УДК 347

О. Б. Березина[1, 2]*

Искусственный интеллект и противодействие экстремизму

[1] Новосибирский государственный технический университет, г. Новосибирск, Российская Федерация
[2] Сибирский государственный университет геосистем и технологий, г. Новосибирск, Российская Федерация
* e-mail: obberezina@mail.ru

Аннотация. В статье рассматривается вопрос использования искусственного интеллекта в качестве средства противодействия экстремизму. Демонстрируются возможности использования современных технологий для оценки высказываний экстремистской направленности.

Ключевые слова: искусственный интеллект, нейронные сети, экстремистские материалы, профилактика экстремизма

O. B. Berezina[1, 2]*

Artificial intelligence and extremism counteraction

[1] Novosibirsk State Technical University, Novosibirsk, Russian Federation
[2] Siberian State University of Geosystems and Technologies, Novosibirsk, Russian Federation
* e-mail: obberezina@mail.ru

Abstract. This paper reviews the application of artificial intelligence to the extremism counteraction and demonstrates the possibilities of modern technologies in detecting extremist statements in a text.

Keywords: artificial intelligence, neural networks, extremist materials, prevention of extremism

Введение

Обеспечение общественной безопасности традиционно определяется как основополагающая цель правоохранительной деятельности государства, вследствие чего активно создается постоянная потребность в формировании новых, отвечающих требованиям актуальности и обладающих высокой результативностью способов противодействия противоправным проявлениям вообще и экстремистской деятельности в частности. Так, федеральный закон от 25 июля 2002 г. N 114-ФЗ «О противодействии экстремистской деятельности» в качестве направлений противодействия экстремистской деятельности устанавливает выявление, предупреждение и пресечение экстремистской деятельности общественных и религиозных объединений, иных организаций, физических лиц [1].

Таким образом, большое значение в современных реалиях приобретают разнообразные способы детекции проявлений экстремизма в окружающей действительности, среди которых наиболее актуальным выступает использование технологий искусственного интеллекта для определения материалов экстремистской направленности, размещенных в сети Интернет.

Результаты

Искусственный интеллект является продуктом технологического развития XX века, когда объединение различных научных направлений (таких как нейрофизиология, математика, кибернетика и др.) определило возникновение совершенно нового подхода в информационных технологиях. Идея о том, что искусственный интеллект может помочь автоматизировать процесс принятия решений и кратно снизить нагрузку на людей, сформировала абсолютно новый подход к задаче выявления потенциально опасной информации. В целом, применение искусственного интеллекта в борьбе с противоправными проявлениями реализуется в различных направлениях, таких как:

– осуществление идентификации личности;
– установление авторства [3];
– детекция компьютерных вирусов и вредоносных программ;
– установление первоисточника информации в сети Интернет;
– анализ деятельности преступных сообществ;
– отслеживания интернет трафика и т.п.

Но в контексте противодействия экстремизму, наиболее эффективным представляется использование искусственного интеллекта для автоматизации процессов распознавания объектов, особенно в сфере лингвистики. Благодаря быстрой обработке данных на основе специально подобранных алгоритмов, складываются широкие возможности для формирования интеллектуальных систем информационной безопасности, способных обнаруживать и определять различного рода угрозы, в том числе и экстремистского толка.

В соответствии со статьей 1 вышеуказанного закона экстремистские материалы – это предназначенные для распространения либо публичного демонстрирования документы, либо информация на иных носителях, призывающие к осуществлению экстремистской деятельности либо обосновывающие, или оправдывающие необходимость осуществления такой деятельности, в том числе труды руководителей национал-социалистической рабочей партии Германии, фашистской партии Италии, выступления, изображения руководителей групп, организаций или движений, признанных преступными в соответствии с приговором Международного военного трибунала для суда и наказания главных военных преступников европейских стран оси (Нюрнбергского трибунала), выступления, изображения руководителей организаций, сотрудничавших с указанными группами, организациями или движениями, публикации, обосновывающие или оправдывающие национальное и (или) расовое превосходство либо оправдывающие практику совершения военных или иных преступлений, направленных на полное или частичное уничтожение какой-либо этнической, социальной, расовой, национальной или религиозной группы [2]. Таким образом, закладываются рамочные характеристики оценки информационных материалов на предмет признания их экстремистскими. Что позволяет использовать искусственный интеллект как инструмент экспертной оценки текстовых материалов на предмет отнесения их к экстремистским.

