УДК 004

Мартынов Д.В.
Черников В.С
студенты 4 курс
Институт инженерных и цифровых технологий БелГУ
Белгородский государственный национальный исследовательский университет
Россия, г. Белгород

ЗАЩИТА ЛИЧНОЙ ИНФОРМАЦИИ В ЧАТАХ С НЕЙРОННЫМИ СЕТЯМИ: АНАЛИЗ ПРОБЛЕМ И МЕТОДОВ ЗАЩИТЫ ПРИ ХРАНЕНИИ ИНФОРМАЦИИ ПОЛЬЗОВАТЕЛЕЙ
PROTECTION OF PERSONAL INFORMATION IN CHATS WITH NEURAL NETWORKS: ANALYSIS OF PROBLEMS AND METHODS OF PROTECTION WHEN STORING USER INFORMATION

В данной статье проводится анализ уязвимостей, связанных с хранением конфиденциальной информации в чатах с нейросетями и рассматриваются возможные методы защиты данных пользователей. В анализе приведены примеры потенциальных сценариев утечки информации. Рассматриваются также и возможности удаления информации из чатов с нейросетями. Результаты анализа подчеркивают важность дальнейших исследований в области безопасности данных в подобных чатах и необходимость принятия эффективных мер для обеспечения конфиденциальности информации.

Ключевые слова: Нейронные сети, глубокое обучение, конфиденциальность данных, защита персональной информации, уязвимости нейронных сетей, обработка текстовой информации, безопасность в чат-ботах, социальная инженерия.

This article analyzes the vulnerabilities associated with storing confidential information in chats with neural networks and discusses possible methods for protecting user data. The analysis provides examples of potential information leakage scenarios. The possibilities of deleting information from chats with neural

networks are also being considered. The results of the analysis highlight the importance of further research into data security in such chats and the need to take effective measures to ensure the confidentiality of information.

Keywords: Neural networks, deep learning, data confidentiality, protection of personal information, vulnerabilities of neural networks, text information processing, security in chatbots, social engineering.

В условиях стремительного развития цифровых технологий пользователи всё чаще пользуются помощью нейронных сетей для решения своих профессиональных и повседневных задач. Одной из наиболее популярных нейросетей является ChatGPT, разработанная компанией OpenAI, которая совсем недавно добавила в неё возможность архивировать чаты. Со слов американского инженера Тибора Блахо чат-бот научится запоминать всю информацию о пользователе, которую он когда-либо ему сообщил. Например, если написать нейросети, что через месяц нужно будет сдавать экзамен по определённой дисциплине, она запомнит эту информацию и будет ссылаться на неё в дальнейшем. Данное исследование направлено на изучение уязвимостей, связанных с хранением персональной информацией нейросетями, и проработку стратегий по защите конфиденциальных данных.

Обзор литературы:

1. "Deep Learning" от Goodfellow, Bengio и Courville

Содержание: книга представляет собой обширное руководство по глубокому обучению и искусственным нейронным сетям. Она охватывает теоретические основы и практические аспекты глубокого обучения, включая секции, посвященные хранению и обработке данных.

- Плюсы: широкий обзор методов глубокого обучения и понимание основных принципов работы нейросетей.
- Минусы: отсутствие фокуса именно на проблемах безопасности и уязвимостях хранения данных в нейронных сетях.

2. "Adversarial Machine Learning" от David C. Montana

Содержание: книга обсуждает техники атак на машинное обучение, включая атаки на нейронные сети. Освещает уязвимости, связанные с обманом и взломом нейросетей, что может быть важным для понимания проблем безопасности в них.

- Плюсы: обзор методов атак и защиты, специфических для машинного обучения.
- Минусы: может быть технически сложной для понимания без предварительных знаний в области машинного обучения и криптографии.

3. "Privacy-Preserving Machine Learning" от Shokri и Shmatikov

Содержание: книга ориентирована на проблемы приватности и безопасности в машинном обучении. Освещает техники и методы для обеспечения конфиденциальности данных в контексте машинного обучения, включая нейросети.

- Плюсы: обзор методов защиты данных и понимание важности конфиденциальности в машинном обучении.
- Минусы: может быть более фокусирована на методах защиты данных, чем на конкретных уязвимостях нейросетей.

Методология

Анализ потенциальных сценариев утечки информации:

• Исследование возможности доступа злоумышленников к аккаунту пользователя и последующее взаимодействие с нейросетью для извлечения конфиденциальной информации.

• Рассмотрение сценариев, в которых злоумышленники могут попытаться манипулировать запросами к нейросети для получения конфиденциальной информации.

Возможности удалить информацию из нейросети:

• Изучение механизмов удаления или забывания информации в нейросетях и в ChatGPT в частности.

• Анализ возможности полного удаления конфиденциальных данных после запроса пользователя и обсуждение реальной эффективности таких механизмов.

Разработка стратегий защиты:

• Разработка методов и стратегий для защиты данных в случае возможного вмешательства злоумышленников.

Выводы:

• Формирование выводов относительно возможности и эффективности защиты данных в нейросети ChatGPT от манипуляций злоумышленников и предложение путей улучшения безопасности хранения информации.

1) Анализ потенциальных сценариев утечки информации

Злоумышленники могут различными методами получить доступ к аккаунту пользователя, содержащему диалоги с различными чат-ботами и, соответственно, полезной конфиденциальной информацией.

