Разработка модели автоматизированной системы разметки набора данных для обнаружения деструктивных тем в Интернет-обсуждениях

Е.Н. Павличева[1], Н.С. Чикунов[1]

[1]*Московский государственный технологический университет «СТАНКИН», Вадковский переулок 3а, Москва, Россия, 127055*

Аннотация

В статье проводится анализ актуальности выявления деструктивных сообщений в Интернете и необходимость использование данных для обучения моделей на русском языке. Предлагается модель для автоматизированного сбора и разметки текста. Рассматривается алгоритм обработки данных для построения классификатора

Ключевые слова

Распознавание деструктивных сообщений, сбор и разметка данных, обработка текстов, социальные сети

1. Введение

Одной из актуальных проблем современного общества является выявление деструктивных сообщений и тем в Интернете, обусловленное увеличением доли общественных коммуникаций в цифровом пространстве. Автоматическая классификация деструктивных обсуждений является популярным направлением исследований в настоящее время, которая требует применение все новых методов искусственного интеллекта. [1, 2] Компания Jigsaw ни раз предоставляла свои размеченные данные для проводимых соревнований на платформе Kaggle, однако русскоязычных открытых наборов для исследований этой темы нет [3].

2. Основная часть

Онлайн-комментарии общедоступны, и с каждым днем число выборок данных увеличивается, однако без маркировки этих данных их можно использовать только для обучения системы без учителя, например, кластеризации или уменьшения размерности, а контролируемые подходы к обучению требуют размеченных данных [4].

В довольно дорогостоящем ручном процессе обучения системы специалисты по разметке данных проверяют каждый комментарий на наличие деструктивности в них. Из-за присущей естественному языку неоднозначности результат классификации может различаться у разных специалистов. Кроме того, комментарий может быть воспринят как деструктивный в одном контексте, но не деструктивный в другом. Различные руководства по разметке, низкое согласие между специалистами и общее низкое качество аннотаций являются одной из текущих проблем исследования в области классификации деструктивных комментариев [5].

Один из вариантов решения проблемы – использование обученного классификатора для разметки большей части требуемого набора данных, который позволяет сократить количество человеческих ресурсов и материальных средств, а дальнейший анализ размеченных данных поможет контролировать точность автоматической маркировки [6].

Авторами предлагается использование модели системы автоматизированной разметки набора данных для исследуемых интернет-обсуждений с целью дальнейшего построения классификатора. Для системы требуются два неразмеченных набора данных (собранных, например, с помощью парсинга):

1. Содержащие номера и полный текст темы обсуждений.

2. Содержащие порядковые номера и полный текст комментарии, а также номера тем обсуждения, к которым относятся данные обсуждения.

Второй набор данных позволит обучить систему для определения степени деструктивности текста, которая в свою очередь разметит оставшиеся комментарии из набора после ручной разметки.

Для разработки системы определения деструктивности в тексте, исходные комментарии необходимо обработать в соответствии со следующим алгоритмом, представленном на рисунке.

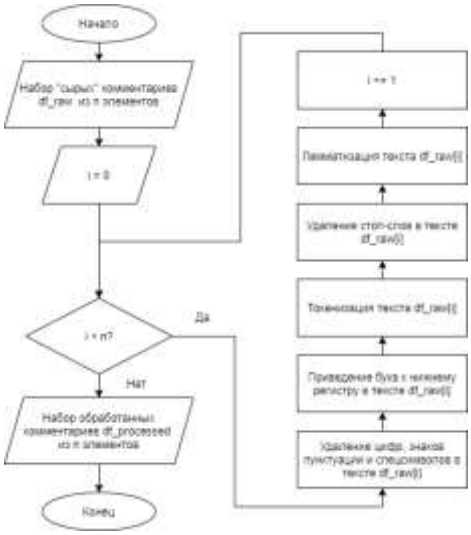

Рисунок 1: Алгоритм обработки комментариев

На размеченных вручную комментариев осуществляется обучение классификатора на основе методов word2vec для определения наличия деструктивности в остальных неразмеченных текстов в наборе данных.

Таким образом, на этом шаге системой решается общая задача определения наличия деструктивности в тексте, однако она не позволяет определять деструктивность тем обсуждений для преждевременной реакции модераторов до появления первых комментариев.

Результаты определения деструктивности комментариев необходимо связать с первым набором, содержащим темы обсуждений, однако для просмотра общего результата исследования, их необходимо представить в виде ключевых слов на основе тематического моделирования, предварительно обработав текст согласно описанному ранее алгоритму.

3. Заключение

Полученные наборы данных могут использоваться для разных прикладных задач по урегулированию общения на различных платформах, но он не позволяет отразить причины, почему конкретные темы или комментарии являются деструктивными – для этого следует применить модель для более узких классов, чтобы отслеживать более специфичные для платформы виды деструктивности (например, разрешена ненормативная лексика, но не приветствуются угрозы).

4. Литература

[1] Challenges for Toxic Comment Classification: An In-Depth Error Analysis [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1809.07572 (дата обращения 05.01.2021).

[2] Чикунов, Н.С. Разработка и исследование системы интеллектуального анализа текста на наличие «Токсичности» / Н.С. Чикунов, Е.Н. Павличева // XIV Международная отраслевая научно-техническая конференция «Технологии информационного общества». – 2020. – С. 173-175.

[3] Algorithms and insults: Scaling up our understanding of harassment on Wikipedia [Электронный ресурс]. – Режим доступа: https://diff.wikimedia.org/2017/02/07/scaling-understanding-of-harassment/ (дата обращения 05.01.2021).
[4] Yahoo Has a Tool that Can Catch Online Abuse Surprisingly Well [Электронный ресурс]. – Режим доступа: https://www.technologyreview.com/2016/07/26/158644/yahoo-has-a-tool-that-can-catch-online-abuse-surprisingly-well/ (дата обращения 05.01.2021).
[5] Risch, J. Toxic Comment Detection in Online Discussions / J. Risch, R. Krestel // Deep Learning-Based Approaches for Sentiment Analysis. – 2020. – P. 85-109.
[6] Чернышева, Е.Н. Формирование цифровой компетентности в сетевом сообществе / Е.Н. Чернышева, Е.Н. Павличева, Н.С. Чикунов // XXI век: итоги прошлого и проблемы настоящего. – 2020. – Т. 9, № 4(52). – С. 62-67.