Система обучения искусственного интеллекта совершенствуется быстрыми темпами, что обуславливает высокую вероятность правильной оценки информационной направленности конкретных материалов. Если раньше для каждого слова была возможна оценка только по частоте употребления [4], то после 2013 года все большее значение приобретает идея контекстной оценки слова, которая позволяет выделять его лексические значения по его употребляемости в различных предложениях [5]. «Похожие слова употребляются в похожем контексте» [6] – концепция, которая лежит в основе обучения современных искусственных нейронных сетей, что позволяет достигать контекстного понимания целых предложений или даже абзацев текста. Так, используя так называемый «двунаправленный механизм внимания», нейросеть может произвести актуализацию смысла слова (или предложения) и определить его тональность в каждом конкретном случае [7]. Например, значение фразы «Убить их всех» не воспринимается нейросетью само по себе как однозначное проявление ненависти, поскольку его смысл может актуализироваться и иначе, в контексте сообщения садовода-любителя, столкнувшегося с нашествием муравьев и интересующегося, каким способом он мог бы от них избавиться. Подобное контекстное понимание всего высказывания, каким бы большим по объему оно не было, является выгодным преимуществом современных текстовых нейросетевых моделей.

Актуальным примером применения таких технологий стало сообщение ГРЧЦ РФ о тестировании систем «Вепрь» и «Окулус», целью работы которых является обнаружение потенциальных угроз в интернете. Созданное решение выявляет нарушения законодательства РФ не только в текстах, но и в изображениях, и видеоматериалах. Кроме того, эти алгоритмы также могут использоваться еще и для прогнозирования последующего распространения деструктивных материалов [8].

Заключение

Таким образом, использование искусственного интеллекта как механизма противодействия экстремизму в современных условиях информационного общества выступает одной из наиболее актуальных и эффективных мер. Представляется, что при условии сохранения возможности принятия итоговых юридически значимых решений в вопросе отнесения материалов к экстремистским за правоприменителем, безусловно формируются широкие перспективы применения искусственного интеллекта как технического способа определения материалов экстремистского содержания и действий экстремистской направленности. Выделяя именно своеобразную сигнальную функцию искусственного интеллекта, когда он практически безошибочно указывает на конкретные материалы, которые могут быть экстремистски ориентированы, можно говорить о наличии эффективных механизмов противодействия экстремистским проявлениям в обществе, связанным с использованием сети Интернет.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. О противодействии экстремистской деятельности: Федер. закон Рос. Федерации от 25 июля 2002 г. N 114-ФЗ: принят Гос. Думой Федер. Собр. Рос. Федерации 27 июня 2002 г.: одобр. Советом Федерации Федер. Собр. Рос. Федерации 10 июля 2002 г. // Рос. газ. - 2002. - 30 июля, N 138-139.

2. О противодействии экстремистской деятельности: Федер. закон Рос. Федерации от 25 июля 2002 г. N 114-ФЗ: принят Гос. Думой Федер. Собр. Рос. Федерации 27 июня 2002 г.: одобр. Советом Федерации Федер. Собр. Рос. Федерации 10 июля 2002 г. // Рос. газ. - 2002. - 30 июля, N 138-139. Ст. 1.

3. Батура Т. В. Формальные методы определения авторства текстов // Вестник НГУ. Серия: Информационные технологии. 2012. № 4. URL: https://cyberleninka.ru/article/n/formalnye-metody-opredeleniya-avtorstva-tekstov (дата обращения: 01.03.2023).

4. Лукашевич Н.В., Четверкин И.И. Извлечение и использование оценочных слов в задаче классификации отзывов на три класса // Вычислительные методы и программирование. 2011. С. 73–81.

5. Mikolov, T., Sutskever, I., Chen, K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. 2013. № 26.

6. Ферс Дж. Техника семантики // Новое в лингвистике. 1962. № 2.

7. Лукашевич Н.В. Автоматический анализ тональности текстов. Проблемы и методы // Интеллектуальные системы. Теория и приложения. 2022. Т. 26 N 1. URL: http://intsysjournal.org/pdfs/26-1/1-5-Lukashevich.pdf (дата обращения: 01.03.2023).

8. Систему "Вепрь" для выявления угроз в интернете запустят во второй половине 2023 года // ТАСС. URL: https://tass.ru/obschestvo/17091419 (дата обращения: 01.03.2023).