Основными способами получения несанкционированного доступа являются:

• Фишинг и социальная инженерия: может быть осуществлена атака с использованием различного вредоносного ПО или проведение манипуляций в рамках задействования социальной инженерии для получения учётных данных пользователей.

• Эксплуатация уязвимостей безопасности: могут использоваться уязвимости в ОС устройств, хранящих учётные данные. Также доступ злоумышленникам «подарят» вовсе ненастроенные методы дополнительной защиты, например, с помощью биометрии и слабые пароли.

• Внедрение в обмен сообщениями: злоумышленники могут перехватывать информацию, которой пользователь обменивается с нейросетью.

• Использование украденных сеансов: одним из способов входа в нужную учётную запись является подмена cookie-файлов. При такой манипуляции система будет думать, что пользователь «уже авторизован» и предоставлять доступ к нужному функционалу.

Рассмотренные выше сценарии наиболее опасны в совокупности, например: злоумышленник может с помощью социальной инженерии втереться в доверие, заставить с помощью манипуляций запустить вредоносное ПО и далее использовать прочие способы кражи данных в зависимости от ситуации.

2) Возможности удалить информацию из нейросети

Нейросеть не представляет собой, например, жёсткий диск, информацию с которого можно удалить без возможности восстановления, затерев всё нулями. Из-за особенностей их архитектуры и работы данная задача является не такой простой.

Увы, подробной информации о механизмах «забывания» конфиденциальных данных разработчики того же ChatGPT не предоставили, исходя из этого помочь в решении вопроса могут следующие действия:

• Удаление истории чата: это один из наиболее очевидных способов удаления информации. Пользователь может запросить удаление всей истории чата из базы данных нейросети. Стоит учитывать, что удаление часто не означает полное избавление от информации, так как резервные копии или анонимизированные данные могут остаться сохранёнными. Однако этот способ поможет усложнить доступ злоумышленника к важным данным.

• Удаление аккаунта: нет аккаунта – нет проблем. Радикальный способ, но надёжный, как швейцарские часы.

Можно сказать, что оба способа довольно радикальные, но даже они не являются решением вопроса. Стоит помнить, что еще задолго до масштабного распространения нейросетей всё, что попадало в интернет, оставалось там навсегда.

3) Разработка стратегий защиты

Для пользователя существует ряд средств и дополнительных мер, которые могут помочь в защите его данных и предотвращении несанкционированного доступа к аккаунту и информации, в том числе и в контексте взаимодействия с нейросетями:

- Сильные пароли и механизм двухфакторной аутентификации (2FA): использование длинных и сложных паролей для аккаунтов вместе с механизмами двухфакторной аутентификации (например, коды, отправляемые на устройство пользователя) существенно повышает безопасность. Регулярное изменение паролей и ограничение доступа к аккаунту через управление правами доступа также помогают минимизировать риск утечки данных.
- Обучение пользователей и осведомленность о безопасности: предупреждён – значит вооружён. Знания и ответственный подход помогут защитить данные в первую очередь. Если пользователь осведомлён о многообразии способов защиты, он будет использовать как можно большее их количество для предотвращения утечек своих данных.
- Проверка настроек конфиденциальности и безопасности: регулярная проверка и обновление настроек конфиденциальности и безопасности аккаунта в системе, включая ограничение доступа к определенным данным или функциям, может усилить защиту.
- Избегание потенциально вредных ресурсов и бдительность: при использовании интернета потенциально опасные сайты заметно выделяются на фоне большинства обычных. Достаточно лишь немного внимательности и критического мышления при использовании того или иного ресурса, и даже до антивируса дело не дойдет. Бдительность также поможет предотвратить возможные атаки с помощью социальной инженерии.

Выводы:

- Безопасность данных в нейросетях, таких как ChatGPT, представляет сложную задачу из-за особенностей их работы и архитектуры.

- Механизмы "забывания" информации в нейросетях не раскрыты и точно не так просты, как удаление файла из обычного хранилища.
- Пользователи должны использовать множество методов для сохранения безопасности своих данных, однако полная гарантия защиты не может быть обеспечена из-за специфики работы нейросетей.

В целом, обеспечение безопасности данных в нейросетях представляет сложную задачу, и эффективные методы защиты требуют как технических решений от разработчиков, так и осознанного подхода со стороны пользователей при обращении с данными и взаимодействии с подобными сетями.

СПИСОК ЛИТЕРАТУРЫ

1. Соколов, М.М. Нейронные сети и их применение в кибербезопасности: Сборник научных статей. - Москва: Издательство "Наука", 2019. - 300 с.
2. Д. Элдер, Д. Джонсон, "Машинное обучение для хакеров". – Санкт-Петербург: Издательство "Прогресс", 2020. - 180 с.
3. Козлов, Н.Н. Методы обнаружения атак в сетях на основе искусственных нейронных сетей. - Москва: Издательский дом "Кодекс", 2018. - 250 с.
4. К. Мерфи, "Введение в вероятность и машинное обучение",. - Санкт-Петербург: Издательство "БХВ-Петербург", 2017. - 220 с.
5. Новиков, В.В. Применение сверточных нейронных сетей для обнаружения вторжений в компьютерные сети. - Москва: Издательство "Лань", 2016. - 280 с.