Моделирование и анализ данных
2023. Том 13. № 1. С. 19–24
DOI: https://doi.org/10.17759/mda.2023130102
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2023. Vol. 13, no. 1, pp. 19–24
DOI: https://doi.org/10.17759/mda.2023130102
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 81.332

Система детоксификации текста в диалоговых переписках

Суворов М.Д.*
Московский авиационный институт
(национальный исследовательский университет) (МАИ)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-8376-0448
e-mail: msuvorov7@gmail.com

Виноградов В.И.**
Московский авиационный институт
(национальный исследовательский университет) (МАИ)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0003-3773-9653
e-mail: vvinogradov@inbox.ru

Работа направлена на повышение культурного уровня переписки в диалоговых системах. Ключевая особенность работы – ориентированность на использование в режиме реального времени и обеспечение устойчивой детоксификации с учетом специфики диалогового общения (опечатки, шумовые символы, транслит и прочее). Предлагаемое решение использует нейросетевой подход и программную обработку для получения эмбеддингов токенов и последующим решением задачи классификации. В отличие от традиционных фильтров сообщений поставлена задача сохранить смысл исходного текста, очистив его от токсичного контента. Работоспособность системы можно проверить на базе мессенджера «Телеграм», в котором модель представлена в виде бота. Сама система развернута на базе Serverless технологии от облачного провайдера, что позволяет ей адаптироваться к пиковым нагрузкам и быть при этом простой в поддержке.

Ключевые слова: детоксификация текста, нейронные сети, бессерверные технологии.

Для цитаты:
Суворов М.Д., Виноградов В.И. Система для детоксификации текста в диалоговых переписках // Моделирование и анализ данных. 2023. Том 13. № 1. С. 19–24. DOI: https://doi.org/10.17759/mda.2023130102

Суворов Максим Дмитриевич, магистрант, Московский авиационный институт (национальный исследовательский университет) (МАИ), ORCID: https://orcid.org/0000-0002-8376-0448, e-mail: msuvorov7@gmail.com

Виноградов Владимир Иванович, кандидат физико-математических наук, доцент кафедры «Математическая кибернетика», Московский авиационный институт (национальный исследовательский университет) (МАИ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0003-3773-9653, e-mail: vvinogradov@inbox.ru

1. ВВЕДЕНИЕ

Основной целью данного проекта является повышение культурного уровня текстовых писем в диалоговых переписках. Зачастую, люди не соблюдают моральные и этические нормы в цифровом пространстве, а иногда сами рвутся оскорбить собеседника в переписке. Для фильтрации подобного контента предназначены классификаторы токсичных сообщений, которые удаляют обнаруженные подозрительные сообщения. Такие системы имеют важный недостаток в том, что обнаруженное сообщение удаляется полностью, даже если в нем было лишь одно токсичное слово. Данный подход является слишком радикальным, поскольку при его неправильной классификации и отнесения к токсичному классу сообщение удаляется полностью и информация, передаваемая в нем, теряется.

Для использования разрабатываемой модели в диалоговых системах с потенциально высоким трафиком система с самого начала задумывалась как надежная, масштабируемая, поддерживаемая и адаптивная. Для выполнения поставленных требований необходимо предусмотреть мониторинг и использовать соответствующие технологии его реализующие. Чтобы упростить разработку и больше сосредоточиться на реализуемом продукте, было принято решение использовать услуги облачного провайдера для развертывания системы на его платформе.

Все вышесказанное можно перевести в задачу тегирования или в задачу классификации токенов сообщения. Но в отличии от обычной задачи классификации сообщения целиком, создаваемая система должна восстановить из полученных токенов и предсказаний исходное сообщение и скрыть только оскорбительные слова. А слово может быть представлено несколькими токенами, и один токен может включать несколько слов.

2. ОБУЧАЮЩИЕ ДАННЫЕ

Используемым в настоящей работе датасетом является набор русскоязычных твитов [1]. Разметку было решено получить полуавтоматическим способом через более слабые модели, использующие методы ближайшего соседа и логистической регрессии. Первоначально был составлен словарь из токсичных слов, затем на малой подвыборке были получены метки слов в зависимости от нахождения в словаре. На полученном сэмпле обучались алгоритмы целевой модели, и затем они применялись к неразмеченным данным. Итеративно процесс повторялся несколько раз, расширяя словарь и заново переобучая базовые модели.

