УДК 004.056

РАЗРАБОТКА МОДЕЛИ СИСТЕМЫ АВТОМАТИЗИРОВАННОЙ ПРОВЕРКИ ТЕКСТОВ НА ПРЕДМЕТ СОДЕРЖАНИЯ ИНФОРМАЦИИ ОГРАНИЧЕННОГО ДОСТУПА

Меренков Д.Н. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – кандидат технических наук, доцент ФБИТ Коржук В.М.
(федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В данной работе рассматривается разработка модели системы автоматизированной проверки текстов на предмет содержания информации ограниченного доступа. Результаты исследования показывают, что предлагаемое решение, основанное на использовании предварительно обученной модели BERT для определения класса подаваемых на вход текстовых данных, может эффективно выполнять задачу классификации и демонстрируют потенциальные возможности практического применения в областях, где требуется проведение идентификационной экспертизы.

Введение.

Потребность в автоматизированной проверке текстов на предмет содержания информации ограниченного доступа быстро растет в современном обществе, особенно среди физических лиц и организаций, попадающих под действие законодательства Российской Федерации в области экспортного контроля.

Существующие подходы к решению проблемы обнаружения информации ограниченного доступа в текстовых массивах данных, в основном базируются на «ручном» анализе и проверках, которые могут требовать большого количества временных затрат, быть трудоемкими и подверженными влиянию человеческого фактора. Однако, достижения в области машинного обучения и обработки естественного языка позволяют разработать более эффективное решение этой проблемы.

Целью данной работы является разработка модели системы автоматизированной проверки текстов на предмет содержания информации ограниченного доступа для уменьшения финансовых издержек, а также для увеличения точности распознавания такой информации.

Основная часть.

Разрабатываемая модель системы предполагает подход, который можно разделить на четыре основных этапа: подготовка набора данных, предварительная обработка входного корпуса текстовых данных, обучение модели и определение метки класса входного текста.

Первый этап – подготовка данных, включает в себя сбор и маркировку некоторого определенного объема текстовых данных, которые будут использоваться для обучения модели машинного обучения. Второй этап заключается в предварительной обработке данных. Данные очищаются и преобразуются таким образом, чтобы они могли быть обработаны моделью машинного обучения. Третий этап, обучение модели, включает в себя использование предварительно обученной модели BERT и ее точную настройку на основе помеченных текстовых данных. Процесс тонкой настройки включает в себя добавление слоя классификатора к модели и обучение его на основе маркированных данных. На четвертом

этапе используется обученная модель машинного обучения для определения одного из семи классов, к которому относятся текстовые данные.

В дополнение к четырем основным этапам предлагаемое решение также имеет два дополнительных шага: выделение ключевых слов в тексте на основе существующих словарей и выбор наиболее компетентного эксперта для предоставления отчета. Первый шаг подразумевает использование существующих словарей ключевых слов для определения и выделения слов в тексте, которые имеют отношение к классу информации. Второй шаг включает в себя использование результатов классификации для определения наиболее компетентного эксперта с целью проведения экспертизы, опираясь на предоставленный отчет, и вынесения заключительного решения.

Выводы.

Эффективность разрабатываемой модели зависит от ряда факторов, включая качество и размер набора данных, используемого для обучения, неоднозначность входных данных, а также размер и качество предварительно обученной модели.

Предлагаемое решение использует предварительно обученную модель BERT для классификации текстовых данных по семи различным классам, включая товары и технологии двойного назначения, ядерные материалы и химические вещества, которые могут быть использованы при создании оружия. Модель системы была протестирована и сравнена с другими моделями машинного обучения, такими как одномерная сеть CNN, сеть LSTM и RNN сеть. Было установлено, что BERT обладает наилучшей производительностью.

Автоматизация процесса проведения идентификационной экспертизы с целью определения наличия информации ограниченного доступа потенциально может сократить временные и финансовые затраты, необходимые для проведения анализа в «ручном» режиме, повысить точность и надежность результатов, а также стать ценным инструментом для организаций и физических лиц, попадающих под действие экспортного контроля.

Меренков Д.Н. (автор)

Коржук В.М. (научный руководитель)