АНАЛИЗ КОРПУСОВ ТЕКСТОВ ТЕРРОРИСТИЧЕСКОЙ И АНТИПРАВОВОЙ НАПРАВЛЕННОСТИ[1]

Лаврентьев А.М.[2], Смирнов И.В.[3], Соловьев Ф.Н.[4], Суворова М.И.[5], Фокина А.И.[6], Чеповский А.М.[7]

Цель исследования: *разработка методики создания и автоматического анализа специальных корпусов текстов для последующего применения их в качестве обучающих выборок и определения дифференцирующих признаков в задачах классификации текстов.*

Метод: *применялись инструменты анализа корпусной платформы TXM, расширенной разработанными процедурами вычисления дополнительных характеристик текстов, таких как буквосочетания, псевдоосновы, именные группы, глагольные группы.*

Полученные результаты: *показано, что разработанные средства расширения корпусной платформы TXM позволяют эффективно решать задачи анализа текстов специальной тематики, созданный корпус текстов экстремистской тематики может использоваться в качестве обучающей выборки для задач классификации текстов, делается вывод об использовании буквосочетаний как универсальных дифференцирующих признаков наряду с классическими лингвистическими характеристиками текстов.*

Ключевые слова: *корпусная лингвистика, автоматический анализ текстов, платформа TXM, псевдоосновы, именные группы, глагольное управление, выявление экстремистских текстов*

DOI: *10.21681/2311-3456-2019-4-54-60*

1. Введение

В связи с развитием средств электронных коммуникаций актуальны разработки методов противодействия распространению экстремизма в сети. Например, в работе [3] рассматриваются вопросы борьбы с экстремизмом и терроризмом в интернете, анализируется российское законодательство в данной области, анализируются судебные дела и прокурорский надзор за исполнением законов о противодействии киберэкстремизму и кибертерроризму, даются рекомендации по оптимизации прокурорского надзора в данном направлении. В исследованиях психологов [5, 6] анализировались методы пропаганды и вербовки, используемые членами Исламского государства в отношении молодежи: как вербуют новых сторонников, каких коммуникативных стратегий придерживаются.

Работ по автоматическому обнаружению противоправных текстов не так много. Однако авторы многих работ признают необходимость создания соответствующего программного обеспечения для экспертов [10, 20].

Для анализа реальных данных из социальных сетей разрабатывается прикладное программные обеспечение. В работе [7] описываются методы обнаружения в сети электронных сообщений, документов, web-ресурсов, содержащих экстремистскую информацию, а также поиска пользователей и сообществ в социальных сетях, распространяющих такую информацию. Предложенные методы основаны на выделении ключевых слов образца, из которых формируются поисковые запросы для социальной сети и тематики документа-образца. В работе [4] описана автоматизированная система, которая по составленным запросам загружает из интернета выдачу, кластеризует документы по нескольким антиправовым тематикам (наркоторговля, терроризм, экстремизм), отсеивая то, что не соответствует ни

1 Работа выполнена при финансовой поддержке РФФИ в рамках научных проектов № 16-29-09546, № 18-00-00606 (18-00-00233) и № 19-07-00806.

2 Лаврентьев Алексей Михайлович, кандидат филологических наук, Ph.D., Институт истории представлений и идей нового времени НЦНИ и Высшей нормальной школы Лиона, г. Лион, Франция. E-mail: alexei.lavrentev@ens-lyon.fr

3 Смирнов Иван Валентинович, кандидат физико-математических наук, заведующий отделом, Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Россия. E-mail: ivs@isa.ru

4 Соловьев Фёдор Николаевич, научный сотрудник, Институт физико-технической информатики, г. Протвино, Московская область, Россия. E-mail: the0@yandex.ru

5 Суворова Маргарита Игоревна, научный сотрудник, Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия. E-mail: suvorova@isa.ru.

6 Фокина Алина Игоревна, бакалавр, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: aifokina@edu.hse.ru

7 Чеповский Андрей Михайлович, доктор технических наук, профессор, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: achepovskiy@hse.ru

одной тематике. Затем эксперт проверяет те материалы, которые были отобраны системой. Отметим, что в большинстве прикладных исследований тестов из социальных сетей (например, [17]) западные исследователи используют примитивные программные продукты для вычисления частотных характеристик.