Поскольку данные, поступающие на вход системы, будут направлены на ее поломку, требуется всегда выдавать эмбеддинги для новых слов. Для решения этой проблемы

была выбрана модель, основанная на BPE-кодировании, основанная на представлении эмбеддинга слова как суммы эмбеддингов n-грамм его букв. Подобной моделью можно назвать FastText, который для любого слова может выдать семантически близкое представление.

3. МОДЕЛЬ КЛАССИФИКАТОРА

В качестве классификатора были опробованы архитектуры со сверточными, линейными, рекуррентными слоями. Общий вывод, который удалось сделать – чем сложнее архитектура, тем быстрее сеть переобучается и тем больше хороших слов попадают в «токсичный» класс. Например, сложно было отучить модель воспринимать такие слова как «копать», «мазь», «да» как нетоксичные. Архитектура такой системы представлена на рис. 1.

На вход подается строка с сообщением, которая разбивается на токены в блоке токенизатора. Затем каждый токен попадает в обработчик, который представляет собой набор правил по очистке и фильтрации ненужных символов, например, удаление смайликов, незначимых спецсимволов, дублирований. Далее каждый токен получает свое представление в FastText блоке, чтобы потом быть обработанным классификатором. Сам классификатор представляет собой нейросеть, состоящую из слоев GRU, функции активации ReLU и линейного слоя на выходе. Получив показатели принадлежности положительному классу, из токенов собирается цензурированное сообщение в блоке детокенизатора.

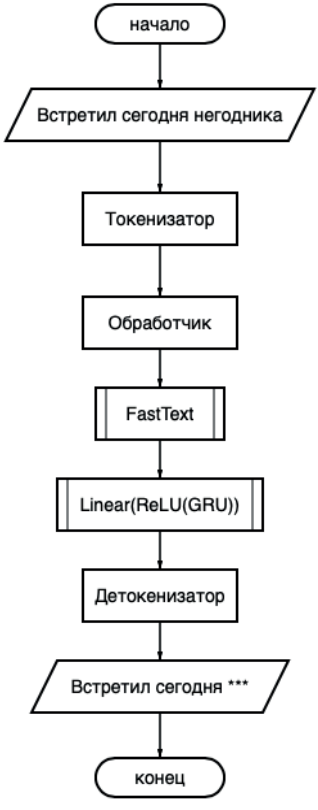

Рис. 1. Архитектура классификатора

Для оценки качества классификации выбрана метрика balanced accuracy из-за сильного дисбаланса классов нетоксичных и токсичных токенов. Предпочтение отдано именно этой метрике, поскольку ее значения коррелировали с наглядными результатами работы модели.

4. РАЗВЕРТЫВАНИЕ МОДЕЛИ

В качестве облачного провайдера выбран «Яндекс» из-за выгодных квот на предлагаемой им serverless технологии. Аренда виртуальной машины и кластера с базой данных – это недешевое удовольствие для системы на старте ее жизни. Serverless технология – это достаточно молодая область, но крайне

перспективная. Ее особенность в том, чтобы предоставить пользователю услугу как сервис. Пользователь не занимается развертыванием системы, ее настройкой и поддержкой. Все это берет на себя провайдер. Пользователь же платит только за непосредственно используемые вычислительные мощности, характеризуемые, например, числом запросов к базе данных, числом вызовов функции, объемом используемого хранилища данных. В отличии от виртуальных хостов такие сервисы активны лишь во время обращения к ним. Все остальное время их ресурсы находятся в общем пуле ресурсов и масштабируются провайдером по мере необходимости.

Для реализации мониторинга за сдвигом распределения данных и таргета нужно где-то эти данные хранить. Как сказано выше, разворачивать полноценный кластер с БД – это дорого. Но нашлась альтернатива в виде serverless YDB, которая предлагала широкие возможности по масштабированию системы и экономила средства пользователя. Простые операции по вставке записей позволили хранить поступающие в модель данные и не терять в скорости ответа. Для аналитики и долгосрочного хранения можно организовать более «прохладное» хранилище, куда переливаются данные не чаще одного раза в час. В построенной системе такой перезаливкой занимается инструмент Data Transfer от Яндекса.

Таким образом, поступающие данные накапливаются в YDB (но нужно не забывать, что установлен лимит по TTL записей в БД – Time ti live) по минимальной стоимости, основной кластер может быть выключен и «поднят» только по мере необходимости перезаливки или непосредственной работы аналитиков.

