В связи с развитием средств электронных коммуникаций актуальны разработки методов противодействия распространению в сети Интернет противоправных текстов. Требуется анализировать методы пропаганды и вербовки, используемые противниками государства в отношении молодежи, выявлять коммуникативные стратегии информационного противоборства. Для решения сформулированных задач необходимо разрабатывать методики автоматического обнаружения противоправных текстов как больших объемов, так и коротких сообщений и комментариев.

Используемые методы анализа текстов обычно основаны на выделении ключевых слов из образца текста. В данном исследовании используются различные расширения стандартной корпусной платформы для изучения специализированных корпусов текстов. Исследованы различных лингвистических признаков, которые оказывают влияние на показатели классификации противоправных текстов с применением методов анализа корпусов текстов.

Использование методов корпусного анализа для выявления дифференцирующих признаков текстов на естественных языках ранее не применялось для выявления противоправных текстов.

Целью исследования является 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. Применены инструменты анализа корпусной платформы TXM, расширенные процедурами вычисления дополнительных характеристик текстов, таких как буквосочетания, псевдоосновы, именные группы, глагольные группы.

Основной задачей является разработка средств расширения корпусной платформы TXM, которые позволят эффективно решать задачи анализа текстов противоправной тематики, создавать и анализировать корпуса текстов экстремистской тематики, делать выводы об использовании в качестве дифференцирующих признаков различных лингвистических характеристик текстов. Данный подход позволяет анализом текстовых конструкций повышать полноту и гибкость корпусного анализа.

Для исследования различных наборов текстов (подкорпусов), их отношений между собой используются такие характеристики как специфичность и анализ соответствий, которые являются эффективными инструментами количественной оценки соотношений специальных подкорпусов относительно друг друга и всего корпуса в целом. 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. Этот «профиль» может быть использован для диагностики нового текста.

Другим подходом к анализу разделенного на части (подкорпуса) по определенному критерию корпуса является анализ соответствий. Анализ соответствий используется для анализа взаимной «близости» или «удаленности» подкорпусов на основе анализа частот совместного появления значений переменных (словоформ, начальных форм, псевдооснов, именных групп, буквосочетаний). На основе анализа соответствий также разработана методика выделения списка конкретных переменных, характерных для конкретной тематики.

Определено разделение в пространстве признаков подкорпусов по их тематическим направленностям для выявления дифференцирующих признаков. Данный подход позволяет создать методику разделения подкорпусов по тематикам и создавать обучающие выборки для задач машинного обучения.