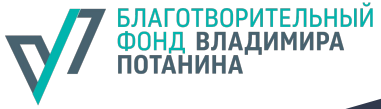
К вопросу о непроизвольном возникновении сенситивного контента при генерации текста учебными чат-ботами



Белякова Ирина
Дрожащих Наталия
Михалькова Елена
Пащенко Людмила

1. Зачем чат-боты нужны в образовании

Образование будущего – онлайн образование на учебных платформах в интернете.

Возможность выбора и комбинирования любого количества дисциплин любой тематики.

Индивидуальный подход.

Образовательная платформа тестирует абитуриента, предлагает курсы, записывает на них, высылает задания, проверяет их выполнение, оказывает поддержку (в т.ч. психологическую).

Чат-бот — это программа с искусственным интеллектом, которая имитирует человеческий диалог в формате “вопрос пользователя – ответ системы”. Взаимодействие происходит мгновенно. Чат-боты используют шаблоны ответов, ищут ответы на заданных интернет-сайтах.

Преимущества программ с искусственным интеллектом

1. Никогда не устают и не теряют терпения.
2. Сохраняют историю общения с учеником. Например, учащийся в любой момент может перечитать свои ответы.
3. Поддерживают мотивацию учащихся в процессе онлайн-обучения, подсказывая, что делать, развлекая, в то же время передавая знания.
4. Могут сделать процесс онлайн-обучения более продуктивным, предлагая персонализированные программы.

Учебные чат-боты в Telegram

@ucheba_bot бот, помогающий ориентироваться в многообразии учебных онлайн курсов

@ias16bot бот, имеющий знания по всем наукам, отвечает на вопросы, тестирует интеллектуальный уровень пользователей

@Wikipedia_voice_bot бот с функцией голосового поиска по «Википедии». Можно запрашивать необходимую информацию, не отвлекаясь от повседневных занятий

Диалоговые технологии активно внедряются в сферу образования.

Успешность этой коммуникации в образовательной сфере зависит от корректности, эмпатийности и сенситивности искусственного интеллекта.

2. Возникновение сенситивного контента

Сенситивный контент – языковые выражения, которые считаются оскорбительными, вводящими в заблуждение или спорными (оскорбления, обсценная и потенциально “обидная” лексика, угрозы, манипуляция, хейтизм, кибербуллинг, номинации насилия, оружия, азартных игр, наркотиков).

Возникновение сенситивного контента: кейсы

Случайные сочетания

1. Парейдолическая иллюзия осмысленного текста.
2. Случайная генерация осмысленной последовательности (теорема о бесконечных обезьянах).

Алгоритмическая предвзятость

1. Искажения в данных.
2. Структура алгоритма.
3. Использование программы не по назначению.

Действительно оскорбительные тексты

Программа, которая создана имитировать человеческое поведение, как и человек, может действительно оскорбить пользователя.

Что нужно учесть в разработке чат-ботов? На примере учащихся с ОВЗ

Физические недостатки

- лица с нарушениями зрения (слепые, слабовидящие);
- лица с нарушениями речи;
- лица с нарушениями опорно-двигательного аппарата (ДЦП)

Психологические недостатки

- лица с нарушениями эмоционально-волевой сферы;
- лица с множественными нарушениями (сочетание 2-х или 3-х нарушений).
- лица, страдающие психическими расстройствами

Речевые нарушения

- лица, имеющие нарушения в чтении (дислексия)
- лица, имеющие нарушения в письменной речи (дисграфия)

В классе преподаватели планируют занятия с учетом особенностей студентов. Мы разделили их на три сферы (группы особенностей).

1. Когнитивные способности и учебные способности (learning skills)
2. Социальные, эмоциональные и ментальные особенности
3. Коммуникация и взаимоотношения

Когнитивные способности и учебные способности (learning skills)

Эта сфера охватывает широкий спектр специальных образовательных потребностей студентов, которые отстают от своих сверстников в обучении и испытывают трудности в понимании учебного материала, в организации учебного процесса.

Социальные, эмоциональные и ментальные особенности

У студентов с нарушениями в этой сфере отмечается тревожность, депрессия, самокритикование, переедание или наоборот отказ от еды и т.д.

Коммуникация и взаимоотношения

Эта сфера учитывает особенности нарушения речи, языка и коммуникации, например, трудности понимания и использования языка.

3. Подготовка учебного контента для чат-бота: Как предусмотреть сенситивный контент

Отбор учебного материала

1. Адаптация существующих программ.
2. Генерация специального учебного контента.
3. Сокращение учебной программы до основных элементов.
4. Упрощение изучаемого лексического, грамматического материала.
5. Подбор упрощенных текстов.
6. Обновление программы обучения в зависимости от реакции/успешности обучаемых.

Предъявление учебного материала при интеракции чат-бота с пользователем-1

1. Предъявление абстрактного материала через конкретную лексику.
2. Употребление слов в прямом значении, избегание использования слов в переносном значении.
3. Включение мультимедийного контента для мультисенсорной стимуляции, эмоциональной поддержки.
4. Замедление темпа предъявления материала.