В предыдущих работах [1, 2, 15] нами был создан корпус текстов, содержащий противоправные тексты семи категорий (терроризм, идеологические тексты, религиозная ненависть, сепаратизм, национализм, агрессия и призывы к беспорядкам, фашизм) и нейтральные тексты со схожей лексикой. Были предложены [8, 9] различные расширения стандартной корпусной платформы для изучения специализированных корпусов текстов. В данной работе проводятся исследования применения всех разработанных методов для анализа корпусов противоправных текстов.

2. Задача распознавания тематики текстов

Задача автоматического обнаружения противоправных текстов является задачей построения систем классификации больших объемов текстовой информации, которая состоит в определении принадлежности входного текста одному или нескольким классам. С математической точки зрения это задача распознавания образов в алгебраической постановке [11, 13], а значит, когда обучение завершено, принадлежность незнакомого текста к одному из классов устанавливается также путем статистического анализа обнаруженных в тексте признаков.

Выбор дифференцирующих признаков имеет важное значение для итогового качества классификаторов. Необходимость ограничения множества дифференцирующих признаков определяется тем, что увеличение используемого числа признаков требует соответствующего увеличения обучающей выборки. Для достаточно точной оценки требуется, чтобы каждый используемый признак встретился в обучающей выборке хотя бы несколько раз, что практически невозможно обеспечить, если брать в качестве множества признаков все возможные слова русского языка или же ограничившись только множеством всех существительных.

Таким образом, задача отбора дифференцирующих признаков имеет определяющее значение для задачи классификации [11, 12]. Для текстов на естественных языках в качестве дифференцирующих признаков можно рассматривать: существительные; существительные и прилагательные; существительные, прилагательные и глаголы; существительные и именные группы; глаголы и глагольные группы; псевдоосновы словоупотреблений текста, полученные алгоритмами аналитического морфологического анализа. Очевидно, что учет различных морфологических признаков оказывает различное влияние на показатели классификации. Для некоторых тематик могут оказывать положительное влияние на показатели классификации существительные и именные группы, а на определение других тематик отрицательное влияние оказывает учет глагольных групп.

Поэтому создание экспертами обучающих по заданным тематикам корпусов текстов с последующим их анализом с целью исследования различных наборов признаков является ключевым в решении задачи тематической классификации текстов.

3. Расширение платформы корпусного анализа

В данной работе мы опираемся на платформу TXM (http://textometrie.org), являющуюся эффективным программным комплексом корпусного анализа, позволяющим проводить анализ корпусов (анализ соответствий, кластеризация, построение лексических таблиц, поиск сложных лексических конструкций, выделение подкорпусов по различным параметрам). Платформа TXM использует словоупотребления в качестве структурных единиц анализа. Для повышения эффективности таких используемых TXM методов, как анализ специфичности и анализ соответствий, целесообразно ввести в рассмотрение новые единицы анализа, опирающиеся на процедуры автоматизированной обработки текстов на естественных языках, описанные в [8, 9, 12, 13].

В настоящих исследованиях мы использовали интегрированный в TXM программный пакет TreeTagger [21], предоставляющий возможность совместного морфологического анализа слов предложения на основе статистической модели, путем сопоставления словоупотреблений, снабжённых специальными метками, кодирующими морфологические характеристики. При предобработке всех русскоязычных текстов мы осуществляем автоматический морфологический анализ словоформ на основе словарной компьютерной морфологии, описанной в [13].

Для определения дифференцирующих признаков коротких текстов, характерных для экстремистской тематики в социальных сетях, мы используем аналитический метод выделения псевдооснов, так как он позволяет обрабатывать отсутствующие в стандартных словарях формы. Используемый способ выделения псевдооснов представляет собой метод структурных схем, описанный подробно в [13, 16]. Суть метода состоит в получении псевдоосновы словоформы путем рассмотрения и отбрасывания ее словоизменительных аффиксов. Данный подход позволяет анализировать текстовые конструкции, опираясь не только на точные словоформы, и тем самым повышает полноту и гибкость корпусного анализа.

