyeva@isa.ru

Dmitry A. Devyatkin. Researcher, ISA FRC CSC RAS. Graduated from Rybinsk State Aviation Technology Academy after Pavel Solovyov in 2011. Author of 21 scientific papers. Main research interests are machine learning, full-text clustering, data mining, scientometrics. E-mail: devyatkin@isa.ru.

Maria V Kobozeva. Junior research fellow, ISA FRC CSC RAS. Graduated from Moscow State University in 2014 and from Russian State University for the Humanities in 2016. Author of 5 scientific papers. Research interests: computational linguistics, natural language processing, discourse analysis. E-mail: kobozeva@isa.ru

Ivan V. Smirnov. PhD, associate professor, head of laboratory, ISA FRC CSC RAS. Author of 52 scientific papers. Research interests: natural language processing, data and text mining. E-mail: ivs@isa.ru.

Fedor N. Solovyev. Junior research fellow in the Institute of Computing for Physics and Technology. Graduated from Moscow Physical Technical Institute (state university) in 2014. Research interests: computational linguistics, natural language processing, pattern recognition. E-mail: the0@yandex.ru

Andrey M. Chepovskiy. Doctor of Technical Science, professor of chair of information security National Research University «Higher School of Economics». Graduated from Moscow Power Engineering Institute in 1979. Professor of chair of applied mathematics and modeling of systems at Moscow Polytechnic University. Author of 100 publications. Research interests: natural language processing, information retrieval, mathematical cybernetics, cybernetics. E-mail: achepovskiy@hse.ru