5. ДОСТОИНСТВА И НЕДОСТАТКИ

Готовая система тестировалась не только на отложенном наборе данных, но и при реальном диалоговом общении людей. Так удалось выявить некоторые неточности, связанные с ложным определением токсичных слов. Явный пример: «мы уехали отдыхать в город Пиза». Название города расценивается моделью как нечто, подлежащее скрытию. Причины данной ошибки можно назвать две: отсутствие в обучающей выборке употребительных примеров со словом «Пиза» и некачественный эмбеддинг от модели FastText. Первую причину можно устранить, добавив подобные предложения в обучающие данные, чтобы сеть научилась лучше улавливать семантику спорных слов (таких, как «хрен» или «очко», например). Вторая причина связана с применением квантизации, дабы ускорить процесс инференса и уложиться в лимиты провайдера.

Противоположным примером можно назвать сценарий использования редко употребительных или устаревших нецензурных слов, когда системой пропускаются токсичные слова. Такие примеры нетривиально придумать во время диалога, но их можно встретить в специальных словарях для телевидения, которые используют для цензурирования.

Не остались без проверки и наиболее распространенные оскорбления. Обычно, это образованные слова от нескольких типичных матерных корней. Успешно прошли

через систему такие примеры как «прихлебатель – это угодливый человек», «из чего сделана похлебка?», «постарайся это не употреблять». Модель достаточно устойчива, несмотря на опечатки и шумовые символы, но существуют разнообразные способы, которые скрывают от реализованной модели истинный смысл. К этим способам можно отнести использование верхнего регистра, транслит, разбиение букв слова пробельными символами или же наоборот, склейка слов без пробелов. Обучение модели быть устойчивой к подобным сценариям является нетривиальной задачей, но для этого можно начать с добавления некоторых простых правил-обработчиков.

6. ВЫВОД

Результатом работы является программное обеспечение, способное очищать текстовые сообщения от токсичного контента, развернутое на платформе мессенджера «Telegram» в виде бота @toxic_segmenter_bot. Данная площадка выбрана как популярное место общения разных групп общества.

В процессе разработки возникали сложные вопросы, касающиеся дизайна системы, проектных ограничений, развертывания системы. И по мере ответа на каждый из вопросов сложилось понимание того, что человек все равно сможет найти способы обойти систему, будь то картинки, аудио записи, смайлики или монолитный текст. Гонка за идеалом может быть бесконечной. Альтернативой является предложить людям осознанно отказаться от токсичного контента в своей жизни.

Литература

1. *Рубцова Ю.В.* Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора // Инженерия знаний и технологии семантического веба. – 2012. – Т. 1. – С. 109–116.

Text Detoxification System in Dialogue Conversations

Maxim D. Suvorov*
Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia
ORCID: https://orcid.org/0000-0002-8376-0448
e-mail: msuvorov7@gmail.com

Vladimir I. Vinogradov**
Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia
ORCID: https://orcid.org/0000-0003-3773-9653
e-mail: vvinogradov@inbox.ru

The work is aimed at improving the cultural level of correspondence in dialog systems. The key feature of the work is its focus on real–time use and ensuring sustainable detoxification, taking into account the specifics of dialog communication (typos, noise symbols, transliteration, etc.). The solution offers the use of a neural network approach and software processing to obtain embeds of tokens and the subsequent solution of the classification problem. Unlike traditional message filters, the task is to preserve the meaning of the source text by clearing it of toxic content. The operability of the system can be checked on the basis of the Telegram messenger, in which the model is presented in the form of a bot. The system itself is deployed on the basis of Serverless technology from a cloud provider, which allows it to adapt to peak loads and at the same time be easy to maintain.

Keywords: detoxification of text, neural networks, serverless.

For citation:
Suvorov M.D., Vinogradov V.I. Text Detoxification System in Dialogue Conversations. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2023. Vol. 13, no. 1, pp. 19–24. DOI: 10.17759/mda.2023130102 (In Russ., abstr. in Engl.).

***Maxim D. Suvorov,** student, Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia, ORCID: https://orcid.org/0000-0002-8376-0448, e-mail: msuvorov7@gmail.com

****Vladimir I. Vinogradov,** PhD, Associate Professor of the Department of Mathematical Cybernetics, Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia, ORCID: https://orcid.org/0000-0003-3773-9653, e-mail: vvinogradov@inbox.ru

References

1. Rubtsova Yu.V. Automatic construction and analysis of the corpus of short texts (microblogging posts) for the task of developing and training a tone classifier // Knowledge engineering and semantic web technologies. – 2012. – Vol. 1. – pp. 109–116.

Получена 17.01.2023
Принята в печать 17.02.2023

Received 17.01.2023
Accepted 17.02.2023