Был предложен [1, 8, 9] ряд расширений, позволяющих дополнить и усложнить анализ корпусов, включающий: автоматический морфологический анализ словоформ и приведение их к канонической форме, выделение псевдооснов, выделение именных и глагольных групп и комбинирование результатов работы предлагаемых расширений. Конечной целью дополнений к платформе TXM является создание механизмов для исследования применимости различных дифференцирующих признаков при решении задачи классификации текстов и создания тематических корпусов текстов.

Дополнительную информацию о специфическом содержании текста можно почерпнуть, анализируя не только словоформы, но и именные группы и глагольные группы целиком. В отличие от отдельных слов, выделенные именные и глагольные группы несут информацию о конкретных отдельных аспектах содержания текста.

Именная группа определяется нами как группа слов, у которой главное слово существительное, а другие слова связаны с ним подчинительными синтаксическими связями. Определенную трудность при выделении именных групп представляет разрешение омонимической неопределенности, проистекающей из множественности морфологических разборов отдельных словоупотреблений, которая как правило имеет место быть. Наш метод выделения именных групп предполагает рассмотрение всего множества возможных морфологических разборов каждого слова. Используемый нами алгоритм подробно описан в [13]. Алгоритм состоит из трех этапов: установление подчинительных синтаксических связей в предложении между парами слов; установление синтаксических связей внутри конструкций с однородными членами; выделение именных групп как цепочки последовательно связанных подчинительными связями слов.

Для расширения характеристик текстов проводилось выделение глагольных групп (словосочетаний, главным словом которых является глагол), то есть установление связей выделенных именных групп с глаголами. Данная задача решается нами при помощи анализа глагольного управления. Каждому глаголу сопоставлен набор ограничений, накладываемых им на зависимые словосочетания. Такой набор есть парадигма глагольного управления, если мы отвлекаемся от самого глагола и рассматриваем набор его ограничений сам по себе

Результаты морфологического анализа и процедуры выделения именных групп позволяют, используя словарь глагольного управления, выявить синтаксические связи для определения глагольных групп. Выделение глагольных групп в предложении осуществляется путем анализа всех возможных пар (глагол, именная группа) предложения на предмет соответствия именной группы парадигме управления соответствующего глагола, а именно поиска в парадигме ограничения, которому удовлетворяет рассматриваемая именная группа. Для анализа глагольного управления был использован электронный словарь глагольного управления, в который вошли первые две тысячи наиболее частотных глаголов русского языка по материалам Национального корпуса русского языка (ruscorpora.ru). Словарь глагольного управления содержит парадигмы глагольного управления, состоящие из ограничений употреблений именных групп.

Описанные в данном разделе характеристики (словарные морфологические характеристики, псевдоосновы, именные группы, глагольные группы), а также двухбуквеные и трехбуквенные сочетания использовались нами как дополнительные признаки текстов на естественном языке. Все эти признаки (характеристики) использовались по аналогии с анализом словоформ в стандартном анализе на базе платформы TXM.

4. Корпусной анализ

Для анализа подкорпусов и их отношений между собой в TXM используется две характеристики: показатель специфичности и анализ соответствий. Эффективным и наглядным инструментом количественной оценки соотношений специальных подкорпусов относительно друг друга и всего корпуса в целом является показатель специфичности [19]. Анализ специфичности позволяет составить своего рода «профиль» подкорпуса, определенного на основании некоторого свойства (тематики текста, идеологической направленности текста) путем выявления наиболее характерных или нехарактерных для него признаков (лексем, псевдооснов, именных и глагольных групп и т.п.). Этот «профиль» может быть использован для диагностики нового текста. Другим подходом к анализу разделенного на части (подкорпуса) по определенному критерию корпуса является анализ соответствий. Методика анализа соответствий, используемая TXM, была предложена Ж.-П. Бензекри [14] и реализована в пакете FactoMineR [21]. Анализ соответствий демонстрирует взаимную «близость» или «удаленность» подкорпусов на основе анализа частот совместного появления значений переменных (словоформ, начальных форм, псевдооснов, именных групп, буквосочетаний).

Приведем примеры результатов исследований. На рисунках 1-3 приведены результаты анализа соответствий для трех признаков: «каноническая форма слова» (рис. 1), «четырехбуквенные сочетания» (рис. 2) и «псевдооснова» (рис. 3). Все графики приведены для второй и третей по статистической значимости величинам признаков.