Предъявление учебного материала при интеракции чат-бота с пользователем-2

5. Многократное предъявление материала.
6. Отсутствие критики (только позитивный фидбэк), похвала.
7. Определение тем, где может возникнуть сенситивный контент. (В таких темах нужен особый подход к подаче материала.)
8. Предъявление материала с помощью четких инструкций.

Тестирование

1. Диагностика нарушений речевой, коммуникативной, эмоционально-волевой сферы у студентов.
2. Тестирование студентов: усвоение материала, впечатления/фидбэк о взаимодействии с ботом.
3. Тестирование бота: работа алгоритма в разных учебных ситуациях.

Над проектом работают

Дрожащих Наталия Владимировна

Доктор филологических наук, доцент

Белякова Ирина Евгеньевна

Кандидат филологических наук, доцент

Пащенко Людмила Петровна

Кандидат педагогических наук, доцент

Михалькова Елена Владимировна

Кандидат филологических наук, доцент, магистр информатики

Список литературы:

1. Астафурова К., Кирьянов Р. Сбербанк объяснил отправку кода с призывом «убивать евреев» ошибкой // РБК. 21 ноября 2019. URL: https://www.rbc.ru/society/21/11/2019/5dd668689a7947794acbbf1f
2. Довлатов Сергей. Соло на Ундервуде. В книге: Записные Книжки. New York: журн. «Слово — Word», 1990.
3. Boden M. A. Creativity and artificial intelligence. Artificial Intelligence 103 (1998), p. 347-356. URL: https://www.sciencedirect.com/science/article/pii/S0004370298000551
4. Crawford, K. Artificial Intelligence's White Guy Problem. The New-York Times. 25 June 2016. URL: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html
5. Dai W., Yu T., Liu Z., Fung P. Kungfupanda at SemEval-2020 Task 12: BERT-Based Multi-Task Learning for Offensive Language Detection. arXiv:2004.13432 - April 2020 // https://arxiv.org/abs/arXiv:2004.134327
6. Introna, L.D. Maintaining the reversibility of foldings: Making the ethics (politics) of information technology visible. Ethics Inf Technol 9, 11–25 (2007). https://doi.org/10.1007/s10676-006-9133-z
7. Kahlbaum, Karl Ludwig. Die Sinnesdelierien. In: Allgemeine Zeitschrift für Psychiatrie und psychisch-gerichtliche Medicin 1866 ; 23 : 1-86.
8. Lee, Peter. Learning from Tay’s introduction. Mar 25, 2016. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/
9. Modha S., Majumder P., Patel D. DA-LD-Hildesheim at SemEval-2019 Task 6: Tracking Offensive Content with Deep Learning using Shallow Representation // Proceedings of the 13th International Workshop on Semantic Evaluation. 2019. Pp. 577-581. doi.org/10.18653/v1/s19-2103

10. Petrenko M., Folk C., Hempelmann C. Automated Ontologized Oppositeness. In: Book of Abstracts. 30th ISHS Conference Humour: Positively (?) Transforming Tallinn University, Tallinn, Estonia. 25-29 June 2018. P. 99.
11. Seuss H., Dankerl P., Ihle M., Grandjean A., Hammon R., Kaestle N., Fasching P.A., Maier C., Christoph J., Sedlmayr M., Uder M., Cavallaro A., Hammon M. Semi-automated De-identification of German Content Sensitive Reports for Big Data Analytics. Fortschr Röntgenstr. 2017. Vol. 189. Pp. 661–671.
12. Sigurbergsson G.I., Derczynski L. Offensive Language and Hate Speech Detection for Danish. arXiv:1908.04531 - August 2019 // https://arxiv.org/abs/arXiv:1908.04531.
13. Vyshnav M.T., Sachin Kumar S., Soman K.P. Offensive Language Detection: A Comparative Analysis. arXiv:2001.03131 - January 2020 // https://arxiv.org/abs/arXiv:2001.03131.
14. Whittaker, Zack. OpenAI built a text generator so good, it's considered too dangerous to release. February 17, 2019. https://techcrunch.com/2019/02/17/openai-text-generator-dangerous/
15. Xu Y., Jiao Y., Chen S., Li Y. Research on Detection Method of Unhealthy Message in Social Network. In: Sun X., Pan Z., Bertino E. (Eds.). Artificial Intelligence and Security. ICAIS 2019. Lecture Notes in Computer Science. 2019. Vol. 11632. Springer, Cham. Doi: 10.1007/978-3-030-24274-9_45.
16. Yan X., Zhao X., Yang G. A Tibetian and Uygur Sensitive Word Tracking System // Z. Zhong (Ed.). Proceedings of the International Conference on Information Engineering and Applications (IEA) 2012. Lecture Notes in Electrical Engineering 219. Doi: 10.1007/978-1-4471-4853-1_40. Springer-Verlag. London. 2013. Pp. 307-
17. Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N., Kumar R. Predicting the Type and Target of Offensive Posts in Social Media. arXiv:1902.09666 - February 2019 // https://arxiv.org/abs/arXiv:1902.09666