Швехтеров Максим Романович, Студент,
Поволжский государственный университет
телекоммуникаций и информатики

Секлетова Наталья Николаевна,
Кандидат педагогических наук, доцент,
Поволжский государственный университет
телекоммуникаций и информатики

СОВРЕМЕННЫЕ МЕТОДЫ СИНТЕЗА ТЕКСТОВЫХ ДАННЫХ: ОТ ШАБЛОНОВ К ГЕНЕРАТИВНЫМ МОДЕЛЯМ И ИХ ПРИМЕНЕНИЕ В ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Аннотация. В статье рассматривается эволюция методов синтеза данных в NLP: от шаблонных алгоритмов до современных нейросетей (GPT). Проводится анализ их эффективности и роли в автоматизации создания контента. Особое внимание уделено практическому применению генеративных моделей для повышения информационной безопасности, создания надежных паролей и персонализации пользовательского опыта.

Ключевые слова: Синтез данных, NLP, нейронные сети, генеративные модели, информационная безопасность, персонализация.

Обработка естественного языка (Natural Language Processing, NLP) за последнее десятилетие претерпела фундаментальные изменения, перейдя от жестких правил к адаптивным нейросетевым моделям. Одной из наиболее актуальных задач в этой области является синтез данных – автоматическая генерация текстового контента, сохраняющего семантическую и синтаксическую структуру естественного языка. Данная технология находит широкое применение не только в диалоговых системах, но и в сфере информационной безопасности [1].

Традиционные подходы к синтезу текста базировались на правилах и статистических методах. Как отмечают С. Бёрд и соавторы, классические инструменты, такие как NLTK, позволяли эффективно работать с частотным анализом и n-граммами [2]. Однако генерация на основе n-грамм обладала существенным недостатком: отсутствием долговременной памяти. Модель могла предсказать следующее слово, но теряла контекст предложения, что делало синтезированный текст бессвязным на длинных дистанциях. Р. Риз также подчеркивает, что подходы, основанные на жестких правилах, требовали колоссальных трудозатрат на создание шаблонов и плохо масштабировались [3].

Революция в синтезе данных произошла с внедрением глубокого обучения. Й. Гольдберг в своей работе указывает, что переход к векторным представлениям слов (embeddings) позволил моделям улавливать семантическую близость понятий, а не просто их статистическое соседство [4]. Дальнейшее развитие привело к появлению рекуррентных нейронных сетей (RNN) и архитектур LSTM, которые, как описывают Б. Макмахан и Д. Рао, смогли удерживать контекст значительно лучше статистических предшественников [5].

Современным стандартом в NLP стали трансформеры (Transformers). Д. Ротман отмечает, что механизмы внимания (Self-Attention) в таких моделях, как BERT и GPT, позволяют учитывать зависимость каждого слова от всех остальных слов в тексте одновременно. Это обеспечило беспрецедентное качество синтеза данных, неотличимых от написанных человеком [6].

В контексте информационной безопасности эти достижения открывают новые возможности. Во-первых, это создание синтетических датасетов. Алекс Томас в работе по Spark NLP указывает на проблему нехватки размеченных данных для обучения систем обнаружения угроз [7]. Синтез позволяет генерировать бесконечное количество примеров «вредоносных» текстов (фишинг, спам) для тренировки классификаторов без риска утечки реальных пользовательских данных.

Во-вторых, генеративные модели применяются для повышения стойкости аутентификации. Вместо случайного набора символов нейросети могут генерировать мнемонические парольные фразы, которые легко запомнить пользователю, но сложно подобрать методом перебора, так как они имеют высокую энтропию.

В-третьих, методы NLP, описанные Д. Джурафски и Дж. Мартином, используются для анализа аномалий. Языковые модели обучаются на нормальном поведении пользователей (логах, сообщениях) и способны с высокой точностью выявлять девиации, свидетельствующие об инсайдерских угрозах или компрометации учетных записей [8].

Таким образом, интеграция трансформерных архитектур в процессы синтеза и анализа данных становится ключевым фактором развития проактивных систем киберзащиты.

Список литературы:

1. Лейн Х., Хапке Х., Ховард К. Обработка естественного языка в действии. – СПб.: Питер, 2020.
2. Bird S., Klein E., Loper E. Natural Language Processing with Python. – O'Reilly Media, 2009.
3. Риз Р. Обработка естественного языка на Java. – М.: ДМК Пресс, 2016.
4. Гольдберг Й. Нейросетевые методы в обработке естественного языка. – М.: ДМК Пресс, 2019.
5. Макмахан Б., Рао Д. Знакомство с PyTorch: глубокое обучение и обработка естественного языка. – М.: ДМК Пресс, 2020.
6. Rothman D. Transformers for Natural Language Processing. – Packt Publishing, 2021.
7. Thomas A. Natural Language Processing with Spark NLP. – O'Reilly Media, 2020.
8. Jurafsky D., Martin J.H. Speech and Language Processing. – 3rd ed. draft, 2020.