Очевидно разделение в пространстве признаков подкорпусов по их тематическим направленностям для всех признаков. Видно не только разделение по тематике подкорпусов, но и их отделение от нейтрального корпуса. Эти данные указывают на хорошее разделение подкорпусов по тематикам, что указывает на корректность созданного корпуса и подтверждает возможность использовать данный корпус в качестве обучающей выборки для задач машинного обучения. Рисунки показывают, что средства анализа платформы TXM эффективны для анализа корпуса по самым различным характеристикам.

Приведенные примеры иллюстрируют такой важный результат, как возможность использования всех исследованных признаков для решения задач тематической классификации текстов. Если признать эквивалентность результатов на приведенных рисунках (рис. 1 - 3), то можно говорить о возможности решения многих задач тематической классификации без глубокого синтаксического и морфологического анализов. Эффективными могут быть такие признаки как «трехбуквенные сочетания» и «четырехбуквенные сочетания». Данные наблюдения формулируются только для задач тематической классификации специального типа текстов.

УДК 81’33: 519.76 DOI: 10.21681/2311-3456-2019-4-54-60

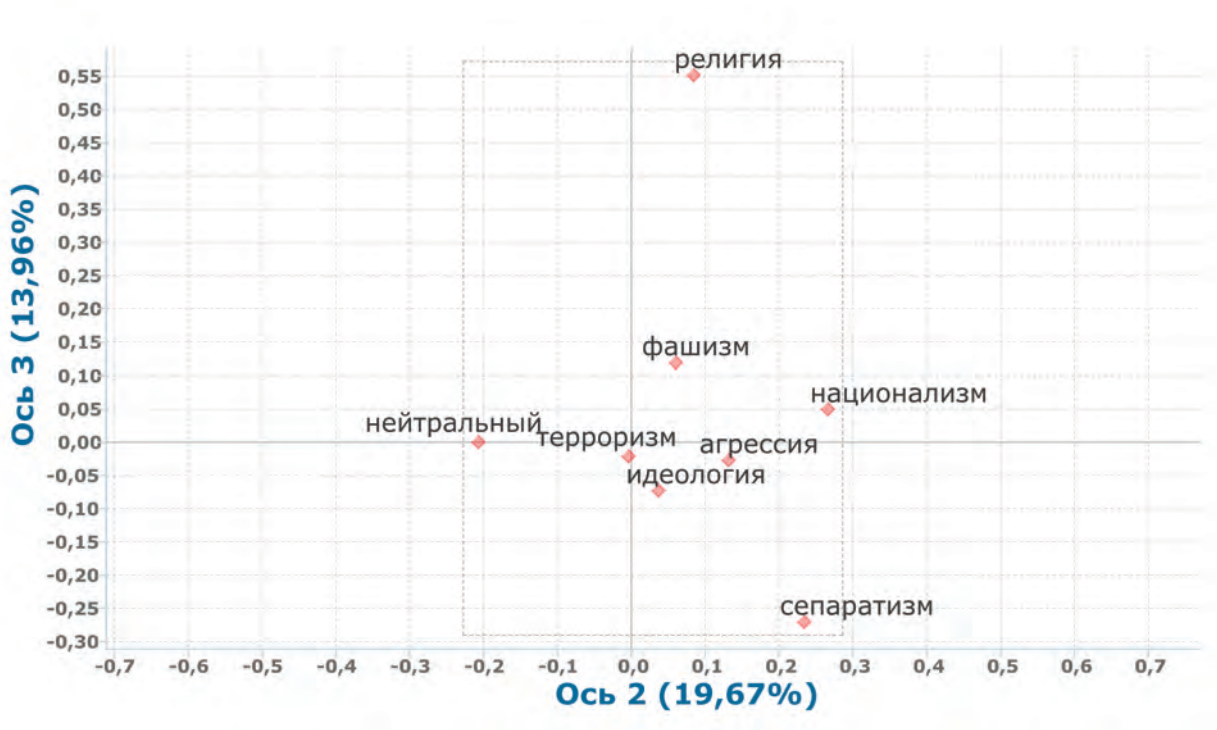

Рис. 1. Анализ соответствий для признаков «каноническая форма слова»

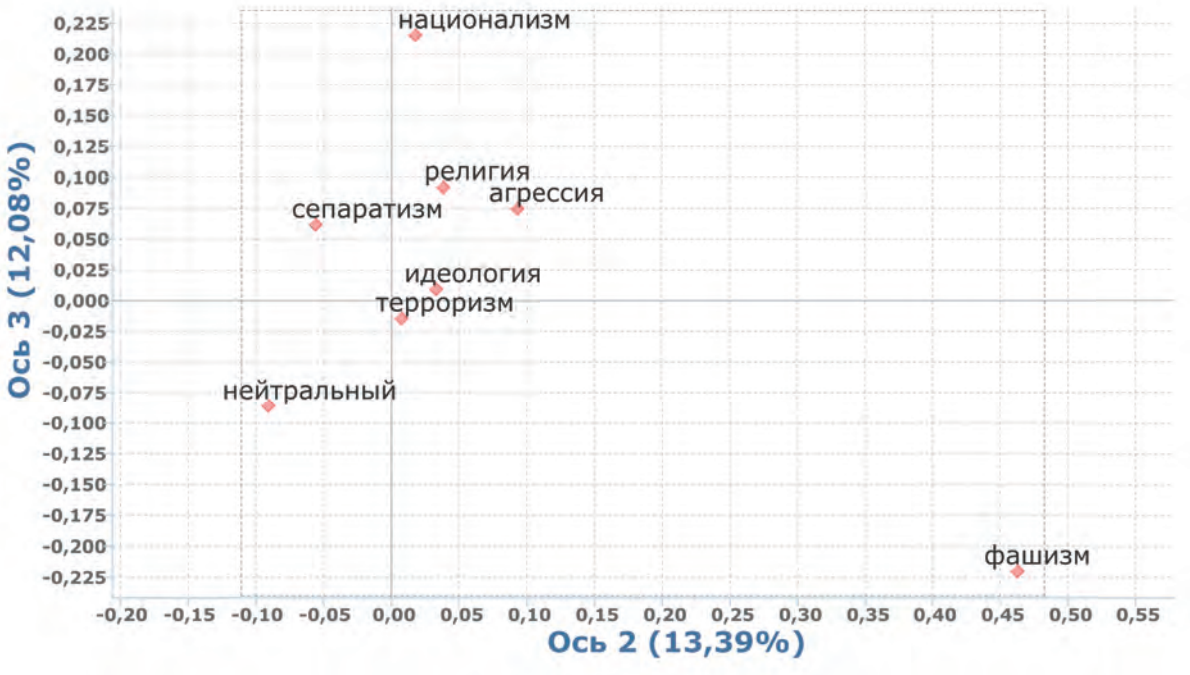

Рис. 2. Анализ соответствий для четырехбуквенных признаков

Рис. 3. Анализ соответствий для признаков «псевдооснова»

5. Выводы

Расширение платформы корпусного анализа ТХМ путем интеграции в нее инструментов автоматической обработки текста показало, что такая интеграция позволяет создать новые возможности статистического анализа текстов.

Исследования влияния различных дифференцирующих признаков и их комбинаций для тематической классификации подкорпусов текстов показывают, что наиболее универсальными мультиязычными дифференцирующими признаками являются трех и четырехбуквенные сочетания символов естественного языка. Псевдоосновы можно рассматривать как эффективные признаки при решении задач классификации текстовых массивов.

Выявленные особенности подкорпусов и противопоставленности нейтрального подкорпуса остальным демонстрируют то, что сформированный корпус может быть использован для машинного обучения в задачах классификации текстов на предмет выявления заданного содержания с целью их углубленного экспертного анализа.

Литература

1. 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.
2. 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.
3. Борисов С.В., Васнецова А.С., Жафяров А.Г. К вопросу о противодействии кибертерроризму и киберэкстремизму //Вестник Академии Генеральной прокуратуры Российской Федерации. – 2015. – Т. 45. – №. 1. – С. 49-55.
4. Вартан А.Ю. Классификация ресурсов из сети Интернет по направлениям наркоторговля, терроризм, экстремизм //Вестник Югорского государственного университета. – 2015. – №. S2 (37).
5. Злоказов К.В. Восприятие экстремистского текста субъектами с различным уровнем деструктивной установки // Политическая лингвистика. 2014. № 1 (47). С. 265-272.
6. Злоказов К.В., Софронова А.Ю. Образы коммуникаторов и стратегии воздействия при пропаганде идей террористической организации «исламское государство» // Политическая лингвистика. 2015. № 2 (52). С. 247-253.
7. Красняков Е.И., Машечкин И.В., Петровский М.И., Царев Д.В. Методы машинного обучения для обнаружения активности экстремистского характера в сети интернет // Тезисы докладов научной конференции «Ломоносовские чтения». 2017. С. 110-111.
8. 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.
9. 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.

10. Лапунова Ю.А., Голяндин Н.П. Распространение идеологии экстремизма и терроризма в киберпространстве: проблемы и пути их решения //Труды Академии управления МВД России. 2017. № 3.(43). С. 100 – 104.
11. Поляков И. В., Соловьев Ф. Н., Чеповский А. А., Чеповский А. М. Задача распознавания для текстов на естественных языках. М. : Национальный открытый университет «ИНТУИТ», 2017
12. 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.
13. Чеповский А. М. Информационные модели в задачах обработки текстов на естественных языках. Второе издание, переработанное. М.: Национальный открытый университет «ИНТУИТ», 2015.
14. Benzécri J.-P. L’analyse des données: l’analyse des correspondances. 2nd ed. Vol. 2. Paris: Dunod, 1979.
15. 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.
16. Egorova E., Chepovskiy A., Lavrentiev A. A structural pattern based method for automated morphological analysis of word forms in a natural language // Journal of Mathematical Sciences. Moscow: Plenum Publishers. 2016. Vol. 214. No. 6. P. 802-813.
17. 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.
18. Heiden S. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme // 24th Pacifc Asia Conference on Language, Information and Computation - PACLIC24 / Ed. R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto and Y. Harada. Institute for Digital Enhancement of Cognitive Development, Waseda University, Sendai, Japan. 2010 P. 389—398. URL: http://halshs.archiies-ouiertes.fr/halshs-00549764
19. Lafon P. Sur la variabilité de la fréquence des formes dans un corpus // Mots. 1980. № 1. P. 127-165.
20. 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. – С. 145-162.
21. Lê S., Josse J., & Husson F. FactoMineR: an R package for multivariate analysis // Journal of statistical software.2008. № 25 (1) P. 1-18.
22. Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. Manchester, UK. 1994 URL: http://www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf

ANALYSIS OF CORPUS OF EXTREMIST TEXTS AND UNLAWFUL TEXTS

Lavrentiev A.M.,[8], Smirnov I.V.[9], Suvorova M.I.[10], Solov’ev F.N.[11], Fokina A.I.[12] Chepovskiy A. M.[13]

The purpose of the study: development of a technique of creation and automatic analysis of special corpora for their subsequent application as the training datasets and detecting the differentiating characters in problems of text classification.

Method: tools of the analysis of the case TXM platform expanded with the developed procedures of calculation of additional characteristics of texts, such as combinations of letters, pseudo-bases, noun phrases, verb phrases were used.

Results: it is shown that the developed extenders of the case TXM platform allow to solve effectively problems of the analysis of texts of special subject, the created corpus of extremist subject can be used as the training selection for problems of classification of texts, the conclusion about use of combinations of letters as the universal differentiating characters along with classical linguistic characteristics of texts is drawn.

Keywords: corpus linguistics, automated text analysis, TXM platform, stemming, noun phrases, verbal dependencies, detecting extremist texts

8 Alexey Lavrentiev, Ph.D., IHRIM Research Lab, CNRS & ENS de Lyon, Lyon, France. E-mail: alexei.lavrentev@ens-lyon.fr

9 Ivan Smirnov, Ph.D., head of department Federal Research Center «Computer Science and Control», RAS, Moscow, Russia. E-mail: ivs@isa.ru

10 Fedor Solov’ev, research scientist, Institute of Physical and Technical Informatics, Protvino Moscow Region, Russia. E-mail: the0@yandex.ru

11 Suvorova M.I. – research scientist, Federal Research Center «Computer Science and Control» RAS, Moscow, Russia. E-mail: suvorova@isa.ru.

12 Alina Fokina, bachelor, National Research University Higher School of Economics, Moscow, Russia. E mail: aifokina@edu.hse.ru

13 Andrey Chepovskiy, Dr. Sc. (Eng.), Professor, National Research University Higher School of Economics, Moscow, Russia. E-mail: achepovskiy@hse.ru

References

1. 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.
2. Ananyeva M.I., Kobozeva M.V., Polyakov I. V., Solov›yev F. N., Chepovskiy A.M. O probleme viyavleniya extremistskoy napravlennosti v tekstakh // Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnyye tekhnologii. 2016. T. 14. № 4. C. 5-13.
3. Borisov S.V., Vasnetsova A.S., Zzafyarov A.G. K voprosu o protivodeystvii kiberterrorizmu i kiberekstremizmu //Vestnik Akademii General›noy prokuratury Rossiyskoy Federatsii. – 2015. – T. 45. – №. 1. – S. 49-55.
4. Vartan A.Yu. Klassifikatsiya resursov iz seti Internet po napravleniyam narkotorgovlya, terrorizm, ekstremizm //Vestnik Yugorskogo gosudarstvennogo universiteta. – 2015. – №. S2 (37).
5. Zlokazov K.V. Vospriyatie ekstremistskogo teksta sub»ektami s razlichnym urovnem destruktivnoy ustanovki // Politicheskaya lingvistika. 2014. № 1 (47). S. 265-272.
6. Zlokazov K.V., Sofronova A.Yu. Obrazy kommunikatorov i strategii vozdeystviya pri propagande idey terroristicheskoy organizatsii «islamskoe gosudarstvo» // Politicheskaya lingvistika. 2015. № 2 (52). S. 247-253.
7. Krasnyakov E.I., Mashechkin I.V., Petrovskiy M.I., Tsarev D.V. Metody mashinnogo obucheniya dlya obnaruzheniya aktivnosti ekstremistskogo haraktera v seti internet // Tezisy dokladov nauchnoy konferentsii «Lomonosovskie chteniya». 2017. S. 110-111.
8. Lavrentiev 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.
9. Lavrentiev 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.
10. Lapunova U.A., Golyandin N.P. Rasprostranenie ideologii ekstremizma i terrorizma v kiberprostranstve: problemy i puti ih resheniya // Trudy Akademii upravleniya MVD Rossii. 2017. № 3.(43). S. 100 – 104.
11. Polyakov I. V., Solov'yev F. N., Chepovskiy A. A., Chepovskiy A. M. Zadacha raspoznavaniya dlya tekstov na yestestvennyh yazykah. M. : Natsional'nyy otkrytyy universitet «INTUIT», 2017
12. Polyakov I. V., Sokolova T. V., Chepovskiy A. A., Chepovskiy A. M. Problema klassifikatsii tekstov i differentsiruyushchiye priznaki // Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnyye tekhnologii. 2015 T. 13 № 2 S. 55-63.
13. Chepovskiy A. M. Informatsionnyye modeli v zadachah obrabotki tekstov na yestestvennyh yazykah. Vtoroye izdaniye, pererabotannoye. M.: Natsional›nyy otkrytyy universitet "INTUIT", 2015.
14. Benzécri J.-P. L'analyse des données: l'analyse des Correspondances. 2nd ed. Vol. 2. Paris: Dunod, 1979.
15. 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.
16. Egorova E., Chepovskiy A., Lavrentiev A. A structural pattern based method for automated morphological analysis of word forms in a natural language // Journal of Mathematical Sciences. Moscow: Plenum Publishers. 2016. Vol. 214. No. 6. P. 802-813.
17. 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.
18. Heiden S. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme // 24th Pacific Asia Conference on Language, Information and Computation - PACLIC24 / Ed. R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto and Y. Harada. Institute for Digital Enhancement of Cognitive Development, Waseda University, Sendai, Japan. 2010 P. 389—398. URL: http://halshs.archiies-ouiertes.fr/halshs-00549764
19. Lafon P. Sur la variabilité de la fréquence des formes dans un corpus // Mots. 1980. № 1. P. 127-165.
20. 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. – C. 145-162.
21. Lê S., Josse J., & Husson F. FactoMineR: an R package for multivariate analysis // Journal of statistical software.2008. № 25 (1) P. 1-18.
22. Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. Manchester, UK. 1994 URL: http://www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf